Causal Inference with VAEs

Wu Lin

March 04, 2020
Outline

Background
   Casual Graphical Model (Review)

VAE
   Problem Formulation
   Objective Function
   Network Architecture

Results
   Results on a toy example
   Results on the Twins dataset
Example: kidney stones (Peters et al. (2017))

<table>
<thead>
<tr>
<th>Recovery = 1</th>
<th>Treatment = A</th>
<th>Treatment = B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\frac{273}{350} = 0.78$</td>
<td>$\frac{289}{350} = 0.83$</td>
</tr>
</tbody>
</table>

Is Treatment B better?
Example: kidney stones (Peters et al. (2017))

<table>
<thead>
<tr>
<th></th>
<th>Treatment = A</th>
<th>Treatment = B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of Stones = Small</td>
<td>( \frac{81}{87} = 0.93 )</td>
<td>( \frac{234}{270} = 0.87 )</td>
</tr>
<tr>
<td>(( \frac{357}{700} = 0.51 ))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size of Stones = Large</td>
<td>( \frac{192}{263} = 0.73 )</td>
<td>( \frac{55}{80} = 0.69 )</td>
</tr>
<tr>
<td>(( \frac{343}{700} = 0.49 ))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovery = 1</td>
<td>( \frac{273}{350} = 0.78 )</td>
<td>( \frac{289}{350} = 0.83 )</td>
</tr>
</tbody>
</table>
Example: kidney stones (Peters et al. (2017))

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

$P:$

![Diagram](attachment:image.png)
Example: kidney stones (Peters et al. (2017))

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

<table>
<thead>
<tr>
<th></th>
<th>$P: T \rightarrow R \rightarrow S$</th>
<th>$P_{do(T=A)}: R \rightarrow T \rightarrow S$</th>
</tr>
</thead>
</table>

Is Treatment B better? $P:$

$P_{do(T=A)}$:
Example: kidney stones (Peters et al. (2017))

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

Is Treatment B better?

$$P :$$

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

$$E_{do(T=A)}[R], E_{do(T=B)}[R],$$

where $R \in \{0, 1\}$.

$$E_{do(T=A)}[R] := \sum_{R} P_{do(T=A)}(R) \times R$$

**Notations:**

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

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

$$P_{do(T=A)}(S, T = A) := P(S|do(T = A))$$
Example: kidney stones (Peters et al. (2017))

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

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

Identities:

$P(S) = P_{\text{do}(T=A)}(S)$, $P(R|S, T = A) = P_{\text{do}(T=A)}(R|S, T = A)$
Example: kidney stones

\[
\mathbb{E}_{\text{do}(T=A)}[R] = P_{\text{do}(T=A)}(R = 1) \\
= \sum_{w} P_{\text{do}(T=A)}(R = 1, S = w) \\
= \sum_{w} P_{\text{do}(T=A)}(R = 1|S = w)P_{\text{do}(T=A)}(S = w) \\
= \sum_{w} P_{\text{do}(T=A)}(R = 1|S = w, T = A)P_{\text{do}(T=A)}(S = w) \\
= \sum_{w} P(R = 1|S = w, T = A)P(S = w) \\
= 0.832
\]

Similarly, we have \( \mathbb{E}_{\text{do}(T=B)}[R] = 0.782 \)
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 (e.g., medication), where it is binary.
- $y$: an outcome (e.g., mortality)
- $z$: an unobserved confounder (e.g., socio-economic status)
- $x$: noisy views of $z$ (e.g., income and place of residence)

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

**Fact:** $P(z|x) \neq P(z|t=1, x)$

**Identities:**
- $P_{do(t=1)}(y|t=1, z) = P(y|t=1, z)$
- $P_{do(t=1)}(x|z) = P(x|z)$
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)

\[ P : \]
\[ P_{do(t=1)} : \]

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

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_{do(t=1)}: $

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

Identities:

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

$P_{do(t=1)}(z) = P(z); \quad P_{do(t=1)}(x|z) = P(x|z)$
Proxy Variable

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

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

$$P_{do(t=1)}(z, x) = P_{do(t=1)}(x|z)P_{do(t=1)}(z)$$
$$= P(x|z)P(z)$$
$$= p(z, x)$$
Proxy Variable

