Carl von Ossietzky Universität Oldenburg
Fakultät II Informatik, Wirtschafts- und Rechtswissenschaften
Department für Informatik

CAUSAL DYNAMIC MODELING WITH BAYESIAN METHODS

---

von der Fakultät fur Informatik, Wirtschafts- und
Rechtswissenschaften der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

angenommene Dissertation von

Juliane Weilbach

geboren am 23.04.1993 in Mannheim

Gutachter	Prof. Dr. Martin Fränzle
	Assoc. Prof. Dr. Melih Kandemir
Betreuer	Dr. Sebastian Gerwinn
Tag der Einreichung	06. September 2024
Tag der Verteidigung	13. Dezember 2024

Declaration

I hereby declare that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare that my thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date	Signature

Abstract

Finding the root causes of faulty processes is an integral part of our lives. For this purpose, counterfactual formulations from the field of causal inference can be utilized to answer the following question: "Would the observed failure have occurred if a different system behaviour had been present?". Addressing such counterfactual questions provides insights into potential system improvements by estimating the effects of hypothetical interventions on prior observation of system behaviour.

We aim to address this issue within an industrial setting for practical applications. First, the hypothetical nature of counterfactual distributions makes them inherently ambiguous, which is particularly challenging in continuous settings. If the underlying system is dynamic, as it is typically the case in industrial processes like production plants, this increases complexity, particularly with regard to scalability. Furthermore, the process may contain unobserved variables that influence it, which may significantly impact the outcomes of counterfactual analyses. In summary, in an industrial context, several key challenges are focused on in this work: uncertainties while estimating the counterfactual distribution, counterfactual reasoning for identifying the root cause within a dynamic system, and the impact of latent confounding variables. We address these challenges within the realm of causal inference, taking advantage of the powerful properties of causal models, such as performing hypothetical interventions in a system.

To address the challenge of uncertainties in counterfactual distribution estimation, we propose a Bayesian method that accounts for the ambiguities involved, allowing for more reliable decision-making. In order to automatically identify root causes in a dynamic system, we develop an auto-regressive model with sampling-based estimation of dynamic processes. Compared to most non-causal approaches, this allows the root cause of the failure to be identified using only a single faulty observation by leveraging the counterfactual properties. Finally, we sketch ideas for potential extensions of deconfounding approaches to the identification of root causes when unobserved variables influence the causal relationships in a dynamic system.

Together, these methods address key challenges in practical applications, enhancing the robustness of counterfactual reasoning while enabling faster and more accurate root cause identification, ultimately simplifying quality control processes.

Zusammenfassung

Die Suche nach den Ursachen fehlerhafter Prozesse ist ein wesentlicher Bestandteil unseres Lebens. Zu diesem Zweck können kontrafaktische Formulierungen aus dem Bereich der kausalen Inferenz verwendet werden, um die folgende Frage zu beantworten: „Wäre der beobachtete Fehler aufgetreten, wenn ein anderes Systemverhalten vorgelegen hätte?“. Die Beantwortung solcher kontrafaktischen Fragen liefert Erkenntnisse über mögliche Systemverbesserungen, indem die Auswirkungen hypothetischer Eingriffe auf ein zuvor beobachtetes Systemverhalten abgeschätzt werden.

Unser Ziel ist es, diese Frage in einem industriellen Umfeld für praktische Anwendungen zu beantworten. Zunächst sind kontrafaktische Verteilungen aufgrund ihres hypothetischen Charakters von Natur aus mehrdeutig, was in kontinuierlichen Umgebungen eine besondere Herausforderung darstellt. Wenn das zugrunde liegende System dynamisch ist, wie es typischerweise bei industriellen Prozessen wie Produktionsanlagen der Fall ist, erhöht dies die Komplexität, insbesondere im Hinblick auf die Skalierbarkeit. Darüber hinaus kann der Prozess unbeobachtete Variablen enthalten, die ihn beeinflussen, was die Ergebnisse der kontrafaktischen Analysen erheblich beeinflussen kann. Zusammenfassend lässt sich sagen, dass sich diese Arbeit im industriellen Kontext auf mehrere zentrale Herausforderungen konzentriert: Unsicherheiten bei der Schätzung der kontrafaktischen Verteilung, kontrafaktische Argumentation zur Identifizierung der Grundursache in einem dynamischen System und die Auswirkungen latenter Störfaktoren. Wir befassen uns mit diesen Herausforderungen im Bereich der kausalen Inferenz, indem wir die starken Eigenschaften von Kausalmodellen nutzen, wie z. B. die Durchführung hypothetischer Eingriffe in ein System.

Um die Herausforderung der Unsicherheiten bei der Schätzung der kontrafaktischen Verteilung zu bewältigen, schlagen wir eine Bayes'sche Methode vor, die die damit verbundenen Mehrdeutigkeiten berücksichtigt und eine zuverlässigere Entscheidungsfindung ermöglicht. Um die Ursachen in einem dynamischen System automatisch zu identifizieren, entwickeln wir ein autoregressives Modell mit stichprobenbasierter Schätzung dynamischer Prozesse. Im Vergleich zu den meisten nicht-kausalen Ansätzen ermöglicht dies die Identifizierung der Fehlerursache mit nur einer einzigen fehlerhaften Beobachtung, indem die kontrafaktischen Eigenschaften genutzt werden. Abschließend skizzieren wir Ideen für mögliche Erweiterungen von Deconfounding-Ansätzen zur Identifizierung von Grundursachen, wenn unbeobachtete Variablen die kausalen Beziehungen in einem dynamischen System beeinflussen.

Zusammengefasst adressieren diese Methoden wichtige Herausforderungen in der Praxis, indem sie die Robustheit der kontrafaktischen Argumentation erhöhen und gleichzeitig eine schnellere und genauere Identifizierung der Ursachen ermöglichen, was letztlich die Prozesse in der Qualitätskontrolle vereinfacht.

Acknowledgments

First and foremost, I want to express my deepest, heartfelt thanks to my parents, Gaby and Dietrich, and my brothers, Christian and Maximilian. Your steady support and constant encouragement have meant the world to me. You have always been listening, and I am forever grateful for that. Equally, my heartfelt thanks go to my life partner, Driton, whose presence and support have been my constant source of strength, especially through the toughest moments - it means so much to me.

I also want to thank my Bosch PhD colleagues, Mirjam and Laura, for making this journey so unforgettable. You have brought joy and friendship into this journey, and I will always cherish the moments we have shared.

A special thank you goes to my Bosch supervisor, Dr. Sebastian Gerwinn, for his dedicated and professional guidance. I am also deeply thankful to Prof. Dr. Martin Fränzle for giving me the opportunity to write this dissertation, for the enriching discussions, and for his ongoing support. Lastly, I am profoundly grateful to Assoc. Prof. Dr. Melih Kandemir, thank you for agreeing to be the second reviewer.

Please note that in this work, Generative AI was used to check spelling and grammar.

Contents

Acknowledgments **vii**

1 Introduction **1**

2 Background **5**

2.1 Causal Graphical Model 6

2.2 Structural Causal Model 11

2.3 Interventional distribution 13

2.3.1 Do-calculus 14

2.4 Counterfactual distribution 15

2.4.1 Twin networks 17

2.5 Potential Outcomes 18

2.5.1 Potential Outcomes and Counterfactuals 18

2.6 Assumptions in Causal Reasoning 19

2.7 Average Treatment Effect 20

2.8 Method Background 21

2.8.1 Warped Gaussian Process 22

2.8.2 Residual Networks 24

3 Previous Work **27**

3.1 Uncertainties in Structural Causal Models 27

3.2 Counterfactuals in Dynamic Systems 29

3.3 Unobserved Confounding in Dynamic Systems 31

3.4 Continuous-Time Dynamic Causal Models 32

4 Estimation of Counterfactual Interventions under Uncertainties 35

4.1 Uncertain Structural Causal Model 36

4.2 Algorithmic recourse 39

4.3 Counterfactual Inference with a Bayesian Warped Gaussian Process . 40

4.4 Experiments 47

4.4.1 Illustrative example 47

4.4.2 Benchmark Experiments 49

4.4.3 Interventional distribution 51

4.5 Conclusion 55

5 Counterfactual-based Root Cause Analysis in Dynamic Systems 57

5.1 Interventional Dynamic SCM 58

5.2 Abducted and Counterfactual SCM 59

5.3 Root cause 60

5.4 Shapley Value 61

5.5 Method for identifying root causes 62

5.6 Experiments 65

5.6.1 Experimental datasets 66

5.6.2 Evaluation 68

5.7 Conclusion 76

6 Confounding 77

6.1 Static Deconfounder 79

6.2 Time Series Deconfounder 81

6.2.1 Factor model 81

6.2.2 Outcome model 83

6.3 Deconfounding Dynamic Root Cause Analysis 84

6.3.1 Deconfounding with a Counterfactual SCM 85

7 Conclusion **87**

8 Appendix **89**

8.1 Estimation of Counterfactual Interventions under Uncertainties . . . 89

8.1.1 Hyperparameter Setup 89

8.2 Counterfactual-based Root Cause Analysis in Dynamic Systems . . . 90

8.2.1 Hyperparameter Setup 90

Bibliography **101**

Notation

N^j	exogenous variable
$\mathbf{Y}^F$	factum
$\mathbf{Y}^j$	observational variables for node j
$\mathbf{Y}_t^j$	jth observable of a system at time t
$\mathcal{M}$	Structural Causal Model
$do(X = x)$	Intervene in the system by setting random variable X to realization x
f^j	functional relationship (causal link) of j
$\mathbf{Y}^{PA(j)}$	observed $PA(j)$ parents of the node j

