arXiv:2304.00513v1 [stat.ME] 2 Apr 2023

TSCI：针对无效工具变量进行因果推断的两阶段曲率识别

David Carl Corinne Emmenegger、Peter B¨uhlmann和 Zijian Guo [2]

[1]苏黎世联邦理工学院统计学研讨班 ¨[2]罗格斯大学统计系

摘要

TSCI 在 R统计计算环境中实现了基于观察性数据、存在无效工具变量时的处理效应估计。现有工具变量方法依赖于可能过强且无法检验的识别假设，这限制了其实际应用。TSCI 不需要经典的工具变量识别条件，即使所有工具变量均无效，也能有效发挥作用。TSCI 实现了一种两阶段算法。第一阶段使用机器学习处理处理模型中的非线性关系和交互作用。第二阶段以数据自适应的方式选取一个空间，用于捕捉工具变量违背。随后将这些违背投影出去，以估计处理效应。

关键词：内生性、工具变量、非参数处理模型、处理效应、R。

1 引言：无效工具变量

从观察性研究中推断因果处理效应可能会因未测量混杂因素而受到内生性的影响。一种常见的解决方法是使用工具变量（IVs），以分离处理变量中与未测量混杂因素不相关的变异。然而，要实现有效推断，这些工具变量必须满足严格且无法检验的假设，即下文的 (A2)–(A3)。在给定协变量的条件下，工具变量

(A1) 必须与处理变量具有足够强的关联，(A2) 不得与未测量混杂因素相关，且 (A3) 不得直接影响结果变量。

在实践中，(A2) 和 (A3) 可能存在疑问或无法检验，而实证分析往往依赖外部知识来验证它们，这可能容易出错。因此，开发不依赖这些假设的方法至关重要，因为错误地

依赖这些假设可能会引入严重偏差并使推断无效。我们考虑违反 (A2) 或 (A3) 的工具变量，并将此类工具变量称为无效工具变量。为应对无效工具变量，现有以下配有 R 统计计算环境软件（R Core Team，2021年）的方法。

- 假设工具变量对处理的效应与工具变量对结局的直接效应近似正交；相关软件可在 Bowden等人（2015）的补充材料中获取。

- 假设处理模型存在异方差误差；相关软件可在 Tchetgen Tchetgen等人（2021）的 GitHub 代码库中获取。
- 从一组可能无效的工具变量中筛选有效工具变量；相关软件可在以下 R 软件包中获取：RobustIV（Koo等人，2022b），用于实现稳健因果推断（Guo等人，2018；Guo，2021）；controlfunctionIV（Koo等人，2022a），用于实现控制函数方法（Li和Guo，2020）；以及 CIIV（Frank Windmeijer，2021），用于实现置信区间方法（Windmeijer等人，2021）。有关 RobustIV 和 controlfunctionIV 的实现细节，请参见 Koo等人（2023）。

在 Stata（StataCorp，2021年）模块 SIVREG（Farbmacher，2018年）中也可以找到工具变量筛选的实现。关于这些方法的更多参考文献，也请参见 Guo 和 B¨uhlmann（2022年）。

Guo 和 B¨uhlmann（2022年）提出的两阶段曲率识别（TSCI）方法不受限制（A1）–（A3）的约束，所有工具变量都可能无效。其关键假设相当温和：违背与工具变量和处理之间的关联源自不同的函数形式。也就是说，排除它们恰好都呈线性等情形。该方法使用机器学习来学习可能存在非线性的处理模型。我们的 R 软件包 TSCI 为这一方法提供了软件实现。

对于有效工具变量，已有 R 软件包可利用机器学习从观察性数据中估计处理效应，例如 DoubleML（Chernozhukov等人，2018年；Bach等人，2021年），以及结合额外正则化的机器学习方法 dmlalg（Emmenegger 和 B¨uhlmann，2021年；Emmenegger，2021年）。不过，这些方法采用工具变量与处理之间的线性拟合。相比之下，TSCI 能够处理无效工具变量，并使用机器学习来捕捉处理模型中工具变量与协变量之间复杂的非线性关系和交互项。

本文结构如下。第 2 节介绍统计模型，并从理论层面讨论 TSCI 算法的关键步骤。第 3 节说明如何利用 TSCI 和经验数据，在存在无效工具变量的情况下，根据观察性数据估计并推断处理效应。第 4 节总结 TSCI 的主要功能，第 5 节最后作出结论。

2 两阶段曲率识别

随后，我们介绍 Guo 和 B¨uhlmann（2022年）提出的 TSCI 方法。我们有$i = 1, \ldots\ldots, n$ 个独立同分布观测值。对于工具变量 Z_i和外生基线协变量 X_i，我们考虑如下处理模型

$$D_i = f(Z_i, X_i) + \delta_i, \quad \mathsf{E}[\delta_i | Z_i, X_i] = 0 \tag{1}$$

对于某个未知函数 f，我们考虑结果模型

$$Y_i = \beta D_i + g(Z_i, X_i) + \epsilon_i, \quad \mathsf{E}[\epsilon_i | X_i, Z_i] = 0$$

对于某个未知函数 g。误差项 δ_i 和 ϵ_i 可能相关，从而引入内生性。如果工具变量 Z_i 有效，函数 g 就不会直接依赖于 Z_i。我们的目标是估计处理对结局的效应 β，并对其进行推断。变量 X_i 和 Z_i 可能是多维的。所有其他变量，包括 β，都是一维的。引入 $\phi(X_i) = \mathsf{E}[g(Z_i, X_i)|X_i]$，我们可以将结果模型改写为

$$Y_i = \beta D_i + h(Z_i, X_i) + \phi(X_i) + \epsilon_i, \quad \mathsf{E}[\epsilon_i | X_i, Z_i] = 0$$

其中 $h(Z_i, X_i) = g(Z_i, X_i) - \mathsf{E}[g(Z_i, X_i)|X_i]$。如果工具变量 Z_i有效，则有 $h \equiv 0$，因为 g 不会直接依赖于 Z_i。反过来，如果 h 非零，则 (A2) 或 (A3) 不成立。因此， h 是工具变量违背程度的度量，我们称其为违反函数。

TSCI 估计 β 的关键思想，是利用处理模型 f 与违反函数 h在函数形式上的差异。假设 f 可以用希尔伯特空间 S的某个子空间 S 良好近似，且 $g = h + \phi$可以用由一组基函数 V 张成的某个子空间 Span(V) $\subseteq S$ 良好近似。为应对工具变量违背，我们需要从结果模型中投影掉 g中与 f“非正交”的部分。也就是说，无须投影掉 g的全部内容，即整个 Span(V)；只需投影到一个使 f 与 g “正交”的空间即可。具体而言，我们选择一个满足 S $\cap$Span(V) $\subseteq$ T $\subseteq S$的空间 T，并将内容投影到 T 的正交补上。我们称 T 为违反空间，此类 T 的一个例子是 Span(V) 本身。由于实践中 T 未知，TSCI 会以数据自适应的方式从用户指定的候选空间中选出一个。只要将~~f投影到 T 的~~正交补后，使用 S 对其进行近似仍保留足够的变异性，我们就可以利用这些投影值来克服工具变量违背。特别是，如果 h 关于 Z_i 呈线性关系，而 f关于 Z_i呈二次关系，我们可以利用 f 中关于 Z_i 的非线性部分来估计 β。请注意，在 Guo 和 B¨uhlmann（2022年）的工作中，T 并未被明确定义，而是取为 Span(V)。

