用于缺失混杂因素因果推断的随机神经网络

Yaxin Fang *麻醉学、围术期与疼痛医学系 斯坦福大学 帕洛
阿尔托，加利福尼亚州 94305，美国yxfang@stanford.edu

Faming LiangD
f S
系统计学 普渡大学 西拉法叶，印
第安纳州 47907，美国
fmliang@purdue.edu

arXiv:2603.01230v1 [stat.CO] 1 Mar 2026

摘要

未测量混杂是基于观察性数据进行因果推断的一项根本障碍。潜变量方法通过插补未观测混杂因素来应对这一挑战，但许多方法缺乏明确的基于模型的可识别性保证，也难以扩展到更丰富的因果结构。我们提出基于随机神经网络的混杂因素插补（CI-StoNet），利用随机神经网络对因果有向无环图的条件结构进行参数化，并通过自适应随机梯度哈密顿蒙特卡洛插补潜在混杂因素。在满足稳定单元处理值假设和重叠性，并假设数据生成过程的结构组件可由容量受控的稀疏深度神经网络类充分逼近的条件下，我们证明了该类中的模型可识别性，并确立了固定干预下平均潜在结果的一致估计。尽管潜在混杂因素只能在保持处理—结果联合分布不变的重新参数化意义下识别，但因果估计目标在这一观察等价类中保持不变。我们还刻画了重叠性对估计准确度的影响。模拟数据和基准数据集上的实证结果表明，该方法表现准确；此外，该框架可自然扩展到具有重叠性诊断和基于自助法的不确定性量化的代理变量及多重原因情形。

1 引言

从观察性研究中进行因果推断，是遗传学、经济学和社会科学等领域广泛关注的课题。在潜在结果框架下（Rubin，1974年），识别因果效应的一项基本条件是强可忽略性条件（Rosenbaum 和 Rubin，1983年）：

$$\boldsymbol{A} \perp\!\!\!\perp \{\boldsymbol{Y}(\boldsymbol{a}) : \boldsymbol{a} \in \mathcal{A}\} \mid \boldsymbol{Z}, \tag{1}$$

其中，$\boldsymbol{A}$ 表示取值属于空间 $\mathcal{A}$ 的处理变量，$\boldsymbol{Y}(\cdot)$表示结果函数，$\boldsymbol{Z}$ 表示混杂因素。混杂变量是指同时影响处理和结果的变量。强可忽略性假设要求所有混杂因素都被观测到。在观察性研究中，这一条件很少能完全满足，因此因果效应估计存在产生重大偏差的风险。

*实验代码可在此处获取：https://github.com/nixay/Stochastic-Neural-Networks-for-Causal-Inference-with-Missing-Confounders

应对缺失混杂因素问题的一种策略是将其建模为潜变量。Wang & Blei (2018) 提出使用潜因子模型为多个原因获取潜在表示，从而在不存在单原因混杂因素的假设下捕获多原因混杂因素。Kallus et al. (2018) 在代理变量设定下处理这一问题，通过矩阵分解获取代理变量的低秩成分，并将其作为真实混杂因素的近似。Louizos 等（2017）也借助代理变量处理这一问题，并提出因果效应变分自编码器（CEVAE），从代理变量、处理和结果的观测分布中推断缺失混杂因素。这些工作代表了利用观察性数据进行因果推断的重要进展；然而，它们也存在明显局限。例如，Imai & Jiang (2019) 指出，Wang & Blei (2018) 本质上将替代混杂因素建模为处理的确定性函数，因此它最终收敛到观测处理的函数，而非真实混杂因素。Kallus et al. (2018) 主要关注线性回归设定，因此除非有大量代理变量对应少数潜变量，否则其在非线性模型中的适用性有限。Rissanen & Marttinen (2021) 考察了 Louizos 等（2017）中因果效应估计量的一致性，并表明当潜变量设定错误或数据分布过于复杂时，该估计量无法正确估计原因效应。

本文提出一种潜变量填补方法，以解决混杂因素缺失问题。这种新方法建立在随机神经网络（StoNet）（Sun & Liang，2022年；Liang 等人，2022年）和稀疏深度学习理论（Sun 等人，2022年）的基础上，有效克服了现有方法的局限性。

核心思想是在 StoNet 中编码因果有向无环图的马尔可夫分解，并通过由此得到的条件分布插补潜在混杂因素。重要的是，我们的理论保证基于模型：在稳定单元处理值假设和重叠性成立，且假设处理机制和结果机制属于可由稀疏深度神经网络良好逼近的函数类时，因果泛函 $\mathbb{E}[Y(\boldsymbol{a})]$ 可被唯一确定，并且可以一致估计。

StoNet 使用自适应随机梯度MCMC算法进行训练（Liang 等人，2022年；Deng 等人，2019年），该算法能够同时插补缺失混杂因素并估计稀疏 StoNet 参数。我们将所提出的方法称为基于随机神经网络的混杂因素插补（CI-StoNet）。总之，该方法在解决缺失混杂变量问题方面具有以下优势：

(i) **准确的因果效应估计：** 这一特性得益于 StoNet 处理缺失数据的固有能力、稀疏深度学习的一致性，以及自适应随机梯度马尔可夫链蒙特卡罗算法提供的收敛保证。

(ii) **复杂非线性建模**。 CI-StoNet继承了深度神经网络（DNN）的通用逼近特性，能够在各种应用中有效地建模复杂非线性关系。

(iii) **结构灵活性**。 CI-StoNet的马尔可夫架构具有结构灵活性，可用于表示因果有向无环图中的各种依赖模式。它支持对每个DNN模块进行局部更新，促进模块化设计，并能轻松适应不同的因果关系。

2 缺失混杂因素的CI-StoNet

2.1 CI-StoNet方法

本节介绍CI-StoNet方法，或者更广义地说，介绍一种用于在存在缺失混杂因素时执行因果推断的深度学习框架。

考虑图 1 所示的简单混杂情形，其中涉及处理 $\boldsymbol{A} \in \{\boldsymbol{a}_1$、……、$\boldsymbol{a}_m\}$、缺失/潜在混杂因素 $\boldsymbol{Z}$，以及一个结果 $\boldsymbol{Y}$。相应的模型为

$$\begin{aligned} \boldsymbol{A} &= g_1(\boldsymbol{Z}, \boldsymbol{e}_a), \\ \boldsymbol{Y} &= g_2(\boldsymbol{Z}, \boldsymbol{A}) + \boldsymbol{e}_y, \end{aligned} \tag{2}$$

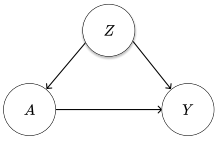


图 1：简单混杂

其中 $g_1(\cdot)$ 和 $g_2(\cdot)$ 是未知函数，可以是非线性且高度复杂的； $\boldsymbol{e}_a$ 和 $\boldsymbol{e}_y$ 是随机误差。本文假设 $\boldsymbol{e}_y \sim N(\mathbf{0}, \sigma_y^2 I_{d_y})$，其中 d_y 表示 $\boldsymbol{Y}$ 的维数。 $\boldsymbol{e}_a$ 的分布设定具有灵活性。如果 $\boldsymbol{A}$ 是连续变量，则可以假设 $\boldsymbol{e}_a$ 服从高斯分布或其他任意连续分布。如果 $\boldsymbol{A}$ 是混合型变量，则可以相应地指定 $\boldsymbol{e}_a$ 各分量的分布。该情形将 Wang & Blei (2018) 中考虑的多个原因作为一种特殊情况纳入其中，此时 $\boldsymbol{a}_i$ 是一个多维向量。由于 $\boldsymbol{Z}$ 缺失，我们根据以下条件分布对其进行插补：

$$\pi(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{Y}) \propto \pi(\boldsymbol{Z})\pi(\boldsymbol{A}|\boldsymbol{Z})\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{A}) \propto \pi(\boldsymbol{Z}|\boldsymbol{A})\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{A}), \tag{3}$$

其中 $\pi(\cdot)$ 表示适当语境中的分布或条件分布 .

公式（3）的后半部分表明，从数学上讲， $\boldsymbol{A}$、 $\boldsymbol{Z}$和$\boldsymbol{Y}$ 可以解释为随机模型的外生输入、潜在状态和输出。受此观点启发，我们提出使用 CI-StoNet 进行插补（其结构见图2），其公式如下：

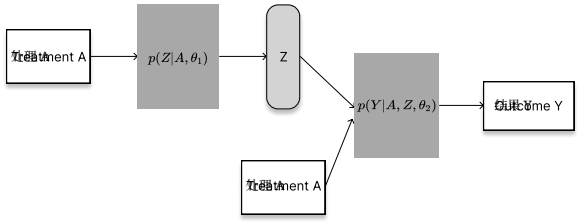


图2：简单混杂情形下 CI-StoNet 的示意图，其中白色矩形表示来自观测数据的变量；浅灰色圆角矩形表示待填补的潜变量；深灰色矩形表示用于学习相应条件分布的神经网络模块。

$$\begin{aligned} \boldsymbol{Z} &= \mu_1(\boldsymbol{A},\boldsymbol{\theta}_1) + \boldsymbol{e}_z, \\ \boldsymbol{Y} &= \mu_2(\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\theta}_2) + \boldsymbol{e}_y, \end{aligned} \tag{4}$$

其中， $\mu_1(\cdot)$和 $\mu_2(\cdot)$是两个神经网络函数，分别由 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$参数化；$\boldsymbol{e}_z \sim N(\mathbf{0}, \sigma_z^2 I_{d_z})$；并且 $\boldsymbol{e}_y$定义为的 $\boldsymbol{Z}$。此外，式（2）中。两个神经网络通过潜变量相互连接，我们对模型（2）和（4）作出如下假设：

假设1. *(i)* $\boldsymbol{e}_a \perp\!\!\!\perp \boldsymbol{e}_y$, $\boldsymbol{e}_a \perp\!\!\!\perp \boldsymbol{Z}$, $\boldsymbol{e}_y \perp\!\!\!\perp (\boldsymbol{Z}, \boldsymbol{A})$; *(ii)* 存在稀疏深度神经网络 $\mu_1(\cdot)$和 $\mu_2(\cdot)$，使得式 *(4)* 成立， $\boldsymbol{e}_z \sim N(\mathbf{0}, \sigma_z^2 I_{d_z})$， $\boldsymbol{e}_y \sim N(0, \sigma_y^2 I_{d_y})$, $\boldsymbol{e}_z \perp\!\!\!\perp \boldsymbol{e}_y$，且 $\boldsymbol{e}_z \perp\!\!\!\perp \boldsymbol{A}$；

假设1的第 (i) 部分确保强可忽略性 (1) 成立。第 (ii) 部分通过将处理机制和结果机制限制在一个稀疏深度神经网络族中（其中 $\boldsymbol{Z}$ 为随机变量），来规定工作模型类。这一结构性限制使理论分析易于处理，并定义了伪真参数 $\boldsymbol{\theta}^*$，该参数使其与真实数据生成分布之间的 KL 散度最小。更多细节请参见第 2.2.1 节。

值得注意的是，式 (4) 中的函数表达式并不意味着存在因果机制 $\boldsymbol{A} \to \boldsymbol{Z}$。例如，雨（$\boldsymbol{Z}$）会导致湿地（$\boldsymbol{A}$），但湿地不会导致雨。同样，由于不存在因果机制 $\boldsymbol{A} \to \boldsymbol{Z}$， $\boldsymbol{Z}$ 不能被解释为中介变量，尽管式 (4) 在数学结构上与中介模型相似。目前，我们假设处理 $\boldsymbol{A}$ 与结果 $\boldsymbol{Y}$ 之间的因果路径上不存在中介变量，从而排除对 $\boldsymbol{Z}$ 作用的任何潜在误解。不过，如果确实存在中介变量，那么与总因果效应估计以及 $\boldsymbol{Z}$ 的解释相关的问题，将在第 2.2 节末尾讨论。

在缺失数据框架下，可以通过求解以下方程来训练 CI-StoNet，该方程表示费舍尔恒等式的贝叶斯版本（Song等（2020））：

$$\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\boldsymbol{A},\boldsymbol{Y}) = \int \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{A},\boldsymbol{Y}) \pi(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{Y},\boldsymbol{\theta}) d\boldsymbol{Z}, \tag{5}$$

其中， $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$， $\boldsymbol{Z}$缺失， $\pi(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{A},\boldsymbol{Y}) \propto \pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2)\pi(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}_1)\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\theta}_2)$，$\pi(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{Y},\boldsymbol{\theta}) \propto \pi(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{\theta}_1)\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\theta}_2)$，而 $\pi(\boldsymbol{\theta}_1)$ 和 $\pi(\boldsymbol{\theta}_2)$ 分别表示施加于 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$ 上的先验分布。本文假设 $\boldsymbol{\theta}$ 的各个分量先验独立，并服从以下混合高斯先验（Sun等人，2022年）：

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{K_n} (1-\lambda_n)\phi(\boldsymbol{\theta}_i/\sigma_0) + \lambda_n \phi(\boldsymbol{\theta}_i/\sigma_1), \tag{6}$$

其中， λ_n 为混合比例， K_n 为 StoNet 中连接的总数（即 $\boldsymbol{\theta}$ 的维数）， $\phi(\cdot)$ 表示标准正态分布的密度函数，而 σ_0 和 σ_1 分别是两个高斯分量的标准差。

恒等式 (5) 进一步表明，目标方程

$$\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|\boldsymbol{A}, \boldsymbol{Y}) = 0, \tag{7}$$

可以使用自适应随机梯度马尔可夫链蒙特卡罗算法求解，该算法在潜变量填补和参数更新之间迭代交替进行。本文采用自适应随机梯度哈密顿蒙特卡洛（随机梯度哈密顿蒙特卡洛）（Liang 等人，2022年），如算法 1 所示，来求解方程 (7)。

算法 1: 自适应随机梯度哈密顿蒙特卡洛

0. 设置先验超参数： λ_n、 σ_0 和 σ_1。

1.（潜变量填补）通过哈密顿蒙特卡洛更新，从 $\pi(\boldsymbol{Z}|\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{\theta})$ 模拟 $\boldsymbol{Z}$：

$$\begin{aligned}\mathbf{v}^{(k+1)} &= (1-\epsilon_{k+1}\eta)\mathbf{v}^{(k)} + \epsilon_{k+1}\nabla_{\boldsymbol{Z}} \log \pi(\boldsymbol{Z}^{(k)}|\boldsymbol{A}, \boldsymbol{\theta}_1^{(k)}) + \epsilon_{k+1}\nabla_{\boldsymbol{Z}} \log \pi(\boldsymbol{Y}|\boldsymbol{Z}^{(k)}, \boldsymbol{A}, \boldsymbol{\theta}_2^{(k)})) \\ &\quad + \sqrt{2\epsilon_{k+1}\eta}\mathbf{e}^{(k+1)}, \\ \mathbf{Z}^{(k+1)} &= \mathbf{Z}^{(k)} + \epsilon_t \mathbf{v}^{(k)},\end{aligned}$$

其中 $\boldsymbol{e}^{(k+1)} \sim N(\mathbf{0}, I_{d_z})$， d_z是 $\boldsymbol{Z}$的维度， ϵ_{k+1}是学习率。

2.（参数更新）给定 $\boldsymbol{Z}^{(k+1)}$，分别更新 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$：

$$\begin{aligned}\boldsymbol{\theta}_1^{(k+1)} &= \boldsymbol{\theta}_1^{(k)} + \gamma_{k+1}\nabla_{\boldsymbol{\theta}_1} \log \pi(\boldsymbol{Z}^{(k+1)}|\boldsymbol{A}, \boldsymbol{\theta}_1^{(k)}) + \gamma_{k+1}\nabla_{\boldsymbol{\theta}_1} \log \pi(\boldsymbol{\theta}_1^{(k)}), \\ \boldsymbol{\theta}_2^{(k+1)} &= \boldsymbol{\theta}_2^{(k)} + \gamma_{k+1}\nabla_{\boldsymbol{\theta}_2} \log \pi(\boldsymbol{Y}|\boldsymbol{Z}^{(k+1)}, \boldsymbol{A}, \boldsymbol{\theta}_2^{(k)}) + \gamma_{k+1}\nabla_{\boldsymbol{\theta}_2} \log \pi(\boldsymbol{\theta}_2^{(k)}).\end{aligned}$$

在模型（4）中， σ_z 和 σ_y 均为标量。它们可以作为超参数在模拟中指定，同时对下游推断的影响很小。值得注意的是，由于神经网络的通用逼近性质， σ_z 在模型（4）中本质上是不可识别的。在推断阶段，见公式（9），我们为 σ_z^2 提供了一个贝叶斯估计量，以便对缺失混杂因素进行插补。具体而言，我们施加逆伽马先验 $\sigma_z^2 \sim$ InvGamma(α, β)，从而得到贝叶斯估计量：

$$\hat{\sigma}_z^2 = \frac{\beta + \frac{1}{2}\sum_{j=1}^{n}(z_j - \mu_1(\boldsymbol{A}_j, \boldsymbol{\theta}_1))^2}{\frac{n}{2} + \alpha - 1}, \tag{8}$$

其中，对于平坦先验，我们设定 $\alpha = \beta = 1$， $\boldsymbol{A}_j$ 表示样本 j中 $\boldsymbol{A}$ 的取值。在模拟中，还可以按照（8）随迭代进行更新，同时对算法性能的影响很小。

为实现因果推断，我们提出以下额外假设，这些假设是进行因果效应识别的标准条件：

假设 2. *1.* **稳定单元处理值假设（SUTVA）**：一个受试者的潜在结果独立于分配给另一受试者的处理；也就是说，受试者之间不存在干扰，且每种分配的处理都只有一个版本。

2. **重叠性**：替代混杂因素 $\boldsymbol{Z}$ 满足重叠条件： $p(A \in \boldsymbol{A}|\boldsymbol{Z}) > 0$ 对于所有具有正测度的集合 $\boldsymbol{A}$，即 $p(\boldsymbol{A}) > 0$。

Un在假设1和假设2下，可以通过以下步骤估计因果效应 :

因果效应估计。 算法 1 收敛后，使用学习到的参数$\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\theta}}_1^*, \hat{\boldsymbol{\theta}}_2^*)$，针对每个观测 i，从 $\pi(\boldsymbol{z} \mid \boldsymbol{a}_i; \hat{\boldsymbol{\theta}}_1^*)$ 中抽取 $\mathcal{M}$ 个样本 $\{\boldsymbol{z}_i^{(l)}\}_{l=1}^{\mathcal{M}}$。期望

结果 $\mathbb{E}\{Y(\boldsymbol{a}) \mid \boldsymbol{\theta}^*\} = \int \mu_2(\boldsymbol{z}, \boldsymbol{a}; \boldsymbol{\theta}_2^*)\ \pi(\boldsymbol{z} \mid \boldsymbol{\theta}_1^*)\ d\boldsymbol{z}$ 随后通过蒙特卡洛平均进行近似

$$\widehat{\mathbb{E}(Y(\boldsymbol{a})|\hat{\boldsymbol{\theta}}^*)} = \frac{1}{n\mathcal{M}} \sum_{i=1}^{n} \sum_{l=1}^{\mathcal{M}} \mu_2(\boldsymbol{z}_i^{(l)}, \boldsymbol{a}, \hat{\boldsymbol{\theta}}_2^*). \tag{9}$$