List of Figures

1 Ladder of Causality(Figure taken from [1]) 7

2 A directed acyclic graph, also referred to as *causal graph*. Here, Z is an (observed) confounder, affecting the treatment variable X and the outcome Y. Each (observed) endogenous variable has a (latent) exogenous variable N influencing it. 8

3 Chain 8

4 Confounder or Fork 8

5 Collider 9

6 Summary graph $\mathcal{G}$ (on the left) and corresponding unrolled graph $\mathcal{G}_t$ over discrete timesteps t (on the right). 12

7 Residual Neural Network building block. Here, $\mathcal{F}$ is a residual function with a mutable number of layers, x being the input. (Image taken from [2]) 24

8 Illustration of the SCM in Eq.2 42

9 The predictive distribution of the BW-GP (left) and the GP (right) of the modelled interventional distribution. 48

10 Illustration of the modelled counterfactual distribution of our BW-GP (left) and a GP (right). 49

11 On the left, we plot the ground truth, and on the right, the Bayesian Warped Gaussian Process model. The colouring corresponds to the classes the classifier yields in the recourse task (blue are the negative and orange are the positive classified points). 52

12 On the left, the Gaussian Process is plotted, and on the right, the CVAE model. The colouring corresponds to the classes the classifier yields in the recourse task (blue is the negative and orange are the positive classified points). 52

13 This figure shows an overview of the individual steps of our method. 62

14 The figure shows the counterfactual samples for the FHN system with injected root cause at $(j = x_1, t = 24)$. The injected root cause disrupts the system observation heavily (black dashed line). However, the counterfactual intervention performed by our model *NLin(*S_t^j, N_t^j*)* corrects the failure in both dimensions, such that it lies inside the threshold region (orange area). 69

15 The root cause was injected at a random node $j = x_1$ at $t = 6$ with varying constants in $[1, 10]$. The horizontal axis shows the injected constant in relation to the noise standard deviation denoted by σ. We report how many root causes could be identified in %. 70

16 The root cause was injected at a random node $j = x_1$ at $t = 24$ with varying constants in $[0.1, 2.0]$. The x-axis shows the injected constant in relation to the noise standard deviation denoted by σ. We report how many root causes could be identified in %. On the left, for *NLin* it can be seen that the model intervening on the structure and the noise achieves a significantly higher identification score for larger added constants. This could be due to a large root cause, in this setting leading to a factum with high distance to the normal data, which may lead to a divergence over time of the normal behaviour system $\mathcal{M}$. On the right, we illustrated *Lin* and as expected, it is inadequate for addressing the complexities of the non-linear problem. 71

17 With the geographical knowledge of the river flow, a summary graph can be inferred (Figure taken from [3]). 74

18 We show five counterfactual samples (for each station) of our model *NLin(*S_t^j, N_t^j*)* with the intervention at the predicted root cause at 08:30 on 16.03.2019. Additionally, we illustrate the resulting Shapley values for each time point, showing that right before the failure occurs the Shapley values increase. 75

19 *Static multi-cause confounder* (Figure taken from [4]). Note that the substitute confounders Z_i are pre-treatment variables and should not contain any mediators along the assigned causes and the outcome because mediators cannot be identified by looking only at the assigned causes [4]. 80

20 *Time Series Deconfounder* (Figure taken from [5]). The Time Series Deconfounder rules out the existence of other multi-cause confounders like V_t which are not captured by Z_t by assuming conditional independence between A_{tk} given Z_t and X_t. As in the static approach of [4], they assume no hidden single cause confounders L_t. 83

21 Extended causal graph with hidden confounder v. Note that variable names in the summary graph have been changed to prevent notation misuse. 86

List of Tables

1 Experimental results of a three variable causal model in a recourse setting with 100 individuals. We compare our model ($\mathcal{M}_{\text{BW-GP}}$) against the reproduced baselines `LIN`, `GP`, `CVAE` of [6]. 50

2 Experimental results of a seven variable semi-synthetic causal model on 100 facta in a recourse setting. 50

3 We show the **Accuracy** for the setup $\sigma = 500$ of the linear synthetic system (see Fig. 15) with varying number of *removed or added edges* of the summary graph $\mathcal{G}$ used by the models. 72

4 We report the **Accuracy** over 20 facta of the summary graph on a linear system and the FHN oscillator. In the lower part of the table we present the experimental results of the EasyRCA benchmark [7] comparing the accuracy for one factum over 30 graphs for different factum lengths T (here, normal data has the same size T). [1] 73

5 Optimal Hyperparameters found with BOHB on a validation set for each SCM and classifier setting. 90

6 Hyperparameter setup of *NLin* for the performed experiments. . . . 90

1 Introduction

As modern industries become increasingly complex and interconnected, promptly identifying and addressing potential failures has become more critical. In manufacturing, where products are passed from one station to the next, every inefficiency, defect, or downtime event can majorly affect productivity and profitability. When a failure occurs at a station, the root cause may have occurred at the station itself or at previous stations at a specific point in time. In the worst case, the expert has to consider each station individually, which makes an automated root cause analysis particularly valuable. It is, therefore, desirable to discover the root cause of a fault in a production line at an early stage. To effectively identify root causes, we use causal analysis rather than relying solely on correlation-based reasoning. Depending exclusively on correlations may lead to misleading or spurious conclusions, as correlations do not imply causation.

Furthermore, causal analysis enables us to change a system through an intervention, which is essential for making informed decisions and predicting the consequences of an intervention without actually performing it. A conceivable intervention in a manufacturing plant could, for example, change the state of a single station and, as a result, influence all following stations. Based on the concept of interventions, causal models offer the capability to calculate counterfactual reasoning, allowing to explore "what-if" questions and hypothetical interventions on a past observation. In a manufacturing plant identifying the failure-causing station, one could ask the counterfactual question: "Would the observed failure also occur if we had replaced the behaviour of a station at a certain point in time with its *normal* behaviour?". As we address this question in a dynamic context, this poses new challenges like the effect of an intervention may change depending on when in time it is applied or delayed effects, where the influence of an intervention is recognizable with a considerable time delay. This temporal complexity also opens up practical problems, particularly in terms of scalability, where, in addition to the complexity due to the increased number of variables, the number of possible interventions scales with the number of points in time where the failure could have occurred. To perform counterfactual reasoning, the analytical or at least computational expressions of the functional couplings between the variables are required [8]. As causal relationships (we call them also functional relationships) between different aspects of a system might not be fully known to an expert and, therefore, have to be inferred from observational data, there are uncertainties surrounding the identified relationships. Within the setting of counterfactual reasoning, it turns out that additional inherent ambiguities occur, which cannot be constrained by the observational data but will lead to different counterfactual results [8]. Ambiguities in the causal relationships may also occur due to hidden confounding variables, which simultaneously influence both the treatment and the outcome. As a result, the assumption of no present hidden confounders can lead to an inaccurate representation of the system, thereby introducing bias into the estimated effect of an intervention [4]. This may lead to incorrect conclusions, making it essential to account for potential confounders to ensure the validity and reliability of causal analysis.

In the context of ambiguities, the current work does not address the uncertainty in counterfactual distributions while accounting for the uncertainty of functional couplings in a continuous and nonlinear context. Furthermore, the related work focuses on additive external influences causing failures rather than structural influences in a static and often linear setting, which limits its practicability (for a review of related work, see Chapter 3).

Research questions

To overcome the previously mentioned shortcomings, we identified the following research questions:

1. How to represent different uncertainties in functional couplings within an uncertain Structural Causal Model?

2. How to perform counterfactual reasoning for root cause analysis in a dynamic system?

3. How to incorporate unobserved confounding variables in estimating counterfactual reasoning for dynamic root cause analysis?

