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

MAX-PLANCK-GESELLSCHAFT

Problem

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

$$X \to Y \qquad \text{or} \qquad Y \to X$$

Related work

- Conditional-independence based methods, e.g., PC or IC.

  Spirtes, Glymour, Scheines. Causation, prediction, and search. 2000.
  Pearl. Causality: reasoning and inference. 2009.

  - $X \rightarrow Y$ and $Y \rightarrow X$ Markov equivalent

Related work

- Conditional-independence based methods, e.g., PC or IC.
  Spirtes, Glymour, Scheines. Causation, prediction, and search. 2000.
  Pearl. Causality: reasoning and inference. 2009.
  - $X \to Y$ and $Y \to X$ Markov equivalent
- Methods restricting the function class, e.g., ANM, LINGAM.
  Hoyer, Janzing, Mooij, Peters, Schölkopf. Nonlinear causal discovery with ANMs. NIPS 2008.
  Peters, Mooij, Janzing, Schölkopf. Causal discovery with continuous ANMs. JMLR 2014.
  Shimizu, Hoyer, Hyvärinen, Kerminen. A linear non-Gaussian acyclic model for causal discovery. JMLR 2006.
  - ANM: $Y = f(X) + N, \;\; X \perp\!\!\!\perp N$

Related work

- Conditional-independence based methods, e.g., PC or IC.

  Spirtes, Glymour, Scheines. Causation, prediction, and search. 2000.
  Pearl. Causality: reasoning and inference. 2009.

  - $X \to Y$ and $Y \to X$ Markov equivalent

- Methods restricting the function class, e.g., ANM, LINGAM.

  Hoyer, Janzing, Mooij, Peters, Schölkopf. Nonlinear causal discovery with ANMs. NIPS 2008.
  Peters, Mooij, Janzing, Schölkopf. Causal discovery with continuous ANMs. JMLR 2014.
  Shimizu, Hoyer, Hyvärinen, Kerminen. A linear non-Gaussian acyclic model for causal discovery. JMLR 2006.

  - 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.
  Janzing, Mooij, Zhang, Lemeire, Zscheischler, Daniusis, Steudel, and Schölkopf. Information-geometric approach to inferring causal directions. AI 2012.

  - IGCI proposed for deterministic relations: $Y = f(X)$

Related work

- Conditional-independence based methods, e.g., PC or IC.
  Spirtes, Glymour, Scheines. Causation, prediction, and search. 2000.
  Pearl. Causality: reasoning and inference. 2009.
  - $X \to Y$ and $Y \to X$ Markov equivalent
- Methods restricting the function class, e.g., ANM, LINGAM.
  Hoyer, Janzing, Mooij, Peters, Schölkopf. Nonlinear causal discovery with ANMs. NIPS 2008.
  Peters, Mooij, Janzing, Schölkopf. Causal discovery with continuous ANMs. JMLR 2014.
  Shimizu, Hoyer, Hyvärinen, Kerminen. A linear non-Gaussian acyclic model for causal discovery. JMLR 2006.
  - 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.
  Janzing, Mooij, Zhang, Lemeire, Zscheischler, Daniusis, Steudel, and Schölkopf. Information-geometric approach to inferring causal directions. AI 2012.
  - IGCI proposed for deterministic relations: $Y = f(X)$

This talk: causal discovery in the *non-deterministic* case based on the **postulate of independence**.

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.

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 $P(Y|X)$.

Independence between causal mechanism and distribution of cause

- Postulate: if $X \to Y$, 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 $P(Y|X)$.

Janzing and Schölkopf. Causal inference using the algorithmic Markov condition. IEEE Trans. on Information Theory 2010.

Lemeire and Dirkx. Causal models as minimal descriptions of multivariate systems. 2006.

Postulate of independence (abstract)

- if $X \to Y$:

$$\boxed{P(Y|X) \text{ ``}\perp\!\!\!\perp\text{'' } P(X)}$$

Postulate of independence (abstract)

- if $X \rightarrow Y$:

$$\boxed{P(Y|X) \text{ ``}\perp\!\!\!\perp\text{''} \; P(X)} \qquad \text{implying} \qquad P(X|Y) \text{ ``}\not\perp\!\!\!\perp\text{''} \; P(Y)$$

Postulate of independence (abstract)

- if $X \to Y$:

$$\boxed{P(Y|X) \text{ ``}\perp\!\!\!\perp\text{'' } P(X)} \qquad \text{implying} \qquad P(X|Y) \text{ ``}\not\!\perp\!\!\!\perp\text{'' } P(Y)$$

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

Postulate of independence (abstract)

- if $X \to Y$:

$$\boxed{P(Y|X) \text{ ``}\perp\!\!\!\perp\text{''} \ P(X)} \qquad \text{implying} \qquad P(X|Y) \text{ ``}\not\!\perp\!\!\!\perp\text{''} \ P(Y)$$

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

$$\boxed{f \text{ ``}\perp\!\!\!\perp\text{''} \ P(X)}$$

Postulate of independence (abstract)

