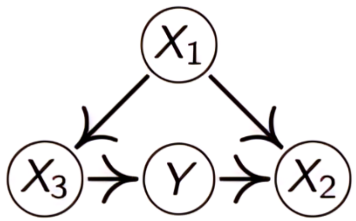
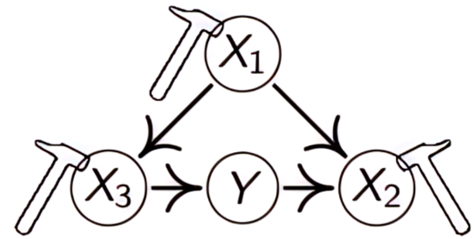
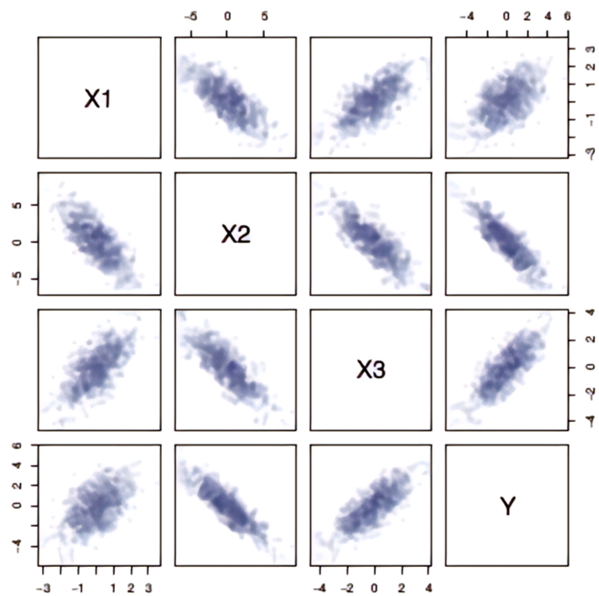
ICP背景

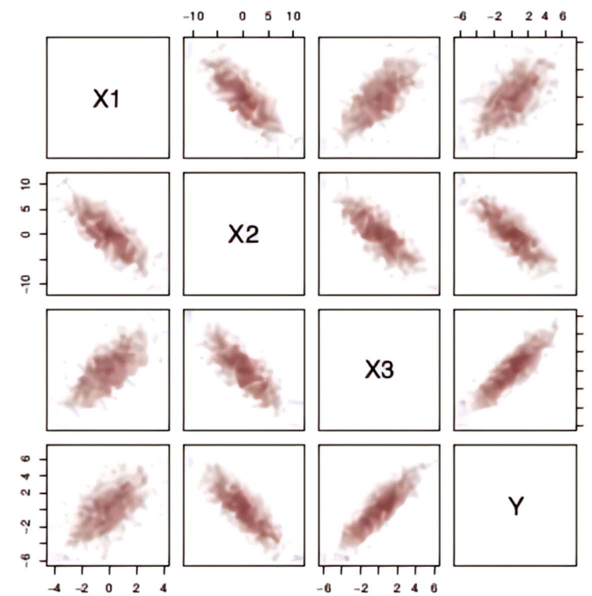
unknown:





known:





在数据不独立同分布情况下，推断Y的父节点集，最终实现不变预测

[Jonas Peters, University of Copenhagen, Mini-course on Causality, Laboratory for Information & Decision Systems (LIDS) and Models, Inference & Algorithms of the Broad Institute, MIT, 2017]

线性回归下结果

东南大学 SOUTHEAST UNIVERSITY

□ 假设检验：

H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$ H_1: 至少有一个$\beta_i \neq 0$, $i = 1, 2, ..., k$

```
> linmod <- lm(Y~X)
> summary(linmod)
```

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.000322   0.025858   0.012     0.99
X1          -0.444534   0.034306 -12.958   <2e-16 ***
X2          -0.402398   0.016471 -24.430   <2e-16 ***
X3           0.603502   0.025642  23.536   <2e-16 ***
```

若 $p \leq \alpha$, 则拒绝原假设; 反之, 则不拒绝。

线性回归下不能准确得到Y的父节点，无法实现跨环境预测

[Jonas Peters, University of Copenhagen, Mini-course on Causality, Laboratory for Information & Decision Systems (LIDS) and Models, Inference & Algorithms of the Broad Institute, MIT, 2017]

ICP结果