To answer these questions, we developed a dynamic counterfactual reasoning framework designed to automatically identify root causes within a system while accounting for inherent ambiguities. Based on the stated research questions, the work is structured as follows. First of all, in Chapter 2, we summarize the foundational causal concepts and methodologies on which this work is based. Following this, we review the related literature in Chapter 3. In Chapter 4, we address the first research question by incorporating uncertainty into counterfactual reasoning, which is based on our published work in Weilbach et al. [9]. We then progress to a dynamic setup in Chapter 5, targeting the second research question by performing root cause analysis using counterfactual reasoning in a dynamic context (based on our published work in Weilbach et al. [10]). Finally, in Chapter 6, we discuss how the third research question could be addressed by sketching potential approaches for integrating unobserved confounders into a dynamic setting.

2 Background

This chapter aims to equip the reader with a comprehensive overview of the foundational concepts of causal inference. It is structured around Judea Pearl's causal ladder [1], which consists of three rungs, namely *Association-Intervention-Counterfactuals* categorising the different levels of causal reasoning. We start by introducing causal graphical models, which serve as the backbone of causal reasoning, providing a structure to represent and analyse cause-and-effect relationships. Building on the causal graphical models, we explore Structural Causal Models (SCMs), which extend these by specifying the causal relationships. SCMs provide a comprehensive framework for understanding how interventions affect outcomes and modelling complex causal structures.

Subsequently, we examine the interventional distribution, a key concept in causal inference that describes the distribution of outcomes under specific interventions. This distribution forms the basis for evaluating the effects of actions in both observed and hypothetical scenarios, further solidifying our understanding of the second rung of the causal ladder. As we move to the third rung of the causal ladder, we introduce the counterfactual distribution addressing counterfactual "what if" questions like 'What would have been the outcome if the circumstances had been different?'.

We then discuss the Potential Outcomes Framework, an alternative approach to counterfactual inference that originated in a medical context that focuses on comparing the outcomes under different treatment scenarios for an individual. This framework naturally leads to the concept of the Average Treatment Effect (ATE), which quantifies the average causal effect across a population. As the described causal concepts rely on assumptions to ensure that the conclusions drawn from causal models are reliable, we take a look at the most relevant assumptions for this work. Finally, the chapter concludes with a discussion of the key methods that we use in this dissertation for causal inference.

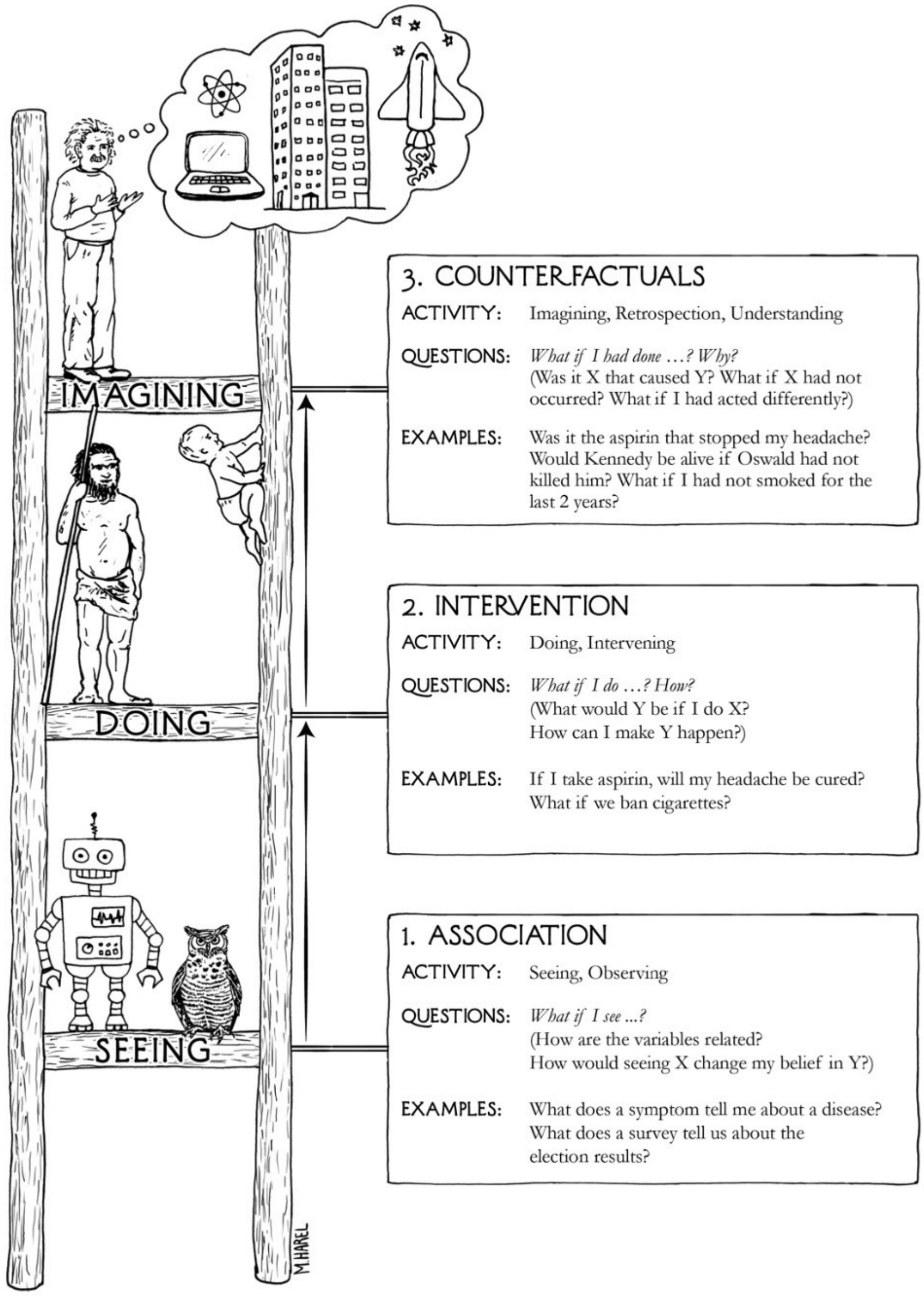
2.1 Causal Graphical Model

In this work, we consider causal models as described in the *Causal Ladder* Fig. 1 that give an overview of the levels of causal reasoning [1]. A causal model usually has an underlying *directed acyclic graph* (DAG) to represent the causal relationships between variables in a system. A DAG offers the advantage of clear visualisation and intuitive interpretation. The edges of a causal graphical model are considered causal links, and therefore, these models offer additional properties to a probabilistic graphical model.

The first rung of the causal ladder represents the learning from association, which is usually specified by conditional probabilities. For example, in a causal graphical model as shown in Fig. 2, the joint probability of the random variables factorises in conditional distributions (without accounting for the exogenous variables N at this point): $P(X, Y, Z) = P(Z) \cdot P(X|Z) \cdot P(Y|X, Z)$. The directed edges illustrate the conditional dependence of the observed random variables.

Incorporating causal graphical structures ensures correct assumptions about the causal relationships, helps avoid biases, and guides model construction. For this reason, in the following we will take a closer look at the most important graphical structures in a causal graph.

A Chain represents a sequence of random variables in which each node is conditionally dependent on its predecessor in the chain. In terms of conditional independence,



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done ...? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do ...? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Figure 1: Ladder of Causality(Figure taken from [1])

a chain can be "blocked" by conditioning on a node. For example, in the causal chain in Fig. 3, conditioning on Z blocks the influence of X on Y, making X and

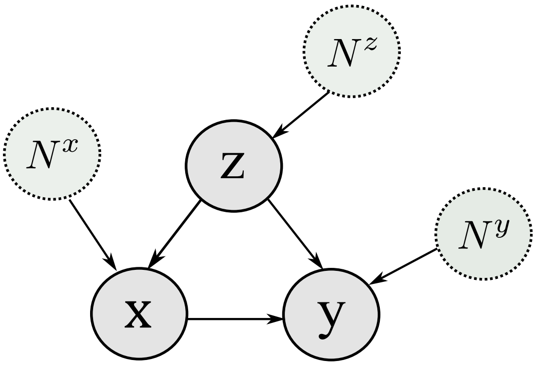


Figure 2: A directed acyclic graph, also referred to as *causal graph*. Here, Z is an (observed) confounder, affecting the treatment variable X and the outcome Y. Each (observed) endogenous variable has a (latent) exogenous variable N influencing it.

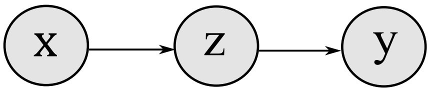


Figure 3: Chain

Y conditionally independent given Z. The variable Z in a chain is also called a *mediator*.

A Confounder (in a graphical context also called fork) is a structure that indicates that the two child nodes (X and Y) are conditionally independent given the parent node Z. However, without conditioning on Z, X and Y may be correlated due to their common cause, Z.

