A Study on Causation and Causal Models

submitted in partial fulfillment of the requirements
for the course of MA-308 (Mini Project)
in
3rd Year (6th Semester) of

Master of Science

(Five Year Integrated Program)
in Mathematics
submitted by

Pansuriya Tarang Bharatbhai I20MA005
Shaikh Khalid Shammi I20MA008
Dharmik Patel I20MA020

under the supervision of
Dr. Raj Kamal Maurya
Assistant Professor

DEPARTMENT OF MATHEMATICS
SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY
SURAT-395007, GUJARAT, INDIA

May 2023

Department of Mathematics
Sardar Vallabhbhai National Institute of Technology
(An Institute of National Importance, NITSER Act 2007)
Surat-395007, Gujarat, India.

APPROVAL SHEET

Report entitled "**A study on Causation and Causal Models**" by Tarang Pansuriya, Khalid Shaikh, Dharmik Patel is approved for the completion of course MA-308 (Mini Project) for the degree of Master of Science in Mathematics.

Dr. Jayesh M. Dhodiya
(**Examiner-1**)

Dr. Sudeep Singh Sanga
(**Examiner-2**)

Dr. Raj Kamal Maurya
(**Supervisor**)

Date: 09/05/2023
Place: Surat

Department of Mathematics
Sardar Vallabhbhai National Institute of Technology
(An Institute of National Importance, NITSER Act 2007)
Surat-395007, Gujarat, India.

DECLARATION

We hereby declare that the report entitled “A study on Causation and Causal Models” is a genuine record of work carried out by us and no part of this report has been submitted to any University or Institution for the completion of any course.

Pansuriya Tarang Bharatbhai
Admission No.: I20MA005
Department of Mathematics
Sardar Vallabhbhai National Institute of Technology
Surat-395007

Shaikh Khalid Shammi
Admission No.: I20MA008
Department of Mathematics
Sardar Vallabhbhai National Institute of Technology
Surat-395007

Dharmik Patel
Admission No.: I20MA020
Department of Mathematics
Sardar Vallabhbhai National Institute of Technology
Surat-395007

Date: 09/05/2023
Place: Surat

Department of Mathematics
Sardar Vallabhbhai National Institute of Technology
(An Institute of National Importance, NITSER Act 2007)
Surat-395007, Gujarat, India.

CERTIFICATE

This is to certify that the course's report entitled "A study on Causation and Causal Models" submitted by Pansuriya Tarang Bharatbhai, Shaikh Khalid Shammi and Dharmik Patel in fulfilment for the completion of course of MA 308: Mini Project at Sardar Vallabhbhai National Institute of Technology, Surat is record of their work carried out under my supervision and guidance.

Dr. Raj Kamal Maurya
Assistant Professor
Department of Mathematics
Sardar Vallabhbhai National Institute of Technology
Surat-395007

Date: 09/05/2023
Place: Surat

Acknowledgment

As students of 5 Years Integrated M.Sc. in Mathematics at Sardar Vallabhbhai National Institute of Technology, Surat, we are very grateful. First, we would like to express our genuine appreciation to our supervisor Dr. Raj Kamal Maurya for his guidance for our work. To work with him is a great opportunity and a pleasure to us. We are thankful for Prof. Anupam Shukla, Director of SVNIT, Dr. Jayesh M. Dhodiya, Head of the Department of Mathematics, and all other Faculties, Research Scholars, and Non-Teaching staff of our department for their regular inspiration and co-activity.

(Pansuriya Tarang Bharatbhai)

(Shaikh Khalid Shammi)

(Dharmik Patel)

Preface

Causation is a concept that has intrigued scholars from various disciplines for centuries. It is fundamental to understanding how events and phenomena are related, and it plays a crucial role in predicting outcomes and designing effective interventions. With the recent advances in data science and artificial intelligence, there has been renewed interest in studying causation and causal models. This project aims to provide a comprehensive study of causation and causal models, including their definitions, types, and applications. It also delves into the challenges of inferring causation from observational data and explores the various methods that have been developed to address these challenges.

List of Figures

1.1 Directed Acyclic Graph 2
1.2 Ladder of Causation 4
1.3 ‘do’ operator 5

2.1 Intervention of Causal factors 7
2.2 Conditional probability for x_n 8
2.3 Encoder Decoder architecture 10
2.4 Generation of complex distribution 11
2.5 Base Distribution 12
2.6 Causal Relationship between variables 13
2.7 Output before intervention 14
2.8 Output after intervention 15

Contents

1 Introduction 1

1.1 Causation . . . 1

1.2 Correlation and Causation . . . 1

1.3 Preliminaries . . . 2

1.3.1 Directed Acyclic Graphs . . . 2

1.3.2 Dynamic Bayesian Networks . . . 2

1.3.3 Structural Equation Model . . . 3

1.3.4 Ladder of Causation . . . 3

1.3.5 Intervention . . . 4

1.4 Out of Distribution Data Generation . . . 5

2 Out of Distribution Data Generation 6

2.1 Causal Representation Learning . . . 6

2.2 Preliminaries . . . 6

2.3 Identification of Causal Variables . . . 6

2.3.1 Causal Factors . . . 6

2.3.2 Function for Observed Data . . . 6

2.3.3 Disentangling Causal Factors . . . 7

2.3.4 Minimal Causal Variables . . . 7

2.3.5 Learning Minimal Causal Variables . . . 7

2.4 Generative Models . . . 8

2.4.1 Autoregressive Model . . . 8

2.4.2 Variational Autoencoder (VAE) . . . 9

2.4.3 Normalizing Flows . . . 11

2.5 Identification from Temporal Sequences . . . 12

2.5.1 Variational Autoencoder . . . 12

2.5.2 Normalizing Flow . . . 12

2.6 Experiment . . . 13

2.6.1 Dataset . . . 13

2.6.2 Model . . . 14

Conclusion and Future Scope 16

Chapter 1

Introduction

1.1 Causation

Causation refers to the relationship between two variables where a change in one variable (the cause) results in a change in another variable (the effect). The study of causation is essential for understanding how events, actions, and outcomes are connected, and how we can manipulate these connections to achieve specific goals. In science, causality is a fundamental concept that allows us to explain and predict the behavior of natural phenomena. By identifying causal relationships, we can develop theories, test hypotheses, and design experiments to validate or refute our assumptions. Similarly, in philosophy, causation has been a central topic of inquiry for centuries, as it raises fundamental questions about the nature of reality, causation as a concept, and the limits of human knowledge.

Causal inference, on the other hand, is a statistical method used to determine the causal effect of a particular treatment or intervention. It is commonly used in fields such as medicine, economics, and social sciences, where it is often difficult or unethical to conduct randomized experiments. By applying various statistical techniques, researchers can draw causal inferences from observational data, which can help inform policy decisions, improve treatment outcomes, and advance our understanding of complex systems (refer [10]).

