some Causal Methods

liu xin

2018.11.23

Outline

- Additive noise model(ANM)
- Post-Nonlinear causal model(PNL)
- ANM Mixture model
- cause-effect pair challenge

Additive noise model(ANM)

$$y = f(x) + n$$

Test whether x and y are statistically independent.

yes, no causal.

no, then:

test model $y = f(x) + n$:

- do a nonlinear regression of y on x (get an estimate $\hat{f}$ of f)
- residuals $\hat{n} = y - \hat{f}(x)$
- Test whether $\hat{n}$ is independent of x. If so, we accept the model

similarly test the reverse model $x = g(y) + n$

Post-Nonlinear causal model(PNL)

$$x_2 = f_2(f_1(x_1) + e)$$

$$e = f_2^{-1}(x_2) - f_1(x_1)$$

$$y_1 = x_1$$

$$y_2 = g_2(x_2) - g_1(x_1)$$

Post-Nonlinear causal model(PNL)

joint density: $\mathbf{y} = (y_1, y_2)^T$ $\quad p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})/|\mathbf{J}|$

Jacobian matrix: $\mathbf{J} = \left[\partial(y_1, y_2)/\partial(x_1, x_2)\right]$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -{g_1}' & {g_2}' \end{bmatrix}$$

$$|\mathbf{J}| = |g_2'|$$

Post-Nonlinear causal model(PNL)

joint entropy :

$$H(\mathbf{y}) = -E\{\log p_{\mathbf{y}}(\mathbf{y})\} = -E\{\log p_{\mathbf{x}}(\mathbf{x}) - \log|\mathbf{J}|\} = H(\mathbf{x}) + E\{\log|\mathbf{J}|\}.$$

mutual information:

$$\begin{aligned} I(y_1, y_2) &= H(y_1) + H(y_2) - H(\mathbf{y}) \\ &= H(y_1) + H(y_2) - E\{\log|\mathbf{J}|\} - H(\mathbf{x}) \\ &= -E\{p_{y_1}(y_1)\} - E\{p_{y_2}(y_2)\} - E\{\log|g_2'|\} - H(\mathbf{x}), \end{aligned}$$

minimize $I(y_1, y_2)$ ing gardient-descent methods

then, test if they are independent

ANM Mixture model

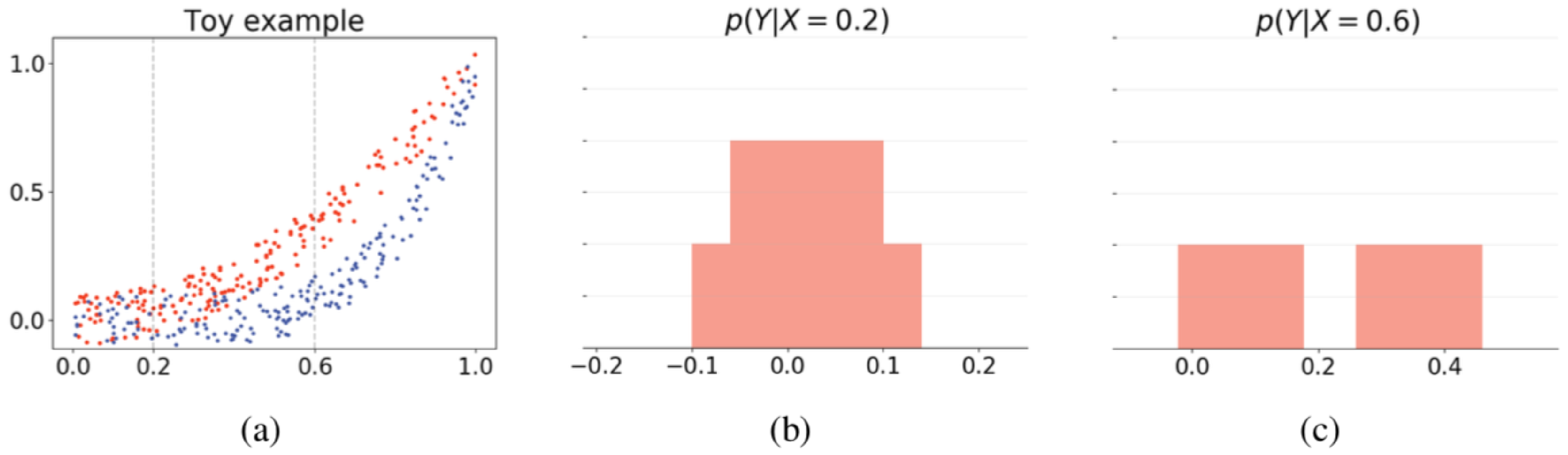

Figure 1: Example illustrating the failure of ANM on the inference of a mixture of ANMs (a) the distribution of data generated from $M_1 : Y = X^2 + \epsilon$ (red) and $M_2 : Y = X^5 + \epsilon$ (blue), where $X \sim U(0,1)$ (x-axis) and $\epsilon \sim U(-0.1, 0.1)$; (b) Conditional $p(Y|X = 0.2)$; (c) Conditional $p(Y|X = 0.6)$. It is obvious that when the data is generated from a mixture of ANMs, the consistency of conditionals is likely to be violated which leads to the failure of ANM.

