arXiv:2603.00041v1 [cs.LG] 9 Feb 2026

计量经济学与因果结构学习在时间序列政策决策中的比较：来自英国新冠肺炎政策的证据

Bruno Petrungaro[1*]和 Anthony C. Constantinou[1]1*贝叶斯人工智能研究实验室、MInDS研究组、伦敦玛丽女王大学电子工程与计算机科学学院，Mile End Road，伦敦，E1 4NS，英国。

通讯作者。电子邮箱：b.petrungaro@qmul.ac.uk；贡献作者：a.constantinou@qmul.ac.uk；

摘要

因果机器学习（ML）能够恢复图结构，从而为我们提供潜在因果关系的信息。目前大多数进展都集中在没有明确时间顺序的横截面数据上，而从时间序列数据中恢复因果结构仍是因果机器学习领域持续研究的课题。除了传统的因果机器学习方法（如基于评分和基于约束的算法）之外，本研究还评估了一些计量经济学方法；有人认为，这些方法能够从时间序列数据中恢复因果结构。它们已在经济学、生物学和其他领域用于此类目的。这些方法之所以得到应用，可以归因于多年来计量经济学领域对因果关系，尤其是时间序列的高度关注。这也为比较计量经济学方法与传统因果机器学习算法的因果发现表现提供了可能。我们旨在了解计量经济学能否为因果机器学习提供借鉴，并提供代码，将这些计量经济学方法的结果转换为应用最广泛的贝叶斯网络 R 语言库 bnlearn 可用的格式。我们以英国 COVID-19 的真实案例为例，探究这些算法在支持政策决策方面的优势与挑战。我们从图结构、模型维度以及恢复因果效应的能力等方面评估了四种计量经济学方法，并将结果与十一种因果机器学习算法的结果进行比较。主要发现之一是，基于评分的因果机器学习方法通常能够恢复更多可识别效应，但通常也会生成更稠密的图。基于收缩思想的方法能够揭示合理的干预目标，但没有任何一类方法在所有指标上都占据优势。总体而言，评估结果

表明，各类方法各有所长。计量经济学方法为时间结构提供了明确的规则，而因果机器学习算法则通过探索更大的图结构空间进行更广泛的发现；这往往会生成更稠密的图，从而捕捉到更多可识别的因果关系。

关键词：有向无环图、结构学习、时间序列、因果推断、模型平均、COVID-19

1 引言

聚焦关联可能会误导政策设计与评估；我们需要确定一个变量的变化是否会导致另一个变量发生变化。传统机器学习（ML）模型并不适合回答这个问题，因为它们关注关联，无法模拟假设行动的影响以辅助决策。我们必须依赖这些模型无法捕捉的重要假设或因果假设。

因果机器学习（因果机器学习）关注无监督学习算法，旨在从观测性数据和干预性数据中还原因果关系。在本研究中，因果机器学习特指因果结构学习算法，其重点是学习现有变量之间的依赖结构；若要将该结构解释为因果模型，还需要额外的因果假设。这可以看作是这些变量之间因果关系的图谱。结构学习算法传统上分为两类：基于评分的算法和基于约束的算法。基于评分的学习侧重于在可能的结构空间中搜索，并优化一个反映学习所得结构与数据拟合程度的评分；基于约束的学习则使用条件独立性检验来推断因果结构。结合这两类学习方法的混合算法通常被视为结构学习的第三类。

过去几十年来，因果结构学习算法在方法上取得了显著进步。然而，与传统的关联式机器学习方法相比，这些算法在现实世界中的影响仍然较为有限。这是因为，从观察性数据中发现因果关系仍然极其困难；而且，与关联式机器学习不同，因果发现中的微小错误可能会在因果推断中造成巨大的负面影响。这或许可以解释，为何政策建议领域的文献仍主要来自计量经济学。也就是说，尽管其他领域无疑为最优决策的发展作出了贡献，经济学仍是对这些主题研究最深入的学科之一。

[1] 是计量经济学中一场影响广泛的讨论的绝佳例子，该讨论偏好采用类似随机对照试验（RCT）的研究设计。不过， [1]也提到了机器学习领域的发展及其潜力。工具变量、回归不连续设计和双重差分方法仍是计量经济学家进行因果推断的首选工具。这些方法催生了令人振奋的研究，例如

[2], 采用回归不连续设计，估算移民合法化对意大利移民犯罪率的因果效应，得出的结论是，合法化降低了获得合法身份的移民的犯罪率。然而，与因果机器学习一样，这些方法也存在局限性。 [1] 概述了与识别策略薄弱或存在缺陷相关的问题，以及这类分析难以推广至其他情境的问题。

因果机器学习在时间序列数据领域的发展有限，这为评估这些方法在此类条件下的现实场景中对政策决策的支持效果，并将其与常用于此目的的计量经济学方法进行比较，提供了宝贵机会。本研究将探讨时间序列数据中的因果关系，这为此类分析增加了一层复杂度。在大数据时代，海量数据不断随时间生成、演变，并对外部冲击（例如政策变化）作出响应。本文选取英国的 COVID-19 作为案例，详情见第 3 节。将此作为我们的研究目标，意味着我们将评估以下方面：

- 这些计量经济学方法从观察性时间序列数据中恢复因果关系的能力。这涉及了解这些方法在因果机器学习最基本的目标——即从观察性数据中恢复因果关系——方面的表现。
- 由这些因果关系所隐含的恢复模型结构，识别并正确估计因果效应方向的能力；也就是说，确定效应是增加还是减少。

2 背景

与计量经济学相比，因果机器学习在从时间序列数据中恢复因果关系方面的发展一直较为有限。我们能够识别出以下与时间序列数据因果机器学习相关的研究：

- [3] 介绍了两种在线因果结构学习算法：在线快速因果推断（OFCI）和快速在线快速因果推断（FOFCI）。OFCI 是快速因果推断（FCI）的在线版本，经过修改以处理潜变量以及动态、随时间变化的因果结构。其工作方式是在新数据点到来时修正相关性，并在数据表明发生显著变化时重新学习结构。FOFCI 是 OFCI 的改进版本，旨在通过利用从先前模型中学到的因果关系来降低学习成本并加快学习速度，同时保持模型拟合。
- [4]介绍了 DOCL，一种专注于现实世界环境的因果结构学习算法，在这些环境中因果结构可能发生不可预测的变化。DOCL 实时处理流式数据，并适应从序列或有序数据中学到的因果结构及底层概率关系的变化。
- [5]提出了局部平稳结构跟踪（LoSST）结构学习算法，该算法假设数据生成过程具有局部平稳性。LoSST 会实时动态跟踪并适应因果结构及其关系的变化。

尽管这些研究与本文相关，但它们旨在识别因果结构的变化。相比之下，我们关注的是分布偏移：图保持不变，但季节性、政策以及行为驱动的变化等输入会改变观测到的分布，却不意味着出现了新的因果联系。

计量经济学提供了丰富的时间序列方法，可用于解决这一问题。第 4 节将详细说明这些方法如何帮助恢复因果结构。不过，显然，本文介绍的时间序列分析方法与传统因果机器学习方法的运作方式相似。这是因为，系数不显著意味着，在控制模型中的其他变量后，没有证据表明预测变量与目标变量之间存在依赖关系。这一结果与预测变量和目标变量条件独立的情况相符。这一理念与传统的因果发现方法类似，后者会检验变量之间的条件独立性。然而，如果遗漏了重要的混杂因素，基于回归的推断有时会反映虚假的关联。另一种计量经济学方法是搜索网络模块，本质上就是识别彼此相关的节点群组。恢复因果结构的方法所做的正是这件事。

学习到结构后，可以将模型转换为动态贝叶斯网络（动态贝叶斯网络）。动态贝叶斯网络是一种概率图模型，专门用于表示变量及其在不同时间点之间的依赖关系。与标准贝叶斯网络类似，动态贝叶斯网络通过有向无环图将变量表示为节点，并将变量间的依赖关系表示为有向弧。然而，与每个变量仅由单个节点表示的标准贝叶斯网络不同，动态贝叶斯网络使用多个节点表示每个变量，每个节点对应同一变量在不同时间点的取值。在考虑所有其他相关的过去变量后，如果连续时间点上的变量与过去变量不独立，则在动态贝叶斯网络中会在这些变量之间绘制有向弧。

就参数化而言，离散贝叶斯网络通常使用条件概率表来表示依赖关系，而连续贝叶斯网络则采用条件概率分布。在某些假设下，贝叶斯网络或动态贝叶斯网络可被赋予因果解释，从而将有向无环图转化为因果贝叶斯网络。因果贝叶斯网络对本研究至关重要，因为它们使我们能够对假设性干预的效应进行建模并估计这些效应，从而实现不确定性下的最优决策。

尽管贝叶斯网络具有实用价值，但大多数应用研究都针对横截面数据（在某一特定时间点从多个实体收集的数据），或者假设数据是由某个过程的多个独立实现生成的（将重复横截面数据合并后视为一个横截面数据集）。这是因为贝叶斯网络结构学习算法通常并非为捕捉系统随时间的演变、构成该系统的变量，或这些变量之间的相互关系如何随时间变化而设计的。

3 COVID-19 案例研究数据

为评估这些方法在支持政策决策方面的实用性，我们选取了备受争议的 COVID-19 应对政策。由 SARS-CoV-2 病毒引发的 COVID-19 大流行，在全球范围内带来了重大的公共卫生和政策挑战。

例如，疫苗通常被视为控制病毒性疾病传播的主要策略，因此，开发出有效疫苗后，人们也对 COVID-19 抱有同样的期待。许多人认为，在疫苗问世之前，与控制这种疾病传播有关的其他措施都只是临时之举。这或许是因为受[6], 启发的传统传染病流行模型假定，人口中的任何成员都可能感染其他任何人，或被其他任何人感染。许多模型忽略了人群内部实际的相互作用以及疾病的传播方式。然而，人们很可能一直抱有这种期待，是因为实施社交距离措施的经济成本很高。