1.2 Correlation and Causation

Causation and correlation are two related but distinct concepts. While correlation measures the degree of association between two variables, causation involves identifying a direct relationship between the variables, with one variable causing a change in the other. Distinguishing between the two is essential because assuming causation based on correlation alone can lead to erroneous conclusions. Correlation merely suggests that the two variables are related in some way. In contrast, causation provides a more complete and accurate understanding of the relationship between two variables by identifying a direct mechanism or process that connects the cause to the effect (refer [1]).

Causation is generally considered a more robust and valuable concept for several reasons. Firstly, it provides a more precise and accurate understanding of the relationship between two variables, allowing researchers to make more precise predictions and develop more effective interventions or treatments. Secondly, it is more useful for identifying the underlying causes of complex phenomena, such as diseases, social problems, or economic trends. By identifying the causal factors, researchers and policymakers can develop strategies to address the root causes of the problem, rather than simply treating the symptoms. Thirdly, it is essential for making informed decisions in areas such as medicine, public health, and policy, where the consequences of decisions can be significant. Finally, causal inference provides a powerful tool for research in situations where randomized experiments are not feasible or ethical.

In conclusion, while correlation is a useful tool for identifying relationships between variables, causation is generally considered a more powerful and valuable tool for understanding complex phenomena and making informed decisions. By identifying the direct causal mechanisms underlying relationships between variables, researchers and decision-makers can make more accurate predictions and develop more effective strategies to achieve their goals.

1.3 Preliminaries

1.3.1 Directed Acyclic Graphs

A directed acyclic graph (DAG) is a visual representation of the causal relationships between a set of variables in a system. DAGs are useful tools for studying causation because they can help identify which variables are causes and which are effects. In a DAG, each variable is represented by a node, while the causal relationships between variables are represented by directed arrows.

The direction of the arrows is important because it shows the direction of causation. If an arrow points from variable X to variable Y, then X is a cause of Y (as shown in Figure 1.1). DAGs are acyclic, meaning they don't have any loops or cycles because a cycle would imply that a variable causes itself, which is impossible.

DAGs can also be used to identify confounding variables, which are variables that are related to both the cause and the effect, but are not part of the causal pathway. By identifying and controlling for confounding variables, researchers can more accurately assess the causal relationship between the variables of interest.

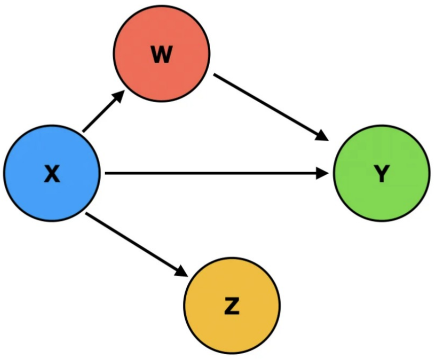

Figure 1.1: Directed Acyclic Graph

1.3.2 Dynamic Bayesian Networks

A Dynamic Bayesian Network (DBN)([5]) is a graphical model with the probability that represents a sequence of variables over time, where the variables can influence each other in a causal manner. DBNs are widely used in various fields, such as finance, bioinformatics, and robotics, to model dynamic processes. For example, consider a stock market prediction system that uses a DBN. The system could model the stock prices of different companies as a set of variables, with each variable representing the price at a specific time. The DBN would capture the causal relationships between the stock prices of different companies over time, considering factors such as company earnings, market trends, and investor sentiment. The DBN could be used to predict future stock prices by using the current stock prices and other relevant information as inputs. By updating the DBN over time with new data, the system could adapt to changes in the market and improve its predictions. In a DBN, the conditional probabilities between variables can change over time. To capture this, DBNs often use time-varying parameters, such as transition probabilities, that describe how the variables change over time. These parameters can be learned from data using techniques such as the Expectation-Maximization algorithm.

1.3.3 Structural Equation Model

Structural Equation Modeling(SEM) is a powerful statistical technique that allows researchers to analyze complex relationships among variables and gain a deeper understanding of the underlying mechanisms of social and psychological phenomena. It allows researchers to specify a theoretical model of relationships among variables and then estimate the parameters of that model using observed data. SEM consists of two main components: the measurement and structural models. The measurement model shows the relationships between observed variables and their corresponding latent constructs. In contrast, the structural model specifies the relationships among the latent constructs and any observed variables that are not part of the measurement model.

SEM is particularly useful for handling complex models with multiple latent constructs and observed variables. It also enables researchers to test the fit of the model to the data using various fit indices such as the Chi-square test and Tucker-Lewis Index (TLI).

For instance, we may investigate the relationship between socioeconomic status, stress levels, and health outcomes. Using SEM, we can specify a model in which socioeconomic status is a latent construct, stress levels are an observed variable, and health outcomes are another latent construct. We can then estimate the path coefficients and assess the model's goodness of fit to the data.By utilizing SEM in our research project, we can test our hypotheses and derive valuable insights into the relationships among the variables of interest.

1.3.4 Ladder of Causation

Pearl's ladder of causation is a framework that helps researchers understand the different levels of causation that are relevant to causal inference. There are three main levels in this framework: association, intervention, and counterfactual (as shown in Figure 1.2 from [2]). At the base of the ladder is the level of association, which involves identifying statistical relationships between variables. This level focuses on describing how variables are related to each other, but it does not provide information about causation. It is important to note that identifying associations between variables is not enough to infer causation, as correlation does not imply causation. Nonetheless, associations between variables can guide further investigation.

Moving up the ladder, we encounter the level of intervention. This level involves manipulating a variable to observe its effects on the outcome of interest. The goal of an intervention is to establish a cause-and-effect relationship between the variable being manipulated and the outcome of interest. Randomized controlled trials are a common method of intervention that can help establish causal relationships between variables.

At the top of the ladder is the level of counterfactuals, which involves comparing what happened in reality to what would have happened under a different intervention. This allows researchers to estimate the causal effect of an intervention by comparing the outcome in the actual system to the outcome that would have occurred under a different scenario. Counterfactuals are often used when

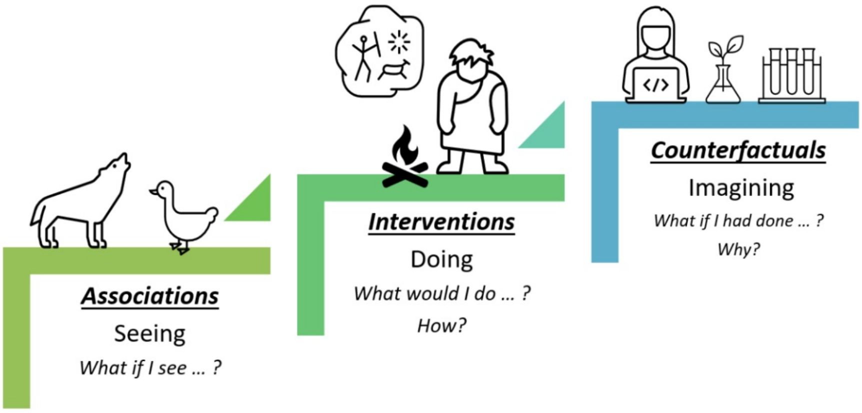

Figure 1.2: The image represents Pearl's ladder of Causation and all of it's three level with conceptual understanding

interventions are not feasible or ethical. However, counterfactual analysis relies on assumptions about the system and may not always accurately capture the complex interactions between variables.