为估计处理效应 β，我们需要近似两个对象：函数 f 和违反空间 T，它可能取决于用于近似 f 的方法。函数 f 使用机器学习（或基展开）进行估计；T 则通过检验用户指定的候选项来选取，要求由此得到的（广义）工具变量强度足够大。具体而言，这些候选项在用于近似 h 的基上有所不同，但用于近似 ϕ 的基相同。TSCI 算法分为以下两个阶段：

- 使用机器学习预测处理模型 (1) 中的处理。
- 使用来自上述预测步骤的“帽子矩阵”对结局进行重缩放，将重缩放后的响应和预测处理投影到违反空间的正交补空间（在那里，工具变量违背不再构成问题），然后进行普通最小二乘法。

如果使用机器学习来近似f，则需要额外的数据拆分策略。在经典两阶段最小二乘法（ TSLS）算法（Theil，1953a,b）中，第二步不需要对响应进行重缩放。TSCI 之所以需要重缩放，是因为与 TSLS 不同，“帽子矩阵”可能不是正交的。接下来的章节将详细介绍样本分割、 ϕ的基展开，以及对 Span(V)、 f的估计，最后介绍 β。随后，我们用粗体表示观测值的按行拼接；例如，$\mathbf{h}$ 表示向量（$h(Z_1,\ X_1)\ldots,\ h(Z_n,\ X_n))^\top$。

2.1 样本分割与 f的机器学习估计量

若使用随机森林或提升法等灵活的机器学习算法来拟合处理模型，则样本分割对于消除过拟合造成的偏差至关重要。样本分割可防止过拟合；因为在过拟合的情况下，处理变量的内生性可能无法被完全消除。我们将样本索引划分为两个集合 A_1 和 A_2。默认情况下，它们的大小分别约为 $2n/3$ 和 $n/3$，且不失一般性，令 $A_1 = \{1, 2, \ldots, 2n/3\}$。通过对数据 A_2 应用某种机器学习算法，再在 A_1上进行预测，可构造 $i \in A_1$ 对应的 $f(Z_i, X_i)$ 估计值。我们将这些预测称为 $\mathbf{f}_{\mathcal{A}_1}$。也就是说，此步骤不使用处理变量 $\mathbf{D}_{\mathcal{A}_2}$ ^，其中 $\mathbf{D}_{\mathcal{A}_2}$ 包含 D 中属于 A_2 的元素。估计得到的 $\mathbf{f}_{\mathcal{A}_1}$^ 必须表示为“帽子矩阵”与处理变量向量的乘积，记为 $\mathbf{f}_{\mathcal{A}_1} = \Omega\mathbf{D}_{\mathcal{A}_1}$，其中 $\Omega \in \mathbb{R}^{|\mathcal{A}_1|\times|\mathcal{A}^1|}$ 不一定是正交矩阵。若将随机森林解释为加权最近邻方法，并使用由此产生的加权函数进行预测，则这种表示适用于随机森林（Lin a nd Jeon，2006）。事实上，对于包含 S 棵决策树的随机森林，Ω 中对应于 $i, j \in A_1$ 的 (i, j) 项为

$$\mathbf{\Omega}_{i,j} = \frac{1}{S}\sum_{s=1}^{S}\omega_j(Z_i, X_i, \theta_s) \quad \text{for weights} \quad \omega_j(z, x, \theta_s) = \frac{\mathbb{1}_{\{(Z_j,X_j)\in\mathcal{R}_{l(z,x,\theta_s)}\}}}{\sum_{k\in\mathcal{A}_1}\mathbb{1}_{\{(Z_k,X_k)\in\mathcal{R}_{l(z,x,\theta_s)}\}}},$$

其中，$R_{l(z,x,\theta_s)}$ 表示包含 (z, x) 的第 s 棵树的叶节点，而 θ_s 表示决定第 s 棵树如何生长的随机参数。对于以回归树为基学习器的提升法（B¨uhlmann and Hothorn, 2007; B¨ uhlmann and Yu, 2006），也有类似的表示。

在 TSCI 中，实现了三种计算$\mathbf{\Omega}$的方法：随机森林、以回归树为基学习器的 L_2 提升法，以及多项式基展开。前两种方法较为灵活，能够捕捉包含非线性关系 f 的复杂函数

和交互作用。这是理想的，因为如果能够捕捉到 f 中的更多变异（机器学习算法通常可以做到这一点），那么得到的工具变量强度就会更大。（广义）工具变量强度衡量的是在 ~~A_1 中，处理的变异与估计~~的处理误差方差之比，并且该比值是在调整违规空间候选项之后计算的（Guo 和 B¨uhlmann，2022年）。如果能捕捉到 f 中与违背正交的更多变异，那么在投影去除违反空间后， f 中仍会保留足够的变异。保留变异能够提高该方法的效率，也让我们有更大的空间来检验工具变量无效性。检验无效性时，我们的目标是选择能够解决该违背的最小可能违反空间。检验更大的空间需要更高的工具变量强度。在特殊情况下，如果 f 中与违背正交的部分可以用多项式很好地捕捉，那么使用多项式方法近似 f 会有所帮助。此时，违规空间序列的选择会特别简单，因为只需考虑由相应多项式基中的元素张成的子空间。然而，多项式展开不适用于二元工具变量。

除了 TSCI 中实现的三种 Ω 估计方法外，用户也可以提供单独的矩阵。此外，只有在使用机器学习方法估计 f 时才需要样本分割；采用多项式基的方法则不需要。下文使用样本分割的记号。

2.2 违反空间选择与基展开

我们使用一组基函数 $\vec{}$V 来近似违反函数 h，并使用另一组基函数 $\vec{}$W 来近似 ϕ（更准确地说，是其中与 f 不正交的部分，即 S ∩Span(V)），定义在 A_1 上。

我们以依赖于数据的方式选择违反空间。给定一列 $Q+1$ 个理想情况下彼此嵌套的基函数集合 $\{V_q\}_{0\le q\le Q}$，使得对于每个 $0 \le q \le Q$，都有 $V_q \subset (\vec{}V \cup \vec{}W)$，且 $V_0 = \vec{}W$，我们针对每个候选空间 V_q， $0 \le q \le Q_{\max} \le Q$ 估计处理效应 β；其中 $Q_{\max}$ 表示最后一个候选空间，并且所有 $q \le Q_{\max}$ 个候选空间均被认为具有足够的工具变量强度。随后，对每个处理效应估计量进行检验，将其与所有后续估计量比较，以判断差异是否显著。最后，选择第一个候选违反空间 $V_{q_{\text{comp}}}$；相对于所有 $V_{q'}$， $q_{\text{comp}} < q' \le Q_{\max}$，均未发现显著差异，其中下标“comp”表示比较。也就是说，选择满足以下条件的最小空间：从该空间开始，无法拒绝处理效应估计无偏这一假设。该方法还会检验 (A2) 或 (A3) 是否确实遭到违反，因为第一个候选空间 $V_0 = \vec{}W$ 表示 (A2) 或 (A3) 均未被违反。因此，如果未选择 V_0，则说明存在违反，因为无违反这一假设已被拒绝。我们使用自助法检验来检验工具变量强度是否超过某个阈值，从而判断工具变量是否足够强（Guo 和 B¨uhlmann，2022年）。该检验也适用于异方差误差。其明确形式见 Guo 和 B¨uhlmann（2022年）。