由于传统模型没有考虑人群内部的实际相互作用，而且社交距离措施的成本很高，因此问题随之而来：哪些干预最能有效降低感染率？COVID-19 不会是最后一次大流行，因此这个关键问题是：由于大流行初期通常无法及时获得疫苗或治疗方法，我们能否利用关于限制人群内部互动方式的知识，更快地控制大流行？在因果推断框架中通过模拟假设性干预来建模这些影响， [7] 或许有助于支持未来大流行期间的决策。

我们使用的是常规收集的公开汇总数据，这些数据在英国大流行期间被广泛使用。有关数据收集方式的详细说明，请参阅 [8]。本研究收集的其他相关数据来自 [8]中所引述的相同来源。这些来源包括 [9]和 [10]。本研究使用的数据集包含 46 列和 866 行，其中每列对应一个变量，每行数据对应每日的一个结果。数据的第一列为时间指标“日期”，其时间范围为 2020年1月30日至2022年6月13日。该数据集包含连续变量和分类变量，也含有缺失数据值。变量说明见表1。

表 1：COVID-19 数据集

变量	**描述**
Date	记录观测值的日期。
超额死亡率（EM）	报告死亡人数与预测死亡人数之间的百分比差异。
学校	学校是开放、部分开放还是关闭。
口罩（口罩）	大流行期间的口罩令：可选、强制或无强制令。
封锁严格程度（封锁严重程度）	封锁严格程度：严格、中等、宽松或仅采取有限措施/社交距离措施。
主要新冠变异株（变体）	主要 COVID-19 变异株。

变量	描述（续）
航班 7日 移动平均值	航班的7日移动平均值。
OpenTable 餐厅（餐厅）	伦敦餐厅预订的OpenTable指数。
谷歌 居家办公（居家办公）	谷歌大伦敦居家办公指数。
谷歌工作场所（工作场所）	谷歌大伦敦工作场所活动指数。
苹果步行（步行）	苹果在伦敦的步行活动指数。
谷歌公园（公园）	谷歌在大伦敦的公园到访指数。
谷歌零售与休闲（零售与休闲）	大伦敦零售与休闲的谷歌指数。
谷歌杂货 & 药-房（杂货与药-房）	大伦敦杂货与药房的谷歌指数。
谷歌 公共交通 站点（公共交通）	交通站点活动的谷歌指数。
伦敦交通局地铁（地铁）	伦敦交通局地铁活动指数。
伦敦交通局公交车（公交车）	伦敦交通局公交车活动指数。
Citymapper 出行（出行）	Citymapper 出行活动指数。
季节	冬季、秋季、夏季和春季。
PCR检测	日期的检测数量。
PCR 检测能力	日期当日的能力水平。
抗体检测	日期当日的检验数量。
抗体检测能力	日期当日的能力水平。
支柱1能力	英国国家医疗服务体系/英国卫生安全局在该日期的能力。
支柱2能力	英国政府在该日期的能力。
支柱3能力	该日期的抗体检测能力。
支柱4能力	日期的监测检测能力。
支柱1检测	英国国家医疗服务体系/英国卫生安全局在该日期的检测。
支柱2检测	英国政府在该日期的检测。
支柱3检测	日期的抗体检测。
支柱4检测	日期的监测检测。
所有 4 个支柱的检测	日期的检测总量。
新增病例	日期当天COVID-19检测呈阳性的人数。
新增感染	日期当天新增感染人数。
再次感染	日期当天新增再感染人数。
住院人数	因 COVID-19 入院的患者。
住院患者	住院的 COVID-19 患者。
COVID-19 患者 in MVBs	使用机械通气床位的 COVID-19 患者（MVBs）。
疫苗接种（总计）	当日接种的疫苗总剂数。
疫苗接种（第一剂）	当日接种的第一剂疫苗。

变量	说明（续）
疫苗接种（第二剂）	日期接种的第二剂疫苗。
疫苗接种（第三剂）	日期接种的第三剂疫苗。
第一剂接种率	报告的第一剂接种率。
第二剂接种率	报告的第二剂接种率。
第三剂接种率	报告的第三剂接种率。
COVID-19死亡人数证书上	按死亡日期统计的死亡证明上注明感染COVID-19的每日死亡人数。

3.1 数据缺失与插补

数据集共包含46个变量，其中40个变量至少有一个缺失值，约占所有变量的86.96%。总体而言，共有5,667个缺失值，占数据值的14.23%。表2总结了各变量的缺失数据情况。主要信息包括：

- 疫苗接种数据（尤其是第三剂变量）的缺失值比例最高，约为71%至74%。这种情况并不意外，因为疫苗接种是在大流行后期才开始的。不过，数据本身似乎有所延迟。英国的疫苗接种始于2020年底，但数据却从2021年底才开始。

- 出行指数（如Apple步行出行指数和Citymapper行程）的缺失百分比也相对较高（分别约为28%和37%）。

- 与COVID检测和监测相关的变量（PCR和抗体检测）的缺失百分比处于中等水平，通常在7%至14%之间。
- 关键的大流行指标，如新增病例、住院人数和COVID-19死亡人数，缺失百分比相对较低。

- 缺失数据的比例随时间增加，较晚的观测值显示出更高的缺失率。

表2：各变量缺失数据汇总

变量	缺失数量	缺失百分比 (%)
Date	0	0.00
超额死亡率	8	0.92
学校	0	0.00
口罩	0	0.00
封锁严格程度	0	0.00
主要新冠变异株	5	0.58
航班7日移动平均值	0	0.00
OpenTable伦敦餐厅预订指数	144	16.63
Google大伦敦居家办公流动性指数	39	4.50
Google大伦敦工作场所流动性指数	39	4.50
苹果伦敦步行移动性指数	247	28.52

变量	缺失计数	缺失百分比(%)
Google大伦敦公园出行指数	39	4.50
Google大伦敦零售与娱乐出行指数	39	4.50
Google大伦敦杂货店与药店出行指数	39	4.50
Google公共交通站点出行指数	49	5.66
伦敦交通局地铁出行指数	54	6.24
伦敦交通局公交出行指数	54	6.24
Citymapper 出行指数	319	36.84
季节	0	0.00
PCR检测	107	12.36
PCR 检测能力	90	10.39
抗体检测	121	13.97
抗体检测能力	70	8.08
支柱一：英国国家医疗服务体系和英国卫生安全局能力	70	8.08
支柱二：英国政府能力	69	7.97
支柱三：抗体检测能力	120	13.86
支柱四：监测能力	70	8.08
支柱一：英国国家医疗服务体系和英国卫生安全局检测	86	9.93
支柱二：英国政府检测	86	9.93
支柱 3 抗体检测	122	14.09
支柱 4 监测检测	62	7.16
所有 4 个支柱的检测	86	9.93
新增病例	25	2.89
新增感染	25	2.89
再次感染	132	15.24
住院人数	63	7.27
住院患者	64	7.39
MVB 中的患者	85	9.82
疫苗接种总数	355	40.99
疫苗接种第一剂	355	40.99
疫苗接种第二剂	355	40.99
疫苗接种第三剂	639	73.79
第一剂接种率	354	40.88
第二剂接种率	354	40.88
第三剂接种率	617	71.25
死亡证明上的 COVID-19 死亡人数	10	1.15

为开展结构学习，我们对缺失数据进行了插补。由于数据属于时间序列，我们采用状态空间模型进行插补，因为该模型在估算未观测值时会考虑时间动态。它将时间序列数据表示为由状态方程和观测方程组成的系统。状态方程描述了.....的演变

系统随时间变化的潜在“状态”，而观测方程则将这些状态与观测数据联系起来。

卡尔曼滤波器（[11]）是一种用于估计状态空间模型中状态的递归算法。当噪声服从高斯分布时，它是最优估计量。卡尔曼滤波器分为两个步骤：

- 预测：使用状态方程预测下一个状态及其不确定性。
- 更新：利用观测数据校正预测状态。

当数据点缺失时，卡尔曼滤波器仅使用预测步骤估计缺失值，并跳过该时间点的更新步骤。为使这一方法能够应用于分类数据，我们创建了虚拟变量，对其进行插补，然后再将其转换回原来的形式。重新转换（从虚拟变量转换为分类数据）时，会将插补概率最高的类别作为缺失数据点最可能所属的类别。需要明确的是，为确保分类变量的每个类别都由一个虚拟变量表示，该模型没有添加常数项。插补是在 R 中使用 imputeTS 软件包（[12]）的 na kalman 函数完成的。然而，插补后最后五行仍然包含缺失值。这或许可以用前面观察到的现象来解释：大多数变量在时间序列末尾的缺失率较高。因此，分析中省略了最后五行。

3.2 数据离散化

除本研究用于执行结构学习的一种计量经济学方法外，其他方法都需要数值数据。因此，我们会将类别转换为数字，以便使用数值数据，通过所有纳入考虑的算法进行结构学习。不过，之后我们会使用离散化数据集为贝叶斯网络设置参数。之所以这样选择，是因为我们所使用的贝叶斯网络软件（bnlearn）在实现上有所限制；该软件遵循一项常见假设，即连续节点只能作为离散变量的子节点。离散化可以避免这些限制，并确保推断过程中所有节点都能得到一致处理。

鉴于大多数变量表示强度等级，例如衡量交通使用情况的指数或住院患者人数，我们采用了将所有连续变量离散为三个类别的做法：低、中、高。我们使用k均值聚类算法进行离散化，具体而言，使用通过kmeans R函数实现的2均值变体（[13]）。我们对聚类中心进行排序，并据此确定断点位置。随后，使用cut函数根据这些断点将原始数据划分为若干分箱。

4 方法论

本节概述了用于学习贝叶斯网络结构并将其转换为动态贝叶斯网络的方法。我们首先介绍所用的计量经济学方法：最小绝对收缩与选择算子（LASSO）、最小角回归（LAR）、James-Stein收缩（JS）以及模块化网络统计推断（SIMONE）。随后介绍所评估的因果机器学习算法，并将其分为基于约束、基于评分和混合方法。最后，我们展示

