ATE estimations from generated RCT data

This notebook examines the use of Bayesian Networks for estimating Average Treatment Effects (ATE) in Randomized Controlled Trials (RCTs) within the Neyman-Rubin potential outcome framework from generated data.

Context

The potential outcomes framework consists of having a set of n independent and identically distributed units. An observation on the i-th unit is given by the tuple (T_i, X_i, Y_i) where:

- $T_i \in \{0, 1\}$ is the treatment assignment.
- $X_i \in \mathbb{R}^d$ is the covariate vector.
- $Y_i \in \mathbb{R}$ is the observed outcome of the treatment.

Under the framekwork's assumption, the *observed* outcome can be expressed as:

$$Y_i = Y_i(T_i) = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

with $Y_i(1)$ and $Y_i(0)$ representing the treated and untreated *potential* outcomes, respectively.

We aim to quantify the effect of a given treatment on the population, namely the quantity $\Delta_i = Y_i(1) - Y_i(0)$. Althought this value cannot be directed calculated due to the presence of counterfactuals (since we cannot observe both the treated and untreated outcomes for a given unit), there exists methods for approximating its expected value, the Avereage Treatment Effect τ:

$$\tau = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_i\right] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

In this notebook, we consider the setting of a Randomized Controlled Trial (RCT), where the ignorability assumption holds:

- $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}$

Here, the *observed outcome* Y_i is dependent of T_i, and independent of other treatments, but the *potential outcomes* $Y_i(0)$ and $Y_i(1)$ are independent of T_i.

We will present estimators of τ using Baysian Networks on generated data through three different methods:

- "Exact" Computation
- Parameter Learning
- Structure Learning

Generated Data

Consider two generative models:

- A linear generative model described by the equation:

$$Y = 3X_1 + 2X_2 - 2X_3 - 0.8X_4 + T(2X_1 + 5X_3 + 3X_4)$$

- And a non-linear generative model described by the equation:

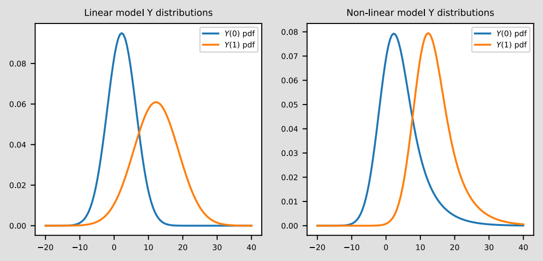
$$Y = 3X_1 + 2X_2^2 - 2X_3 - 0.8X_4 + 10T$$

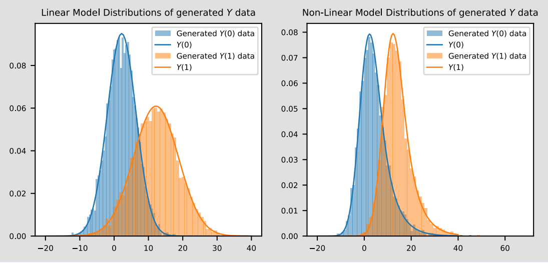
Where $(X_1, X_2, X_3, X_4) \sim \mathcal{N}_4((1, 1, 1, 1), I_4)$, $T \sim \mathcal{Ber}(1/2)$ and (X_1, X_2, X_3, X_4, T) are jointly independent in both of the models.

Data from the models can be obtatined by the functions given below.

Furthermore, the expected values of $Y(0)$ and $Y(1)$ can be explicitly calculated, providing us the theoretical ATE which will serve as a point of reference for the estimations.

Both models have an ATE of $\tau = 10$, their probability density functions are plotted below.



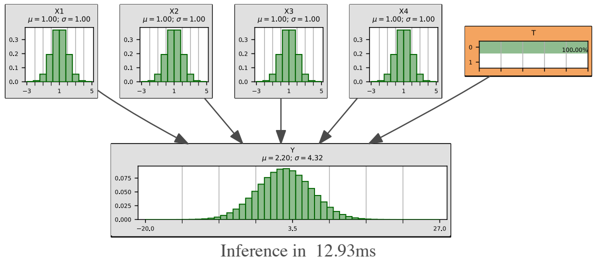


1 - "Exact" Computation

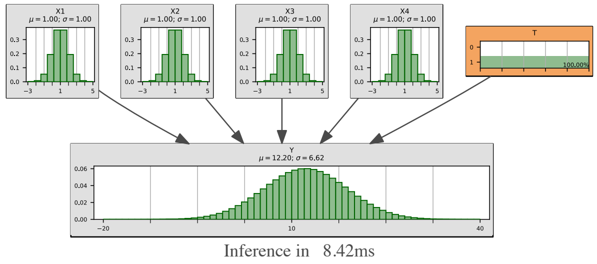
Exact theoretical expected values can be calculated using Bayesian Networks by inputting the data-generating distribution directly into the network. However, since pyAgrum does not support continuous variables as of July 2024, a discretization of continuous distributions is necessary. Consequently, the calculated value will not be exact in a strict sense, but with a sufficient number of discrete states, a close enough approximation can be achieved.