```
> ExpInd
```

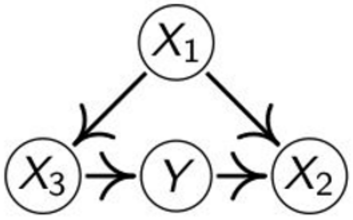
不同环境标识

[1]111111111111111111111111111111...2222222222222...

```
> icp <- ICP(X,Y,ExpInd)
```

	LOWER BOUND	UPPER BOUND	MAXIMIN EFFECT	P-VALUE
Variable_1	-0.11	0.10	0.00	1.0000
Variable_2	-0.33	0.00	0.00	1.0000
Variable_3	0.47	1.05	0.47	0.0012 **

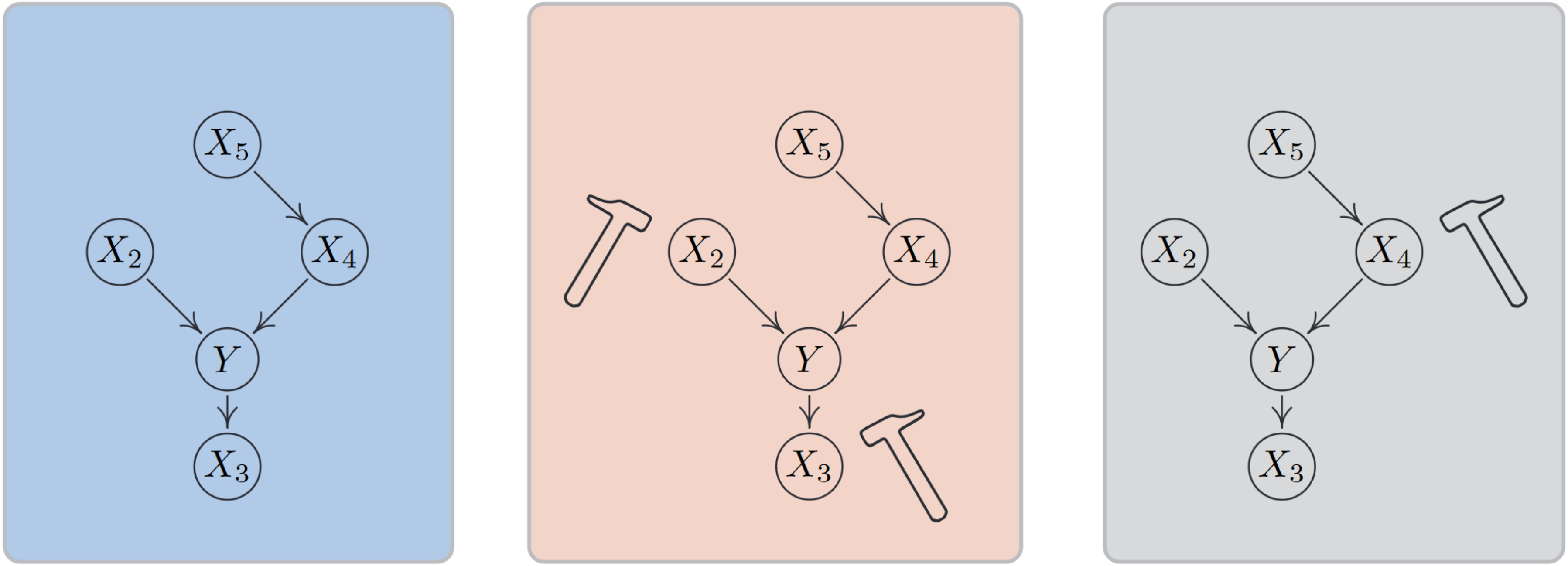
若 $p \leq \alpha$, 则拒绝原假设; 反之, 则不拒绝。



ICP框架下，Y基于父节点X_3实现不变性预测

[Jonas Peters, University of Copenhagen, Mini-course on Causality, Laboratory for Information & Decision Systems (LIDS) and Models, Inference & Algorithms of the Broad Institute, MIT, 2017]

ICP思想-Invariance



$$P(Y \mid PA(Y)) = P(Y \mid X_2, X_4)$$ **跨环境不变，用于预测Y**

$$\left.\begin{array}{c} P(Y \mid X_2, X_3, X_4) \\ P(Y \mid X_3) \\ P(Y \mid X_5) \end{array}\right\} \text{变}$$

找到Y的父节点，实现跨环境不变预测

[Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals[J]. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2016: 947-1012]

ICP理论

□ 不变性原理：

Let S^* be the indices of parents (Y). There exists γ^* with support S^* that satisfies

$$
\begin{aligned}
\text{for all } e \in \mathrm{E}: \quad & X^e \text{ has an arbitrary distribution and} \\
& Y^e \mid X^e_{S^*} = x \text{ invariant.} \\
& Y^e = X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}
\end{aligned}
$$

We say:

" S^* satisfies invariant prediction." $\Longleftrightarrow$ "$H_{0,S^*}(\mathrm{E})$ is true."

□ 目标：

给定不同环境数据 $e \in E$，找到 S^*

□ 方法：

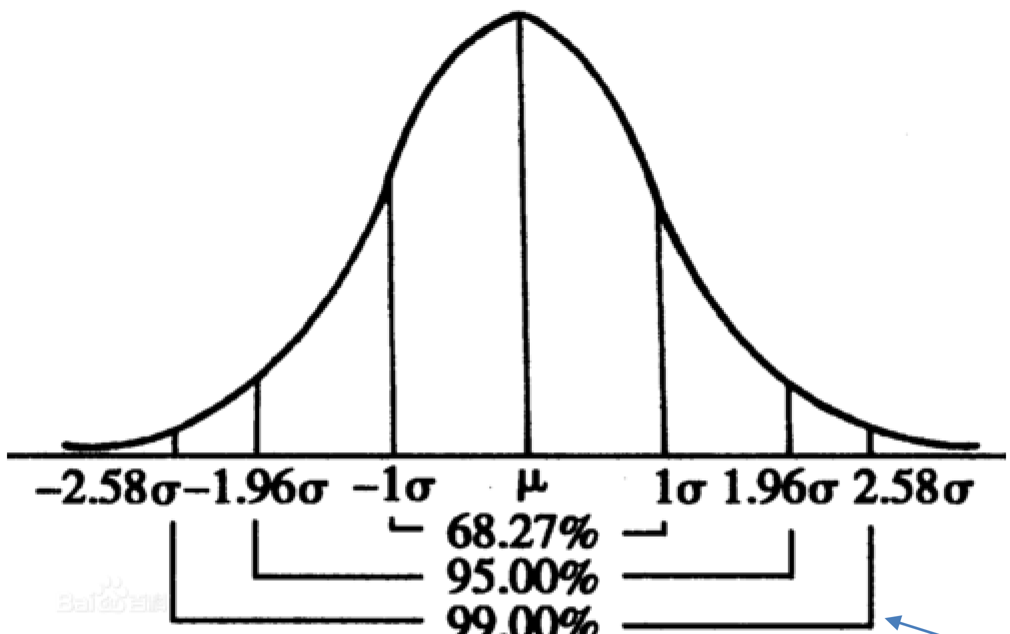
穷举 S，**检测** S是否会使得**$P(Y|S)$在不同环境下一致**。若某个S使得$P(Y|S)$在不同环境下一致，那么S**可能**就是PA_Y。当出现多个集合符合要求，取交集：

$$\hat{S}(E) := \bigcap_{S: H_{0,S}(E) \text{ not rejected}} S$$

所有满足不变预测的S求交，得最终输出

假设检验流程

$H_0\colon u = 3$



犯第一类错误的概率 $P(\text{拒绝}H_0|H_0\text{为真}) = \alpha$

利用“小概率事件在少量实验中是几乎不可能出现”，证明原假设错误

[茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程[M]. 高等教育出版社, 2011.]

Hypothesis testing is not all you need

□ 统计中：

- ➢ **从样本估计总体统计性质。**

□ 机器学习中很少进行假设检验原因：

- ➢ 结果为特定参数、特定情况下所得，不具有可重复性。
- ➢ 深度学习发展。实验成本太高。

□ 因果发现中：

- ➢ 马尔可夫性+忠实性。　　**基于条件独立性的因果发现**
- ➢ 数据分布特定假设。　　**基于因果函数的因果发现**

ICP中的假设检验

□ 检测$P(Y|S)$在不同环境下是否一致：

H_0: S满足不变预测　　　　H_1: S不满足不变预测

$\Updownarrow$ **线性高斯假设下**

$$\text{for all } e \in \mathrm{E}: \quad Y^e = X_S^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X_S^e$$

$\Updownarrow$

由于不同环境下的ε^e具有相同的分布，

可验证不同环境下ε^e的均值和方差是否相同，

通过两次假设检验从样本得总体统计性质。

H_0^1: 不同环境下总体均值相同

T检验：$T = \dfrac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

H_0^2: 不同环境下总体方差相同

F检验： $F = \dfrac{s_1^2}{s_2^2}$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}} \qquad s_1 = \sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_{1i} - \bar{X}_1\right)^2}$$

$\boldsymbol{H_0^1}$: 不同环境下总体均值相同

$\boldsymbol{T}$检验： $T = \dfrac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

$\boldsymbol{H_0^2}$: 不同环境下总体方差相同

$\boldsymbol{F}$检验: $F = \dfrac{s_1^2}{s_2^2}$

$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}$ $\quad s_1 = \sqrt{\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_{1i} - \bar{X}_1\right)^2}$

Bonferroni correction：用于解决多重假设检验中犯第一类错误的概率增加的问题。

比如：有三个环境，则每个环境需进行2次假设检验，共进行6次假设检验。当满足 $6\min_e p^e > \alpha$ 时，才能接受原假设H_0: S满足不变性

[Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals[J]. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2016: 947-1012]

ICP相关定理

□ 总体上环境越多，结果越好：

$$S(E_1) \subseteq S(E_2) \subseteq S^* \quad \text{if} \quad E_1 \subseteq E_2$$

$$\Downarrow \text{更泛化}$$

如果有**更多、更强**的干预措施，干预在**更好的位置**，数据中有更多的**异质性**，**可识别性**就会提高。

□ 因果不变集的置信度保证：

$$P\left(\hat{S}(E) \subseteq S^*\right) \geq 1 - \alpha$$

$$\Uparrow$$

$$P\left(\hat{S}(E) \subseteq S^*\right) \boxed{\geq} P(H_{0,S^*} \text{ accepted }) \boxed{\geq} 1 - \alpha$$

右可推左　　　犯第一类错误的概率 $= \alpha$

[Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals[J]. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2016: 947-1012]

ICP优缺点

口 优：

- ICP不需要知道干预的**具体位置和类型**。只假设干预不改变$P(Y \mid PA(Y))$。
- 能得到**因果预测集的置信度**和**因果系数的置信区间。**
- 不依赖**忠实性**或其他**可识别性假设**（与其他因果发现方法对比）。

口 缺：

- 不适用于**非线性**框架。
- 特征X需要**预处理**好（不能end2end），且**维数不高。**
- ICP需要**环境已知**（哪些数据来自哪一环境）。

[Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals[J]. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2016: 947-1012]

ICP≠因果发现

东南大学 SOUTHEAST UNIVERSITY

□ 因果发现：

- 传统因果发现在**数据独立同分布**下，推断**整个图结构。**
- 多数仅提供**点估计**，缺乏**统计置信度保证。**
- 依赖**忠实性**或其他**可识别性假设。**

样本估计总体

不能推广

□ ICP：

- ICP在数据**不独立同分布**下，推断目标变量***Y*的父节点。**
- 能得到**因果预测集的置信度**和**因果系数的置信区间。**
- **不依赖忠实性**或其他**可识别性假设。**

样本估计总体