1.3.5 Intervention

In this project, we have extensively used the 'Interventions' from the ladder of causation. Intervention can be understood by the A/B test, where we divide the samples into the Test and Controlled groups. From a medical perspective, we will not use any new medication or treatment with the Control group, whereas, for the test group, we will use new medicine or treatment. We compare the results and see whether we significantly improved in the test group. Intervention is conceptually the same thing.
When researchers intervene on a variable in a model, they change the system and fix the value of the variable being manipulated. This allows them to observe the effects of the intervention on the system and determine whether there is a causal relationship between the manipulated variable and the outcome of interest. However, as mentioned earlier, interventions can also have spillover effects on other variables in the system, which need to be carefully considered and accounted for to assess the impact of the intervention accurately.

Mathematically, there is an operator called 'do operator'(Figure 1.3)(refer [**?**]), which is used for the intervention in field causation. The do operator is a notation used in causal modeling to represent the effect of an intervention or manipulation of a variable in a system. It allows researchers to simulate the impact of a hypothetical intervention on a variable of interest and make predictions about the resulting changes in the other variables in the system.
When we intervene on a variable in a causal model, we fix its value to a specific value and observe the resulting changes in the other variables. Using the do operator, we can specify the variable being intervened upon and the value it is being set to. This allows us to make counterfactual statements about what would happen if a variable were to be intervened upon in a particular way and to study the causal effects of interventions in a system.
Figure 1.3 shows the application of 1^{st} rule of do-calculus. Thus, $P(y|do(x_0), z) = P(y|do(x_0))$ since $(Y \perp Z|X)_{G_{\overline{X}}}$, where $G_{\overline{X}}$ is the graph with all edges directed towards X removed.

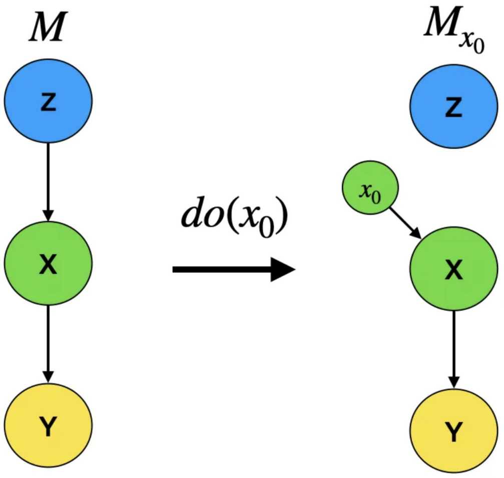

Figure 1.3: Representation of do-operator

1.4 Out of Distribution Data Generation

In a causal model, we aim to identify the causal relationships between different variables in a system. To achieve this, we typically train the model on a set of data that is representative of the underlying distribution of the variables in the system.
However, in real-world scenarios, we may encounter situations where the data available to us is different from the original training distribution. This is where out-of-distribution data generation can be useful. Out-of-distribution data generation involves creating new, diverse data points that fall outside of the original training distribution. This can be achieved using a variety of techniques.
By generating out-of-distribution data, we can test the causal model on new and diverse scenarios, which can help us to better understand the generalisability and robustness of the model. This can also help us to identify previously unknown causal relationships that were not evident in the original training data (refer [9]).
Moreover, out-of-distribution data generation can help us better prepare for real-world scenarios where we may encounter data that falls outside the original training distribution. By testing the causal model on out-of-distribution data during development, we can better understand its limitations and ensure that it can perform well even in unforeseen situations. It makes the model more robust and realistic since we have trained it for unforeseen circumstances.

Chapter 2

Out of Distribution Data Generation

2.1 Causal Representation Learning

The field of causal representation learning is concerned with creating representations of causal factors from high-dimensional observations. In simpler settings, causal factors are often assumed to be scalar values. However in complex setting, estimating every scalar causal variable becomes impractical. For example, in images, it is more natural to consider factors like position and rotation of objects as multidimensional variables instead of scalar values, refer this [6].

2.2 Preliminaries

We consider our model to be a Dynamic Bayesian Network (G) (Refer [3]). There are K causal factors involved in G. Thus, $G = (V, E)$, where V is the set of nodes related to a causal variable and each edge in E represents a causal relation.

The causal factor C_i is instantiated at each step and is represented as C_i^t. C_i^t depends only on its causal parents at time $t-1$, denoted as C_j^{t-1} and the variable itself at the previous time step i.e., C_i^{t-1}. Thus, the model is $C_i^t = f_i(pa_G(C_i^t), \epsilon_i)$, where $i \in [[1..K]]$, $t \in [[0..T]]$ and $pa_G(C_i^t) \subseteq \{C_1^{t-1}, ..., C_K^{t-1}\}$.

The binary variable $I^t \in \{0,1\}^K$ indicates that C_i^t is intervened upon if $I_i^t = 1$. We assume that there exists a confounding variable R^t, called the regime variable (refer [4]) which allows I_i^t to be affected by I_j^t, $i \neq j$ as shown in Figure 2.1. The new graph formed $G' = (V', E')$ is called the augmented DAG where, $V' = \{\{C_i^t\}_{i=1}^K \cup \{I_i^t\}_{i=1}^K \cup R^t\}_{t=1}^T$ and $E = \{\{pa_G(C_i^t) \rightarrow C_i^t\}_{i=1}^K \cup \{I_i^t \rightarrow C_i^t\}_{i=1}^K \cup \{R^t \rightarrow I_i^t\}_{i=1}^K\}_{t=1}^T$.

2.3 Identification of Causal Variables

2.3.1 Causal Factors

A causal factor C_i is a M_i dimensional variable. Thus, $C_i \in \mathcal{D}_i^{M_i}$, where $\mathcal{D}$ is $\mathbb{R}$ for continuous values or $\mathbb{Z}$ for discrete values. As a result the causal factor space is $\mathcal{C} = \mathcal{D}_1^{M_1} \times \mathcal{D}_2^{M_2} \times ... \times \mathcal{D}_K^{M_K}$ (Refer [6], [8]).