由.....构建的常识图 [8], 该图是评估本研究所实现算法性能的基准之一。

假设我们有一个带有有向无环图 G 的动态贝叶斯网络，它描述了一个离散时间随机过程 $X = X_i(t)$; $i = 1, .., k$; $t = 1, ..., T$，其取值属于 R^k，包含 k 个变量，跨越 t 个时间点。若模型满足以下假设，则可以将其表示为动态贝叶斯网络：

1. 随机过程 X 是一阶马尔可夫的。这意味着，时间 t 的任何变量都仅通过上一期的观测值依赖于过去。2. $\forall t > 0$，在时间 t 观测到的随机变量 $X(t) = (X_1(t) ..., X_k(t))$，在给定时间 $t-1$ 的随机变量 $X(t-1)$ 的条件下相互独立。这意味着，在任意给定时间同时观测到的变量，在给定上一期观测值的条件下相互独立。3. 任意 X_i 的 $X_i = (X_i(1) ..., X_i(t))$ 都不能表示为任意 $j \neq i$ 的 X_j 的线性组合。当 k 个变量线性独立时，这确保了 G 的唯一性。

计量经济学方法使用向量自回归（VAR）过程对动态贝叶斯网络中的依赖关系进行建模。如果多元时间序列 Y_t遵循 p 阶 VAR 过程，则：

$$Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} + \epsilon_t \tag{1}$$

其中：

- Y_t 是时间 t 的时间序列变量向量。
- c是一个常数向量。
- A_i 是与时间序列变量向量的 i_{th} 阶滞后项相关的系数矩阵。
- ϵ_t 是时间 t时误差项的向量。[1]

4.1 计量经济学方法

4.1.1 最小绝对收缩与选择算子（LASSO）

[15]着重于在高维图模型中应用 LASSO 进行变量选择。 [16] 最初提出 LASSO，旨在处理预测变量数量远超观测值数量的高维数据分析问题。 [15] 使这一方法得以扩展并用于构建图模型。有向无环图（DAG）可以表示变量之间的条件依赖关系，而 [15] 中的过程可用于学习该图的结构。他们着重研究使用 LASSO 的邻域选择方案来处理稀疏高维图，并证明随着样本量增大，该方法以较高概率恢复真实的图。使用 LASSO 的基本思想是，现实世界中的网络应当是稀疏的，也就是说，每个变量只受到少数其他变量的直接影响。与原始 LASSO 一样，这是一种约束估计程序：通过对系数之和施加 L_1

[1]有关本方法论讨论的更多详情，请参阅 [14] 。

范数惩罚项，使某些系数趋于精确的零。只有显著偏离零的系数才定义了显著的依赖关系。

我们使用 lars 软件包在 R 中以如下方式实现（[17]）：1. 我们编写了一个函数，对每个变量（即数据框中的每一列，不包括 Date 列）应用 LASSO。2. 在拟合任何模型之前，我们使用最小-最大缩放方法对数据进行归一化，以确保所有变量都处于可比较的尺度上。3. 之后，我们通过移除一段观测值来对齐预测变量和目标变量。根据当前循环中该列是作为预测变量还是目标变量，我们移除前 p 个观测值或最后 p 个观测值。如何选择 p 将在第 6 节中说明。如本节开头所述，p 是 VAR 过程的阶数。4. 在本次循环迭代中，除目标变量外的每一列都将被视为预测变量，其中也包括滞后目标变量。5. 对于每个回归，我们都会确定最佳正则化参数值。我们通过交叉验证来实现这一点，从而在模型拟合与复杂度之间取得平衡。使用最佳正则化参数拟合模型，并获取所有预测变量相对于目标变量的系数估计值。6. 如果这些系数估计值不为零，我们就从预测变量到目标变量创建一条有向边。我们遍历各个系数，为每个系数非零的预测变量添加一条边，前提是添加该边不会形成环。

使用 lars 软件包 ([17]) 中的 LASSO 实现时，采用了以下超参数：

- **10折交叉验证**：这是默认的超参数，也是从业者常用的选择。

•**模式** $=$**“fraction”**：这会指示函数使用解路径的分数阶 L_1 范数参数化，而非采用离散步长；由此得到一个连续参数，可在估计过程中进行调节，进而形成平滑且易于解释的惩罚强度连续谱。该分数用于指示算法可以利用总可用 L_1 范数的多少。

- **类型** $=$ **”lasso”**：这可确保函数使用 LASSO，因为它也可以使用其他方法。

4.1.2 最小角回归（LAR）

LAR ([18]) 旨在为线性回归确定一个合适的预测变量子集，尤其适用于预测变量数量超过观测值数量的情况。这使我们能够对多元时间序列中的相互依赖关系进行建模；在这类场景中，候选模型会产生大量潜在预测变量。因此，在多元时间序列情境下，使用 LAR 来生成图模型的理由，与 LASSO 的用途相同。本质上，两者解决的是同一个问题，只是方式略有不同。

估计过程可分为以下步骤：1. 将所有系数设为零。2. 找出与响应变量相关性最高的预测变量。

3. 将该预测变量的系数向其最小二乘值调整（使其与不断变化的残差的相关性增加）。4. 当另一个预测变量与残差的相关性达到同等程度时，沿着与这两个预测变量等角的方向前进。5. 按照这种方式继续，根据预测变量与残差的相关性进行调整。

该方法与有向无环图的关系与前一种方法并无不同。与 LASSO 一样，该方法会返回一个稀疏有向无环图。这是因为这两种方法都遵循这样一个原则：当潜在原因很多时，我们应当预期只有少数原因会对其他变量产生显著影响。我们也在 R 中使用 lars 软件包 ([17]) 实现了该方法。实现逻辑遵循上文介绍的 LASSO 步骤。超参数需要调整为 **type =“lar”**。

4.1.3 James-Stein收缩（詹森-香农散度）

[19] 提出了一种简单的算法，可将相关网络（无向网络）转换为能够表示因果关系的有向无环图。该算法基于这样一个思路，近似搜索所有可能有向无环图的完整过程：在因果网络中，当两个变量相关时，去除其他所有节点对这两个节点的影响（偏相关）后，可以根据给定的顺序为它们之间的直接连边确定方向。影响方向由时间顺序决定；这意味着滞后变量在时间上必须先于响应变量。在时间序列分析中，偏相关有助于通过控制其他序列的影响，理解两个不同时间序列之间的直接关系。这对于我们的多元时间序列情境尤其有用，因为隔离单个变量的影响至关重要。该方法在 R 中通过协方差矩阵的高效估计量实现：允许估计的相关系数向零收缩，并使其估计方差趋近其中位数。R 中用于实现该方法的软件包是 GeneNet [20]。该方法的超参数如下：

- **方法 = ”动态”**：选择”静态”（默认选项）时，无法考虑时间依赖关系。
- **cutoff.ggm = 0.05**：它有助于筛除节点之间不显著的连接。我们选择标准的 5% 显著性水平。
- 我们保留默认设置，即使用 Student t 检验计算每个偏相关的 p 值。Student t 检验是在正态性条件下的经典精确检验，适用于样本量不大的情况，本研究正是如此。在分析过程中，我们无法针对不同变量更改检验方法，因此对混合数据集采用这一经典检验。
- **方法.ggm = ”概率”**：这意味着根据边的后验概率来选择边，也就是说，根据数据判断连接真实存在的概率。基于概率的筛选更为稳健，因为它纳入了显著性检验，而不是仅仅根据相关性值设定阈值。

4.1.4 模块化网络统计推断（SIMONE）

SIMONE ([21]) 提出了一种用于推断网络中模块的统计框架（及其在 R 中的实现）。这些模块或簇的特点是同一簇内的节点之间连接密集，而不同簇的节点之间连接较少。

该方法采用概率模型，本质上根据节点的模块归属来刻画节点之间连接的似然。该方法还引入了一种正则化形式，以抑制过于复杂的模块。为了找出最可能的模块，他们提出了一种考虑正则化的优化过程，以最大化模型对观测数据的拟合。他们在模拟数据和真实世界数据上检验了所提出的方法。该方法在模拟情形中有效恢复了真实模块，但在真实世界数据上效果较差，只能揭示有意义的模块。

SIMONE 的 R 实现具有以下超参数：

- **clusters.crit = "BIC"**：这意味着 SIMONE 将选择使 BIC 分数最优的网络结构，其定义见第5节。我们认为这一超参数选择是合适的，因为后文所述基于分数的结构学习算法同样优化该分数。
- **clustering = TRUE**：SIMONE 将找出具有相似连接的节点，并利用该信息指导网络结构的估计。这一超参数选择对时间序列数据至关重要，因为随时间相互作用的变量可能更多。将该超参数设为 true 意味着，诸如流动性和医院负担等随时间呈现模块化的变量可以得到恰当建模，从而在网络中获得更好的正则化与可解释性。
- **类型 = "时间过程"**：该超参数是我们分析的基础；它用于告知 SIMONE 使用 VAR 过程来寻找网络结构。这是此方法中专为时间序列数据设计的设置。

4.1.5 模型平均

上文讨论了用于学习动态贝叶斯网络结构的四种不同计量经济学方法。不同方法学习得到的结构往往彼此之间存在很大差异。为解决这一问题，我们还对计量经济学方法学习得到的四个图进行模型平均，以获得总体结构。模型平均的结果将在分析中与各计量经济学方法独立学习得到的四个图一并考察。模型平均所采用的流程与 [22]，中的流程相同，可概述如下：

1. 根据边在学习得到的图中出现的次数（四次以内）对其进行排序。2. 按照排序结果将边添加到总体图中。3. 需要设置一个超参数，用于指定一条边必须在给定的估计结构（本例中为四个）中出现多少次才能保留。为确定该截断值，我们对平均后的结构计算 BIC 分数并进行优化（见第5节）。也就是说，我们评估每个可能的截断值