估计量的合理性论证见附录 A5.3。随后即可据此推导出处理效应估计量。

2.2 理论保证

2.2.1 基于模型的可识别性

沿用 Rissanen & Marttinen (2021) 的做法，我们区分非参数可识别性与基于模型的可识别性。非参数可识别性关注的是，能否仅凭观测分布恢复因果效应；而基于模型的可识别性则考察，在受限的模型类中，因果效应是否唯一。我们的结果针对后者。具体而言，我们通过以下方式限制潜变量模型类：

假设 3. （混杂机制）结构条件均值函数$m_A(\boldsymbol{z}) = \mathbb{E}[\boldsymbol{A}|\boldsymbol{Z} = \boldsymbol{z}]$ 和 $m_Y(\boldsymbol{a}, \boldsymbol{z})$ $= \mathbb{E}[\boldsymbol{Y}|\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{Z} = \boldsymbol{z}]$ 属于定义在有界域上的函数类 $\boldsymbol{F}$，并且存在一列稀疏深度神经网络 $\mu_{A,n}$ 和 $\mu_{Y,n}$ 满足假设 *A10*，且

$$\|\mu_{A,n} - m_A\|_{L^2} + \|\mu_{Y,n} - m_Y\|_{L^2} \leq \omega_n$$

其中， ω_n 是一个在 $n \to \infty$时收敛至 *0* 的序列

许多函数类 $\mathcal{F}$都满足此条件，详细讨论见附录 A5.4。在此类限制下，令 $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta_n\}$ 表示在工作因果有向无环图和稀疏深度神经网络设定（假设1-(ii)）下，由 CI-StoNet 诱导的观测数据分布族。对于目标泛函 $\psi_{\boldsymbol{\theta}}(\boldsymbol{a}) = \mathbb{E}_{P_{\boldsymbol{\theta}}}[Y(\boldsymbol{a})],$，我们称 $\psi_{\boldsymbol{\theta}}(\boldsymbol{a})$ 关于 $P_{\boldsymbol{\theta}}$ 是基于模型可识别的，如果任意两个参数值 $\boldsymbol{\theta}$、 $\boldsymbol{\theta}'$ 诱导出相同的观测数据分布（例如，简单混杂下的 $p_{\boldsymbol{\theta}}(\boldsymbol{A}, \boldsymbol{Y})$，或代理变量设定下的 $p_{\boldsymbol{\theta}}(\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{X})$），则对所有 $\boldsymbol{a}$，都有 $\psi_{\boldsymbol{\theta}}(\boldsymbol{a}) = \psi_{\boldsymbol{\theta}'}(\boldsymbol{a})$。

在我们的设定中，由于损失不变变换， $\boldsymbol{Z}$ 和 $\boldsymbol{\theta}$ 可能并不唯一，但因果泛函在每个观测等价类内保持不变；因此，下文给出的定理 A1（在此类变换意义下的参数恢复）和定理 1（$\hat{}$ $\psi(\boldsymbol{a})$）的一致性）共同确立了指定函数族内 $\mathbb{E}[Y(\boldsymbol{a})]$ 的基于模型的识别与相合估计。当真实机制不属于稀疏深度神经网络函数族时，定理2量化了由此产生的误设偏差。

2.2.2 理论分析

令 P_0 表示$(\boldsymbol{A}, \boldsymbol{Z}, \boldsymbol{Y})$的真实联合分布， $P_{\boldsymbol{\theta}}$ 为由参数 $\boldsymbol{\theta}$的CI-StoNet所诱导的分布。定义伪真实随机神经网络参数

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}(P_0, P_{\boldsymbol{\theta}}) := \arg\min_{\boldsymbol{\theta}} \int \log \frac{dP_0}{dP_{\boldsymbol{\theta}}} dP_0.$$

令 $\psi(P_0) = \int m_Y(\boldsymbol{a}, \boldsymbol{z}) p_{P_0}(\boldsymbol{z}) d\boldsymbol{z}$ 为真实潜在结果， $\psi(P_{\boldsymbol{\theta}^*}) = \mathbb{E}(Y(\boldsymbol{a})|\boldsymbol{\theta}^*) = \int \mu_2(\boldsymbol{a}, \boldsymbol{z})$ $p_{P_{\boldsymbol{\theta}^*}}(\boldsymbol{z}) d\boldsymbol{z}$ 为伪真实潜在结果，且 $\psi(\hat{} P_{\boldsymbol{\theta}}) = \mathbb{E}(Y(\boldsymbol{a})|\hat{}\boldsymbol{\theta}_n^*) = \frac{1}{n\mathcal{M}} \sum_{i=1}^{n} \sum_{l=1}^{\mathcal{M}} \mu_2(\boldsymbol{z}_i^{(l)}, a, \hat{}$ $\boldsymbol{\theta}_2^*)$。误差分解表明：

$$\|\psi(\hat{P}_{\boldsymbol{\theta}}) - \psi(P_0)\| \leq \underbrace{\|\psi(\hat{P}_{\boldsymbol{\theta}}) - \psi(P_{\boldsymbol{\theta}^*})\|}_{\text{estimation error}} + \underbrace{\|\psi(P_0) - \psi(P_{\boldsymbol{\theta}^*})\|}_{\text{misspecification error}}.$$

定理 1. （估计误差）假设假设*1–2*以及引理 *A1*和定理 *A1*（见补充材料 *A5*）中的条件均成立。则

$$\|\psi(\hat{P}_{\boldsymbol{\theta}}) - \psi(P_{\boldsymbol{\theta}^*})\| \xrightarrow{p} 0, \quad as\ \mathcal{M} \to \infty\ and\ n \to \infty.$$

备注 1. 如定理 *1* 的证明所示，估计量 *(9)* 的一致性源于真实稀疏 *StoNet* 的存在，以及$\hat{\boldsymbol{\theta}}_n^*$的一致性。值得注意的是，由于$\hat{\boldsymbol{\theta}}_n^*$ 并非唯一（参见备注 *A1*），插补得到的潜在混杂因素可能与其真实值不同。然而，$\pi(\boldsymbol{z}|\boldsymbol{A},\hat{\boldsymbol{\theta}}_1^*)$ 仍然是 $\pi(\boldsymbol{z}|\boldsymbol{A},\boldsymbol{\theta}_1^*)$的相合估计量（就密度函数而言），但$\hat{\boldsymbol{\theta}}_n^*$ 需经过损失不变变换。尽管如此，这并不影响估计量 *(9)* 的一致性，这是一个显著的性质。

定理2。（模型设定错误）固定一个处理水平 a。令 P_0 表示 $\boldsymbol{A}, \boldsymbol{Z}, \boldsymbol{Y}$ 的真实（完整数据）分布()，并令 $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta_n\}$ 表示在样本量为 n 时，由采用稀疏 *DNN* 架构的 *CI-StoNet* 工作模型所诱导的完整数据分布族。假设结构条件均值函数 $m_A(\boldsymbol{z}) = \mathbb{E}_{P_0}[\boldsymbol{A} \mid \boldsymbol{Z} = \boldsymbol{z}]$和 $m_Y(\boldsymbol{a}, \boldsymbol{z}) = \mathbb{E}_{P_0}[\boldsymbol{Y} \mid \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{Z} = \boldsymbol{z}]$ 满足假设 *3* 和假设*A10*，且其逼近率为 ω_n；此外，假设稀疏 *DNN* 架构满足假设 *A13*。进一步假设，对于任意固定的处理a，模型的结果回归一致有界：$\sup_{\boldsymbol{z}} |\mu_2(\boldsymbol{a}, \boldsymbol{z}; \boldsymbol{\theta}^*)| \le C_{\mu_2}$。

定义因果泛函 $\psi(P)$：$= \mathbb{E}_P[Y(\boldsymbol{a})]$。特别地，

$$\psi(P_0) = \int m_Y(\boldsymbol{a}, \boldsymbol{z})\, p_{P_0}(\boldsymbol{z})\, d\boldsymbol{z}, \qquad \psi(P_{\boldsymbol{\theta}}) = \int \mu_2(\boldsymbol{a}, \boldsymbol{z}; \boldsymbol{\theta})\, p_{P_{\boldsymbol{\theta}}}(\boldsymbol{z}) d\boldsymbol{z}.$$

于是存在常数 $C_1, C_2 < \infty$（与 n 无关），使得

$$KL(P_0, P_{\boldsymbol{\theta}^*}) \le C_1\, \omega_n^2, \qquad \|\psi(P_0) - \psi(P_{\boldsymbol{\theta}^*})\| \le C_2\, \omega_n.$$

补充材料 A5.5 给出了简单混杂情形（定理 A2）和基本代理情形（定理 A3）的有限样本误差分析。分析表明，当重叠性不佳 $P(\boldsymbol{A}|\boldsymbol{Z})$ 时，潜在结果的估计误差界会变差。

值得注意的是，CI-StoNet 通过使用稀疏深度神经网络对处理机制和结果机制进行近似，对混杂过程施加的结构限制相对宽松。与 Wang & Blei (2018) 不同，它不依赖于“无单一原因混杂因素”这一假设，因此能够在统一框架内兼容多原因混杂和单一原因混杂。此外，与 Louizos 等（2017）的变分自编码器方法不同，后者在使用大型神经网络时可能产生缺乏一致性保证的潜在表示；CI-StoNet 则利用稀疏深度学习理论和贝叶斯正则化，在指定的模型类中确保参数估计一致性，从而针对伪真参数提供具有理论依据的因果效应估计。

此外，我们注意到，算法 1 的步骤 (i) 中插补的潜变量不能用于因果效应估计，因为它可能包含与碰撞变量有关的信息。图3(a) 展示了这一概念，其中碰撞变量 $\boldsymbol{C}$ 同时受到 $\boldsymbol{A}$ 和 $\boldsymbol{Y}$ 的影响。在步骤 (i) 中，我们根据 $\boldsymbol{A}$ 和 $\boldsymbol{Y}$对 $\boldsymbol{Z}$ 进行插补。如果存在碰撞变量，插补的潜变量可能会在 $\boldsymbol{A}$ 和 $\boldsymbol{Y}$之间引入虚假关联，从而可能导致因果效应估计产生偏差。为缓解这一问题，我们专门从 $\pi(\boldsymbol{Z}|\boldsymbol{A}, \hat{\boldsymbol{\theta}}_1^*)$中插补潜变量，确保分析中不包含任何与碰撞变量相关的信息。

在第2.1节中，我们假设不存在中介变量，以便将 $\boldsymbol{Z}$ 清晰地解释为潜在混杂变量。然而，如果确实存在中介变量 $\boldsymbol{M}$，如图3(b)所示，插补的潜变量 $\boldsymbol{Z}$ 可能会无意中包含与 $\boldsymbol{M}$相关的信息。数学上，条件分布可表示为：

$$\pi(\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{Z}) = \int \pi(\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{Z}, \boldsymbol{M})\, \pi(\boldsymbol{M} \mid \boldsymbol{A})\, d\boldsymbol{M},$$

这表明，在未观测到 $\boldsymbol{M}$ 的情况下，其效应会被吸收到 $\boldsymbol{Z}$ 中，使二者

(a) 碰撞变量 (b) 中介变量

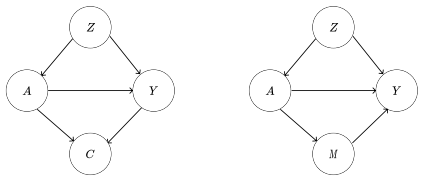


图3：因果结构的其他示例：(a) 存在碰撞变量，以 C 表示；(b) 存在中介变量，以 M表示。

在 CI-StoNet 框架中在统计上无法区分。在这种情况下，仍可依据（9）估计处理效应，其中 $\boldsymbol{Z}$ 作为潜在调整变量，有助于估计总因果效应。尽管这会排除通路特异性解释，但并不会使总因果效应的估计失效。不过，如果根据领域知识或实验设计已知 $\boldsymbol{M}$ 是中介变量，且 $\boldsymbol{A}$ 与 $\boldsymbol{M}$ 之间以及 $\boldsymbol{M}$ 与 $\boldsymbol{Y}$ 之间均不存在未测量混杂，则可应用前门准则（Pearl, 2009年）。在这种情况下，可将 $\boldsymbol{M}$ 纳入 CI-StoNet 的潜在混杂层，通过前门调整识别直接因果效应。

2.3 模拟研究

作为概念验证示例，我们通过一项模拟研究评估了 CI-StoNet，该研究针对 $\boldsymbol{A}$ 和 $\boldsymbol{Y}$ 采用非线性数据生成过程，并涵盖可分离混杂与不可分离混杂两种情形。我们将潜在混杂因素 $Z_1, \ldots, Z_6$ 生成为独立的标准高斯随机变量，然后使用逆累积分布函数从该分布中独立抽取 $A_1, \ldots, A_9$：

$$p(\boldsymbol{a}_i|\boldsymbol{Z}) = \frac{\operatorname{expit}(\xi(\boldsymbol{Z})\boldsymbol{a}_i)}{\int_{-1}^{1} \operatorname{expit}(\xi(\boldsymbol{Z})\boldsymbol{a}_i)} 1_{\{-1 \le \boldsymbol{a}_i \le 1\}}, \quad i = 1, \ldots, 9,$$

其中 $\operatorname{expit}(\xi(\boldsymbol{Z})\boldsymbol{a}_i) = \exp\{\xi(\boldsymbol{z})\boldsymbol{a}_i\}/1+\exp\{\xi(\boldsymbol{z})\boldsymbol{a}_i\}$，且 $\xi(\boldsymbol{Z}) = \sum_{i=1}^{2} \beta_i \sin z_i + \sum_{j=3}^{4} \beta_j \cos z_j + \sum_{k=5}^{6} 1+ {}^{1}\exp\{-\beta_k z_k + 0.5\}$。我们设定 $f_1(\boldsymbol{A}) = \boldsymbol{\theta}^T \boldsymbol{A}^{\otimes 2}$ 和 $f_2(\boldsymbol{A}) = \sum_{i<j} \boldsymbol{a}_i \boldsymbol{a}_j$，其中 $\boldsymbol{A}^{\otimes 2}$表示逐元素平方运算，并在两种设定下生成 Y：（i）可分离混杂。处理与混杂变量分别影响结果：$Y = f_1(\boldsymbol{A}) - \boldsymbol{\theta}_0 f_2(\boldsymbol{A}) + \xi(\boldsymbol{Z}) + \epsilon$，其中 $\boldsymbol{\theta}_0 \sim U(-1, 1)$ 且 $\epsilon \sim N(0, 1)$。（ii）不可分离混杂。处理与混杂变量之间存在交互作用：$Y = f_1(\boldsymbol{A}) - \xi(\boldsymbol{Z}) f_2(\boldsymbol{A}) + \xi(\boldsymbol{Z}) + \epsilon$，其中 $\epsilon \sim N(0, 1)$。

对于每种设置，实验均在 10 个模拟数据集上进行，每个数据集包含 1000 个训练样本、500 个验证样本和 500 个测试样本。边际处理效应使用测试集计算。图 S1 比较了这 10 个数据集上的真实边际处理效应与估计边际处理效应。图中显示，大多数估计的边际效应与真实边际效应之间的差异在半个标准差以内，这表明 CI-StoNet 能够以较小的偏差估计每种处理的边际效应。

3 用于代理变量的 CausalStoNet

在某些应用中，可以获得未观测混杂因素的代理变量，不过这些代理变量可能存在噪声，或只能提供部分信息。以这类代理变量为条件可以减少混杂偏差，但无法将其完全消除，因此很自然地会将它们用作缺失混杂变量的替代变量。Kuroki & Pearl（2014年）在关于测量误差的特定假设下，利用矩阵调整或谱方法建立了识别结果。Tchetgen 等人（2020年）和 Miao 等人（2018年）提出的近端因果推断框架表明，当观测到两类代理变量——处理和结果混杂代理变量——时，可以识别因果效应。相比之下，Louizos 等（2017年）提出了一种变分自编码器方法，利用基本代理变量学习潜在混杂因素表示。

考虑如图 4(a) 所示的带有基本代理变量的因果结构。该因果结构表明，$\boldsymbol{Z}$ 可以基于以下条件分布进行插补：

$$\pi(\boldsymbol{Z}|\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{X}) \propto \pi(\boldsymbol{Z})\pi(\boldsymbol{X}|\boldsymbol{Z})\pi(\boldsymbol{A}|\boldsymbol{Z})\pi(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{A}) \propto \pi(\boldsymbol{Z}|\boldsymbol{X})\pi(\boldsymbol{A}|\boldsymbol{Z})\pi(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{A}). \tag{10}$$

条件分布 (10) 的分解进一步表明 StoNet 结构：

$$\boldsymbol{Z} = \mu_1(\boldsymbol{X}, \boldsymbol{\theta}_1) + \boldsymbol{e}_z, \quad \boldsymbol{A} = \mu_2(\boldsymbol{Z}, \boldsymbol{\theta}_2, \boldsymbol{e}_a), \quad \boldsymbol{Y} = \mu_3(\boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\theta}_3) + \boldsymbol{e}_y, \tag{11}$$

其中 $\boldsymbol{e}_z$ 和 $\boldsymbol{e}_y$ 是高斯随机误差，而 $\boldsymbol{e}_a$ 的形式可根据处理类型确定。这些随机误差彼此独立，且

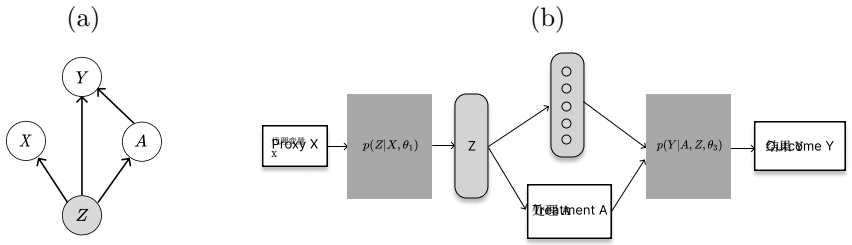


图 4：(a) 因果有向无环图：不依赖于代理变量；(b) 代理变量设定下 CI-StoNet 的示意图：白色矩形表示来自观测数据的变量；浅灰色圆角矩形表示隐藏神经元；深灰色矩形表示用于学习各自条件分布的网络模块。

也独立于 $\boldsymbol{X}$。图 4(b) 展示了相应的 CI-StoNet 结构。或者，我们可以考虑以下 StoNet 模型：

$$\boldsymbol{Z} = \mu_1(\boldsymbol{X}, \boldsymbol{\theta}_1) + \boldsymbol{e}_z, \quad \boldsymbol{A} = \mu_2(\boldsymbol{Z}, \boldsymbol{\theta}_2) + \boldsymbol{e}_a, \quad \boldsymbol{Y} = \mu_3(\boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\theta}_3) + \boldsymbol{e}_y, \tag{12}$$

其中 $\boldsymbol{e}_z$、$\boldsymbol{e}_a$ 和 $\boldsymbol{e}_y$ 为高斯随机误差。值得注意的是，模型 (11) 与模型 (12) 渐近等价，即使当 $\boldsymbol{A}$ 为二元向量时亦然。在二元情形下，二者的等价性由如下结果所支持：如 Liang (2003) 与 Duda 等人 (2001) 所示，$\mu_2(\boldsymbol{Z}, \boldsymbol{\theta}_2)$ 在 $n \to \infty$ 时收敛于概率函数 $P(\boldsymbol{A} = \mathbf{1}|\boldsymbol{Z}, \boldsymbol{\theta}_2)$。