2.3.2 Function for Observed Data

The observations X^t are given by the functional relation $X^t = h(C_1^t, C_2^t, ..., C_K^t, E_0^t)$, where E_0^t is noise separate from the causal factors affecting the observed data. The functional relation is defined as

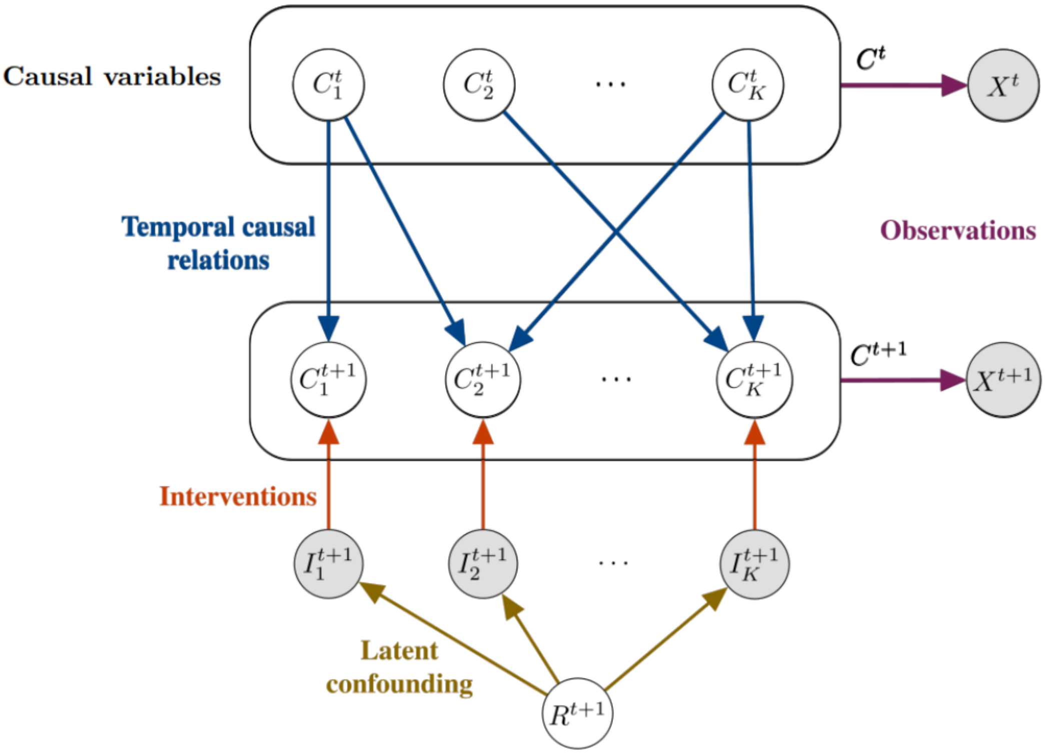

Figure 2.1: Causal factor C_i^{t+1} is affected by interventions at $t+1$ and causal factors $C_j^t,\ j \in [[1..K]]$ at time t

$h : \mathcal{C} \times \mathcal{E} \to \mathcal{X}$. We assume that h is bijective. This allows us to find causal factors uniquely from observed data using inverse mapping f.

2.3.3 Disentangling Causal Factors

In order to uniquely identify causal factors given the observed data and interventions it is necessary that the there do not exist causal factors C_i and C_j such that C_i and C_j are always jointly intervened on or not intervened on at all.

2.3.4 Minimal Causal Variables

Let for each causal factor $C_i \in \mathcal{D}^{M_i}$, $i \in [[1..K]]$, there exists an invertible function s_i such that $s_i : \mathcal{D}_i^{M_i} \to \mathcal{D}_i^{var} \times \mathcal{D}_i^{inv}$. Thus, s_i splits the variable C_i into two parts, one which is variant under interventions and one which is invariant.

$$s_i(C_i^t) = (s_i^{var}(C_i^t),\ s_i^{inv}(C_i^t)) \tag{2.1}$$

Invertibility of s_i allows back and forth from $\mathcal{D}^{M_i}$ to $\mathcal{D}_i^{var} \times \mathcal{D}_i^{inv}$. Splitting the causal variable C_i into variable and invariant parts allows us to concentrate on the variable part.

2.3.5 Learning Minimal Causal Variables

The dataset $\mathcal{D}$ is a tuple of $\{x^t, x^{t+1}, I^{t+1}\}$, where $x^t, x^{t+1} \in \mathbb{R}^N$. They are observations at time $t = t,\ t = t+1$ and interventions at time $t+1$ applied on causal factors at time $t+1$. The causal factors have dimensions $M_1, M_2, ..., M_K$. To represent them properly we construct a latent space $\mathcal{Z}$

with dimension $M \geq \dim(\mathcal{E}_i) + \sum_{i=1}^{K} M_i$ i.e., $\mathcal{Z} \subseteq \mathbb{R}^M$.

We define functions $g_\theta : \mathcal{X} \rightarrow \mathcal{Z}$ which maps the observations to the latent space and $\psi : [[1..M]] \rightarrow [[0..K]]$ which maps the latent space dimensions to causal factors. We define $\psi(x_j) = 0$, to represent that the dimension z_j does not map to any causal variable. This is useful for representing $s_i^{inv}(C_i^t)$ and noise $\mathcal{E}_0^t$.

Let $p_\phi(z^{t+1}|z^t, I^{t+1})$ be the prior distribution, where, $z^t = g_\theta(x^t)$ and $z^{t+1} = g_\theta(x^{t+1})$, $z^t, z^{t+1} \in \mathcal{Z}$. We have,

$$p_\phi(z^{t+1}|z^t, I^t) = \prod_{i=0}^{K} p_\phi(z_{\psi_i}^{t+1}|z^t, I_i^{t+1}) \tag{2.2}$$

Thus, we have maximize likelihood,

$$p_{\phi,\theta}(x^{t+1}|x^t, I^{t+1}) = \left| \frac{\partial g_\theta(x^{t+1})}{\partial x^{t+1}} \right| p_\phi(z^{t+1}|z^t, I^{t+1}) \tag{2.3}$$

2.4 Generative Models

In this section we provide a brief overview of the generative processes used by us. This will serve as an overview of the inner workings of these models which is essential in understanding the functioning of the models.

2.4.1 Autoregressive Model

Representation

These are models which use observations from earlier time steps to predict outcome at current time step. Assume that our dataset $\mathcal{D}$ has n-dimensional datapoints. For the sake of simplicity we shall assume that $\mathbf{x} = \{0,1\}^n$.

Using chain rule,

$$p(\mathbf{x}) = \prod_{i=1}^{n} p\left(x_i|x_1, x_2, ..., x_{n-1}\right) = \prod_{i=1}^{n} (x_i|\mathbf{x}_{<i}),$$

where $\mathbf{x}_{<i} = \{x_1, x_2, ..., x_{n-1}\}$ represents vector of random variables with index less than i.

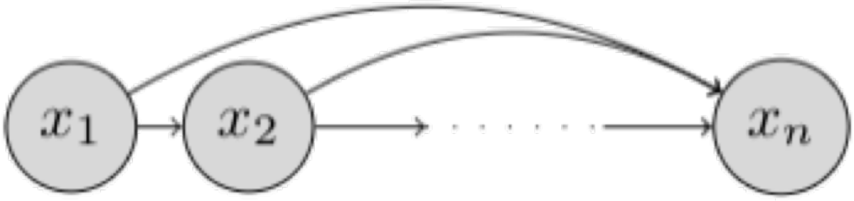

Figure 2.2: Conditional probability for x_n

This kind of Bayesian networks which make no assumptions in conditional independence are known as Autoregressive models. Here an order is fixed for these random variables $x_1, x_2, ..., x_n$ so that x_n

depends on all the previous random variables $x_1, x_2, ..., x_{n-1}$ as shown in Figure 2.2.