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

\[
P_{do(t=1)}(z, x) = P_{do(t=1)}(x|z) P_{do(t=1)}(z) \\
= P(x|z) P(z) \\
= p(z, x)
\]

\[
P_{do(t=1)}(z|x) = \frac{P_{do(t=1)}(z, x)}{\int P_{do(t=1)}(z, x) dz} \\
= \frac{P(z, x)}{\int P(z, x) dz} \\
= P(z|x)
\]
Proxy Variable

\[ P_{do(t=1)}(y|t = 1, x) = P_{do(t=1)}(y|x) \]

\[ = \int P_{do(t=1)}(y, z|x)dz \]

\[ = \int P_{do(t=1)}(y|z, x)P_{do(t=1)}(z|x)dz \]

\[ = \int P_{do(t=1)}(y|z)P_{do(t=1)}(z|x)dz \]

\[ = \int P(y|t = 1, z)P(z|x)dz \]
Proxy Variable

\[ P_{do(t=1)}(y|t=1, x) = P_{do(t=1)}(y|x) \]

\[ = \int P_{do(t=1)}(y,z|x)dz \]

\[ = \int P_{do(t=1)}(y|z,x)P_{do(t=1)}(z|x)dz \]

\[ = \int P_{do(t=1)}(y|z)P_{do(t=1)}(z|x)dz \]

\[ = \int P(y|t=1,z)P(z|x)dz \]

\[ P(y|t=1, x) = \int P(y,z|t=1, x)dz \]

\[ = \int P(y|t=1, z, x)P(z|t=1, x)dz \]

\[ = \int P(y|t=1, z)P(z|t=1, x)dz \]
Issues of proxy variables

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

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

Proxy variables \((x)\) are not ordinary confounders \((z)\).
The goal of casual inference

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

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

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

\[ \text{ATE} := \mathbb{E} [\text{ITE}(k)] \]

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

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

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

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

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

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.

$$P(z|x) \approx q(z|x)$$

$$= \sum_t \int q(z, t, y|x)dy$$

$$= \sum_t \int q(z|t, y, x)q(t, y|x)dy$$

$$= \sum_t \int q(z|t, y, x)q(y|t, x)q(t|x)dy$$

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

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

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

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

\[
p(z_i) = \prod_{j=1}^{D_z} \mathcal{N}(z_{ij} | 0, 1); \quad p(x_i | z_i) = \prod_{j=1}^{D_x} p(x_{ij} | z_i) \]
\[
p(t_i | z_i) = \text{Bern}(\sigma(f_1(z_i)))
\]

where \( p(x_i | z_i) \) is an appropriate distribution to model the proxy \( x_i \).

Bernoulli Output:

\[
p(y_i | t_i, z_i) = \text{Bern}(\pi = \hat{\pi}_i) \quad \hat{\pi}_i = \sigma(t_i f_2(z_i) + (1 - t_i) f_3(z_i))
\]

Continuous Output:

\[
p(y_i | t_i, z_i) = \mathcal{N}(\mu = \hat{\mu}_i, \sigma^2 = \nu) \quad \hat{\mu}_i = t_i f_2(z_i) + (1 - t_i) f_3(z_i)
\]

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

TARnet is also used in the inference network.
Inference Network

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

\[ \bar{\mu}_i = t_i \mu_{t=0,i} + (1 - t_i) \mu_{t=1,i} \]
\[ \bar{\sigma}_i^2 = t \rho_{t=0,i}^2 + (1 - t_i) \rho_{t=1,i}^2 \]
\[ \mu_{t=0,i}, \sigma_{t=0,i}^2 = g_2 \circ g_1(x_i, y_i) \]
\[ \mu_{t=1,i}, \sigma_{t=1,i}^2 = g_3 \circ g_1(x_i, y_i) \]
\[ q(t_i | x_i) = \text{Bern}(\pi = \sigma(g_4(x_i))) \]

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:

\[ q(y_i | x_i, t_i) = \mathcal{N}(\mu = \bar{\mu}_i, \sigma^2 = \bar{\nu}) \]
\[ \bar{\mu}_i = t_i (g_6 \circ g_5(x_i)) + (1 - t_i) (g_7 \circ g_5(x_i)) \]

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.

\[
\begin{align*}
    z_i &\sim \text{Bern}(0.5) \\
    x_i | z_i &\sim \mathcal{N}(z_i, \sigma^2 z_1 z_i + \sigma^2 z_0 (1 - z_i)) ;
    t_i | z_i &\sim \text{Bern}(0.75 z_i + 0.25(1 - z_i)) \\
    y_i | t_i, z_i &\sim \text{Bern}(\text{Sigmoid}(3(z_i + 2(2t_i - 1))))
\end{align*}
\]

The results obtain by the proposed method:

![Graph showing absolute ATE error as a function of log(Nsamples).](image)
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 obtained by the proposed method:
