未观测混杂因素检测：核化回归方法

Yikai Chen[1*]、**Yunxin Mao**[1*]、**Chunyuan Zheng**[2]、**Hao Zou**[3]、**Shanzhi Gu**[1]、**Shixuan Liu**[1]、、**Yang Shi**[4]、、**Wenjing Yang**[1]、**Kun Kuang**[5]、**Haotian Wang**[1†]

[1]国防科技大学计算机科学学院，中国长沙 a

[2]北京大学数学科学学院，中国北京 [3]清华大学计算机科学系，中国北京[4]北京大学计算机科学学院，中国北京[5]浙江大学计算机科学与技术学院，中国杭州

chenyikai1998@nudt.edu.cn

arXiv:2601.00200v1 [stat.ML] 1 Jan 2026

摘要

检测未观测混杂因素对于在观察性研究中实现可靠的因果推断至关重要。现有方法要么要求满足线性假设，要么依赖多个异质环境，因此在非线性单环境设定中的适用性有限。为弥补这一不足，我们提出核回归混杂因素检测（KRCD），一种用于在单环境条件下检测非线性观测数据中未观测混杂的新方法。KRCD 利用再生核希尔伯特空间对复杂依赖关系进行建模。通过比较标准核回归与高阶核回归，我们推导出一个检验统计量，其显著偏离零表明存在未观测混杂。从理论上看，我们证明了两个关键结果：首先，在无限样本下，当且仅当不存在未观测混杂因素时，回归系数才相等。其次，有限样本下的差异收敛至均值为零、方差可解析的高斯分布。在合成基准数据集和 Twins 数据集上进行的大量实验表明，KRCD 不仅优于现有基线方法，而且计算效率也更高。

1 引言

因果效应估计在医疗保健（Prosperi et al. 2020; Dahabreh and Bibbins-Domingo 2024; Wang et al. 2024）、公共政策（Matthay and Glymour 2022）以及金融决策（K¨uhne et al. 2022; Vanderschueren 2024）等多个高风险领域中发挥着核心作用。随机对照试验（随机对照试验）虽然仍是因果推断的黄金标准，但在实际应用中往往受伦理问题或可扩展性有限的制约。因此，观察性研究已成为一种主要替代方案，因其具有更广泛的外部效度且成本更低而备受重视（Pearl 2009; Listl, J¨urges, and Watt 2016）。

然而，观察性研究的可靠性关键取决于无混杂假设——这一假设要求所有会混杂处理-结果关系的变量都得到充分测量并纳入考量。当无混杂假设不成立时，虚假相关会削弱基于观测数据得出的因果结论（Angrist、Imbens 和 Rubin，1996年）。直观地说，在减轻未观测混杂因素的影响之前（Lee 和 Lemieux，2010），我们首先需要确定这类混杂因素是否存在，而这正是本文的主要关注点。因此，检测未观测混杂因素是否存在，对于指导后续的方法学决策具有基础性作用。

检测未观测混杂的方法大致可分为两类。第一类需要满足线性模型假设，即所有变量之间存在线性函数关系（Schultheiss, B¨uhlmann, and Yuan 2024; Janzing and Sch¨olkopf 2018; Liu and Chan 2018）。尽管这类假设简单明了，却无法检验，并会严重削弱在非线性数据上的检测性能。第二类方法需要具有机制变化的结构因果模型（结构因果模型）的多个环境，并利用因果机制变化，通过传播特征检测混杂因素（Karlsson and Krijthe 2023;Mameche, Vreeken, and Kaltenpoth 2024; Reddy and N Balasubramanian 2024）。然而在实践中，要获得足够异质且满足因果不变性和机制独立性假设的环境颇具挑战。值得注意的是，这两类方法都对数据结构作出了假设，由此引出了一个重要的开放问题：

> 在单数据环境中，能否检测非线性情形下的未观测混杂因素？

为此，我们提出一种有效的方法，称为**“核回归混杂因素**检测</style”（KRCD），用于检测从单一环境中采集的非线性观测数据中的未观测混杂因素。具体而言，所提出的 KRCD 方法包含三个主要步骤：非线性结构建模、核化回归和假设构造。为了表征非线性数据结构，我们的 KRCD 首先将处理变量（T）、结果变量（Y）和观测协变量（X）映射到再生核希尔伯特空间（RKHS），从而能够以线性内积表示底层的

*这些作者贡献相同。

†通讯作者。版权所有 © 2026，美国人工智能促进协会（www.aaai.org）。保留所有权利。

非线性结构函数。随后，所提出的 KRCD 在 RKHS 中对 Y 关于（T，X）进行两次回归：第一次是标准核化回归，第二次是由 $\|T,\ X\|^2$ 加权的高阶回归，旨在捕捉表明存在未观测混杂的非线性依赖关系。最后，我们构造假设检验：如果两种回归系数之间的归一化差异足够小，则拒绝存在未观测混杂因素的假设。本文的主要贡献可概括如下：

- 我们研究如何在单环境设定下检测非线性数据中的未观测混杂因素。为此，我们提出了一种新方法 KRCD，通过比较再生核希尔伯特空间中的标准核回归与高阶核回归，实现对未观测混杂的高效检测。

- 从理论上讲，我们证明，在无限样本下，当且仅当不存在未观测混杂因素时，标准核回归与高阶核回归的系数才会一致。对于有限样本，两者之差渐近收敛于一个均值为零、渐近方差可解析求得的高斯分布。

- 通过在合成基准数据集和 Twins 数据集上开展大量实验，我们的方法在检测未观测混杂因素方面的表现稳定优于基线方法。

2 相关工作

2.1 单环境设定下的未观测混杂因素检测

检测未观测混杂因素的传统方法——包括工具变量（IV）回归（Angrist、Imbens 和 Rubin，1996年）以及 Durbin-Wu-Hausman 检验（Hausman，1978年）——依赖外部工具变量或估计量比较（Wang 等人，2022；Kou 等人，2025年）。后续创新方法，如 ROCD（Schultheiss、B¨uhlmann 和 Yuan，2024年），源自重加权策略（Buja 等人 2019a,b），通过利用非高斯高阶矩，绕过了对工具变量的要求，并提供了一种无需重采样的 OLS/HOLS 比较框架（Wang 等人，2023a）。对于高维线性模型，谱方法（如 Liu 和 Chan，2018年；Janzing 和 Sch¨olkopf，2018年）通过对回归诱发关联进行谱分析，或检测协方差结构的偏离来发现混杂；其中，Janzing 和 Sch¨olkopf（2018年）明确将混杂与多元线性模型扰动联系起来。在线性模型之外，早期解决非线性混杂问题的尝试依赖较强的结构假设（例如非参数工具变量（Newey 和 Powell，2003））或特定的非高斯分布（例如 LiNGAM（Shimizu 等人，2006年）），但缺乏针对未观测混杂因素的通用非参数检验。尽管这些方法推动了无需工具变量的检测，它们仍从根本上受限于线性模型；而我们的 KRCD 方法克服了这一局限，将混杂因素检测扩展到了非线性关系。

2.2 多环境设定下的未观测混杂因素检测

超越线性约束，近期方法利用不同环境中的因果机制变化，在不依赖因果充分性假设的情况下检测未观测混杂因素。理论研究证明，在机制变化下出现的条件独立性违背可作为混杂的可检验迹象（Karlsson 和 Krijthe，2023）。随后的研究提出了可操作的框架：基于信息论的检测度量（Mameche、Vreeken 和 Kaltenpoth，2024），以及通过机制变化对混杂强度进行非参数量化（Reddy 和 N Balasubramanian，2024）。这些方法将环境变化视为自然实验，以提取混杂信号（Yin 等人，2023年；Wang，2025年；Wang 等人，2023b）。联合因果推断（Mooij、Magliacane 和 Claassen，2020）以及异质性条件下的因果发现（Perry、Von K¨ugelgen 和 Sch¨olkopf，2022）也有助于检测混杂因素。然而，此类方法通常要求存在明确的分布变化或辅助实验数据（Kallus, Puli, and Shalit，2018；Hu 等人，2024），因此难以应用于纯观测、单一情境的设置。相比之下，KRCD 能够在单环境观测数据中稳健地检测混杂，同时保持非参数灵活性，无需多源数据。

2.3 核化因果推断

再生核希尔伯特空间（再生核希尔伯特空间）的理论基础由 Aronszajn 关于再生性质的开创性工作奠定（Aronszajn 1950年），表示定理（Kimeldorf and Wahba 1971年）则使通过核展开进行高效正则化成为可能。该框架支持三种关键的因果策略：（1）核均值嵌入（核均值嵌入）（Hazlett 2016；Muandet 等，2017）用于协变量平衡，但假设不存在未观测混杂因素；（2）基于再生核希尔伯特空间的工具变量回归（Singh, Sahani, and Gretton 2019年；Mastouri 等，2021）可处理已观测的混杂，但需要外部工具变量；（3）潜在状态识别（Zhang, Gong, and Sch¨olkopf 2015年；Pfister 等，2018）无需明确检测混杂因素，即可推断未测量变量。尽管这些方法增强了非参数灵活性，但仍受因果充分性假设或辅助数据依赖的限制。

关键在于，现有的再生核希尔伯特空间方法都无法直接检测标准观测数据中的非线性未观测混杂。当前技术要么忽略潜在偏差（核均值嵌入），要么依赖无法检验的工具变量（核工具变量），要么缺乏偏差量化能力（独立性检验）（Li 等，2023年；Wang 等，2025年a、b）。为弥补这一空白，我们通过一种新颖的表示定理变体，将再生核希尔伯特空间理论与 ROCD框架（Schultheiss、B¨uhlmann 和 Yuan，2024年）统一起来。这一统一为非线性混杂检测构建了数据自适应核基，从而形成了我们的 KRCD 方法。KRCD 首次通过将 ROCD 的高阶矩结构嵌入再生核希尔伯特空间，实现了对非线性未观测混杂因素的直接检测，既不需要工具变量，也不需要多环境数据。

3 预备知识与问题设定

在本文中，我们分别用 T、 X、 Y 和 U 表示处理、观测协变量、结果和未观测混杂因素。我们遵循潜在结果框架，将 $T = t$ 这一分配下 Y 的潜在结果表示为 $Y(t)$。我们用大写字母表示随机变量（例如 T 和 Y），用小写字母表示它们的实现值（例如， t 和 y）。

3.1 先前的问题设定

现有的未观测混杂因素检测方法分为两种受限范式：

- **线性约束：**单环境方法普遍需要参数化函数假设才能实现可识别性。这使得它们无法应用于具有复杂依赖关系的一般非线性系统。
- **多环境依赖：**多环境方法普遍要求结构因果模型中存在可观测的因果机制变化。这从根本上使其无法应用于单环境观测研究。