为简化计算，本文采用模型 (12)。可以为该模型构造类似于 (7) 的梯度方程，并使用类似于算法 1 的自适应 SGHMC 算法求解。令 $\{\boldsymbol{z}_i^{(l)}: l = 1, 2, \ldots\ldots, \mathcal{M}\}$ 表示从 $\pi(\boldsymbol{z}|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}_1^*)$ 模拟得到的样本。随后，可通过蒙特卡洛平均估计期望结果函数 $\mathbb{E}(Y(\boldsymbol{a}))$，即

$$\widehat{\mathbb{E}(Y(\boldsymbol{a})|\hat{\boldsymbol{\theta}}_3^*)} = \frac{1}{n\mathcal{M}} \sum_{i=1}^{n} \sum_{l=1}^{\mathcal{M}} \mu_3(\boldsymbol{z}_i^{(l)}, \boldsymbol{a}, \hat{\boldsymbol{\theta}}_3^*). \tag{13}$$

3.1 数值实验

为简单起见，我们在实验中考虑单一二元处理。我们将 CI-StoNet 与以下基线方法进行比较：

(i) 针对平均处理效应（ATE）设计的方法：双重选择估计量（**DSE**）（Belloni 等人，2014年）、近似残差平衡估计量（**ARBE**）（Athey 等人，2018年）、目标最大似然估计量（**TMLE**）（van der Laan 和 Rubin，2006年），以及针对无混杂处理的深度正交网络（**DONUT**）（Hatt 和 Feuerriegel，2021年）。

(ii) 为异质性处理效应设计：**X-学习器** (Künzel 等人, 2017年)、**Drag-onnet**(Shi 等人, 2019年)、因果多任务深度集成 (**CMDE**) (Jiang 等人, 2023年))、因果多任务高斯过程 (**CMGP** (Alaa 和 van der Schaar, 2017年))、因果效应变分自编码器 (**CEVAE**) (Louizos 等人, 2017年)、生成对抗网络 (**GANITE**) (Yoon 等人, 2018年) 以及反事实回归网络 (**CFRNet**(Shalit 等人, 2017年))。对于第 (ii) 部分的基线方法，我们使用 Jiang 等人 (2023年) 在 GitHub 上的代码。
在性能评估中，我们考虑两项指标：(i) 平均处理效应（ATE）的估计准确度，以 ATE 估计值的平均绝对误差（MAE）衡量；(ii) 条件平均处理效应（CATE）的估计准确度，以异质性效应估计精度（PEHE）衡量。

3.1.1 模拟示例

本例旨在比较非线性处理效应和非线性结果函数问题中的各种方法。我们按照 A1.2 节所述的流程生成了 10 个数据集，每个数据集包含 2000 个训练样本、500 个验证样本和 500 个测试样本。表 1 显示，CI-StoNet 能够准确估计异质性处理效应，并且优于基线方法。

表 1：使用代理变量估计异质性处理效应的不同方法比较。其中，PEHE 基于 10 个数据集计算；“样本内 PEHE”基于训练样本和验证样本计算；“样本外 PEHE”基于测试样本计算。

	样本内 PEHE	样本外 PEHE
CI-StoNet	**0.3614(0.0328)**	**0.3731(0.0350)**
CMDE	0.9019(0.0746)	0.9059 (0.0699)
CMGP	1.8823(0.0836)	2.2116 (0.1682)
CEVAE	0.6190(0.0350)	0.6246 (0.0384)
GANITE	1.2099(0.0558)	1.1797 (0.0499)
X-learner-RF	0.8308(0.0200)	1.4272 (0.0132)
X-learner-Bart	0.6489(0.0168)	0.6570 (0.0151)
CFRNet-Wass	1.7127(0.1668)	1.7258 (0.1667)
CFRNet-MMD	2.0238(0.0537)	2.0250 (0.0582)
DragonNet	0.4217(0.0356)	0.4305 (0.0361)

3.1.2 基准数据集

我们在一些基准数据集上评估了 CI-StoNet，包括 Twins 数据集以及来自 2019年大西洋因果推断会议（ACIC）数据挑战赛的 10 个数据集。A1.3 节报告的结果表明，CI-StoNet 优于基线方法。

4 结论

本文将 StoNet 与自适应随机梯度马尔可夫链蒙特卡罗相结合，构建了一个实用且具有坚实理论基础的缺失混杂因素因果推断框架。CI-StoNet 对因果有向无环图所蕴含的依赖结构进行编码，并通过带有贝叶斯正则化的稀疏深度学习估计参数。尽管潜在混杂因素只能在损失不变变换意义下实现可识别，但在稀疏 StoNet 类中，由此诱导的因果效应是定义明确的。该框架可自然扩展至存在多个原因和代理变量的情形。

在这一框架下，因果泛函在模型意义下是可识别的：观测上等价的参数化会得到相同的因果效应。此外，相对于伪真参数，总误差可分解为统计估计误差和模型设定错误，从而能够对收敛性和模型近似进行明确刻画。

尽管具有上述优势，本研究仍存在一些局限。首先，CI-StoNet 的结构和参数估计依赖于对底层因果有向无环图的正确设定。例如，在存在多重处理的情形下，未被识别的中介变量可能会被吸收到学习得到的替代混杂因素中，从而潜在地导致因果效应估计产生偏差，因为该模型是通过 $(\boldsymbol{A}、\boldsymbol{Y}、\boldsymbol{Z})$ 的联合分布定义的。不过，如果能够正确识别中介变量，则可以相应地修改 CI-StoNet 结构，以避免这一问题。其次，当前的 CI-StoNet 形式并未为因果效应估量提供原则性且基于模型的不确定性量化。尽管该方法本身无法为因果泛函提供有效的后验区间，我们在附录 A4 中提供了一种实用的后处理程序，用于为估计得到的因果效应构建自助法置信区间。通过采用原始的 StoNet 框架（Liang 等人，2022年），可以实现更为形式化的不确定性量化；该框架通过其完整的贝叶斯形式实现推断。

最后，CI-StoNet 的马尔可夫结构为建模范围广泛的因果结构提供了很大的灵活性。A2 节将 CI-StoNet 扩展到了两种代理变量情形：（i）结果取决于代理变量；（ii）处理取决于代理变量。详情请参见该节。

参考文献

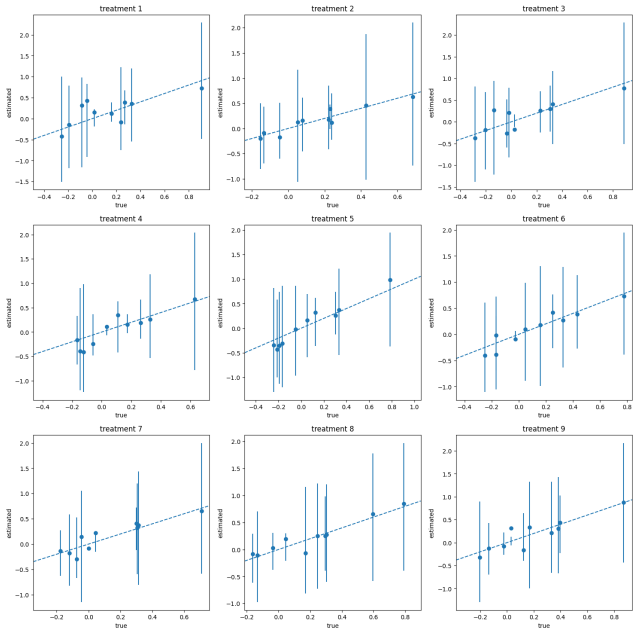
Ahmed M. Alaa 和 Mihaela van der Schaar。使用多任务高斯过程对个体化治疗效应进行贝叶斯推断。神经信息处理系统会议，2017年。URL https://api.semanticscholar.org/CorpusID:15589829。Susan Athey、Guido W. Imbens 和 Stefan Wager。近似残差平衡：高维情形下平均处理效应的去偏推断。皇家统计学会会刊*B*辑：统计方法学，80(4):597–623，2018年。Alexandre Belloni、Victor Chernozhukov 和 Christian Hansen。对高维控制变量中进行选择后的处理效应进行推断。经济研究评论，81: 608–650，2014年。Helmut Bölcskei、Philipp Grohs、Gitta Kutyniok 和 Philipp Petersen。使用稀疏连接深度神经网络进行最优逼近。*CoRR*，abs/1705.01714，2019年。Helmut Bolcskei、Philipp Grohs、Gitta Kutyniok 和 Philipp Petersen。使用稀疏连接深度神经网络进行最优逼近。*SIAM*数据科学数学杂志，1(1):8–45，2019年。Wei Deng、Xiao Zhang、Faming Liang 和 Guang Lin。用于稀疏深度学习的自适应经验贝叶斯方法。神经信息处理系统进展，2019: 5563–5573，2019年。Richard O. Duda、Peter E. Hart 和 David G. Stork。模式分类，第*2*版。威利，纽约，2001年。URL https://api.semanticscholar.org/CorpusID:361680。Tobias Hatt 和 Stefan Feuerriegel。通过正交正则化估计平均处理效应。载于*ACM*信息与知识管理国际会议第*30*届会议论文集，第680–689页，2021年。Kosuke Imai 和 Zhichao Jiang。评论：多个原因带来的挑战。*Journal ofthe American Statistical Association*，2019年。Ziyang Jiang、Zhuoran Hou、Yiling Liu、Yiman Ren、Keyu Li 和 David Carlson。使用多任务深度集成估计因果效应。载于第*40*届国际机器学习会议（*ICML*）论文集，PMLR，第680–689页，2023年。Nathan Kallus、Xiaojie Mao 和 Madeleine Udell。通过矩阵分解处理含噪与缺失协变量的因果推断。神经信息处理系统进展，2018年。Sören R. Künzel、Jasjeet S. Sekhon、Peter J. Bickel 和 Bin Yu。使用机器学习估计异质性处理效应的元学习器。美国国家科学院院刊，116:4156 – 4165，2017年。URL https://api.semanticscholar.org/CorpusID:73455742。Manabu Kuroki 和 Judea Pearl。因果推断中的测量偏差与效应恢复。*Biometrika*，101，2014年。F. Liang、B. Jia、J. Xue、Q. Li 和 Y. Luo。针对高维缺失数据问题及其他问题的插补正则化优化算法。皇家统计学会会刊*B*辑，80(5):899–926，2018a。F. Liang、Q. Li 和 L. Zhou。用于筛选药物敏感基因的贝叶斯神经网络。*Journal of the American Statistical Association*，113:955–972，2018b。Faming Liang。一种有效的贝叶斯神经网络分类器及其与支持向量机的比较研究。神经计算，15:1959–1989，2003年。URL https://api.semanticscholar.org/CorpusID:14168018。Siqi Liang、Yan Sun 和 Faming Liang。使用随机神经网络进行非线性充分降维。第*36*届神经信息处理系统会议（*NeurIPS 2022*），2022年。

Christos Louizos、Uri Shalit、Joris M. Mooij、David A. Sontag、Richard S. Zemel 与 Max Welling。基于深度潜变量模型的因果效应推断。见神经信息处理系统，2017 年。URL https://api.semanticscholar.org/CorpusID:260564。Wang Miao、Zhi Geng 与 Eric J. Tchetgen Tchetgen。利用未测量混杂因素的代理变量识别因果效应。*Biometrika*，105，2018年。S.F. Nielsen。随机 EM 算法：估计与渐近结果。*Bernoulli*，6：457–489，2000年。Judea Pearl。统计学中的因果推断：综述。*Statistics Surveys*，3：96–146，2009年1月。doi: 10.1214/09-SS057。Philipp Petersen 与 Felix Voigtlaender。使用深度 ReLU 神经网络对分段光滑函数的最优逼近。神经网络，108：296–330，2018年。Nicholas G. Polson 与 Veronika Ročková。稀疏深度学习的后验集中。见第*32*届神经信息处理系统国际会议，NIPS'18，第938–949页，美国纽约州雷德胡克，2018年。Curran Associates Inc.。Severi Rissanen 与 Pekka Marttinen。对使用深度潜变量模型进行因果估计的一致性的批判性审视。神经信息处理系统进展，2021年。Paul R. Rosenbaum 与 Donald B. Rubin。倾向得分在因果效应观察性研究中的核心作用。*Biometrika*，70：41–55，1983年。URL https://api.semanticscholar.org/CorpusID:49190930。Donald B. Rubin。在随机与非随机研究中估计处理的因果效应。*Journal of Educational Psychology*，66：688–701，1974年。URL https://api.semanticscholar.org/CorpusID:52832751。Johannes Schmidt-Hieber。使用带 ReLU 激活函数的深度神经网络进行非参数回归。*arXiv:1708.06633*，2017年。Uri Shalit、Fredrik D. Johansson 与 David Sontag。估计个体治疗效应：泛化界与算法。载于 Doina Precup 与 Yee Whye Teh（编），第*34*届国际机器学习会议（*ICML*）论文集，*PMLR*第6卷，第4709–4718页，国际会议中心，澳大利亚悉尼，2017年8月6–11日。Claudia Shi、David M. Blei 与 Victor Veitch。调整神经网络以估计治疗效应。见*NeurIPS*，2019年。Q. Song、Y. Sun、M. Ye 与 F. Liang。面向大规模贝叶斯变量选择的扩展随机梯度马尔可夫链蒙特卡罗。*Biometrika*，107(4)：997–1004，2020年。Y. Sun、Q. Song 与 F. Liang。一致稀疏深度学习：理论与计算。*Journal of the American Statistical Association*，117(540)：1981–1995，2022年。Yan Sun 与 Faming Liang。核展开随机神经网络。*Journal of theRoyal Statistical Society Series B*，84(2)：547–578，2022年。Eric J Tchetgen Tchetgen、Andrew Ying、Yifan Cui、Xu Shi 与 Wang Miao。近端因果学习导论。*arXiv* 预印本，*arXiv:2009.10982*，2020年。Mark J. van der Laan 与 Daniel Rubin。目标最大似然学习。*TheInternational Journal of Biostatistics*，2(1)：11，2006年。Yixin Wang 与 David M. Blei。多重原因的福祉。*Journal of the AmericanStatistical Association*，114：1574 – 1596，2018年。URL https://api.semanticscholar.org/CorpusID:21694910。Jinsung Yoon、James Jordon 与 Mihaela van der Schaar。GANITE：使用生成对抗网络估计个体化治疗效应。见国际学习表征会议，2018年。URL https://api.semanticscholar.org/CorpusID:65516833。

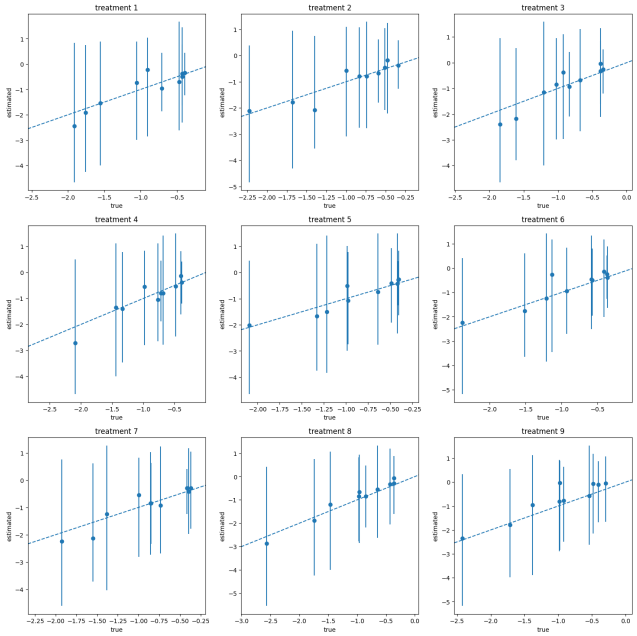
A 附录

A1 补充示例

A1.1 第2.3节模拟研究的图



（a）可分离混杂



（b）不可分离混杂

Figu图 S1：模拟研究中的 CI-StoNet 结果。点表示真实的边际效应边际效
量和误差条表示边际效应估计量的一个标准误。

A1.2 模拟示例

这些模拟示例用于比较不同方法在处理效应和结果函数均为非线性的情况下的性能。我们按照以下流程生成了十个数据集，每个数据集包含 2000 个训练样本、500 个验证样本和 500 个测试样本。

1. 从 $N(0, 1)$ 中独立生成混杂变量 $\boldsymbol{z}_i = (z_{i,}1, \cdots z_{i,}5)$。2. 从区间$[-10, 10]$内截断的 $N(\mu_i, 1)$ 中独立生成 $\gamma_{i,}r_{i,}1, \cdots r_{i,}100$，其中 $\mu_i = \frac{1}{5}\sum_{k=1}^{5} z_{i,k}$。设置代理变量 $\boldsymbol{x}_i = (x_{i,}1, \ldots, x_{i,}100)$，其中$x_{i,j} = \frac{\gamma_i + r_i}{\sqrt{2}}$，且 $\boldsymbol{x}$ 和 $\boldsymbol{z}$ 通过 μ相关。3. 倾向得分 $p(\boldsymbol{z}_i) = \frac{1}{4}(1 + \beta_{2,}4(\frac{1}{3}(\Phi(z_{i,}1) + \Phi(z_{i,}3) + \Phi(z_{i,}5))))$，其中 $\beta_{2,}4$ 是形状参数为 (2, 4) 的贝塔分布的累积分布函数，Φ 表示标准正态分布的累积分布函数。这确保了 $p(\boldsymbol{z}_i) \in [0.25, 0.5]$，从而提供了充分的重叠性。因此，处理 A_i 由成功概率为 $p(\boldsymbol{z}_i)$ 的伯努利分布生成。我们对处理组和对照组进行重采样，以确保数据集中处理组和对照组的样本数量均衡。

4. 为模拟结果，我们设定

$$\boldsymbol{y}_i = c(\boldsymbol{z}_i) + (\tau + \eta(\boldsymbol{z}_i))A_i + \sigma_y e_i,$$
$$c(\boldsymbol{z}_i) = \frac{5z_{i3}}{1 + z_{i4}^2} + 2z_{i5},$$

其中 $\eta(\boldsymbol{z}_i) = f(z_{i1})f(z_{i2}) - E(f(z_{i1})f(z_{i2}))$，且 $f(w) = \frac{2}{1+\exp(-w+0.5))}$。也就是说，我们将处理效应设定为 $\tau(\boldsymbol{z}_i) = \tau + \eta(\boldsymbol{z}_i)$，该效应对不同个体是同质的。我们在设定 $\tau = 3$、$\sigma_y = 0.25$ 和 $e_i \sim N(0, 1)$ 下生成了样本。

A1.3 基准数据集

我们在若干基准数据集上比较所提方法的性能，包括 Twins 数据集以及来自大西洋因果推断会议（ACIC）2019年数据挑战赛的 10 个数据集。

ACIC 2019 数据集 我们首先处理了 10 个 ACIC 2019 数据集。本实验重点比较 CI-StoNet 与专为平均处理效应估计设计的基线方法。结果汇总于表 S1，表明 CI-StoNet 优于这些基线方法。

表 S1：在 10 个 ACIC 2019 数据集上的平均处理效应估计，括号中的数值表示平均绝对误差的标准差，并包含额外的基准结果

方法	样本内	样本外
CI-StoNet	**0.0669 (0.0166)**	**0.0709 (0.0133)**
CMDE	0.0802 (0.0166)	0.0877 (0.0246)
CMGP	0.1252 (0.0156)	0.1349 (0.0170)
CEVAE	0.0773 (0.0152)	0.0875 (0.0154)
GANITE	0.1622 (0.0390)	0.1747 (0.0425)
X-learner-RF	0.1720 (0.0257)	0.1903(0.0253)
X-learner-Bart	0.0738 (0.0251)	0.0817(0.0248)
CFRNet-Wass	0.1024 (0.0241)	0.1099(0.0256)
CFRNet-MMD	0.1105 (0.0258)	0.1208(0.0246)

