Estimating Causal Effects with the Neural Autoregressive Density Estimator

Sergio Garrido  
Department of Transport  
Technical University of Denmark  
Lyngby, Denmark  
shgm@dtu.dk

Stanislav Borysov  
Department of Transport  
Technical University of Denmark  
Lyngby, Denmark  
stabo@dtu.dk

Jeppe Rich  
Department of Transport  
Technical University of Denmark  
Lyngby, Denmark  
rich@dtu.dk

Francisco Pereira  
Department of Transport  
Technical University of Denmark  
Lyngby, Denmark  
pereira@dtu.dk

Abstract

Estimation of causal effects is fundamental in situations where the underlying system will be subject to active interventions. Part of building a causal inference engine is defining how variables relate to each other, that is, defining the functional relationship between variables given conditional dependencies. In this paper, we deviate from the common assumption of linear relationships in causal models by making use of neural autoregressive density estimators and use them to estimate causal effects within the Pearl’s do-calculus framework. Using synthetic data, we show that the approach can retrieve causal effects from non-linear systems without explicitly modeling the interactions between the variables.

1 Introduction

Making well-informed recommendations with respect to interventions is important in many areas. As an example, when assigning a specific treatment to a sick patient, introducing a new tax policy, or taking an action of a satellite in space. A popular framework for understanding how interventions affect outcomes is that of causal models. Causal models are a powerful mathematical tool to estimate the effects of such interventions by making assumptions of the causes and effects in the studied system. Such causal models can be developed using both experimental data, for example, by randomized control trials, or expert knowledge and observational data, for example, “natural experiments” in economics.

Causal modeling can be thought of as inference engines (Pearl and Bareinboim, 2014). The engine takes causal assumptions and queries as input. And gives answer to those queries as outputs. The input assumptions can be taken in the form of relations between variables, namely, conditional dependencies which together constitute a model of the data. Alternatively, graphical and distributional assumptions can be given as input to the model in order to perform causal discovery from the data. Causal discovery deals with finding the causal relationships between variables. Recent research on recovering causal relationships have focused on using neural networks, for example, Goudet et al. (2018) which uses a deep generative model or in Parascandolo et al. (2017) which recovers independent mechanisms from transformed datasets using adversarial training. After the assumptions on the relationships of variables is done, an estimation procedure follows. The estimation procedure requires assumptions on the functional form of the relationships between variables in order to answer...
the queries that the causal model wants to answer. This paper focuses on the estimation of the queries and the assumptions of the functional form.

There are two main approaches to do causal inference. The potential outcomes approach [Rubin, 1974; Imbens and Rubin, 2015] is based on the fact that it is not possible to see both a treated and a control outcome for the same unit in question, for instance, a patient. As a result, causality is approached as a missing data problem. The usual assumptions used to estimate causal effects (within the potential outcomes framework) are the Stable Unit Treatment Value Assumption (SUTVA) and the Strong ignorability. The former states that the outcome of an individual will not depend on whether other individuals are treated or controlled. The latter states that, given observed covariates, the assignment of the treatment is not dependent on non-observed covariates that affect both the treatment and the outcome of interest. These variables are called confounders in the causality literature. In contrast, Pearl’s causality framework assumes that the causal graph of the system is known or hypothesized. By using a mathematical framework called do-calculus (Pearl, 2009), it is possible to measure the causal effects of one variable on another. It is not guaranteed, though, that the causal effects are identifiable, for example, if certain confounding variables are not observed.

One key point we address in this paper is Pearl’s suggestion to move away from the assumptions of simple parametric equations in the causal model. These assumptions are often based on unrealistic simplifications and are not strictly required to estimate the models. For example, the assumption that a variable of interest is a linear function of its causes or the parameters associated with these. Even though the simplification makes the estimation and, in some cases, the interpretation easier, such simplifications can often bias the estimates of interest. We believe that advancements in deep learning (Goodfellow et al., 2016) and, particularly deep generative modeling, can help researchers move away from these assumptions. Therefore, we propose to use flexible functional forms based on neural networks to model these relationships instead. Particularly, we propose to use the neural autoregressive density estimators (NADE) Larochelle and Murray (2011); Uria et al. (2013) to model distributions required for causal inference.