It can be seen that the computation and storage of all these conditional probabilities is very expensive. As a result we assume that the conditionals arrive from a Bernoulli distribution which depends on a function of all the previous variables. Thus,

$$p_{\theta_i}(x_i|\mathbf{x}_{<i}) = Br\left(f_i(x_1, x_2, ..., x_{i-1})\right),$$

where θ_i denotes parameters to define mean and function f_i maps $x_1, x_2, ..., x_{i-1}$ to the mean of the distribution.

Learning

Generative models use KL divergence to determine the similarity between the true distribution and the distribution developed by our model. Thus,

$$\underset{\theta \in \mathcal{M}}{min}\ d_{KL}(p_{data}, p_\theta) = \mathbb{E}_{\mathbf{x} \sim p_{data}}\left[log\ p_{data}(x) - log\ p_\theta(x)\right]$$

To approximate the expectation over the unknown p_{data}, we assume that points in the dataset $\mathcal{D}$ are sampled i.i.d. from p_{data}. This gives the objective function as,

$$\mathcal{L}(\theta|\mathcal{D}) = \underset{\theta \in \mathcal{M}}{max} \frac{1}{|D|} \sum_{\mathbf{x} \in \mathcal{D}} log\ p_\theta(x)$$

2.4.2 Variational Autoencoder (VAE)

Dimensionality reduction

Dimensionality reduction is a machine learning technique that involves reducing the number of features used to describe data. This can be achieved by either selecting a subset of existing features or by creating a new set of features based on the original ones. Dimensionality reduction can be useful in various applications such as data visualization, storage, and computation. Although there are many different methods for dimensionality reduction, most of them follow a similar framework.

Some methods to perform dimensionality reduction:

- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Random Forests
- Linear Discriminant Analysis (LDA)

The process of creating new features is referred to as the encoder, while the reverse process is called the decoder. In this framework, dimensionality reduction is seen as a type of data compression, where the encoder compresses the data from the original space to a new space (also known as the latent space), and the decoder decompresses the data. However, depending on the initial data distribution, the latent space dimension, and the encoder's definition, this compression can be lossy, meaning that some information is lost during the encoding process and cannot be recovered during decoding.

Autoencoders

Autoencoders are neural networks used for dimensionality reduction by setting up an encoder and a decoder. These networks are trained using an iterative optimization process where the encoder compresses the data and the decoder decompresses it. The overall architecture creates a bottleneck that allows only the main structured part of the information to pass through and be reconstructed. The encoder and decoder are defined by their network architectures, and the parameters of these networks are optimized through gradient descent to minimize reconstruction error.

If the encoder and decoder are linear and have only one layer, they are similar to PCA in that they search for the best linear subspace to project the data onto with minimal information loss. However, there can be multiple encoder/decoder pairs that provide the optimal reconstruction error, and the new features do not have to be independent.

If the encoder and decoder are deep and non-linear, the autoencoder can perform high dimensionality reduction while keeping reconstruction loss low. However, this often results in a lack of interpretable and exploitable structures in the latent space, and careful control of the dimension of the latent space and the depth of the autoencoder is necessary depending on the purpose of the dimensionality reduction.

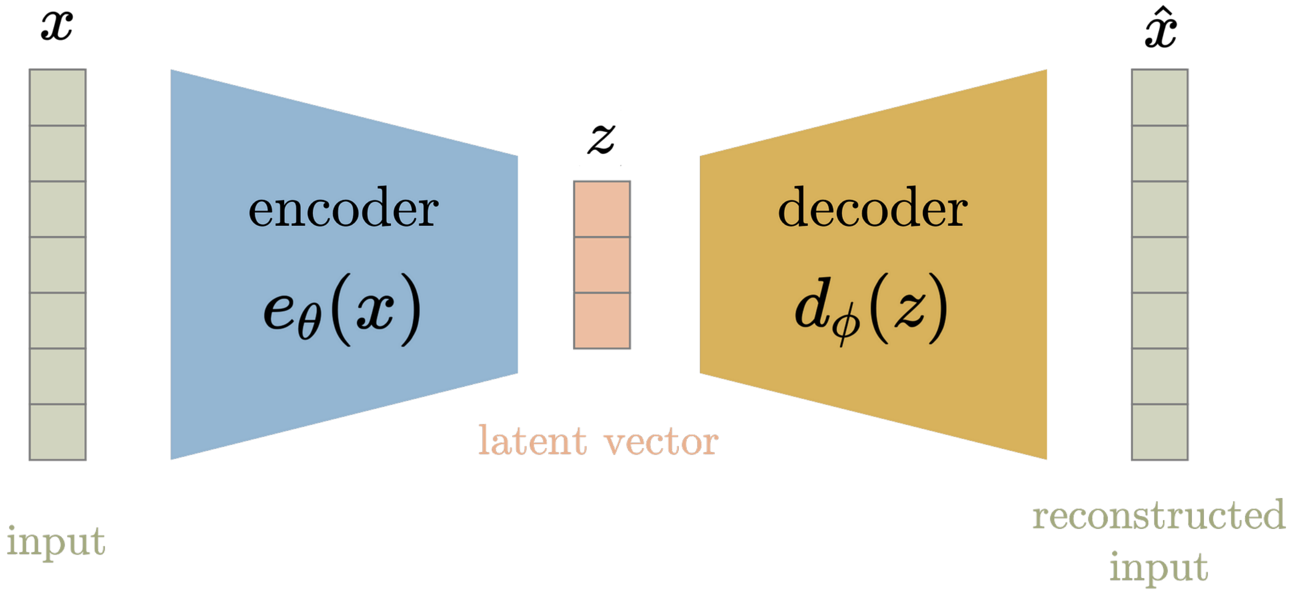

Figure 2.3: Encoder Decoder architecture

Variational Autoencoder

Variational autoencoders were developed to address the issue of generating novel content using autoencoder decoders. They are a type of autoencoder that incorporates explicit regularization into the training process to create a consistent latent space. Unlike traditional autoencoders, which encode input as a single point, variational autoencoders encode input as a distribution over the latent space. These distributions are typically normal, and the encoder is trained to provide the mean and covariance matrix that define these Gaussian distributions. The variance control in the encoding process enables local regularity of the latent space, while the mean control guarantees global regularity. A variational autoencoder's loss function comprises a reconstruction term and a regularization term. The reconstruction term aims to optimize the encoding-decoding process's performance, while the regularization term ensures that the latent space organization is regularized by making the distributions returned by the encoder close to a standard normal distribution. The regularization term is expressed as the Kulback-Leibler divergence between the returned distribution and a standard Gaussian.

2.4.3 Normalizing Flows

Autoregressive models can manage likelihoods more easily, but do not offer a straightforward method for feature learning. Conversely, variational autoencoders can learn feature representations, but their ability to calculate marginal likelihoods is limited. To overcome these drawbacks, this section presents normalizing flows, a technique that blends the best features of both approaches. Normalizing flows provide the ability to learn features and calculate manageable marginal likelihoods.