双胞胎数据。我们分析了美国1989年至1991年的双胞胎出生真实世界数据集。处理变量为二元变量，其中“1”表示出生时体重较重的双胞胎。

该数据集包含46个变量，包括临床信息和父母的社会经济地位，我们将其视为潜在混杂因素的代理变量。结果变量为二元变量，其中“1”表示双胞胎在出生后第一年内死亡。我们将每对双胞胎的记录视为潜在结果，从而能够得出真实的平均处理效应。经过数据预处理后，我们得到一个包含4,821个样本的数据集。在这个最终数据集中，体重较轻和较重的双胞胎死亡率分别为16.9%和14.42%，真实平均处理效应为 -2.48%。

我们采用三折交叉验证进行实验：将数据集划分为三个子集，使用其中两个子集训练模型，并用剩余的一个子集估计平均处理效应。表 S2（左栏）报告了三折的平均处理效应均值及其标准差。与基线方法相比，CI-StoNet 得出的平均处理效应估计（以均方根误差衡量）更加稳定。

表 S2：不同方法在双胞胎数据上估计平均处理效应（ATE）的比较。括号中的数字表示 ATE 绝对误差的标准差，RMSE 表示均方根误差。

	含混杂变量$gestat10$		缺失混杂变量 $gestat10$	
方法	ATE 绝对误差	RMSE	ATE 绝对误差	RMSE
CI-StoNet	0.0099(0.0089)	**0.0133**	**0.0135(0.0071)**	**0.0153**
DSE	0.0157(0.0176)	0.0236	0.0211(0.0193)	0.0286
ARBE	0.0152 (0.0201)	0.0252	0.0168(0.0257)	0.0307
目标最大似然估计(Lasso)	0.0855 (0.0599)	0.1044	0.0932(0.0791)	0.1222
目标最大似然估计（集成）	0.1042 (0.0779)	0.1301	0.1238(0.0607)	0.1379
DONUT	0.0490 (0.0128)	0.0506	0.0490(0.0124)	0.0505
CMDE	0.0108(0.0905)	0.0911	0.0635(0.0905)	0.1106
CEVAE	0.0249(**0.0002**)	0.0249	0.0327(0.0633)	0.0712
GANITE	0.3519 (0.1533)	0.3838	0.4198(0.2278)	0.4776
X-learner-RF	**0.0056** (0.0257)	0.0252	0.0157(0.0257)	0.0301
X-learner-Bart	0.0194 (0.0192)	0.0273	0.0251(0.0312)	0.0400
CFRNet-Wass	0.0189 (0.0425)	0.0465	0.0211(0.0254)	0.0330
CFRNet-MMD	0.0439 (0.0146)	0.0463	0.0619(0.0158)	0.0639

最后，为了更有力地证明所提出的方法在存在缺失混杂因素时仍表现良好，我们进行了实验，故意省略一个重要的混杂变量 $gestat10$（孕 gestational age）。在对数据集进行预处理时，我们遵循 Louizos 等（2017）的做法，聚焦于出生体重低于 2 kg 的同性别双胞胎对，并使用变量 $gestat10$ 生成“伪处理分配”。由于$gestat10$ 也是影响新生儿死亡率的重要因素，因此它是一个重要的混杂变量。我们从数据集中移除了$gestat10$。表 S2（右侧面板）的结果表明，CI-StoNet 在存在缺失混杂因素时表现稳健。在这种情形下，它在 ATE 绝对误差和均方根误差两项指标上均优于所有基线方法，说明 CI-StoNet 借助潜在混杂因素插补获得了优越性能。

A2 扩展至其他因果结构

CI-StoNet 中嵌入的马尔可夫结构使其能够灵活地对各种因果结构进行建模。在本节中，我们将 CI-StoNet 扩展为能够处理涉及代理变量的其他因果结构。具体而言，我们考虑两种情形：结果取决于代理变量，以及处理取决于代理变量。

A2.1 结果取决于代理变量

当结果依赖于代理变量时，见图 S2(a)， $\boldsymbol{Z}$ 的插补基于以下分解：

$$\pi(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{Y},\boldsymbol{X}) \propto \pi(\boldsymbol{Z})\pi(\boldsymbol{X}|\boldsymbol{Z})\pi(\boldsymbol{A}|\boldsymbol{Z})\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{A},\boldsymbol{X}) \propto \pi(\boldsymbol{Z}|\boldsymbol{X})\pi(\boldsymbol{A}|\boldsymbol{Z})\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{A},\boldsymbol{X}).$$

相应地，CI-StoNet 的结构可安排如下：

$$\begin{aligned} \boldsymbol{Z} &= \mu_1(\boldsymbol{X}, \boldsymbol{\theta}_1) + \boldsymbol{e}_z, \\ \boldsymbol{A} &= \mu_2(\boldsymbol{Z}, \boldsymbol{\theta}_2) + \boldsymbol{e}_a, \\ \boldsymbol{Y} &= \mu_3(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\theta}_3) + \boldsymbol{e}_y, \end{aligned} \tag{A1}$$

其中 $\boldsymbol{e}_z$、$\boldsymbol{e}_a$ 和 $\boldsymbol{e}_y$ 表示高斯随机误差。相应的示意图见图 S2(b)。

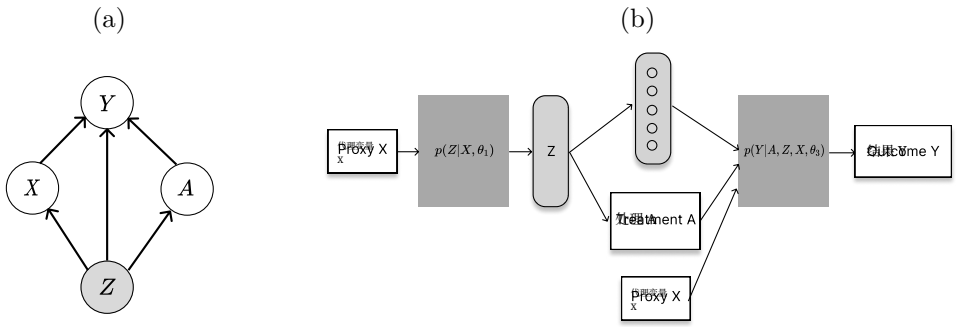


图 S2：(a) 因果结构：结果取决于代理变量；(b) 结果取决于代理变量这一情形下的 CI-StoNet 结构。

A2.2 取决于代理变量的处理

当处理取决于代理变量时，见图 S3(a)，$\boldsymbol{Z}$ 的插补基于以下分解：

$$\pi(\boldsymbol{Z}|\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{X}) \propto \pi(\boldsymbol{Z})\pi(\boldsymbol{X}|\boldsymbol{Z})\pi(\boldsymbol{A}|\boldsymbol{Z}, \boldsymbol{X})\pi(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{A}) \propto \pi(\boldsymbol{Z}|\boldsymbol{X})\pi(\boldsymbol{A}|\boldsymbol{Z}, \boldsymbol{X})\pi(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{A}).$$

CI-StoNet 的结构可安排如下：

$$\begin{aligned} \boldsymbol{Z} &= \mu_1(\boldsymbol{X}, \boldsymbol{\theta}_1) + \boldsymbol{e}_z, \\ \boldsymbol{A} &= \mu_2(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\theta}_2) + \boldsymbol{e}_a, \\ \boldsymbol{Y} &= \mu_3(\boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\theta}_3) + \boldsymbol{e}_y, \end{aligned} \tag{A2}$$

其中，$\boldsymbol{e}_z$、$\boldsymbol{e}_a$ 和 $\boldsymbol{e}_y$ 表示高斯随机误差。相应的示意图见图 S3(b)。

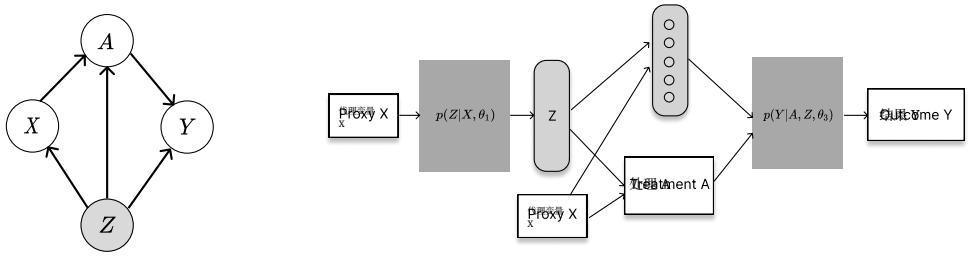


图 S3：当处理取决于代理变量时的 (a) 因果结构和 (b) CI-StoNet 结构。

这两种模型均可使用自适应随机梯度马尔可夫链蒙特卡罗算法进行训练，并可根据$\pi(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}_1^*)$中的插补混杂因素估计相应的因果效应。

对于图 S2(a) 和 S3(a) 所示的因果结构，X 是代理变量，Z 表示缺失混杂变量，而 A 和 Y 分别表示处理变量和结果变量。白色节点表示观测变量，浅灰色节点表示未观测变量。对于图 S2(b) 和 S3(b) 所示的 CI-StoNet 结构，白色矩形表示观测数据中的变量；浅灰色圆角矩形表示隐藏神经元；深灰色矩形表示用于学习相应条件分布的网络模块。

A3 DAG 误设实验

我们进行了一项实验，其中真实的数据生成过程包含处理依赖型代理变量和结果依赖型代理，但我们使用基本代理变量拟合数据。数据生成过程如下

1. 混杂变量：$Z = (z_1, z_2, z_3, z_4, z_5) \sim N(0, I_5)$ 2. 代理变量：对于 $j \in \{1. \ldots, 50\}$，$X_j = h_1(Z) + \epsilon_X$，其中 $h_1(Z) = \sin(z_1) + 0.5z_2^2 + 0.3(z_3 + z_4 + z_5)$，$\epsilon_X \sim N(0, 0.5I_{50})$ 3. 处理：令 $h_2(Z) = 0.7z_1 + 0.3z_2 - 0.2z_3$, $g(X) = \frac{1}{d_X}\sum_{j=1}^{d_X} |X_j|$。对于结果依赖型代理和基本代理变量，$A \sim \text{Bernoulli}(\text{expit}(h_2(Z)))$。对于处理依赖型代理变量，$A \sim \text{Bernoulli}(\text{expit}(h_2(Z) + 0.5g(X)))$。4. 结果：令 $f(Z) = z_1^2 + 0.5z_2z_3$, $\tau(Z) = 3 + 0.5\sin(z_1)$，$q(X) = \frac{1}{d_X}\sum_{j=1}^{d_X} |X_j|$。对于处理依赖型代理变量和基本代理变量，$Y = f(Z) + A \cdot \tau(Z) + \epsilon_Y$。对于结果依赖型代理，$Y = f(Z) + A \cdot \tau(Z) + \gamma q(X) + \epsilon_Y$。$\epsilon_Y \sim N(0, 0.52)$。

真实平均处理效应为3。对于每种情景，生成10个模拟数据集，每个数据集包含2000年训练样本、500个验证样本和500个测试样本。我们使用基本代理结构（公式15）对全部三种情景进行建模，其中基本代理变量情景提供正确的基线，而结果依赖情景和处理依赖情景则用于展示 DAG 误设。样本内平均绝对误差在训练集和验证集上计算，样本外平均绝对误差在测试集上计算。

表S3：DAG 误设下平均处理效应估计的平均绝对误差

	样本内平均绝对误差	样本外平均绝对误差
基本代理变量	0.0682 (0.0193)	0.0702 (0.0207)
依赖结果的代理变量	0.0705 (0.0188)	0.0976 (0.0236)
依赖处理的代理变量	0.0911 (0.0156)	0.1239 (0.0178)

A4 诊断与不确定性量化

潜在重叠的诊断 虽然 Z 是潜在的，我们仍可通过从其后验分布中抽样来进行稳健性诊断。我们展示在简单混杂设定下对潜在重叠进行诊断的步骤。给定重叠阈值 $0 < \alpha < 1$以及倾向评分模型$\hat{e}(A|Z)$，一个简单的压力测试为：

1. For each unit i, draw $\boldsymbol{Z}_i^{(b)} \sim p_{\hat{\boldsymbol{\theta}}}(\boldsymbol{Z}_i|\boldsymbol{A}_i, \boldsymbol{Y}_i)$, b=1, …, B
2. 对于每一次抽样，计算倾向评分以及重叠性的汇总统计量

$$S_\alpha^{(b)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{e}_i^{(b)} < \alpha \text{ or } > 1 - \alpha\}$$

3. 将均值 $S_\alpha = \frac{1}{B}\sum_b S_\alpha^{(b)}$报告为诊断结果。

该诊断并不能“验证”潜在重叠假设，因为从理论上讲，这些假设无法通过检验加以验证；但它可以让从业者在 Z 的后验抽样下，检测重叠性的违背情况或脆弱性，其方式类似于在协变量完全观测时评估重叠性。

自助法置信区间 正如插补正则化优化论文（Liang 等人，2018a）中所述，“对插补 Z进行蒙特卡洛平均”所产生的不确定性，仅反映缺失数据（即 Z）所导致的不确定性，并不等同于因果效应本身的不确定性。要估计因果效应的不确定性，还需要额外的方法，例如自助法，或计算观测信息矩阵的方法。

要推导自助法置信区间，请考虑以下计算开销较低的自助法流程：

1. 在完整数据集上拟合 CI-StoNet，并保存最终剪枝后的参数。参数 $\hat{\boldsymbol{\theta}}$ 将保存为所有自助重复样本的热启动。2. 抽取大小为 n 的自助样本，即有放回地从 1 到 n 中独立同分布均匀抽取索引$\{i_1^b, \ldots, i_n^b\} \sim$，然后构建自助数据集 $D^b = \{(A_{i_j^b}, X_{i_j^b}, Y_{i_j^b})\}_{j=1}^n$。3. 使用热启动初始化模型。通过复制 $\hat{\boldsymbol{\theta}}$来初始化该次重复的模型参数。4. 运行一个简短的随机梯度下降 + 随机梯度哈密顿蒙特卡洛插补阶段，使参数能够调整到D^b。5. 计算自助法 ATE 估计值。使用训练后的参数 $\hat{\boldsymbol{\theta}}^b$，计算 $\tau^b = \hat{E}^b[Y(1)] - \hat{E}^b[Y(0)]$。

对于 DAG 模型设定错误实验中的基本代理变量情形（详情见 A3），对每个生成的数据集进行 100 次此类自助法迭代。请注意，平均处理效应的真实值为 3。

表 S4：基本代理变量设定下的自助法界限

数据生成过程 随机种子	$\hat{\tau}$	$L_{\text{自助法}}$	$U_{\text{自助法}}$
0	3.1365	3.0897	3.1581
1	3.0057	2.9758	3.0574
2	3.0074	2.9836	3.0720
3	3.1582	3.1066	3.2113
4	2.9718	2.9609	3.0404
5	3.0123	2.9774	3.0801
6	3.0515	3.0263	3.1113
7	3.0556	3.0128	3.0904
8	3.1846	3.1181	3.2119
9	3.0241	2.9787	3.0471

A5 理论证明

估计量 (9) 的一致性可通过几个步骤建立。首先，我们证明通过算法 1 得到的估计量 $\boldsymbol{\theta}^{(k)} = \{\boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)}\}$ 依概率收敛到 (7) 的一个解，记为 $\hat{\boldsymbol{\theta}}_n^*$（见引理 A1）。随后讨论如何处理 $\hat{\boldsymbol{\theta}}_n^*$ 的非唯一性。接下来，我们证明 $\hat{\boldsymbol{\theta}}_n^*$ 是 $\boldsymbol{\theta}^*$ 的相合估计量，其中 $\boldsymbol{\theta}^*$ 是 (4) 中定义的稀疏 StoNet 的真实参数向量（见定理 A1）。在此结果的基础上，我们建立估计量 (9) 的一致性（见定理 1）。以下给出这些结果。

A5.1 收敛性 of $\boldsymbol{\theta}^{(k)}$

使用 IRO 算法训练 CI-StoNet 时，每次迭代都需要使用完整数据集，这使得该算法难以扩展到大规模神经网络。在

相比之下，自适应 SGHMC 算法可以在参数更新时使用小批量数据。正如 Liang et al. (2022) 所示，自适应 SGHMC 算法在以下条件下求解方程（7）。

记号：令 $\boldsymbol{D}$ 表示包含 n 个观测的数据集，令 D_i 表示 $\boldsymbol{D}$ 中的第 i 个观测。对于 StoNet，D_i 同时包含该观测的输入变量和输出变量。对于 CI-StoNet，D_i 包含处理和结果，即 $D_i = \{\boldsymbol{A}_i, \boldsymbol{Y}_i\}$。为简化记号，我们将与 D_i 对应的潜变量重新记为 Z_i，并用 $f_{D_i}(Z_i, \boldsymbol{\theta})$ 表示 $= -\log \pi(Z_i|D_i, \boldsymbol{\theta})$，即 Z_i 的负对数密度函数。令 $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$，令 $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$ 为 $\boldsymbol{Z}$ 的一个实现，令 $F_{\boldsymbol{D}}(\boldsymbol{Z}, \boldsymbol{\theta})$ 表示 $= \sum_{i=1}^{n} f_{D_i}(Z_i, \boldsymbol{\theta})$，并令 $H(\boldsymbol{Z}, \boldsymbol{\theta})$ 表示 $= \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{Z}|\boldsymbol{A}, \boldsymbol{\theta})$。为研究算法 1 中所述自适应 SGHMC 算法的收敛性，我们作出以下假设：

假设 A4. *(i)*（有界性）函数 $F_{\boldsymbol{D}}(\cdot,\cdot)$ 取非负实数值，且存在常数 $A, B \geq 0$，使得 $|F_{\boldsymbol{D}}(0, \boldsymbol{\theta}^*)| \leq A$, $\|\nabla_{\boldsymbol{Z}} F_{\boldsymbol{D}}(0, \boldsymbol{\theta}^*)\| \leq B$, $\|\nabla_{\boldsymbol{\theta}} F_{\boldsymbol{D}}(0, \boldsymbol{\theta}^*)\| \leq B$，以及 $\|H(0, \boldsymbol{\theta}^*)\| \leq B$。

(ii)（光滑性）$F_{\boldsymbol{D}}(\cdot,\cdot)$ 是 M-光滑的，且 $H(\cdot,\cdot)$ 是 M-利普希茨的：存在某个常数 $M > 0$，使得对任意 $\boldsymbol{Z}, \boldsymbol{Z}' \in \mathbb{R}^{d_z}$ 以及任意 $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$，

$$\|\nabla_{\boldsymbol{Z}} F_{\boldsymbol{D}}(\boldsymbol{Z}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{Z}} F_{\boldsymbol{D}}(\boldsymbol{Z}', \boldsymbol{\theta}')\| \leq M\|\boldsymbol{Z} - \boldsymbol{Z}'\| + M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|,$$
$$\|\nabla_{\boldsymbol{\theta}} F_{\boldsymbol{D}}(\boldsymbol{Z}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} F_{\boldsymbol{D}}(\boldsymbol{Z}', \boldsymbol{\theta}')\| \leq M\|\boldsymbol{Z} - \boldsymbol{Z}'\| + M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|,$$
$$\|H(\boldsymbol{Z}, \boldsymbol{\theta}) - H(\boldsymbol{Z}', \boldsymbol{\theta}')\| \leq M\|\boldsymbol{Z} - \boldsymbol{Z}'\| + M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