- if $X \rightarrow Y$:

$$\boxed{P(Y|X) \text{ ``}\perp\!\!\!\perp\text{'' } P(X)} \qquad \text{implying} \qquad P(X|Y) \text{ ``}\not\!\perp\!\!\!\perp\text{'' } P(Y)$$

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

$$\boxed{f \text{ ``}\perp\!\!\!\perp\text{'' } P(X)} \qquad \text{implying} \qquad f^{-1} \text{ ``}\not\!\perp\!\!\!\perp\text{'' } P(Y)$$

Postulate of independence (abstract)

- if $X \to Y$:

$$\boxed{P(Y|X) \text{ ``}\perp\!\!\!\perp\text{''}\ P(X)} \qquad \text{implying} \qquad P(X|Y) \text{ ``}\not\!\perp\!\!\!\perp\text{''}\ P(Y)$$

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

$$\boxed{f \text{ ``}\perp\!\!\!\perp\text{''}\ P(X)} \qquad \text{implying} \qquad f^{-1} \text{ ``}\not\!\perp\!\!\!\perp\text{''}\ P(Y)$$

This **asymmetry** between cause and effect can be useful for causal discovery

Postulate of independence (abstract)

- if $X \to Y$:

$$\boxed{P(Y|X) \text{ “}\perp\!\!\!\perp\text{” } P(X)} \qquad \text{implying} \qquad P(X|Y) \text{ “}\not\!\perp\!\!\!\perp\text{” } P(Y)$$

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

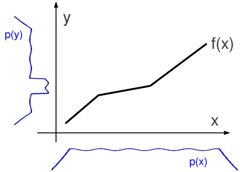
$$\boxed{f \text{ “}\perp\!\!\!\perp\text{” } P(X)} \qquad \text{implying} \qquad f^{-1} \text{ “}\not\!\perp\!\!\!\perp\text{” } P(Y)$$

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

Janzing and Schölkopf. Causal inference using the algorithmic Markov condition. IEEE Trans. on Information Theory 2010.

Asymmetry

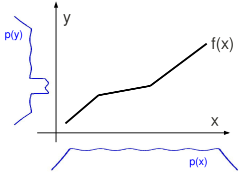
If $X \to Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\!\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\!\perp\!\!\!\perp$" $P(Y)$

Asymmetry

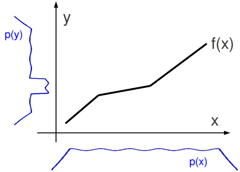
If $X \to Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
Formal asymmetry	**IGCI**: $\mathrm{Cov}(\log f', p_X) = 0$ whereas $\mathrm{Cov}(\log {f^{-1}}', p_Y) \geq 0$	

Asymmetry

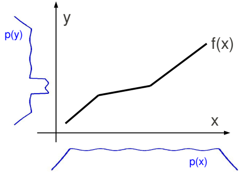
If $X \to Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
Formal asymmetry	**IGCI**: $\mathrm{Cov}(\log f', p_X) = 0$ whereas $\mathrm{Cov}(\log f^{-1'}, p_Y) \geq 0$	?

Asymmetry

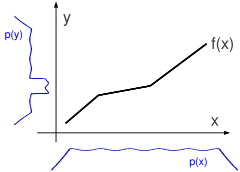
If $X \to Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
Formal asymmetry	**IGCI**: $\mathrm{Cov}(\log f', p_X) = 0$ whereas $\mathrm{Cov}(\log {f^{-1}}', p_Y) \geq 0$	? It is difficult to explicitly formalize independence between $P(Y\|X)$ and $P(X)$

Asymmetry

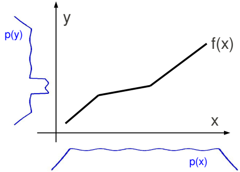
If $X \to Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\!\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\!\perp\!\!\!\perp$" $P(Y)$
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?		

Asymmetry

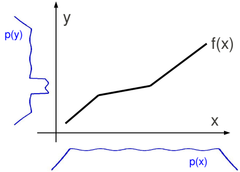
If $X \to Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
Formal asymmetry	**IGCI**: $\mathrm{Cov}(\log f', p_X) = 0$ whereas $\mathrm{Cov}(\log f^{-1'}, p_Y) \geq 0$	? 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

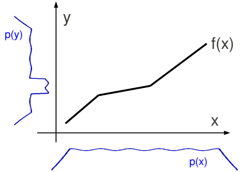
If $X \to Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
Formal asymmetry	**IGCI**: $\mathrm{Cov}(\log f', p_X) = 0$ whereas $\mathrm{Cov}(\log {f^{-1}}', p_Y) \geq 0$	? 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	

Asymmetry

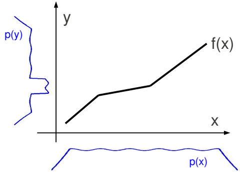
If $X \rightarrow Y$:



	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
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 $X \rightarrow Y$:

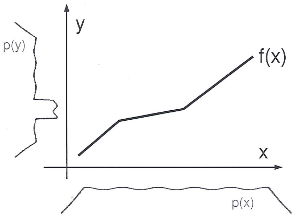


	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
Formal asymmetry	**IGCI**: $\mathrm{Cov}(\log f', p_X) = 0$ whereas $\mathrm{Cov}(\log f^{-1'}, p_Y) \geq 0$	? 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

The inspiration for the last asymmetry came from:

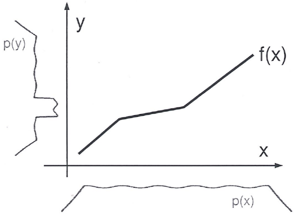
Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij. On causal and anticausal learning. ICML 2012.