（即保留至少在一个、两个、三个或全部四个结构中出现的边），并选择使 BIC 分数最大化的值。

4.2 因果机器学习算法

结构学习是一种无监督学习过程。其传统方法主要基于组合优化，大多数算法旨在寻找离散值邻接矩阵。该矩阵表示边的存在（及其方向）或不存在。

第 1 节所述的三类学习方法各有其假设、优点和局限性。对于哪一类最优或最合适，目前尚无共识。此外，即使属于同一类学习方法，不同算法生成的图往往也大相径庭。这种现象部分可归因于许多算法对输入数据采用了不同的假设，并依赖不同的目标函数。所考察的大多数目标函数具有分数等价性，即对于属于同一马尔可夫等价类的图结构，它们会给出相同的分数。这一特性意味着，无法从观察性数据中恢复所有有向关系。

生成唯一 DAG 结构的算法采用不具分数等价性的目标函数，或者仅从得分最高的等价类中随机生成一个 DAG。该等价类由完全部分有向无环图（CPDAG）表示，其中包含仅凭观察性数据无法确定方向的无向边（以及有向边）。在本研究中遇到这一问题时，我们将使用 R 中 bnlearn 的 cextend命令（[23]），将 CPDAG 一致地扩展为 DAG。

本研究使用的算法如下：

- PC-Stable [24]：这是一种基于约束的 PC 派生算法 [25]。它从一个完全连接的无向图开始，检查每对变量在给定条件下是否相互独立；若是，则移除连接它们的边。完成这些检验后，算法会尝试确定剩余边的方向。

- Grow-Shrink (GS) [26]：这是一种基于约束的算法，用于识别一个节点与网络中其余部分相互独立所需的最小节点集，以此构建网络结构。这个最小集合称为马尔可夫毯。利用它可以识别每个节点的邻居（父节点和子节点），以及其子节点的父节点。随后逐步识别边，以表示数据中变量之间的依赖关系。

- 增量关联马尔可夫毯 (IAMB) [27]：这是 GS 算法的一个改进版本，采用由马尔可夫毯的增长和修剪阶段组成的两步流程。修剪阶段能够更高效地识别每个变量的马尔可夫毯。
- Fast-IAMB [28]：IAMB 的改进版本，旨在通过减少条件独立性检验的数量来进一步提高效率。
- 交错增量关联马尔可夫毯（Inter-IAMB） [27]：与 IAMB 的其他变体不同，它采用前向逐步选择，有助于提高每个节点的马尔可夫毯候选集的准确性。

- 带错误发现率的 IAMB（IAMB-FDR）（[29]）：在变量选择过程中纳入错误发现率，以减少假阳性的数量。
- H2PC [30]：一种混合算法，先学习变量周围的局部结构，然后使用爬山算法（HC）学习结构。

- 2受限极大化（RSMAX2） [31]：基于独立性检验，为每个目标变量确定一小组候选父节点变量。这缩小了需要搜索的潜在图空间。

- 爬山法（HC） [32]：一种基于评分的算法，从空图开始，逐步添加、删除或重新定向边，以最大化目标函数。该算法很可能只能找到局部最大值。

- 最大-最小爬山法（MMHC） [33]：一种混合算法，首先约束每个节点的父节点集合，然后在缩减后的搜索空间中应用爬山算法（HC）来寻找最优结构。

- 禁忌搜索 [32]：与爬山算法（HC）不同，禁忌搜索（同样基于评分）有时会做出降低目标评分的更改，以尝试改进局部最大值解。

基于约束和混合的算法要求用户输入用于评估变量间条件独立性的统计检验。所选检验为“mi-g-sh”。根据 bnlearn ([23])，该检验是一种基于 James-Stein 估计量、用于互信息的收缩估计量。显著性水平保持标准值 0.05。基于评分和混合的算法要求用户输入一个评分，以评估所学习结构的质量。所选评分为“ebic-g”。根据 bnlearn ([23])，“ebic-g”在 BIC 分数的基础上增加了第二个惩罚项，用于惩罚稠密网络。

所选检验和评分都遵循惩罚稠密网络的原则。对于比较这些方法与本研究采用的计量经济学方法而言，这似乎是公平的做法，因为后者也采用了收缩。与以往相关研究 ([8],[22]) 一样，其他所有超参数均采用默认设置。文献中尚不清楚如何系统地比较具有不同类型超参数的各类算法所得的结果。鉴于算法数量众多，我们建议感兴趣的读者查阅 [23] 中的默认超参数。

4.3 知识图谱

图 1 展示了本研究所假设的知识图谱。在下方的可视化图示中，为简化呈现，对部分变量进行了分组：

- 表 1 中的 12 个移动性指数已归为「流动性指数」一类。
- 表 1 中与检验相关的 13 个变量已归为「检验」一类。

有关知识图谱构建方式的详细说明，请参阅[8]。不过，由于我们使用的变量多于 [8], ，且整合这些变量的逻辑并不明晰，因此有些变量未纳入知识图谱。具体而言：

- 记录这些观测值的日期。
- 与每日接种的疫苗数量相关的变量。
- 仅纳入了所有 *4* 个支柱的检测。能力与检测之间的进一步关系将由算法来发现。

本研究使用的知识图谱基于 [8],中所述的知识图谱，并反映了与 COVID-19 大流行相关的变量之间的关系。选择该知识图谱用于本研究，是因为它与本研究相关，且有助于说明我们的方法。 [8] 阐明了他们在对反馈回路进行建模时面临的问题，而本文以该研究所定义的静态关系为基础，并将其扩展至时间关系。图1 展示了本研究使用的知识图谱，并遵循 [8]中定义的原则。这一动态过程始于大流行：感染率和住院率上升，随后政府政策才开始实施。

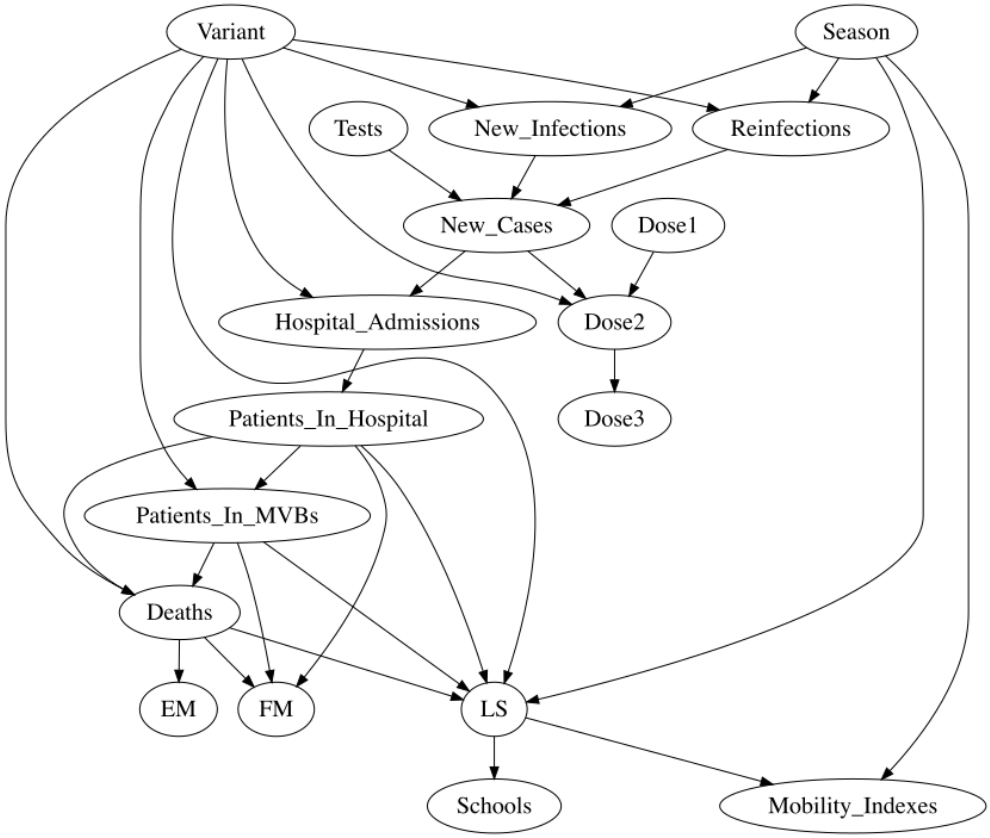


图 1：本研究基于 [8] 构建的知识图谱。

5 评估

使用各自的库运行计量经济学和因果机器学习算法后，我们通过 bnlearn 库（[23]）将它们生成的有向无环图转换为贝叶斯网络模型。由于这些算法生成的都是有向无环图结构，因此转换过程

非常简单。为了将这些模型扩展为动态贝叶斯网络，我们在数据集中加入了各变量的滞后版本。随后，我们通过五种不同的评估方法分析每种算法的输出。这些方法包括：

- 使用结构汉明距离（SHD）分数来考察学习得到的图与知识图谱之间的相似性，该分数用于量化两个图之间的差异（[33]）。
- 自由参数的数量，即向图中添加每条边所产生的额外参数数量。这是常用的指标，可用于衡量图复杂度。
- 边的数量，用于反映图密度。
- 模型选择评分：贝叶斯信息准则（BIC）。我们使用 [34], 提供的 bnlearn 实现，其定义如下：

$$BIC(G, D) = \sum_{i=1}^{p} [\log Pr(X_i | \prod_{X_i}) - \frac{|\Theta_{X_i}|}{2} \log n] \tag{2}$$

其中，G 表示图，D 表示数据，X_i 表示节点，$\prod_{X_i}$ 表示节点X_i的父节点，$|\Theta_{X_i}|$ 表示条件概率表中的自由参数数量，n 表示样本量。

- 对数似然（LL）用于衡量生成的图对数据的拟合程度。

评估所学习结构的优势和局限性，以支持政策决策，其复杂程度高于单独评估结构本身。一种验证方法是将这些结构预测的政策效果与相应政策的实际结果进行比较。由于各方对效应大小尚未达成共识，评估 COVID-19 政策尤其具有挑战性，但人们对效应方向已有共识。例如，封锁措施必然会降低而非提高感染率。所学习的结构使我们能够模拟各种政策情景，并确定哪些干预措施——尤其是旨在减少社会互动的干预措施——最能降低感染率。例如，如果我们的结构一贯得出限制互动会减少病毒传播的结论，这将有助于证明它们在为未来政策决策提供依据方面的可靠性。具体而言，我们将通过以下方式评估生成结构的能力：