The contributions of this paper are the following:

- We propose a new way to estimate causal effects using Pearl’s do-calculus. Using this framework, we can estimate arbitrary interventions of which the back-door criterion and the front-door adjustment, explained in Section 4 are special cases.
- The proposed estimation method scales to high dimensions and complex relations between the variables as long as the structure of the relationships between the variables are known.
- The resulting model preserves the properties of a generative model. It is therefore easy to create new data from the model or to impute missing values by doing ancestral sampling.

2 Related work

Estimating causal effects using linear functions has been predominant in the social sciences literature (Angrist, 1990; Miguel and Kremer, 2004; Angrist and Pischke, 2008). A common assumption in these studies, is that the outcome variable is modelled as a linear function of the parameters and the covariates. An exception is when there is additional domain expertise as how the covariates relate with the outcome, in that case, non-linear functions of the covariates can be used. This way of modeling causal effects is restrictive and often only applies to the estimation of aggregated causal effects.

The idea of estimating causal effects beyond these predominant approaches is not new. For example, Hill (2011) proposes a non-parametric Bayesian method—the Bayesian Additive Regression Trees (BART)—to estimate Average Treatment Effects (ATE). Johansson et al. (2016), propose using a feedforward neural network to learn representations from the data and estimate Individual Treatment Effects (ITE). Alaa and van der Schaar (2017) propose using multi-task Gaussian processes to estimate ITE. Yoon et al. (2018) propose using a variation of a Generative Adversarial Network, in order to estimate ITE. The approach proposed in this paper can be viewed as a generalization of the idea by Hartford et al. (2017), where deep neural networks are used to estimate causal effects using instrumental variables (deepIV). Further, Bennett et al. (2019) extend the deepIV approach to a Generalized Method of Moments approach for instrumental variables, however both papers are only concerned with the estimation of causal effects using instrumental variables.
The literature related to autoregressive generative models is rich. As mentioned in the Introduction, we are interested in those approaches that use neural networks to learn useful representations from the data. Here, we can highlight Bengio and Bengio (1999) who introduced the idea of using neural networks to estimate the parameters of distributions of any variable given previously observed variables. Larochelle and Murray (2011); Uria et al. (2013); Germain et al. (2015) extend this idea in order to, for example, share weights between the neural networks, model real valued variables (as opposed to binary) or mask some weights of the networks making their training more efficient. To the best of our knowledge, neural autoregressive models have not been used to perform causal inference.

3 Methods

Consider a set of $J$ random variables $X_1, ..., X_J$ representing a system of interest. Further, let lowercase letter variables $x_1, ..., x_J$ denote realizations of these random variables. Mathematically, the causal effect of an intervention $X_j = x_j$ on an outcome variable $X_o$ can be defined as $P(X_o \mid do(X_j = x_j))$ or $P_{do(X_j = x_j)}(X_o)$ (Pearl, 2009). In these equations, the do operator, “do($X_j = x_j$)”, expresses that the variable variable $X_j$ is forced to be $x_j$. This distribution allows us to estimate other quantities of interest, for example, the Average Treatment Effect (ATE) defined as

$$ATE = E[X_o \mid do(X_j = x_j)] - E[X_o \mid do(X_j = \hat{x}_j)].$$

(1)

It also allows the estimation of more general causal effects such as the Individual Treatment Effect, defined for a specific individual with characteristics $x_c$,

$$ITE = E[X_o \mid do(X_j = x_j), x_c] - E[X_o \mid do(X_j = \hat{x}_j), x_c].$$

(2)

3.1 Assumptions

Since this paper deals with the estimation of causal models, the three following assumptions should be made:

- The underlying causal graph of the process is known.
- The distribution family of every variable in the process is known. This is not so restrictive as we will see in Section 4.1.3.
- The system can be decomposed as a factorization of conditional distributions. In other words, there are no feedback loops or cycles in the causal graph.

These assumptions, either implicitly or explicitly, are in accordance with the common practice when working with causal effects. However, it is worth stressing that, in contrast to the majority of models in the literature, we do not make any assumptions about the functional form of the relationships between the variables and the corresponding parameters. These assumptions are relaxed exactly for the reason of being able to estimate more robust causal effect models for which these relationships are not known in advance.

3.2 Autoregressive density estimators

Before introducing the relationship with causal models, we briefly explain the concept of neural autoregressive estimators. These models (Bengio and Bengio, 1999; Larochelle and Murray, 2011; Germain et al, 2015) are a class of generative models, which seek to estimate the full distribution of the variables. Given that the underlying causal model can be represented as a causal graph, these joint distributions can be defined by the conditional factorization of the joint distribution

$$P(X) = \prod_j P(X_j \mid PA(X_j)).$$

(3)

This equation expresses that the joint probability distribution can be formulated as the product of the individual distributions conditioned on their parents, $PA(X_j)$. Even though Bengio and Bengio (1999) suggest that these models can be constructed on the basis of probabilistic models, most of the literature has moved away from this approach. A common approach is to prove that the densities estimated by these models are robust with respect to changes in the factorization. There is a body of
evidence from the deep generative model literature that suggests that this is indeed possible (Uria et al., 2013; Germain et al., 2015; Papamakarios et al., 2017). However, we are interested in causal models and, as a consequence, further assumptions on the distribution factorization are required. In other words, we cannot just set an arbitrary factorization of conditional distributions as in the body of neural autoregressive estimator literature, but instead assume a “causal factorization” as it is called in Parascandolo et al. (2017). The individual conditional distributions are also called Individual Causal Mechanisms (ICM) or Markov Kernels Peters et al. (2017).

To parametrize the conditional distributions on the right hand side of the factorization in Eq. (3), we propose to use neural networks. A neural network takes the parents of a variable $X_j$, call this $PA(X_j)$, and estimate parameters $\Theta^j$ of a distribution for the variable of interest. The modelled parametric distributions can contain both categorical random variables (Bengio and Bengio, 1999; Larochelle and Murray, 2011; Germain et al., 2015) and continuous random variables (Bengio and Bengio, 1999; Uria et al., 2013). Let $W^j_l$ and $b^j_l$ denote the neural network weights and biases to be learned, associated with the layer $l$, and $X_j$ the variable of interest. In addition, $h^j_l$ represents the outputs of the hidden layer neurons and $f^j_l$ a non-linear transformation. Mathematically, we can write the neural network with $L$ hidden layers as

$$
\begin{align*}
    h^j_1 &= f^j_1(PA(X_j)W^j_1 + b^j_1), \\
    \vdots \\
    h^j_l &= f^j_l(h^j_{l-1}W^j_l + b^j_l), \\
    \vdots \\
    \Theta^j &= f^j_L(h^j_{L-1}W^j_L + b^j_L)
\end{align*}
$$

(4)

These neural networks are trained jointly, using the negative log-likelihood in Equation (3) as the loss function. Figure 1 shows an example of such a model for a causal graph of the “kidney stone” example (Charig et al., 1986; Peters et al., 2017). The causal graph is represented above. This graph contains assumptions of conditional dependencies between variables, for example, that $R$ is dependent of both $KS$ and $T$. Below, our proposed approach is represented. Each neural network takes parents of each variable as input and outputs the parameters of the respective variable. In this case, the output is a single neuron, since each random variable follows a Bernoulli distribution with parameter $p$.

