Causal Inference with VAEs

Wu Lin

March 04, 2020

Outline

Example: kidney stones (Peters et al. (2017))

	Treatment= A	Treatment= B
Recovery=1	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$

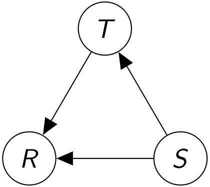
Is Treatment B better?

Example: kidney stones (Peters et al. (2017))

	Treatment= A	Treatment= B
Size of Stones=Small ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Size of Stones=Large ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
Recovery=1	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$

Example: kidney stones (Peters et al. (2017))

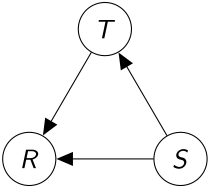
	Treatment= A	Treatment= B
Size of Stones=Small ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Size of Stones=Large ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
Recovery=1	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
P :		



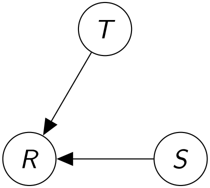
Example: kidney stones (Peters et al. (2017))

	Treatment$=$ A	Treatment$=$ B
Size of Stones=Small ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Size of Stones=Large ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
Recovery=1	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$

P :



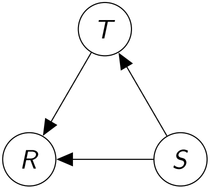
$P_{\mathrm{do}(T=A)}$:



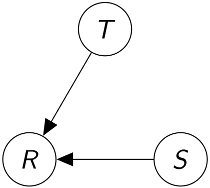
Example: kidney stones (Peters et al. (2017))

	Treatment= A	Treatment= B
Size of Stones=Small ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Size of Stones=Large ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
Recovery=1	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$

P :



$P_{\mathrm{do}(T=A)}$:



Question: Compute $\mathbb{E}_{\mathrm{do}(T=A)}[R]$, $\mathbb{E}_{\mathrm{do}(T=B)}[R]$, where $R \in \{0, 1\}$.

$$\mathbb{E}_{\mathrm{do}(T=A)}[R] := \sum_{R} P_{\mathrm{do}(T=A)}(R) \times R$$

Notations:

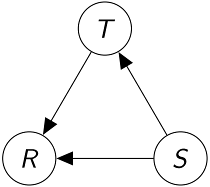
$$P_{\mathrm{do}(T=A)}(S) := P(S|\mathrm{do}(T = A))$$

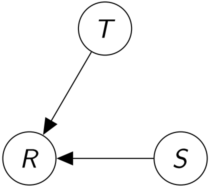
$$P_{\mathrm{do}(T=A)}(S|T = A) := P(S|\mathrm{do}(T = A))$$

$$P_{\mathrm{do}(T=A)}(S, T = A) := P(S|\mathrm{do}(T = A))$$

Example: kidney stones (Peters et al. (2017))

	Treatment= A	Treatment= B
Size of Stones=Small ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Size of Stones=Large ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
Recovery=1	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
P :	$P_{\mathrm{do}(T=A)}$:	





Question: Compute $\mathbb{E}_{\mathrm{do}(T=A)}[R]$, $\mathbb{E}_{\mathrm{do}(T=B)}[R]$, where $R \in \{0, 1\}$.

Identities:

$$P(S) = P_{\mathrm{do}(T=A)}(S), \quad P(R|S, T = A) = P_{\mathrm{do}(T=A)}(R|S, T = A)$$

Example: kidney stones

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{do}(T=A)}[R] \\
=&P_{\mathrm{do}(T=A)}(R=1) \\
=&\sum_w P_{\mathrm{do}(T=A)}(R=1, S=w) \\
=&\sum_w P_{\mathrm{do}(T=A)}(R=1|S=w)P_{\mathrm{do}(T=A)}(S=w) \\
=&\sum_w P_{\mathrm{do}(T=A)}(R=1|S=w, T=A)P_{\mathrm{do}(T=A)}(S=w) \\
=&\sum_w P(R=1|S=w, T=A)P(S=w) \\
=&0.832
\end{aligned}
$$

