Causal discovery with Unsupervised inverse REgression (CURE)

Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, Bernhard Schölkopf

Max-Planck-Institute for Intelligent Systems, Tübingen, Germany

Advances in Causal Inference Workshop, UAI 2015

16th July 2015
Problem

- Causal discovery in the two-variable case, assuming no confounders: given a sample from $P(X, Y)$, infer whether

$$X \rightarrow Y \quad \text{or} \quad Y \rightarrow X$$
Related work

- Conditional-independence based methods, e.g., PC or IC.
    - $X \rightarrow Y$ and $Y \rightarrow X$ Markov equivalent

- Methods restricting the function class, e.g., ANM, LINGAM.
  - Peters, Mooij, Janzing, Schölkopf. Causal discovery with continuous ANMs. JMLR 2014.

- Methods based on the postulate of independence of causal mechanisms, e.g., IGCI.
  - Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf. Inferring deterministic causal relations. UAI 2010.

Upon this talk: causal discovery in the non-deterministic case based on the postulate of independence.
Related work

- **Conditional-independence based methods, e.g., PC or IC.**
  - $X \rightarrow Y$ and $Y \rightarrow X$ Markov equivalent

- **Methods restricting the function class, e.g., ANM, LINGAM.**
  Peters, Mooij, Janzing, Schölkopf. Causal discovery with continuous ANMs. JMLR 2014.
  - ANM: $Y = f(X) + N$, $X \perp N$
Related work

- **Conditional-independence based methods, e.g., PC or IC.**
  
  
  - $X \rightarrow Y$ and $Y \rightarrow X$ Markov equivalent

- **Methods restricting the function class, e.g., ANM, LINGAM.**
  
  Peters, Mooij, Janzing, Schölkopf. Causal discovery with continuous ANMs. JMLR 2014.  
  
  - ANM: $Y = f(X) + N$, $X \perp \perp N$

- **Methods based on the postulate of independence of causal mechanisms, e.g., IGCI.**
  
  Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf. Inferring deterministic causal relations. UAI 2010.  
  
  - IGCI proposed for deterministic relations: $Y = f(X)$
Related work

- Conditional-independence based methods, e.g., PC or IC.
  
    
    - $X \rightarrow Y$ and $Y \rightarrow X$ Markov equivalent

- Methods restricting the function class, e.g., ANM, LINGAM.
  
  - Peters, Mooij, Janzing, Schölkopf. Causal discovery with continuous ANMs. JMLR 2014.
    
    - ANM: $Y = f(X) + N, \; X \perp \perp N$

- Methods based on the postulate of independence of causal mechanisms, e.g., IGCI.
  
  - Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf. Inferring deterministic causal relations. UAI 2010.
    
    - IGCI proposed for deterministic relations: $Y = f(X)$

This talk: causal discovery in the non-deterministic case based on the postulate of independence.
Postulate: if $X \rightarrow Y$, then $P(X)$ and $P(Y|X)$ are “independent”, in the sense that $P(X)$ contains no information about $P(Y|X)$ and vice versa.
Independence between causal mechanism and distribution of cause

- Postulate: if $X \rightarrow Y$, then $P(X)$ and $P(Y|X)$ are “independent”, in the sense that $P(X)$ contains no information about $P(Y|X)$ and vice versa.

- This “independence” can be violated in the backward direction: $P(Y)$ and $P(X|Y)$ may contain information about each other, because they both inherit properties from $P(X)$ and $P(Y|X)$. 


Lemeire and Dirkx. Causal models as minimal descriptions of multivariate systems. 2006.
Independence between causal mechanism and distribution of cause

Postulate: if $X \rightarrow Y$, then $P(X)$ and $P(Y|X)$ are “independent”, in the sense that $P(X)$ contains no information about $P(Y|X)$ and vice versa.

This “independence” can be violated in the backward direction: $P(Y)$ and $P(X|Y)$ may contain information about each other, because they both inherit properties from $P(X)$ and $P(Y|X)$.


Lemeire and Dirkx. Causal models as minimal descriptions of multivariate systems. 2006.
Postulate of independence (abstract)

▶ if $X \rightarrow Y$:

$$P(Y|X) \perp \perp P(X)$$

This asymmetry between cause and effect can be useful for causal discovery, but needs to be precisely defined.

Postulate of independence (abstract)

- if $X \rightarrow Y$:

  $P(Y|X) \perp \perp P(X)$ implying $P(X|Y) \not\perp \not\perp P(Y)$

This asymmetry between cause and effect can be useful for causal discovery, but needs to be precisely defined.

Postulate of independence (abstract)

► if $X \rightarrow Y$:

$$P(Y | X) \perp \!\!\!\!\!\! \perp P(X)$$  implying  $$P(X | Y) \perp \!\!\!\!\!\! \perp P(Y)$$

► if $X \rightarrow Y$ deterministically ($Y = f(X)$ as opposed to $Y = f(X, E)$):

This asymmetry between cause and effect can be useful for causal discovery, but needs to be precisely defined.

Postulate of independence (abstract)

- if $X \rightarrow Y$:
  \[ P(Y|X) \perp \perp P(X) \]
  implying
  \[ P(X|Y) \not\perp \not\perp P(Y) \]

- if $X \rightarrow Y$ deterministically ($Y = f(X)$ as opposed to $Y = f(X, E)$):
  \[ f \perp \perp P(X) \]
Postulate of independence (abstract)

- if $X \rightarrow Y$:
  
  $P(Y|X) \perp \perp P(X)$ implying $P(X|Y) \not\perp \not\perp P(Y)$

- if $X \rightarrow Y$ deterministically ($Y = f(X)$ as opposed to $Y = f(X, E)$):
  
  $f \perp \perp P(X)$ implying $f^{-1} \not\perp \not\perp P(Y)$

This asymmetry between cause and effect can be useful for causal discovery, but needs to be precisely defined.


Eleni Sgouritsa

CURE 5
Postulate of independence (abstract)

- if $X \rightarrow Y$:

  $P(Y|X) \perp \! \! \! \! \perp P(X)$ implying $P(X|Y) \not\perp \! \! \! \! \perp P(Y)$

- if $X \rightarrow Y$ deterministically ($Y = f(X)$ as opposed to $Y = f(X, E)$):

  $f \perp \! \! \! \! \perp P(X)$ implying $f^{-1} \not\perp \! \! \! \! \perp P(Y)$

This asymmetry between cause and effect can be useful for causal discovery.
Postulate of independence (abstract)

- if $X \rightarrow Y$:

  $P(Y|X) \perp \perp P(X)$ \hspace{1cm} implying \hspace{1cm} $P(X|Y) \not\perp \not\perp P(Y)$

- if $X \rightarrow Y$ deterministically ($Y = f(X)$ as opposed to $Y = f(X, E)$):

  $f \perp \perp P(X)$ \hspace{1cm} implying \hspace{1cm} $f^{-1} \not\perp \not\perp P(Y)$

This asymmetry between cause and effect can be useful for causal discovery, but needs to be precisely defined.