3.2 我们的问题：非线性单环境（NSE）中的未观测混杂因素检测

我们的目标是检测未观测混杂因素是否存在U 于非线性单环境设定中。在零假设 H_0 下，不存在未观测混杂因素，其结构方程为：

$$H_0 : \begin{cases} T = f_T(X) + \epsilon_T, \\ Y = g(T, X) + \epsilon_Y. \end{cases} \tag{1}$$

相反，在备择假设 H_1 下，存在未观测混杂因素，其特征如下：

$$H_1 : \begin{cases} T = f_T(X) + h_T(U) + \epsilon_T, \\ Y = g(T, X) + h_Y(U) + \epsilon_Y, \end{cases} \tag{2}$$

其中， f_T、 g、 h_T 和 h_Y 是非线性函数， ϵ_T 和 ϵ_Y 是外生噪声项。为便于后续分析，我们定义 $Z = (T, X)^\top$，其中 $\mathbb{E}[\epsilon_T] = \mathbb{E}[\epsilon_Y] = 0$。

注 1 (与先前问题设定的区别**)。**我们提出的方法与以往研究有着根本区别，主要体现在两个方面：

(1) 非参数化表述：我们不对 f_T、 g、 h_T、 h_Y 的函数形式（例如线性形式）或 ϵ_T、 ϵ_Y 的分布作参数化假设，因此我们的方法适用于一般的非线性系统。

*(2)*单环境设定：不同于依赖多个环境的方法（例如随机机制变化或显式干预），我们研究更具挑战性的情形：只能获取来自单一环境的观测数据。

这使我们的工作成为首个利用单环境观测数据检测未观测混杂因素的非参数方法。

4 用于混杂因素检测的核回归

本节提出核回归混杂因素检测法（KRCD）。这是一种新颖且强大的方法，利用再生核希尔伯特空间（RKHS）理论，将残差正交诊断条件（ROCD）框架扩展到高维情境下非参数地检测未观测的非线性混杂。该方法显著增强了基于矩的方法检测线性关系以外关系的能力。

4.1 动机

在可能存在非线性关系的 NSE 设置下，传统的线性检测方法存在根本性局限。这促使我们重新审视 ROCD 框架的核心前提：

回归导向混杂因素检测（ROCD）。在正确的线性设定下，ROCD 利用普通最小二乘（OLS）与高阶最小二乘（HOLS）估计量的渐近等价性。系统性差异表明模型设定有误（例如存在未观测混杂）。该方法：

- 计算普通最小二乘（OLS）和高阶最小二乘（HOLS）系数（Y关于 X，T）。
- 通过标准化差异检验系数是否相等；显著性意味着存在未观测混杂因素。

然而，ROCD仅限于线性回归，无法处理非线性数据。为了将这一范式扩展到非线性情形，我们为非线性函数 $g(Z)$ 定义两类最小二乘估计量：

$$g^{OLS} = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}\left[\|Y - g(Z)\|^2 \right], \tag{3}$$

$$g^{HOLS} = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}\left[\|Y - g(Z)\|^2 \, \|Z\|^2 \right], \tag{4}$$

其中， $\mathcal{H}$ 是一个适当的函数空间。

挑战。 在线性情形下， g 位于有限维欧几里得空间中， g^{OLS} 和 g^{HOLS} 均有显式的向量表示。然而，对于无限维函数空间中的非线性 g，如何表示 g 以进行比较分析和后续检测，仍是一项根本性的理论与计算挑战。

4.2 函数回归的表示定理

为应对这一挑战，我们利用再生核希尔伯特空间中的表示定理。该方法将无限维优化问题转化为关于系数的有限维凸优化，从而能够捕捉表明存在未观测混杂的复杂非线性依赖关系。

再生核希尔伯特空间（RKHS）。（Aronszajn，1950年）再生核希尔伯特空间 $\mathcal{H}_k$ 是一个具有再生核 k 的函数空间，满足 $\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$。其特征映射$\phi(x) = k(\cdot, x)$ 将输入嵌入到 $\mathcal{H}_k$ 中。

表示定理。(Kimeldorf and Wahba 1971年**)** 对于由核 k 诱导的再生核希尔伯特空间 $\mathcal{H}_k$ 中的 g，正则化经验风险的最小化器具有如下形式：$\hat{g}(\cdot) = \sum_{i=1}^{N} \alpha_i \phi_i(\cdot)$,

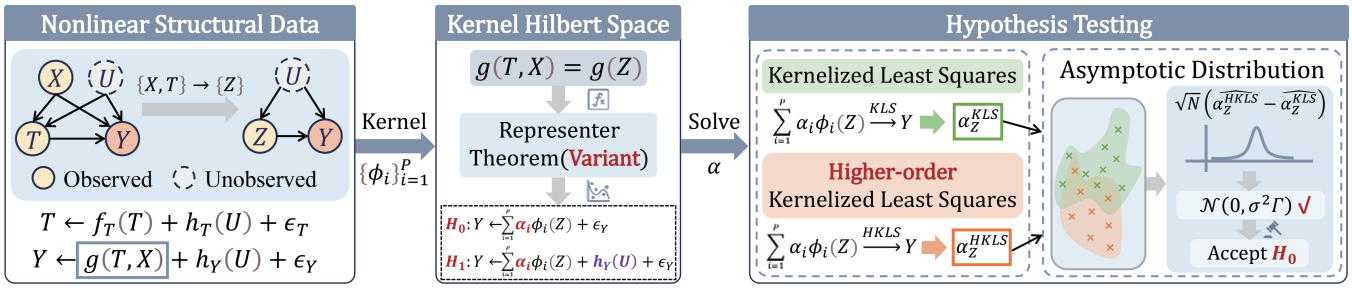


图 1：所提方法的工作流程（KRCD）。

其中 $\phi_i(\cdot) = k(x_i, \cdot)$，$\{x_i\}_{i=1}^N$ 为训练样本点，$\alpha_i \in \mathbb{R}$ 为系数。

遗憾的是，我们观察到此类表示定理无法直接适用于我们的检测问题，因为 $\boldsymbol{\alpha} \in \mathbb{R}^N$ 的维数会随着样本量 N 的增加而变化。因此，随着 N 显著增大（在现实分析中十分常见），通过检查回归系数 $\hat{g}(\cdot)$ 的两个表示元 $\in \mathcal{R}^N$ 来评估所设计检测方法的计算可行性（即复杂度）将变得极其困难，甚至无法实现。

为弥补这一空白，我们建立了一个在具有固定且有限维数的子空间中运作的新型表示定理变体：$P < N$（P 为特征规模）：

定理 1 $X \in \mathbb{R}^{N\times d}$（表示定理变体）。设 为秩为 的设计矩阵，且让$r \le \{\phi_i\}_{i=1}^P$（$r \le P < N$）构成一个 P 维子空间 $\mathcal{S}_P \subseteq \mathcal{H}_k$（再生核希尔伯特空间）的基。考虑正则化经验风险最小化问题：

$$\hat{g}(\cdot) = \underset{g\in\mathcal{S}_P}{\operatorname{argmin}} \left(\frac{1}{N}\sum_{i=1}^{N} \ell(y_i, g(\boldsymbol{x}_i)) + \xi\Omega(\|g\|_{\mathcal{H}_k}) \right),$$

采用凸损失 ℓ、凸非减正则项Ω和 $\xi > 0$。那么，最小化器 $\hat{g}$ 可表示为：

$$\hat{g}(\cdot) = \sum_{i=1}^{P} \alpha_i \phi_i(\cdot), \tag{5}$$

而系数向量 $\boldsymbol{\alpha} \in \mathbb{R}^P$ 则完整编码了基于样本的解。

注 2. 该定理（详细证明见附录 *A.1*）使我们能够将针对 g^{OLS} 和 g^{HOLS} 的无限维函数估计问题，转化为关于 $\boldsymbol{\alpha} \in \mathbb{R}^P$的固定维数优化问题。

4.3 理论基础

借助定理5，我们在一个由对称正定核的特征函数中选取的基 $\{\phi_i\}_{i=1}^P$ 所张成的 P维子空间 $\mathcal{S}_P$内，构造核化最小二乘（KLS）和高阶核化最小二乘（HKLS）估计量。因此，式（3）和（4）中的函数回归可简化为 P 维的 $\boldsymbol{\alpha}_Z^{KLS}$和 $\boldsymbol{\alpha}_Z^{HKLS}$，它们通过求解以下最小化问题得到：

1. **核化最小二乘（KLS）**通过最小化期望平方误差损失：

$$\boldsymbol{\alpha}_Z^{KLS} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^P}{\operatorname{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2 \right]. \tag{6}$$

2. **高阶核最小二乘法（HKLS）**最小化期望的 $\|Z\|^2$加权平方误差损失：

$$\boldsymbol{\alpha}_Z^{HKLS} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^P}{\operatorname{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2 \|Z\|^2 \right]. \tag{7}$$

KRCD 方法的核心原理。 基于这些重新表述，我们现在可以通过逐坐标检查并比较 $\boldsymbol{\alpha}_Z^{KLS}$和 $\boldsymbol{\alpha}_Z^{HKLS}$，检测观测背后非线性且未观测到的混杂因素。具体而言，将 $\boldsymbol{\alpha}_Z^{KLS}$的第 i个坐标记为 $\boldsymbol{\alpha}_Z^{KLS}[i]$，我们证明，$\boldsymbol{\alpha}_Z^{KLS}[i]$ 和 $\boldsymbol{\alpha}_Z^{HKLS}[i]$ 逐坐标相等足以表明存在未观测到的混杂：

定理 2. 在由方程(1)和(2)所定义的因果模型下，总体估计量的等价性满足以下结论：

$$H_0 : \forall i \in [P], \quad \boldsymbol{\alpha}_Z^{KLS}[i] = \boldsymbol{\alpha}_Z^{HKLS}[i]$$
(No unobserved confounder U)

$$H_1 : \exists i \in [P], \quad \boldsymbol{\alpha}_Z^{KLS}[i] \neq \boldsymbol{\alpha}_Z^{HKLS}[i]$$
(Unobserved confounder U present)

附录 A.2 给出了在 H_0 下证明相等、在 H_1 下证明不等的证明。总之，这一定理为 KRCD 奠定了严谨的总体层面理论基础：$\boldsymbol{\alpha}_Z^{KLS}$ 与 $\boldsymbol{\alpha}_Z^{HKLS}$ 之间的系统性差异，可作为未观测混杂存在的有效统计指标。