Similarly, we have $\mathbb{E}_{\mathrm{do}(T=B)}[R] = 0.782$

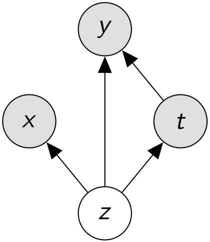
Proxy Variable

From the above example, we can see S is a confounder.
Some confounders are hard to measure: personal preferences, socio-economic status.
We can use some proxy variables to measure these confounders.
Socio-economic status: zip code and job type

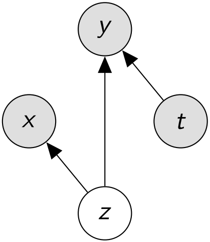
Proxy variable

t: a treatment (eg, medication), where it is binary.
y: an outcome (eg, mortality)
z: an unobserved confounder (eg, socio-economic status)
x: noisy views of z (eg, income and place of residence)

P :



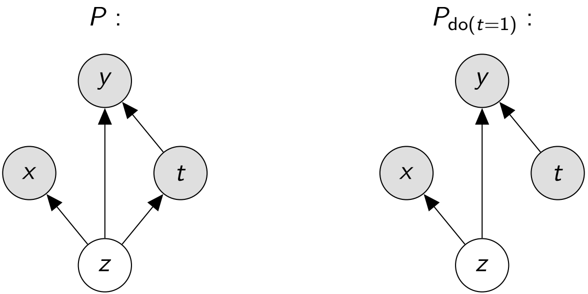
$P_{\mathrm{do}(t=1)}$:



Question: $P_{\mathrm{do}(t=1)}(y|t=1,x) \stackrel{?}{=} P(y|t=1,x)$

Proxy variable

t: a treatment (eg, medication), where it is binary.
y: an outcome (eg, mortality)
z: an unobserved confounder (eg, socio-economic status)
x: noisy views of z (eg, income and place of residence)

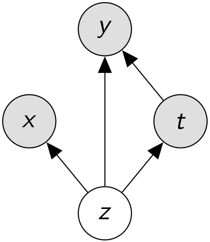


Question: $P_{\mathrm{do}(t=1)}(y|t=1,x) \stackrel{?}{=} P(y|t=1,x)$
Fact: $P(z|x) \neq P(z|t=1,x)$

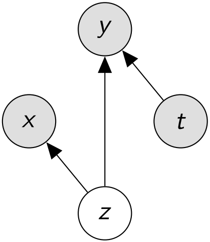
Proxy variable

t: a treatment (eg, medication), where it is binary.
y: an outcome (eg, mortality)
z: an unobserved confounder (eg, socio-economic status)
x: noisy views of z (eg, income and place of residence)

P :



$P_{\mathrm{do}(t=1)}$:



Question: $P_{\mathrm{do}(t=1)}(y|t=1,x) \stackrel{?}{=} P(y|t=1,x)$
Identities:

$$P_{\mathrm{do}(t=1)}(y|t=1,z) = P(y|t=1,z)$$

$$P_{\mathrm{do}(t=1)}(z) = P(z); \qquad P_{\mathrm{do}(t=1)}(x|z) = P(x|z)$$

Proxy Variable

Note that $P_{\mathrm{do}(t=1)}(z|x) = P(z|x)$ due to the following equations.

Proxy Variable

Note that $P_{\mathrm{do}(t=1)}(z|x) = P(z|x)$ due to the following equations.

$$
\begin{aligned}
P_{\mathrm{do}(t=1)}(z, x) &= P_{\mathrm{do}(t=1)}(x|z) P_{\mathrm{do}(t=1)}(z) \\
&= P(x|z) P(z) \\
&= p(z, x)
\end{aligned}
$$

Proxy Variable

Note that $P_{\mathrm{do}(t=1)}(z|x) = P(z|x)$ due to the following equations.