Asymmetry

If $X \rightarrow Y$:

<table>
<thead>
<tr>
<th>Abstract asymmetry</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f \perp \perp P(X)$ whereas $f^{-1} \not\perp \not\perp P(Y)$</td>
<td>$P(Y</td>
<td>X) \perp \perp P(X)$ whereas $P(X</td>
</tr>
</tbody>
</table>

## Asymmetry

If $X \rightarrow Y$:

![Graph showing the relationship between $X$ and $Y$](image)

<table>
<thead>
<tr>
<th></th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract asymmetry</strong></td>
<td>$f$ “⊥” $P(X)$ whereas $f^{-1}$ “⊥” $P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td><strong>Formal asymmetry</strong></td>
<td><strong>IGCI:</strong> $\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$</td>
<td></td>
</tr>
</tbody>
</table>
## Asymmetry

If $X \rightarrow Y$:

![Diagram showing the relationship between $X$ and $Y$](image)

<table>
<thead>
<tr>
<th>Abstract asymmetry</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f \perp P(X)$</td>
<td>$f^{-1} \not\perp P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td>$f^{-1} \not\perp P(Y)$</td>
<td></td>
<td>$P(X</td>
</tr>
</tbody>
</table>

**IGCI:**
- $\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$
- $? \quad ?$
## Asymmetry

If $X \rightarrow Y$:

![Graph showing function $f(x)$ and distributions $p(x)$ and $p(y)$]

<table>
<thead>
<tr>
<th></th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract asymmetry</strong></td>
<td>$f$ &quot;⊥&quot; $P(X)$ whereas $f^{-1}$ &quot;∥&quot; $P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td><strong>Formal asymmetry</strong></td>
<td><strong>IGCI:</strong> $\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$</td>
<td>? It is difficult to explicitly formalize independence between $P(Y</td>
</tr>
</tbody>
</table>
## Asymmetry

If $X \rightarrow Y$:

<table>
<thead>
<tr>
<th>Abstract asymmetry</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$ “⊥” $P(X)$ whereas $f^{-1}$ “∥” $P(Y)$</td>
<td>$f$ “⊥” $P(X)$ whereas $f^{-1}$ “∥” $P(Y)$</td>
<td></td>
</tr>
<tr>
<td>IGCI: $\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$</td>
<td>$\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$</td>
<td></td>
</tr>
</tbody>
</table>

? It is difficult to explicitly formalize independence between $P(Y|X)$ and $P(X)$

---

Eleni Sgouritsa

CURE
Asymmetry

If $X \to Y$:

<table>
<thead>
<tr>
<th>Abstract asymmetry</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f \ &quot;\perp\perp&quot; \ P(X)$ whereas $f^{-1} \ &quot;\not\perp\not\perp&quot; \ P(Y)$</td>
<td>$P(Y</td>
<td>X) \ &quot;\perp\perp&quot; \ P(X)$ whereas $P(X</td>
</tr>
</tbody>
</table>

| Formal asymmetry   | IGCI: $\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$ | ? It is difficult to explicitly formalize independence between $P(Y|X)$ and $P(X)$ |

| Alternative asymmetry? | $f$ can’t be estimated from $p_X$ |
Asymmetry

If $X \rightarrow Y$:

![Graph showing the relationship between $X$, $f(X)$, $Y$, and $p(x)$ and $p(y)$]

<table>
<thead>
<tr>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract asymmetry</strong></td>
<td></td>
</tr>
<tr>
<td>$f$ “⊥” $P(X)$ whereas</td>
<td>$P(Y</td>
</tr>
<tr>
<td>$f^{-1}$ “⊥” $P(Y)$</td>
<td>$P(X</td>
</tr>
<tr>
<td><strong>Formal asymmetry</strong></td>
<td></td>
</tr>
<tr>
<td><strong>IGCI:</strong></td>
<td></td>
</tr>
<tr>
<td>Cov($\log f'$, $p_X$) = 0</td>
<td></td>
</tr>
<tr>
<td>Cov($\log f^{-1}'$, $p_Y$) $\geq$ 0</td>
<td></td>
</tr>
<tr>
<td><strong>Alternative asymmetry?</strong></td>
<td></td>
</tr>
<tr>
<td>$f$ can’t be estimated from $p_X$ whereas</td>
<td></td>
</tr>
<tr>
<td>$f^{-1}$ may be estimated from $p_Y$</td>
<td></td>
</tr>
</tbody>
</table>

Formal asymmetry

\[ \text{Cov}(\log f', p_X) = 0 \] whereas \[ \text{Cov}(\log f^{-1}', p_Y) \geq 0 \]

\[ P(Y|X) \text{ “⊥” } P(X) \] whereas \[ P(X|Y) \text{ “⊥” } P(Y) \]

It is difficult to explicitly formalize independence between $P(Y|X)$ and $P(X)$.

The inspiration for the last asymmetry came from:

## Asymmetry

If $X \rightarrow Y$:

$$f(x)$$

**Deterministic** $Y = f(X)$ whereas $f^{-1}$ “$\perp \perp$” $P(Y)$

| Abstract asymmetry                  | $f$ “$\perp$” $P(X)$ whereas $f^{-1}$ “$\perp \perp$” $P(Y)$ | $P(Y|X)$ “$\perp$” $P(X)$ whereas $P(X|Y)$ “$\perp \perp$” $P(Y)$ |
|-------------------------------------|---------------------------------------------------------------|-----------------------------------------------------------------|
| **Formal asymmetry**                | **IGCI**: $\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$ | ? It is difficult to explicitly formalize independence between $P(Y|X)$ and $P(X)$ |
| **Alternative asymmetry?**          | $f$ can’t be estimated from $p_X$ whereas $f^{-1}$ may be estimated from $p_Y$ | **CURE**: $p_Y|X$ can’t be estim. from $p_X$ whereas $p_X|Y$ may be estimated from $p_Y$ |
## Asymmetry

If $X \rightarrow Y$:

![Diagram showing the relationship between $X$ and $Y$ with $f(x)$ and $p(x)$.]  

| Abstract asymmetry | Deterministic $Y = f(X)$: $f$ “⊥” $P(X)$ whereas $f^{-1}$ “⊥” $P(Y)$ | Non-deterministic $Y = f(X, E)$: $P(Y | X)$ “⊥” $P(X)$ whereas $P(X | Y)$ “⊥” $P(Y)$ |
|---------------------|----------------------------------------------------------|-----------------------------------------------------------------|
| IGCI:               | $\text{Cov}(\log f', p_X) = 0$ whereas $\text{Cov}(\log f^{-1}', p_Y) \geq 0$ | $? \text{ It is difficult to explicitly formalize independence between } P(Y | X) \text{ and } P(X)$ |
| Alternative asymmetry? | $f$ can’t be estimated from $p_X$ whereas $f^{-1}$ may be estimated from $p_Y$ | **CURE:** $p_Y | X$ can’t be estim. from $p_X$ whereas $p_X | Y$ may be estimated from $p_Y$ |

The inspiration for the last asymmetry came from:

Idea of CURE for the simpler deterministic case

How can \( f^{-1} \) be estimated based only on \( p_Y \)?

\[
Y = f(X) = f \circ F^{-1}(X(u)) = h(X(u)) = F_Y(y)
\]

Use \( h^{-1} \) as an estimate for \( f^{-1} \).
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F^{-1}(X = h(Y)) = h^{-1}(y) = F_Y(y)$

Use $h^{-1}$ as an estimate for $f^{-1}$

$f$ cannot be estimated based only on $p_X$.

$X = f^{-1}(Y) = f^{-1} \circ F^{-1}(Y = g(Y)) = g(Y)$

Eleni Sgouritsa

CURE 7
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$$
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F^{-1}_X(X_u) = h(X_u)$

$h^{-1}(y) = F_Y(y)$
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

$h^{-1}(y) = F_Y(y)$

Use $h^{-1}$ as an estimate for $f^{-1}$
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

$h^{-1}(y) = F_Y(y)$

Use $h^{-1}$ as an estimate for $f^{-1}$

$f$ cannot be estimated based only on $p_X$
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

Use $h^{-1}$ as an estimate for $f^{-1}$

$f$ cannot be estimated based only on $p_X$

$X = f^{-1}(Y) = f^{-1} \circ F_Y^{-1}(Y_u) = g(Y_u)$
Estimate $p_{X|Y}$ based on $p_{Y}$

$X \rightarrow Y$
Goal: estimate $p_{X|Y}$ based on $p_{Y}$

observed: $y \in \mathbb{R}^{N}$, unobserved: $x \in \mathbb{R}^{N}$
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$

Goal: estimate $p_{X|Y}$ based on $p_Y$

observed: $y \in \mathbb{R}^N$, unobserved: $x \in \mathbb{R}^N$

▶ Model:
  ▶ GP (marg.) likelihood: $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$     $\theta = (\ell, \sigma_f, \sigma_n)$
Estimate $p_X|Y$ based on $p_Y$

$X \rightarrow Y$
Goal: estimate $p_X|Y$ based on $p_Y$

observed: $y \in \mathbb{R}^N$, unobserved: $x \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$ $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x) = \prod_{i=1}^N \mathcal{U}(0, 1)$ and hyperparameters’ prior
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$

Goal: estimate $p_{X|Y}$ based on $p_Y$

observed: $y \in \mathbb{R}^N$, unobserved: $x \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$ \hspace{1cm} $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x) = \prod_{i=1}^{N} \mathcal{U}(0, 1)$ and hyperparameters’ prior

- Estimate $p_{X|Y}$:
Estimate $p_{X|Y}$ based on $p_{Y}$

Goal: estimate $p_{X|Y}$ based on $p_{Y}$

observed: $y \in \mathbb{R}^N$, unobserved: $x \in \mathbb{R}^N$

- **Model:**
  - GP (marg.) *likelihood*: $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$ \hspace{1cm} $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s *prior*: $p(x) = \prod_{i=1}^{N} \mathcal{U}(0, 1)$ and hyperparameters’ prior

- **Estimate $p_{X|Y}$:**

\[
\hat{p}_{X|Y}^y : (x, y) \mapsto p(x|y, y) = \int p(x|y, y, x, \theta) p(x, \theta|y) dx d\theta
\]
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$

Goal: estimate $p_{X|Y}$ based on $p_Y$

observed: $y \in \mathbb{R}^N$, unobserved: $x \in \mathbb{R}^N$

- **Model:**
  - GP (marg.) **likelihood:** $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$
  - latent’s **prior:** $p(x) = \prod_{i=1}^N U(0, 1)$ and hyperparameters’ prior
- **Estimate $p_{X|Y}$:**
  \[
  \hat{p}_{X|Y}^y : (x, y) \mapsto p(x|y, y) = \int p(x|y, y, x, \theta) p(x, \theta|y) \, dx \, d\theta
  \]
  GP posterior
Goal: estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$

observed: $y \in \mathbb{R}^N$, unobserved: $x \in \mathbb{R}^N$

Model:

- GP (marg.) likelihood: $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$  \hspace{1cm} $\theta = (\ell, \sigma_f, \sigma_n)$

- latent’s prior: $p(x) = \prod_{i=1}^N U(0, 1)$ and hyperparameters’ prior

Estimate $p_{X|Y}$:

$\hat{p}_{X|Y}^y : (x, y) \mapsto p(x|y, y) = \int p(x|y, y, x, \theta) \ p(x, \theta|y) \ dx d\theta$

GP posterior

$(N+3)$-dimens.
Estimate $p_{X|Y}$ based on $p_Y$

Goal: estimate $p_{X|Y}$ based on $p_Y$

observed: $y \in \mathbb{R}^N$, unobserved: $x \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$  \hspace{1cm} $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x) = \prod_{i=1}^{N} U(0, 1)$ and hyperparameters’ prior

- Estimate $p_{X|Y}$:

  \[
  \hat{p}_{X|Y}^y : (x, y) \mapsto p(x|y, y) = \int p(x|y, y, x, \theta) \underbrace{p(x, \theta|y)}_{\text{GP posterior}} \, dx d\theta
  \]

  \[
  \approx \frac{1}{M} \sum_{i=1}^{M} p(x|y, y, x^i, \theta^i)
  \]
Estimate $p_{X|Y}$ based on $p_{Y}$

$X \to Y$

Goal: estimate $p_{X|Y}$ based on $p_{Y}$

observed: $y \in \mathbb{R}^{N}$, unobserved: $x \in \mathbb{R}^{N}$

- Model:
  - GP (marg.) likelihood: $p(y|x, \theta) = \mathcal{N}(y; 0, K_{x,x} + \sigma_n^2 I_N)$  $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x) = \prod_{i=1}^{N} \mathcal{U}(0, 1)$ and hyperparameters' prior

- Estimate $p_{X|Y}$:

  $\hat{p}_{X|Y}^y : (x, y) \mapsto p(x|y, y) = \int p(x|y, y, x, \theta) \underbrace{p(x, \theta|y)}_{\text{GP posterior}} \, dx \, d\theta$

  $\approx \frac{1}{M} \sum_{i=1}^{M} p(x|y, y, x^i, \theta^i)$

Grey: $(x, y)$
Red: $(x^i, y)$
Blue: $p(x|y = 0.559, y, x^i, \theta^i)$
Empirical data: $y \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

1. Estimate $p_{X|Y}$ by $\hat{p}_{y|X|Y}(y)$ (using only $y$)

2. Estimate $p_{Y|X}$ by $\hat{p}_{x|Y|X}(x)$ (using only $x$)

3. Check which estimation is better

4. Infer $X \rightarrow Y$ if $D_{X|Y} < D_{Y|X}$, otherwise infer $Y \rightarrow X$

$D_{X|Y} = -\log \prod_{j=1}^{N} \hat{p}_{y|X|Y}(x_j, y_j)$

$D_{Y|X} = -\log \prod_{j=1}^{N} \hat{p}_{x|Y|X}(y_j, x_j)$
Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}^y_{X|Y} \) (using only \( y \))
Empirical data: $y \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

1. Estimate $p_{X|Y}$ by $\hat{p}^y_{X|Y}$ (using only $y$)
2. Estimate $p_{Y|X}$ by $\hat{p}^x_{Y|X}$ (using only $x$)

$D_{X|Y} = -\log \prod_{j=1}^{N} \hat{p}^y_{X|Y}(x_j, y_j) \prod_{j=1}^{N} \hat{p}^x_{Y|X}(x_j, y_j)$

Sgouritsa, Janzing, Hennig, Schölkopf. Inference of cause and effect with unsupervised inverse regression. AISTATS 2015
Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X|Y}^y \) (using only \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y|X}^x \) (using only \( x \))
3. Check which estimation is better

\[ D_{X|Y} = -\log \prod_{j=1}^N \hat{p}_{X|Y}^y \] 

\[ D_{Y|X} = -\log \prod_{j=1}^N \hat{p}_{Y|X}^x \]
Empirical data: \( \mathbf{y} \in \mathbb{R}^N, \mathbf{x} \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X|Y}^y \) (using **only** \( \mathbf{y} \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y|X}^x \) (using **only** \( \mathbf{x} \))
3. Check which estimation is better
4. Infer \( X \rightarrow Y \) if 1. better \( (D_{X|Y} < D_{Y|X}) \), otherwise infer \( Y \rightarrow X \)
Empirical data: \( y \in \mathbb{R}^N, \, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X|Y}^y \) (using only \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y|X}^x \) (using only \( x \))
3. Check which estimation is better
4. Infer \( X \rightarrow Y \) if 1. better \( (D_{X|Y} < D_{Y|X}) \), otherwise infer \( Y \rightarrow X \)

\[
D_{X|Y} = -\log \frac{\prod_{j=1}^{N} \hat{p}_{X|Y}^y (x_j, y_j)}{\prod_{j=1}^{N} \hat{p}_{X|Y}^x (x_j, y_j)}
\]
Empirical data: $y \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

1. Estimate $p_{X|Y}$ by $\hat{p}_{X|Y}^y$ (using only $y$)
2. Estimate $p_{Y|X}$ by $\hat{p}_{Y|X}^x$ (using only $x$)
3. Check which estimation is better
4. Infer $X \rightarrow Y$ if 1. better ($D_{X|Y} < D_{Y|X}$), otherwise infer $Y \rightarrow X$

$$D_{X|Y} = -\log \frac{\prod_{j=1}^N \hat{p}_{X|Y}^y(x_j, y_j)}{\prod_{j=1}^N \hat{p}_{X|Y}^x(x_j, y_j)}$$

$$D_{Y|X} = -\log \frac{\prod_{j=1}^N \hat{p}_{Y|X}^x(y_j, x_j)}{\prod_{j=1}^N \hat{p}_{Y|X}^y(x_j, y_j)}$$
Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X|Y}^y \) (using only \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y|X}^x \) (using only \( x \))
3. Check which estimation is better
4. Infer \( X \rightarrow Y \) if 1. better \( (D_{X|Y} < D_{Y|X}) \), otherwise infer \( Y \rightarrow X \)

\[
D_{X|Y} = -\log \frac{\prod_{j=1}^N \hat{p}_{X|Y}^y(x_j, y_j)}{\prod_{j=1}^N \hat{p}_{X|Y}^x(y_j, x_j)}
\]

\[
D_{Y|X} = -\log \frac{\prod_{j=1}^N \hat{p}_{Y|X}^x(y_j, x_j)}{\prod_{j=1}^N \hat{p}_{Y|X}^y(y_j, x_j)}
\]

with

\[
\hat{p}_{X|Y}^y : (x, y) \mapsto p(x|y, y) = \frac{1}{M} \sum_{i=1}^M p(x|y, y, x', \theta^i)
\]

\[
\hat{p}_{X|Y}^x : (x, y) \mapsto p(x|y, y, x, \theta)
\]
Examples

True DAG: $X \rightarrow Y$

- Often get “good” MCMC samples even when the data are generated by non-Gaussian noise or non-additive noise or non-uniform input.
Examples

True DAG: $X \rightarrow Y$

- Often get “good” MCMC samples even when the data are generated by non-Gaussian noise or non-additive noise or non-uniform input.

- Often get “bad” MCMC samples when trying to predict based on the distribution of the cause.
Results: simulated data
Results: real data (81 cause-effect pairs)

Sample size: 200

Cause-effect pairs dataset: Mooij, Peters, Janzing, Zscheischler, Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. 2014.
Assumption: independence of causal mechanisms $P(cause)$ and $P(effect|cause)$. 

CURE: Estimate $p_{X|Y}$ based on $p_Y$ 
Estimate $p_{Y|X}$ based on $p_X$ 
Infer $X \rightarrow Y$ if the first estimation is better. Otherwise, infer $Y \rightarrow X$. 

Results seem promising, sampling computationally expensive.
Assumption: independence of causal mechanisms $P(\text{cause})$ and $P(\text{effect}|\text{cause})$.

This independence introduces an asymmetry between cause and effect used for causal discovery.
Assumption: independence of causal mechanisms $P(\text{cause})$ and $P(\text{effect}|\text{cause})$.

This independence introduces an asymmetry between cause and effect used for causal discovery.

CURE:
- Estimate $p_{X|Y}$ based on $p_Y$
Conclusion

- Assumption: independence of causal mechanisms $P(cause)$ and $P(effect|cause)$.
- This independence introduces an asymmetry between cause and effect used for causal discovery.
- CURE:
  - Estimate $p_{X|Y}$ based on $p_Y$
  - Estimate $p_{Y|X}$ based on $p_X$
Assumption: independence of causal mechanisms $P(\text{cause})$ and $P(\text{effect}|\text{cause})$.

This independence introduces an asymmetry between cause and effect used for causal discovery.

CURE:
- Estimate $p_{X|Y}$ based on $p_Y$
- Estimate $p_{Y|X}$ based on $p_X$
- Infer $X \rightarrow Y$ if the first estimation is better. Otherwise, infer $Y \rightarrow X$.

Results seem promising, sampling computationally expensive.
Assumption: independence of causal mechanisms $P(\text{cause})$ and $P(\text{effect}|\text{cause})$.

This independence introduces an asymmetry between cause and effect used for causal discovery.

CURE:
- Estimate $p_{X|Y}$ based on $p_Y$
- Estimate $p_{Y|X}$ based on $p_X$
- Infer $X \rightarrow Y$ if the first estimation is better. Otherwise, infer $Y \rightarrow X$.

Results seem promising, sampling computationally expensive.
Supervised inverse regression

Unlike standard supervised GP regression, the predictive distribution of supervised inverse regression

\[ p(x|y, y, x, \theta) \propto p(y, y|x, x, \theta)p(x|x, \theta) = \mathcal{N}(y, y; 0, K_{(x,x),(x,x)} + \sigma_n^2 I_N) \]

is not Gaussian.

Green: \( p(y|x, x, y, \theta) \), Blue: \( p(x|y, y, x, \theta) \)
Supervised inverse regression

Unlike standard supervised GP regression, the predictive distribution of supervised inverse regression

\[
p(x|y, y, x, \theta) \propto p(y, y|x, x, \theta)p(x|x, \theta) = \mathcal{N}(y, y; 0, K_{(x,x),(x,x)} + \sigma_n^2 I_N)
\]

is not Gaussian.

Green: \( p(y|x, x, y, \theta) \), Blue: \( p(x|y, y, x, \theta) \)
What about the non-deterministic (noisy) case?

<table>
<thead>
<tr>
<th>If $X \rightarrow Y$:</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract asymmetry</td>
<td>$f$ “⊥” $P(X)$ whereas</td>
<td>$P(Y</td>
</tr>
<tr>
<td></td>
<td>$f^{-1}$ “⊣” $P(Y)$</td>
<td>$P(X</td>
</tr>
</tbody>
</table>

$\text{Eleni Sgouritsa}$
What about the non-deterministic (noisy) case?

<table>
<thead>
<tr>
<th>If $X \rightarrow Y$:</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract asymmetry</td>
<td>$f \perp \perp P(X)$ whereas $f^{-1} \perp \perp P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td>Alternative asymmetry?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What about the non-deterministic (noisy) case?

If $X \rightarrow Y$:

<table>
<thead>
<tr>
<th></th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>$f$ “⊥” $P(X)$ whereas $f^{-1}$ “∥” $P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td>asymmetry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternative</td>
<td>$f$ can’t be estimated from $p_X$ whereas $f^{-1}$ may be estimated from $p_Y$</td>
<td><strong>CURE:</strong> $p_{Y</td>
</tr>
<tr>
<td>asymmetry?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What about the non-deterministic (noisy) case?

<table>
<thead>
<tr>
<th>If $X \rightarrow Y$:</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract asymmetry</strong></td>
<td>$f \leftarrow\rightarrow P(X)$ whereas $f^{-1} \leftarrow\rightarrow P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td><strong>Alternative asymmetry?</strong></td>
<td>$f$ can’t be estimated from $p_X$ whereas $f^{-1}$ may be estimated from $p_Y$</td>
<td><strong>CURE:</strong> $p_Y</td>
</tr>
<tr>
<td><strong>Estimate from $p_Y$</strong></td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
What about the non-deterministic (noisy) case?

<table>
<thead>
<tr>
<th>If $X \rightarrow Y$:</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract asymmetry</td>
<td>$f$ “⊥” $P(X)$ whereas $f^{-1}$ “⊥” $P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td>Alternative asymmetry?</td>
<td>$f$ can’t be estimated from $p_X$ whereas $f^{-1}$ may be estimated from $p_Y$</td>
<td>CURE: $p_Y</td>
</tr>
<tr>
<td>Estimate from $p_Y$</td>
<td>$Y = f(X) = h(X_u)$ estimate $h^{-1}$ from $p_Y$</td>
<td></td>
</tr>
</tbody>
</table>
What about the non-deterministic (noisy) case?

<table>
<thead>
<tr>
<th>If $X \rightarrow Y$:</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract asymmetry</td>
<td>$f \ &quot;\perp\perp&quot; \ P(X)$ whereas $f^{-1} \ &quot;!!!\perp\perp!!!\perp&quot; \ P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td>Alternative asymmetry?</td>
<td>$f$ can’t be estimated from $p_X$ whereas $f^{-1}$ may be estimated from $p_Y$</td>
<td><strong>CURE:</strong> $p_{Y</td>
</tr>
<tr>
<td>Estimate from $p_Y$</td>
<td>$Y = f(X) = h(X_u)$ estimate $h^{-1}$ from $p_Y$</td>
<td>$Y = f(X) + E = h(X_u) + E$</td>
</tr>
</tbody>
</table>
What about the non-deterministic (noisy) case?

<table>
<thead>
<tr>
<th>If $X \rightarrow Y$:</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract asymmetry</strong></td>
<td>$f \ “\perp” \ P(X)$ whereas</td>
<td>$P(Y</td>
</tr>
<tr>
<td></td>
<td>$f^{-1} \ “\not\perp” \ P(Y)$</td>
<td>$P(X</td>
</tr>
<tr>
<td><strong>Alternative asymmetry?</strong></td>
<td>$f$ can’t be estimated from $p_X$ whereas</td>
<td><strong>CURE:</strong></td>
</tr>
<tr>
<td></td>
<td>$f^{-1}$ may be estimated from $p_Y$</td>
<td>$p_{Y</td>
</tr>
<tr>
<td><strong>Estimate from $p_Y$</strong></td>
<td>$Y = f(X) = h(X_u)$</td>
<td>$p_{X</td>
</tr>
<tr>
<td></td>
<td>estimate $h^{-1}$ from $p_Y$</td>
<td></td>
</tr>
</tbody>
</table>

Eleni Sgouritsa
What about the non-deterministic (noisy) case?