- 该结构能够识别的因果效应数量。与人口互动相关的变量有十二个[2]，与感染率相关的变量有三个 [3]。因此，算法生成的每个结构最多可用于识别 36 个因果效应（即 3×12）。
- 因果效应的方向符合知识预测的次数；也就是说，效应是增加还是减少。这意味着我们将统计提高交互作用限制（或等效地，减少交互作用限制）意味着感染率降低的次数。

[2]这些指标分别是航班（7日移动平均值）、OpenTable餐厅预订（伦敦）指数、谷歌居家办公（大伦敦）移动性指数、谷歌工作场所（大伦敦）移动性指数、苹果步行（伦敦）移动性指数、谷歌公园（大伦敦）移动性指数、谷歌零售与休闲（大伦敦）移动性指数、谷歌杂货店与药店（大伦敦）移动性指数、Google公共交通站点出行指数、伦敦交通局地铁出行指数、伦敦交通局公交出行指数、Citymapper 出行指数[3]新增病例、新增感染、再次感染

6 VAR设定与诊断检验

本节第一部分介绍如何估计 VAR 过程的阶数，即时间序列过程。随后，我们分析 VAR 模型残差的正态性，并检验是否存在序列相关。我们讨论这些传统假设对分析的影响。这些诊断检查对于理解估计过程的可靠性，以及确定我们能从 VAR 模型中得出哪些有效推断至关重要。

6.1 最优向量自回归模型阶数

我们使用R中vars软件包的VARselect函数（[35]）来选择最优VAR过程阶数 p，依据如下准则：

- 赤池信息准则（AIC）（[36]）奖励拟合优度，但同时对参数数量加入惩罚项，以避免过拟合。
- 施瓦茨准则（SC）或贝叶斯信息准则（BIC）（[37]）对参数数量的惩罚比AIC更重，因而往往得到更为简约的模型。
- 汉南–奎因信息准则（汉南-奎因准则）（[38]）是AIC与BIC之间的折中，其惩罚项的量级介于二者之间。
- 最终预测误差（FPE）（[39]）基于一步预测误差。

对于从1到所选最大阶数的每个模型阶数，VARselect函数都会拟合一个向量自回归模型，并计算这些信息准则。随后，选择使各项准则达到最小值的阶数。最初，我们没有设置最大滞后阶数。然而，表3显示，使用的最大滞后阶数为17。我们将在下文讨论这一点。

每项准则都有其优缺点，并不存在适用于所有情形的正确选择。预计将比较各项准则的结果，并结合其他诊断工具和领域知识，最终确定模型阶数。

表3：滞后阶数选择准则

准则	公式	最佳滞后阶数
AIC	$-2ln(L) + 2k$	17
SC	$-2ln(L) + kln(n)$	1
HQ	$-2ln(L) + 2klnln(n)$	17
FPE	$\frac{\hat{\sigma}_p^2}{n}\frac{n+p+1}{n-p-1}$	17

其中， L 是模型的似然， k 是估计参数的数量， n 是观测值的数量，而$\hat{\sigma}_p^2$ 是误差项的估计方差。

AIC、汉南-奎因准则和 FPE 得出的向量自回归模型最优阶数为 17 阶滞后，而施瓦茨准则的最优阶数为 1 阶滞后。我们在 COVID-19 案例研究中运行 VARselect 时，将最大可能阶数设为 17 以上（也就是说，我们运行 VARselect 时，最大

可能阶数设为 18、19，以此类推），结果所有指标都开始将最大可能滞后阶数选为最优滞后阶数（也就是说，所有指标都开始选择 18 阶、19 阶等模型阶数）。这是一个问题，因为这种参数爆炸会使该向量自回归模型的估计变得不可靠或难以执行。图 2 有助于我们直观地了解这一问题。对此现象有三种可能的解释：

- 数据存在结构突变和异常值。向量自回归模型将尝试通过增加滞后阶数来弥补这一问题。
- 我们必须认识到，可能存在遗漏变量，例如潜在混杂因素，以及数据集中未能充分捕捉的其他动态过程。模型将尝试通过选择更多滞后阶数来弥补这一问题。
- 在某种程度上，这些指标侧重于模型在训练数据上的表现。因此，增加复杂度有利于实现这一目的，但不利于模型在样本外数据上的表现。

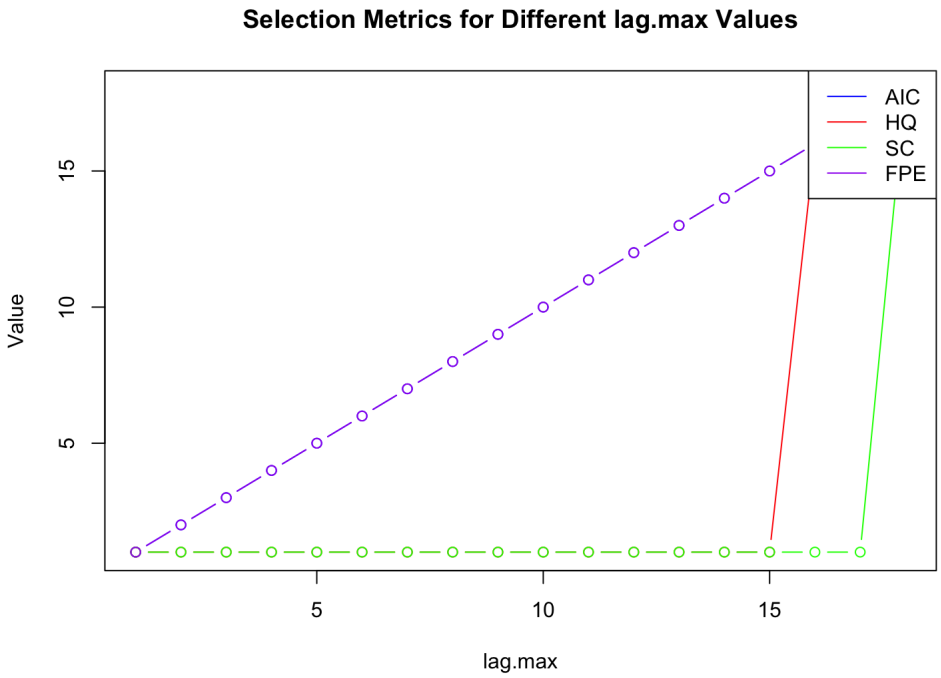


图 2：在每种可能的最大模型阶数下，各项指标对应的最优向量自回归阶数。

基于此，我们采用了一阶 VAR 模型（即包含一个滞后阶数）。这样可以使模型更加简约，尤其是在有 46 个变量和 866 个观测值的情况下。从原始变量中创建的滞后变量越多，变量数量就越接近样本量。鉴于为所有变量生成一阶滞后变量后，将有 92 个潜在变量可纳入模型

，我们假设 $VAR(1)$ 是最优选择。虽然在 COVID-19 大流行期间讨论如何支持政策决策时，这种选择可能显得有些奇怪，但观察新增感染 的偏自相关函数（PACF）可以发现，在滞后阶数为 1 时存在显著的正偏自相关。尽管更高阶的滞后仍然显著，但 PACF 的数值在滞后阶数为 1 之后大幅下降。PACF 衡量的是当前观测值与先前滞后观测值之间的相关性，同时控制其间的滞后阶数。滞后阶数为 1 时，PACF 等于自相关函数。这有助于我们更好地理解，尽管我们可能预期政策的影响需要超过一个时期才会显现，但上一时期包含了本分析所需的大部分信息。这也合乎情理，因为变化是渐进的，今天的新增感染人数与昨天的新增感染人数密切相关。新增感染 的 PACF 如下图 3 所示。

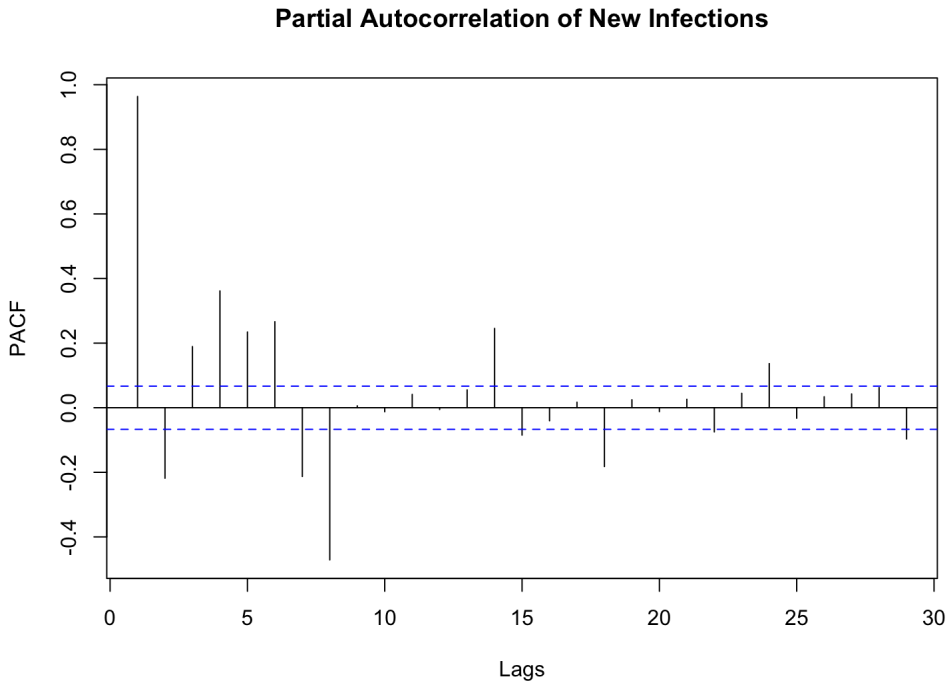


图 3：新增感染的偏自相关。

6.2 残差正态性

Jarque-Bera 检验（[40]）是一种拟合优度检验，其依据的假设是：在原假设成立时，样本所来自总体的偏度和峰度均与正态分布的相应值相同（即偏度为 $= 0$，超额峰度为 $= 0$）。检验统计量 JB 的计算公式为：