ANM Mixture model

form: $$Y = f(X;\theta) + \epsilon,$$ e set $\Theta = \{\theta_1, \cdots, \theta_C\}$

Lemma 1. *Let $X \rightarrow Y$ and they follow an ANM-MM. If there exists a backward ANM in the anti-causal direction, i.e.*

$$X = g(Y) + \tilde{\epsilon},$$

the cause distribution (p_X), the noise distribution (p_ϵ), the nonlinear function (f) and its parameter distribution (p_θ) should jointly fulfill the following ordinary differential equation (ODE)

$$\xi''' - \frac{G(X,Y)}{H(X,Y)}\xi'' = \frac{G(X,Y)V(X,Y)}{U(X,Y)} - H(X,Y), \tag{2}$$

where $\xi := \log p_X$, and the definitions of $G(X,Y)$, $H(X,Y)$, $V(X,Y)$ and $U(X,Y)$ are provided in supplementary due to the page limitation.

ANM Mixture model

preliminaries:

Dual PPCA: given $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$ latent representation $\mathbf{x}_n$

$$\mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\epsilon}_n, \quad \boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I})$$

objective function:

$$\text{log-likelihood } \mathcal{L} = -\frac{DN}{2}\ln(2\pi) - \frac{D}{2}\ln(|\mathbf{K}|) - \frac{1}{2}\operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T\right)$$

$$\mathbf{K} = \mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I} \text{ and } \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$$

GP-LVM: nonlinear relation $\boldsymbol{\Phi} = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)]^T$ map

$$\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \beta^{-1}\mathbf{I}$$

Lawrence, Neil. "Probabilistic non-linear principal component analysis with Gaussian process latent variable models." Journal of machine learning research 6.Nov (2005): 1783-1816.

ANM Mixture model

Proposed method

1. $$\mathbf{y}_n = \tilde{\mathbf{W}}\tilde{\mathbf{x}}_n + \boldsymbol{\epsilon}_n, \quad n = 1, \ldots, N \qquad \tilde{\mathbf{x}}_n = \left[\mathbf{x}_n^T, \boldsymbol{\theta}_n^T\right]^T$$

$$p(\tilde{\mathbf{W}}) = \prod_{i=1}^{D} \mathcal{N}(\tilde{\mathbf{w}}_{i,:}|\mathbf{0}, \mathbf{I})$$

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y}, \beta) = -\frac{DN}{2}\ln(2\pi) - \frac{D}{2}\ln\left(|\tilde{\mathbf{K}}|\right) - \frac{1}{2}\operatorname{tr}\left(\tilde{\mathbf{K}}^{-1}\mathbf{Y}\mathbf{Y}^T\right)$$

$$\tilde{\mathbf{K}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \boldsymbol{\beta}^{-1}\mathbf{I} = [\mathbf{X}, \boldsymbol{\Theta}]\,[\mathbf{X}, \boldsymbol{\Theta}]^T + \boldsymbol{\beta}^{-1}\mathbf{I} = \mathbf{X}\mathbf{X}^\mathbf{T} + \boldsymbol{\Theta}\boldsymbol{\Theta}^\mathbf{T} + \boldsymbol{\beta}^{-1}\mathbf{I}$$

2. $$\tilde{\mathbf{K}} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\beta}^{-1}\mathbf{I} = \mathbf{K}_X \circ \mathbf{K}_\theta + \boldsymbol{\beta}^{-1}\mathbf{I}$$

3. $$\underset{\boldsymbol{\Theta},\Omega}{\arg\min}\,\mathcal{J}(\boldsymbol{\Theta}) = \underset{\boldsymbol{\Theta},\Omega}{\arg\min}\left[-\mathcal{L}(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y}, \Omega) + \lambda \log \mathrm{HSIC_b}(\mathbf{X}, \boldsymbol{\Theta})\right]$$

Algorithm 1: Causal Inference

input : $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$ - the set of observations of two r.v.s;
λ - parameter of independence

output : The causal direction

Standardize observations of each r.v.;
Initialize β and kernel parameters;
Optimize (8) in both directions, denote the the value of HSIC term by $\text{HSIC}_{X \to Y}$ and $\text{HSIC}_{Y \to X}$, respectively;
if $\underline{\text{HSIC}_{X \to Y} < \text{HSIC}_{Y \to X}}$ **then**
    The causal direction is $X \to Y$;
else if $\underline{\text{HSIC}_{X \to Y} > \text{HSIC}_{Y \to X}}$ **then**
    The causal direction is $Y \to X$;
else
    No decision made.
end

cause-effect pair challenge

data: cause-effect pairs

target: 1 (x→y); -1 (y→x); 0(other)

feature:

mean and variance normalization

information-theoretic measures:

discrete entory and joint entropy

$$H(X) = -\sum_x p(x)\log(p(x))$$

add bias correction term

$$\hat{H}_m(X) = -\sum_x \frac{n_x}{N}\log(\frac{n_x}{N}) + \frac{M-1}{2N}$$

cause-effect pair challenge

information-theoretic measures:

discrete conditional entory: H(y|x) = H(x,y)-H(x)

discrete mutual information:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I_j(X;Y) = \frac{I(X;Y)}{H(X,Y)} \qquad I_h(X;Y) = \frac{I(X;Y)}{min(H(X),H(Y))}$$

gaussian and uniform divergence:

$$D_g(X) = D(X||G) = H(X) - H(G) = H(X) - \frac{1}{2}\log(2\pi e)$$

$$X_u = \frac{X - min(X)}{max(X) - min(X)} \qquad D_u(X) = D(X_u||U) = H(X_u) - H(U) = H(X_u)$$

cause-effect pair challenge

Hilbert Schmidt Independence Criterion (HSIC)

Pearson correlation

Moments and mixed moments: $m_{1,2} = E[xy^2]$ and $m_{1,3} = E[xy^3]$

skewness, kurtosis

then, feature selection and classification (Gradient Boosting)

AUC score : 0.82