<table>
<thead>
<tr>
<th>If $X \rightarrow Y$:</th>
<th>Deterministic $Y = f(X)$</th>
<th>Non-deterministic $Y = f(X, E)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract asymmetry</td>
<td>$f \perp \perp P(X)$ whereas $f^{-1} \perp \perp P(Y)$</td>
<td>$P(Y</td>
</tr>
<tr>
<td>Alternative asymmetry?</td>
<td>$f$ can’t be estimated from $p_X$ whereas $f^{-1}$ may be estimated from $p_Y$</td>
<td><strong>CURE:</strong> $p_{Y</td>
</tr>
<tr>
<td>Estimate from $p_Y$</td>
<td>$Y = f(X) = h(X_u)$ estimate $h^{-1}$ from $p_Y$</td>
<td>$Y = f(X) + E = h(X_u) + E$ estimate $p_{X_u</td>
</tr>
</tbody>
</table>
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X_u|Y}$ based on $p_Y$
(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X|Y}$ based on $p_Y$
(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{x_u,x_u} + \sigma_n^2 I_N)$
  \[ \theta = (\ell, \sigma_f, \sigma_n) \]
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X_u|Y}$ based on $p_Y$

(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- **Model:**
  - GP (marg.) *likelihood*: $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{x_u,x_u} + \sigma_n^2 I_N)$ \hspace{1cm} $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s *prior*: $p(x_u) = \prod_{i=1}^N U(0, 1)$ and hyperparameters’ prior
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{Xu|Y}$ based on $p_Y$

(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{xu}.x_u + \sigma_n^2 I_N)$ \quad $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x_u) = \prod_{i=1}^N \mathcal{U}(0, 1)$ and hyperparameters’ prior

- Estimate $p_{Xu|Y}$:
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X_u|Y}$ based on $p_Y$
(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{x_u,x_u} + \sigma_n^2 I_N)$  \hspace{1em} $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x_u) = \prod_{i=1}^{N} \mathcal{U}(0, 1)$ and hyperparameters’ prior

- Estimate $p_{X_u|Y}$:
  \[
  \hat{p}_{X_u|Y}^y : (x_u, y) \mapsto p(x_u|y, y) = \int p(x_u|y, y, x_u, \theta) p(x_u, \theta|y) dx_u d\theta
  \]
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X_u|Y}$ based on $p_Y$

(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{x_u,x_u} + \sigma_n^2 I_N)$  \hspace{1cm} $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x_u) = \prod_{i=1}^N \mathcal{U}(0, 1)$ and hyperparameters’ prior

- Estimate $p_{X_u|Y}$:
  \[
  \hat{p}_{X_u|Y}^y : (x_u, y) \mapsto p(x_u|y, y) = \int p(x_u|y, y, x_u, \theta) p(x_u, \theta|y) dx_u d\theta
  \]
  \hspace{1cm} GP posterior
Estimate $p_X|Y$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X|Y}$ based on $p_Y$

(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- **Model:**
  - GP (marg.) **likelihood:** $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{x_u,x_u} + \sigma_n^2 I_N)$ \hspace{1cm} $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s **prior:** $p(x_u) = \prod_{i=1}^N \mathcal{U}(0, 1)$ and hyperparameters’ prior
  - Estimate $p_{X_u|Y}$:
    \[
    \hat{p}_{X_u|Y}^y : (x_u, y) \mapsto p(x_u|y, y) = \int p(x_u|y, y, x_u, \theta) \ p(x_u, \theta|y) \ dx_u d\theta
    \]
    
    **GP posterior** \hspace{1cm} $(N+3)$-dimens.
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X_u|Y}$ based on $p_Y$

(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- **Model:**
  - GP (marg.) likelihood: $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{x_u,x_u} + \sigma_n^2 I_N)$, $\theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x_u) = \prod_{i=1}^N \mathcal{U}(0, 1)$ and hyperparameters’ prior

- **Estimate $p_{X_u|Y}$:**
  
  \[
  \hat{p}_{X_u|Y}^y : (x_u, y) \mapsto p(x_u|y, y) = \int p(x_u|y, y, x_u, \theta) \, p(x_u, \theta|y) \, dx_u d\theta \\
  \approx \frac{1}{M} \sum_{i=1}^M p(x_u|y, y, x_u^i, \theta^i) 
  \]

  GP posterior (N+3)-dimens.
Estimate $p_{X|Y}$ based on $p_Y$

$X \rightarrow Y$ with $Y = f(X) + E = h(X_u) + E$

Goal: estimate $p_{X_u|Y}$ based on $p_Y$
(then use this as an estimate of $p_{X|Y}$)

observed: $y \in \mathbb{R}^N$, unobserved: $x_u \in \mathbb{R}^N$

- Model:
  - GP (marg.) likelihood: $p(y|x_u, \theta) = \mathcal{N}(y; 0, K_{x_u}x_u + \sigma_n^2 I_N)$ $\quad \theta = (\ell, \sigma_f, \sigma_n)$
  - latent’s prior: $p(x_u) = \prod_{i=1}^{N} \mathcal{U}(0, 1)$ and hyperparameters’ prior
- Estimate $p_{X_u|Y}$:
  \[ \hat{p}^y_{X_u|Y} : (x_u, y) \mapsto p(x_u|y, y) = \int p(x_u|y, y, x_u, \theta) \frac{p(x_u, \theta|y)}{\int \frac{p(x_u, \theta|y)}{p(x_u|y, y, x_u, \theta)} \, dx_u \, d\theta} \]
  
  \[ \approx \frac{1}{M} \sum_{i=1}^{M} p(x_u|y, y, x_u^i, \theta^i) \]
  
  Grey: $(x, y)$
  Red: $(x_u^i, y)$
  Blue: $p(x_u|y = 0.559, y, x_u^i, \theta^i)$
Empirical data: $y \in \mathbb{R}^N$, $x \in \mathbb{R}^N$
Empirical data: $y \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

1. Estimate $p_{X|Y}$ by $\hat{p}_{X|Y}^y$ (using only $y$)

2. Estimate $p_{Y|X}$ by $\hat{p}_{Y|X}^x$ (using only $x$)

3. Check which estimation is better

4. Infer $X \rightarrow Y$ if $1.$ better ($D_{X|Y} < D_{Y|X}$), otherwise infer $Y \rightarrow X$

$D_{X|Y} = -\log \prod_{j=1}^{N} \hat{p}_{X,Y|Y}(x_j, y_j) \prod_{j=1}^{N} \hat{p}_{X,Y|Y}(x_j, y_j)$

$D_{Y|X} = -\log \prod_{j=1}^{N} \hat{p}_{X,Y|X}(y_j, x_j) \prod_{j=1}^{N} \hat{p}_{X,Y|X}(y_j, x_j)$
Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}^y_{X_u|Y} \) (using **only** \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}^x_{Y_u|X} \) (using **only** \( x \))
Empirical data: \( y \in \mathbb{R}^N, \ x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X|Y}^y \) (using \textbf{only } \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y|X}^x \) (using \textbf{only } \( x \))
3. Check which estimation is better

\[
D_{X|Y} = -\log \prod_{j=1}^N \hat{p}_{X|Y}(x_j, y_j) \\
D_{Y|X} = -\log \prod_{j=1}^N \hat{p}_{Y|X}(y_j, x_j)
\]

with \( \hat{p}_{X|Y}^y : (x_u, y) \mapsto p(x_u|y, y) = 1 \)