'Normalizing Flow', can be understood as:

- 'Normalizing' referring to the property that the probability density function becomes normalized after an invertible transformation is applied to it.
- 'Flow' implies that these invertible transformations can be merged to generate more intricate transformations.

Assume our functions form the iteration,

$$\begin{aligned} z_1 &= f_1(z_0) = mz_0 + b \\ z_2 &= f_2(z_1) = e^{z_1} + d = e^{mz_0+b} + d \\ &\vdots \end{aligned}$$

This process represented in Figure 2.4 displays how a simple distribution can be transformed into a complex distribution.

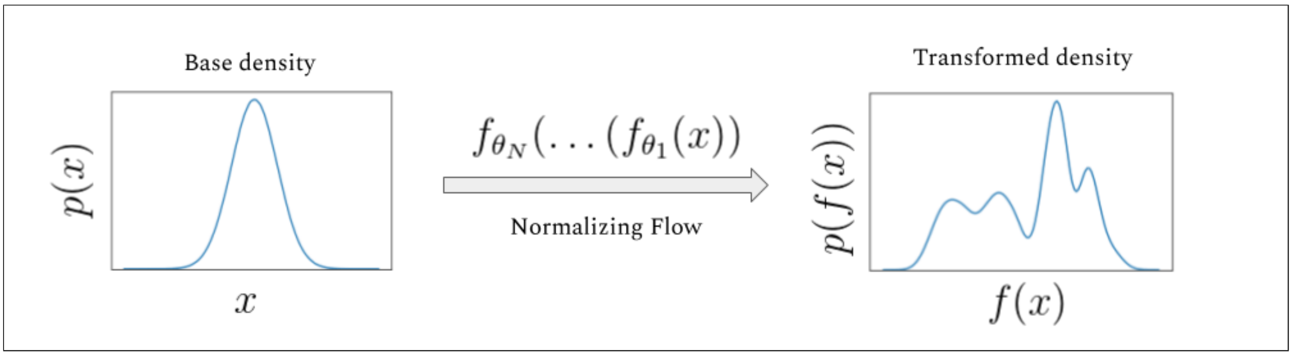

Figure 2.4: Generating complex distributions by using back-to-back change of variable functions

Change of Variables

Change of variables between random variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ is given by mapping $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $\boldsymbol{X} = f(\boldsymbol{Z})$ and $\boldsymbol{Z} = f^{-1}(\boldsymbol{X})$. Thus,

$$p_X(x) = p_Z(f^{-1}(x)) \left| \frac{\partial f^{-1}(x)}{\partial x} \right|$$

- $\boldsymbol{x}$ and $\boldsymbol{z}$ are continuous and $\dim(\boldsymbol{x}) = \dim(\boldsymbol{z})$

- Since $\boldsymbol{Z} = f^{-1}(\boldsymbol{X})$ we have,

$$p_X(x) = p_Z(z)\left|\frac{\partial f(z)}{\partial z}\right|^{-1}$$

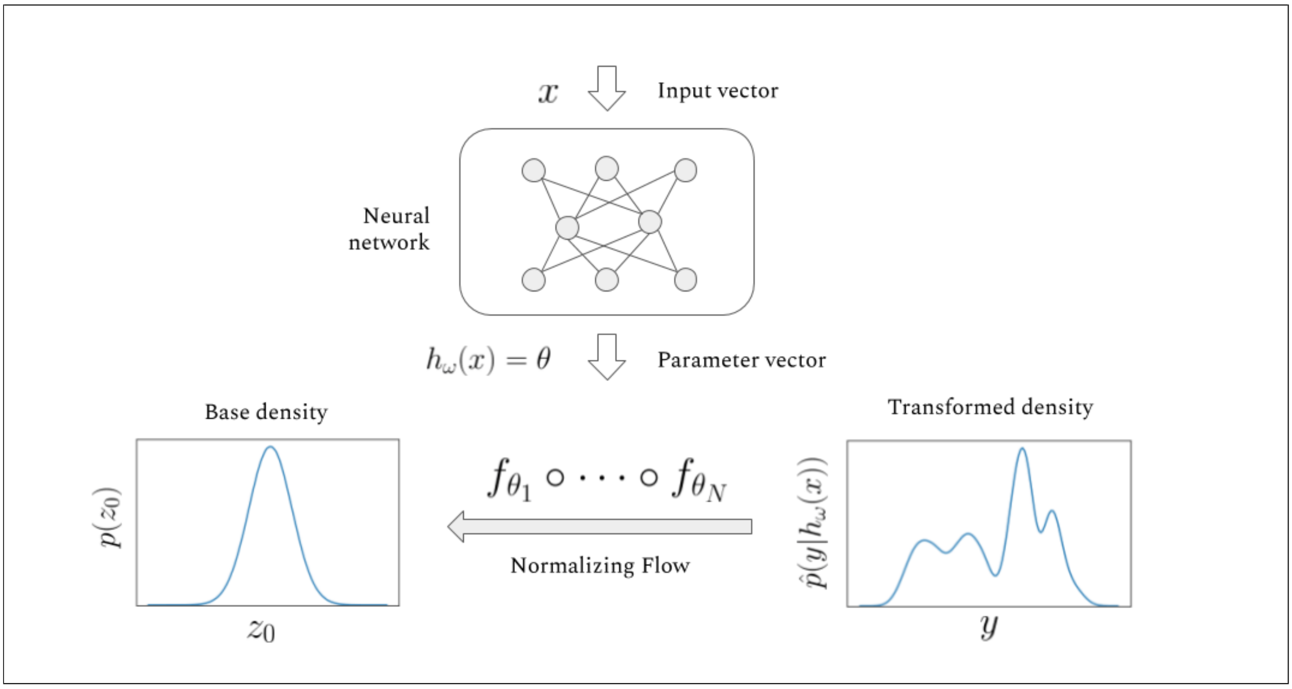

Figure 2.5: Using invertibility of change of variable functions for calculating base distribution

2.5 Identification from Temporal Sequences

2.5.1 Variational Autoencoder

We learn a variational autoencoder (Refer [7]) for our model. We design functions q_θ and f_θ as encoder and decoder which are used to approximate g_θ. The Evidence Lower Bound (ELBO) is given as,

$$\mathcal{L}_{ELBO} = -\mathbb{E}_{z^{t+1}}\left[\log p_\theta(x^{t+1}|z^{t+1})\right] + \mathbb{E}_{z^t,\psi}\left[\sum_{i=0}^{K} D_{KL}\left(q_\theta(z_{\psi_i}^{t+1}|x^{t+1}) \parallel p_\phi(z_{\psi_i}^{t+1}|z^t, I_i^{t+1})\right)\right] \tag{2.4}$$

In this approach, the invertible map g_θ is implemented by a deep network architecture that is trained to encode and decode high-dimensional observations to low-dimensional feature vectors, without any disentanglement. During training, we introduce small Gaussian noise to the latent variables to prevent the latent distribution from collapsing to single delta peaks. We do not impose any specific prior, which allows for complex marginal distributions.