Idea of CURE for the simpler deterministic case

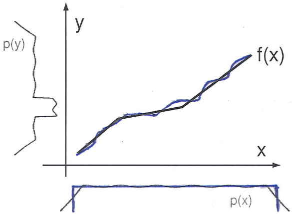


How can f^{-1} be estimated based **only** on p_Y?

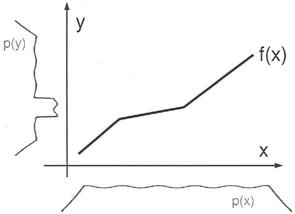
Idea of CURE for the simpler deterministic case



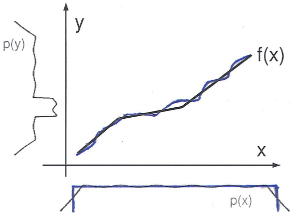
How can f^{-1} be estimated based **only** on p_Y?



Idea of CURE for the simpler deterministic case

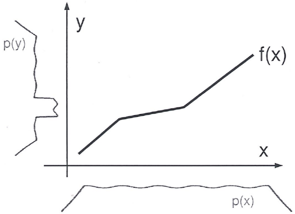


How can f^{-1} be estimated based **only** on p_Y?

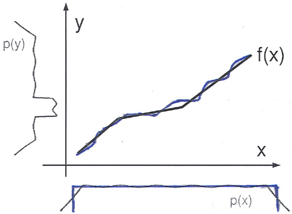


$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

Idea of CURE for the simpler deterministic case



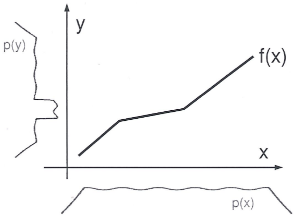
How can f^{-1} be estimated based **only** on p_Y?



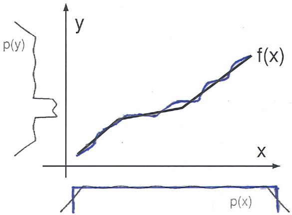
$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

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

Idea of CURE for the simpler deterministic case



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

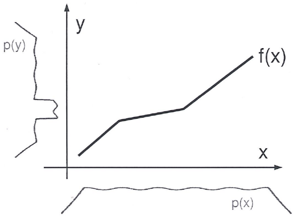


$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

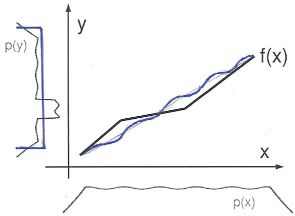
$h^{-1}(y) = F_Y(y)$

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

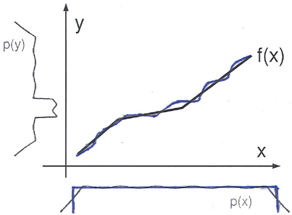
Idea of CURE for the simpler deterministic case



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



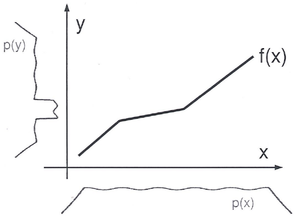
f cannot be estimated based **only** on p_X



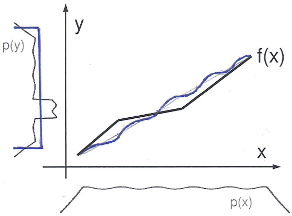
$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

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

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

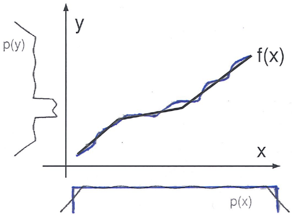


How can f^{-1} be estimated based **only** on p_Y?



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

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



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

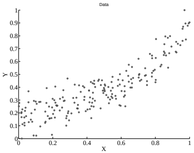
$h^{-1}(y) = F_Y(y)$

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

Estimate $p_{X|Y}$ based on p_Y

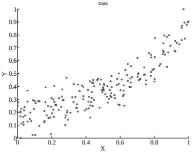
$X \rightarrow Y$
Goal: estimate $p_{X|Y}$ based on p_Y

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



Estimate $p_{X|Y}$ based on p_Y

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

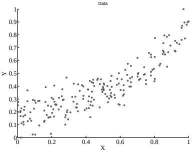


observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x} \in \mathbb{R}^N$

- Model:
  - GP (marg.) **likelihood**: $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, K_{\mathbf{x},\mathbf{x}} + \sigma_n^2 I_N)$ $\quad \boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$

Estimate $p_{X|Y}$ based on p_Y

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

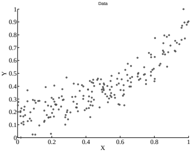


observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x} \in \mathbb{R}^N$

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

Estimate $p_{X|Y}$ based on p_Y

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

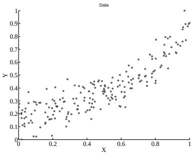


observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x} \in \mathbb{R}^N$

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

Estimate $p_{X|Y}$ based on p_Y

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



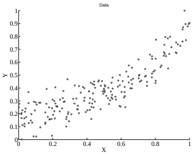
observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x} \in \mathbb{R}^N$

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

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

Estimate $p_{X|Y}$ based on p_Y

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



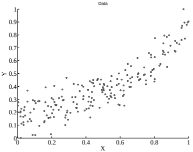
observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x} \in \mathbb{R}^N$

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

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

Estimate $p_{X|Y}$ based on p_Y

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



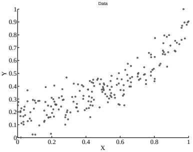
observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x} \in \mathbb{R}^N$

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

$$\hat{p}^{\mathbf{y}}_{X|Y} : (x, y) \mapsto p(x|y, \mathbf{y}) = \int p(x|y, \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \underbrace{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}_{\substack{\text{GP posterior} \\ \text{(N+3)-dimens.}}} d\mathbf{x} d\boldsymbol{\theta}$$

Estimate $p_{X|Y}$ based on p_Y

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



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

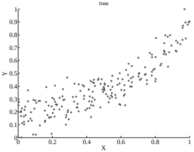
- Model:
  - GP (marg.) **likelihood**: $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, K_{\mathbf{x},\mathbf{x}} + \sigma_n^2 I_N) \qquad \boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$
  - latent's **prior**: $p(\mathbf{x}) = \prod_{i=1}^N \mathcal{U}(0,1)$ and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

$$\hat{p}^{\mathbf{y}}_{X|Y} : (x, y) \mapsto p(x|y, \mathbf{y}) = \int p(x|y, \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \underbrace{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}_{\substack{\text{GP posterior} \\ \text{(N+3)-dimens.}}} d\mathbf{x} d\boldsymbol{\theta}$$

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

Estimate $p_{X|Y}$ based on p_Y

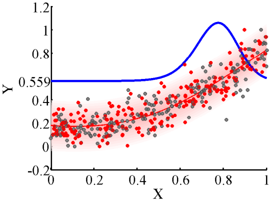
$X \rightarrow Y$
Goal: estimate $p_{X|Y}$ based on p_Y



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

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

$$\hat{p}^{\mathbf{y}}_{X|Y} : (x, y) \mapsto p(x|y, \mathbf{y}) = \int p(x|y, \mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \underbrace{p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}_{\substack{\text{GP posterior} \\ \text{(N+3)-dimens.}}} d\mathbf{x} d\boldsymbol{\theta}$$



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

Grey: $(\mathbf{x}, \mathbf{y})$
Red: $(\mathbf{x}^i, \mathbf{y})$
Blue: $p(x|y = 0.559, \mathbf{y}, \mathbf{x}^i, \boldsymbol{\theta}^i)$

CURE

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

CURE

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

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

CURE

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

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

CURE

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

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

CURE

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

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

CURE

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

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

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

CURE

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

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

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

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

CURE

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

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

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

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

with

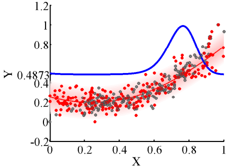
$$\hat{p}^{\mathbf{y}}_{X|Y} : (x, y) \mapsto p(x|y, \mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} p(x|y, \mathbf{y}, \mathbf{x}^i, \boldsymbol{\theta}^i)$$

$$\hat{p}^{\mathbf{x},\mathbf{y}}_{X|Y} : (x, y) \mapsto p(x|y, \mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$$

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

Examples

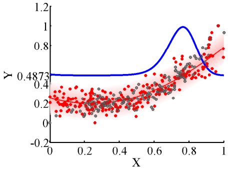
True DAG: $X \rightarrow Y$



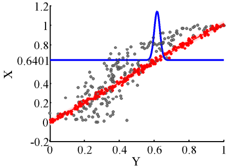
- Often get "good" MCMC samples even when the data are generated by non-Gaussian noise or non-additive noise or non-uniform input.

Examples

True DAG: $X \rightarrow Y$

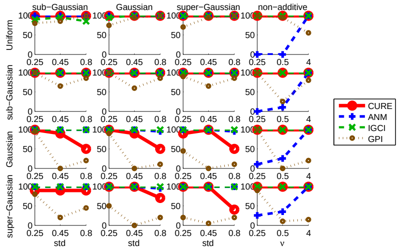


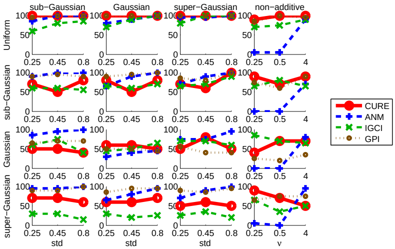
- Often get "good" MCMC samples even when the data are generated by non-Gaussian noise or non-additive noise or non-uniform input.