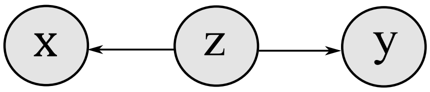


Figure 4: Confounder or Fork

In a Collider structure like in Fig. 5, X and Y are marginally independent of each other. However, they become conditionally dependent given the collider Z. In a graphical model, conditioning on a collider Z (or any of its descendants) opens a path between X and Y, potentially leading to spurious correlations (known as *collider bias*).

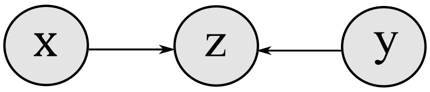


Figure 5: Collider

Reichenbach's Common Cause Principle states that if two random variables X and Y are statistically dependent, then this dependency must be due to one of the following reasons: X causes Y, Y causes X, or there exists a third variable Z —a common cause — that influences both X and Y. In the context of the confounder structure, Z serves as the common cause that explains the correlation observed between X and Y.

d-separation (or directed separation) offers a rule-based approach to determine conditional independence directly from the topology of the DAG without requiring explicit knowledge of the underlying probability distributions. For example, in the confounder structure of Fig. 4, if X and Y are d-separated by Z, then X and Y are conditionally independent given Z. In this example, we ensure this by adjusting for Z. To build an adjustment set of nodes (nodes we need to condition on to block spurious information flow) for a directed graph, the backdoor or frontdoor criterion may be used (introduced by [11]).

Backdoor Criterion

The idea behind the backdoor criterion is to adjust for the confounder Z (in Fig. 2) that could create a spurious relationship between treatment X and the result Y. Conditioned on Z, one can isolate the causal effect of X on Y. For the backdoor criterion to be applicable, the following conditions must be met:

- No Confounders After Conditioning on Z: The identification of a set of variables Y (confounders) that "block" all backdoor paths from X to Y. A backdoor path is a path that starts with an arrow that points to X and then reaches Y.

- Blocking Backdoor Paths: The set of variables Z should block every path between X and Y that contains an arrow that points to X, except for the direct path from X to Y

The backdoor criterion is used when a confounder can be identified and observed. In contrast, the frontdoor criterion is used when a confounder is not observed.

Frontdoor Criterion

The frontdoor criterion is a concept of causal inference used to identify and estimate causal effects if unobserved confounders are present. It provides a way to estimate the effect of a treatment or intervention on an outcome by using an intermediate variable (often called a *mediator*) that lies on the causal pathway between the treatment and the outcome. In Fig. 2 imagine an additional mediator node M in the path from X to Y such that $X \rightarrow M$ and $M \rightarrow Y$. The idea behind this is to decompose the causal effect of X on Y into two components:

- The effect of X on M: This can be estimated directly because there are no unobserved confounders allowed between X and M.

- The effect of M on Y: This can also be estimated because M blocks any backdoor paths from X to Y, making the relationship between M and Y unconfounded.

By examining the graphical structure of a causal graph, we can infer how certain nodes might need to be conditioned to draw proper conclusions. However, the precise nature of the influence that a variable X has on a variable Y remains unclear. Specifically, we do not yet know whether manipulating X will result in an increase or decrease in Y.

2.2 Structural Causal Model

An important advantage of causal models is that they can be represented by direct graphical models, which offer greater interpretability. However, when building a new car or designing autonomous driving functions, the underlying causal graph may not be sufficient to define the relationship of the components or entities involved. In this regard, a Structural Causal Model (SCM) serves as a mathematical structure in causal inference that allows the representation and examination of causal connections among variables. Each equation typically includes a functional form that describes how the variable is determined by its direct causes. In order to capture the stochastic component of the system, exogenous variables are integrated into the structural equations.

Definition 2.2.1. *(Structural causal model (SCM)) [8]*
An SCM $\mathcal{M}(\mathcal{S}, P_N, \mathcal{G})$ is defined by a set of structural equations $\mathcal{S}$, an acyclic graph $\mathcal{G} = (\mathcal{Y}, \mathcal{E})$ with nodes $\mathcal{Y}$ and edges $\mathcal{E}$, and a set of independent noise variables $N^j \sim P_{N^j}, j \in \mathcal{Y}$. The structural equations for each node j are given by:

$$S^j := Y^j = f^j(Y^{PA(j)^{\mathcal{G}}}, N^j)$$

describing the functional relationship f^j of the observational variables $\mathbf{Y}^j$, where $\mathcal{S} = \cup_{j \in \mathcal{E}} \{S^j\}$ is a set of structural collections, and $PA(j)^{\mathcal{G}} \subseteq \mathcal{E}$ denotes the parents of the node j according to the graph $\mathcal{G}$.

For a given SCM, observations $Y = (Y_1, .., Y_d)$ can be generated by sampling $N^j \sim P_{N^j}$ and subsequently applying the functional couplings. We denote this observational

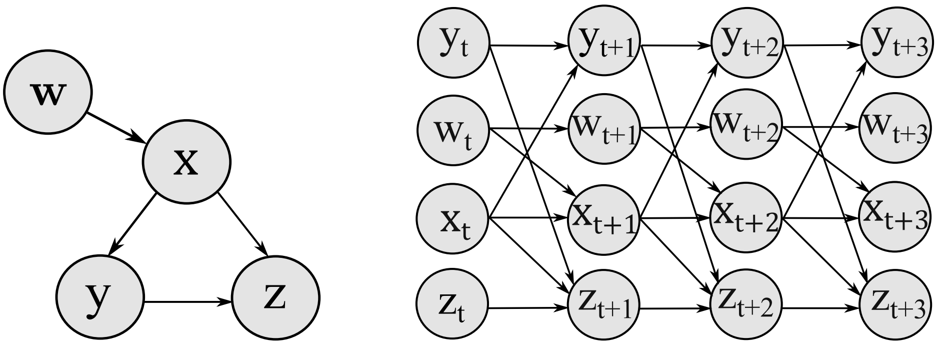


Figure 6: Summary graph $\mathcal{G}$ (on the left) and corresponding unrolled graph $\mathcal{G}_t$ over discrete timesteps t (on the right).

distribution by $P_{\mathcal{G}}$. An extension of Structural Causal Models (SCMs) is dynamic SCMs that capture causal relationships in dynamic systems, where variables evolve over time. While traditional SCMs focus on static relationships between variables, dynamic SCMs explicitly model how variables change over time and how interventions at different time points affect the system's behaviour.

To formally describe dynamic SCMs, we extend the definition 2.2.1 to the dynamic case by unrolling a causal graph over time, as seen in Figure 6. We follow the notation of [8] and define the structural equations for a node Y_t^j by:

Definition 2.2.2. *(Dynamic SCM)*

In analogy to a static SCM, a dynamic SCM $\mathcal{M}(\mathcal{S}_t, P_{N_t}, \mathcal{G}_t)$ is given by an acyclic graph $\mathcal{G}_t$ and exogenous noise influences $N_t^j \sim P_{N_t^j}$ independent over each point in time t and variable j. $\mathcal{G}_t$ refers to a graph consisting of an unrolled version of a summary graph $\mathcal{G}$ (see Fig. 6).

$$S_t^j := Y_t^j = Y_{t-1}^j + f^j(Y_{t-1}^{PA(j)}, Y_{t-1}^j, N_t^j)$$

with $\mathbf{Y}^{PA(j)}$ being the parents of node j according to the summary graph $\mathcal{G}$ excluding the node itself. A notable difference to static SCMs is that the functional coupling f is constant over time.

We model autoregressive dynamic behaviour by assuming discretised time steps. This is a legitimate approach, as the (industrial) problems considered in this work provide measurements at discrete points in time. As long as the intervals between the points in time are sufficiently small, dynamic behaviour can be sufficiently mapped. However, there are several works that focus on continuous dynamical systems in a causal context, see Chapt. 3.3.

2.3 Interventional distribution

In the second level of the causal ladder, in Figure 1, manipulations enable us to intervene in a system and observe the resulting changes, which is a critical feature of causal models that makes them particularly powerful. It is crucial to emphasize that interventions are fundamentally distinct from conditioning within a causal model. An intervention modifies the causal structure by changing the relationship between a variable and its parents. This can be performed as *hard intervention* (sometimes also referred to as structural intervention) by breaking the causal link to the parent nodes or as *soft intervention* (sometimes also referred to as parametric intervention) by manipulating the structural equation with remaining edges to the parents, allowing for a more nuanced form of manipulation. This indicates that a hard intervention has the consequence that edges in the causal graph coming from the parents are eliminated completely [8].