(iii)（耗散性）对任意 $\boldsymbol{\theta} \in \Theta$，函数 $F_{\boldsymbol{D}}(\cdot\ \boldsymbol{\theta}^*)$ 是 (m, b)-耗散的：存在某些常数 $m > \frac{1}{2}$ 和 $b \geq 0$，使得 $\langle \boldsymbol{Z}, \nabla_{\boldsymbol{Z}} F_{\boldsymbol{D}}(\boldsymbol{Z}, \boldsymbol{\theta}^*)\rangle \geq m\|\boldsymbol{Z}\|^2 - b$。

(iv)（梯度噪声）存在常数 $\varsigma \in [0, 1)$，使得对于任意 $\boldsymbol{Z}$ 和 $\boldsymbol{\theta}$, $\mathbb{E}\|\nabla_{\boldsymbol{Z}} \hat{} F_{\boldsymbol{D}}(\boldsymbol{Z}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{Z}} F_{\boldsymbol{D}}(\boldsymbol{Z}, \boldsymbol{\theta})\|^2 \leq 2\varsigma(M^2\|\boldsymbol{Z}\|^2 + M^2\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 + B^2)$。

假设 A5。步长 $\{\gamma_k\}_{k\in\mathbb{N}}$ 是一个正的递减序列，使得 $\gamma_k \to 0$ 和 $\sum_{k=1}^{\infty} \gamma_k = \infty$。此外，令 $h(\boldsymbol{\theta}) = \mathbb{E}(H(\boldsymbol{Z}, \boldsymbol{\theta}))$，则存在 $\delta > 0$，使得对于任意 $\boldsymbol{\theta} \in \Theta$，$\langle \underline{\boldsymbol{\theta} - \boldsymbol{\theta}^*}, \underline{h(\boldsymbol{\theta}))}\rangle \geq \delta\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$，且 $\liminf_{k\to\infty} 2\delta \frac{\gamma_k}{\gamma_{k+1}} + \frac{\gamma_{k+1} - \gamma_k}{\gamma_{k+1}} > 0$。

假设 *A6.*（泊松方程的解）对于任意 $\boldsymbol{\theta} \in \Theta$、$\boldsymbol{z} \in \mathfrak{Z}$ 以及函数 $V(\boldsymbol{z}) = 1 + \|\boldsymbol{z}\|$，存在定义在 $\mathfrak{Z}$ 上的函数 $\mu_{\boldsymbol{\theta}}$，它求解泊松方程 $\mu_{\boldsymbol{\theta}}(\boldsymbol{z}) - T_{\boldsymbol{\theta}}\mu_{\boldsymbol{\theta}}(\boldsymbol{z}) = H(\boldsymbol{\theta}, \boldsymbol{z}) - h(\boldsymbol{\theta})$，其中 $T_{\boldsymbol{\theta}}$ 表示一个概率转移核，且 $T_{\boldsymbol{\theta}}\mu_{\boldsymbol{\theta}}(\boldsymbol{z}) = \int_{\mathfrak{Z}} \mu_{\boldsymbol{\theta}}(\boldsymbol{z}')T_{\boldsymbol{\theta}}(\boldsymbol{z}, \boldsymbol{z}')d\boldsymbol{z}'$，使得

$$H(\boldsymbol{\theta}_k, \boldsymbol{z}_{k+1}) = h(\boldsymbol{\theta}_k) + \mu_{\boldsymbol{\theta}_k}(\boldsymbol{z}_{k+1}) - \mathcal{T}_{\boldsymbol{\theta}_k}\mu_{\boldsymbol{\theta}_k}(\boldsymbol{z}_{k+1}), \quad k = 1, 2, \ldots. \tag{A3}$$

此外，对于所有 $\boldsymbol{\theta}$、$\boldsymbol{\theta}' \in \Theta$ 和 $\boldsymbol{z} \in \mathfrak{Z}$，我们有 $\|\mu_{\boldsymbol{\theta}}(\boldsymbol{z}) - \mu_{\boldsymbol{\theta}'}(\boldsymbol{z})\| \leq \varsigma_1\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|V(\boldsymbol{z})$ 以及 $\|\mu_{\boldsymbol{\theta}}(\boldsymbol{z})\| \leq \varsigma_2 V(\boldsymbol{z})$，其中 $\varsigma_1 > 0$ 和 $\varsigma_2 > 0$ 为某些常数。

引理 A1.（定理 *S1*，*Liang et al. (2022)*）设假设 *A4-A6* 成立。对于算法 *1*，若我们设定 $\epsilon_k = C_\epsilon/(c_e + k^\alpha)$ 和 $\gamma_k = C_\gamma/(c_g + k^\alpha)$，其中 $\alpha \in (0, 1)$、$C_\epsilon > 0$、$C_\gamma > 0$、$c_e \geq 0$ 和 $c_g \geq 0$ 为某些常数，则存在一次迭代 k_0 和常数 $\Lambda_0 > 0$，使得对于任意 $k > k_0$，

$$\mathbb{E}(\|\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}_n^*\|^2) \leq \Lambda_0\gamma_k, \tag{A4}$$

其中，$\hat{\boldsymbol{\theta}}_n^*$ 表示公式 *(7)* 的一个解，即 $\hat{\boldsymbol{\theta}}_n^* \in L = \{\boldsymbol{\theta}: \nabla_{\boldsymbol{\theta}}$ 对数 $\pi(\boldsymbol{\theta}|\boldsymbol{A}, \boldsymbol{Y}) = 0\}$；且 $\Lambda_0 = \Lambda_0' + 6\sqrt{6}C_{\boldsymbol{\theta}}^{1/2}((3M^2 + \varsigma_2)C_{\boldsymbol{Z}} + 3M^2C_{\boldsymbol{\theta}} + 3B^2 + \varsigma_2^2)^{1/2}$，其中 Λ_0'、$C_{\boldsymbol{\theta}}$ 和 $C_{\boldsymbol{Z}}$ 为某些正常数。

证明。引理 A1 是对 Liang et al. (2022) 中定理 S1 的重述，因此省略其证明。

□

有关常数 $C_{\boldsymbol{\theta}}$ 和 $C_{\boldsymbol{Z}}$ 的推导，请参阅 Liang et al. (2022) 的引理 S1；这些常数表示 $\boldsymbol{\theta}^{(k)}$ 的收敛对 StoNet (4) 结构的依赖。由 l_2-收敛（A4）可立即得到 $\|\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}_n^*\| \xrightarrow{p} 0$ as $k \to \infty$，其中 $\xrightarrow{p}$ 表示依概率收敛。

备注 A1. 对于神经网络，已知其损失函数在连接权重的某些变换下保持不变，例如对层内隐藏神经元重新排序，或同时改变特定权重和偏置的符号或尺度；详细讨论请参阅 *Liang* 等人（*2018b*）和 *Sun* 等人（*2022*）。因此，解 $\hat{\boldsymbol{\theta}}_n^*$ 并不唯一，所有此类解都可视为属于一个由损失不变变换定义的唯一解等价类。该等价类构成参数空间的一种简化表示，其中每个成员对应一个不同的网络（即无法通过损失不变操作转换为另一个网络），且其损失值可能不同。本文所建立的一致性结果专门适用于这一简化后的神经网络空间。

A5.2 $\hat{\boldsymbol{\theta}}_n^*$ 的一致性

A5.2.1 IRO 算法的一致性

IRO 算法IRO 算法（Liang 等人，2018a）以初始权重设定 $\hat{\boldsymbol{\theta}}^{(0)} = (\hat{\boldsymbol{\theta}}_1^{(0)}, \hat{\boldsymbol{\theta}}_2^{(0)})$ 开始，随后在潜在混杂因素的插补与用于参数更新的正则化优化之间交替迭代：

- **插补：**从预测分布中模拟 $\boldsymbol{z}_{(t+1)\,i}$：其中 t 表示迭代次数，且 $\boldsymbol{\sigma}_{CI}^2 = (\sigma_z^2, \sigma_y^2)$。
$$\pi(\boldsymbol{z}_i \mid \boldsymbol{y}_i, \boldsymbol{a}_i, \hat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\sigma}_{CI}^2) \propto \pi(\boldsymbol{z}_i \mid \boldsymbol{a}_i, \hat{\boldsymbol{\theta}}_1^{(t)}, \sigma_z^2)\pi(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{a}_i, \hat{\boldsymbol{\theta}}_2^{(t)}, \sigma_y^2)$$

- **正则化优化：**给定伪完整数据 $\{(\boldsymbol{y}_i, \boldsymbol{z}_{(t+1)\,i}, \boldsymbol{a}_i) : i = 1, 2, \ldots, n\}$，通过最大化惩罚对数似然函数来更新 $\hat{\boldsymbol{\theta}}^{(t+1)}$，如下所示：

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \log \pi(\boldsymbol{y}_i, \boldsymbol{z}_i^{(t+1)} | \boldsymbol{a}_i, \boldsymbol{\theta}, \boldsymbol{\sigma}_{CI}^2) - \frac{1}{n} \log P_{\lambda_n}(\boldsymbol{\theta}) \Big\}. \tag{A5}$$

惩罚函数 $\frac{1}{n} \log P_{\lambda_n}(\boldsymbol{\theta})$ 满足若干条件（见假设 A9），使得 $\hat{\boldsymbol{\theta}}^{(t+1)}$ 在各次迭代上一致地构成工作参数的相合估计量

$$\begin{aligned} \boldsymbol{\theta}_*^{(t+1)} &= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\hat{\boldsymbol{\theta}}^{(t)}} \log \pi(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\sigma}_{CI}^2) \\ &= \arg\max_{\boldsymbol{\theta}} \int \log \pi(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\sigma}_{CI}^2) \pi(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{a}, \hat{\boldsymbol{\theta}}^{(t)}, \sigma_z^2) \pi(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{\theta}^*, \sigma_y^2) d\boldsymbol{z} d\boldsymbol{y}, \end{aligned} \tag{A6}$$

其中 $\boldsymbol{\theta}^*$ 表示 CI-StoNet 模型的真实参数值。

参数估计的一致性 参数估计一致性的主要证明建立在 Liang 等人（2018a）提出的理论框架之上。令 $\tilde{\boldsymbol{x}} = (\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{Z})$ 为完整数据，即观测变量与潜变量的集合。定义

$$\begin{aligned} G_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) &= \int \log \pi(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\sigma}_{CI}^2) \pi(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{a}, \hat{\boldsymbol{\theta}}^{(t)}, \sigma_z^2) \pi(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{\theta}^*, \sigma_y^2) d\boldsymbol{z} d\boldsymbol{y}, \\ \hat{G}_n(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}^{(t)}) &= \frac{1}{n} \sum_{i=1}^{n} \log \pi(\boldsymbol{y}_i, \boldsymbol{z}_i | \boldsymbol{a}_i, \boldsymbol{\theta}, \boldsymbol{\sigma}_{CI}^2), \quad \boldsymbol{z}_i \sim \pi(\boldsymbol{z} | \boldsymbol{y}_i, \boldsymbol{a}_i, \hat{\boldsymbol{\theta}}^{(t)}, \sigma_z^2), \end{aligned}$$

引理 A2.（定理 *1*；*Liang* 等人（*2018a*））令 T 表示 *IRO* 算法的总迭代次数。在温和的正则条件（见 *Liang* 等人（*2018a*）中的假设 *1–3*）下，对任意 T，当 $\log(T) = o(n)$ 时，下列一致大数定律成立：

$$\sup_{\hat{\boldsymbol{\theta}}^{(t)} \in \boldsymbol{\theta}^T} \sup_{\boldsymbol{\theta} \in \Theta} \left| \hat{G}_n(\boldsymbol{\theta} \mid \tilde{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}^{(t)}) - G_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) \right| \xrightarrow{p} 0, \tag{A7}$$

随着样本量 $n \to \infty$。

假设 *A7*。对于每个 $t = 1, 2, \ldots, T$， $G_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ 在 $\boldsymbol{\theta}_*^{(t)}$ 处取得唯一最大值（除损失不变变换外）；对于任意 $\epsilon > 0$，$\sup_{\boldsymbol{\theta} \in \Theta \setminus B_t(\epsilon)} G_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ 存在，其中 $B_t(\epsilon) = \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*^{(t)}\| < \epsilon\}$。令 $\delta_t = G_n(\boldsymbol{\theta}_*^{(t)} \mid \hat{\boldsymbol{\theta}}^{(t)}) - \sup_{\boldsymbol{\theta} \in \Theta \setminus B_t(\epsilon)} G_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$，$\delta = \min_{t \in \{1,2,\ldots,T\}} \delta_t > 0$ 成立。

假设 A7 限制了全局最大值点附近 $G_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(t)})$ 的形状，确保其既不间断，也不会过于平坦。鉴于神经网络模型的不可识别性，假设 A7 隐含地假设每个 $\boldsymbol{\theta}$ 在损失不变变换下都是唯一的，例如对同一层中的隐藏神经元重新排序，或同时改变某些权重和偏置的符号或尺度；进一步讨论参见 Liang 等人（2018b）和 Sun 等人（2022）。或者，可以将最优解视为属于某个等价类，并在适当的损失不变变换下考虑，同时将唯一性假设应用于该等价类。

此外，考虑由以下公式定义的映射 $M(\boldsymbol{\theta})$：

$$M(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}'} \mathbb{E}_{\boldsymbol{\theta}} \log \pi(\boldsymbol{Y}, \boldsymbol{Z} | \boldsymbol{a}, \boldsymbol{\theta}', \boldsymbol{\sigma}_{CI}^2).$$

正如 Liang 等人（2018a）和 Nielsen（2000）所论证的，合理的假设是，该映射是压缩映射，因为对该映射进行递归应用，即令

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \boldsymbol{\theta}_*^{(t+1)} = M(\hat{\boldsymbol{\theta}}^{(t)}),$$

导致目标期望 $\mathbb{E}_{\hat{\boldsymbol{\theta}}^{(t)}} \quad \pi\ \boldsymbol{Y}\ \boldsymbol{Z} | \boldsymbol{a}\ \boldsymbol{\theta}\ \boldsymbol{\sigma}_{CI}^2 \quad t = \log(,,,)$ 对 ,, 单调增加。 $1\ 2\ \ldots$

假设 A8。映射 $M(\boldsymbol{\theta})$ 是可微的。令 $\rho_n(\boldsymbol{\theta})$ 为 $\partial M(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ 的最大奇异值。存在一个数 $\rho^* < 1$，使得 $\rho_n(\boldsymbol{\theta}) \leq \rho^*$ 对所有 $\boldsymbol{\theta} \in \Theta$、充分大的 n 以及几乎每一个 $(\boldsymbol{A}, \boldsymbol{Y})$ 的观测序列成立。

假设 A9。惩罚函数 $\frac{1}{n} \log P_{\lambda_n}(\boldsymbol{\theta})$ 在集合 $\{\boldsymbol{\theta}_*^{(t)} : t = 1, 2, \ldots, T\}$ 上随着 $n \to \infty$ 一致收敛到 0，其中 λ_n 是正则化参数，其取值可以依赖于样本量 n。

引理 A3。（定理 *4*；*Liang* 等人（*2018a*））假设引理 *A2* 的条件以及假设 *A7–A9* 成立，并且 $\sup_{n,t} \mathbb{E}\|\hat{\boldsymbol{\theta}}_n^{(t)}\| < \infty$ 成立。则当 t 充分大时，对几乎每一个 $(\boldsymbol{A}, \boldsymbol{Y})$-序列，随着 $n \to \infty$，有 $\|\hat{\boldsymbol{\theta}}_n^{(t)} - \boldsymbol{\theta}^*\| \xrightarrow{p} 0$。

A5.2.2 假设 A9 的验证

为验证假设 A9，我们证明以下引理。

引理 A4. 设 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{K_n})^T$。假设 $\boldsymbol{\theta}$ 的所有分量先验相互独立，且服从以下混合高斯先验 *(6)*。假设 $K_n \succ n$，$\boldsymbol{\theta}$ 以某个常数 $c > 1$ 为水平是稀疏的，$m_n \prec {}_c{}^{n}{}_{\log(K_n/n)}$，并且对于某个常数 $\delta_n = o(1)$，$|\boldsymbol{\theta}_i| : \boldsymbol{\theta}_i \neq 0, i = 1, 2, \ldots, K_n$ 的 $\min > \delta_n$。若我们设定 $\sigma_1 = O(1)$，并令 (λ_n, σ_0) 满足以下条件：

$$\begin{aligned} &(\frac{n}{K_n})^c \prec \lambda_n \prec \frac{n}{K_n}, \\ &(\frac{n}{K_n})^c \prec \sigma_0 \prec \min\left\{1 - \frac{n}{K_n}, \frac{\delta_n}{\sqrt{c\log(K_n) - (c-1)\log(n)}}\right\}, \end{aligned} \tag{A8}$$

那么以下结果成立：

$$\frac{1}{n}\left|\log \pi(\boldsymbol{\theta}) + K_n \log\left(\sqrt{2\pi}\sigma_0\right)\right| \to 0, \quad as\ n \to \infty. \tag{A9}$$

证明。 经过直接计算可得

$$|\log \pi(\boldsymbol{\theta}) + K_n \log(\sigma_0)| \lesssim K_n |\log(1-\lambda_n)| + (K_n - m_n)\frac{\sigma_0 \lambda_n}{\sigma_1(1-\lambda_n)} + m_n \left|\log\left(\frac{\sigma_0\lambda_n}{1-\lambda_n}\right)\right| - m_n \frac{\delta_n^2}{2\sigma_1^2} + \frac{m_n(1-\lambda_n)\sigma_1}{\lambda_n \sigma_0} e^{-\frac{\delta_n^2}{2}(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2})}.$$

为确保 $K_n|\log(1-\lambda)| \prec n$，我们设定

$$\lambda_n \prec 1 - e^{-n/K_n} \asymp \frac{n}{K_n}. \tag{A10}$$

为确保 $m_n \left|\log\left(\frac{\sigma_0\lambda_n}{1-\lambda_n}\right)\right| \prec n$，我们设定

$$\sigma_0 \succ (\frac{n}{K_n})^c \succ e^{-n/m_n}, \quad \lambda_n \succ (\frac{n}{K_n})^c \succ e^{-n/m_n}. \tag{A11}$$

为确保 $(K_n - m_n)\frac{\sigma_0\lambda_n}{\sigma_1(1-\lambda_n)} \prec n$，我们设定

$$\sigma_0 \prec 1 - \frac{n}{K_n} \prec \frac{n}{K_n}\frac{(1-\lambda_n)}{\lambda_n}. \tag{A12}$$

为确保 $\frac{m_n(1-\lambda_n}{\lambda_n\sigma_0}\sigma_1 e^{-\frac{\delta_n^2}{2}})(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}) \prec n$，我们设定

$$\sigma_0 \prec \frac{\delta_n}{\sqrt{c\log(K_n) - (c-1)\log(n)}} \prec \frac{\delta_n}{\sqrt{|\log(n\lambda_n/m_n))|}}. \tag{A13}$$

由于 $\delta_n \prec o(1)$ 且 $m_n \prec n$，因此我们有 $m_n \frac{\delta_n^2}{2\sigma_1^2} \prec n$。