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{3}$$

其中 n 为样本量， S 为样本偏度， K 为样本峰度。如果样本偏度和超额峰度显著偏离零，则有证据拒绝原假设，转而支持备择假设，即数据并非来自正态分布总体。在 R 中，可通过 moments 软件包的 jarque.bera.test 函数实现此检验（[41]）。这些结果列于表4。

表 4: Jarque-Bera 检验结果

变量	X平方	df	p值
EM	57375	2	< 0.001
学校	1886013	2	< 0.001
FM	2292604	2	< 0.001
LS	340414	2	< 0.001
变体	1751490	2	< 0.001
航班	869.26	2	< 0.001
餐厅	18932	2	< 0.001
居家办公	140.54	2	< 0.001
工作场所	221.52	2	< 0.001
步行	1705.8	2	< 0.001
公园	327.34	2	< 0.001
零售与休闲	8459.3	2	< 0.001
杂货店与药店	21820	2	< 0.001
公共交通	957.71	2	< 0.001
Tube	8336.4	2	< 0.001
Bus	1097.3	2	< 0.001
出行	30853	2	< 0.001
季节	391019	2	< 0.001
PCR检测	315.31	2	< 0.001
PCR 检测能力	551227	2	< 0.001
抗体检测	25300	2	< 0.001
抗体检测能力	1140233	2	< 0.001
支柱1能力	3207734	2	< 0.001
支柱2能力	601332	2	< 0.001
支柱3能力	3969140	2	< 0.001
支柱4能力	2101.4	2	< 0.001
支柱1检测	79.16	2	< 0.001
支柱2检测	2486.6	2	< 0.001
支柱3检测	33161	2	< 0.001
支柱4检测	1126.1	2	< 0.001
所有 4 个支柱的检测	2354.6	2	< 0.001
新增病例	29868	2	< 0.001

表 4：Jarque-Bera 检验结果（续）

变量	X平方	df	p值
新增感染	28258	2	< 0.001
再次感染	29486	2	< 0.001
住院人数	1181.2	2	< 0.001
住院患者	454.62	2	< 0.001
MVB中的新冠患者	231.69	2	< 0.001
疫苗接种（总计）	1672.3	2	< 0.001
疫苗接种（第一剂）	25080	2	< 0.001
疫苗接种（第二剂）	13233	2	< 0.001
疫苗接种（第三剂）	39094	2	< 0.001
第一剂接种率	22.49	2	< 0.001
第二剂接种率	53.617	2	< 0.001
第三剂接种率	180116	2	< 0.001
死亡证明上的 COVID-19 死亡人数	1364.9	2	< 0.001

如表 4所示，所有序列的 p 值都非常接近于零，因此在所有情况下都拒绝原假设。这表明，我们的向量自回归模型中所有变量的残差均不服从正态分布，这违反了 VAR 过程误差服从正态分布的假设。然而，用于估计向量自回归模型的普通最小二乘法（OLS）在一致性和无偏性方面的性质并不依赖于正态性假设。对于本研究而言，这已足够，因为有向无环图中边的存在是根据向量自回归模型的系数推导得出的。由于估计量具有一致性，我们知道，随着样本量趋于无穷大，系数估计量会收敛于参数的真值。由于估计量是无偏的，我们知道，平均而言，估计量不会高估或低估参数的真实值。因此，我们认为所得到的系数是可靠的。然而，对于 JS 方法而言，未能满足正态性意味着 Student t 检验不再成立，这是一个局限；我们应预期 JS 方法会产生一些假阳性边。

6.3 序列相关检验

Breusch-Godfrey 检验（[42] 和 [43]）用于检测回归模型残差中的自相关。以下是 Breusch-Godfrey 检验过程的说明：

a. 估计模型并获取残差：

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \cdots + \beta_k X_{k,t} + u_t$$

b. 运行辅助回归：

$$u_t = \alpha_0 + \alpha_1 X_{1,t} + \cdots + \alpha_k X_{k,t} + \delta_1 u_{t-1} + \cdots + \delta_p u_{t-p} + v_t$$

其中 p 表示用于自相关检验的滞后阶数。

c. 计算 LM统计量：

$$LM = T \cdot R^2$$

其中， T 表示样本量， R^2 表示辅助回归的决定系数。在无自相关的原假设下，该统计量服从自由度为 p 的卡方分布。

在 R 中，此检验按向量自回归模型的每个方程分别实施，步骤如下：a. 从向量自回归模型中选择一个方程。b. 使用 lm函数拟合该方程，以提取残差（误差）。c. 使用 lmtest程序包中的 bgtest函数（[44]）对拟合后的模型应用 Breusch-Godfrey 检验。

该循环会针对向量自回归模型中的每个变量自动执行这一过程，逐一检测是否存在序列相关。检验结果见表 5。

表5：Breusch-Godfrey 检验结果

Test	LM检验	df	p值
EM	3.253	1	0.007
学校	0.341	1	0.559
FM	0.061	1	0.806
LS	0.002	1	0.962
变体	0.898	1	0.343
航班	288.03	1	< 0.001
餐厅	6.837	1	0.009
居家办公	32.444	1	< 0.001
工作场所	28.705	1	< 0.001
步行	22.703	1	< 0.001
公园	3.882	1	0.049
零售与休闲	0.099	1	0.753
杂货店与药店	0.135	1	0.713
公共交通	1.096	1	0.295
Tube	2.962	1	0.085
Bus	15.658	1	< 0.001
出行	0.002	1	0.963
季节	0.507	1	0.476
PCR检测	2.548	1	0.11
PCR 检测能力	1.871	1	0.171
抗体检测	37.832	1	< 0.001
抗体检测能力	2.102	1	0.147

表5：Breusch-Godfrey 检验结果（续）

Test	LM检验	df	p值
支柱1能力	1.362	1	0.243
支柱2能力	1.529	1	0.216
支柱3能力	1.144	1	0.285
支柱4能力	24.712	1	< 0.001
支柱1检测	19.349	1	< 0.001
支柱2检测	0.01	1	0.92
支柱3检测	167.59	1	< 0.001
支柱4检测	2.836	1	0.092
所有 4 个支柱的检测	0.008	1	0.93
新增病例	119.92	1	< 0.001
新增感染	112.91	1	< 0.001
再次感染	142.35	1	< 0.001
住院人数	0.085	1	0.771
住院患者	8.496	1	0.004
MVB中的新冠患者	74.709	1	< 0.001
疫苗接种（总计）	3.656	1	0.056
疫苗接种（第一剂）	0.386	1	0.534
疫苗接种（第二剂）	17.615	1	< 0.001
疫苗接种（第三剂）	3.192	1	0.074
第一剂接种率	138.42	1	< 0.001
第二剂接种率	261.06	1	< 0.001
第三剂接种率	6.466	1	0.011
死亡证明上的 COVID-19 死亡人数	47.388	1	< 0.001

如果p值小于0.05的显著性水平，则表明残差存在序列相关。这些结果表明，关于VAR过程残差行为的假设不成立。不过，系数估计仍然是无偏的，这足以满足基于系数大小的边选择方法的要求。尽管如此，这仍被视为一项局限。

7 结果

第一个小节聚焦于结构学习，将学习得到的结构与知识图谱进行比较，并考察描述学习所得结构的拟合度和复杂度的指标。后续小节讨论政策评估，重点了解这些结构在决策中的应用质量。

7.1 图形与建模指标

数值结果汇总于表 6。模型平均结构（“平均值”）纳入了四种算法中至少两种算法包含的所有边。 [4]这种方法增加了边的数量，但其 BIC 分数并非计量经济学方法中最高；SIMONE 的贝叶斯信息准则和对数似然得分均居首位。 [5]

SHD统计将学习得到的图转换为知识图谱所需的更改次数。这些更改包括添加、删除或重新定向一条边。结果显示，许多算法的SHD值都很高，与 [8]的发现相似。例如，LASSO学习得到的结构包含161条边，但相对于知识图谱，其SHD为246。这表明LASSO得出的结构与知识图谱之间的重合度有限；也就是说，要将LASSO图转换为共同知识结构，必须删除LASSO的大部分边，然后添加知识图谱中的大部分边。根据这些结果可以得出一个有力的结论：知识图谱与学习得到的图并不一致。不过，图4 显示，计量经济学方法学习得到的图之间也存在显著差异。

自由参数的数量因算法而异，差异显著，这符合预期，因为各算法学习得到的结构存在很大差异。尽管平均模型比SIMONE更复杂，但其拟合效果较弱。在计量经济学方法中，SIMONE的拟合效果最佳。平均模型仅优于LASSO、LAR和詹森-香农散度，但未能超过SIMONE。表6 中表现较差的是知识结构，其LL得分相对较低。这并不意外，因为学习算法会明确优化对数据的拟合，而知识图谱并非通过拟合数据得出。

因果机器学习算法的 SHD 值大多低于计量经济学方法，只有两种基于评分的算法 HC 和 TABU 例外。这是因为这些基于评分的方法生成的边明显多于其他算法，导致其结构比本研究测试的任何其他方法都复杂得多，但对数似然得分也明显更好。相反，基于约束的方法学习到的图相对稀疏，但仍取得了有竞争力的贝叶斯信息准则和相当不错的对数似然得分。

计量经济学方法在贝叶斯信息准则和对数似然得分方面处于劣势，这可能是因为它们会固有地施加时间约束。在我们的一阶向量自回归设定中，时间 t 的每个节点只能有时间 $t-1$ 的父节点，因此得到的动态贝叶斯网络只包含从过去指向现在的边。在下一节中，这一点至关重要；我们将讨论这些方法在为政策决策提供依据方面的效用，并指出，如果算法无法辨别关系的正确时间顺序，其在政策分析中的价值就有限。