4.4 有限样本下的假设检验

尽管定理 6 通过比较回归表示函数 $\boldsymbol{\alpha}_Z^{HKLS}$ 和 $\boldsymbol{\alpha}_Z^{KLS}$，为未观测混杂检测提供了理论指导，但有限样本下的算法实现仍面临一项重大障碍。

具体而言，式 (6) 和 (7) 中定义的最小化问题假设样本无限（即总体期望 $\mathbb{E}$），而 $\boldsymbol{\alpha}_Z^{HKLS}$ 和

$\boldsymbol{\alpha}_Z^{KLS}$ 的有限样本估计可能偏离其总体对应值。因此，直接比较这些系数的有限样本估计可能导致错误检测。

为克服这一问题，我们首先基于经验样本 $\{(Z_j,\ Y_j)\}_{j=1}^N$ 推导经验估计量 $\boldsymbol{\alpha}_Z^{KLS}$ 和 $\boldsymbol{\alpha}_Z^{HKLS}$ 的闭式解。利用核基 $k(Z_i,\ \cdot)$（设 $\phi_i(\cdot) = k(Z_i,\ \cdot)$，$\Omega(\|g\|_{\mathcal{H}_k}) = \sum_{i=1}^P \alpha_i^2$，以及 $P < N$），我们得到：

定理 3。 经验 *KLS* 和 *HKLS* 估计量为：

$$\begin{aligned}\widehat{\boldsymbol{\alpha}_Z^{KLS}} &= (K_{ZZ}K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\boldsymbol{Y},\\ \widehat{\boldsymbol{\alpha}_Z^{HKLS}} &= (K_{ZZ}\Psi K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\Psi\boldsymbol{Y},\end{aligned} \tag{8}$$

其中，$K_{ZZ} \in \mathbb{R}^{P\times N}$，且 $[K_{ZZ}]_{ij} = k(Z_i,\ Z_j)$、$\boldsymbol{Y} = [Y_1,\ldots,Y_N]^\top$，以及 $\Psi = \text{diag}(\|Z_1\|^2,\ \ldots,\ \|Z_N\|^2)$。

如前所述，多种风险因素，例如抽样变异和结果噪声，将导致差向量 $\hat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\alpha}_Z^{HKLS}} - \widehat{\boldsymbol{\alpha}_Z^{KLS}}$ 发生扰动；该差向量被用作未观测混杂的诊断标志。因此，下面给出最后一个定理，以说明 $\hat{\boldsymbol{\delta}}$ 的渐近正态性，并进一步指导有限样本的假设检验构造：

定理 4 (渐近分布：$\hat{\boldsymbol{\delta}}$)。 假设标准正则条件（见附录 *A.3*）成立，且因果模型(1)、(2)正确。在 H_0 ($\boldsymbol{\alpha}^{KLS} = \boldsymbol{\alpha}^{HKLS}$) 下，当 $N \to \infty$ 时：

$$\sqrt{N}\hat{\boldsymbol{\delta}} = \sqrt{N}(\widehat{\boldsymbol{\alpha}_Z^{HKLS}} - \widehat{\boldsymbol{\alpha}_Z^{KLS}}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2\Gamma), \tag{9}$$

其中，渐近协方差矩阵 $\sigma^2\Gamma$ 可一致估计。

这一渐近特征使得能够对 H_0 进行实际的假设检验。具体而言，已确立的渐近正态性允许构造分量级标准化检验，其零分布近似为标准正态分布。通过 Bonferroni 校正处理多重检验后，若任何标准化分量差异在调整后的显著性水平上表现出统计显著偏离，则拒绝不存在未观测混杂的零假设。

4.5 复杂度分析

我们推导出所提出的 KRCD 方法的总体复杂度为 $O\left(N^2d + P^3 + P^2N\right)$，以样本量 N、参数维度 P 和特征维度 d 表示（详细推导见附录 A.4）。

算法1：核回归混杂因素检测

输入：观测协变量 $X \in \mathbb{R}^{N\times d}$、处理 $T \in \mathbb{R}^N$、结果 $Y \in \mathbb{R}^N$、参数维度 P、正则化参数 $\lambda = N\xi$。

输出：检测结论

$Z \leftarrow [T, X]$
$K \leftarrow \text{rows}_{1:P}(\text{Kernel Function}(Z))$
$K_\psi \leftarrow K \odot (\text{diag}(ZZ^\top))^\top$
$\alpha_{KLS} \leftarrow \left(KK^\top + \lambda I\right)^{-1} KY$
$\alpha_{HKLS} \leftarrow \left(K_\psi K^\top + \lambda I\right)^{-1} (K_\psi Y)$
$V_0 \leftarrow (K_\psi K^\top + \lambda I)^{-1}K_\psi - (KK^\top + \lambda I)^{-1}K$
$V \leftarrow N \cdot (V_0 V_0^\top)$
$\sigma^2 \leftarrow \frac{1}{N}(Y - K^\top\alpha_{KLS})^\top(Y - K^\top\alpha_{KLS})$
for $j = 1$ **to** P **do**
  $p_j \leftarrow 2(1 - \Phi(|\frac{\sqrt{N}(\alpha_{HKLS,j} - \alpha_{KLS,j})}{\sqrt{\sigma^2 V_{jj}}}|))$
end for
Apply Bonferroni correction: $\delta_j \leftarrow \mathbb{1}[p_j < 0.05/P]$
if $\sum \delta_j = 0$ **then**
  return "Support null hypothesis"
else
  return "Reject null hypothesis"
end if

5 实验

基准测试。 为全面评估 KRCD 检测未观测混杂的能力，我们使用合成数据集和 Twins 数据集（Almond、Chay 和 Lee，2005）开展了模拟研究。除非另有明确说明，所有实验均采用非高斯独立噪声分量，默认正则化参数为 $\lambda = 10^{-8}$。每种参数组合重复 30 次，样本量为 1000 个观测，并按照惯例以 $\alpha = 0.05$ 作为统计显著性阈值。

基线方法。 本文聚焦于最先进的混杂检测方法，以进行比较分析，包括：ROCD（Schultheiss、B¨uhlmann 和 Yuan，2024）、ME-ICM（Karlsson 和 Krijthe，2023）和 CNF（Reddy 和 N Balasubramanian，2024）。（基线方法的详情见附录 B）

评估协议。 我们在多种场景下评估所提出的 KRCD，包括（场景详情见附录 C）：

- 非线性单环境。 所有数据均由固定因果模型下相同的非线性结构因果机制生成。
- 非线性多环境。不同环境在共享的因果框架下呈现不同的非线性因果机制。

评估指标包括：

- *ROC* 曲线和 *AUC*。不同阈值下的灵敏度和假阳性率；AUC 用于量化这种权衡的整体表现。

• 检出率。该方法对未观测混杂因素的检出率，用于量化灵敏度（存在混杂因素时）和假阳性率（不存在混杂因素时）。

• 运行时间。 在固定混杂强度下完成所有检测运行所需的总时间，用于评估计算效率。

研究问题。 实证评估旨在回答以下三个研究问题:

• 在我们的NSE问题中，所提出的KRCD方法是否优于检测基线方法?
• 在多环境设定下，所提出的KRCD方法是否优于检测基线方法?
• KRCD能否在检测隐匿混杂时实现更高的计算效率，胜过基线方法?
• 当真实结构函数不属于核空间时，KRCD能否表现出稳健性?
• KRCD 的性能对正则化参数的变化有多敏感?

5.1 我们的 NSE 设置中的性能

为验证 KRCD 在我们的 NSE 设置中的性能，我们在 Twins 数据集上开展了系统性实验，其中所有环境共享相同的因果机制。

如图 2a 所示，KRCD 在零混杂情况下仍保持稳健，并且在混杂强度达到阈值（≥ 0.5）后表现出较高的灵敏度。图 2c 表明，KRCD 仅用 200 个样本就能检测到未观测混杂，凸显了其样本效率。

图 2b 比较了单环境下的非线性混杂检测。ME-ICM 需要多个环境，因此在此完全失效。ROCD 基于线性假设，会将非线性误判为混杂，导致其在零混杂情况下出现大量假阳性；而在强混杂情况下，由于真实混杂因素掩盖了非线性特征，其检测率也会下降。相比之下，KRCD 在零混杂情况下的检测率保持在 0%（控制第一类错误），并且在混杂强度为 ≥ 1 时检测率超过 95%。

这些结果证实了 KRCD 出色的单环境性能：能够抵御零混杂，在强度超过阈值后具有较高的敏感性，样本需求量低，并且明显优于依赖线性假设或多环境数据的方法。

5.2 多环境下的性能

为验证 KRCD 在多环境设定下的性能，我们进行了系统性实验，其中各环境的因果机制存在分布变化，但因果结构保持一致。由于不同基线方法对环境的定义各不相同，我们分别在 Twins 和合成数据集上进行了评估。

与 ME-ICM 的比较。在 Twins 数据集上，KRCD 的根本区别在于将环境变量作为协变量整合，而非采用 ME-ICM 的显式划分方式。图 3a 显示，两种方法的表现不相上下，说明 KRCD 具有出色的多环境适应能力。

数据集	方法	运行时间（秒/迭代）
Twins	KRCD	68.748
	ME-ICM	2657.838
合成	KRCD	0.184
	CNF-1	0.072
	CNF-2	0.048
	CNF-3	0.046

表1：KRCD 与基线方法的运行时间比较

与 CNF 的比较。在合成数据集上，KRCD 在检测未观测到的混杂因素方面优于 CNF，尤其是在混杂较弱时，图 3b 为此提供了证据。这种性能差距是因为 CNF 的分块方法虽然计算效率高，却会丢失信息，并降低对较弱混杂因素的灵敏度。

5.3 计算效率分析 s

为了评估 KRCD 的计算效率，我们比较了实验中用于图 3a 和图 3b 的所有基线方法的运行时间，该时间定义为进行 100 次试验所需的时间。表1显示，在 Twins 数据集上，KRCD 比 ME-ICM 快得多，显著缩短了计算时间。在合成数据集上，KRCD 的效率与 CNF 相近。

值得注意的是，KRCD 在不同数据集上的效率显著变化，直接源于其参数维度 P：在 Twins 数据集上设置 $P = 999$ 优先考虑建模精度，而在合成数据集上设置 $P = 40$ 则可最大限度地提高计算效率。这种可调性使 KRCD 能够通过调整 P，在预测性能与计算需求之间灵活权衡，从而确保其适应不同的资源约束。