作为对 (A10)-(A13) 的总结，我们可以按 (A8) 中所述设置 (λ_n, σ_0)，从而确保 (A9) 成立。 □

定理 A1. 假设引理 *A2* 中给出的正则条件以及假设 *A7-A8*（见补充材料 *A5*）均成立。此外，假设 $\boldsymbol{\theta}$ 的维度记为 K_n，并随 n 以多项式速率 $K_n = O(n^\zeta)$ 增长，其中 $\zeta > 1$ 为某个常数；同时，真实 *StoNet* 是稀疏的，其非零连接数为 $m_n \prec c^{n}{}_{\log(K_n/n)}$，其中 $c > 1$ 为某个常数。设置先验 *(6)* 的超参数，使其满足以下条件：

$$(\frac{n}{K_n})^c \prec \lambda_n \prec \frac{n}{K_n}, \quad \sigma_1 = O(1), \quad (\frac{n}{K_n})^c \prec \sigma_0 \prec \min\left\{1 - \frac{n}{K_n}, \frac{\delta_n}{\sqrt{c\log(K_n) - (c-1)\log(n)}}\right\}. \tag{A14}$$

则当 $n \to \infty$ 时，$\|\hat{}\boldsymbol{\theta}_n^* - \boldsymbol{\theta}^*\| \xrightarrow{p} 0$ 成立，其中 $\boldsymbol{\theta}^*$ 表示 *StoNet (4)* 的真实参数，而 $\hat{}\boldsymbol{\theta}_n^*$ 在损失不变变换意义下成立。

证明。 由于 $\hat{}\boldsymbol{\theta}_n^*$ 是方程 (7) 的一个解，它是关于不完整数据（将 $\boldsymbol{Z}$ 视为缺失）的 $\boldsymbol{\theta}$ 的最大后验 （MAP）估计量。根据引理 A3，我们立即得到它关于 $\boldsymbol{\theta}^*$ 的一致性，即

$$\|\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}^*\| \xrightarrow{p} 0, \quad \text{as } n \to \infty. \tag{A15}$$

在引理 A3 的各项条件中，我们只需验证假设 A9，因为其余条件通常都满足。回想一下，我们在计算 $\boldsymbol{\theta}$ 的 MAP 时采用了混合高斯先验 (6)。根据引理 A4，假设 A9 成立。证明完毕。 □

注 A2. 在定理 *A1* 中，我们假设真实的稀疏 *StoNet* 的规模为 $m_n = o(n)$。这一假设可由 *Bölcskei* 等（*2019*）、*Schmidt-Hieber*（*2017*）以及 *Petersen* 与 *Voigtlaender*（*2018*）建立的理论加以论证；这些研究表明，具有此规模的深度神经网络足以逼近许多类函数，包括仿射函数、分段光滑函数以及 α-*Hölder* 光滑函数。关于这一问题的讨论，参见 *Sun* 等人（*2022*）。此外，*Sun* 等人（*2022*）表明，在样本量 n 变大时，具有此规模的稀疏神经网络搭配混合高斯先验（*6*），足以实现所需的函数逼近和后验一致性。我们的理论允许K_n 随 n多项式增长，这通常适用于深度神经网络。

A5.3 定理 1 的证明

估计量（9）的论证 为论证合并蒙特卡洛平均值，固定一个可积测试函数 φ，并定义 $\bar{\varphi}_i := \frac{1}{M}\sum_{l=1}^{M}\varphi(\boldsymbol{z}_i^{(l)})$。给定 $\boldsymbol{A}_i = \boldsymbol{a}_i$，抽样值 $\boldsymbol{z}_i^{(l)} \sim p(\cdot \mid \boldsymbol{a}_i)$ 为独立同分布（或遍历）的，因此

$$\bar{\varphi}_i \xrightarrow{p} \mathbb{E}\{\varphi(\boldsymbol{z}) \mid \boldsymbol{a}_i\} \qquad \text{as } \mathcal{M} \to \infty,$$

其中，$\xrightarrow{p}$表示依概率收敛。因此，对于任意固定的 n,

$$\frac{1}{n\mathcal{M}}\sum_{i=1}^{n}\sum_{l=1}^{\mathcal{M}}\varphi(\boldsymbol{z}_i^{(l)}) = \frac{1}{n}\sum_{i=1}^{n}\bar{\varphi}_i \xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{\varphi(\boldsymbol{z}) \mid \boldsymbol{a}_i\}, \qquad \text{as } \mathcal{M} \to \infty.$$

根据假设 2，$\{\boldsymbol{a}_i\}_{i=1}^{n}$ 可假定为独立同分布，弱大数定律表明

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{\varphi(\boldsymbol{z}) \mid \boldsymbol{a}_i\} \xrightarrow{p} \mathbb{E}_A[\mathbb{E}\{\varphi(\boldsymbol{z}) \mid \boldsymbol{a}\}] = \int \varphi(\boldsymbol{z})\, p(\boldsymbol{z})\, d\boldsymbol{z}, \qquad \text{as } n \to \infty.$$

定理 1 的证明

证明。 考虑联合密度函数:

$$\pi(\boldsymbol{Z}, \boldsymbol{Y} | \boldsymbol{A}, \boldsymbol{\theta}^*) = \pi(\boldsymbol{Z}|\boldsymbol{A}, \boldsymbol{\theta}_1^*)\pi(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\theta}_2^*),$$

在假设真实模型是由 $\boldsymbol{\theta}^*$ 参数化的稀疏 StoNet (4) 的条件下，我们有

$$\mathbb{E}(Y(\boldsymbol{a})|\boldsymbol{\theta}^*) = \int \boldsymbol{y}\pi(\boldsymbol{z}|\boldsymbol{\theta}_1^*)\pi(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{a}, \boldsymbol{\theta}_2^*)d\boldsymbol{z}d\boldsymbol{y} = \int \mu_2(\boldsymbol{z}, \boldsymbol{a}, \boldsymbol{\theta}_2^*)\pi(\boldsymbol{z}|\boldsymbol{\theta}_1^*)d\boldsymbol{z}.$$

Let $\boldsymbol{z}_i^{(l)}$，对于 $l = 1$、2、.....、$\mathcal{M}$，表示从 $\pi(\boldsymbol{z}|\boldsymbol{a}_i$、$\hat{}\boldsymbol{\theta}_1^*)$ 中抽取的$\mathcal{M}$ 个独立样本。令

$$\widehat{\mathbb{E}(Y(\boldsymbol{a})|\hat{\boldsymbol{\theta}}_n^*)} = \frac{1}{n\mathcal{M}}\sum_{i=1}^{n}\sum_{l=1}^{\mathcal{M}}\mu_2(\boldsymbol{z}_i^{(l)}, \boldsymbol{a}, \hat{\boldsymbol{\theta}}_2^*).$$

根据蒙特卡洛平均的标准性质（与估计量 (9) 的论证相同），我们有

$$\|\widehat{\mathbb{E}(Y(\boldsymbol{a})|\hat{\boldsymbol{\theta}}_n^*)} - \mathbb{E}(Y(\boldsymbol{a})|\hat{\boldsymbol{\theta}}_n^*)\| \xrightarrow{p} 0, \quad \text{as } n, \mathcal{M} \to \infty. \tag{A16}$$

另一方面，根据引理 A1 中所证明的 $\hat{}\boldsymbol{\theta}_n^* = (\hat{}\boldsymbol{\theta}_1^*, \hat{}\boldsymbol{\theta}_2^*)$ 关于 $\boldsymbol{\theta}^*$ 的一致性，我们有

$$\|\mathbb{E}(Y(\boldsymbol{a})|\hat{\boldsymbol{\theta}}_n^*) - \mathbb{E}(Y(\boldsymbol{a})|\boldsymbol{\theta}^*)\| \xrightarrow{p} 0, \quad \text{as } n \to \infty, \tag{A17}$$

因为 $\mu_2(\cdot)$ 关于参数连续（神经网络模型的假设）。

结合 (A16) 和 (A17) 中的收敛结果，我们有

$$\widehat{\|\mathbb{E}(Y(\boldsymbol{a})|\hat{}\boldsymbol{\theta}_n^*) - \mathbb{E}(Y(\boldsymbol{a})|\boldsymbol{\theta}^*)\|} \le \|\mathbb{E}(Y(\boldsymbol{a})|\hat{}\boldsymbol{\theta}_n^*) - \mathbb{E}(Y(\boldsymbol{a})|\hat{}\boldsymbol{\theta}^*)\| + \|\mathbb{E}(Y(\boldsymbol{a})|\hat{}\boldsymbol{\theta}^*) - \mathbb{E}(Y(\boldsymbol{a})|\boldsymbol{\theta}^*)\| \xrightarrow{p} 0,$$

其中 $n \to \infty$ 和 $\mathcal{M} \to \infty$。至此，证明完毕。 □

A5.4 定理 2 的证明

我们可以利用 Sun 等人（2022年）所发展的理论，为因果效应的模型设定错误提供依据。

假设 A10。 （重述 *Sun* 等人（*2022*年）的假设*A.2*）稀疏*DNN*模型 $\mu_{A,n}$ 和 $\mu_{Y,n}$ 满足以下条件:

1. 网络结构满足：

$$r_n H_n \log n \;+\; r_n \log \overline{L} \;+\; s_n \log p_n \;\le\; C_0 n^{1-\varepsilon},$$

其中，$0 < \varepsilon < 1$ 是一个较小的常数，$r_n = |\gamma^*|$ 表示 γ^*的连通度，$L =$ 最大值$_{1\le j\le H_n-1} L_j$ 表示最大隐藏层宽度，而 s_n 表示 γ^*的输入维度。

2. 网络权重具有多项式有界性：

$$\|\boldsymbol{\beta}^*\|_\infty \le E_n, \qquad E_n = n^{C_1}$$

对于某个常数 $C_1 > 0$。

例如，仿射系统函数（Bolcskei 等人（2019年））、输入维度固定的分段光滑函数（Petersen & Voigtlaender（2018年）），以及有界的 α-Holder 光滑函数（Polson & Ročková（2018年））。假设 A10 阐明了能够逼近未知结构均值函数的稀疏深度神经网络类别，因此这些网络在本文中被视为“真实稀疏深度神经网络”。非正式地说，假设 A10 要求稀疏深度网络的活跃权重数和相关输入数的增长速度低于样本量，且其权重至多随 n 呈多项式增长。这确保“真实”网络属于一个容量受控函数类。通过相应地改变网络结构和网络权重，我们可以针对不同函数类建立不同的逼近误差上界序列ω_n $\mathcal{F}$，更多详情请参见 Sun 等人（2022年）第2.2节的讨论。

证明。 令 $\mathcal{G}_n$ 表示与假设 A10 相容的稀疏深度神经网络类，我们对每个样本量 n 将伪真稀疏深度神经网络 $(\mu^*_{A,n},\mu^*_{Y,n})$ 定义为 $\mathcal{G}_n$ 内逼近误差的极小化元：

$$(\mu^*_{A,n}, \mu^*_{Y,n}) \in \arg\min_{(\mu_1,\mu_2)\in\mathcal{G}_n} \{\|\mu_{A,n} - m_A\|_{L^2} + \|\mu_{Y,n} - m_Y\|_{L^2}\}$$

根据假设 3，存在序列 $\omega_n \to 0$ 使得

$$\{\|\mu_{A,n} - m_A\|_{L^2} + \|\mu_{Y,n} - m_Y\|_{L^2}\} \lesssim \omega_n$$

其中 ω_n 按所选函数 $\mathcal{F}$ 的选定速率缩放（例如，对于维数为 d 的 α -赫尔德函数，$\omega_n \asymp n^{-\alpha/(2\alpha+d)}$ 至多相差对数因子）。

令 P_0 表示 $(\boldsymbol{A}, \boldsymbol{Z}, \boldsymbol{Y})$ 的真实联合分布。对稀疏深度神经网络的任意参数 η，令 Q_η 为相应广义线性模型（高斯或逻辑斯蒂）在回归函数μ_η下诱导的联合分布，并令 $P_{\boldsymbol{\theta}}$ 为参数 $\boldsymbol{\theta}$ 的 CI-StoNet 所诱导的分布。遵循 Sun 等人（2022），我们定义 Kullback–Leibler 散度

$$d_0(q,p) = \int q \log \frac{q}{p},$$

以及一族距离 $d_t(q,p) = \frac{1}{t}\int q[(\frac{q}{p})^t - 1]$，对任意 $t > 0$ 均成立；当 t 趋近于 0 时，该距离减小至 d_0。对于均值函数为 m_1 和 m_2 的高斯回归，他们证明，在噪声尺度固定且回归函数有界并且已知 σ^2 的条件下，

$$d_0(p_{m_1}, p_{m_2}) = \frac{1}{2\sigma^2}(m_1 - m_2)^2,$$

对于逻辑斯蒂回归，他们推导出了上界

$$d_1(p_{m_1}, p_{m_2}) \le \frac{1}{2}(m_1 - m_2)^2 + O((m_1 - m_2)^3).$$

由于 $d_0(p_{m_1}, p_{m_2}) \le d_1(p_{m_1}, p_{m_2})$，我们有 $d_0(p_{m_1}, p_{m_2}) \le C(m_1 - m_2)^2$ 其中某个常数 $C > 0$ 仅取决于 μ 的一致有界性。于是可得

$$\mathrm{KL}(P_0, Q_\eta) \lesssim \|\mu_\eta - \mu\|^2_{L^2},$$

其中 μ_η 是 Q_η 中基于深度神经网络的回归函数， μ^* 是真实回归函数。综上，若函数类 $\mathcal{F}$ 可由稀疏深度神经网络以 L_2 中的速率 ω_n 逼近，则存在稀疏深度神经网络参数 η_n ，使得

$$\mathrm{KL}(P_0, Q_{\eta_n}) \lesssim \omega_n^2.$$

在 StoNet 的定义中，确定性部分 m_1、 m_2 本身就是神经网络。因此，对于任何满足假设 A10 的稀疏深度神经网络架构，我们可以通过选择 $\boldsymbol{\theta}$ ，将同一架构嵌入 StoNet，使 StoNet 的确定性部分与该深度神经网络一致，从而得到一个满足 $Q_\eta = P_{\boldsymbol{\theta}(\eta)}$的映射 $\eta \to \boldsymbol{\theta}(\eta)$。因此，定义一个伪真实随机神经网络参数

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}(P_0, P_{\boldsymbol{\theta}}),$$

则我们有

$$\mathrm{KL}(P_0, P_{\boldsymbol{\theta}^*}) \le \mathrm{KL}(P_0, P_{\boldsymbol{\theta}(\eta)}) = \mathrm{KL}(P_0, Q_\eta) \lesssim \omega_n^2.$$

现在考虑简单混杂情形（回顾第4节中的设定）。令 $\psi(P_0) = \int m_Y(\boldsymbol{a}, \boldsymbol{z}) p_{P_0}(\boldsymbol{z}) d\boldsymbol{z}$ 以及 $\psi(P_{\boldsymbol{\theta}^*}) = \int \mu_2(\boldsymbol{a}, \boldsymbol{z}) p_{P_{\boldsymbol{\theta}^*}}(\boldsymbol{z}) d\boldsymbol{z}$，真实因果估计目标与伪真实因果估计目标之间的逼近误差为:

$$\begin{aligned}
\|\psi(P_0) - \psi(P_{\boldsymbol{\theta}^*})\| &= \| \int m_Y(\boldsymbol{a}, \boldsymbol{z}) p_{P_0}(\boldsymbol{z}) d\boldsymbol{z} - \int \mu_2(\boldsymbol{a}, \boldsymbol{z}) p_{P_{\boldsymbol{\theta}^*}}(\boldsymbol{z}) d\boldsymbol{z} \| \\
&\le \underbrace{\| \int [m_Y(\boldsymbol{a}, \boldsymbol{z}) - \mu_2(\boldsymbol{a}, \boldsymbol{z})] p_{P_0}(\boldsymbol{z}) d\boldsymbol{z} \|}_{A_1} + \underbrace{\| \int \mu_2(\boldsymbol{a}, \boldsymbol{z}) [p_{P_0}(\boldsymbol{z}) - p_{P_{\boldsymbol{\theta}^*}}(\boldsymbol{z})] d\boldsymbol{z} \|}_{A_2}
\end{aligned}$$

A_1 可以被界定为

$$A_1 \le \|m_Y(\boldsymbol{a}, \boldsymbol{z}) - \mu_2(\boldsymbol{a}, \boldsymbol{z})\|_{L_2(P_0(\boldsymbol{z}))} \lesssim \omega_n.$$

第一行由柯西–施瓦茨不等式得到，第二行由稀疏深度神经网络逼近结果以及随机神经网络与深度神经网络的等价性得到。

假设 $|\mu_2(\boldsymbol{a}, \boldsymbol{z})| \le C_{\mu_2}$，则

$$\begin{aligned}
A_2 &\le C_{\mu_2} \int |p_{P_0(\boldsymbol{z})} - p_{\boldsymbol{\theta}^*}(\boldsymbol{z})| d\boldsymbol{z} \\
&\le 2C_{\mu_2} \sqrt{\frac{1}{2} \mathrm{KL}(P_{0,Z}, P_{\boldsymbol{\theta}^*,Z})} \\
&\le \sqrt{2} C_{\mu_2} \sqrt{\mathrm{KL}(P_0, P_{\boldsymbol{\theta}^*})} \\
&\lesssim \omega_n.
\end{aligned}$$

从第一行到第二行，我们使用 Pinsker 不等式；从第三行到最后一行，我们使用了先前的结果，即 $\mathrm{KL}(P_0, P_{\boldsymbol{\theta}^*}) \lesssim \omega_n^2$。从第二行到第三行，我们对 KL 散度使用数据处理不等式（在边缘化下的单调性）：

$$\mathrm{KL}(P_{0,Z}, P_{\boldsymbol{\theta}^*,Z}) \le \mathrm{KL}(P_0, P_{\boldsymbol{\theta}^*}).$$

综合以上各部分，我们得到

$$\|\psi(P_0) - \psi(P_{\boldsymbol{\theta}^*})\| \le A_1 + A_2 \lesssim \omega_n + \omega_n \lesssim \omega_n.$$

因此，我们将干扰函数的逼近误差传播到了因果估计目标，并证明因果估计目标的逼近误差为 $O(\omega_n)$。 □

我们论文中的定理 1 控制了估计误差，这一结果由稀疏深度神经网络的性质所保证。模型设定错误仅取决于稀疏深度神经网络类的逼近能力（深度、宽度、稀疏率）。因此，先验仅通过控制模型容量间接影响模型设定错误。

A5.5 重叠膨胀的有限样本分析

假设 A11（有界性与重叠性）。对于每个处理 $\boldsymbol{a} \in A$，令 $m^*(\boldsymbol{z}, \boldsymbol{a})$ 表示真实结果干扰函数，$\hat{m}(\boldsymbol{z}, \boldsymbol{a})$ 为其估计量。假设：

1. 结果有界性：

$$\sup_z |m^*(\boldsymbol{z}, \boldsymbol{a})| \le C_m, \qquad \sup_z |\hat{m}(\boldsymbol{z}, \boldsymbol{a})| \le C_m.$$

2. 潜在倾向得分（正性）界限：存在常数 $0 < \kappa_z(\boldsymbol{a}) \le K_z(\boldsymbol{a}) < \infty$，使得对于相关支持集中的所有 z，

$$\kappa_z(\boldsymbol{a}) \le p^*(\boldsymbol{a} \mid \boldsymbol{z}) \le K_z(\boldsymbol{a}), \qquad \kappa_z(\boldsymbol{a}) \le \hat{p}(\boldsymbol{a} \mid \boldsymbol{z}) \le K_z(\boldsymbol{a}).$$