[4]由于条件概率表的规模过大，bnlearn 报错，因此无法对 LASSO 和 LAR 进行参数化。所以，只考虑大于 0.4 而非大于 0 的系数。出于同样的原因，模型平均结构只能考虑添加至少出现在两个结构中的边。[5]需要指出的是，在 R 中使用 bnlearn 的 cextend命令（[23]）后，GS 的边无法定向以构建有向无环图。因此，这里不展示 GS 的结果。

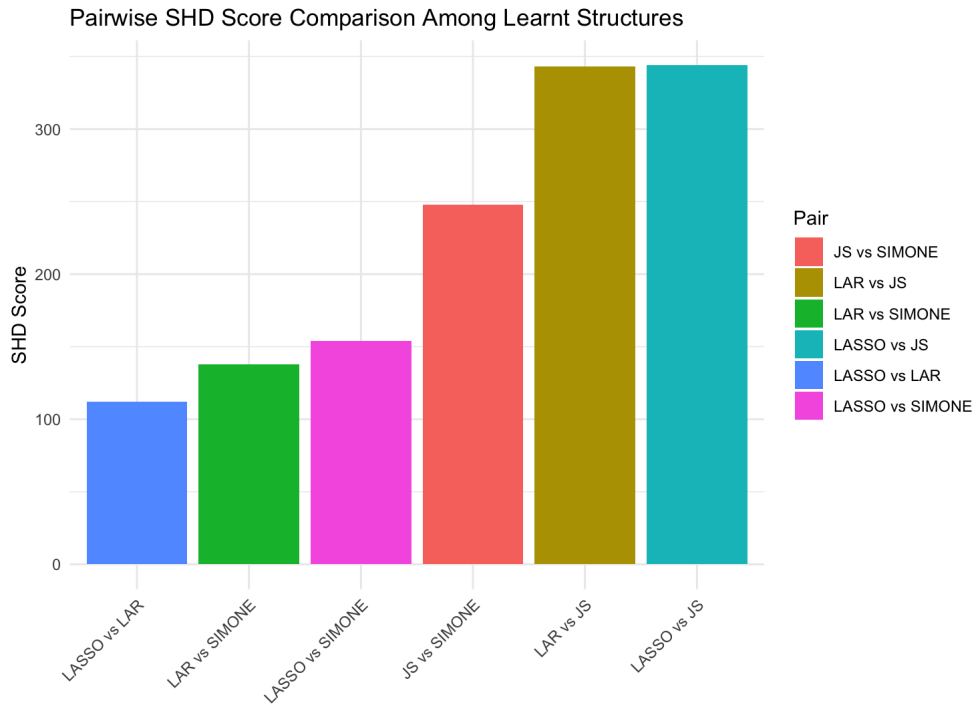


图 4：学习得到的计量经济结构之间的 SHD 比较。该图使用 ggplot2 绘制（[45]）。

7.2 政策评估

在因果贝叶斯网络中模拟干预，相当于操纵模型中的一个变量。在本研究中，这相当于例如实施一项旨在降低某个出行指数的政策干预，并估计该行动对未来 COVID-19 病例的影响。我们使用 Pearl 的 [7] do算子来模拟干预，以进行因果效应估计；其中 $P(Y|do(X = x))$ 表示在操纵 X 并将其设为状态 x 时，观察到 Y 的概率。我们使用 R软件包causaleffect（[46]）计算因果效应；该软件包实现了 [47] 和 [48], ，并通过 bnlearn（[23]）及其 cpquery 函数（[23]）实现这些操作，而该函数基于蒙特卡洛粒子滤波器。

由于连续变量的数据被离散化为三种不同状态，我们按如下方式计算平均因果效应（ACE）：

$$\frac{1}{2}\left[(P(Y|\text{do}(X=2)) - P(Y|\text{do}(X=3))) + (P(Y|\text{do}(X=1)) - P(Y|\text{do}(X=3)))\right] \tag{4}$$

平均因果效应表示结果变量（Y）的概率随着处理（X）从基准状态发生变化而产生的平均变化，并对处理变量各次转变的效应取平均值

表6：结构学习结果

算法	SHD	自由参数数量	边数	BIC	LL
LASSO	246	258	161	-40213	-39341
LAR	246	263	163	-39683	-38795
JS	246	725	195	-53467	-51017
SIMONE	151	109	66	-36914	-36546
平均值	209	238	126	-39435	-38631
基于知识的	0	407	91	-61946	-60571
PC-Stable	147	101	57	-35192	-34851
GS	NA	NA	NA	NA	NA
IAMB	155	103	65	-35103	-34755
Fast-IAMB	153	101	63	-34851	-35192
Inter-IAMB	155	103	65	-35103	-34755
IAMB-FDR算法	148	101	58	-35192	-34851
RSMAX2算法	103	189	13	-65153	-64514
MMHC	108	197	18	-62878	-62212
H2PC	102	189	12	-65229	-64591
HC	738	1662	697	-24749	-19134
TABU	740	1660	699	-24733	-19124

。计算 $P(Y \mid \text{do}(X = x))$ 时， X 指时间 $t-1$ 的干预变量值，而 Y 指时间t的结果变量。因此，以下报告的平均因果效应代表平均一步前瞻值。

公式（4）计算将 X 从设定为3的基线降低的平均因果效应，其中3表示这些变量所能取的最高值所对应的状态。此处， X表示与人口互动相关的变量，而 Y 表示与感染率相关的变量，其定义见第5节。 X 将被设为3，因为我们希望了解将 X 从高水平人口互动变为低水平人口互动是否会降低高水平 Y的似然。

表7 下方给出了与已识别因果效应数量以及效应方向与常识相符次数这两项评估指标相关的结果。表8 给出了已识别且方向与既有知识相符的平均因果效应。

如表7所示，能够识别出大量因果效应的算法只有 HC 和 TABU（各识别出 27 个效应），其中分别有 14 个和 12 个效应的方向与关于大流行的常识所暗示的方向一致（例如，人口互动减少意味着 COVID-19病例减少）。JS 识别出 2 个效应，且两者的方向都符合预期。HC 和 TABU 得出的可识别效应最多，这可以用它们生成的图更稠密来解释；可识别性与模型稀疏性之间的这种权衡值得进一步研究。然而，与计量经济学方法不同，这些因果机器学习算法并未受到必须遵循时间序列时间顺序的约束。尽管学习得到的图在静态意义上是无环的，但其中可能存在与潜在时间顺序不符的边。因此，表8 仅报告满足以下所有条件的平均因果效应（ACE）：(i) 在学习得到的图中可识别；(ii) 方向与先验知识/常识一致；以及 (iii)

这些效应来自由强制遵循过去到现在时间结构的计量经济学方法所推导出的动态贝叶斯网络（这里指基于 JS 的模型）。

根据这些标准，我们的数据集中出现了两种此类干预：降低 Citymapper行程指数和减少 OpenTable 餐厅活动量，两者都会降低高水平再次感染的概率。有趣的是，TABU 和爬山算法也识别出了这些影响。其影响幅度与詹森-香农散度的结果相近。对于 Citymapper行程，TABU 的结果为 -0.02，爬山算法的结果为 -0.01。对于 OpenTable 餐厅活动量，TABU 的结果为 -0.003，爬山算法的结果为 -0.001。我们将此视为稳健性检验。

这些影响的幅度不大，但与预期机制相符，这表明减少外出流动性和室内社交活动会降低传播风险。然而，各种算法对詹森-香农散度识别出的影响似乎具有一致性。这表明，在所评估的 COVID-19 干预中，限制这些互动是英国数据中降低感染率最有效的措施。这一结论并不容易解释。例如，一个关键问题是，人们为何进行这些出行？有证据表明，尽管家庭成员仅占接触者的6%，但却造成了40%的 COVID-19 传播（[49]）。此外，距离较远的个体需要更长时间的接触才会感染（[49]）。这一点与 OpenTable 餐厅的结果相结合，表明与熟人近距离接触的出行会增加感染风险，因为这类互动更可能发生在近距离环境中。

表 7：政策评估结果

算法	识别出的因果效应数量	方向与已知信息一致的次数
LASSO	0	0
JS	2	2
SIMONE	0	0
LAR	0	0
平均值	0	0
基于知识的	0	0
PC-Stable	0	0
GS	NA	NA
IAMB	0	0
Fast-IAMB	0	0
Inter-IAMB	0	0
IAMB-FDR算法	0	0
RSMAX2算法	0	0
MMHC	0	0
H2PC	0	0
HC	27	14
TABU	27	12

8 结论

本研究考察如何识别或假设因果关系（而非仅仅是相关性），并将其用于政策评估。研究比较了计量经济学

表8：选定出行干预对高再次感染的平均因果效应（平均因果效应），由基于詹森-香农散度的动态贝叶斯网络估计。仅报告（i）可从学习得到的图中识别且（ii）方向与常识一致的效应。

原因	效应	平均因果效应
Citymapper行程	再次感染	-0.04
OpenTable 餐厅	再次感染	-0.003

时间序列方法与因果机器学习方法，评估它们在因果发现和因果建模方面的有效性。结果表明，计量经济学方法具有一定优势，有助于应对因果机器学习在时间序列情境中的现有限制；但两种方法各有取舍，目前尚不能认为其中任何一种普遍优于另一种。

通过采用一种评估策略，既纳入与学习因果结构相关的指标，也纳入这些结构在计算因果效应方向（即正向或负向）以支持政策决策方面的成效，我们发现：

- 计量经济学方法学习得到的结构存在显著差异。
- 模型复杂度差异很大；JS 的复杂度明显高于 LASSO/LAR，而 SIMONE 则相对稀疏。
- 在计量经济学方法中，SIMONE 的 BIC 和对数似然表现最佳；平均模型并未优于 SIMONE。知识图谱在 BIC/对数似然方面表现最差，且拥有最多的自由参数。
- 在政策评估中，JS 识别出两个相关的因果效应，且二者均与常识一致。基于评分的算法（HC、TABU）识别出更多因果效应（各 27 个），但代价是生成极其稠密的图，且未施加时间限制；其中只有一部分（14 个、12 个）与知识图谱一致。总体而言，证据好坏参半，并非一致支持某一类方法。
- 相较于传统的因果机器学习，计量经济学方法通常具有更高的 SHD 和更多的边（SIMONE 除外），且在 BIC/LL 指标上表现较差；需要注意的是，传统算法并未施加时间顺序限制。
- 将计量经济学方法中的显式时间结构和模块性理念融入因果机器学习，或可减少可识别性所需的边，并提升政策实用性。
- 有必要进一步探索基于评分的方法如何对其模型进行正则化，因为它们仍会生成极其稠密的图。或许计量经济学方法也能为这一领域提供一些启示。