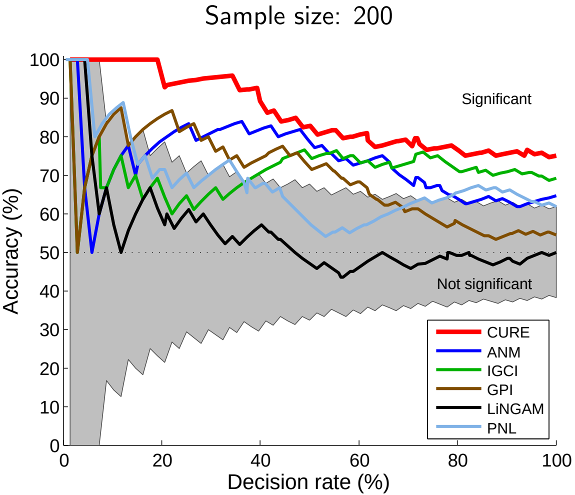
- Often get "bad" MCMC samples when trying to predict based on the distribution of the cause.

Results: simulated data





Results: real data (81 cause-effect pairs)



Cause-effect pairs dataset: Mooij, Peters, Janzing, Zscheischler, Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. 2014.

Conclusion

- Assumption: independence of causal mechanisms $P(\text{cause})$ and $P(\text{effect}|\text{cause})$.

Conclusion

- 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.

Conclusion

- 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 $P(\text{cause})$ 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

Conclusion

- 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 $X \to Y$ if the first estimation is better. Otherwise, infer $Y \to X$.

Conclusion

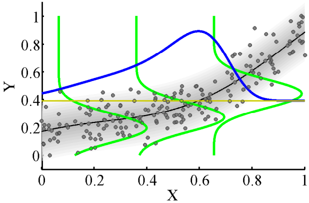
- Assumption: independence of causal mechanisms $P(\text{cause})$ and $P(\text{effect}|\text{cause})$.
- This independence introduces an asymmetry between cause and effect used for causal discovery.
- CURE:
  - Estimate $p_{X|Y}$ based on p_Y
  - Estimate $p_{Y|X}$ based on p_X
  - Infer $X \to Y$ if the first estimation is better. Otherwise, infer $Y \to X$.
- Results seem promising, sampling computationally expensive.

Supervised inverse regression

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

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

is not Gaussian.



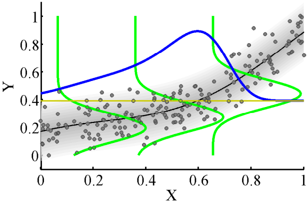
Green: $p(y|x,\mathbf{x},\mathbf{y},\boldsymbol{\theta})$, Blue: $p(x|y,\mathbf{y},\mathbf{x},\boldsymbol{\theta})$

Supervised inverse regression

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

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

is not Gaussian.



Green: $p(y|x, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$, Blue: $p(x|y, \mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$

What about the non-deterministic (noisy) case?

If $X \to Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$

What about the non-deterministic (noisy) case?

If $X \to Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
Alternative asymmetry?		

What about the non-deterministic (noisy) case?

If $X \to Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
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

What about the non-deterministic (noisy) case?

If $X \to Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
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
Estimate from p_Y		

What about the non-deterministic (noisy) case?

If $X \to Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
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
Estimate from p_Y	$Y = f(X) = h(X_u)$ estimate h^{-1} from p_Y	

What about the non-deterministic (noisy) case?

If $X \rightarrow Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
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
Estimate from p_Y	$Y = f(X) = h(X_u)$ estimate h^{-1} from p_Y	$Y = f(X) + E = h(X_u) + E$

What about the non-deterministic (noisy) case?

If $X \to Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
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
Estimate from p_Y	$Y = f(X) = h(X_u)$ estimate h^{-1} from p_Y	$Y = f(X) + E = h(X_u) + E$ estimate $p_{X_u\|Y}$ from p_Y

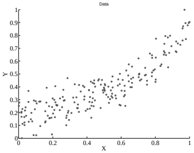
What about the non-deterministic (noisy) case?

If $X \rightarrow Y$:

	Deterministic $Y = f(X)$	Non-deterministic $Y = f(X, E)$
Abstract asymmetry	f "$\perp\!\!\!\perp$" $P(X)$ whereas f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$	$P(Y\|X)$ "$\perp\!\!\!\perp$" $P(X)$ whereas $P(X\|Y)$ "$\not\perp\!\!\!\perp$" $P(Y)$
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
Estimate from p_Y	$Y = f(X) = h(X_u)$ estimate h^{-1} from p_Y	$Y = f(X) + E = h(X_u) + E$ estimate $p_{X_u\|Y}$ from p_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 $p_{X|Y}$)

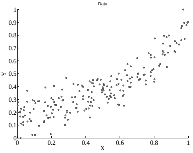


observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x_u} \in \mathbb{R}^N$

Estimate $p_{X|Y}$ based on p_Y

$X \to Y$ 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 $p_{X|Y}$)