3. 观测倾向得分界限：存在常数 $0 < \kappa_x(\boldsymbol{a}) \le C_{ax}(\boldsymbol{a}) < \infty$，使得对于所有代理变量 x_i，

$$\kappa_x(\boldsymbol{a}) \le p^*(\boldsymbol{a} \mid \boldsymbol{x}_i) \le C_{ax}(\boldsymbol{a}), \qquad \kappa_x(\boldsymbol{a}) \le \hat{p}(\boldsymbol{a} \mid \boldsymbol{x}_i) \le C_{ax}(\boldsymbol{a}).$$

4. 边际处理密度界限：对于每个处理取值 $\boldsymbol{a} \in A$，都存在$\kappa_A(\boldsymbol{a}) > 0$，使得

$$p^*(\boldsymbol{a}) \ge \kappa_A(\boldsymbol{a}), \qquad \hat{p}(\boldsymbol{a}) \ge \kappa_A(\boldsymbol{a}).$$

定义 $\kappa(\boldsymbol{a}) := \min\{\kappa_z(\boldsymbol{a}), \kappa_x(\boldsymbol{a}), \kappa_A(\boldsymbol{a})\}$ 以及 $K(\boldsymbol{a}) := K_z(\boldsymbol{a})$。

引理 A5（简单混杂中的贝叶斯桥）。固定 $\boldsymbol{a} \in \mathcal{A}$ 满足 $p(\boldsymbol{a}) > 0$。则

$$p(\boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{a})\, w(\boldsymbol{z}, \boldsymbol{a}), \qquad w(\boldsymbol{z}, \boldsymbol{a}) := \frac{p(\boldsymbol{a})}{p(\boldsymbol{a} \mid \boldsymbol{z})}.$$

特别地，在假设 *A11* 下，

$$0 < w(\boldsymbol{z}, \boldsymbol{a}) \le \frac{K(\boldsymbol{a})}{\kappa(\boldsymbol{a})}, \qquad |w(\boldsymbol{z}, \boldsymbol{a}) - 1| \le \max\left\{1 - \frac{\kappa_A(\boldsymbol{a})}{K(\boldsymbol{a})},\ \frac{K(\boldsymbol{a})}{\kappa(\boldsymbol{a})} - 1\right\}.$$

同样的陈述对估计模型成立，其中 $\hat{w}(\boldsymbol{z}, \boldsymbol{a}) := \hat{p}(\boldsymbol{a})/\hat{p}(\boldsymbol{a} \mid \boldsymbol{z})$。

证明。贝叶斯法则给出 $p(\boldsymbol{z} \mid \boldsymbol{a}) = p(\boldsymbol{a} \mid \boldsymbol{z})p(\boldsymbol{z}) / p(\boldsymbol{a})$，因此 $p(\boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{a}) \cdot p(\boldsymbol{a}) / p(\boldsymbol{a} \mid \boldsymbol{z})$。在假设 A11 下，$p(\boldsymbol{a} \mid \boldsymbol{z}) \ge \kappa(\boldsymbol{a})$ 且 $p(\boldsymbol{a} \mid \boldsymbol{z}) \le K(\boldsymbol{a})$。此外，$p(\boldsymbol{a}) = \int p(\boldsymbol{a} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \le K(\boldsymbol{a})$，并且由假设 A11-(4)，$p(\boldsymbol{a}) \ge \kappa_A(\boldsymbol{a})$。因此 $w(\boldsymbol{z}, \boldsymbol{a}) \le K(\boldsymbol{a})/\kappa(\boldsymbol{a})$ 且 $w(\boldsymbol{z}, \boldsymbol{a}) \ge \kappa_A(\boldsymbol{a})/K(\boldsymbol{a})$，这蕴含了关于 $|w - 1|$ 的界。 □

定理 A2（估计量 (9) 的简单混杂界限）. 固定 $\boldsymbol{a} \in A$。令 $m^*(\boldsymbol{z}, \boldsymbol{a}) = \mathbb{E}[Y \mid \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{Z} = \boldsymbol{z}]$ 且 $\hat{m}(\boldsymbol{z}, \boldsymbol{a}) = \mu_2(\boldsymbol{z}, \boldsymbol{a}; \hat{\boldsymbol{\theta}}_2^*)$。设假设 *A11* 成立。定义总体因果估计目标

$$\psi^*(\boldsymbol{a}) := \int m^*(\boldsymbol{z}, \boldsymbol{a})\, p^*(\boldsymbol{z})\, d\boldsymbol{z}.$$

假设 (9) 中使用的潜在抽样满足

$$\boldsymbol{z}_i^{(l)} \overset{i.i.d.}{\sim} \hat{p}(\boldsymbol{z} \mid \boldsymbol{a}_i), \qquad i = 1, \dots, n,\ l = 1, \dots, \mathcal{M},$$

在给定$(\boldsymbol{a}_i)_{i=1}^n$ 与拟合模型的条件下。考虑该估计量

$$\hat{\psi}_{n,\mathcal{M}}(\boldsymbol{a}) := \frac{1}{n\mathcal{M}} \sum_{i=1}^{n} \sum_{l=1}^{\mathcal{M}} \hat{m}(\boldsymbol{z}_i^{(l)}, \boldsymbol{a}).$$

令相应的（基于模型的）条件插件目标为

$$\Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}}; \boldsymbol{a}) := \frac{1}{n} \sum_{i=1}^{n} \int \hat{m}(\boldsymbol{z}, \boldsymbol{a})\, \hat{p}(\boldsymbol{z} \mid \boldsymbol{a}_i)\, d\boldsymbol{z},$$

并将 $\psi^*(\boldsymbol{a})$ 改写为

$$\psi^*(\boldsymbol{a}) = \Phi^*_{n,\mathrm{bridge}}(\boldsymbol{a}) := \frac{1}{n}\sum_{i=1}^{n}\int m^*(\boldsymbol{z},\boldsymbol{a})\, p^*(\boldsymbol{z}\mid \boldsymbol{a}_i)\, w^*(\boldsymbol{z},\boldsymbol{a}_i)\, d\boldsymbol{z},$$

其中

$$w^*(\boldsymbol{z},\boldsymbol{a}_i) := \frac{p^*(\boldsymbol{a}_i)}{p^*(\boldsymbol{a}_i\mid \boldsymbol{z})}.$$

对于任意 $\delta\in(0,1)$，以至少 $1-\delta$ 的概率，

$$|\hat{\psi}_{n,\mathcal{M}}(\boldsymbol{a}) - \psi^*(\boldsymbol{a})| \le \underbrace{C_m\sqrt{\frac{2\log(2/\delta)}{n\mathcal{M}}}}_{\textit{Monte Carlo}} + \underbrace{\left|\Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a}) - \Phi^*_{n,\mathrm{cond}}(\boldsymbol{a})\right|}_{\textit{estimation under } p(\boldsymbol{z}|\boldsymbol{a})} + \underbrace{\left|\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a}) - \Phi^*_{n,\mathrm{bridge}}(\boldsymbol{a})\right|}_{\textit{Bayes-bridge (overlap-inflated) bias}},$$

其中 C_m 来自假设 ***A11-(1)***，并且

$$\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a}) := \frac{1}{n}\sum_{i=1}^{n}\int m^*(\boldsymbol{z},\boldsymbol{a})\, p^*(\boldsymbol{z}\mid \boldsymbol{a}_i)\, d\boldsymbol{z}.$$

此外，贝叶斯桥接偏差项满足

$$\begin{aligned}\left|\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a}) - \Phi^*_{n,\mathrm{bridge}}(\boldsymbol{a})\right| &= \frac{1}{n}\sum_{i=1}^{n}\left|\int m^*(\boldsymbol{z},\boldsymbol{a})\, p^*(\boldsymbol{z}\mid \boldsymbol{a}_i)\,\big(w^*(\boldsymbol{z},\boldsymbol{a}_i)-1\big)\, d\boldsymbol{z}\right| \\ &\le C_m\cdot \max_{1\le i\le n}\max\left\{1-\frac{\kappa_A(\boldsymbol{a}_i)}{K(\boldsymbol{a}_i)},\ \frac{K(\boldsymbol{a}_i)}{\kappa(\boldsymbol{a}_i)}-1\right\}.\end{aligned}$$

证明。 我们证明所述分解及其界。

步骤 1（误差分解）。加上再减去 $\Phi_{n,\ \mathrm{cond}}(\hat{}\boldsymbol{\theta};\boldsymbol{a})$ 以及 $\Phi^*_{n,\ \mathrm{cond}}(\boldsymbol{a})$：

$$\begin{aligned}&|\hat{\psi}_{n,\mathcal{M}}(\boldsymbol{a}) - \psi^*(\boldsymbol{a})| = |\hat{\psi}_{n,\mathcal{M}}(\boldsymbol{a}) - \Phi^*_{n,\mathrm{bridge}}(\boldsymbol{a})| \\ &\le |\hat{\psi}_{n,\mathcal{M}}(\boldsymbol{a}) - \Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a})| + |\Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a}) - \Phi^*_{n,\mathrm{cond}}(\boldsymbol{a})| + |\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a}) - \Phi^*_{n,\mathrm{bridge}}(\boldsymbol{a})|.\end{aligned}$$

控制每一项后，即可得到所示的三项界。

步骤 2（蒙特卡洛项）。给定 $(\boldsymbol{a}_i)_{i=1}^n$ 和拟合模型，根据假设，抽样值 $\boldsymbol{z}_i^{(l)} \overset{i.i.d.}{\sim} \hat{}p(\boldsymbol{z}\mid \boldsymbol{a}_i)$ 在 $(i,\ l)$ 上相互独立，且 $\hat{}m(\boldsymbol{z}_i^{(l)},\ \boldsymbol{a}) \in [-C_m, C_m]$。因此，根据霍夫丁不等式，对任意 $\epsilon>0$，

$$\mathbb{P}\Big(\Big|\hat{\psi}_{n,\mathcal{M}}(\boldsymbol{a}) - \Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a})\Big| > \epsilon\Big) \le 2\exp\left(-\frac{n\mathcal{M}\,\epsilon^2}{2C_m^2}\right).$$

令右侧等于 δ，可得以至少 $1-\delta$ 的概率，

$$\left|\hat{\psi}_{n,\mathcal{M}}(\boldsymbol{a}) - \Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a})\right| \le C_m\sqrt{\frac{2\log(2/\delta)}{n\mathcal{M}}}.$$

步骤 3（贝叶斯桥偏差界）。根据定义，

$$\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a}) - \Phi^*_{n,\mathrm{bridge}}(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\int m^*(\boldsymbol{z},\boldsymbol{a})\, p^*(\boldsymbol{z}\mid \boldsymbol{a}_i)\,\big(1-w^*(\boldsymbol{z},\boldsymbol{a}_i)\big)\, d\boldsymbol{z}.$$

取绝对值并使用 $\|m^*(\cdot\ \boldsymbol{a})\|_\infty \le C_m$,

$$\left|\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a})-\Phi^*_{n,\mathrm{bridge}}(\boldsymbol{a})\right| \le \frac{C_m}{n}\sum_{i=1}^{n}\int p^*(\boldsymbol{z}\mid \boldsymbol{a}_i)\,|w^*(\boldsymbol{z},\boldsymbol{a}_i)-1|\,d\boldsymbol{z} \le C_m\cdot\sup_i\sup_{\boldsymbol{z}}|w^*(\boldsymbol{z},\boldsymbol{a}_i)-1|.$$

在假设 A11 下，对于任意 a 以及所有 z, $\kappa(\boldsymbol{a}) \le p^*(\boldsymbol{a}\mid \boldsymbol{z}) \le K(\boldsymbol{a})$。此外，$p^*(\boldsymbol{a}) = \int p^*(\boldsymbol{a}\mid \boldsymbol{z})\,p^*(\boldsymbol{z})d\boldsymbol{z} \le K(\boldsymbol{a})$，并且根据假设 A11-(4)，$p^*(\boldsymbol{a}) \ge \kappa_A(\boldsymbol{a})$。因此，

$$\frac{\kappa_A(\boldsymbol{a})}{K(\boldsymbol{a})} \le \frac{p^*(\boldsymbol{a})}{p^*(\boldsymbol{a}\mid \boldsymbol{z})} \le \frac{K(\boldsymbol{a})}{\kappa(\boldsymbol{a})},$$

因此，对于任意 i （以 $\boldsymbol{a}_i$ 替代 $\boldsymbol{a}$），

$$\sup_{\boldsymbol{z}}|w^*(\boldsymbol{z},\boldsymbol{a}_i)-1| = \sup_{\boldsymbol{z}}\left|\frac{p^*(\boldsymbol{a}_i)}{p^*(\boldsymbol{a}_i\mid \boldsymbol{z})}-1\right| \le \max\left\{1-\frac{\kappa_A(\boldsymbol{a}_i)}{K(\boldsymbol{a}_i)},\ \frac{K(\boldsymbol{a}_i)}{\kappa(\boldsymbol{a}_i)}-1\right\}.$$

步骤 4（估计误差项——$p(\boldsymbol{z}|\boldsymbol{a})$）。剩余项 $\left|\Phi_{n,\ \mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a})\ -\Phi^*_{n,\ \mathrm{cond}}(\boldsymbol{a})\right|$ 恰好是条件潜在分布下插件泛函的估计误差：

$$\Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a})-\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\int\left[\hat{m}(\boldsymbol{z},\boldsymbol{a})\hat{p}(\boldsymbol{z}\mid \boldsymbol{a}_i)-m^*(\boldsymbol{z},\boldsymbol{a})p^*(\boldsymbol{z}\mid \boldsymbol{a}_i)\right]d\boldsymbol{z}.$$

如有需要，可以通过加上并减去 $m^*(\boldsymbol{z},\ \boldsymbol{a})\hat{}p(\boldsymbol{z}\mid \boldsymbol{a}_i)$，并使用$\|m^*(\cdot\ \boldsymbol{a})\|_\infty \le C_m$, $\|\hat{}m(\cdot\ \boldsymbol{a})\|_\infty \le C_m$, 进一步对其进行界定：

$$\left|\Phi_{n,\mathrm{cond}}(\hat{\boldsymbol{\theta}};\boldsymbol{a})-\Phi^*_{n,\mathrm{cond}}(\boldsymbol{a})\right| \le \epsilon_m(\boldsymbol{a}) + \frac{2C_m}{n}\sum_{i=1}^{n}\|\hat{p}(\boldsymbol{z}\mid \boldsymbol{a}_i)-p^*(\boldsymbol{z}\mid \boldsymbol{a}_i)\|_{\mathrm{TV}},$$

其中 $\epsilon_m(\boldsymbol{a}) := \sup_{\boldsymbol{z}}|\hat{}m(\boldsymbol{z},\ \boldsymbol{a})\ - m^*(\boldsymbol{z},\ \boldsymbol{a})|$。

结合步骤 1–3，可得所述的三项高概率界；步骤 3 则给出了明确的重叠性膨胀贝叶斯桥偏差界。证明完毕。 □

假设 A12 (代理变量条件独立性**)**。 在基本代理变量设定下，假设

$$\boldsymbol{A} \perp \boldsymbol{X} \mid \boldsymbol{Z},$$

等价地，对所有 $(\boldsymbol{a},\boldsymbol{z},\boldsymbol{x})$，$p(\boldsymbol{a}\mid \boldsymbol{z},\boldsymbol{x}) = p(\boldsymbol{a}\mid \boldsymbol{z})$。

因此，我们得到代理变量贝叶斯恒等式：

$$p(\boldsymbol{z}\mid \boldsymbol{x},\boldsymbol{a}) = \frac{p(\boldsymbol{a}\mid \boldsymbol{z})\,p(\boldsymbol{z}\mid \boldsymbol{x})}{p(\boldsymbol{a}\mid \boldsymbol{x})}. \tag{A18}$$

引理 *A6*（代理变量条件 $p(\boldsymbol{z}\mid \boldsymbol{x},\boldsymbol{a})$ 的全变差界）。假设假设 *A11* 和假设 *A12*。固定 $\boldsymbol{x}$。令 $f(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x}) := p(\boldsymbol{a}\mid \boldsymbol{z})p(\boldsymbol{z}\mid \boldsymbol{x})$ 且 $\hat{}\ f(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x}) := \hat{}p(\boldsymbol{a}\mid \boldsymbol{z})\hat{}p(\boldsymbol{z}\mid \boldsymbol{x})$，其归一化常数为 $g := p(\boldsymbol{a}\mid \boldsymbol{x}) = \int f(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x})\ d\boldsymbol{z}$ 且 $\hat{}g := \hat{}p(\boldsymbol{a}\mid \boldsymbol{x}) = \int\hat{}f(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x})\ d\boldsymbol{z}$。那么

$$\|\hat{}p(\boldsymbol{z}\mid \boldsymbol{x},\boldsymbol{a})-p^*(\boldsymbol{z}\mid \boldsymbol{x},\boldsymbol{a})\|_{\mathrm{TV}} \le \underbrace{2\,p^*(\boldsymbol{a}\mid \boldsymbol{x})\|\hat{}f-f\|_1}_{\frac{1}{}} + \underbrace{\|\hat{}2\,p^*(\boldsymbol{a}\mid \boldsymbol{x})\ p(\ )\hat{}(\boldsymbol{a}\mid \boldsymbol{x})}_{f\|_1}|\hat{}p(\boldsymbol{a}\mid \boldsymbol{x})-p^*(\boldsymbol{a}\mid \boldsymbol{x})|.$$

此外，利用假设 *A11*，可得粗略界

$$\|\hat{f}-f\|_1 \le K(\boldsymbol{a})\,\|\hat{p}(\boldsymbol{z}\mid \boldsymbol{x})-p^*(\boldsymbol{z}\mid \boldsymbol{x})\|_1 + \mathbb{E}_{p^*(\boldsymbol{z}|\boldsymbol{x})}\left[|\hat{p}(\boldsymbol{a}\mid \boldsymbol{z})-p^*(\boldsymbol{a}\mid \boldsymbol{z})|\right].$$

证明。 记 $\hat{}p(\boldsymbol{z}\mid \boldsymbol{x},\boldsymbol{a}) = \hat{}f(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x})/\hat{}g$以及 $p^*(\boldsymbol{z}\mid \boldsymbol{x},\boldsymbol{a}) = f(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{x})/g$，其中 $g = p^*(\boldsymbol{a}\mid \boldsymbol{x})$，且 $\hat{}g = \hat{}p(\boldsymbol{a}\mid \boldsymbol{x})$。那么

$$\left\|\frac{\hat{f}}{\hat{g}}-\frac{f}{g}\right\|_1 \le \left\|\frac{\hat{f}-f}{g}\right\|_1 + \left\|\hat{f}\left(\frac{1}{\hat{g}}-\frac{1}{g}\right)\right\|_1 = \frac{1}{g}\|\hat{f}-f\|_1 + \|\hat{f}\|_1\left|\frac{1}{\hat{g}}-\frac{1}{g}\right|.$$

利用 $\left|\frac{1}{\hat{g}}-\frac{1}{g}\right|=\frac{|g-\hat{g}|}{\hat{g}g}$，并除以 2，即可得到所述全变差界。对于 $\|\hat{f}-f\|_1$ 的粗略界，先加上再减去 $\hat{p}(\boldsymbol{a}\mid\boldsymbol{z})p^*(\boldsymbol{z}\mid\boldsymbol{x})$：

$$\|\hat{f}-f\|_1 \le \int \hat{p}(\boldsymbol{a}\mid\boldsymbol{z})\,|\hat{p}(\boldsymbol{z}\mid\boldsymbol{x})-p^*(\boldsymbol{z}\mid\boldsymbol{x})|\,d\boldsymbol{z} + \int p^*(\boldsymbol{z}\mid\boldsymbol{x})\,|\hat{p}(\boldsymbol{a}\mid\boldsymbol{z})-p^*(\boldsymbol{a}\mid\boldsymbol{z})|\,d\boldsymbol{z},$$