### 3.3 Connection to Pearl’s Do-Calculus

The way to connect the model described in the previous subsection to the do-calculus framework is by realizing that Equation (3) is closely related to the truncated factorization formula (Pearl, 2009). For an intervention where $X_j$ takes a single value, \( \text{do}(X_j = x_j') \), this corresponds to

\[
P(X_{o \neq j} \mid \text{do}(X_j = x_j')) = \begin{cases} 
\prod_{o \neq j} P(X_j \mid PA(X_j)) & \text{if } x_j = x_j' \\
0 & \text{otherwise}
\end{cases} \tag{5}
\]

The only condition required to make such a model causal is the assumption that the parent-child relationship in the distribution is a cause-effect relationship. Based on this assumption, we can perform interventions on variables provided that the parameters of the likelihood function are properly estimated. In practice, an intervention can be described using the following steps:

1. Set the variable to be intervened, e.g. $X_j$, to the value or distribution of interest.
2. Perform a forward pass on the estimated networks with the value or values of variable $X_j$. Since each variable is conditioned only on its parents, values that are not affected “downstream” in the causal graph by $X_j$ will remain unchanged.
3. In the output parameters of the distributions, change the conditional distribution of the intervened variable $P(X_j \mid PA(X_j))$ to 1, if an intervention of a single value (an atomic intervention) was performed. If the intervention is a distributional intervention, change $P(X_j \mid PA(X_j))$ for such intervention.
Following these steps will yield the truncated factorization. It will not, however, estimate the causal effects associated with the intervention. In other words, the neural estimator will approximate the causal conditionals which are fundamental to get the causal effects. In order to get causal effects, it is required that we use the rules of the do-calculus in Pearl (2009, 2012) and Pearl and Mackenzie (2018).

4 Simulation studies

4.1 The kidney stone example: the backdoor adjustment formula

The kidney stone example (Charig et al., 1986) is a classical example of the Simpson’s paradox (Simpson, 1951) used in the causality literature (Peters et al., 2017) to prove that conditioning is not the same as intervening. The setup is simple. There are three binary variables: Kidney Size, $KS$ (small or large); Treatment, $T$ (A or B); and Recovery status $R$ (recovered or not). The causal model is known beforehand as introduced in Section 3.2 and in Figure 1, and the observed values, for the binary case, are shown in Table 1. From this table it is possible to verify that, if the confounder $KS$ is not taken into account, the treatment effects will be biased. However, from the table it is not possible to infer the relationship between the confounder and the treatment variable. We simulated data from this causal graph using the distribution in Table 1 and modifications of it. Moreover, we used the proposed method to estimate the ground truth causal effects of interventions and to check this against their analytical equivalents. The idea behind the modifications from the basic setting is to test whether the proposed method would work on continuous treatment, confounding, and outcome variables; and in non-linear scenarios were the proposed method overcomes the problem of a linear assumption.
Table 1: Kidney stone example using the values from [Charig et al. (1986); Peters et al. (2017)]

| Treatment type | A      | B      |
|----------------|--------|--------|
| Small          | 93% (81/87) | 87% (234/270) |
| Large          | 73% (192/263) | 69% (55/80) |
| Total          | 78% (273/350) | 83% (289/350) |

The treatment effects of the kidney stones example are calculated analytically using a result known as the “backdoor adjustment formula” ([Pearl, 2009]). This result can be derived by using the rules of do-calculus in combination with the estimated probability distribution:

\[
P(R = 1 \mid do(T = A)) = \sum_{ks} P(R = 1 \mid T = A, KS = ks)P(KS = ks)
\]

\[
P(R = 1 \mid do(T = B)) = \sum_{ks} P(R = 1 \mid T = B, KS = ks)P(KS = ks)
\]

\[
ATE = E[R = 1 \mid do(T = A)] - E[R = 1 \mid do(T = B)] \tag{6}
\]

All of the terms in Equation (6) can be retrieved from the proposed estimation framework.