Employing a 10-bin discretization for the covariates and a 60-bin discretization for the outcome, the Bayesian Network estimator accurately approximates the true distribution for both the treated and untreated outcome.

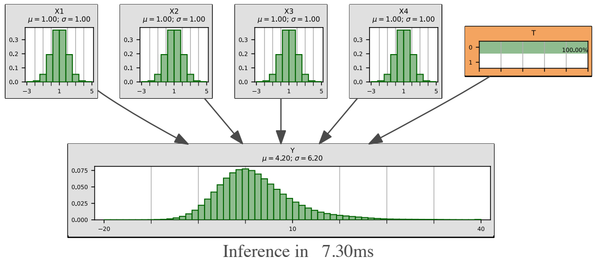
```
BN{nodes: 6, arcs: 5, domainSize: 10^6.07918, dim: 1180037, mem: 9Mo 159Ko 336o}
```



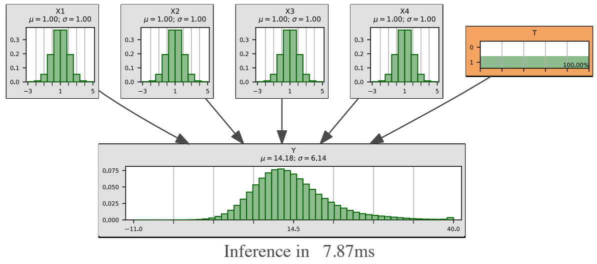
Linear Y(0)



Linear Y(1)



Non-Linear Y(0)



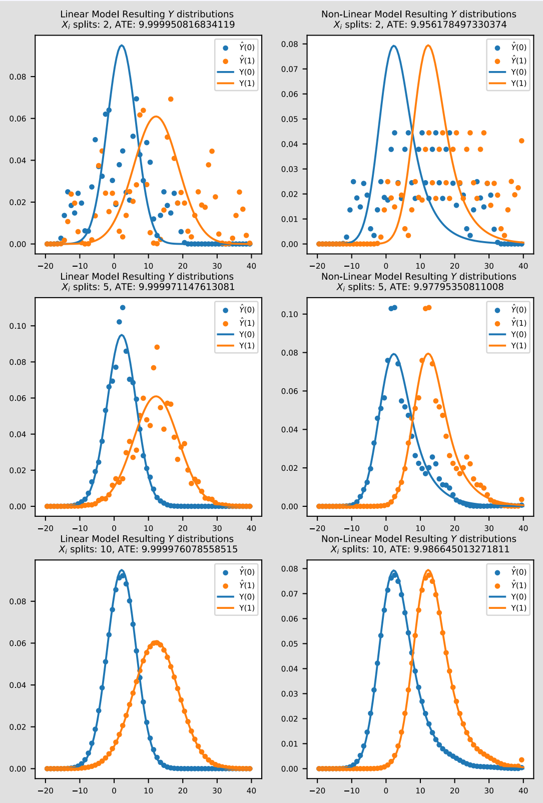
Non-Linear Y(1)

From now on, obtaining the average treatment effect is straightforward using pyAgrum's LazyPropagation exact inference. This involves setting the evidence of the treatment T to 0 and then to 1 to compute the respective posterior CPTs. The ATE is subsequently obtained by performing manipulations on the difference between the CPTs.

```
BN{nodes: 6, arcs: 5, domainSize: 10^6.07918, dim: 1180037, mem: 9Mo 159Ko 336o}
ATE(lin_exbn) = 9.999976078558516

BN{nodes: 6, arcs: 5, domainSize: 10^6.07918, dim: 1180037, mem: 9Mo 159Ko 336o}
ATE(nl_exbn) = 9.986645013271815
```

Let's examine how the fineness of covariate discretization impacts the outcome distribution.

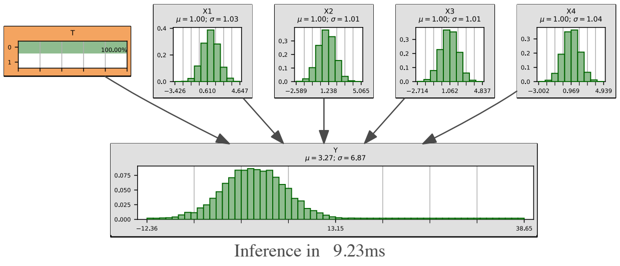


Finer discretization with a greater number of bins results in improved approximations of the probability density functions. However, despite the use of rough discretization, the estimations of the ATE remain remarkably accurate.

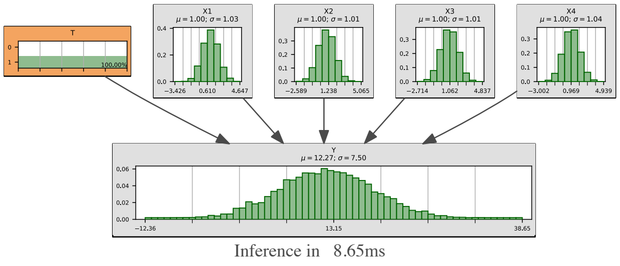
2 - Parameter Learning

Given the data generating function defined above, parameter learning methods can be employed to infer the underlying distribution based on the given structure of the Bayesian network. The default Mutual Information based Inference of Causal Networks (MIIC) algorithm utilized in the `BNLearner` class effectively performs this task.

```
Filename       : /tmp/tmpnm9nbuu8.csv
Size           : (10000,6)
Variables      : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types  : False
Missing values : False
Algorithm      : MIIC
Score          : BDeu  (Not used for constraint-based algorithms)
Correction     : NML  (Not used for score-based algorithms)
Prior          : Smoothing
Prior weight   : 0.000001
```

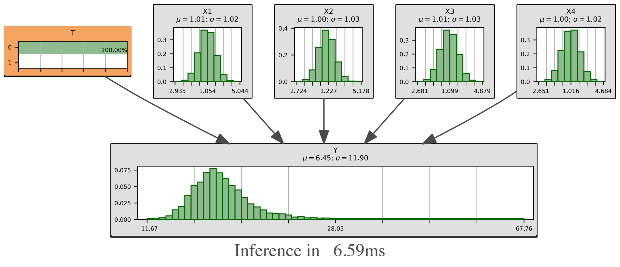


Linear Y(0)

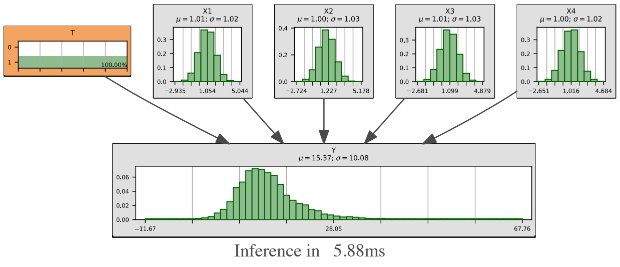


Linear Y(1)

```
Filename       : /tmp/tmpdllb_717.csv
Size           : (10000,6)
Variables      : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types  : False
Missing values : False
Algorithm      : MIIC
Score          : BDeu  (Not used for constraint-based algorithms)
Correction     : NML  (Not used for score-based algorithms)
Prior          : Smoothing
Prior weight   : 0.000001
```

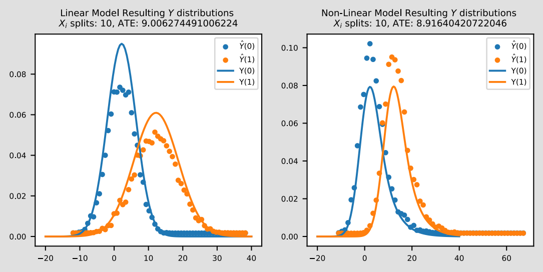


Non-Linear Y(0)

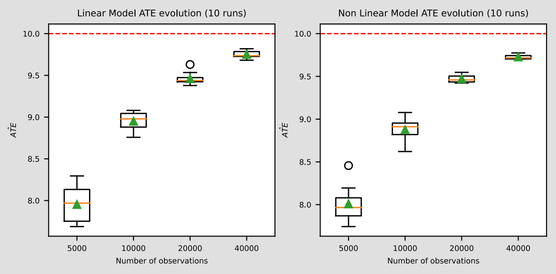


Non-Linear Y(1)

We observe that the inferred outcome distribution generally matches the exact distribution. However, the ATE seems to be biased, as it is consistently smaller.



This underestimation can be further observed with varying numbers of observations in both models.

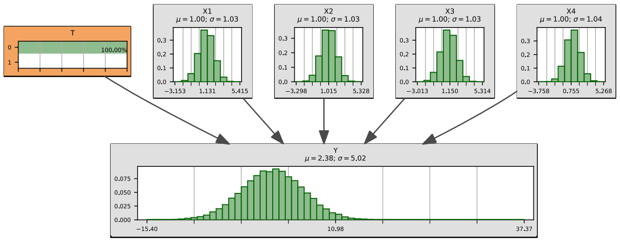


With an increasing number of observations, we observe a convergence of the estimation towards the true value of the ATE and a corresponding reduction in variance.