$$\begin{aligned} P_{\mathrm{do}(t=1)}(z,x) &= P_{\mathrm{do}(t=1)}(x|z)P_{\mathrm{do}(t=1)}(z) \\ &= P(x|z)P(z) \\ &= p(z,x) \end{aligned}$$

$$\begin{aligned} P_{\mathrm{do}(t=1)}(z|x) &= \frac{P_{\mathrm{do}(t=1)}(z,x)}{\int P_{\mathrm{do}(t=1)}(z,x)dz} \\ &= \frac{P(z,x)}{\int P(z,x)dz} \\ &= P(z|x) \end{aligned}$$

Proxy Variable

$$\begin{aligned}
P_{\mathrm{do}(t=1)}(y|t=1,x) &= P_{\mathrm{do}(t=1)}(y|x) \\
&= \int P_{\mathrm{do}(t=1)}(y,z|x)dz \\
&= \int P_{\mathrm{do}(t=1)}(y|z,x)P_{\mathrm{do}(t=1)}(z|x)dz \\
&= \int P_{\mathrm{do}(t=1)}(y|z)P_{\mathrm{do}(t=1)}(z|x)dz \\
&= \int P(y|t=1,z){\color{red}P(z|x)}dz
\end{aligned}$$

Proxy Variable

$$
\begin{aligned}
P_{\mathrm{do}(t=1)}(y|t=1,x) &= P_{\mathrm{do}(t=1)}(y|x) \\
&= \int P_{\mathrm{do}(t=1)}(y,z|x)dz \\
&= \int P_{\mathrm{do}(t=1)}(y|z,x)P_{\mathrm{do}(t=1)}(z|x)dz \\
&= \int P_{\mathrm{do}(t=1)}(y|z)P_{\mathrm{do}(t=1)}(z|x)dz \\
&= \int P(y|t=1,z){\color{red}P(z|x)}dz
\end{aligned}
$$

$$
\begin{aligned}
P(y|t=1,x) &= \int P(y,z|t=1,x)dz \\
&= \int P(y|t=1,z,x)P(z|t=1,x)dz \\
&= \int P(y|t=1,z){\color{red}P(z|t=1,x)}dz
\end{aligned}
$$

Issues of proxy variables

From above expressions, we have
$P_{\mathrm{do}(t=1)}(y|t=1,x) \neq P(y|t=1,x)$

However, $P_{\mathrm{do}(t=1)}(y|t=1,z) = P(y|t=1,z)$

Proxy variables (x) are not ordinary confounders (z).

The goal of casual inference

We would like to estimate the individual treatment effect (ITE)

$$\mathrm{ITE}(k) := \mathbb{E}_{\mathrm{do}(t=1)}\left[y|x=k\right] - \mathbb{E}_{\mathrm{do}(t=0)}\left[y|x=k\right]$$

where we assume t is a binary variable.
Similarly, we would like to estimate the average treatment effect (ATE):

$$\mathrm{ATE} := \mathbb{E}\left[\mathrm{ITE}(k)\right]$$

We can approximate $\mathbb{E}_{\mathrm{do}(t=1)}\left[y|x=k\right]$ as

$$\mathbb{E}_{\mathrm{do}(t=1)}\left[y|x=k\right] \approx \frac{1}{M}\sum_{i=1}^{M} y_i$$

where y_i is independently drawn from $P_{\mathrm{do}(t=1)}(y|t=1, x=k)$.

Estimation problem

Clearly, we need to know $P_{\text{do}(t=1)}(y|t=1, x=k)$.

Recall that $P_{\text{do}(t=1)}(y|t=1, x) = \int P(y|t=1, z)P(z|x)dz$

Our goal is to estimate the posterior $P(z|x)$ from the graph P.

Why VAE ?

We would like to make predictions given only x_i is observed without performing inference on the graph P at the test time.
The framework of VAE can achieve that thanks to the amortized inference.
Note that when only x_i is observed, we have to inference t_i, z_i, and y_i.
To this end, we consider the following structured inference network.

$$q(z, t, y, x) = q(z|t, y, x)q(y|t, x)q(t|x)$$

The key idea

A VAE can be used to approximate $P(z|x)$ as shown below.

$$\begin{aligned}
P(z|x) &\approx q(z|x) \\
&= \sum_t \int q(z,t,y|x)dy \\
&= \sum_t \int q(z|t,y,x)q(t,y|x)dy \\
&= \sum_t \int q(z|t,y,x)q(y|t,x)q(t|x)dy
\end{aligned}$$

Now, our goal is to build an inference network to learn $q(z|t,y,x)$, $q(y|t,x)$, and $q(t|x)$ simultaneously.