\( \hat{p}_{Y|X}^x : (x_u, y) \mapsto p(x_u|y, y, x_u, \theta) \)
Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X_u|Y}^y \) (using only \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y_u|X}^x \) (using only \( x \))
3. Check which estimation is better
4. Infer \( X \rightarrow Y \) if 1. better \((D_{X|Y} < D_{Y|X})\), otherwise infer \( Y \rightarrow X \)
Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X_u|Y}^y \) (using only \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y_u|X}^x \) (using only \( x \))
3. Check which estimation is better
4. Infer \( X \rightarrow Y \) if 1. better (\( D_{X|Y} < D_{Y|X} \)), otherwise infer \( Y \rightarrow X \)

\[
D_{X|Y} = -\log \frac{\prod_{j=1}^N \hat{p}_{X_u|Y}^y(x_j, y_j)}{\prod_{j=1}^N \hat{p}_{X_u|Y}^x(x_j, y_j)}
\]
Empirical data: $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{x} \in \mathbb{R}^N$

1. Estimate $p_{X|Y}$ by $\hat{p}^y_{X_u|Y}$ (using only $\mathbf{y}$)
2. Estimate $p_{Y|X}$ by $\hat{p}^x_{Y_u|X}$ (using only $\mathbf{x}$)
3. Check which estimation is better
4. Infer $X \rightarrow Y$ if 1. better ($D_{X|Y} < D_{Y|X}$), otherwise infer $Y \rightarrow X$

$$D_{X|Y} = -\log \frac{\prod_{j=1}^N \hat{p}^y_{X_u|Y}(x_j, y_j)}{\prod_{j=1}^N \hat{p}^{x,y}_{X_u|Y}(x_j, y_j)}$$

$$D_{Y|X} = -\log \frac{\prod_{j=1}^N \hat{p}^x_{Y_u|X}(y_j, x_j)}{\prod_{j=1}^N \hat{p}^{y,x}_{Y_u|X}(y_j, x_j)}$$
Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( p_{X|Y} \) by \( \hat{p}_{X_u|Y}^y \) (using only \( y \))
2. Estimate \( p_{Y|X} \) by \( \hat{p}_{Y_u|X}^x \) (using only \( x \))
3. Check which estimation is better
4. Infer \( X \rightarrow Y \) if 1. better (\( D_X|Y < D_Y|X \)), otherwise infer \( Y \rightarrow X \)

\[
D_{X|Y} = - \log \frac{\prod_{j=1}^{N} \hat{p}_{X_u|Y}^y(x_j, y_j)}{\prod_{j=1}^{N} \hat{p}_{X_u|Y}^x(x_j, y_j)}
\]

\[
D_{Y|X} = - \log \frac{\prod_{j=1}^{N} \hat{p}_{Y_u|X}^x(y_j, x_j)}{\prod_{j=1}^{N} \hat{p}_{Y_u|X}^y(y_j, x_j)}
\]

with

\[
\hat{p}_{X_u|Y}^y : (x_u, y) \mapsto p(x_u|y, y) = \frac{1}{M} \sum_{i=1}^{M} p(x_u|y, y, x_{u_i}^i, \theta^i)
\]

\[
\hat{p}_{X_u|Y}^x : (x_u, y) \mapsto p(x_u|y, y, x, \theta)
\]
Formalization of independence in the deterministic case

If $X \rightarrow Y$ with $Y = f(X):$ \[
\begin{array}{c}
\text{f "⊥" } P(X) \\
\end{array}
\]
Formalization of independence in the deterministic case

If $X \rightarrow Y$ with $Y = f(X)$: $f \perp \perp P(X)$ implying $f^{-1} \not\perp \not\perp P(Y)$
Formalization of independence in the deterministic case

If $X \rightarrow Y$ with $Y = f(X)$: $f \perp P(X)$ implying $f^{-1} \not\perp P(Y)$
Formalization of independence in the deterministic case

If \( X \rightarrow Y \) with \( Y = f(X) \):

\[
f \perp \! \! \! \perp P(X) \quad \text{implying} \quad f^{-1} \not\perp \! \! \! \perp P(Y)
\]

Asymmetry

- **Postulate:**

\[
\text{Cov}(\log f', p_X) = 0
\]

Peaks of \( p_X \) do **not correlate** with the slope of \( f \).
Formalization of independence in the deterministic case

If $X \rightarrow Y$ with $Y = f(X)$: $f \perp \perp P(X)$ implying $f^{-1} \not\perp \not\perp P(Y)$

Asymmetry

- **Postulate:**
  \[
  \text{Cov}(\log f', p_X) = 0
  \]
  Peaks of $p_X$ do **not correlate** with the slope of $f$.

- **Implication:**
  \[
  \text{Cov}(\log f^{-1}', p_Y) \geq 0
  \]
  Peaks of $p_Y$ **correlate** with the slope of $f^{-1}$.
Formalization of independence in the deterministic case

If $X \rightarrow Y$ with $Y = f(X)$: $f \perp \perp P(X)$ implying $f^{-1} \nRightarrow P(Y)$

Asymmetry

- **Postulate:**
  $$\text{Cov}(\log f', p_X) = 0$$
  Peaks of $p_X$ do **not correlate** with the slope of $f$.

- **Implication:**
  $$\text{Cov}(\log f^{-1}', p_Y) \geq 0$$
  Peaks of $p_Y$ **correlate** with the slope of $f^{-1}$.

- Interpret $\log f'$ and $p_X$ as random variables on $[0, 1]$.
- $f$ a nonlinear monotonously increasing bijection of $[0, 1]$.