4.1.1 Study Case I: Binary Variables

To estimate causal effects with binary outcomes, we sample 5000 observations from the distribution shown in Table 1. Table 2 summarizes the simulated data distribution which, as expected, is consistent with the original distribution. The applied neural networks are presented in Figure 1. Each neural network takes the parents of the outcome variable as input and produce the parameters of a distribution as the output. In this case, all the variables are modelled as random variables with Bernoulli distributions and, therefore, the output is a single parameter \( p \) from which we can evaluate the likelihood for each variable. Because of the ease of this task, neural networks with only 2 hidden layers with 4 neurons each were used.

Table 2: The probability from the generated data. The table follows the same behavior of Table 1, where the aggregated probability of Recovery given B is higher than the probability of Recovery given A; while the individual probabilities favor A

| Generated data probabilities |        |
|------------------------------|--------|
| \( p(R = 1 \mid B) \)       | 82.70% |
| \( p(R = 1 \mid A) \)       | 77.52% |
| \( p(R = 1 \mid \text{large}, A) \) | 71.73% |
| \( p(R = 1 \mid \text{large}, B) \) | 66.66% |
| \( p(R = 1 \mid \text{small}, A) \) | 95.00% |
| \( p(R = 1 \mid \text{small}, B) \) | 87.74% |
| ATE (eq. 6)                  | 6.19%  |

Table 2 presents the obtained causal conditionals. If we compare these results with those in Table 2, it can be verified that the probabilities are recovered to a high degree. The fact that we are able to recover the probabilities from the simulated data is not surprising as knowing the structure of the causal graph boils down to a conditional density estimation or, equivalently, a supervised learning problem. For a linear model, the resulting ATE is 6.43%. This value is very close to the analytical one since, in this model, the probability of the outcome does not include complicated non-linear relationships between the covariates.

4.1.2 Study Case II: Continuous response variables

In the second simulation study, we generated the confounding variable and the treatment variable as before but now changed the recovery variable to be normally distributed with \( R \sim N(R \mid \mu = \)
Table 3: We are able to recover the probabilities from the generated data, and estimate the ATE with relatively high precision.

| Basic model results                  |
|--------------------------------------|
| $p(\text{large})$                   | 48.97%            |
| $p(R = 1 \mid \text{large}, A)$     | 75.58%            |
| $p(R = 1 \mid \text{large}, B)$     | 69.68%            |
| $p(R = 1 \mid \text{small}, A)$     | 95.68%            |
| $p(R = 1 \mid \text{small}, B)$     | 88.94%            |

**ATE (eq. (6))** 6.32%

$T * 4 + \exp(KS), \sigma = 2)$. The analytical causal effect of the treatment variable is 4. In this case, the output of the neural network corresponding to the Recovery variable has two units, one for the mean and one for the variance of the normal distribution. In that way, we are able to estimate the likelihood of our points given the assumption of a normally distributed recovery variable. Table 4 shows the results of the proposed method. The treatment effect from the generated data is the same as the analytical one. The estimate deviates by 2.8% of the total effect. This suggest that the proposed method can also estimate linear effects. The estimated effect with a linear model is 3.73%, this, as above, is expected since there is no non-linear relation between the parents and the outcome variable.

Table 4: Estimates of the ATE under a continuous recovery model.

| Continuous case ATE (eq. (6))        |
|--------------------------------------|
| Analytical ATE                       | 4               |
| Generated ATE                        | 4.00            |
| Estimated ATE                        | 3.89            |

4.1.3 Study Case III: Continuous confounding variables

In the third simulation study, we generated the confounding variable as a continuous variable. Since the confounding variable represents a measure of size, it only makes sense to simulate it as a positive real number. In this simulation case, we are also interested in testing the impact of the second assumption in Section 3.1 (“The distribution family of every variable in the process is known.”). To do this, we generated two different data sets with different distributions of the confounding variable—gamma and log-normal. The gamma distribution cannot be learnt in a straightforward way using neural networks, so we used the log-normal parametrization for both cases. That is, in the first neural network of Figure 1, the output is now a mean and a variance with which we can evaluate the likelihood of the kidney stone sizes.