Objective function

Given observations, $\{x_j, y_j, t_j\}_{j=1}^N$, the Evidence Lower BOund (ELBO) is

$$\sum_{j=1}^{N} \log P(t_j, x_j, y_j)$$

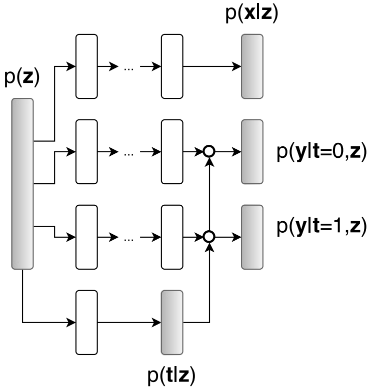
$$= \sum_{j=1}^{N} \log \int q(z_j|t_j, x_j, y_j) \frac{P(t_j, x_j, y_j, z_j)}{q(z_j|t_j, x_j, y_j)} dz_j$$

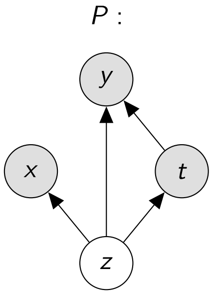
$$\geq \sum_{j=1}^{N} \mathbb{E}_{q(z_j|t_j,x_j,y_j)} [\log P(t_j, x_j, y_j, z_j) - \log q(z_j|t_j, x_j, y_j)] = \underline{\mathcal{L}}$$

To estimate $q(y|t,x)$ and $q(t|x)$, additional terms are included in the objective function of the VAE denoted by $\mathcal{F}_{\text{CEVAE}}$.

$$\mathcal{F}_{\text{CEVAE}} = \underbrace{\underline{\mathcal{L}}}_{\text{max the ELBO}} + \underbrace{\sum_{j=1}^{N} [\log q(y_j|x_j, t_j) + \log q(t_j|x_j)]}_{\text{max the log-likelihood}}$$

Model Network





We can read the model factorization from graph P as
$P(t, x, y, z) = P(z)P(t|z)P(x|z)P(y|t, z)$

Model Network

$$p(z_i) = \prod_{j=1}^{D_z} \mathcal{N}(z_{ij}|0,1); \quad p(x_i|z_i) = \prod_{j=1}^{D_x} p(x_{ij}|z_i)$$

$$p(t_i|z_i) = \mathrm{Bern}(\sigma(f_1(z_i)))$$

where $p(x_i|z_i)$ is an appropriate distribution to model the proxy x_i.
Bernoulli Output:

$$p(y_i|t_i,z_i) = \mathrm{Bern}(\pi = \hat{\pi}_i) \qquad \hat{\pi}_i = \sigma(t_i f_2(z_i) + (1-t_i) f_3(z_i)),$$

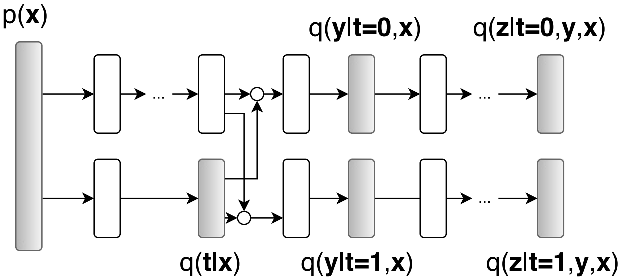
Continuous Output:

$$p(y_i|t_i,z_i) = \mathcal{N}(\mu = \hat{\mu}_i, \sigma^2 = \hat{v}) \qquad \hat{\mu}_i = t_i f_2(z_i) + (1-t_i) f_3(z_i)$$

Each $f_k(\cdot)$ is a neural network parametrized by its own parameters.

TARnet (Shalit et al. (2017)) is used to model the individual treatment effect.

