非线性ICP背景-ICP问题

东南大学 SOUTHEAST UNIVERSITY

Definition 1 (Environmental variables). *We know or assume that the variables E are neither descendants nor parents of Y in the causal DAG of (Y, X, E). If this is the case, we call E environmental variables.*

$$H_{0,S}: \quad Y \perp\!\!\!\perp E \mid X_S. \qquad\qquad Y = f(PA(Y), N_Y)$$

- ➢ **不同环境误差均值和方差相同，但分布不同：**

Example 1. *Consider a discrete environmental variable E. If in E = 1 we have*

$$Y = 2X + N, N \perp\!\!\!\perp X,$$

and in $E = 2$

$$Y = 2X + M, M \perp\!\!\!\perp X,$$

- ➢ **数据生成机制非线性：**

Example 2 (Linear model and nonlinear data). *Consider the following SCM, in which* X_2 *and* X_3 *are direct causes of* Y.

$$\begin{aligned} X_1 &\leftarrow E + \eta_X \\ X_2 &\leftarrow \sqrt{3X_1 + \eta_{X_1}} \\ X_3 &\leftarrow \sqrt{2X_1 + \eta_{X_2}} \\ Y &\leftarrow X_2^2 - X_3^2 + \eta_Y \end{aligned} \quad \Longrightarrow \quad Y \perp\!\!\!\perp E \mid X_1$$

考虑条件独立性测试，将 ICP 扩展到非线性情况

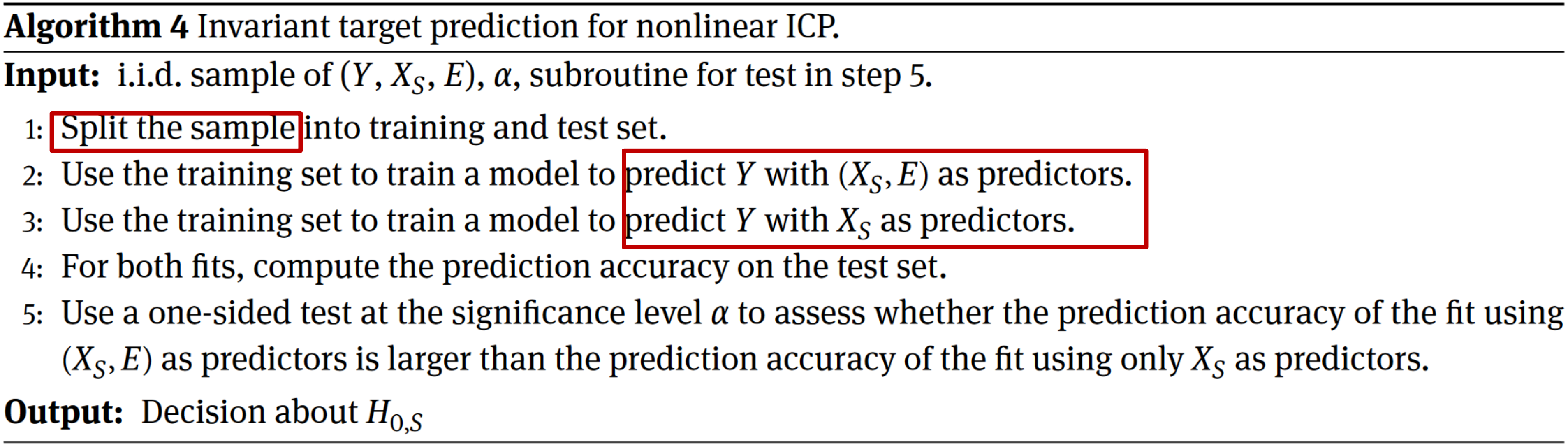
[Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models[J]. Journal of Causal Inference, 2018, 6(2).]

条件独立性测试

□ **作用：用于检验非参数、非线性下的** $H_{0,S}: \quad Y \perp\!\!\!\perp E \mid X_S.$

$$Y = f(PA(Y), N_Y)$$

➢ **不变目标预测：**



Algorithm 4 Invariant target prediction for nonlinear ICP.

Input: i.i.d. sample of (Y, X_S, E), α, subroutine for test in step 5.

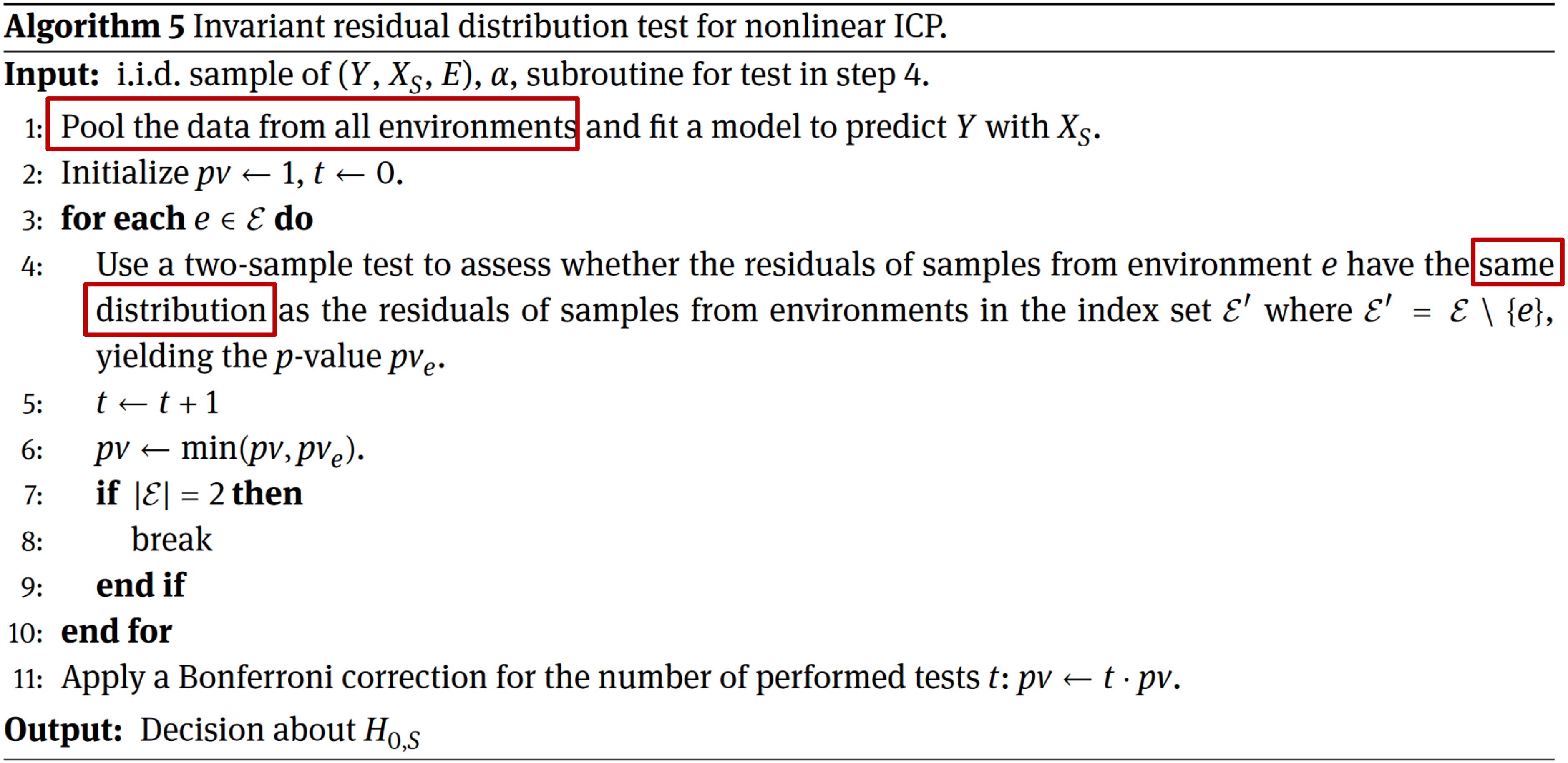
```
1: Split the sample into training and test set.
2: Use the training set to train a model to predict $Y$ with $(X_S, E)$ as predictors.
3: Use the training set to train a model to predict $Y$ with $X_S$ as predictors.
4: For both fits, compute the prediction accuracy on the test set.
5: Use a one-sided test at the significance level $\alpha$ to assess whether the prediction accuracy of the fit using
   $(X_S, E)$ as predictors is larger than the prediction accuracy of the fit using only $X_S$ as predictors.
```

Output: Decision about $H_{0,S}$

通过比较两种预测的ACC，判断条件独立性

[Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models[J]. Journal of Causal Inference, 2018, 6(2).]

➢ 不变残差分布测试：结果稳定



Algorithm 5 Invariant residual distribution test for nonlinear ICP.

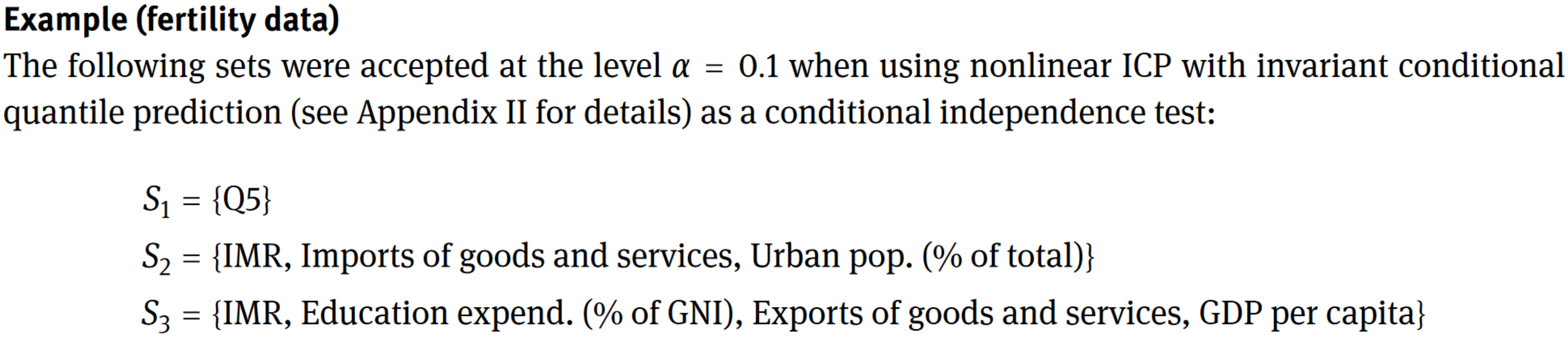
```
Input: i.i.d. sample of $(Y, X_S, E)$, $\alpha$, subroutine for test in step 4.
 1: Pool the data from all environments and fit a model to predict $Y$ with $X_S$.
 2: Initialize $pv \leftarrow 1$, $t \leftarrow 0$.
 3: for each $e \in \mathcal{E}$ do
 4:     Use a two-sample test to assess whether the residuals of samples from environment $e$ have the same distribution as the residuals of samples from environments in the index set $\mathcal{E}'$ where $\mathcal{E}' = \mathcal{E} \setminus \{e\}$, yielding the $p$-value $pv_e$.
 5:     $t \leftarrow t + 1$
 6:     $pv \leftarrow \min(pv, pv_e)$.
 7:     if $|\mathcal{E}| = 2$ then
 8:         break
 9:     end if
10: end for
11: Apply a Bonferroni correction for the number of performed tests $t$: $pv \leftarrow t \cdot pv$.
Output: Decision about $H_{0,S}$
```

通过检验不同环境中残差的分布是否相同，判断条件独立性

[Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models[J]. Journal of Causal Inference, 2018, 6(2).]

定义集

□ 作用：解决高度相关的变量无法区分问题



Example (fertility data)

The following sets were accepted at the level $\alpha = 0.1$ when using nonlinear ICP with invariant conditional quantile prediction (see Appendix II for details) as a conditional independence test:

$$S_1 = \{\text{Q5}\}$$

$$S_2 = \{\text{IMR, Imports of goods and services, Urban pop. (\% of total)}\}$$

$$S_3 = \{\text{IMR, Education expend. (\% of GNI), Exports of goods and services, GDP per capita}\}$$

Q5与IMR高度相关，可相互替代，可能导致排除两个变量

$$\hat{S} = \emptyset$$

□ 定义：

a defining set $\hat{D} \subseteq \{1, \ldots, p\}$ has the properties:
(i) $S \cap \hat{D} \neq \emptyset$ for all S such that $H_{0,S}$ is accepted.
(ii) there exists no strictly smaller set D' with $D' \subset \hat{D}$ for which property (i) is true.

$$P(S^* \cap \hat{D} = \emptyset) \leq P(H_{0,S^*} \text{ rejected}) \leq \alpha.$$

$$S_1 = \{\text{Q5}\}$$

$$S_2 = \{\text{IMR, Imports of goods and services, Urban pop. (\% of total)}\}$$

$$S_3 = \{\text{IMR, Education expend. (\% of GNI), Exports of goods and services, GDP per capita}\}$$

Example (fertility data)

We obtain seven defining sets:

$\boxed{\hat{D}_1 = \{\text{IMR, Q5}\}}$ → **IMR与Q5至少一个是父变量**

$$\hat{D}_2 = \{\text{Q5, Education expenditure (\% of GNI), Imports of goods and services}\}$$

$$\hat{D}_3 = \{\text{Q5, Education expenditure (\% of GNI), Urban pop. (\% of total)}\}$$

$$\hat{D}_4 = \{\text{Q5, Exports of goods and services, Imports of goods and services}\}$$

$$\hat{D}_5 = \{\text{Q5, Exports of goods and services, Urban pop. (\% of total)}\}$$

$$\hat{D}_6 = \{\text{Q5, GDP per capita, Imports of goods and services}\}$$

$$\hat{D}_7 = \{\text{Q5, GDP per capita, Urban pop. (\% of total)}\}$$

通过引入定义集的概念，确保至少能找到一个父变量

[Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models[J]. Journal of Causal Inference, 2018, 6(2).]

非线性ICP优缺点

□ 优：

- 通过考虑**条件独立性测试**，突破ICP中**线性高斯假设**，且环境变量从**离散扩展到连续**，求得**非线性非参数下**的父节点。
- 通过引入“**定义集**”的概念，解决了**变量高度相关**的情况下，所得父节点集为空的问题。
- 非线性ICP在**线性和非线性情况下**结果均较好；线性情况下ICP效果更好。

□ 缺：

- 依赖**因果充分性**假设。
- 当父节点集包含**两个以上的变量**时，条件独立性测试**结果不好**。

[Heinze-Deml C, Peters J, Meinshausen N. Invariant causal prediction for nonlinear models[J]. Journal of Causal Inference, 2018, 6(2).]