在 fin ite 样本中，某些违背可能无法通过这种方法检测到。这可 y

如果调整所选违反空间后，违背造成的剩余偏差足够小，以致处理估计值之间没有显著差异，就可能出现这种情况。因此，TSCI 还实现了一种更保守的选择合适违反空间的方法。我们将上文所述的前一种方法称为比较方法，将后一种方法称为保守方法。如果工具变量足够强，保守方法不会选择序列中第一个未发现显著差异的违反空间候选项，而是选择其后继项，因此该方法更加保守。

我们将适合用于近似 ϕ 的基的选择留给用户。如果用户不熟悉如何选择合适的基，可以使用 TSCI_poly 中实现的多项式基展开方法，省略协变量，或将它们视为可能无效的额外工具变量。

2.3 估计处理效应

最后，我们将用于近似 $\mathbf{h}_{\mathcal{A}_1}$ 的基函数与用于逐行近似 $\boldsymbol{\phi}_{\mathcal{A}_1}$ 的基函数拼接起来，组成违背矩阵 $\mathbf{V}_{\mathcal{A}_1}$。因此，$\mathbf{V}_{\mathcal{A}_1}$ 可视为包含 S ∩Span(V) 的一组基，我们将包含用于近似 $\boldsymbol{\phi}_{\mathcal{A}_1}$的基的 $\mathbf{V}_{\mathcal{A}_1}$子矩阵记为 $\mathbf{W}_{\mathcal{A}_1}$ 。处理效应估计量通过将 $\boldsymbol{\Omega}\mathbf{Y}_{\mathcal{A}_1}$ 和 $\mathbf{f}_{\mathcal{A}_1} = \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}$ 投影到由 $\mathbf{V}_{\mathcal{A}_1} = \boldsymbol{\Omega}\mathbf{V}_{\mathcal{A}_1}$张成空间的正交补上，然后对投影后的量进行普通最小二乘法得到。使用 $\boldsymbol{\Omega}$ 对违背矩阵 $\mathbf{V}_{\mathcal{A}_1}$ 进行缩放，是因为 $\mathbf{Y}_{\mathcal{A}_1}$ 也使用$\boldsymbol{\Omega}$进行了缩放；后者是必要的，因为 $\boldsymbol{\Omega}$ 不一定是正交的。$\mathbf{V}_{\mathcal{A}_1}$ 张成的空间是 T 的一个近似。此外，为了校正因内生性导致的 $\boldsymbol{\epsilon}_{\mathcal{A}_1}$ 与 $\mathbf{D}_{\mathcal{A}_1}$ 之间任何残留的相关性，还要从该效应量估计量中减去 Guo 和 B¨uhlmann（2022年）明确给出的偏差校正项，从而得到最终估计量

$$\hat{\beta} = \frac{\mathbf{Y}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}}{\mathbf{D}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}} - (\text{bias term}). \tag{2}$$

为阐明该估计量的原理，我们将式（2）中的第一项改写为

$$\begin{aligned} \beta \cdot \frac{\mathbf{D}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}}{\mathbf{D}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}} + \frac{\mathbf{h}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}}{\mathbf{D}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}} + \\ \frac{\boldsymbol{\phi}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}}{\mathbf{D}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}} + \frac{\boldsymbol{\epsilon}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}}{\mathbf{D}_{\mathcal{A}_1}^\top \boldsymbol{\Omega}^\top \mathbf{P}^\perp_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}\mathbf{D}_{\mathcal{A}_1}}. \end{aligned} \tag{3}$$

(3)中的第一项可简化为真实因果效应 β。(3)中的第二项是由无效工具变量导致的偏差ˆ，然而，如果 $\mathbf{h}_{\mathcal{A}_1}$ 中不与$\mathbf{f}_{\mathcal{A}_1}$ 正交的部分

能够由违背矩阵$\mathbf{V}_{\mathcal{A}_1}$的列空间良好近似，则该偏差较小。类似地，如果 $\phi_{\mathcal{A}_1}$ 中不与$\hat{\mathbf{f}}_{\mathcal{A}_1}$ 正交的部分能够由$\mathbf{V}_{\mathcal{A}_1}$ 的列空间良好近似，或者更具体地说，能够由子矩阵$\mathbf{W}_{\mathcal{A}_1}$的列空间良好近似，则(3)中的第三项较小。最后，(2)中的偏差校正项用于校正由(3)中的第四项引起的潜在过拟合偏差，因为在存在内生混杂时， $\epsilon_{\mathcal{A}_1}$ 与$\mathbf{D}_{\mathcal{A}_1}$ 相关。在过拟合的情况下，即使在将$\tilde{\epsilon}_{\mathcal{A}_1}$ 和$\mathbf{D}_{\mathcal{A}_1}$ 按$\mathbf{\Omega}$ 重新缩放并随后投影到$\mathbf{V}_{\mathcal{A}_1}$的正交补之后，这种相关性仍可能部分残留。

2.4 推断

$\hat{}$ 如果$\mathbf{V}_{\mathcal{A}_1}$ 能够很好地逼近$\mathbf{h}_{\mathcal{A}_1}$ 和 $\phi_{\mathcal{A}_1}$ 中与$\mathbf{f}_{\mathcal{A}_1}$ 不正交的部分，且工具变量足够强，则可以证明，在满足$\widehat{}$某些正则条件时，$(\hat{\beta}-\beta)/\widehat{\text{SE}}(\mathbf{V}_{\mathcal{A}_1}) \xrightarrow{d} \text{N}(0,1)$，其中 SE($\mathbf{V}_{\mathcal{A}_1}$) 表示处理效应估计量标准差的一个估计量。用户可以选择采用 Guo 和 B¨uhlmann（2022年）提出的标准差估计量，或采用另一种自助法（见附录 B）。采用自助法的动机在于，它有望提供更准确的有限样本检验和置信区间，这与经典理论一致，也有一些实证证据支持。无论采用哪种方法，都会使用处理效应估计量的渐近高斯近似来计算置信区间和 p 值。

2.5 抵消样本分割的随机性

将样本随机分为两部分，这可能会影响处理效应估计量。因此，整个过程会重复多次。不同重复中的处理效应估计值通过取中位数进行汇总，用户可以选择两种方案之一来计算汇总后的 p 值和置信区间（以及下文所述“DML”方案中的标准误）。第一种方法在 TSCI 中称为“FWER”，通过控制家族错误率来获得有效的 p 值，最初由 Meinshausen 等人（2009年）提出。所得的 p 值可能较为保守。第二种方法在 TSCI 中称为“DML”，直接汇总各次重复的标准误估计量，最初由 Chernozhukov 等人（2018年）提出。