This manipulation or intervention gives rise to a new (interventional) distribution. Interventional distributions represent the distribution of outcomes that would be observed if a particular intervention or action were taken on a system. For example, causal model interventions can be used to find a suitable therapy plan for patients to achieve a desired outcome without actually performing the treatment physically [8]. Formally, an intervention can be carried out in various ways, for example, by adding noise to a variable, by entirely removing a variable from the causal graph, or by setting a variable to a specific value. After an intervention is performed, the new state of the intervened on variable is propagated through the causal graph. This involves applying the structural equations while taking into account the intervened variables. To formally distinguish between intervening and conditioning, we use the *do-operator* to manipulate a variable [1] For example, the (hard) intervention on a variable X by setting it to a specific value x is denoted by $do(X = x)$. Sequentially the interventional distribution of an outcome variable Y when the variable X is set to a specific value x is described by $P(Y|do(X = x))$. The do-calculus consists of three main rules, each serving a specific purpose in transforming and simplifying causal expressions.

2.3.1 Do-calculus

The do-calculus is a set of rules developed by Judea Pearl to facilitate causal reasoning in complex systems when dealing with interventions [11]. The choice of which do-calculus rule to apply is directly informed by the graph's structure. We illustrate these rules using the graph introduced in Fig 2.

Definition 2.3.1. *Rule 1 (Insertion/Deletion of Observations) This rule applies when certain irrelevant variables can be ignored.*

$$P(Y|do(X), Z) = P(Y|do(X))$$

If Y is independent of Z given X in the intervened on graph, then the effect of Z on Y can be ignored.

Definition 2.3.2. *Rule 2 (Action/Observation) This rule is used when the effect of an intervention is equivalent to an observation.*

$$P(Y|do(X), Z) = P(Y|X, Z)$$

If we condition on Z, the relationship between X and Y remains the same whether X is observed naturally or manipulated through an intervention. This holds because Z blocks any unmeasured confounders, making the effect of X on Y equivalent whether X is set or observed.

Definition 2.3.3. *Rule 3 (Insertion/Deletion of Actions) This rule relies on identifying and blocking all back-door paths and corresponds to the backdoor criterion 2.1*

$$P(Y|do(X), do(Z)) = P(Y|do(Z))$$

If Z blocks all back-door paths from X to Y (which it does in this case since Z directly influences both X and Y), then the effect of intervening on X does not provide additional information about Y once we have already intervened on Z. This means that the intervention on Z alone is sufficient to determine the effect on Y.

In summary, intervening involves actively manipulating or changing the value of a variable to observe its causal effect on another variable of interest. Suppose an intervention is performed by setting a variable to a value. In that case, this variable receives no input from its parents any more, meaning the underlying graphical model is changed by cutting the connections between the variable and its parents. This is particularly useful in scenarios where interventions can only be carried out theoretically, for instance because of ethical reasons.

2.4 Counterfactual distribution

The highest and third rung allows for imagining or counterfactual reasoning. A counterfactual distribution is a hypothetical distribution that represents what the outcome would have been for a particular individual under different circumstances.

An (industrial) example for a counterfactual question would be "Would the observed failure also occur if we had changed the behaviour of the system at a certain point in time?". However, only *one* outcome for each individual can be observed, meaning that the counterfactual is unobservable. The formal method for reasoning about counterfactuals (introduced by [12]) based on an SCM consists of three steps: *Abduction-Action-Prediction.* We introduce the procedure using the causal graphical model of Fig 2. We assume that an SCM $\mathcal{M}$ has been defined like in Sec 2.2. Since a counterfactual is interlinked to a particular observation, we first observe a so-called factum $\mathbf{Y}^F$ generated from the underlying SCM (see Definition 2.2.1). For this individual $\mathbf{Y}^F$ we can now calculate the three steps as follows:

Abduction

In the abduction step, we infer the state of the latent variables of the causal model that explain the observation $\mathbf{Y}^F$. To estimate these latent variables, a noise posterior distribution $P(\mathbf{N} \mid \mathbf{Y}^{\mathbf{F}})$ can be calculated with Bayesian Inference [13]. The noise posterior distribution restricts the exogenous noise influences to the ones which are consistent with the given factum $\mathbf{Y}^F$ within the functional couplings of a given SCM $\mathcal{M}$. The exact form of the posterior distribution depends on the specific model assumptions, such as the functional form and the distributions of the variables involved.

Intervention

The action step refers to the process of intervening in the causal model by changing the value of a variable. Assume we want to change the value of X to a new value x' in the model in Fig 2. We denote this (in this case hard) intervention by $do(X = x')$, as described in Sec. 2.3. Then the structural equation for the node X is modified by the intervention like: $do(X = x') \Rightarrow Y^X = x'$. By performing this action step which modifies the causal model so that X is fixed at the new value x', the causal dependence to the parent Z is broken.

Prediction

In the last step, we use the modified causal model from the action step to estimate the counterfactual outcome Y^{CF}. Given the inferred values of the latent variables $\mathbf{N}$ from the abduction step, and the modified structural equation from the action step, we calculate the values for the descendants of X by following the structural equations of $\mathcal{M}$.

2.4.1 Twin networks

The concept of twin networks, as introduced by Balke and Pearl [14], provides a formal mechanism to evaluate counterfactuals. A twin network consists of two parallel models: one representing the actual (factual) world and the other representing the counterfactual world. In this twin of the factual causal model, an intervention is applied, creating a counterfactual causal model.

The primary goal of the twin network approach is to simplify the computation of counterfactual queries by representing both factual and counterfactual scenarios in a unified graphical model. These two models are interconnected through shared latent variables, ensuring that the underlying causal structure remains consistent between the factual and counterfactual scenarios.

However they require duplicating the entire causal model to create a parallel structure for the counterfactual scenario. This duplication increases the model's complexity, making it more difficult to manage, especially in large-scale or complex systems with many variables. Furthermore, twin networks are limited to two possible networks. However, this problem has previously been addressed in a so-called parallel worlds graph [15]. In contrast, in counterfactuals, as introduced by Pearl [12] are based on a single set of structural equations, without the need to duplicate the entire model. This can lead to simpler computations, especially when duplicating the entire structure, as in twin networks might be computationally expensive.

2.5 Potential Outcomes

Since we consider the Potential Outcomes (PO) Framework in Chapter 6 in the context of latent confounding, we would like to provide a brief introduction in the following. The Potential Outcomes (PO) Framework, also known as the Neyman-Rubin Causal Model [16], [17], [18], is a widely used approach for causal inference in statistics and social sciences. It provides a formal way to define and estimate causal effects by comparing the outcomes that would occur under different interventions or treatments. The outcomes are called potential outcomes because, for any given individual, only one of these outcomes can actually be observed (depending on whether they receive the treatment or not). The potential outcome is denoted with $Y \sim P(Y|do(T = 1))$ if the individual receives the treatment and with $Y \sim P(Y|do(T = 0))$ if the individual does not receive the treatment (i.e., is in the control group).

2.5.1 Potential Outcomes and Counterfactuals

In recent decades, there have been repeated discussions in the causality community about the differences and similarities between the Structural Causal Model and the Potential Outcomes (PO) framework. In [19], Bollen and Pearl explain that the two frameworks are interchangeable and that they are logically equivalent [20]. They explain that the key difference lies in how they represent causal knowledge. In an SCM, causal knowledge is encoded through functional relationships among endogenous (observable) and exogenous (latent) variables, while PO represent it using statistical relationships among hypothetical (or counterfactual) variables, whose values are defined after a treatment is carried out. Further, they show that a theorem or assumption in one framework has a corresponding interpretation in the other. For instance, in an SCM, the assumption that X does not cause Y is represented by the absence of an $X \to Y$ arrow. In contrast, a PO analyst would imagine a hypothetical variable Y_x (representing the value that Y would have if treatment $X = x$ were applied) and state that $Y_x = Y$, indicating that the potential outcome Y_x remains unchanged regardless of x, and equals the observed value Y. Similarly, SCM's assumption of independent disturbances is captured in the PO framework as an independent relationship between counterfactual variables.

In this work, we mainly follow Pearl's counterfactual framework because it offers better interpretability through the representation of causal graphical models.

2.6 Assumptions in Causal Reasoning

A key assumption in causal inference is the absence of unobserved confounders between the treatment variable and the potential outcomes. This assumption, often referred to as *ignorability* or *unconfoundedness*, implies that all variables affecting both the treatment assignment and the outcome have been observed and accounted for [21]. If this assumption is violated, the estimated Average Treatment Effect may be biased, leading to potentially misleading conclusions [22].

The following assumptions are commonly made when estimating potential outcomes in the Neyman-Rubin Causal model, see [23, 18].

Stable Unit Treatment Value Assumption (SUTVA) has two components: first, it requires that the value of the potential outcome for any individual is unaffected by the treatment assignment mechanism; second, it assumes that an individual's potential outcome is unaffected by the treatment exposures of other individuals. This assumption is crucial for ensuring that the potential outcomes are well-defined and that interference between units does not occur [18]. Peters et al. [8] show that SUTVA is satisfied when the data are generated from an SCM.