2.5.2 Normalizing Flow

Once the autoencoder has finished training, its parameters are fixed, and a normalizing flow is used to transform the entangled latent space into a disentangled one. The reversible nature of the flow ensures that no information is lost, and the pretrained decoder can be used to reconstruct observations without any further fine-tuning. This method involves replacing the VAE encoder with a frozen encoder and a normalizing flow applied to the encoded latents, along with using the transition prior structure and target classifier. This approach is especially useful for modeling complex, high-dimensional images with intricate details.

2.6 Experiment

2.6.1 Dataset

We use the Temporal Causal3DIdent dataset. The dataset is a three dimensional rendering of objects. The causal factors included are object position, object rotation, spotlight rotation, object hue, spotlight hue, background hue and object shape.

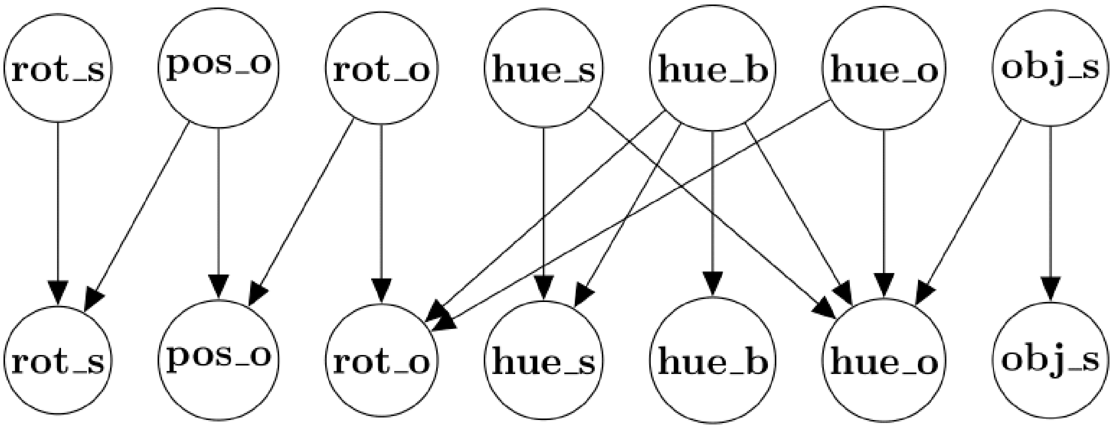

Figure 2.6: Causal relationship between variables from previous time step (t) to current time step ($t + 1$)

The Figure 2.6 shows how factors are affected by the previous time steps like the factor spotlight rotation from time stamp $t + 1$ is dependent on the factors spotlight rotation and object position from time stamp t.

Variable	Representation	Value
Object Position	pos_o	$[x, y, z] \in [-2, 2]^3$
Object Rotation	rot_o	$[\alpha, \beta] \in [0, 2\pi)^2$
Spotlight Rotation	rot_s	$\theta \in [0, 2\pi)$
Object Hue	hue_o	$h_{obj} \in [0, 2\pi)$
Spotlight Hue	hue_s	$h_{light} \in [0, 2\pi)$
Background Hue	hue_b	$h_{bg} \in [0, 2\pi)$
Object Shape	obj_s	$s \in \{teapot,\ cow,\ head,\ horse,\ armadillo,\ dragon,\ hare\}$

Table 2.1: Causal Factors and their values

The continuous variables follow Gaussian distribution. Intervention variables follow Bernoulli distribution.

The dataset contains of combinations of these factors which gives rise to images with different background colours, shapes and orientations of those shapes. Figure 2.7 represent the dataset over 20 time steps with varying features resulting in unique combinations.

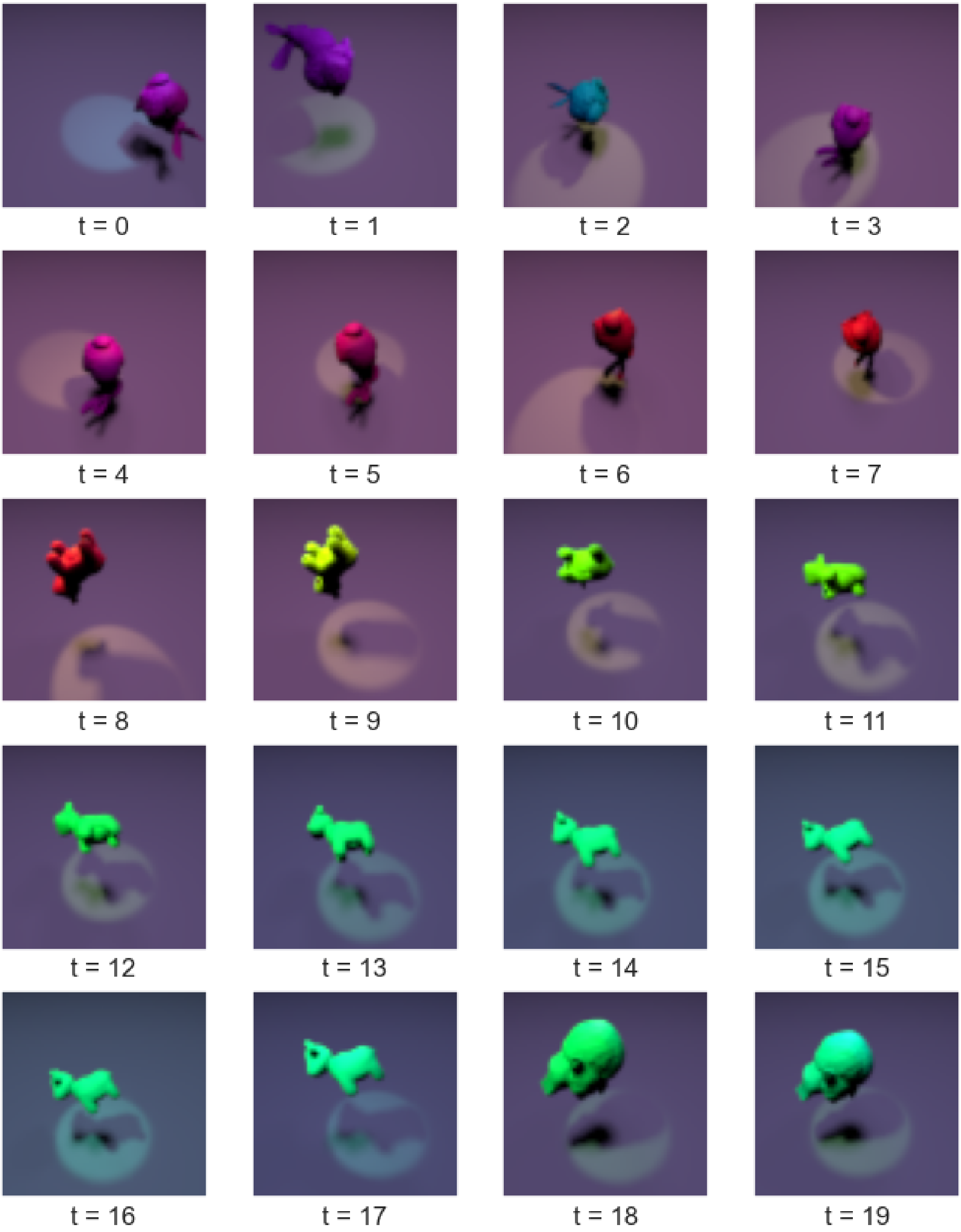

Figure 2.7: Different objects with different hues, background hues, rotation angles, spotlight rotation angles at various time steps

2.6.2 Model

The autoencoder is trained on the dataset to learn the latent representations. The learned representations are classified into the causal factors using a classification based architecture. The normalizing flows then sample from the latent representations and disentangle the causal factors. Using intervention we can manipulate the causal factors to generate out of distribution data.