3 使用 TSCI

随后，我们使用两个实证数据集来说明 TSCI 的应用。

3.1 示例 1

应用 TSCI 的第一步是为处理模型选择一个学习器。TSCI 提供四种内置选项，详见表1。它们最重要的输入参数列于表2。

函数	描述
tsci_forest	使用随机森林拟合处理模型。要求用户指定一组违背空间候选项。
tsci_boosting	使用提升法拟合处理模型。要求用户指定一组违背空间候选项。
tsci_poly	使用多项式基展开拟合处理模型。不要求用户指定一组违背空间候选项。
tsci_secondstage	不适用于处理模型，而是使用用户提供的帽子矩阵。 要求用户指定一组违背空间候选项。

表1：TSCI 用于处理效应大小估计的四项主要功能。

随后，我们使用 Card（1993）的数据集演示如何使用 TSCI。该数据集包含 $n = 3010$ 个观测值，研究旨在使用与四年制大学（nearc4）的距离作为工具变量，估计教育（educ）对对数收入（lwage）的因果效应。作为外生基线协变量 X，我们考虑工作经验（exper、expersq）、种族（race）和地理信息（smsa、smsa1966、south、reg661– reg668）。该数据集可在 R 包 ivmodel（Kang 等，2021）中获取，Guo 和 B¨uhlmann（2022）也曾对其进行分析。

```
R> library("TSCI")
R> library("ivmodel")
R> data("card.data")
```

应用 TSCI 的第一步是从表1中选择一个函数。下面我们演示如何使用 tsci_forest，随后使用 tsci_secondstage。函数 tsci_boosting 和 tsci_poly 的结构与 tsci_forest 类似。应用 TSCI 的第二步是选择一种合适的基线协变量变换，以近似 ϕ。变换后协变量矩阵可传入输入参数 W。该矩阵将包含在每个违反空间候选项中。如果未指定 W，则改为加入未经变换的基线协变量。最后，必须提供一个列表，其中包含用于近似 h 的不同候选项。默认情况下，tsci_forest 会根据此列表创建一个嵌套的违背空间候选项序列：首先将 W 作为第一个违背空间候选项，随后将列表中的下一个元素加入当前违背矩阵，以创建下一个违背空间候选项。例如，若要将工具变量 Z 的最高三次多项式与未经变换的基线协变量 X 一起指定为违背空间候选项，只需向 tsci_forest 传入一个包含元素 {Z，Z^2,Z^3} 的列表。无论 Z 有多少列，此表示法均适用。或者，函数 create_monomials 会通过调用 create_monomials(Z = Z, degree = 3) 自动创建此序列。在我们的 TSCI 演示中，我们将

参数	描述
Y	结果变量的数值向量。
D	处理变量的向量。
Z	工具变量的矩阵或数据框。
X	用于处理模型的协变量矩阵或数据框。
W	基线协变量 X 的（变换后）观测值，用于 结果模型。
违背_空间	用于指定违背空间候选项的列表。
创建_嵌套_序列	逻辑值。如果为 TRUE，则定义违背空间候选项 依次从 W 开始，再逐步加入下一个 违反_空间中的元素。如果为 FALSE，则违反空间的候选 值定义为将 W 加入违反_空间中的元素。
sel_方法	用于估计处理效应的选择方法。 任意一者 “比较”或“保守”；参见第2.2节。
split_prop	用于拟合结果模型的观测值比例。 必须是 (0,1) 中的一个值。
sd_boot	逻辑值。 如果为 TRUE，则使用 自助法确定标准误。
iv_阈值	工具变量强度检验的最小阈值。
阈值_boot	逻辑值。如果为 TRUE，则工具变量强度的阈值将被确定—— 使用自助法调整工具变量强度的估计 估计误差，从而得到更保守的 工具变量强度检验。
nsplits	数据划分次数。
mult_split_method	计算标准误、 p 值的方法，以及用于构 造置信区间的方法（如果进行了多重拆分）。 “FWER”或“双重机器学习”；参见第2.5节。

表2：表1中 TSCI 四个主要函数最重要的参数。

考虑一个长度为二的列表，其中第一个列表元素对应工具变量 nearc4，第二个元素对应 nearc4 与基线协变量的交互作用：

```
R> vio_space <- with(card.data, list(nearc4,
+    nearc4 * cbind(exper, expersq, black, south, smsa, smsa66, reg661,
+    reg662, reg663, reg664, reg665, reg666, reg667, reg668)))
```

```
R> ## alternatively:
R> # vio_space <- create_interactions(card.data$nearc4, as.matrix(X))
```

还有一个函数 create_interactions 可用于创建相同的列表；请参阅上面的 R 代码中的注释。若要使用非嵌套的违背空间候选项，我们需要将 tsci_forest 的输入参数 create_nested_sequence 设为 FALSE。在这种情况下，提供列表中的每个元素都会被单独视为一个违背空间候选项，而不会添加之前的元素。指定违背空间候选项后，我们可以调用 tsci_forest 来估计处理效应：

```
R> set.seed(10)
R> Xname <- c("exper", "expersq", "black", "south", "smsa", "smsa66",
+     "reg661", "reg662", "reg663", "reg664", "reg665", "reg666",+     "reg667",
"reg668")R> fit_forest <- tsci_forest(Y = card.data$lwage, D = card.data$educ,
+     Z = card.data$nearc4, X = card.data[, Xname], vio_space = vio_space)
```

有一种 summary 方法可用于概述最相关的统计量，并且其形式对于 tsci_forest、tsci_boosting、tsci_poly 和 tsci_secondstage 均相同。

```
R> summary(fit_forest)
```

关于数据划分过程的统计量：样本量 A1：2007 样本
量 A2：1003 数据划分次数：500 汇总方法：家族错
误率

关于工具变量有效性的统计量：有效 无效 不可_检验 262
238 0

所选违反空间候选项的处理效应估计值：估计值 标准_误 2.5% 97.5% Pr(>|t|)

TSCI-估计值 0.05752 . 0.02703 0.08757 4e-05
选择方法：比较

处理模型的统计量：估计方法：随机森林

关于违背空间选择的统计量：q_comp q_cons Qmax
q0 262 0 0

```
q1      41     262     0
q2     197     238   500
```

上述汇总输出的第一部分提供了有关数据划分过程的信息。首先给出了用于估计处理模型（A1）和结果模型（A2）的数据点数量。可通过输入参数 split_prop 调整其规模。除非另有指定，否则会进行 10 次数据划分（输入参数 nsplits），以抵消数据划分引入的随机性，并使用“FWER”方法（输入参数 mult_split_method）汇总每次数据划分所得的估计量。接下来，输出展示了检验工具变量无效性的结果。在上述输出中，tsci_forest 执行的 500 次数据划分中，有 262 次判定工具变量有效，238 次判定其无效。也就是说，有 238 次检验发现假设（A2）或（A3）不成立。对于除假设不存在违背的空间之外的所有违背空间候选项，如果工具变量强度过弱，无法估计处理效应，则相应的数据划分次数列在 non_testable 项下。在本例中，500 次数据划分中均未出现这种情况。接下来，输出显示了由选定的违背空间候选项得到的处理效应，以及用于获取帽子矩阵 **Ω**的方法。默认情况下，使用自助法获取标准误。最后，输出显示每个违背空间候选项被比较方法（第一列）和保守方法（第二列）选中的次数（共 500 次），以及每个候选项作为工具变量强度足够时所能支持的最大违背空间候选项的次数（第三列）。如果该违背空间候选项恰好也是比较方法选中的候选项（因此也会被保守方法选中），则应谨慎解读处理效应估计值。事实上，由于无法可靠检验使用更大的违背空间候选项时处理效应估计值是否会显著变化，违反假设（A2）或（A3）造成的偏差可能未得到充分处理。用户可以通过指定输入参数 sel_method，选择使用比较方法还是保守方法估计处理效应。默认使用比较方法，本例也是如此。若要获取更详细的输出摘要，可将 extended_output= TRUE 设为真。随后会针对每个违背空间候选项列出处理效应估计值、估计的工具变量强度，以及判定工具变量强度足够的阈值。如果进行了多次数据划分，由选定的违背空间候选项得到的处理效应估计值可能与这些点估计值均不一致，因为每次数据划分中选中的违背空间候选项可能不同。可以看到，本例确实如此。

```
R> summary(fit_forest, extended_output = TRUE)
```