For the first data set, we used a Gamma distribution with shape 5 and scale 2, which implies a mean equal to 10. The treatment variable is still a binary variable. However, we now use the mean of the confounding variable as a cutoff to decide whether the kidney stone is small or large $T = 1\{KS > 10\}$. After deciding whether the kidney stone is large or small, we assigned the treatment using the same probabilities as in Table 1. Finally, the response variable was defined as a normal distribution $R \sim N(R \mid \mu = T * 4 + KS, \sigma = 2)$. In the second data set, we simulated the confounding variable from a log-normal distribution with $\mu = 2.5$ and $\sigma = 0.25$. We simulated the rest of the variables as in the Gamma case.

Another difficulty in this setting is related to the fact that we have a continuous confounding variable. This implies that the sum in Equation (6) must be changed for an integral. If the integral is mathematically intractable, numerical approximations can be used. We propose using Monte Carlo sampling in order to obtain an approximation of the interventional distribution and the ATE. To do this, we simply sample from the inferred distribution and propagate forward through the neural network to get different estimates of the response variable.
The results of this experiment are shown in Table 5. The results are in line with the rest of the experiments. Even though the analytical treatment effects are not recovered perfectly, they are recovered to a high degree using the proposed autoregressive approach. The second message that these results reveal is that the assumption of knowing the exact parametrization of the distribution is not so important. This is reinforced by the possibility of increasing the complexity of any modeled variable through a mixture of densities (Bishop [1994]) or normalizing flows (Agnelli et al. [2010]; Tabak and Turner [2013]; Rezende and Mohamed [2015]). Both approaches add more flexibility up to a point where the target distribution can be approximated perfectly. With the linear model, for the gamma generated experiment, the causal effect estimated is 3.78% and 4.21% for the log-normal generated set.

Table 5: Estimates of the ATE with a continuous confounding variable generated by a Gamma and a Log-normal distribution.

| Continuous confounding ATE (eq. (6)) | Gamma confounding | Log-normal confounding |
|-------------------------------------|------------------|-----------------------|
| 1 sample                            | 3.77             | 4.15                  |
| 5 samples                           | 3.77             | 4.15                  |
| 25 samples                          | 3.77             | 4.15                  |
| 50 samples                          | 3.77             | 4.15                  |
| Total analytical ATE                | 4                |                       |

4.1.4 Study Case IV: Non-linear relation between the confounding variable and the treatment variable over the recovery variable

In the fourth simulation study, we generated data for both the confounding and the recovery value as continuous variables. In this case, the effect of the confounding and treatment variable on the recovery is non-linear. In fact, this setting allows us to do Individual Treatment Effects (ITE) depending on the confounding variable, which is one of the main contributions of the paper.

The confounding variable was simulated from a Log-normal distribution, $K S \sim \text{LogNormal}(K S \mid \mu = 2.5, \sigma = 0.25)$. The treatment, as in the previous case, was modeled using a Bernoulli distribution with probability $p = \frac{1}{1 + \exp((K S - \mu) / 10)}$ and so $T \sim \text{Bernoulli}(p)$. $p$ was modeled using a transformation of $K S$ in order to avoid extreme probabilities. Finally, the response variable was simulated from a normal distribution, $R \sim N(\frac{K S + 3}{K S + 3}, 1)$. These definitions make treatment effects non-linear, as seen in Figure 2. The results in this figure are not surprising for the linear model, as the linear model is only able to retrieve a pooled estimate of the treatment effects. On the other hand, the neural network model is able to estimate the individual treatment precisely.

Nevertheless, Figure 2 shows some weakness in our approach. As we get further away from the support of the data, the estimated treatment effects get worse. Particularly, it is possible to see that, around $K S = 8$, the estimated linear effect starts deviating from the analytical one. The possibility to estimate conditional treatment effects using neural networks, therefore depends heavily on the ability of the neural network to extrapolate.