Positivity requires that every individual in the population has a positive probability of receiving each treatment level. This ensures that comparisons between treated and untreated groups are valid because there are individuals in each group who could feasibly have received the other treatment.

Consistency states that if a unit is assigned to a particular treatment, the observed outcome is the same as the potential outcome corresponding to that treatment. This means that the potential outcome for an individual is exactly what would be observed if the individual were exposed to that treatment, implying that there are no different versions of the treatment that might lead to different outcomes. Consistency also requires that the potential outcome is uniquely determined by the treatment value and that no unaccounted-for variations exist in how the treatment is implemented.

2.7 Average Treatment Effect

Once an intervention or treatment[1] has been carried out, we are interested in measuring the overall impact of the intervention. Specifically, we want to quantify how the outcome changes on average when the intervention is applied compared to when it is not. The Average Treatment Effect (ATE) provides a formal way to quantify this effect by comparing the expected outcomes under both the treatment and control conditions. In the context of the potential outcomes framework, each individual exposed to treatment has a potential outcome and another potential outcome if the individual is in the control group. The treatment effect, then, is the difference between these two potential outcomes.

---

[1] Note that we use the terms intervention and treatment interchangeably.

Definition 2.7.1. *For an outcome variable Y and a binary treatment variable T the ATE for the potential outcome with a given treatment $Y \sim P(Y|do(T=1))$ and the potential outcome for no treatment has been given $Y \sim P(Y|do(T=0))$ is defined as:*

$$ATE = \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)]$$

However, only one potential outcome for each individual can be observed, meaning that the remaining potential outcome is unobservable. This is known as the fundamental problem of causal inference [23]. If a treatment effect affects individuals differently it is called heterogeneous and can be addressed by dividing the data into subgroups (e.g., men and women). It is then analysed if the average treatment effects the subgroups differently. The per-subgroup ATE is called a *conditional average treatment effect* (CATE) [22].

2.8 Method Background

In this concluding section of the background, we outline the theoretical foundations of the methodologies employed throughout this dissertation, highlighting their respective strengths and limitations. Specifically, given observational data and a known causal graphical structure, these methods first infer the underlying functional relationships. Once the model representing this inferred SCM is fitted, we can utilize it to estimate both interventions and counterfactuals.

Causal Discovery

Although we assume in this work that the causal graph is known, we would like to briefly discuss the area of research dealing with inferring the causal graph from observational data. In the causal discovery community, one of the most prominent approaches is the *PC Algorithm* [24], which is based on conditional independence tests. If time series data is available, *Granger causality* [25], a method that determines a causal relation by verifying if one-time series can predict another one. However, Granger-causality is designed to handle pairs of variables and may result in misleading results when the causal relationship involves more variables. Yet, there are several extensions to address these limitations [26]. A more recent approach is using Large Language Models to infer the causal relationship with the use of metadata, for example, by using the names of variables as additional information [27]. A limitation in the evaluation of causal discovery methods is that for most real-world examples, the true underlying graph is unknown. Therefore, these methods result in a set of candidate causal graphs. Performing causal inference based on this set of causal graphs induces additional uncertainty. This uncertainty would be present in the complete inference chain, which significantly increases the complexity of estimating causal effects. For this reason, inferring the causal graph from observational data is out of the scope of this dissertation. Although it is evaluated how a violation of the known causal graph assumption affects the causal inference model, see Sec. 5.6.2.

In Chapter 4, we utilize a Warped Gaussian Process (WGP), an extended form of traditional Gaussian Processes, for estimating counterfactual interventions under uncertainty. Chapter 5 involves the application of a Residual Network (ResNet) to identify root causes based on these counterfactual estimations. Although various alternative models are available, we have selected these specific methods due to the distinct properties they offer, which are elaborated upon in the following.

2.8.1 Warped Gaussian Process

A Gaussian Process (GP) is a powerful non-parametric method used for regression and classification tasks [28]. It models the relationship between inputs and outputs as a probability distribution over functions. A GP is fully defined by a prior characterized by its mean function $m(x)$ and covariance function (or kernel) $k(x, x')$, which together determine the properties of the functions that the GP can model. We can incorporate expert knowledge into the model by appropriately selecting the prior on the functions. Bayesian inference is employed to update the belief over the functions, resulting in a posterior distribution over functions, which reflects how the GP's understanding of the data has been adjusted based on the observations.

A significant advantage of GPs is their ability to provide uncertainty estimates, which naturally increase as the posterior predictive (representing the model's predictions along with uncertainty estimates for new inputs) is evaluated further from the training data. Although Gaussian Processes are capable of modelling complex relationships, they inherently follow Gaussian distributions. However, in many real-world scenarios, the data may exhibit non-Gaussian characteristics, such as with multi-modal noise distributions. Therefore, Snelson et al. [29] developed Warped Gaussian Processes, which are particularly useful when the data (or the underlying function) to be modelled is not well-represented by the assumptions of a standard Gaussian Process. Warped Gaussian Processes apply a nonlinear warping function to transform the observation space, making the data more compatible with modelling by a GP. This transformation allows the model to capture more complex behaviours, including non-Gaussian characteristics such as multi-modality.

The non-linear warping function allows for a non-Gaussian mapping with non-additive noise. Maroñas et al. [30] apply a Normalizing flow (NF) [31] as non-linear warping since it is a powerful and flexible class of models used for density estimation and generative modelling. An NF is based on the idea of transforming a simple, well-known distribution (such as a Gaussian) into a more complex target distribution by applying a series of invertible and differentiable transformations. These characteristics ensure that the transformation remains bijective, in the sense that forward mapping and the reconstruction of the original space is feasible. An NF covers multi-modal distributions well while accounting for complex densities and maintaining analytical tractability [31].

In summary, WGPs are generalisations of GPs that can model a wider range of data distributions than standard GPs, making them more versatile in practice. However, GPs are computationally expensive depending on the size of the training data, and so are WGPs. The additional learning of the warping function adds additional complexity to the model, which could make it more computationally intensive and harder to interpret. However, to make GPs more efficient there are various extensions, such as sparse GPs [32], but this is outside the scope of this work.

2.8.2 Residual Networks

Residual Networks are a type of deep neural network architecture that was introduced to address the challenges of training very deep networks, particularly the problem of vanishing and exploding gradients [2]. In a traditional feedforward neural network, each layer learns a transformation of its input.

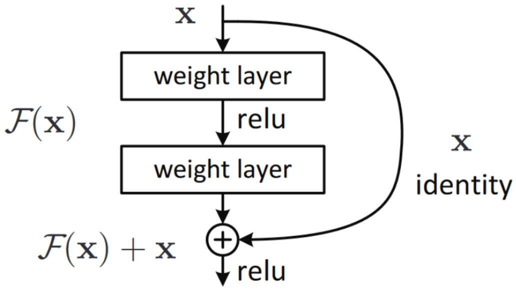

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 7: Residual Neural Network building block. Here, $\mathcal{F}$ is a residual function with a mutable number of layers, x being the input. (Image taken from [2])

In contrast to that, the core idea of ResNets is the residual block (see Fig. 7), which allows us to learn only the differences (or residuals) between the output and the input rather than trying to learn the entire transformation. The learning or training process in ResNets, like other neural networks, involves minimizing the empirical risk or training error, where the network adjusts its parameters (specifically the weights) by minimizing a loss function (e.g. cross-entropy or mean squared error depending on if the task is belonging to classification or regression). The loss function quantifies the difference between the neural network's predictions and the actual target values. The model weights are then updated via backpropagation using gradient descent or its variants [33, 34].

In addition to the standard neural layers, a residual block includes skip connections. These connections allow the input to bypass one or more layers and be added directly to the output of the residual block. In other words, the original input is added directly to the output of the standard layers [2]. These skip connections facilitate better gradient flow during backpropagation, allowing for deeper networks to be trained effectively by mitigating issues like vanishing gradients. In ResNets, the use of skip connections can also be viewed as a form of implicit regularization. We commonly use regularization techniques to prevent overfitting, especially when dealing with limited data. Regularization methods such as weight decay or dropout are used to ensure that the model is not overfitting by penalizing overly complex models. However, skip connections may not always provide a significant advantage and may introduce unnecessary complexity. For example, ResNets can lead to overfitting in small datasets, as the model may struggle to learn the patterns [35, 36].

3 Previous Work

In this chapter, we review the related work based on the research questions stated in Sect.1. First, we review the literature on different uncertainties in functional couplings as well as the ambiguities of a Structural Causal Model. Secondly, we review dynamic counterfactual reasoning for root cause analysis. And finally, how to incorporate an unobserved confounder in such a dynamic model.