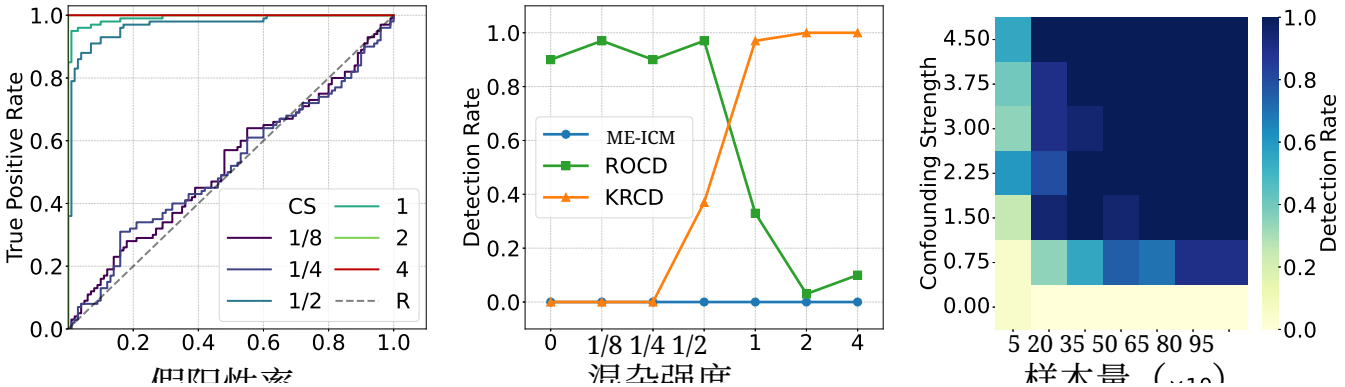
5.4 核设定错误下的稳健性

为评估 KRCD 方法的计算效率，我们比较了图 3a 和图 3b 所示实验中所有基线方法的运行时间，运行时间定义为 100 次试验所需的时间。表1显示，在 Twins 数据集上，KRCD 的速度远快于 ME-ICM，显著缩短了计算时间。在合成数据集上，KRCD 的效率与 CNF 基线方法相近。

如图 3c 所示，KRCD 的两种核配置均严格控制第一类错误（无误报）。即使在混杂程度极低的情况下，多项式-KRCD 也能保持较高的灵敏度；随着混杂强度增加，它仍能实现完全检出。高斯-KRCD 需要更强的信号，但在混杂强度较高时也能实现完全检出。因此，KRCD 稳健可靠，能够在不产生误报的情况下检测未观测混杂，并且在高混杂场景下表现出色。

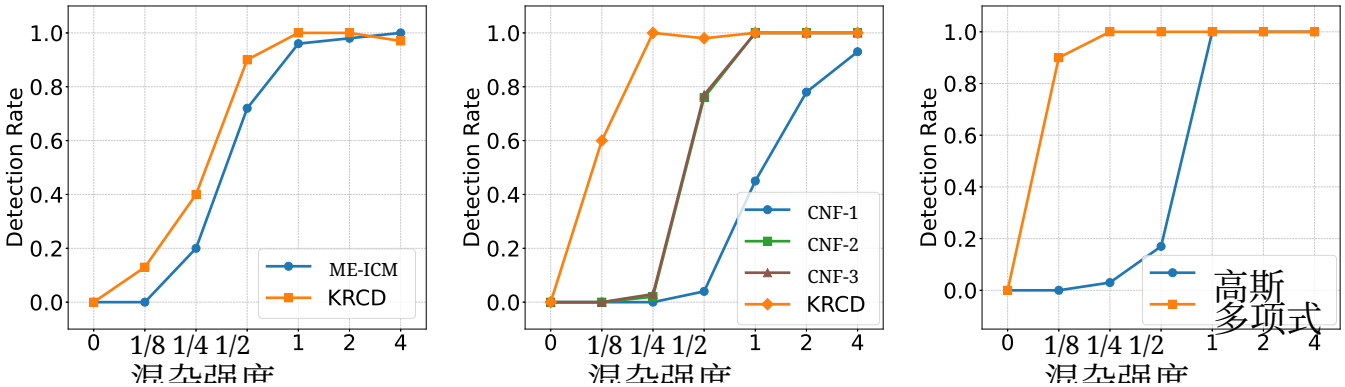
5.5 参数灵敏度分析

算法1中的正则化参数 λ 可防止过拟合，并确保核矩阵求逆过程中的数值稳定性。如表2所示，在广泛的参数范围（10^{-12} 至



(a) (b) (c)

图2：KRCD在非线性单环境中的表现——(a) 不同混杂强度下的ROC曲线。(b) KRCD、ROCD和ME-ICM的检出率。(c) 不同混杂强度和样本量下的检出率。



(a) (b) (c)

图3：(a) KRCD和ME-ICM在非线性多环境中的检出率。(b) KRCD和CNF在非线性多环境中的检出率。(c) 多项式-KRCD和高斯-KRCD的表现。

混杂强度	参数 λ			
	10^{-12}	10^{-8}	10^{-4}	1
0.25	0.468	0.507	0.473	0.404
0.5	0.959	0.980	0.981	0.257
1	0.998	1.000	1.000	0.720
2	1.000	1.000	1.000	1.000

表2：KRCD的AUC性能对 λ的灵敏度

10^{-4}）内，KRCD 对 $\geq$ 0.95 的灵敏度极低，AUC性能保持稳定。在这一宽泛的适用范围内，该方法始终能够实现较高的准确率（混杂强度为 $\geq$0.5 时，AUC 为 $\geq$ 0.95 ），体现出对参数选择的显著稳健性。只有在正则化程度极端时（$\lambda = 1$），性能才会下降；此时过度平滑会掩盖重要的混杂信号，尤其会对具有临床显著效应的中等混杂强度造成不利影响。这种对 λ变化的稳健性证实了 KRCD 在多种场景中的实用性。

6 结论

本文提出核回归混杂因素检测（KRCD）方法，可检测非线性单环境观测数据中的未观测混杂因素，从而弥补现有方法的局限。KRCD 的核心原理是在再生核希尔伯特空间中使用核回归来建模复杂关系；通过比较标准核回归与高阶核回归的结果来构造检验统计量——若该统计量显著偏离零，则表明存在未观测混杂。针对合成基准数据集和 Twins 数据集开展的实验表明，KRCD 不仅在检测性能上显著优于基线方法，而且计算效率也更高。

尽管 KRCD 能够有效检测非线性单环境设置中的未观测混杂，但无法量化混杂强度。一个有前景的研究方向是基于 KRCD 方法开发一种用于量化混杂强度的新框架。

伦理声明

本研究不涉及任何伦理问题。所有实验均使用公开可用的数据集，且不涉及人类受试者或敏感属性。

致谢

本研究部分获得国家自然科学基金委员会项目（项目编号：62372459、62525213和62376243）资助，部分获得国防科技大学青年独立创新科学基金项目（项目编号：ZK25-20）资助，部分获得浙江省“尖兵”“领雁”研发攻关计划（2025C02037）资助。

参考文献

Almond、Chay、K. Y. 和 Lee、D. S.，2005年。低出生体重的代价。《经济学季刊》，120(3)：1031–1083。

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

Aronszajn, N. 1950年。再生核理论。《美国数学会会刊》，68(3)：337–404。Buja, A.; Brown, L.; Berk, R.; George, E.; Pitkin, E.; Traskin, M.; Zhang, K.; and Zhao, L. 2019a。模型作为近似 I。《统计科学》，34(4)：523–544。Buja, A.; Brown, L.; Kuchibhotla, A. K.; Berk, R.; George, E.; and Zhao, L. 2019b。模型作为近似 II。《统计科学》，34(4)：545–565。Dahabreh, I. J.; and Bibbins-Domingo, K. 2024年。医学期刊中基于观察性研究的干预效果因果推断。《美国医学会杂志》，331(21)：1845–1853。Hausman, J. A. 1978年。计量经济学中的设定检验。《计量经济学》：计量经济学会期刊，1251–1271。Hazlett, C. 2016。核平衡：一种用于估计因果效应的灵活非参数加权方法。arXiv:1605.00155。Hu, K.; Yang, W.; Huang, W.; Zhou, X.; Cao, M.; Ren, J.; and Tan, H. 2024。基于序贯融合的时空Transformer跟踪多粒度一致性。发表于*AAAI*人工智能会议论文集，第38卷，12519–12527。

Janzing, D.; 和 Sch¨olkopf, B. 2018. 通过谱分析检测多元线性模型中的混杂。《因果推断杂志》，6(1)：20170013。

Kallus, N.; Puli, A. M.; 和 Shalit, U. 2018. 通过实验奠基消除隐藏的混杂。神经信息处理系统进展，31。

Karlsson, R.; 和 Krijthe, J. 2023. 利用多个环境检测观测数据中的隐藏混杂。神经信息处理系统进展，36：44280–44309。

Kimeldorf, G.；和 Wahba, G. 1971。关于切比雪夫样条函数的一些结果。《数学分析与应用杂志》，33(1)：82–95。Kou, Z.；Qin, S.；Wang, H.；Xie, M.；Chen, S.；Jia, Y.；Liu, T.；Sugiyama, M.；和 Geng, X. 2025。通过学习多标签表示进行带偏置标注的标签分布学习。*arXiv* 预印本 *arXiv:2502.01170*。K¨uhne, F.；Schomaker, M.；Stojkov, I.；Jahn, B.；Conrads-Frank, A.；Siebert, S.；Sroczynski, G.；Puntscher, S.；Schmid, D.；Schnell-Inderst, P.；等。2022。卫生决策中的因果证据：因果推断与卫生决策科学的方法学路径。*GMS*德国医学科学，20：Doc12。Lee, D. S.；和 Lemieux, T. 2010。经济学中的回归不连续设计。《经济文献杂志》，48(2)：281–355。Li, H.；Dai, Q.；Li, Y.；Lyu, Y.；Dong, Z.；Zhou, X.-H.；和 Wu, P. 2023。用于推荐的多重稳健学习。载于《*AAAI*人工智能会议论文集》，第37卷，4417–4425。Listl, S.；J¨urges, H.；和 Watt, R. G. 2016。基于观测数据的因果推断。《社区牙科与口腔流行病学》，44(5)：409–415。Liu, F.；和 Chan, L. 2018。利用谱测度的一阶矩检测高维线性模型中的混杂因素。《神经计算》，30(8)：2284–2318。Mameche, S.；Vreeken, J.；和 Kaltenpoth, D. 2024。识别由因果机制变化导致的混杂。载于《国际人工智能与统计会议》，4897–4905。PMLR。Mastouri, A.；Zhu, Y.；Gultchin, L.；Korba, A.；Silva, R.；Kusner, M.；Gretton, A.；和 Muandet, K. 2021。使用核进行近端因果学习：两阶段估计与矩限制。载于《国际机器学习会议》，7512–7523。PMLR。Matthay, E. C.；和 Glymour, M. M. 2022。社会政策健康影响的流行病学研究中的因果推断挑战与新方向。《当代流行病学报告》，9(1)：22–37。Mooij, J. M.；Magliacane, S.；和 Claassen, T. 2020。来自多个情境的联合因果推断。《机器学习研究杂志》，21(99)：1–108。Muandet, K.；Fukumizu, K.；Sriperumbudur, B.；Sch¨olkopf, B.；等。2017。分布的核均值嵌入：综述与拓展。《机器学习基础与趋势®》，10(1-2)：1–141。Newey, W. K.；和 Powell, J. L. 2003。非参数模型的工具变量估计。《计量经济学》，71(5)：1565–1578。Pearl, J. 2009。《因果关系》。剑桥大学出版社。Perry, R.；Von K¨ugelgen, J.；和 Sch¨olkopf, B. 2022。稀疏机制变化假设下异质环境中的因果发现。《神经信息处理系统进展》，35：10904–10917。Pfister, N.；B¨uhlmann, P.；Sch¨olkopf, B.；和 Peters, J. 2018。联合独立性的基于核的检验。《*The*