Using a linear model, we learn a single causal effect as in the simulation studies above. We get a unique causal effect of 9.88, of course, regardless of the value of $K S$. The problem with this estimation is that the non-linear relation between the confounding variable and the treatment is aggregated and thereby form aggregation bias (Stoker [2016]). This bias can be of any order of magnitude. For example, in our simulation study, if we had a recovery generated by $R \sim N(R \mid 4 * T * \exp(-99l) + K S)$, the aggregated treatment effect would be close to 2, underestimating the effect for small $K S$ and overestimating it for large ones.

4.2 The front door adjustment

In the previous simulation studies all the variables in the causal graph were observed. In fact, because of the simplicity of the graph, in Section 4.1.4 the causal effects collapsed to a conditional estimation. In the last simulation study, we seek to prove that our proposed approach works in conditions where
Comparison between true and estimated conditional Treatment Effects

Linear TE  Neural TE  True TE

5.0  7.5  10.0  12.5  15.0  17.5  20.0  22.5  25.0

Figure 2: Comparison between the linear and neural network models using the back-door adjustment formula. A histogram of $KS$ is represented in the background of the plot.

there are unobserved confounders and, as a result, the back-door criterion does not hold. In Figure 3, a causal graph with those conditions is presented. The dashed line of the $KS$ circle denotes an unobserved variable and, hence, Equation (6) cannot be used. In this case, $KS$ was simulated from standard normal distribution, $T$ was simulated from a normal distribution $T \sim N(\sin(KS), 0.1)$, $Mg$ a mediator variable representing the amount of a certain chemical (in this case Magnesium) in the treatment, was simulated from a normal distribution $Mg \sim N(1 + T^2, 0.1)$ and finally, $R$ was simulated from a normal distribution $R \sim N \left( \sin(KS^2) + \frac{5}{Mg}, 0.1 \right)$.

Figure 3: Graphical model of an extended kidney stone example. In this case, the confounder, $KS$ is not observed.

Given the conditions of the causal graph, the back-door criterion cannot be used and we must rely on other methods in order to estimate the causal effect of $x$ on $y$. As a particular case, the method to identify the causal effect in this graph is called the front-door adjustment (Pearl, 2009): 

$$P(R \mid do(T = \hat{t})) = \sum_{mg} P(Mg = mg \mid T = \hat{t}) \sum_{t'} P(R \mid Mg = mg, T = t') P(T = t').$$  (7)

4.2.1 Auxiliary networks

As can be seen from Equation (7), the neural autoregressive model does not provide all the conditional distributions necessary for the estimation of the treatment effects. In this case, we have to estimate an “auxiliary network” which we can use to estimate the causal effects. For our particular causal graph, and the causal effect identified in 7, the neural network to be estimated takes $Mg$ and $T$ as input and predicts $R$. With that auxiliary network, it is possible to estimate the causal effects.

4.2.2 Results

Since the analytical solution of this causal graph is not readily available, the true causal effects was simulated from binned conditional distributions using the true generative process. We compared
the distributions using two values of $T = 0$ and $T = 0.5$ ($T$ is constrained between -1 and 1). Furthermore, the integrals that arise in Equation (7) are approximated using Monte Carlo sampling as explained in Section 4.1.3.

In Figure 4 we compare the estimated distribution of $R$ given three models and the true distribution. On the left, the linear model finds a single distribution for the different values of $T$, as expected. In the middle, our distribution matches to a high degree that of the simulated distribution. The discrepancies between the true distribution and the real distribution might arise from the different levels of stochasticity we are subject to: the sampling of the true effects, the stochastic optimization procedure and the Monte Carlo approximation of the integral. In the rightmost plot of the same figure, we see that pure conditioning on $T$ and integrating $M_g$ out from the conditional network $P(R \mid X = x, M_g = mg)$ gives results far off from the simulated results. We also included the Wasserstein Distance (WD) in every plot as a metric of comparison between the different models and the true distribution. The proposed approach performs better on both interventional distributions than the competing approaches.