3 - Structure Learning

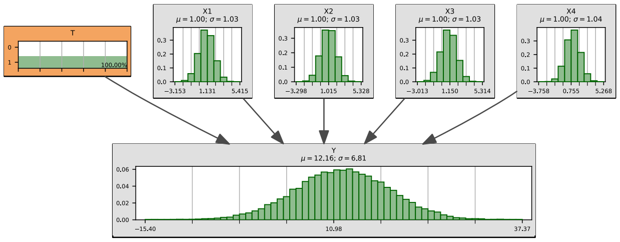
The network's structure and the distributions of the variables can be derived from a sufficiently large dataset through non-parametric learning methods. However, to ensure the integrity of the process, we will impose a slice order on the learner. This ensures that no node is an ancestor of the treatment variable, and no node is a descendant of the outcome variable.

```
Filename               : /tmp/tmpg_0y9hna.csv
Size                   : (40000,6)
Variables              : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types          : False
Missing values         : False
Algorithm              : MIIC
Score                  : BDeu  (Not used for constraint-based algorithms)
Correction             : NML  (Not used for score-based algorithms)
Prior                  : Smoothing
Prior weight           : 0.000001
Constraint Slice Order : {X3:1, X1:1, X4:1, T:0, X2:1, Y:2}
```



Inference in 5.29ms

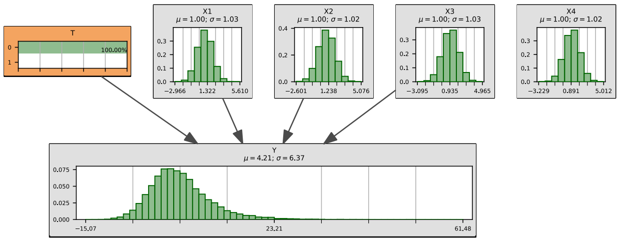
Linear Y(0)



Inference in 5.14ms

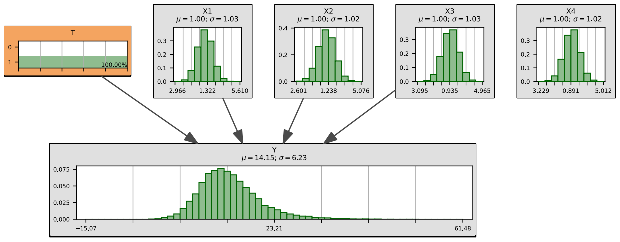
Linear Y(1)

```
Filename               : /tmp/tmprz4y3o19.csv
Size                   : (40000,6)
Variables              : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types          : False
Missing values         : False
Algorithm              : MIIC
Score                  : BDeu  (Not used for constraint-based algorithms)
Correction             : NML  (Not used for score-based algorithms)
Prior                  : Smoothing
Prior weight           : 0.000001
Constraint Slice Order : {X3:1, X1:1, X4:1, T:0, X2:1, Y:2}
```



Inference in 0.73ms

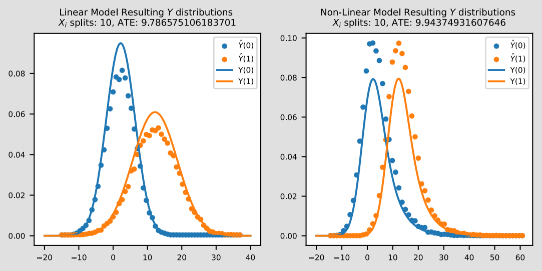
Non-Linear Y(0)

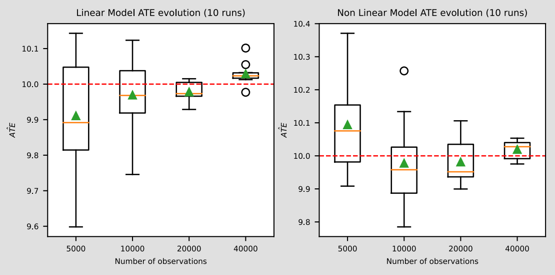


Inference in 0.79ms

Non-Linear Y(1)

With 10,000 observations, structure learning yields a more accurate estimation of the Average Treatment Effect (ATE) compared to parameter learning. This improvement can be attributed to the use of a less complex structure, as opposed to the structure used previously, which features a higher in-degree on the outcome node. The increased number of parameters to be estimated in the outcome model due to this higher in-degree is suboptimal for smaller datasets.


Linear Model Resulting Y distributions
X_i splits: 10, ATE: 9.786575106183701
Ŷ(0)
Ŷ(1)
Y(0)
Y(1)
Non-Linear Model Resulting Y distributions
X_i splits: 10, ATE: 9.94374931607646
Ŷ(0)
Ŷ(1)
Y(0)
Y(1)


Linear Model ATE evolution (10 runs)
ATE
Number of observations
Non Linear Model ATE evolution (10 runs)
ATE
Number of observations