英国皇家统计学会*B*辑：统计方法论，80(1)：5–31。

Prosperi, M.; Guo, Y.; Sperrin, M.; Koopman, J. S.; Min, J. S.; He, X.; Rich, S.; Wang, M.; Buchan, I. E.; and Bian, J. 2020年. 机器学习中的因果推断与反事实预测：实现可操作的医疗保健。《自然·机器智能》, 2(7): 369–375.

Reddy, A. G.; and N Balasubramanian, V. 2024年. 利用因果机制变化检测并测量混杂。神经信息处理系统进展, 37: 61677–61699.

Schultheiss, C.; B¨uhlmann, P.; and Yuan, M. 2024年. 高阶最小二乘：评估线性因果模型的部分拟合优度。《美国统计学会杂志》, 119(546): 1019–1031.

Shimizu, S.; Hoyer,P. O.;Hyv¨arinen, A.;Kerminen, A.; and Jordan, M. 2006年. 用于因果发现的线性非高斯无环模型。《机器学习研究杂志》, 7(10).

Singh, R.; Sahani, M.; and Gretton, A. 2019年。核工具变量回归。神经信息处理系统进展，32。

Vanderschueren,T. 2024年。 基于机器学习与因果推断的运营决策。博士论文，安特卫普大学。

Wang, H.; Chen, Z.; Zhang, H.; Li, Z.; Pan, L.; Li,H.; and Gong, M. 2025年a。通过 Wasserstein 因果平衡实现去偏推荐。《*ACM*信息系统汇刊》。

Wang, H.; Kuang, K.; Chi, H.; Yang, L.; Geng, M.; Huang, W.; and Yang, W. 2023年a。结合调整特征选择的处理效应估计。载于第*29*届*ACM SIGKDD*知识发现与数据挖掘大会论文集，2290–2301。

Wang, H.; Kuang, K.; Lan, L.; Wang, Z.; Huang, W.; Wu, F.; andYang, W. 2023年b。利用因果特征分离实现分布外泛化。《*IEEE*知识与数据工程汇刊》，36(4)：1758–1772。

Wang,H.; Yang,W.; Yang, L.; Wu,A.; Xu,L.; Ren,J.; Wu, F.; and Kuang, K. 2022. 使用混杂工具变量估计个体化因果效应。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘大会论文集，1857–1867。

Wang, M.; Li, J.; Su, H.; Yin, N.; Yang, L.; and Li, S. 2024年。GraphCL：用于半监督医学图像分割的基于图的聚类。*arXiv*预印本 *arXiv:2411.13147*。

Wang, M.; Ren, W.; Zhang, Y.; Fan, Y.; Shi, D.; Jing, L.; and Yin, N. 2025年b。用于图域适应的高斯混合模型。载于 Kwok, J. 编，《style id='1'>第三十四届国际人工智能联合会议论文集，IJCAI-25，1963–1972。国际人工智能联合会议组织。主轨道。

Wang, M.-z. 2025年。SimProF：一种用于无监督域适应的简单概率框架。载于*AAAI*人工智能会议论文集，第39卷，21153–21161。

Yin, N.; Shen, L.; Wang, M.; Lan, L.; Ma, Z.; Chen, C.; Hua, X.-S.; and Luo, X. 2023年。Coco：一种用于无监督域适应图分类的耦合对比框架。载于国际机器学习会议，40040–40053。PMLR。

Zhang, K.; Gong, M.; and Sch¨olkopf, B.2015年. 多源域适应：因果视角。载于*AAAI*人工智能会议论文集，第29卷。

A 理论证明

A.1 表示定理变体

定理5 (表示定理变体)。 设 $X \in \mathbb{R}^{N\times d}$ 为秩为 $r < N$ 的设计矩阵，且$\{\phi_i\}_{i=1}^{P}(r \le P < N)$构成再生核希尔伯特空间中一个 P 维子空间 $\mathcal{S}_P \subseteq \mathcal{H}_k$ 的基。考虑正则化经验风险最小化问题：

$$\hat{g}(\cdot) = \underset{g\in\mathcal{S}_P}{\operatorname{argmin}} \left(\frac{1}{N}\sum_{i=1}^{N} \ell(y_i, g(\boldsymbol{x}_i)) + \xi\Omega(\|g\|_{\mathcal{H}_k}) \right),$$

其中凸损失为 ℓ，凸非减正则项为Ω，且 $\xi > 0$。那么，最小化器$\hat{}g$ 可表示为：

$$\hat{g}(\cdot) = \sum_{i=1}^{P} \alpha_i \phi_i(\cdot),$$

系数向量 $\boldsymbol{\alpha} \in \mathbb{R}^P$ 完全编码了基于样本的解。

证明。 证明由两个主要部分组成：

第 ***1*** 部分：存在性与唯一性

1. 有限维参数化：由于 $\mathcal{S}_P$ 是一个具有基 $\{\phi_i\}_{i=1}^{P}$的 P维子空间，每个 $g \in \mathcal{S}_P$都可表示为 $g = \sum_{i=1}^{P} \alpha_i\phi_i$。 因此：

$$\min_{g\in\mathcal{S}_P} J(g) \equiv \min_{\boldsymbol{\alpha}\in\mathbb{R}^P} \left\{ \frac{1}{N}\sum_{i=1}^{N} \ell\left(y_i, \left(\sum_{j=1}^{P}\alpha_j\phi_j\right)(\boldsymbol{x}_i)\right) + \xi\Omega\left(\left\|\sum_{j=1}^{P}\alpha_j\phi_j\right\|_{\mathcal{H}_k}\right) \right\}$$

2. 凸性与存在性：

- $\ell(y_i, \cdot)$ 是凸的，并且 $f_i(\boldsymbol{\alpha}) = \sum_{j=1}^{P}\alpha_j\phi_j(\boldsymbol{x}_i)$ 关于 $\boldsymbol{\alpha}$线性$\Longrightarrow$ 凸
- 正则化项： $\boldsymbol{\alpha} \mapsto \|\sum_{j=1}^{P}\alpha_j\phi_j\|_{\mathcal{H}_k}$ 是一个范数（基于基的线性无关性），因此是凸的。与凸非递增的Ω复合可保持凸性
- $J(\boldsymbol{\alpha})$ 凸函数之和仍为凸函数
- 强制性： $\|\boldsymbol{\alpha}\| \to \infty \implies \Omega(\|g\|_{\mathcal{H}_k}) \to \infty$

根据凸分析，存在一个最小化器 $\boldsymbol{\alpha}^*$。

3. 唯一性：

- 如果 $\ell(y, \cdot)$ 严格凸： $\ell(y_i, f_i(\boldsymbol{\alpha}))$ 严格凸（线性单射 f_i)
- 如果Ω严格凸：$\Omega(\|g\|_{\mathcal{H}_k})$ 在 $\boldsymbol{\alpha}$中严格凸（线性双射映射）

在这两种情况下， J严格凸 $\implies$ 唯一 $\boldsymbol{\alpha}^*$。

第 ***2*** 部分：拟合能力

令$\hat{}g_{class} = \sum_{k=1}^{N} \beta_k k(\boldsymbol{x}_k, \cdot)$ 为经典的表示定理解。由于 $\text{span}\{\phi_i\}_{i=1}^{P} \supseteq \text{Col}(X)$ （因为$\text{rank}(X) = r \leq P$），对于每个 $\boldsymbol{x}_i$：

$$k(\boldsymbol{x}_i, \cdot) = \sum_{j=1}^{P} \gamma_{ij} \phi_j \quad \textit{(data embedding in } \mathcal{S}_P\textit{)}$$

因此，$\hat{}g_{class}$ 在 $\mathcal{S}_P$ 中具有等价表示：

$$\hat{g}_{class} = \sum_{k=1}^{N} \beta_k k(\boldsymbol{x}_k, \cdot) = \sum_{j=1}^{P} \left(\sum_{k=1}^{N} \beta_k \gamma_{kj} \right) \phi_j = \sum_{j=1}^{P} \tilde{\alpha}_j \phi_j$$

由于 $\hat{}g_{class}$ 在 $\mathcal{H}_k$ 中是最优的，且 $\hat{}g_{class} \in \mathcal{S}_P$ 也是如此，受约束的最小化器 $\hat{}g$ 必须对所有 $i = 1$、.....、N 满足：

$$\hat{g}(\boldsymbol{x}_i) = \langle \hat{g}, k(\boldsymbol{x}_i, \cdot) \rangle = \langle \hat{g}_{class}, k(\boldsymbol{x}_i, \cdot) \rangle = \hat{g}_{class}(\boldsymbol{x}_i)$$

因为任何偏离都会破坏 $\hat{}g_{class}$ 或 $\hat{}g$ 在各自空间中的最优性。因此，相对于该样本而言，$\boldsymbol{\alpha}^*$ 完整地编码了 $\hat{}g$。

A.2 未观测混杂因素检测等价性

定理 6。在因果模型下，考虑零假设 H_0 和备择假设 H_1，关于总体估计量的等价性，有如下结论：

$$\begin{aligned} H_0 :& \forall i \in [P], \quad \boldsymbol{\alpha}_Z^{KLS}[i] = \boldsymbol{\alpha}_Z^{HKLS}[i] \\ & \textit{(No unobserved confounder } U\textit{)} \\ H_1 :& \exists i \in [P], \quad \boldsymbol{\alpha}_Z^{KLS}[i] \neq \boldsymbol{\alpha}_Z^{HKLS}[i] \\ & \textit{(Unobserved confounder } U \textit{ present)} \end{aligned}$$

基于表示定理变体，H_0 和 H_1 下的结构方程可重新整理为以下统一的数学表达式：

$$H_0 : Y = \sum_{i=1}^{P} \alpha_i \phi_i(Z) + \epsilon_Y, \tag{10}$$

$$H_1 : Y = \sum_{i=1}^{P} \alpha_i \phi_i(Z) + h_Y(U) + \epsilon_Y, \tag{11}$$

其中，总体回归系数定义为 $\boldsymbol{\alpha}_Z = [\alpha_1, \ldots, \alpha_p]^\top$。

证明。证明分为两个主要部分：部分 ***1***：

$\boldsymbol{\alpha}_Z^{KLS} = \boldsymbol{\alpha}_Z^{HKLS}$ 在 ***.....*** 下 H_0

给定结构方程 (10)，*KLS* 和 *HKLS* 估计量的形式如下：

$$\begin{aligned} \boldsymbol{\alpha}_Z^{KLS} &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^P}{\text{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2 \right] \\ &= (A^{KLS})^{-1} \boldsymbol{b}^{KLS} \\ &\overset{H_0}{=} \boldsymbol{\alpha}_Z, \\ \boldsymbol{\alpha}_Z^{HKLS} &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^P}{\text{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2 \|Z\|^2 \right] \\ &= (A^{HKLS})^{-1} \boldsymbol{b}^{HKLS} \\ &\overset{H_0}{=} \boldsymbol{\alpha}_Z, \end{aligned}$$

其中，矩阵和向量定义如下：

$$\begin{aligned} A_{ji}^{KLS} &= \mathbb{E}\big[\phi_i(Z)\phi_j(Z)\big], \\ \boldsymbol{b}_j^{KLS} &= \mathbb{E}\big[Y\phi_j(Z)\big], \\ A_{ji}^{HKLS} &= \mathbb{E}\big[\phi_i(Z)\phi_j(Z)\,\|Z\|^2\big], \\ \boldsymbol{b}_j^{HKLS} &= \mathbb{E}\big[Y\phi_j(Z)\,\|Z\|^2\big]. \end{aligned}$$

因此，在零假设 H_0 下，参数 $\boldsymbol{\alpha}_Z^{KLS}$ 和 $\boldsymbol{\alpha}_Z^{HKLS}$ 相等。

Part 2: $\boldsymbol{\alpha}_Z^{KLS} \neq \boldsymbol{\alpha}_Z^{HKLS}$ ***under*** H_1

给定结构方程 (11)，*KLS* 和 *HKLS* 估计量的形式为：

$$\begin{aligned} \boldsymbol{\alpha}_Z^{KLS} &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^P}{\text{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2 \right] \\ &= (A^{KLS})^{-1} \boldsymbol{b}^{KLS} \\ &\overset{H_1}{=} (A^{KLS})^{-1} \boldsymbol{c}^{KLS} + \boldsymbol{\alpha}_Z, \\ \boldsymbol{\alpha}_Z^{HKLS} &= \underset{\boldsymbol{\alpha} \in \mathbb{R}^P}{\text{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2 \|Z\|^2 \right] \\ &= (A^{HKLS})^{-1} \boldsymbol{b}^{HKLS} \\ &\overset{H_1}{=} (A^{HKLS})^{-1} \boldsymbol{c}^{HKLS} + \boldsymbol{\alpha}_Z, \end{aligned}$$

其中，矩阵和向量定义如下：

$$\begin{aligned} A_{ji}^{KLS} &= \mathbb{E}\big[\phi_i(Z)\phi_j(Z)\big], \\ \boldsymbol{b}_j^{KLS} &= \mathbb{E}\big[Y\phi_j(Z)\big], \\ \boldsymbol{c}_j^{KLS} &= \mathbb{E}\big[h_Y(U)\phi_j(Z)\big], \\ A_{ji}^{HKLS} &= \mathbb{E}\big[\phi_i(Z)\phi_j(Z)\,\|Z\|^2\big], \\ \boldsymbol{b}_j^{HKLS} &= \mathbb{E}\big[Y\phi_j(Z)\,\|Z\|^2\big], \\ \boldsymbol{c}_j^{HKLS} &= \mathbb{E}\big[h_Y(U)\phi_j(Z)\,\|Z\|^2\big]. \end{aligned}$$

我们将采用数学归纳法证明 $\boldsymbol{\alpha}_Z^{KLS} \neq \boldsymbol{\alpha}_Z^{HKLS}$。

1. 基例：当 $P=1$,

$$\boldsymbol{\alpha}_Z^{KLS}=\frac{\mathbb{E}\big[Y\phi_1(Z)\big]}{\mathbb{E}\big[\phi_1(Z)\phi_1(Z)\big]},$$
$$\boldsymbol{\alpha}_Z^{HKLS}=\frac{\mathbb{E}\big[Y\phi_1(Z)\,\|Z\|^2\big]}{\mathbb{E}\big[\phi_1(Z)\phi_1(Z)\,\|Z\|^2\big]}.$$

为诊断设定错误，重新加权框架（*Buja* 等人 *2019a,b*）使用仅依赖于 $\bar{x}$ 的权重函数 $w(\bar{x})$ 对回归变量分布进行扰动。对于设定正确的泛函，$\theta(P^w_{Y,\,\bar{X}})=\theta(P_{Y,\,\bar{X}})$ 对所有此类 w 均成立。此处，$\boldsymbol{\alpha}_Z^{HKLS}$ 对应于使用 $w(Z)=\|Z\|^2$ 对 $\boldsymbol{\alpha}_Z^{KLS}$ 重新加权：

$$\boldsymbol{\alpha}_Z^{HKLS}=\theta\left(P^w_{Y,\bar{X}}\right)\quad for\quad w(Z)=\|Z\|^2,$$
$$\boldsymbol{\alpha}_Z^{KLS}=\theta(P_{Y,\bar{X}})\quad for\quad w(Z)=1.$$

由于这些泛函设定有误（取决于 U 和 Z的联合分布），且 $w(Z)=\|Z\|^2$ 会以非平凡的方式改变 P_Z（因为 Z是非高斯的），$\theta(P^w_{Y,\bar{X}})\neq\theta(P_{Y,\bar{X}})$。因此：

$$\boldsymbol{\alpha}_Z^{KLS}\neq\boldsymbol{\alpha}_Z^{HKLS}$$

2. 归纳假设：假设对于所有正整数 $P\le M-1$，不等式 $\boldsymbol{\alpha}_Z^{KLS}\neq\boldsymbol{\alpha}_Z^{HKLS}$均成立。***3.*** 归纳步骤：现在证明，当 $P=M$时，该不等式仍然成立。定义中心化估计量：

$$\boldsymbol{\beta}^{KLS}=(A^{KLS})^{-1}\boldsymbol{c}^{KLS}=\boldsymbol{\alpha}_Z^{KLS}-\boldsymbol{\alpha}_Z,$$
$$\boldsymbol{\beta}^{HKLS}=(A^{HKLS})^{-1}\boldsymbol{c}^{HKLS}=\boldsymbol{\alpha}_Z^{HKLS}-\boldsymbol{\alpha}_Z.$$

当 $\boldsymbol{\beta}^{KLS}=\boldsymbol{\beta}^{HKLS}$时，等价性 $\boldsymbol{\alpha}_Z^{KLS}=\boldsymbol{\alpha}_Z^{HKLS}$成立。根据假设，$\boldsymbol{\beta}_{M-1}^{KLS}\neq\boldsymbol{\beta}_{M-1}^{HKLS}$。反设，对于维数为 M的情形，$\boldsymbol{\beta}_M^{KLS}=\boldsymbol{\beta}_M^{HKLS}=\boldsymbol{\beta}_M$ 成立。

将 M 维系统划分为：

$$A_M^{KLS}=\begin{pmatrix}A_{M-1}^{KLS} & \boldsymbol{b}^{KLS}\\ (\boldsymbol{b}^{KLS})^\top & d^{KLS}\end{pmatrix},$$
$$\boldsymbol{c}_M^{KLS}=\begin{pmatrix}\boldsymbol{c}_{M-1}^{KLS}\\ e^{\mathrm{KLS}}\end{pmatrix},$$
$$A_M^{HKLS}=\begin{pmatrix}A_{M-1}^{HKLS} & \boldsymbol{b}^{HKLS}\\ (\boldsymbol{b}^{HKLS})^\top & d^{HKLS}\end{pmatrix},$$
$$\boldsymbol{c}_M^{HKLS}=\begin{pmatrix}\boldsymbol{c}_{M-1}^{HKLS}\\ e^{HKLS}\end{pmatrix},$$

其中：

$$\boldsymbol{b}^{\mathrm{KLS}}=\begin{pmatrix}\mathbb{E}[\phi_1(Z)\phi_M(Z)]\\ \vdots\\ \mathbb{E}[\phi_{M-1}(Z)\phi_M(Z)]\end{pmatrix},$$
$$d^{\mathrm{KLS}}=\mathbb{E}[\phi_M(Z)^2],$$
$$e^{\mathrm{KLS}}=\mathbb{E}[h_Y(U)\phi_M(Z)],$$
$$\boldsymbol{b}^{\mathrm{HKLS}}=\begin{pmatrix}\mathbb{E}[\phi_1(Z)\phi_M(Z)\|Z\|^2]\\ \vdots\\ \mathbb{E}[\phi_{M-1}(Z)\phi_M(Z)\|Z\|^2]\end{pmatrix},$$
$$d^{\mathrm{HKLS}}=\mathbb{E}[\phi_M(Z)^2\|Z\|^2],$$
$$e^{\mathrm{HKLS}(1)}\mathbb{E}[h_Y(U)\phi_M(Z)\|Z\|^2].$$

令 $\boldsymbol{\beta}_M=\begin{pmatrix}\boldsymbol{\beta}^{(1)}\\ \gamma\end{pmatrix}$，其中 $\boldsymbol{\beta}^{(1)}\in\mathbb{R}^{M-1}$、$\gamma\in\mathbb{R}$。前 $M-1$ 个方程为：

$$A_{M-1}^{KLS}\boldsymbol{\beta}^{(1)}+\boldsymbol{b}^{KLS}\gamma=\boldsymbol{c}_{M-1}^{KLS},$$
$$A_{M-1}^{HKLS}\boldsymbol{\beta}^{(1)}+\boldsymbol{b}^{HKLS}\gamma=\boldsymbol{c}_{M-1}^{HKLS}.$$