关于数据拆分程序的统计量：样本量 A1：2007

样本量 A2：1003 数据划分次数：
500 汇总方法：家族错误率

关于工具变量有效性的统计量：
有效 无效不可_检验 262 238 0

所选违背空间候选项的处理效应估计值：估计值 标准_误差 2.5% 97.5% Pr(>|t|)
TSCI- 估计值 0.05752 . 0.02703 0.08757 4e-05 选择方法：比较

所有违反空间候选项的处理效应估计值：估计值 标准_误差 2.5% 97.5%
Pr(>|t|) TSCI-q0 0.06293 . 0.03389 0.09196 0.00000 TSCI-q1 0.06185 .
0.03243 0.09179 0.00001 TSCI-q2 0.05102 . 0.01925 0.08296 0.00077

处理模型的统计量：估计方法：随机森林

关于违背空间选择的统计量：q_comp q_cons Qmax
q0 262 0 0 q1 41 262 0 q2 197 238 500

关于工具变量强度的统计量：

	工具变量_强度	工具变量_阈值
q0	130.08	40
q1	126.78	40
q2	99.98	40

提取相关信息的其他方法包括 coef（返回处理效应估计值）和 cofint（返回指定置信水平下处理效应估计值的置信区间）。

如果有理由认为未经变换的基线协变量无法充分近似 ϕ sufficiently well，我们可以设置 tsci_forest 的函数参数 W。在这种情况下，原始基线协变量 X 用于拟合处理模型，而变换后的基线协变量 W 用于结果模型。例如，可以考虑使用列-

wise 个体协变量的基样条拼接 :

```
R> suppressPackageStartupMessages(library("fda"))
R> X <- card.data[, c("exper", "expersq")]
R> head(X, 4)   # covariates
```

经验 经验平
方 116 256 2 9
81 316 256 410
100

```
R> nknots <- 2  # number of knots for bspline basis
R> norder <- 3  # order of spline
R> nbasis <- nknots + norder - 2 # number of basis functions
R> W <- matrix(ncol = NCOL(X) * nbasis, nrow = NROW(X))
R> for (j in seq_len(NCOL(X))) {
+    knots <- quantile(unique(X[, j]), seq(0, 1, length = nknots))
+    basis <- create.bspline.basis(rangeval = range(knots), breaks = knots,
+      norder = norder)
+    W[, c(((j - 1) * nbasis + 1) : (j * nbasis))] <- eval.basis(X[, j], basis)
+  }
R> head(W, 4) # basis transformation of covariates
```

```
    [,1]      [,2]      [,3]      [,4]      [,5]       [,6][1,] 0.0926276
0.4234405 0.4839319 0.2663262 0.4994836 0.23419013[2,] 0.3705104 0.4763705
0.1531191 0.7172073 0.2593473 0.02344546[3,] 0.0926276 0.4234405 0.4839319
0.2663262 0.4994836 0.23419013[4,] 0.3194707 0.4914934 0.1890359 0.6576627
0.3066027 0.03573458
```

3.1.1 指定个体帽子矩阵

函数 tsci_secondstage 提供了指定个体帽子矩阵**Ω**的灵活性。如果我们对处理模型有充分了解，从而能够指定一个比随机森林或提升方法能捕捉更多处理变异的模型，这会很有优势；如果我们希望将违背投影到某个特定空间中（例如，使用加性模型以避免需要检验违背中的交互作用），这也会很有用。

我们再次使用 Card（1993）的数据集来演示 tsci_secondstage 的用法。Card（1993）讨论了最初选定的工具变量——大学邻近性——可能因家庭背景而异的情况。在这种情况下，大学邻近性就不是有效的工具变量，因为它除了通过教育产生影响外，还会影响收入。为了演示 tsci_secondstage 的用法，我们将探究大学邻近性与

家庭背景之间的交互作用是否是有效的工具变量。为此，我们将处理模型指定为基线协变量与新工具变量的线性组合。然后，我们可以使用 tsci_secondstage 检验假设（A2）或（A3）是否被违背，并获得处理效应估计量。按照 Card（1993）的做法，我们使用基于种族、1966年的地理信息以及14岁时的家庭构成信息，预测附近没有大学时的教育水平，并将其作为家庭背景的代理变量。有关如何获得这一预测教育水平的详细信息，请参阅 Card（1993）。首先，我们指定帽子矩阵 **Ω** 和违反空间候选项。这里，我们使用普通最小二乘拟合所得的投影矩阵；该拟合通过将教育对交互工具变量和协变量 X（包括截距项）进行回归得到。我们指定违反空间候选项，以检验大学邻近性是否违背假设（A2）或（A3）。由于处理模型设定中未包含交互作用，因此也无需将其纳入违反空间候选项。

```
R> Z <- with(card.data,+            cbind(nearc4,
+                   nearc4 * family_background))
R> vio_space <- list(card.data$nearc4)
R> Xname <- c("exper", "expersq", "black", "south", "smsa", "smsa66",
+    "reg661","reg662", "reg663", "reg664", "reg665", "reg666", "reg667",
+    "reg668", "fatheduc", "fatheduc_na","motheduc", "motheduc_na",
+    "parenteduc", "momdad14", "sinmom14", "step14")
R> X <- card.data[, Xname]R> X$motheduc_na <- as.numeric(X$motheduc_na)
R> X$fatheduc_na <- as.numeric(X$fatheduc_na)
R> A <- cbind(1, as.matrix(Z), as.matrix(X))
R> omega <- A %*% chol2inv(chol(t(A) %*% A)) %*% t(A)
```

接下来，我们调用 tsci_secondstage，并将我们的帽子矩阵传递给输入参数 weigh t.

```
R> fit_secondstage <- tsci_secondstage(Y = card.data$lwage,
+    D = card.data$educ, Z = Z, W = X,+    vio_space = vio_space,
 weight = omega)
R> summary(fit_secondstage, extended_output = TRUE)
```

有关数据拆分程序的统计量：样本量：3010 未执行样
本分割。

关于工具变量有效性的统计量：有效 无效 不可_检验 0 0 1

所选违背空间候选项的处理效应估计值：估计值 标准_误差 2.5% 97.5% Pr(>|t|)
TSCI- 估计值 0.13125 0.03452 0.06359 0.19891 0.00014 选择方法：比较

所有违反空间候选项的处理效应估计值：估计值 标准_误差 2.5% 97.5%
Pr(>|t|) TSCI-q0 0.13125 0.03452 0.06359 0.19891 0.00014 TSCI-q1
0.12508 0.04525 0.03639 0.21377 0.00571

处理模型的统计量：估计方法：用户指定

违背空间选择的统计量：q_comp q_cons Qmax q0 1
1 1 q1 0 0 0

工具变量强度的统计量：工具变量_强
度 工具变量_阈值 q0 40.21 38.30 q1
25.24 34.93

正如摘要输出的最后一部分所示，在此处理模型设定下，工具变量强度不足以可靠地检验假设（A2）或（A3）是否被违背。

3.2 示例 2

随后，我们展示如何使用 TSCI 在第二个实证数据集中评估工具变量的有效性。Angrist和Krueger（1991）估计了美国1920年至1929年间出生的男性接受义务教育对周收入对数的影响（美国商务部，1992）。义务教育法要求学生在达到法定退学年龄之前留在学校。因此，年初出生、入学时年龄较大的学生，可能会在受教育历程中较早的阶段辍学。Angrist和Krueger（1991）估计了由出生季度差异所导致的教育差异（该差异源于义务教育法）对收入的影响。根据大学生数据，Angrist和Krueger（1991）推断，出生季度不会因义务教育以外的原因影响收入。然而，假设（A2）或（A3）可能遭到违背，因为工具变量可能与未测量混杂因素相关，而这些因素同时与教育和收入相关

例如能力及其他因素。因此，我们旨在使用 TSCI 评估该工具变量是否确实有效。为此，我们考虑一种关于工具变量呈线性的违背。

我们考虑以下变量。响应变量 Y 为周工资对数（LWKLYWGE），处理变量 D 为受教育年限（EDUC），工具变量由 30 个虚拟变量构成，表示 3 个出生季度与 10 个出生年份之间的交互项。Angrist和Krueger（1991）使用这一教育工具变量，以纳入教育中的跨年季节性变化。他们的估计值还在工资方程中纳入了 10 个年份虚拟变量，因此，两阶段最小二乘估计中用于识别教育回报的教育变异完全来自出生季节之间的差异。观测到的基线协变量 X 包括 9 列表示出生年份的虚拟变量、年龄（AGE）、婚姻状况（MARRIED；1 表示已婚）、种族二元指标（RACE；1 表示黑人）、在标准大都市统计区工作的指标（SMSA；1 表示市中心），以及 8 个居住地区虚拟变量（NEWENG、MIDATL、ENOCENT、WNOCENT、SOATL、ESOCENT、WSOCENT、MT）。我们已将完整数据（美国商务部，1992）作为数据框 df 加载到当前会话中；所考虑的变量如下：

```
R> # 3 quarter of birth dummies
R> QOB <- as.matrix(df[, c("QTR2", "QTR3", "QTR4")])
R> # 10 year of birth dummies
R> YOB <- as.matrix(df[, paste0("YR", 20:29)])
R> # 30 interaction dummies as instrumental variables
R> Z <- do.call(apply(QOB, 2, FUN = function(quarter)
+    create_interactions(quarter, YOB)[[2]], simplify = F),
+    what = cbind)R> # 9 year of birth dummiesR> YOB <- YOB[, -1]
R> # outcomeR> Y <- df$LWKLYWGER> # treatmentR> D <- df$EDUC
R> # baseline covariates
R> X <- cbind(YOB, df$RACE, df$MARRIED, df$SMSA, df$NEWENG, df$MIDATL,
+    df$ENOCENT, df$WNOCENT, df$SOATL, df$ESOCENT, df$WSOCENT, df$MT)
R> colnames(X) <- c(paste0("YR", 21:29), "race", "married", "smsa", "n

+    "midatl", "enocent", "wnocent", "soatl", "esocent", "wsocent", "m
eweng",
t")
```

注意，Z 和出生年份呈线性相关。因此，对于处理模型，我们可以删去 X 的前 9 列，但结果模型中必须保留这些列。实际上，工具变量应仅为出生年份与季度之间的交互作用，而不应包含主

效应。我们使用 tsci_forest 评估工具变量的有效性。为此，我们考虑一个违规空间序列：其第一个元素仅包含基线协变量，对应于有效工具变量；第二个元素则是工具变量本身，对应于检验线性违规。我们考虑对数据进行 1 次划分（nsplits），并使用 100 棵树（num_trees），最小节点大小为 5（min_node_size）。我们考虑采用自助法近似得到的标准误估计，这是默认设置（sd_boot = TRUE）。

```
R> set.seed(10)R> vio_space <- list(Z)
R> fit_angrist_tsci <- tsci_forest(Y = Y, D = D, Z = Z,
+     X = X[, -seq_len(9)], W = X, vio_space = vio_space, nsplits = 1,
+     min_node_size = 5, num_trees = 100)
```

以下扩展摘要输出表明，工具变量实际上是无效的，因为选择了更大的违反空间（q1）；这一点也由“无效”项下的 1 表示。默认情况下，使用比较方法来选择违反空间。在此违反空间下，处理效应估计量为 0.14669，自助法标准误为 0.00389。因此，在 5% 显著性水平下，该估计量与 0 显著不同。

```
R> summary(fit_angrist_tsci, extended_output = TRUE)
```

有关数据划分过程的统计量：样本量 A1：164799 样本量 A2：82400 数据划分次数：1 聚合方法：双重机器学习

工具变量有效性统计量：
有效 无效不可_检验 0 1 0

所选违背空间候选项的处理效应估计值：估计值 标准_误差 2.5% 97.5% Pr(>|t|) TSCI- 估计值 0.14669 0.00389 0.13906 0.15432 0 选择方法：比较

所有违背空间候选项的处理效应估计值：估计值 标准_误差 2.5% 97.5% Pr(>|t|) TSCI-q0 0.14402 0.00363 0.13690 0.15114 0 TSCI-q1 0.14669 0.00389 0.13906 0.15432 0

处理模型的统计量：估计方法：随机森林

关于违反空间选择的统计量：q_comp q_cons Qmax
q0 0 0 0 q1 1 1 1

关于工具变量强度的统计量：工具变量
_强度 工具变量_阈值 q0 2282.93 40
q1 2134.61 40

4 实践指南

本节总结 tsci_forest、tsci_boosting 和 tsci_poly 之间的差异。如果对 f 知之甚少，可以采用随机森林和提升法机器学习方法。如果认为 f 中与违背正交的部分能够由多项式很好地捕捉，则可以采用多项式方法 tsci_poly。采用机器学习方法时，需要更加谨慎地选择违背空间序列。采用多项式方法时，可以使用多项式基来指定违背空间候选项。

5 总结与讨论

TSCI 的主要目标是提供一种用户友好的两阶段曲率识别（TSCI）实现（Guo 和 B¨ uhlmann，2022年）。TSCI 方法填补了工具变量（IV）回归文献中的一个重要空白：它提供了一种数据驱动的方法，用于检验工具变量的无效性，并提供一种能够抵御此类违背的效应量估计量。具体而言，所有工具变量都可能无效。与现有的无效工具变量处理方法不同，TSCI 只作出一个非常温和的假设：处理模型与工具变量违背具有不同的函数形式。该方法采用机器学习拟合处理模型，从而能够捕捉复杂的非线性关系和交互作用。不过，使用机器学习时，不应将 TSCI 视为黑箱算法，因为需要运用专家知识判断违背可能具有的函数形式。

计算细节

所有使用的程序包均可从综合 R 档案网络（CRAN）获取，网址为 https://CRAN.R-project.org/。第 3.2 节的结果使用 R 4.1.3 得出。TSCI 2.0.1 使用了 ranger 0.13.3（Wright 和 Ziegler，2017年）中的随机森林实现、xgboost 1.6.0.1（Chen 等，2022年）中的提升法实现，以及 Rfast 2.0.6（Papadakis 等，2022年）中的快速矩阵乘法实现。本文中的其他所有结果均使用 R 4.2.3 得出。TSCI 2.0.1 使用了 ranger 0.14.1（Wright 和 Ziegler，2017年）中的随机森林实现、xgboost 1.7.5.1（Chen 等，2022年）中的提升法实现，以及 Rfast 2.0.7（Papadakis 等，2022年）中的快速矩阵乘法实现。此外，本文使用 fda 6.0.5（Ramsay 等，2022年）生成 B 样条基。我们使用 MASS 7.3.58.3（Venables 和 Ripley，2002年）生成多元高斯数据。使用 tsci_forest 计算帽子矩阵的计算复杂度关于观测值数量呈二次关系，而使用 tsci_boosting 时则呈三次关系。因此，对于大型数据集，使用 tsci_boosting 可能会比使用 tsci_forest 慢很多。

致谢

我们感谢袁伟提供随机森林代码的初始实现，该代码是在袁伟担任 ZG 在罗格斯大学的研究助理期间开发的。我们感谢郑梦初（现为 ZG 在罗格斯大学的研究助理）提出的宝贵意见。CE 和 PB 获得了欧洲研究理事会（ERC）的资助，该资助来自欧盟地平线 2020 研究与创新计划（资助协议编号 786461）。ZG 的研究部分得到美国国家科学基金会数学科学部 NSF-DMS 1811857、2015373 以及美国国立卫生研究院 NIH-1R01GM140463-01 的支持；ZG 还感谢苏黎世联邦理工学院数学研究所（FIM）为其访问提供的经费支持。

参考文献

J. D. Angrist 和 A. B. Krueger。义务入学会影响受教育程度和收入吗？《经济学季刊》，106(4):979–1014，1991年。P. Bach、V. Chernozhukov、M. S. Kurz 和 M. Spindler。DoubleML——R 中双重机器学习的面向对象实现，2021年。预印本 arXiv:2103.09603。J. Bowden、G. Davey Smith 和 S. Burgess。使用无效工具变量进行孟德尔随机化：通过埃格回归进行效应估计和偏倚检测。《国际流行病学杂志》，44(2):512–525，2015年6月。

P. B¨uhlmann 和 T. Hothorn。提升算法：正则化、预测与模型拟合。《统计科学》，22(4):477–505，2007年。P. B¨uhlmann 和 B. Yu。稀疏提升。《机器学习研究杂志》，7:1001–1024，2006年。D. Card。利用大学邻近性的地理差异估计受教育程度的回报”。工作论文 4483，美国国家经济研究局，1993年10月。T. Chen、T. He、M. Benesty、V. Khotilovich、Y. Tang、H. Cho、K. Chen、R. Mitchell、I. Cano、T. Zhou、M. Li、J. Xie、M. Lin、Y. Geng、Y. Li 和 J. Yuan。xgboost：极端梯度提升，2022年。URL https://CRAN.R-project.org/package=xgboost。R 软件包版本 1.6.0.1 和 1.7.3.1。V. Chernozhukov、D. Chetverikov、M. Demirer、E. Duflo、C. Hansen、W. Newey 和 J. Robins。用于处理效应和结构参数的双重 / 去偏机器学习。《计量经济学杂志》，21(1):C1–C68，2018年。C. Emmenegger。dmlalg：双重机器学习算法，2021年。URL https://CRAN.R-project.org/package=dmlalg。R 软件包版本 1.0.2。C. Emmenegger 和 P. B¨uhlmann。在部分线性内生模型中对双重机器学习进行正则化。《电子统计学杂志》，15(2):6461–6543，2021年。H. Farbmacher。SIVREG：用于通过自适应 Lasso 处理部分无效工具变量的 Stata 模块，2018年。X. L. Frank Windmeijer。CIIV，2021年。URL https://github.com/xlbristol/CIIV。R 软件包版本 1.0.0。Z. Guo。使用无效工具变量进行因果推断：后选择问题及一种基于搜索与抽样的解决方案。预印本 arXiv:2104.06911，2021年。Z. Guo 和 P. B¨uhlmann。利用机器学习进行两阶段曲率识别：使用可能无效的工具变量进行因果推断，2022年。预印本 arXiv:2203.12808。Z. Guo、H. Kang、T. T. Cai 和 D. S. Small。通过结合投票的两阶段硬阈值法，计算无效工具变量下因果效应的置信区间。《皇家统计学会学报：B辑（统计方法学）》，80(4):793–815，2018年。H. Kang、Y. Jiang、Q. Zhao 和 D. Small。ivmodel：工具变量模型的统计推断与敏感性分析，2021年。URL https://CRAN.R-project.org/package=ivmodel。R 软件包版本 1.9.0。

T. Koo、S. Li 和 Z. Guo。controlfunctionIV：使用可能无效的工具变量的控制函数方法，2022a。网址：https://CRAN.R-project.org/package=controlfunctionIV。R 软件包版本 0.1.0。

T. Koo、Z. Wang、H. Kang、D. Small 和 Z. Guo。《RobustIV：线性模型中的稳健工具变量方法》，2022b。网址：https://CRAN.R-project.org/package=RobustIV。R 软件包版本 0.2.4。T. Koo、Y. Lee、D. S. Small 和 Z. Guo。《Robustiv 和 co ntrolfunctioniv：使用无效工具变量对线性和非线性模型进行因果推断》。预印本 arXiv:2301.04412，2023年。S. Li 和 Z. Guo。《使用可能无效的工具变量对非线性结果模型进行因果推断》，2020年。预印本 arXiv:2010.09922。Y. Lin 和 Y. Jeon。《随机森林与自适应最近邻》。《美国统计学会杂志》，101(474):578–590，2006年。N. Meinshausen、L. Meier 和 P. B¨uhlmann。《高维回归的p值》。《美国统计学会杂志》，104(488):1671–1681，2009年。M. Papadakis、M. Tsagris、M. Dimitriadis、S. Fafalios、I. Tsamardinos、M. Fasiolo、G. Borboudakis、J. Burkardt、C. Zou、K. Lakiotaki 和 C. Chatzipantsiou。《Rfast：高效且极快速的R函数集》，2022年。网址：https://CRAN.R-project.org/package=Rfast。R 软件包版本 2.0.6 和 2.0.7。J. O. Ramsay、S. Graves 和 G. Hooker。《fda：函数型数据分析》，2022年。网址：https://CRAN.R-project.org/package=fda。R 软件包版本 6.0.5。R Core Team。《R：用于统计计算的语言与环境》。R统计计算基金会，奥地利维也纳，2021年。网址：https://www.R-project.org/。StataCorp。《Stata统计软件》。德克萨斯州科利奇站，2021年。E. Tchetgen Tchetgen、B. Sun 和 S. Walter。《GENIUS方法：稳健孟德尔随机化推断》。《统计科学》，36(3):443–464，2021年。H. Theil。《最小二乘法重复应用于完整方程组》。中央计划局，海牙，1953a。油印备忘录。H. Theil。《完整方程组中的估计与同时相关》。中央计划局，海牙，1953b。油印备忘录。美国商务部。《1970年人口普查基础记录的公共使用样本》。华盛顿哥伦比亚特区，1992。数据取自Angrist数据档案，Angrist 和 Krueger（1991）部分，文件NEW7080.rar，https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archiv

e.