在各种方法中，最清晰且最一致的政策信号涉及减少出行，从而降低再感染风险。尽管从一天到下一天的效应大小较为有限，但这与持续的密切接触互动会推动传播这一机制相一致。对于决策者而言，编码时间先后关系并惩罚复杂度的模型，似乎更适合用于对干预措施进行压力测试。

非正态误差、局部残差自相关以及潜在的遗漏变量等局限，均表明存在模型设定错误的风险。离散化

（DBN参数化所必需）可能会削弱信号，而缺乏普遍适用的时间约束则使直接比较更加困难。该知识图谱并非为动态环境而设计，进一步限制了SHD可解释性。

本研究对不同方法的比较得出了可为后续研究提供指导的结果。计量经济学方法未能应对该数据集的复杂度和遗漏变量。显然，我们还需要进一步开发方法，以便在时间序列环境中更准确地识别因果关系，为政策决策提供依据。众所周知，防止人际接触可以阻止大流行病传播，但许多受测算法未能检测到这些关系，这一事实清楚地表明了这一点。

声明

8.1 数据的可用性与获取

本研究所用数据已随论文一并提交。

8.2 竞争利益

作者声明，他们没有已知的、可能被认为影响本文所报告工作的财务利益冲突或个人关系。

8.3 数据使用的伦理与知情同意

本文不包含任何作者开展的涉及人类参与者或动物的研究。

8.4 贡献

Bruno Petrungaro：概念构思、方法设计、软件、分析、撰写初稿、审阅与编辑。
Anthony C. Constantinou：撰写、审阅与编辑、指导。

参考文献

[1] Athey, S., Imbens, G.W.：应用计量经济学的现状：因果关系与政策评估。《Journal of Economic Perspectives》**31**(2), 3–32 (2017年) https://doi.org/10.1257/jep.31.2.3

[2] Pinotti, P.：敲开天堂之门：移民合法化对犯罪的影响。《美国经济评论》**107**(1), 138–68（2017年）https://doi.org/10.1257/aer.20150355

[3] Kocacoban, D., Cussens, J.：存在潜变量时的在线因果结构学习。2019年第18届IEEE国际机器学习与应用会议（ICMLA），392–395（2019年）https://doi.org/10.1109/ICMLA.2019.00073

[4] Kummerfeld, E., Danks, D.：时变因果结构的在线学习。载于：UAI因果结构学习研讨会（2012年）[5] Kummerfeld, E., Danks, D.：追踪时变图结构。神经信息处理系统进展第26卷（2013年）[6]Kermack, W., McKendrick, A.：对流行病数学理论的贡献。伦敦皇家学会会刊A辑115(772)，700–721（1927年）[7] Pearl, J.：用于实证研究的因果图。Biometrika 82，669–710（1995年）[8] Constantinou, A.，等：因果结构学习中的未决问题：以英国COVID-19为例。《专家系统与应用》234（2023年）https: //doi.org/10.1016/j.eswa.2023.121069[9] 冠状病毒（COVID-19）：英国检测。英国政府冠状病毒（COVID-19）数据与洞见官方网站。[10] GOV.UK（2022d）。冠状病毒（COVID-19）：英国医疗保健。英国政府冠状病毒（COVID-19）数据与洞见官方网站。[11] Kalman,R.E.：一种线性滤波与预测问题的新方法。《基础工程学报》82(1)（1960年）https://doi.org/10.1115/1.3662552[12] Moritz, S., Bartz-Beielstein, T.：imputeTS：R中的时间序列缺失值插补。《R期刊》9(1)，207–218（2017年）https://doi.org/10.32614/RJ-2017-009[13] Hartigan, J.A., Wong, M.A.：算法AS 136：一种k均值聚类算法。《英国皇家统计学会学报C辑（应用统计）》28(1)，100–108（1979年）[14] Radhakrishnan, N.，等：R中的贝叶斯网络及其在系统生物学中的应用。Springer（2013年）[15] Meinshausen, N., B¨uhlmann, P.：利用LASSO进行变量选择和高维图建模。《统计年鉴》34，1436–1462（2006年）[16] Tibshirani, R.：通过LASSO实现回归收缩与变量选择。《英国皇家统计学会学报B辑：统计方法学》58(1)，267–288（1996年）[17] Hastie, T., Efron, B.：lars：最小角回归、LASSO与前向逐步回归（检索于2023年12月）[18] Efron, B.，等：最小角回归。《统计年鉴》32(2)，407–451（2004年）

[19] Opgen-Rhein, R., Strimmer, K.：从系统生物学时间进程数据中学习因果网络：针对向量自回归过程的一种有效模型选择程序。BMC Bioinformatics 8(2), 1–8 (2007)[20] Schaefer, J., 等：GeneNet：基因网络的建模与推断（检索于2023年12月）[21] Chiquet, J., 等：SIMONE：模块网络统计推断。Bioinformatics 25(3), 417–418 (2009)[22] Petrungaro, B., Kitson, N.K., Constantinou, A.C.：运用贝叶斯网络结构学习探究脓毒症的潜在成因。Applied Intelligence 55(6), 496 (2025)[23] Scutari, M.：使用bnlearn R软件包学习贝叶斯网络。Journal of Statistical Software 35(3), 1–22 (2010) https://doi.org/10.18637/jss.v035.i03[24] Colombo, D., 等：与排序无关的基于约束的因果结构学习。J. Mach. Learn. Res. 15(1), 3741–3782 (2014)[25] Spirtes, P., 等：一种快速恢复稀疏因果图的算法。Social Science Computer Review 9(1), 62–72 (1991)[26] Margaritis, D.：从数据中学习贝叶斯网络模型结构。博士论文，卡内基梅隆大学计算机科学学院，美国宾夕法尼亚州匹兹堡 (2003)[27] Tsamardinos, I., 等：大规模马尔可夫毯发现算法。载于：FLAIRS 会议，第2卷，第376–380页 (2003)。美国佛罗里达州圣奥古斯丁[28] Yaramakala, S., 等：用于最优特征选择的推测式马尔可夫毯发现。载于：第五届IEEE国际数据挖掘会议（ICDM'05），第4页 (2005)。IEEE[29]Pena, J.M.：在控制错误发现率的条件下学习基因网络的高斯图模型。载于：欧洲进化计算、机器学习与生物信息学数据挖掘会议，第165–176页 (2008)。Springer[30] Gasse, M., 等：一种用于贝叶斯网络结构学习的混合算法及其在多标签学习中的应用。《专家系统与应用》41(15), 6755–6772 (2014)[31] Friedman, N., 等：从海量数据集中学习贝叶斯网络结构："稀疏候选"算法。arXiv预印本 arXiv:1301.6696 (2013)[32] Bouckaert, R.R.：贝叶斯信念网络：从构建到推断。博士论文 (1995)

[33] Tsamardinos, I., 等：最大-最小爬山贝叶斯网络结构学习算法。机器学习 65, 31–78 (2006)[34]Scutari, M., Denis, J.-B.：《贝叶斯网络：R语言实例》，第2版。Chapman and Hall，博卡拉顿 (2021)。ISBN 978-0367366513[35] Pfaff, B.：Var、svar和svec模型：在R软件包vars中的实现。Journal of Statistical Software 27(4) (2008)[36] Akaike, H.：重新审视统计模型识别。《IEEE自动控制汇刊》19(6), 716–723 (1974)[37] Schwarz, G.：估计模型的维数。《统计年鉴》，461–464 (1978)[38] Hannan, E.J., Quinn, B.G.：自回归模型阶数的确定。《皇家统计学会杂志：B辑（方法论）》41(2), 190–195 (1979)[39] Akaike, H.：用于预测的自回归模型拟合。Ann Inst Stat Math 21, 243–247 (1969) https://doi.org/10.1007/BF02532251[40] Jarque, C.M., Bera, A.K.：观测值与回归残差正态性的检验。《国际统计评论》/Revue Internationale de Statistique, 163–172 (1987)[41] Komsta, L., Novomestky, F.：Moments：矩、累积量、偏度、峰度及相关检验。(2022)。R软件包版本0.14.1。https://CRAN.R-project.org/package=moments[42] Godfrey, L.G.：当回归变量包含滞后因变量时，针对一般自回归和移动平均误差模型进行检验。《计量经济学》：计量经济学会期刊，1293–1301 (1978)[43] Breusch, T.S.：动态线性模型中的自相关检验。《澳大利亚经济论文集》17(31), 334–355 (1978)[44] Zeileis, A., Hothorn, T.：回归关系的诊断检验。《R新闻》2(3), 7–10 (2002)[45] Wickham, H.：ggplot2：优雅的图形数据分析。Springer，纽约，美国 (2016)。https://ggplot2.tidyverse.org[46] Tikka, S., Karvanen, J.：使用R软件包causaleffect识别因果效应。Journal of Statistical Software 76(12), 1–30 (2017) https://doi.org/10.18637/jss.v076.i12[47] Shpitser, I., Pearl, J.：条件干预分布的识别。

arXiv预印本 arXiv:1206.6876（2012年）

[48] Shpitser, I.,Pearl, J.：递归半马尔可夫因果模型中联合干预分布的识别。见：全国人工智能会议论文集，第21卷，第1219页（2006年）。加利福尼亚州门洛帕克；马萨诸塞州剑桥；伦敦；AAAI出版社；麻省理工学院出版社；1999年

[49] Ferretti, L.等：基于700万名接触者的SARS-CoV-2传播风险数字化测量。《自然》（2023年）