解得 $\boldsymbol{\beta}^{(1)}$:

$$\boldsymbol{\beta}^{(1)}=(A_{M-1}^{KLS})^{-1}(\boldsymbol{c}_{M-1}^{KLS}-\boldsymbol{b}^{KLS}\gamma),\tag{12}$$
$$\boldsymbol{\beta}^{(1)}=(A_{M-1}^{HKLS})^{-1}(\boldsymbol{c}_{M-1}^{HKLS}-\boldsymbol{b}^{HKLS}\gamma).\tag{13}$$

Equating (12) *and* (13)*:*

$$\begin{aligned}&(A_{M-1}^{KLS})^{-1}(\boldsymbol{c}_{M-1}^{KLS}-\boldsymbol{b}^{KLS}\gamma)\\ =&(A_{M-1}^{HKLS})^{-1}(\boldsymbol{c}_{M-1}^{HKLS}-\boldsymbol{b}^{HKLS}\gamma).\end{aligned}\tag{14}$$

定义

$$\begin{aligned}\boldsymbol{\mu}&=\boldsymbol{\beta}_{M-1}^{KLS}-\boldsymbol{\beta}_{M-1}^{HKLS}\\ &=(A_{M-1}^{KLS})^{-1}\boldsymbol{c}_{M-1}^{KLS}-(A_{M-1}^{HKLS})^{-1}\boldsymbol{c}_{M-1}^{HKLS}\\ &\neq\mathbf{0}.\end{aligned}$$

重排 (14):

$$\boldsymbol{\mu}=\left[(A_{M-1}^{KLS})^{-1}\boldsymbol{b}^{KLS}-(A_{M-1}^{HKLS})^{-1}\boldsymbol{b}^{HKLS}\right]\gamma.$$

令 $\boldsymbol{w}=(A_{M-1}^{KLS})^{-1}\boldsymbol{b}^{KLS}-(A_{M-1}^{HKLS})^{-1}\boldsymbol{b}^{HKLS}$。因此：

$$\boldsymbol{\mu}=\boldsymbol{w}\gamma.$$

由于 $\boldsymbol{\mu}\neq\mathbf{0}$、$\boldsymbol{w}\neq\mathbf{0}$。然而，在非高斯性条件下：

- 该向量 $\boldsymbol{w}$ 包含 $\|Z\|^2$加权的 *HKLS* 分量期望
- 高阶矩无法化简为二阶矩
- 与 Z 之间的依赖关系 U 造成了结构性差异

违反了对齐条件 $\boldsymbol{\mu}=\boldsymbol{w}\gamma$。第 M 个方程：

$$(\boldsymbol{b}^{KLS})^\top\boldsymbol{\beta}^{(1)}+d^{KLS}\gamma=e^{KLS},$$
$$(\boldsymbol{b}^{HKLS})^\top\boldsymbol{\beta}^{(1)}+d^{HKLS}\gamma=e^{HKLS},$$

由于 $\|\boldsymbol{Z}\|^2$ 加权引入了不可约的高阶依赖，这些条件无法同时成立。这一矛盾证明了该定理：

$$\boldsymbol{\alpha}_Z^{KLS}\neq\boldsymbol{\alpha}_Z^{HKLS}\quad for\quad P=M.$$

结论。根据数学归纳法，我们证明：$\boldsymbol{\alpha}_Z^{KLS}\neq\boldsymbol{\alpha}_Z^{HKLS}$在 H_1 条件下。

A.3 渐近分布所需的假设与理论ˆ证明 δ

定理 7. 经验 *KLS* 估计量和 *HKLS* 估计量如下：

$$\widehat{\boldsymbol{\alpha}_Z^{KLS}} = (K_{ZZ}K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\boldsymbol{Y}$$
$$\widehat{\boldsymbol{\alpha}_Z^{HKLS}} = (K_{ZZ}\Psi K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\Psi\boldsymbol{Y}$$

其中 $K_{ZZ} \in \mathbb{R}^{P\times N}$ 与 $[K_{ZZ}]_{ij} = k(Z_i, Z_j)$、$\boldsymbol{Y} = [Y_1, \ldots, Y_N]^\top$ 和 $\Psi = \mathrm{diag}(\|Z_1\|^2, \ldots, \|Z_N\|^2)$。

ˆ**定理 8 (** δ**)**。假设标准正则条件（详见附录 *A.3*）成立，且 H_0 和 H_1 下的因果模型正确。在 H_0（$\boldsymbol{\alpha}^{KLS} = \boldsymbol{\alpha}^{HKLS}$）下，当 $N \to \infty$ 时：

$$\sqrt{N}\hat{\boldsymbol{\delta}} = \sqrt{N}(\widehat{\boldsymbol{\alpha}_Z^{HKLS}} - \widehat{\boldsymbol{\alpha}_Z^{KLS}}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2\Gamma),$$

其中渐近协方差矩阵 $\sigma^2\Gamma$可以一致估计。

我们考虑额外的假设，以便应用中心极限定理。

假设 1（矩阵正则性）。矩阵 K_{ZZ} 满足：

$$\lambda_{\max}\left(K_{ZZ}\Psi K_{ZZ}^\top\right) \le \lambda_{\max},$$
$$\lambda_{\min}\left(K_{ZZ}\Psi K_{ZZ}^\top\right) \ge \lambda_{\min},$$
$$\lambda_{\max}\left(K_{ZZ}K_{ZZ}^\top\right) \le \lambda_{\max},$$
$$\lambda_{\min}\left(K_{ZZ}K_{ZZ}^\top\right) \ge \lambda_{\min},$$

其中 $\lambda_{\max} < \infty$，$\lambda_{\min} > 0$ 为常数。

假设 2 (核与数据的矩条件**)**。核函数 $k(Z_i, Z_j)$ 和数据点 $\{Z_i\}_{i=1}^N$ 满足：

- 核函数有界：

$$\exists C_k > 0 \quad \textit{such that} \quad |k(Z_i, Z_j)| \le C_k \quad \forall Z_i, Z_j.$$

- 数据的矩条件：输入数据满足$\mathbb{E}\left[\|Z_i\|^6\right] < \infty$，且加权矩阵 Ψ的对角元素 $\|Z_i\|^2$ 的四阶矩满足 $\mathbb{E}\left[\|Z_i\|^8\right] < \infty$。

假设 3(误差项条件**)**。误差向量$\boldsymbol{\epsilon}_Y = (\epsilon_{Y,1}, \ldots, \epsilon_{Y,N})^\top$ 满足：

- 独立性：

$$\epsilon_{Y,i} \perp \{z_j\}_{j=1}^N \quad \forall i.$$

- 零均值和有限方差：

$$\mathbb{E}[\epsilon_{Y,i}] = 0, \quad \mathbb{E}[\epsilon_{Y,i}^2] = \hat{\sigma}^2 < \infty.$$

- 四阶矩的存在性：

$$\mathbb{E}[\epsilon_{Y,i}^4] < \infty.$$

假设4 (正则化参数的正则性**)**。正则化参数 ξ 满足：

$$|\xi| \ll \frac{\|K_{ZZ}\Psi K_{ZZ}^\top\|}{N}, \quad |\xi| \ll \frac{\|K_{ZZ}K_{ZZ}^\top\|}{N}.$$

ˆ**证明 (** δ的渐近分布**)**。假设数据遵循零假设 H_0下的模型，

$$\boldsymbol{Y} = K_{ZZ}^\top\boldsymbol{\alpha}_Z + \boldsymbol{\epsilon}_Y,$$

其中，$\mathbb{E}[\boldsymbol{\epsilon}_Y] = \mathbf{0}$、$Var[\boldsymbol{\epsilon}_Y] = \sigma^2 I_N$，且假设*1*至*4*成立。则*HKLS*估计量与*KLS*估计量之差为：

$$\begin{aligned}
&\widehat{\boldsymbol{\alpha}_Z^{HKLS}} - \widehat{\boldsymbol{\alpha}_Z^{KLS}} \\
=&\Big[(K_{ZZ}\Psi K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\Psi \\
&- (K_{ZZ}K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\Big]\Big[K_{ZZ}^\top\boldsymbol{\alpha}_Z + \boldsymbol{\epsilon}_Y\Big] \\
\approx&\Big[(K_{ZZ}\Psi K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\Psi \\
&- (K_{ZZ}K_{ZZ}^\top + N\xi I_P)^{-1}K_{ZZ}\Big]\boldsymbol{\epsilon}_Y.
\end{aligned}$$

当 $N \to \infty$时，我们采用以下概率极限：

$$\frac{1}{N}(K_{ZZ}K_{ZZ}^\top + N\xi I_P) \xrightarrow{P} \mathbb{E}\left[K_{ZZ_i}K_{ZZ_i}^\top\right],$$
$$\frac{1}{N}(K_{ZZ}\Psi K_{ZZ}^\top + N\xi I_P) \xrightarrow{P} \mathbb{E}\left[\|Z_i\|^2 K_{ZZ_i}K_{ZZ_i}^\top\right],$$

由此可得渐近近似：

$$(K_{ZZ}K_{ZZ}^\top + N\xi I_P)^{-1} \approx \frac{1}{N}\left(\mathbb{E}\left[K_{ZZ_i}K_{ZZ_i}^\top\right]\right)^{-1},$$
$$(K_{ZZ}\Psi K_{ZZ}^\top + N\xi I_P)^{-1} \approx \frac{1}{N}\left(\mathbb{E}\left[\|Z_i\|^2 K_{ZZ_i}K_{ZZ_i}^\top\right]\right)^{-1}.$$

将这些结果代入估计量之差，得到：

$$\widehat{\boldsymbol{\alpha}_Z^{HKLS}} - \widehat{\boldsymbol{\alpha}_Z^{KLS}} \approx \frac{1}{N}\sum_{j=1}^N \boldsymbol{v}_j\epsilon_{Y,j},$$

其中 $\boldsymbol{v}_j \equiv \left(\mathbb{E}\left[\|Z_i\|^2 K_{ZZ_i}K_{ZZ_i}^\top\right]\right)^{-1}K_{ZZ_j}\|Z_j\|^2 - \left(\mathbb{E}\left[K_{ZZ_i}K_{ZZ_i}^\top\right]\right)^{-1}K_{ZZ_j}$。因此，

$$\begin{aligned}
\sqrt{N}\hat{\boldsymbol{\delta}} =& \sqrt{N}\left(\widehat{\boldsymbol{\alpha}_Z^{HKLS}} - \widehat{\boldsymbol{\alpha}_Z^{KLS}}\right) \\
\approx& \frac{1}{\sqrt{N}}\sum_{j=1}^N \boldsymbol{v}_j\epsilon_{Y,j} \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \sigma^2\Gamma\right),
\end{aligned}$$

w其协方差矩阵可一致估计为：　　。

$$\hat{\sigma}^2\left(\frac{1}{N}\sum_{j=1}^N v_j v_j^\top\right) \xrightarrow{P} \sigma^2\Gamma.$$