3.1 Uncertainties in Structural Causal Models

In this section, we review the literature related to our work (in Chapter 4), where we account for ambiguities in the underlying SCM as well as for the uncertainties arising from imperfect knowledge of functional couplings due to limited observational data. Specifically, we want to capture the uncertainty about the parameterisation of an SCM in addition to the uncertainty in the functional couplings in a continuous setting. Counterfactual distributions are non-identifiable due to an inherent ambiguity in counterfactual reasoning. This arises from the fact that different SCMs can have distinct parametrizations of stochastic influences, resulting in identical observational and interventional distributions but producing different counterfactual outcomes (see an illustrative example in Sec. 4.3). Although this non-identifiability of counterfactual distributions can be avoided by imposing additional assumptions onto the underlying structural equation and the exogenous noise distributions [11, 37], these additional assumptions are inherently non-testable and specific modelling assumptions are currently only available for discrete variables within the structural causal model [38]. Correspondingly, the identifiability of discrete variable SCMs has been addressed by [39, 40, 41, 42] and [43] by treating the counterfactual distribution directly as a Bayesian variable.

In the counterfactual reasoning, the analytical or at least computational expressions of functional couplings of the SCM (see Def. 2.2) are required [8]. Although these functional couplings can be learned from data, they cannot be uniquely identified due to limited data or intrinsic degrees of freedom of the modelling choices of the functional couplings. In Karimi et al. [44], a Gaussian process is used to capture the uncertainty over the functional couplings. However, they do not account for the inherent uncertainty over the parameterisations of an SCM in a counterfactual.

Given the range of methods focused on ensuring robust counterfactuals, we will explore some of them in more detail in the following. In [45], they quantify the robustness of counterfactual explanations by evaluating how small changes in input data affect the validity of the counterfactual. However, they do not account for uncertainties in the functional couplings; they even assume that these functions are given. The authors of [46] primarily focus on the robustness of counterfactuals under model changes in the context of neural networks (e.g. occurring while retraining the model). They assumed that the neural network model is fully accessible without accounting for uncertainties. Dutta et al. [47] introduce an algorithm that refines counterfactuals generated by existing methods to make them more robust. However, these methods do not consider the underlying causal structure. Note that the term counterfactuals does not necessarily imply using a causal model.

In summary, the current work does not address the ambiguity in counterfactual distributions while accounting for the uncertainty of functional couplings in a continuous and non-linear context. Therefore, we would like to use Chapter 4 to close the gap of nonidentifiability of counterfactual distributions in a continuous setting by explicitly modelling the uncertainty over the parameterisation of an SCM and the uncertainty over the functional couplings.

3.2 Counterfactuals in Dynamic Systems

In this section, we give an overview of the related work to our research question about counterfactuals for dynamic systems. We want to address the identification of root causes in terms of a counterfactual query: "Would the observed failure also occur if we had replaced the behaviour of a sub-system at a certain point in time with its *normal* behaviour?". We notice the following shortcomings in the existing literature:
Static systems: Root cause analysis based on causal inference has been addressed in [48, 49, 3], but only in static environments.
Structural influences: Existing causal inference methods using counterfactuals [49, 3] focus on additive external influences causing failures rather than structural influences. To detect root causes affecting the graphical structure or transition function Assad et. al. [7] propose a method based on the assessment of the direct causal effect. By modelling such causal effects with linear models, Assad et al. [7] shows that the total effect changes if the underlying causal model changes. In turn, they can use this fact to identify structural changes in the causal model. However, their method is limited to linear models and does not include single-time external influences.
Non-linear systems: Existing methods for root cause analysis are typically limited to linear dynamic models. Additionally, existing methods are limited to small systems as they rely on the computation of Shapley values, which scale exponentially with the number of variables. This becomes infeasible in a dynamic setting since the corresponding causal graph – unrolled over time – would have an increasingly large number of nodes.

Suppose more than one observation (as in our counterfactual approach) of anomalous data is available. In that case, the problem of identifying the root causes is also amenable to statistically estimating the correlation or causation of the different variables and time points onto the variable associated with the label "anomalous". To this end, Tonekaboni et al. [50] introduce feature importance in time (FIT), a scoring mechanism to quantify the importance of features in a multivariate time series. The authors propose to assess feature importance based on their predictive power w.r.t. the outcome distribution while accounting for temporal distributional shifts. The approach localizes important features over time and can thus be used to gain useful insights into the behaviour of dynamic systems. However, FIT does not exploit the causal structure of the underlying system and provides correlative explanations for the observed outcomes.

However, the best alignment with the problem setup presented in Chapter 5 is the work by [3], which defines the problem of identifying root causes of system failure as a counterfactual query. With this reformulation, the authors claim to be the first to propose actionable explanations for the anomalous behaviour of underlying systems. In principle, counterfactual reasoning assumes and leverages complete causal knowledge of the underlying system in the form of a SCM. More precisely, the work in [3] assumes linear functional causal models in *invertible* models in which exogenous variables are computable from endogenous system observations. In fact, the authors leverage the default split between endogenous and exogenous variables in a graphical causal model to disentangle a node's inherited impact from its own contribution. They further complement their work with two key contributions. They account for the notion of graded causation [51] and provide order-independent feature scoring using a game-theoretic concept commonly adopted in explainable machine learning [52]; namely Shapley values [53]. With its computational complexity, their approach lacks direct applicability to dynamical systems because Shapley values scale exponentially with the number of variables. This becomes infeasible in a dynamic setting since the corresponding unrolled causal graph would have an increasingly large number of potential candidates; see Sect. 5.4. Furthermore, they assume linear causal relationships. In summary, the existing causal methods for root cause identification are typically limited to linear static settings and focus on failures caused by additive external influences rather than structural influences. In Chapter 5, we address failures from both influences by modelling the dynamic causal system using a Neural Network and deriving corresponding counterfactual distributions over trajectories.

3.3 Unobserved Confounding in Dynamic Systems

The exploration of causal models, particularly those involving hidden confounding variables, has been an important area of research for several decades. In this section, we briefly present the research concerned with latent confounding in order to form the basis for Chapter 6.

One of the fundamental works in this field is the work of Pearl et al. [54]. This work introduced a probabilistic framework for evaluating sequential decision plans in the presence of hidden variables. Specifically, it uses a graphical model structure where hidden variables are incorporated into the causal graph that allows for the adjustment of these hidden variables. Building on this foundation, Elidan et al. [55] continued to explore the implications of hidden variables in causal inference. They search for "structural signatures" of hidden variables — substructures in the causal graph that tend to suggest the presence of a hidden variable. Rosenbaum's work on sensitivity analysis [56] provides another major perspective on dealing with confounders. He focuses on assessing the robustness of causal inferences in observational studies, where randomization is impossible and hidden biases might exist. His sensitivity analysis method quantifies how the estimated treatment effect would change if an unmeasured confounder affected both the treatment and the outcome. Numerous further works focus on the calculation of the treatment effect with the present latent confounding variables in a static setup, e.g. [4, 57, 58, 59, 60]. In fact, some works even consider latent variable modelling and cyclic behaviour, such as [61], [62].

To address time-varying confounders in dynamic settings, Robins [63] introduced Marginal Structural Models (MSMs) to estimate causal effects in longitudinal data. To address time-varying confounders, MSMs use Inverse Probability of Treatment Weighting to create a pseudo-population in which the treatment assignment is independent of the observed confounders. The weighting adjusts for observed confounders and allows for an unbiased estimate of treatment effects. More recently, some approaches have focused on the estimation of causal effects that vary over time with latent time-dependent confounders by first inferring the latent confounders and then using MSM for the estimation of effects (see Chapter 6) [5], [64]. Since existing approaches like [5] assume multiple treatments available at each time step (see Chapter 6), Hatt et al. [65] recently introduced an approach for a single treatment at each time step. Another recent approach by Haussmann et al. [66] focuses on single-arm trials where all patients belong to the treatment group. They propose an identifiable deep latent-variable model that can also account for missing covariate observations. Finally, for further reading, we refer to a comprehensive review of recent developments in causal reasoning with latent variables, see [67]. Overall, while there is substantial research on static confounders, there is limited work on unobserved confounders in dynamic systems. Additionally, some approaches assume idealized conditions, such as the availability of multiple treatments, which may not reflect real-world scenarios. Given the limited attention to complex dynamic environments where unobserved confounders evolve over time, we sketch an idea to address this in Chapter 6.