W. N. Venables 和 B. D. Ripley。《Modern Applied Statistics with S》。施普林格，纽约，第四版，2002年。

F. Windmeijer、X. Liang、F. P. Hartwig 和 J. Bowden。用于选择有效工具变量的置信区间方法。《皇家统计学会学报：B辑（统计方法学）》，83(4):752–776，2021年。

M. N. Wright 和 A. Ziegler。ranger：面向高维数据的随机森林在 C++ 和 R 中的快速实现。《统计软件学报》，77(1):1–17，2017年。

A 在 h 中编码工具变量违背

接下来，我们论证：如果工具变量 Z_i 无效，则 $h \neq 0$。我们的论证表明，如果 g 不依赖于 Z_i，但 Z_i 与未测量混杂因素相关，那么这种情况仍可表示为一个 g 依赖于工具变量的模型。如果 g 依赖于Z_i，此时假设（A2）或（A3）被违背，则该论断成立。因此，让我们考虑一个违背（A2）的工具变量 Z_i。由此，我们有 Cov($f(Z_i, X_i)$ ϵ_i) $\neq 0$。定义 $G(Z_i) = \mathrm{E}[\epsilon_i|Z_i]$。由于 X_i 的外生性，我们有 Cov($f(Z_i, X_i)$ $\epsilon_i - G(Z_i)$) $= 0$，进而有 Cov($f(Z_i)$ $G(Z_i)$) $\neq 0$。如果将 ϵ_i 重新定义为 $\epsilon_i - G(Z_i)$，我们便再次得到一个与工具变量不相关的误差，但此时 g 依赖于 Z_i，即 G。

B 标准差的自助法

为简化记号，我们定义

$$\mathbf{M}(\mathbf{V}_{\mathcal{A}_1}) = \boldsymbol{\Omega}^\top \mathbf{P}^{\perp}_{\widehat{\mathbf{V}}_{\mathcal{A}_1}} \boldsymbol{\Omega}. \tag{4}$$

因此，式（2）中的效应量估计量可分解为

$$\hat{\beta} - \beta = b(\mathbf{V}_{\mathcal{A}_1}) + \frac{\mathbf{D}_{\mathcal{A}_1}^\top \mathbf{M}(\mathbf{V}_{\mathcal{A}_1})\epsilon}{\mathbf{D}_{\mathcal{A}_1}^\top \mathbf{M}(\mathbf{V}_{\mathcal{A}_1})\mathbf{D}_{\mathcal{A}_1}} - \frac{\sum_{i=1}^{n_1}[\mathbf{M}(\mathbf{V}_{\mathcal{A}_1})]_{ii}\hat{\boldsymbol{\delta}}(\mathbf{V}_{\mathcal{A}_1})_i\hat{\boldsymbol{\epsilon}}(\mathbf{V}_{\mathcal{A}_1})_i}{\mathbf{D}_{\mathcal{A}_1}^\top \mathbf{M}(\mathbf{V}_{\mathcal{A}_1})\mathbf{D}_{\mathcal{A}_1}}, \tag{5}$$

其中，$b(\mathbf{V}_{\mathcal{A}_1})$ 是式（3）中的第二项与第三项之和；若选择了合适的违反空间候选项，该项可能较小；式（5）中的第三项是式（2）中偏差校正项的显式形式，其中 $\hat{\boldsymbol{\delta}}(\mathbf{V}_{\mathcal{A}_1}) = \mathbf{D}_{\mathcal{A}_1} - \widehat{f}$ 和

$$\hat{\boldsymbol{\epsilon}}(\mathbf{V}_{\mathcal{A}_1}) = \mathbf{P}^{\perp}_{\widehat{\mathbf{V}}_{\mathcal{A}_1}, \widehat{\mathbf{W}}_{\mathcal{A}_1}} \left(Y - D\frac{\mathbf{Y}_{\mathcal{A}_1}^\top \mathbf{M}(\mathbf{V}_{\mathcal{A}_1})\mathbf{D}_{\mathcal{A}_1}}{\mathbf{D}_{\mathcal{A}_1}^\top \mathbf{M}(\mathbf{V}_{\mathcal{A}_1})\mathbf{D}_{\mathcal{A}_1}} \right).$$

在满足某些正则条件时，偏差校正项引入的额外方差在渐近意义下可忽略不计（Guo 和 B¨uhlmann，2022年）。然而，在有限样本

情形下，将其纳入考虑可能会提高标准误估计量的准确性。我们采用以下自助法来推导此类估计量：对于 $1 \le i \le n_1$，我们分别将处理模型和结果模型的残差中心化，即 $\widetilde{\delta}_i = \widehat{\delta}_i - \bar{\mu}_\delta$ with $\bar{\mu}_\delta = \frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{\delta}_i$ and$\widetilde{\epsilon}_i = [\widehat{\epsilon}(V_{Q_{\max}})]_i - \bar{\mu}_\epsilon$ with $\bar{\mu}_\epsilon = \frac{1}{n_1}\sum_{i=1}^{n_1}[\widehat{\epsilon}(V_{Q_{\max}})]_i$。对于某个自然数$L$ 以及 $1 \le l \le L$，接着，我们为 $1 \le i \le n_1$ 生成 $\delta_i^{[l]} = U_i^{[l]} \cdot \widetilde{\delta}_i$ 和 $\epsilon_i^{[l]} = U_i^{[l]} \cdot \widetilde{\epsilon}_i$，其中 $\{U_i^{[l]}\}_{1 \le i \le n_1}$ 是独立同分布的标准正态随机变量。接下来，我们根据以下方式计算式（5）中第二项和第三项的自助法对应量：

$$N^{(l)} = \frac{\mathbf{D}_{\mathcal{A}_1}^\top \mathbf{M}(\mathbf{V}_{\mathcal{A}_1})\epsilon^{[l]} - \sum_{i=1}^{n_1}([\mathbf{M}(\mathbf{V}_{\mathcal{A}_1})]_{ii}\delta_i^{[l]}\epsilon_i^{[l]})}{D_{\mathcal{A}_1}^\top \mathbf{M}(\mathbf{V}_{\mathcal{A}_1})D_{\mathcal{A}_1}}.$$

然后，我们根据自助法统计量 $\{N^{(l)}\}_{1 \le l \le L}$ 得到效应量估计量的经验标准误，并记为 $\mathrm{SE}_{\mathrm{boot}}$。随后，双侧95%置信区间为

$$(\widehat{\beta} - 1.96 \cdot \mathrm{SE}_{\mathrm{boot}}, \widehat{\beta} + 1.96 \cdot \mathrm{SE}_{\mathrm{boot}}).$$