A.4 计算复杂度的推导1. 核矩阵构造：

- Compute $ZZ^\top$: $(N \times d) \times (d \times N)$
  $\Rightarrow O(N^2 d)$
- Element-wise square: $O(N^2)$
- Extract first P rows: $O(1)$

Total: $O(N^2 d)$

2. **加权核计算**:

- Compute $\mathrm{diag}(ZZ^\top) : O(Nd)$
- Element-wise multiplication $(K \odot \cdots)$: Matrix size $P \times N \Rightarrow O(PN)$

Total: $O(PN)$

3. **参数估计**:

- α_{KLS} :
  - $KK^\top : O(P^2N)$
  - Matrix inversion: $O(P^3)$
  - Matrix-vector products: $O(P^2) + O(PN)$
- α_{HKLS} :
  - $K_\psi K^\top : O(P^2N)$
  - Matrix inversion: $O(P^3)$
  - Matrix-vector products: $O(P^2) + O(PN)$

Total: $O(P^3) + O(P^2N)$

4. **协方差矩阵计算**:

- V_0 :
  - Matrix products: $O(P^2N)$
  - Matrix subtraction: $O(PN)$
- $V = N \cdot (V_0 V_0^\top)$: Matrix multiplication $(P \times N) \times (N \times P)$ $\Rightarrow O(P^2N)$

Total: $O(P^2N)$

5. **残差方差**:

- $K^\top \alpha_{KLS}$: O（PN）
- 残差计算: $O(N)$
- 方差计算: $O(N)$ 总计: $O(PN)$

6. **假设检验**:

- 遍历 P 参数: $O(P)$
- Bonferroni 校正: O

(1) 总计: $O(P)$

7. **组合复杂度**:

$$\text{Total} = O(N^2d) + O(PN) + [O(P^3) + O(P^2N)] + O(P^2N) + O(PN) + O(P)$$

主导项:

- N^2d （核构造）
- P^3 (矩阵求逆)
- P^2N (矩阵乘积)

8. **最终表达式**:

$$O\left(N^2d + P^3 + P^2N\right)$$

B 基线详情

B.1 ROCD

回归导向混杂因素检测（ROCD）利用了在正确设定的线性模型下，普通最小二乘（OLS）与高阶最小二乘（HOLS）估计量的渐近等价性（Schultheiss、B¨uhlmann 和 Yuan，2024年）。只有在模型设定错误时（例如存在未观测混杂或非线性），两者之间才会出现系统性差异，从而能够对结构性违背进行诊断性量化。具体而言，ROCD 包含以下三个连续步骤:

- 通过回归 X，T对 Y，计算普通最小二乘和 HOLS 系数;
- 通过标准化差异检验每个预测变量对应的普通最小二乘和 HOLS 系数是否相等。

值得注意的是，回归系数不相等的显著性表明存在未观测混杂因素，这是由于在模型正确设定下其具有渐近正态性。

B.2 ME-ICM

ME-ICM 方法（Karlsson 和 Krijthe，2023年）用于检测多环境观测数据中处理 T 与结果 Y 之间的未观测混杂。其核心思想利用了独立因果机制（ICM）原则，该原则指出，各变量的因果机制在不同环境中独立变化。未观测混杂表现为对环境特定条件独立性的违反：如果不存在未观测混杂因素 U，则对于环境 k 中任意两个不同的观测（i，j），条件 $T_i^{(k)} \perp\!\!\!\perp Y_j^{(k)} \mid T_j^{(k)})$，$X_i^{(k)}$，$X_j^{(k)}$ 必须成立。因此，检测这种独立性违反的统计检验表明存在未观测混杂。

B.3 CNF

CNF 是一类混杂度量，利用多情境观测数据量化变量之间的混杂强度（Reddy and N Balasubramanian，2024年）。

CNF-1. 通过比较观测分布与干预分布来度量混杂，使用定向信息 $I(X_i \to X_j)$，其定义为 KL 散度 $D_{KL}(\mathbb{P}(X_i|X_j)||\mathbb{P}(X_i|do(X_j)))$。若 $I(X_i \to X_j) > 0$ 和 $I(X_j \to X_i) > 0$均成立，则存在混杂（表明存在双向虚假关联）。

CNF-2. 利用未观测混杂因素 Z 的机制变化如何在不同情境下诱发边缘期望 $E_i^C = \mathbb{E}_{\mathbb{P}^c}[X_i]$、$E_j^C = \mathbb{E}_{\mathbb{P}^c}[X_j]$ 之间的依赖关系。若 X_i 和 X_j 共享混杂因素 Z，其特定情境下的期望 E_i^C、E_j^C将成为统计上相互依赖的随机变量。

CNF-3. 当 X_i 与 X_j 之间的因果方向已知时（例如，$X_i \to X_j$），可使用 CNF-2 的扩展版本。它不使用边缘期望，而是使用条件期望$E_{ji}^C = \mathbb{E}_{\mathbb{P}^c}[X_j|X_i]$ 或 $E_{ij}^C = \mathbb{E}_{\mathbb{P}^c}[X_i|X_j]$。在不同情境下，这些条件期望与效应变量的边缘期望之间存在依赖关系，体现了混杂。

C 场景详情

C.1 非线性单环境。

半合成双胞胎数据集。 该数据生成过程基于真实世界的双胞胎数据集，其中的特征经过最小-最大缩放和均值中心化，归一化至 $[0,1]$ 范围。已发现的混杂因素$\mathbf{X}$ = （birmon，brstate，dfageq）与未观测到的混杂因素$\mathbf{U}$ = （meduc6，mplbir，nprevistq）被区分开来。处理 T的生成方式如下：

$$T = \|\mathbf{X}\|_2^2 + \rho\|\mathbf{U}\|_2^2 + \epsilon_1, \quad \epsilon_1 \sim \text{Uniform}(-0.1, 0.1).$$

随后，使用拼接向量 $\mathbf{Z} = [T,\mathbf{X}]$计算结果 Y：

$$Y = \|\mathbf{Z}\|_2^2 + \rho\|\mathbf{U}\|_2^2 + \epsilon_2, \quad \epsilon_2 \sim \text{Uniform}(-0.1, 0.1).$$

这里， $\|\cdot\|_2^2$ 表示欧几里得范数平方，而 ρ控制隐匿混杂强度（$\rho = 0$ 消除混杂）。

C.2 非线性多环境。

半合成双胞胎数据集。 数据生成过程基于一个真实世界的双胞胎数据集，其中通过均值中心化的最小-最大缩放将特征归一化到 $[0,1]$ 范围。将已发现的混杂因素$\mathbf{X}$ =(birmon, brstate,dfageq)与未观测到的混杂因素$\mathbf{U}$ =(meduc6,mplbir, nprevistq)分开。使用 brstate 变量定义环境 $e \in \mathcal{E}$：

$$\mathcal{E} = \{e : |\{i : \texttt{brstate}_i = e\}| \geq 100\}.$$

对于每个环境 e，采样参数：

$$w_{T,x}^{(e)}, w_{T,u}^{(e)}, w_{Y,x}^{(e)}, w_{Y,u}^{(e)} \sim \text{Uniform}(1, 5),$$

$$w_{Y,T}^{(e)} \sim \text{Uniform}(1, 2).$$

对于环境 e_i 中的样本 i：

$$T_i = \sum_{x\in\mathbf{X}} w_{T,x}^{(e_i)} x_i^2 + \sum_{u\in\mathbf{U}} w_{T,u}^{(e_i)} (3\rho u_i)^2 + \epsilon_{1,i},$$

$\epsilon_{1,\,i} \sim$ 均匀分布$(-0.1,\ 0。1)$,

$$Y_i = w_{Y,T}^{(e_i)} T_i^2 + \sum_{x\in\mathbf{X}} w_{Y,x}^{(e_i)} x_i^2 + \sum_{u\in\mathbf{U}} w_{Y,u}^{(e_i)} (3\rho u_i)^2 + \epsilon_{2,i},$$

$\epsilon_{2,\,i} \sim$ 均匀分布$(-0.1,\ 0。1)$。

合成数据集。 该数据生成过程为因果推断分析合成观测数据集和干预数据集。其核心关系通过连续潜变量 T_0 和 Y_0定义。

已发现混杂因素的生成：

$$X \sim \text{Uniform}(0, 1),$$

未观测混杂因素的生成：

$$U \sim \text{Uniform}(0, 1),$$

潜在处理：

$$T_0 = X^2 + 2.5\rho U^2 + \epsilon_1, \quad \epsilon_1 \sim \text{Uniform}(-0.1, 0.1),$$

潜在结果：

$$Y_0 = \|\mathbf{Z}\|_2^2 + 2.5\rho U^2 + \epsilon_2, \quad \epsilon_2 \sim \text{Uniform}(-0.1, 0.1),$$

其中 $\mathbf{Z} = [T, X]$, ρ 控制混杂强度。二元变量通过阈值化得到：

$$T = \mathbb{I}(T_0 > 1), \quad Y = \mathbb{I}(Y_0 > 1).$$

对于干预数据（do-操作），特定变量会被强制设定，同时保留其他关系。观测数据则维持自然关系：

$$T = \mathbb{I}(T_0 > 1),\ Y = \mathbb{I}(Y_0 > 1).$$

对 T ($do(T = t)$)进行干预会覆盖处理机制：

$$T \leftarrow t,\ Y = \mathbb{I}(Y_0 > 1).$$

其中潜在结果 Y_0 仍取决于自然的T_0。对 Y ($do(Y = y)$) 进行干预会覆盖结果机制：

$$T = \mathbb{I}(T_0 > 1),\ Y \leftarrow y,$$

其中潜在处理 T_0 仍保持其对 X 和 U的自然依赖关系。

D 系统规格

所有实验均仅在中央处理器上运行。系统的详细硬件和软件规格如下：

- **处理器：** AMD EPYC 7642 48 核处理器
- **内存：** 90.0 GB DDR4 内存
- **系统类型：** 64 位操作系统，x64 架构处理器

- **操作系统：** Ubuntu22.04LTS
- **图形处理器：** NVIDIA GeForce RTX 3090 24GB GDDR6X PCIExpress 4.0