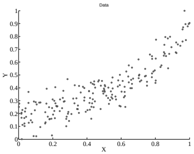
observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x_u} \in \mathbb{R}^N$



- Model:
  - GP (marg.) **likelihood**: $p(\mathbf{y}|\mathbf{x_u}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, K_{\mathbf{x_u},\mathbf{x_u}} + \sigma_n^2 I_N)$ $\quad \boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$

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 $p_{X|Y}$)

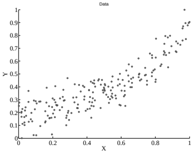


observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x_u} \in \mathbb{R}^N$

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

Estimate $p_{X|Y}$ based on p_Y

$X \to Y$ 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 $p_{X|Y}$)



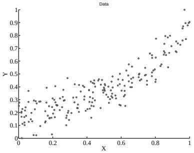
observed: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x_u} \in \mathbb{R}^N$

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

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

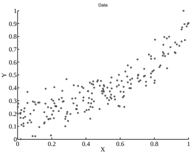


- Model:
  - GP (marg.) **likelihood**: $p(\mathbf{y}|\mathbf{x_u}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, K_{\mathbf{x_u},\mathbf{x_u}} + \sigma_n^2 I_N) \qquad \boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$
  - latent's **prior**: $p(\mathbf{x_u}) = \prod_{i=1}^N \mathcal{U}(0,1)$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

$$\hat{p}^{\mathbf{y}}_{X_u|Y} : (x_u, y) \mapsto p(x_u|y, \mathbf{y}) = \int p(x_u|y, \mathbf{y}, \mathbf{x_u}, \boldsymbol{\theta}) p(\mathbf{x_u}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{x_u} d\boldsymbol{\theta}$$

Estimate $p_{X|Y}$ based on p_Y

$X \to Y$ 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: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x_u} \in \mathbb{R}^N$

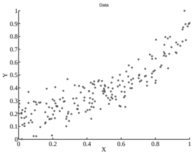
- Model:
  - GP (marg.) **likelihood**: $p(\mathbf{y}|\mathbf{x_u}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, K_{\mathbf{x_u},\mathbf{x_u}} + \sigma_n^2 I_N) \qquad \boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$
  - latent's **prior**: $p(\mathbf{x_u}) = \prod_{i=1}^N \mathcal{U}(0,1)$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

$$\hat{p}^{\mathbf{y}}_{X_u|Y} : (x_u, y) \mapsto p(x_u|y, \mathbf{y}) = \int p(x_u|y, \mathbf{y}, \mathbf{x_u}, \boldsymbol{\theta}) \underbrace{p(\mathbf{x_u}, \boldsymbol{\theta}|\mathbf{y})}_{\text{GP posterior}} d\mathbf{x_u} d\boldsymbol{\theta}$$

Estimate $p_{X|Y}$ based on p_Y

$X \to Y$ 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: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x_u} \in \mathbb{R}^N$



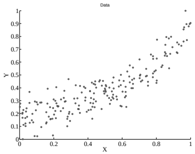
- Model:
  - GP (marg.) **likelihood**: $p(\mathbf{y}|\mathbf{x_u}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, K_{\mathbf{x_u},\mathbf{x_u}} + \sigma_n^2 I_N)$ $\boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$
  - latent's **prior**: $p(\mathbf{x_u}) = \prod_{i=1}^{N} \mathcal{U}(0,1)$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

$$\hat{p}^{\mathbf{y}}_{X_u|Y} : (x_u, y) \mapsto p(x_u|y, \mathbf{y}) = \int p(x_u|y, \mathbf{y}, \mathbf{x_u}, \boldsymbol{\theta}) \underbrace{p(\mathbf{x_u}, \boldsymbol{\theta}|\mathbf{y})}_{\substack{\text{GP posterior} \\ \text{(N+3)-dimens.}}} d\mathbf{x_u} d\boldsymbol{\theta}$$

Estimate $p_{X|Y}$ based on p_Y

$X \to Y$ 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: $\mathbf{y} \in \mathbb{R}^N$, unobserved: $\mathbf{x_u} \in \mathbb{R}^N$



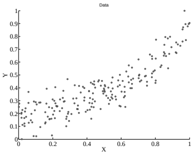
- Model:
  - GP (marg.) **likelihood**: $p(\mathbf{y}|\mathbf{x_u}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, K_{\mathbf{x_u},\mathbf{x_u}} + \sigma_n^2 I_N) \qquad \boldsymbol{\theta} = (\ell, \sigma_f, \sigma_n)$
  - latent's **prior**: $p(\mathbf{x_u}) = \prod_{i=1}^N \mathcal{U}(0,1)$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

$$\hat{p}^{\mathbf{y}}_{X_u|Y} : (x_u, y) \mapsto p(x_u|y, \mathbf{y}) = \int p(x_u|y, \mathbf{y}, \mathbf{x_u}, \boldsymbol{\theta}) \underbrace{p(\mathbf{x_u}, \boldsymbol{\theta}|\mathbf{y})}_{\substack{\text{GP posterior} \\ \text{(N+3)-dimens.}}} d\mathbf{x_u} d\boldsymbol{\theta}$$