Inference Network



Recall that we need to learn $q(t|x)$, $q(y|t,x)$, and $q(z|t,y,x)$.

Inference Network

$$q(z_i|x_i, t_i, y_i) = \prod_{j=1}^{D_z} \mathcal{N}(\mu_j = \bar{\mu}_{ij}, \sigma_j^2 = \bar{\sigma}_{ij}^2)$$

$$\bar{\mu}_i = t_i \mu_{t=0,i} + (1 - t_i)\mu_{t=1,i} \qquad \bar{\sigma}_i^2 = t_i \sigma_{t=0,i}^2 + (1 - t_i)\sigma_{t=1,i}^2$$

$$\mu_{t=0,i}, \sigma_{t=0,i}^2 = g_2 \circ g_1(x_i, y_i) \qquad \mu_{t=1,i}, \sigma_{t=1,i}^2 = g_3 \circ g_1(x_i, y_i)$$

$$q(t_i|x_i) = \text{Bern}(\pi = \sigma(g_4(x_i)))$$

Bernoulli Output:

$$q(y_i|x_i, t_i) = \text{Bern}(\pi = \bar{\pi}_i)$$

$$\bar{\pi}_i = t_i(g_6 \circ g_5(x_i)) + (1 - t_i)(g_7 \circ g_5(x_i))$$

Continuous Output:

$$q(y_i|x_i, t_i) = \mathcal{N}(\mu = \bar{\mu}_i, \sigma^2 = \bar{v})$$

$$\bar{\mu}_i = t_i(g_6 \circ g_5(x_i)) + (1 - t_i)(g_7 \circ g_5(x_i))$$

TARnet is also used in the inference network.

Results on a toy example

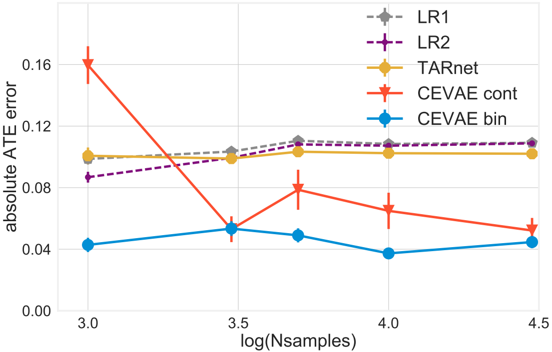
When z_i is a binary variable, we consider the following data generating process.

$$z_i \sim \text{Bern}\,(0.5)$$

$$x_i|z_i \sim \mathcal{N}\left(z_i, \sigma^2_{z_1} z_i + \sigma^2_{z_0}(1-z_i)\right); \quad t_i|z_i \sim \text{Bern}\,(0.75 z_i + 0.25(1-z_i))$$

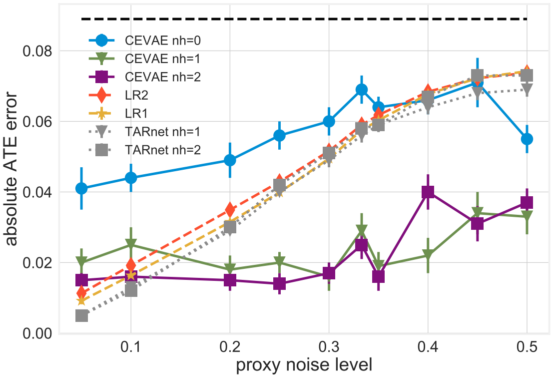
$$y_i|t_i, z_i \sim \text{Bern}\,(\text{Sigmoid}\,(3(z_i + 2(2t_i - 1))))$$

The results obtain by the proposed method:



Results on the Twins dataset

The authors also talk about how to generate the dataset as a benchmark and how to create proxy variables.
The results obtain by the proposed method:



Reference I

- Causal effect inference with deep latent-variable models. NeurIPS 2017. Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., Welling, M.
- Elements of causal inference: foundations and learning algorithms. MIT press 2017. Peters, J., Janzing, D., Schlkopf, B.
- Estimating individual treatment effect: generalization bounds and algorithms. ICML 2017. Shalit, U., Johansson, F. D., Sontag, D.