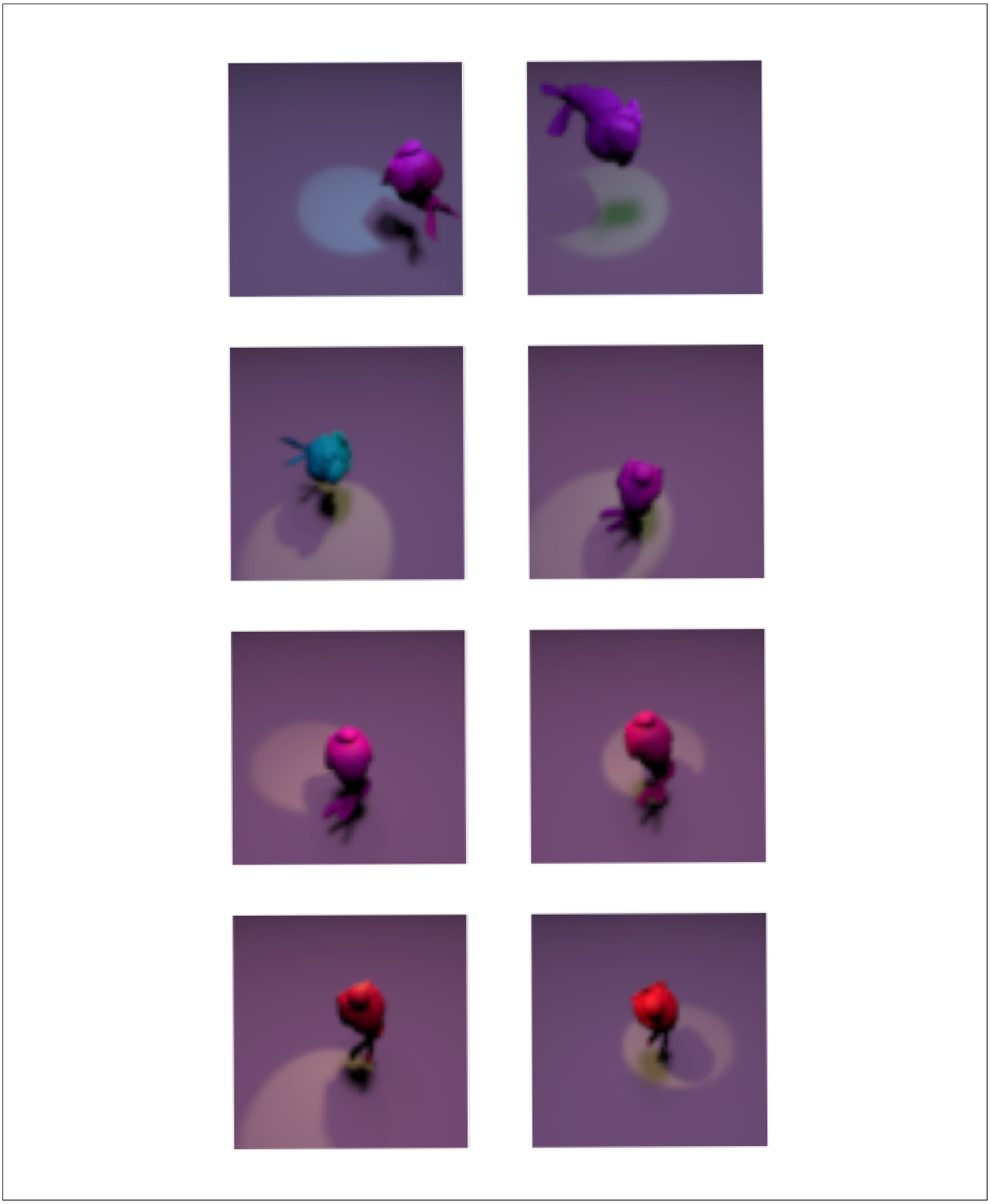

Figure 2.8: Images of hare with different hues, rotation angles and spotlight angles

Intervening on the factors leads to generation of new data points. Figure 2.8 displays a hare which has been intervened on its colour, spotlight rotation and orientation.

Conclusion and Future Scope

In conclusion, we have presented a novel approach to identify causal variables in dynamic systems by disentangling them into variable and invariant parts. We proposed a learning framework that maps observed data to a latent space. Our approach is able to handle interventions and confounding variables by augmenting the causal model with a regime variable. However, there are still some limitations and future work to be done. One limitation of our approach is that it assumes that the functional relation between causal variables and observed data is bijective, which may not always be the case in practice. Additionally, our approach requires knowledge of the dimensionality of each causal variable, which may not be known beforehand. Future work includes exploring more flexible models that can handle non-bijective functional relations and developing methods to automatically determine the dimensionality of each causal variable. Another direction for future work is to investigate the performance of our approach on real-world datasets and compare it with other causal inference methods. Overall, our work contributes to the growing field of causal inference and provides a new perspective on identifying causal variables in dynamic systems.

Bibliography

[1] Peter M Aronow and Fredrik Sävje. The book of why: The new science of cause and effect: Judea pearl and dana mackenzie. new york: Basic books, 2018, isbn: 978-0-46-509760-9., 2020.

[2] Alycia N Carey and Xintao Wu. The fairness field guide: Perspectives from social and formal sciences. *arXiv preprint arXiv:2201.05216*, 2022.

[3] Thomas Dean and Keiji Kanazawa. A model for reasoning about persistence and causation. *Computational intelligence*, 5(2):142–150, 1989.

[4] Vanessa Didelez, Philip Dawid, and Sara Geneletti. Direct and indirect effects of sequential treatments. *arXiv preprint arXiv:1206.6840*, 2012.

[5] Dobroslawa M Grzymala-Busse and Jerzy W Grzymala-Busse. The usefulness of a machine learning approach to knowledge acquisition. *Computational Intelligence*, 11(2):268–279, 1995.

[6] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[8] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

[9] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[10] Judea Pearl. *Bibliography*, page 429–453. Cambridge University Press, 2 edition, 2009. `doi:10.1017/CBO9780511803161.015`.

`https://towardsdatascience.com/causal-inference-962ae97cefda`