并应用 $\hat{p}(\boldsymbol{a}\mid\boldsymbol{z}) \le K(\boldsymbol{a})$。 □

假设 A13 (高斯潜在工作模型稳定性)。 假设潜在变量的 *CI-StoNet* 工作模型服从固定协方差的高斯分布：

$$p(\boldsymbol{z}\mid u;\boldsymbol{\theta}_1)=\mathcal{N}(\mu_1(u;\boldsymbol{\theta}_1),\sigma_z^2 I),$$

其中，$u=\boldsymbol{x}$ 表示代理变量，即 $p(\boldsymbol{z}\mid\boldsymbol{x};\ \boldsymbol{\theta}_1)$。假设 μ_1 关于 $\boldsymbol{\theta}_1$一致满足利普希茨条件：

$$\|\mu_1(u;\hat{\boldsymbol{\theta}}_1)-\mu_1(u;\boldsymbol{\theta}_1^*)\|_2 \le L_1\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^*\|, \qquad \forall u.$$

引理 A7 (Pinsker–高斯全变差界)。 在假设 *A13* 下，对于任意 u，

$$\big\|p(\cdot\mid u;\hat{\boldsymbol{\theta}}_1)-p(\cdot\mid u;\boldsymbol{\theta}_1^*)\big\|_{\mathrm{TV}} \le \frac{L_1}{2\sigma_z}\,\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^*\|.$$

证明。 根据 Pinsker 不等式，$\|P-Q\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|Q)}$。对于具有共同协方差 $\sigma_z^2 I$ 的高斯分布，

$$D_{\mathrm{KL}}\big(\mathcal{N}(\mu^*,\sigma_z^2 I)\,\|\,\mathcal{N}(\hat{\mu},\sigma_z^2 I)\big)=\frac{\|\mu^*-\hat{\mu}\|_2^2}{2\sigma_z^2}.$$

因此

$$\|P-Q\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\cdot\frac{\|\mu^*-\hat{\mu}\|_2^2}{2\sigma_z^2}}=\frac{\|\mu^*-\hat{\mu}\|_2}{2\sigma_z} \le \frac{L_1}{2\sigma_z}\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^*\|,$$

其中最后一步使用了假设 A13。 □

定理 A3 （具有显式重叠膨胀的有限样本界）。 固定 $\boldsymbol{a}\in A$，并假设假设 *A11–A13* 成立。令样本条件因果目标为

$$\Phi_n^*(\boldsymbol{a}):=\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Y(\boldsymbol{a})\mid\boldsymbol{x}_i]=\frac{1}{n}\sum_{i=1}^n\int m^*(\boldsymbol{z},\boldsymbol{a})\,p^*(\boldsymbol{z}\mid\boldsymbol{x}_i)\,d\boldsymbol{z}.$$

定义插件泛函

$$\Phi_n(\hat{\boldsymbol{\theta}};a):=\frac{1}{n}\sum_{i=1}^n\int \hat{m}(\boldsymbol{z},\boldsymbol{a})\,\hat{p}(\boldsymbol{z}\mid\boldsymbol{x}_i)\,d\boldsymbol{z},$$

以及蒙特卡洛近似

$$\hat{\Phi}_{n,M}(\boldsymbol{a}):=\frac{1}{nM}\sum_{i=1}^n\sum_{\ell=1}^M \hat{m}(\boldsymbol{z}_i^{(\ell)},\boldsymbol{a}), \qquad \boldsymbol{z}_i^{(\ell)}\overset{i.i.d.}{\sim}\hat{p}(\boldsymbol{z}\mid\boldsymbol{x}_i).$$

那么对于任意 $\delta\in$ （0，1），以至少 $1-\delta$ 的概率，

$$|\hat{\Phi}_{n,M}(\boldsymbol{a})-\Phi_n^*(\boldsymbol{a})| \le C_m\sqrt{\frac{2\log(2/\delta)}{nM}}\ +\ |\Phi_n(\hat{\boldsymbol{\theta}};\boldsymbol{a})-\Phi_n^*(\boldsymbol{a})|.$$

此外，定义

$$w_i^*(\boldsymbol{z},\boldsymbol{a}):=\frac{p^*(\boldsymbol{a}\mid\boldsymbol{x}_i)}{p^*(\boldsymbol{a}\mid\boldsymbol{z})}, \qquad \hat{w}_i(\boldsymbol{z},\boldsymbol{a}):=\frac{\hat{p}(\boldsymbol{a}\mid\boldsymbol{x}_i)}{\hat{p}(\boldsymbol{a}\mid\boldsymbol{z})},$$

$$\epsilon_{az}(\boldsymbol{a}):=\sup_{\boldsymbol{z}}|\hat{p}(\boldsymbol{a}\mid\boldsymbol{z})-p^*(\boldsymbol{a}\mid\boldsymbol{z})|, \quad \epsilon_{ax}(\boldsymbol{a}):=\frac{1}{n}\sum_{i=1}^n|\hat{p}(\boldsymbol{a}\mid\boldsymbol{x}_i)-p^*(\boldsymbol{a}\mid\boldsymbol{x}_i)|.$$

Then

$$|\Phi_n(\hat{\boldsymbol{\theta}};\boldsymbol{a}) - \Phi_n^*(\boldsymbol{a})| \le \underbrace{\frac{2C_m C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})} \cdot \frac{1}{n}\sum_{i=1}^{n} \|\hat{p}(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a}) - p^*(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})\|_{\mathrm{TV}}}_{\textit{distribution/reconstruction error (overlap inflated)}}$$
$$+ \underbrace{\frac{C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})}\,\epsilon_m(\boldsymbol{a})}_{\textit{outcome nuisance (overlap inflated)}} + \underbrace{\frac{C_m}{\kappa(\boldsymbol{a})}\,\epsilon_{ax}(\boldsymbol{a}) + \frac{C_m C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})^2}\,\epsilon_{az}(\boldsymbol{a})}_{\textit{weight error (overlap inflated)}}.$$

证明。 我们分析误差 $|\hat{\Phi}_{n,M}(\boldsymbol{a}) - \Phi_n^*(\boldsymbol{a})|$,

$$|\hat{\Phi}_{n,M}(\boldsymbol{a}) - \Phi_n^*(\boldsymbol{a})| \le \underbrace{|\hat{\Phi}_{n,M}(\boldsymbol{a}) - \Phi_n(\hat{\boldsymbol{\theta}};\boldsymbol{a})|}_{A_1} + \underbrace{|\Phi_n(\hat{\boldsymbol{\theta}};\boldsymbol{a}) - \Phi_n^*(\boldsymbol{a})|}_{A_2}$$

其中， A_1 表示由蒙特卡洛近似导致的误差， A_2 表示由干扰函数估计导致的误差。

根据霍夫丁不等式，我们有:

$$P(A_1 > \epsilon) \le 2\exp\left\{-\frac{nM\epsilon^2}{2C_m^2}\right\}$$

等价地，以至少 $(1-\delta)$ 的概率,

$$A_1 \le C_m\sqrt{\frac{2\log(2/\delta)}{nM}}$$

根据 (A18)，我们有

$$p^*(\boldsymbol{z} \mid \boldsymbol{x}_i) = p^*(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})\,\frac{p^*(\boldsymbol{a} \mid \boldsymbol{x}_i)}{p^*(\boldsymbol{a} \mid \boldsymbol{z})} = p^*(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})\,w_i^*(\boldsymbol{z}, \boldsymbol{a}),$$

类似地，$\hat{p}(\boldsymbol{z} \mid \boldsymbol{x}_i) = \hat{p}(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})\,\hat{w}_i(\boldsymbol{z}, \boldsymbol{a})$。因此

$$\Phi_n^*(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\int m^*(\boldsymbol{z}, \boldsymbol{a})\,p^*(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})\,w_i^*(\boldsymbol{z}, \boldsymbol{a})\,d\boldsymbol{z},$$
$$\Phi_n(\hat{\boldsymbol{\theta}};\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{m}(\boldsymbol{z}, \boldsymbol{a})\,\hat{p}(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})\,\hat{w}_i(\boldsymbol{z}, \boldsymbol{a})\,d\boldsymbol{z}.$$

对于项 A_2，我们有

$$A_2 \le \frac{1}{n}\sum_{i=1}^{n}\left|\int\left(\hat{m}\,\hat{p}_{ia}\hat{w}_i - m^*\,p_{ia}^*w_i^*\right)d\boldsymbol{z}\right|,$$

其中 $\hat{p}_{ia} = \hat{p}(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})$ 和 $p_{ia}^* = p^*(\boldsymbol{z} \mid \boldsymbol{x}_i, \boldsymbol{a})$。加上并减去 $\hat{m}\,p_{ia}^*\hat{w}_i$，得到$A_2 \le B_1 + B_2$，其中

$$B_1 := \frac{1}{n}\sum_{i=1}^{n}\left|\int \hat{m}(\boldsymbol{z}, \boldsymbol{a})\hat{w}_i(\boldsymbol{z}, \boldsymbol{a})\big(\hat{p}_{ia}(\boldsymbol{z}) - p_{ia}^*(\boldsymbol{z})\big)\,d\boldsymbol{z}\right|,$$

$$B_2 := \frac{1}{n}\sum_{i=1}^{n}\left|\int \big(\hat{m}(\boldsymbol{z}, \boldsymbol{a})\hat{w}_i(\boldsymbol{z}, \boldsymbol{a}) - m^*(\boldsymbol{z}, \boldsymbol{a})w_i^*(\boldsymbol{z}, \boldsymbol{a})\big)p_{ia}^*(\boldsymbol{z})\,d\boldsymbol{z}\right|.$$

界定 B_1。 根据假设 A11,

$$\|\hat{w}_i(\cdot, \boldsymbol{a})\|_\infty \le \frac{\hat{p}(\boldsymbol{a} \mid \boldsymbol{x}_i)}{\inf_z \hat{p}(\boldsymbol{a} \mid \boldsymbol{z})} \le \frac{C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})}.$$

Thus

$$B_1 \le \frac{1}{n}\sum_{i=1}^{n} \|\hat{m}(\cdot,\boldsymbol{a})\|_\infty \|\hat{w}_i(\cdot,\boldsymbol{a})\|_\infty \|\hat{p}_{ia} - p^*_{ia}\|_1 \le \frac{2C_m C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})} \cdot \frac{1}{n}\sum_{i=1}^{n} \|\hat{p}_{ia} - p^*_{ia}\|_{\mathrm{TV}}.$$

界定 B_2。将 $\hat{m}\hat{w}_i - m^* w_i^* = (\hat{m} - m^*) w_i^* + m^*(\hat{w}_i - w_i^*)$ 分解。对于第一部分，

$$\frac{1}{n}\sum_{i=1}^{n} \left| \int (\hat{m} - m^*) w_i^* \, p^*_{ia} \, d\boldsymbol{z} \right| \le \frac{1}{n}\sum_{i=1}^{n} \|w_i^*\|_\infty \, \mathbb{E}_{p^*_{ia}}[|\hat{m} - m^*|] \le \frac{C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})} \, \epsilon_m(\boldsymbol{a}).$$

对于第二部分，利用 $\left|\frac{1}{u} - \frac{1}{v}\right| \le \frac{|u-v|}{u\,v\,\inf(\cdot)^2}$ 和假设 A11，

$$\begin{aligned} |\hat{w}_i(\boldsymbol{z},\boldsymbol{a}) - w_i^*(\boldsymbol{z},\boldsymbol{a})| &= \left| \frac{\hat{p}(\boldsymbol{a} \mid \boldsymbol{x}_i)}{\hat{p}(\boldsymbol{a} \mid \boldsymbol{z})} - \frac{p^*(\boldsymbol{a} \mid \boldsymbol{x}_i)}{p^*(\boldsymbol{a} \mid \boldsymbol{z})} \right| \\ &\le \frac{|\hat{p}(\boldsymbol{a} \mid \boldsymbol{x}_i) - p^*(\boldsymbol{a} \mid \boldsymbol{x}_i)|}{\kappa(\boldsymbol{a})} + \frac{C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})^2} \, |\hat{p}(\boldsymbol{a} \mid \boldsymbol{z}) - p^*(\boldsymbol{a} \mid \boldsymbol{z})|. \end{aligned}$$

因此，

$$\frac{1}{n}\sum_{i=1}^{n} \left| \int m^*(\boldsymbol{z},\boldsymbol{a}) \big(\hat{w}_i(\boldsymbol{z},\boldsymbol{a}) - w_i^*(\boldsymbol{z},\boldsymbol{a})\big) p^*_{ia}(\boldsymbol{z}) \, d\boldsymbol{z} \right| \le \frac{C_m}{\kappa(\boldsymbol{a})} \, \epsilon_{ax}(\boldsymbol{a}) + \frac{C_m C_{ax}(\boldsymbol{a})}{\kappa(\boldsymbol{a})^2} \, \epsilon_{az}(\boldsymbol{a}).$$

结合 B_1 和 B_2 的界，得到所述代理不等式。 □

备注 *A3*（重叠膨胀）。在定理 *A3* 中，有限样本界包含显式的重叠逆因子 $1/\kappa(\boldsymbol{a})$ 和 $1/\kappa(\boldsymbol{a})^2$，，因此，对于固定的干扰项估计误差 $(\epsilon_m, \epsilon_{ax}, \epsilon_{az})$，当 $\kappa(\boldsymbol{a}) \downarrow 0$ 时，界会变差。此外，引理 *A6* 表明，控制 $\|\hat{p}(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{a}) - p^*(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{a})\|_{\mathrm{TV}}$ 可能会引入对 $K(\boldsymbol{a}) = \sup_z p(\boldsymbol{a} \mid \boldsymbol{z})$（及其估计版本）的乘性依赖；当 A 为连续变量时，这一点尤为重要。

A6 实验设置

A6.1 模拟示例

A6.1.1 缺失混杂因素

对于存在缺失混杂因素的情形，网络的隐藏层由两个模块组成。第一个模块以处理变量为输入，并插补潜在混杂因素；第二个模块以插补后的混杂因素与处理变量的拼接向量为输入，对结果进行建模。对于可分离混杂和不可分离情形，第一个模块包含两个层，大小分别为 32 和 6；第二个模块包含两个层，大小分别为 8 和 4。(4) 中噪声项 $\boldsymbol{e}_z$ 和 $\boldsymbol{e}_y$ 的方差分别设为 10^{-5} 和 10^{-3}。训练分为三个阶段——预训练、训练和剪枝后的微调，训练轮次分别为 100、500 和 100。预训练和训练时，网络按照普通的深度神经网络进行训练，但插补学习率 ϵ_k 和网络参数学习率 γ_k 仅从训练阶段开始衰减。训练后，网络经过剪枝，并在微调阶段以较小的学习率进行优化。微调阶段通常是可选的，对整体性能的提升并不显著。

初始插补学习率 ϵ 在不可分离混杂情况下设为 5×10^{-4}，在可分离混杂情况下设为10^{-3}，并按 $\epsilon_k = \frac{\epsilon_k}{1+\epsilon_k \times k^{0.95}}$衰减。初始参数学习率 γ 分别为第一模块的 5×10^{-7} 和第二模块的 5×10^{-6}，并按 $\gamma_k = \frac{\gamma_k}{1+\gamma_k \times k^{0.7}}$衰减。对于混合高斯先验 6，$\lambda_n = 10^{-6}$、$\sigma_0^2 = 10^{-4}$和 $\sigma_1^2 = 10^{-1}$。

A6.1.2 代理变量

对于包含代理变量的情况，网络的隐藏层由三个模块组成。第一个模块以代理变量为输入，并插补潜在混杂因素；第二个模块以插补混杂因素的拼接向量为输入，对处理变量进行建模；第三个模块以处理变量为输入，对结果进行建模。第一个模块包含两层，大小分别为 64 和 32；第二个模块包含一层，大小为 16；第三个模块包含一层，大小为 8。

式（12）中噪声项 $\boldsymbol{e}_z$、 $\boldsymbol{e}_a$ 和 $\boldsymbol{e}_y$ 的方差分别设为 10^{-5}、 10^{-4} 和 10^{-3}。训练分为三个阶段：预训练、训练以及剪枝后的微调，训练轮次分别为 50、100 和 50。

初始插补学习率为 $\epsilon_1 = 10^{-3}$ 和 $\epsilon_2 = 10^{-4}$，并以 $\epsilon_k = \frac{\epsilon_k}{1+\epsilon_k \times k^{0.8}}$ 衰减。三个模块的初始参数学习率分别设为 $\gamma_1 = 5 \times 10^{-6}$、 $\gamma_2 = 5 \times 10^{-5}$ 和 $\gamma_3 = 5 \times 10^{-7}$，并以 $\gamma_k = \frac{\gamma_k}{1+\gamma_k \times k^{0.6}}$ 衰减。对于混合高斯先验（6）， $\lambda_n = 10^{-6}$、 $\sigma_0^2 = 10^{-4}$ 和 $\sigma_1^2 = 10^{-2}$。

A6.2 基准数据集

基准数据集的网络结构与代理变量的网络结构相似。

A6.2.1 ACIC

第一个模块包含两个层，大小分别为 64 和 32；第二个模块包含一个大小为 16 的层，第三个模块包含一个大小为 8 的层。
式（12）中噪声项 $\boldsymbol{e}_z$、 $\boldsymbol{e}_a$ 和 $\boldsymbol{e}_y$ 的方差分别设为 10^{-5}、 10^{-4} 和 10^{-3}。训练分为三个阶段——预训练、训练和剪枝后的微调，轮次分别为 50、100 和 50。

初始插补学习率为 $\epsilon_1 = 5 \times 10^{-3}$ 和 $\epsilon_2 = 5 \times 10^{-4}$，并按 $\epsilon_k = \frac{\epsilon_k}{1+\epsilon_k \times k^{0.8}}$ 衰减。三个模块的初始参数学习率分别设为 $\gamma_1 = 10^{-6}$、 $\gamma_2 = 10^{-5}$ 和 $\gamma_3 = 10^{-7}$，并按 $\gamma_k = \frac{\gamma_k}{1+\gamma_k \times k^{0.6}}$ 衰减。对于混合高斯先验（6）， $\lambda_n = 10^{-6}$、 $\sigma_0^2 = 2 \times 10^{-4}$ 和 $\sigma_1^2 = 10^{-2}$。

A6.2.2 Twins

第一个模块包含两个大小分别为64和32的层，第二个层包含一个大小为16的层，第三个层包含一个大小为8的层。
噪声项 $\boldsymbol{e}_z$、 $\boldsymbol{e}_a$和 $\boldsymbol{e}_y$ 在（12）中的方差分别设为 10^{-3}、 10^{-5}和 10^{-7}。训练包括三个阶段——预训练、训练以及剪枝后的微调，轮次分别为100、1000和200。

初始插补学习率为 $\epsilon_1 = 3 \times 10^{-3}$ 和 $\epsilon_2 = 5 \times 10^{-5}$，并以$\epsilon_k = \frac{\epsilon_k}{1+\epsilon_k \times k^{0.8}}$衰减。三个模块的初始参数学习率分别设为 $\gamma_1 = 10^{-3}$、 $\gamma_2 = 10^{-5}$和$\gamma_3 = 10^{-10}$，并以 $\gamma_k = \frac{\gamma_k}{1+\gamma_k \times k^{0.95}}$衰减。对于混合高斯先验（6）， $\lambda_n = 10^{-6}$、 $\sigma_0^2 = 2 \times 10^{-5}$和 $\sigma_1^2 = 10^{-2}$。