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

Estimate $p_{X|Y}$ based on p_Y

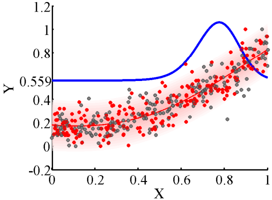
$X \to Y$ with $Y = f(X) + E = h(X_u) + E$
Goal: estimate $p_{X_u|Y}$ based on p_Y
(then use this as an estimate of $p_{X|Y}$)



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

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

$$\hat{p}^{\mathbf{y}}_{X_u|Y} : (x_u, y) \mapsto p(x_u|y, \mathbf{y}) = \int p(x_u|y, \mathbf{y}, \mathbf{x_u}, \boldsymbol{\theta}) \underbrace{p(\mathbf{x_u}, \boldsymbol{\theta}|\mathbf{y})}_{\substack{\text{GP posterior} \\ \text{(N+3)-dimens.}}} d\mathbf{x_u} d\boldsymbol{\theta}$$



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

Grey: $(\mathbf{x}, \mathbf{y})$
Red: $(\mathbf{x_u}^i, \mathbf{y})$
Blue: $p(x_u|y = 0.559, \mathbf{y}, \mathbf{x_u}^i, \boldsymbol{\theta}^i)$

CURE

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

CURE

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

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

CURE

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

1. Estimate $p_{X|Y}$ by $\hat{p}^{\mathbf{y}}_{X_u|Y}$ (using **only** $\mathbf{y}$)
2. Estimate $p_{Y|X}$ by $\hat{p}^{\mathbf{x}}_{Y_u|X}$ (using **only** $\mathbf{x}$)

CURE

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

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

CURE

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

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

CURE

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

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

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

CURE

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

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

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

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

CURE

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

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

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

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

with

$$\hat{p}^{\mathbf{y}}_{X_u|Y} : (x_u, y) \mapsto p(x_u|y, \mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} p(x_u|y, \mathbf{y}, \mathbf{x_u}^i, \boldsymbol{\theta}^i)$$

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

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

Formalization of independence in the deterministic case

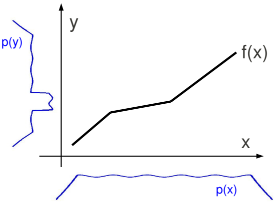
If $X \to Y$ with $Y = f(X)$: $\boxed{f \text{ ``}\perp\!\!\!\perp\text{''} P(X)}$

Formalization of independence in the deterministic case

If $X \to Y$ with $Y = f(X)$: $\boxed{f \text{ ``}\perp\!\!\!\perp\text{''} P(X)}$ implying $f^{-1} \text{ ``}\not\!\perp\!\!\!\perp\text{''} P(Y)$

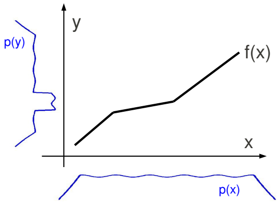
Formalization of independence in the deterministic case

If $X \to Y$ with $Y = f(X)$: $\boxed{f \text{ “}\perp\!\!\!\perp\text{” } P(X)}$ implying $f^{-1} \text{ “}\not\!\perp\!\!\!\perp\text{” } P(Y)$



Formalization of independence in the deterministic case

If $X \to Y$ with $Y = f(X)$: $\boxed{f \text{ “}\perp\!\!\!\perp\text{” } P(X)}$ implying f^{-1} “$\not\perp\!\!\!\perp$” $P(Y)$



Asymmetry

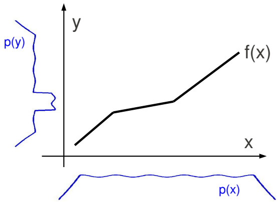
- **Postulate:**

$$\boxed{\mathrm{Cov}(\log f', p_X) = 0}$$

Peaks of p_X do **not correlate** with the slope of f.

Formalization of independence in the deterministic case

If $X \to Y$ with $Y = f(X)$: $\boxed{f \text{ ``}\perp\!\!\!\perp\text{''} P(X)}$ implying f^{-1} "$\not\perp\!\!\!\perp$" $P(Y)$



Asymmetry

- **Postulate:**

$$\boxed{\mathrm{Cov}(\log f', p_X) = 0}$$

Peaks of p_X do **not correlate** with the slope of f.

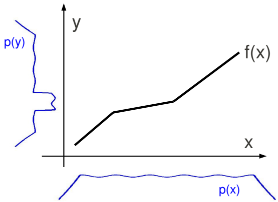
- **Implication:**

$$\boxed{\mathrm{Cov}(\log {f^{-1}}', p_Y) \geq 0}$$

Peaks of p_Y **correlate** with the slope of f^{-1}.

Formalization of independence in the deterministic case

If $X \to Y$ with $Y = f(X)$: $\boxed{f \text{ ``}\perp\!\!\!\perp\text{''} P(X)}$ implying $f^{-1} \text{ ``}\not\!\perp\!\!\!\perp\text{''} P(Y)$



Asymmetry

- **Postulate:**

$$\boxed{\mathrm{Cov}(\log f', p_X) = 0}$$

Peaks of p_X do **not correlate** with the slope of f.