Figure 4: Comparison of three different ways to estimate causal effects. (Left) Linear model. (Center) Our proposed approach. (Right) Simple conditioning. (Above) Comparing against the true $do(X = 0)$, and (Below) against the true $do(X = 0.5)$. The Wasserstein Distance (WD) between the distributions is specified on each plot.

5 Conclusions and future work

We have proposed the neural autoregressive density estimator to estimate the causal conditionals in causal graphs. The motivation to do so is to avoid having to make any assumptions on the way variables related to common children in the graph. We have shown in several simulation examples, with different specifications that the proposed method is able to recover both linear and non-linear relations, superseding any linear model. We have also shown the limitations of our method, by showing that the estimated causal effects degrade as we get far from the data support. Besides the applications to machine learning research, we believe these methods would also deem useful to social sciences where the causal graphs or certain systems have already been proposed and debated, but where the functional forms remain unknown.

An interesting extension of the proposed method is to find architectures that would also allow to compute counterfactuals in addition to the interventions proposed here. Given the assumptions that we made, this could (potentially) be reduced to a conditional density estimation problem.

Further extensions of this model have to do with the uncertainty quantification of the estimates as we would like to get full distributions of the parameters of the inferred distributions, instead of point estimates. Uncertainty estimation in Neural Networks or Bayesian Neural Networks has been
an active area of research which has taken off recently and could provide some approaches to this research direction.

Furthermore, other research directions can point to what can be done with the single networks of the composite neural network. This question is related to the concept of transfer and continuous learning in the machine learning literature, and transportability in causality literature (Pearl and Bareinboim 2014).

References

Agnelli, J., Cadeiras, M., Tabak, E., Turner, C. V., and Vanden-Eijnden, E. (2010). Clustering and classification through normalizing flows in feature space. Multiscale Modeling & Simulation, 8(5):1784–1802.

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. In Advances in Neural Information Processing Systems, pages 3424–3432.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. The American Economic Review, 80(3):313–336.

Angrist, J. D. and Pischke, J.-S. (2008). Mostly harmless econometrics: An empiricist’s companion. Princeton university press.

Bengio, Y. and Bengio, S. (1999). Modeling high-dimensional discrete data with multi-layer neural networks. In Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS’99, pages 400–406, Cambridge, MA, USA. MIT Press.

Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. In Advances in Neural Information Processing Systems, pages 3559–3569.

Bishop, C. M. (1994). Mixture density networks.

Charig, C. R., Webb, D. R., Payne, S. R., and Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed), 292(6524):879–882.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems, 2(4):303–314.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). Made: Masked autoencoder for distribution estimation. In International Conference on Machine Learning, pages 881–889.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. (2018). Learning functional causal models with generative neural networks. In The Springer Series on Challenges in Machine Learning, pages 39–80. Springer International Publishing.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1414–1423. JMLR. org.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240.

Imbens, G. W. and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In International conference on machine learning, pages 3020–3029.
Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 29–37.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, pages 159–217.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In Advances in Neural Information Processing Systems, pages 2338–2347.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., and Schölkopf, B. (2017). Learning independent causal mechanisms. *arXiv preprint arXiv:1712.00961*.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition.

Pearl, J. (2012). The do-calculus revisited. *arXiv preprint arXiv:1210.4852*.

Pearl, J. and Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, pages 579–595.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241.

Stoker, T. M. (2016). *Aggregation (Econometrics)*, pages 1–11. Palgrave Macmillan UK, London.

Tabak, E. G. and Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164.

Uria, B., Murray, I., and Larochelle, H. (2013). Rnade: The real-valued neural autoregressive density-estimator. In Advances in Neural Information Processing Systems, pages 2175–2183.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*. 