Daniusis et al. Inferring deterministic causal relations. UAI 2010.
Semi-supervised learning (SSL)

- Given: $D_l = \{(x_i, y_i) | i = 1, \ldots, l\}$ drawn i.i.d from $P(X, Y)$
  $D_u = \{x_{l+j} | j = 1, \ldots, u\}$ drawn i.i.d from $P(X)$

- Goal: learn a mapping from $X$ to $Y$, i.e. estimate $P(Y|X)$

Semi-supervised learning (SSL)

- Given: \( D_l = \{(x_i, y_i) | i = 1, \ldots, l\} \) drawn i.i.d from \( P(X, Y) \)
  \( D_u = \{x_{l+j} | j = 1, \ldots, u\} \) drawn i.i.d from \( P(X) \)

- Goal: learn a mapping from \( X \) to \( Y \), i.e. estimate \( P(Y|X) \)

- For SSL to work, the distribution of the unlabeled data \( P(X) \) has to carry information relevant to the estimation of \( P(Y|X) \)
Semi-supervised learning (SSL)

- Given: \( D_l = \{(x_i, y_i) | i = 1, \ldots, l\} \) drawn i.i.d from \( P(X, Y) \)
  \( D_u = \{x_{l+j} | j = 1, \ldots, u\} \) drawn i.i.d from \( P(X) \)

- Goal: learn a mapping from \( X \) to \( Y \), i.e. estimate \( P(Y|X) \)

- For SSL to work, the distribution of the unlabeled data \( P(X) \) has to carry information relevant to the estimation of \( P(Y|X) \) ⇒
  - **SSL pointless** if \( X \to Y \), because \( P(X) \) contains no information about \( P(Y|X) \)
  - **SSL can help** if \( Y \to X \), because \( P(X) \) contains information about \( P(Y|X) \)

Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

Use $h^{-1}$ as an estimate for $f^{-1}$

Use $g^{-1}$ as an estimate for $f^{-1}$
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

$h^{-1}(y) = F_Y(y)$

Use $h^{-1}$ as an estimate for $f^{-1}$
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

$h^{-1}(y) = F_Y(y)$

Use $h^{-1}$ as an estimate for $f^{-1}$

$f$ cannot be estimated based only on $p_X$
Idea of CURE for the simpler deterministic case

How can $f^{-1}$ be estimated based only on $p_Y$?

$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

$h^{-1}(y) = F_Y(y)$

Use $h^{-1}$ as an estimate for $f^{-1}$

$f$ cannot be estimated based only on $p_X$

$X = f^{-1}(Y) = f^{-1} \circ F_Y^{-1}(Y_u) = g(Y_u)$

$g^{-1}(x) = F_X(x)$

Use $g^{-1}$ as an estimate for $f$
Gaussian process

- Generalization of the Gaussian probability distribution
- Describes a distribution over functions
- \( f(x) \sim \mathcal{GP}(m(x), k(x, x')) \)
Gaussian process

- Generalization of the Gaussian probability distribution
- Describes a distribution over functions
- \[ f(x) \sim GP(m(x), k(x, x')) \]
- The specification of the covariance function implies a specific distribution over functions:
  \[ \text{cov}(f(x_p), f(x_q)) = k(x_p, x_q) = \exp\left(-\frac{1}{2\ell}(x_p - x_q)^2\right) \]
- Finite number of points: \[ \mathbf{y} \sim \mathcal{N}(\mathbf{0}, K_{x,x} + \sigma_n^2 I_N) \]
- GP regression
Causal discovery: CURE method

Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)
Causal discovery: CURE method

Empirical data: $y \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

1. Estimate $\hat{p}_{X|Y}^y$

2. Estimate $\hat{p}_{Y|X}^x$

Evaluate conditional estimation:

$D_{X|Y} = L_{unsup} X|Y - L_{sup} X|Y = -\frac{1}{N} \sum_{i=1}^{N} \log \hat{p}_{Y|X|u} (x_i, y_i) + \frac{1}{N} \sum_{i=1}^{N} \log \hat{p}_{X|Y|u} (x_i, y_i)$

with $\hat{p}_{X|Y|u}: (x, y) \mapsto p(x|y, y, x, \theta)$

3. Causal discovery: CURE

If $D_{X|Y} < D_{Y|X}$, infer $X \rightarrow Y$, otherwise infer $Y \rightarrow X$
Causal discovery: CURE method

Empirical data: $y \in \mathbb{R}^N$, $x \in \mathbb{R}^N$

1. Estimate $\hat{p}_{X|Y}^y$
   - Evaluate conditional estimation:
     $$D_{X|Y} = L_{X|Y}^{\text{unsup}} - L_{X|Y}^{\text{sup}} = -\frac{1}{N} \sum_{i=1}^N \log \hat{p}_{X|Y}^y(x_i, y_i) + \frac{1}{N} \sum_{i=1}^N \log \hat{p}_{X|Y}^{x,y}(x_i, y_i)$$
     with
     $$\hat{p}_{X|Y}^{x,y} : (x, y) \mapsto p(x|y, y, x, \theta)$$
Causal discovery: CURE method

Empirical data: \( y \in \mathbb{R}^N, x \in \mathbb{R}^N \)

1. Estimate \( \hat{p}^y_{X_u|Y} \)
   
   Evaluate conditional estimation:

   \[
   D_{X|Y} = L_{X|Y}^{\text{unsup}} - L_{X|Y}^{\text{sup}} = -\frac{1}{N} \sum_{i=1}^{N} \log \hat{p}^y_{X_u|Y}(x_i, y_i) + \frac{1}{N} \sum_{i=1}^{N} \log \hat{p}^{x,y}_{X_u|Y}(x_i, y_i)
   \]

   with

   \( \hat{p}^{x,y}_{X_u|Y} : (x, y) \mapsto p(x|y, y, x, \theta) \)

2. Estimate \( \hat{p}^x_{Y_u|X} \)
   
   Evaluate conditional estimation: compute \( D_{Y|X} \)
Causal discovery: CURE method

Empirical data: $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{x} \in \mathbb{R}^N$

1. Estimate $\hat{p}_{\mathbf{x} \mid \mathbf{y}}^{\mathbf{y}}$
   - Evaluate conditional estimation:
   \[
   D_{X \mid Y} = L_{X \mid Y}^{\text{unsup}} - L_{X \mid Y}^{\text{sup}} = -\frac{1}{N} \sum_{i=1}^{N} \log \hat{p}_{X \mid Y}^{\mathbf{y}}(x_i, y_i) + \frac{1}{N} \sum_{i=1}^{N} \log \hat{p}_{X \mid Y}^{\mathbf{x}, \mathbf{y}}(x_i, y_i)
   \]
   with
   \[
   \hat{p}_{X \mid Y}^{\mathbf{x}, \mathbf{y}} : (x, y) \mapsto p(x \mid y, \mathbf{y}, \mathbf{x}, \theta)
   \]

2. Estimate $\hat{p}_{\mathbf{y} \mid \mathbf{x}}^{\mathbf{x}}$
   - Evaluate conditional estimation: compute $D_{Y \mid X}$

Causal discovery: CURE

If $D_{X \mid Y} < D_{Y \mid X}$, infer $X \rightarrow Y$, otherwise infer $Y \rightarrow X$

Sgouritsa, Janzing, Hennig, Schölkopf. Inference of cause and effect with unsupervised inverse regression. AISTATS 2015