- **Implication:**

$$\boxed{\mathrm{Cov}(\log {f^{-1}}', p_Y) \geq 0}$$

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 $[0, 1]$.

Janzing et al. Information-geometric approach to inferring causal directions. AI 2012.
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 $P(Y|X)$

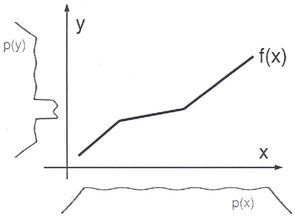
Semi-supervised learning (SSL)

- 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 $P(Y|X) \Rightarrow$
  - **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 $P(Y|X)$

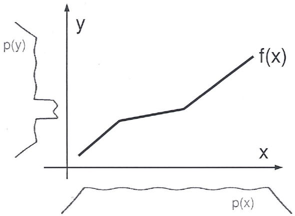
Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij. On causal and anticausal learning. ICML, 2012.

Idea of CURE for the simpler deterministic case

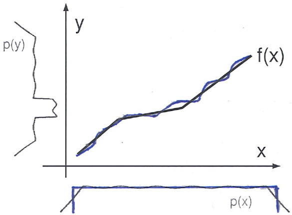


How can f^{-1} be estimated based **only** on p_Y?

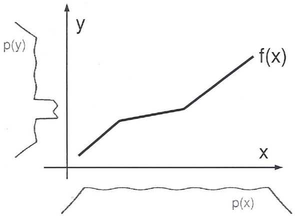
Idea of CURE for the simpler deterministic case



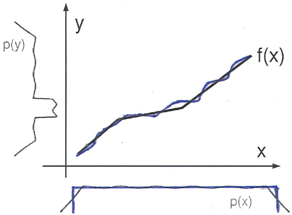
How can f^{-1} be estimated based **only** on p_Y?



Idea of CURE for the simpler deterministic case



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

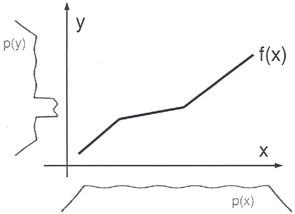


$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

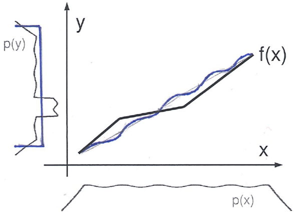
$h^{-1}(y) = F_Y(y)$

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

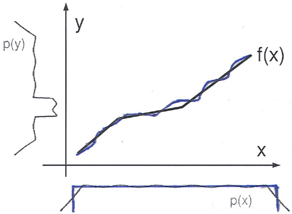
Idea of CURE for the simpler deterministic case



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



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

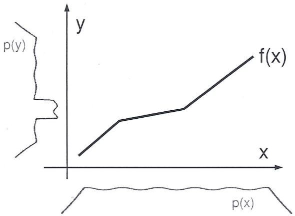


$Y = f(X) = f \circ F_X^{-1}(X_u) = h(X_u)$

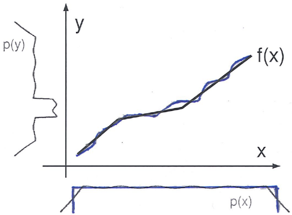
$h^{-1}(y) = F_Y(y)$

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

Idea of CURE for the simpler deterministic case



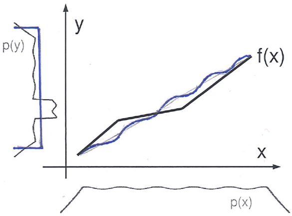
How can f^{-1} be estimated based **only** on p_Y?



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

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

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



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

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

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

$\boxed{\text{Use } g^{-1} \text{ as an estimate for } f}$

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*
- $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$
- The specification of the covariance function implies a specific distribution over functions:
  $cov(f(x_p), f(x_q)) = k(x_p, x_q) = exp(-\frac{1}{2\ell}(x_p - x_q)^2)$
- Finite number of points: $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, K_{\mathbf{x},\mathbf{x}} + \sigma_n^2 I_N)$
- GP regression

Causal discovery: CURE method

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

Causal discovery: CURE method

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

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

Causal discovery: CURE method

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

1. Estimate $\hat{p}^{\mathbf{y}}_{X_u|Y}$
   - Evaluate conditional estimation:

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

with

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

Causal discovery: CURE method

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

1. Estimate $\hat{p}^{\mathbf{y}}_{X_u|Y}$
   - Evaluate conditional estimation:

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

   with

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

2. Estimate $\hat{p}^{\mathbf{x}}_{Y_u|X}$
   - 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 $\hat{p}^{\mathbf{y}}_{X_u|Y}$
   - Evaluate conditional estimation:

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

   with

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

2. Estimate $\hat{p}^{\mathbf{x}}_{Y_u|X}$
   - Evaluate conditional estimation: compute $D_{Y|X}$
3.

Causal discovery: CURE

If $D_{X|Y} < D_{Y|X}$, infer $X \to Y$, otherwise infer $Y \to X$

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