3.4 Continuous-Time Dynamic Causal Models

Although we employ discretised time series in this dissertation, we would like to briefly present some papers on the modelling of continuous time series, as our work could also be extended in this regard. The following methods define a mathematical framework through differential equations in order to define an SCM. In continuous-time causal models based on differential equations, Peters et al. [68] introduce foundational concepts in causal modelling specifically tailored for dynamical systems. The research of [69], [70] focuses on the identifiability of the effects of interventions in time-continuous dynamic systems. Similarly, the approach of Bouwer et al. [71] introduces a method to predict the effects of interventions over time, also accounting for the inherent uncertainty. In addition to estimating interventions and their effects, research is being pursued that addresses the counterfactual in continuous time, such as [72] and the work of Sanchez et al. [73], where by using a diffusion model in the counterfactual they allow for latent variable inference with forward diffusion. However, in practice, observations are available at discrete points in time, and therefore, we focus on time-discretized formulations.

4 Estimation of Counterfactual Interventions under Uncertainties

Due to their hypothetical nature, counterfactual distributions are inherently ambiguous. This ambiguity is particularly challenging in continuous settings in which a continuum of explanations exists for the same observation. In this chapter, we address this problem by following a hierarchical Bayesian approach that explicitly models such uncertainty. In particular, we derive counterfactual distributions for a Bayesian Warped Gaussian Process, see Sect. 2.8.1 thereby allowing for non-Gaussian distributions and non-additive noise. Finally, we illustrate the properties of our approach in a synthetic and semi-synthetic example and show its performance when used within an algorithmic recourse downstream task [9].

Within the context of counterfactual reasoning, there is an inherent uncertainty due to different SCMs featuring disparate parameterisation of stochastic influences but producing the same observational and interventional distributions (see Figure 8). Although this non-identifiability of counterfactual distributions can be avoided by imposing additional assumptions onto the underlying structural equation and the exogenous noise distributions [11, 37], these additional assumptions are inherently non-testable and specific modelling assumptions are currently only available for discrete variables within the structural causal model [38]. Instead of imposing non-testable assumptions on the model structure, we propose to follow a hierarchical Bayesian approach which assigns a prior on different parametrizations that leads to potentially different counterfactual distributions and infers the corresponding posterior from observations. By averaging across different parametrizations, we effectively account for all possible counterfactual distributions consistent with the observations. Specifically, we equip the established Gaussian Process with random transformations by placing a Normalizing Flow on the likelihood function [30]. Such a transformation allows for non-Gaussian distributed descendent node variables. It also provides a means of assessing possibly different SCMs with the same observational distribution. To this end, we extend the method of [44] by replacing the Gaussian Process with a Bayesian transformed Gaussian Process according to [30] using a Bayesian Normalizing flow as likelihood transformation. Using this extended setting, we derive the corresponding counterfactual distribution and show that the resulting distribution over counterfactual estimates can account for the non-uniqueness of counterfactual distributions due to ambiguous parametrizations.

4.1 Uncertain Structural Causal Model

Throughout this Chapter, we assume that there is no latent confounder influencing multiple observational variables, reflected by the independence assumption of $P_N = \prod^j P_{N^j}$ of the exogenous variables N^j. For a given SCM, observations $Y = (Y_1, .., Y_d)$ can be generated by sampling $N^j \sim P_{N^j}$ and subsequently applying the functional couplings. To incorporate imperfect knowledge into the notation of an SCM, we extend Definition 2.2.1 to allow functional couplings to be subject to further uncertainty:

Definition 4.1.1 (Uncertain SCM)**.** *An uncertain structural causal model $\mathcal{M} = (\boldsymbol{S}, P_F, P_N, \mathcal{G})$ additionally contains a distribution P_F which allows for specifying independent distributions over functional couplings within an SCM $\mathcal{M}$:*

$$S^j : Y^j = f^j(Y_{pa(j)}, N^j), \quad N^j \sim P_{N^j}, f^j \sim P_{F^j}$$

Note that, within the above definition, we introduced an additional distribution on functional couplings f^j, which allows us to separately interpret different random effects: exogenous noise and imperfect knowledge of functional mappings. The uncertainty arising from incomplete knowledge of the functional couplings is epistemic and can be reduced by acquiring more data. In contrast, the uncertainty due to exogenous noise is intrinsic to the system's randomness and is categorised as aleatoric uncertainty.

When uncertainty is introduced through parameterisation, it becomes both epistemic and aleatoric since the structural equation is modified by the function g (which dictates the interaction between the sampled function f and exogenous noise). Consequently, the output of g encompasses both epistemic and aleatoric residual uncertainty. However, with an adequately accurate model, the residual uncertainty should predominantly be aleatoric.

To estimate interventional distributions, we would average across both random influences, whereas for counterfactual analysis, we fix the exogenous noise influence and only average across our imperfect knowledge of the functional mappings. With a slight abuse of notation, we do not distinguish between uncertain and deterministic SCMs as deterministic SCMs are a special case of uncertain SCMs by defining a point mass distribution P_F on the deterministic functional couplings. Within an SCM, we denote interventions using the *do*-operator. That is when intervening on a set of variables $Y_{\mathcal{I}} = (Y_{\mathcal{I}_1}, \dots Y_{\mathcal{I}_a})$ to set values θ explicitly for these variables, we substitute the corresponding structural equations by $S_{\mathcal{I}_i} : Y_{\mathcal{I}_i} = \theta_i$ and denote the corresponding derived SCM with $\mathcal{M}[do(Y_{\mathcal{I}} = \theta)]$. With $P_{\mathcal{M}}$ we denote the data-generating distribution from which observations Y can be generated by propagating samples of $Y_{\mathrm{pa}(j)}$ to Y^j via sampling N^j and applying the functional mapping.

Counterfactual analysis estimates hypothetical alternative outcomes that would arise if an individual had made a different decision. It is, therefore, directly linked to a particular observation Y^F generated from the underlying SCM (see Definition 2.2.1). To perform this kind of analysis, in the first *abduction* step [11], a noise posterior distribution $P_{N|Y^F}$ is calculated. This noise posterior distribution restricts the exogenous noise influences to those that are consistent with the given factum Y^F within the functional couplings of a given SCM $\mathcal{M}$. Consequently, for a given SCM $\mathcal{M}$ and factum Y^F, we denote the adapted counterfactual SCM $\mathcal{M}_{|Y^F} = (\mathbf{S}, P_{N|Y^F}, \mathcal{G})$. Calculating the noise posterior depends on both the functional coupling f^j and the noise distribution P_N, which is particularly challenging when functional couplings are also considered to be probabilistic, i.e., $f^j \sim P_{f^j}$. Within this work, we rely on available results for calculating the noise posterior distribution for the case in which functional couplings and noise distributions are modelled with Gaussian distributions, which we state in the following:

Proposition 4.1.1 (Noise posterior of a Gaussian Process [44])**.** *Let a node j of an uncertain SCM in which the functional couplings are distributed according to a Gaussian Process with kernel k^j and additive noise N^j be given by:*

$$Y^j = f^j(Y_{pa(j)}) + N^j; \quad f^j \sim \mathcal{GP}(0, k^j); N^j \sim \mathcal{N}(0, \sigma_j^2)$$

For an observed factum Y^F with $y^{j,F}, y^F_{pa(j)}$ containing descendent and parent observations according to the graph $\mathcal{G}$ and training data $\boldsymbol{Y}^j = \{y^{j,i}\}_i$, $\mathbf{Y}_{pa(j)} = \{Y^i_{pa(j)}\}_i$, the noise posterior $P_{N|Y^F}$ is given by:

$$P(N^j|Y^F) = \mathcal{N}(\mu^j, \Sigma^j), \ \textit{with} \quad \mu^j = \sigma_j^2 \left(\left(K^j + \sigma_j^2 \mathbb{1}\right)^{-1} \left(\boldsymbol{Y}_j, y_j^F\right)\right)_{i_F},$$

$$\Sigma_j^2 = \sigma_j^2 \left(\mathbb{1} - \sigma_j^2 \left(K^j + \sigma_j^2 \mathbb{1}\right)^{-1}\right)_{i_F, i_F}, \ K^j = k^j \left(\left(\boldsymbol{Y}_{pa(j)}, y^F_{pa(j)}\right), \left(\boldsymbol{Y}_{pa(j)}, y^F_{pa(j)}\right)\right)$$

where i_F indicates the index of the factum, i.e. the last entry, as training data and factum are concatenated.

4.2 Algorithmic recourse

To evaluate the proposed method in are more realistic setting, we apply it on an established algorithmic recourse benchmark [44], thereby assessing the impact of the counterfactual distribution on the downstream task as making accurate decisions in such settings based on quantitative results requires handling uncertainties effectively. Optimizing interventions for complex and often intransparent decision processes is a key application of counterfactual analysis and is called algorithmic recourse. However, the graph of the causal dependencies between variables is often not sufficient to perform the necessary optimizations. When using counterfactual distributions to devise recourse actions [44, 74, 75], rendering counterfactual explanations more robust against uncertainties, is an active field of research. For a recent survey on algorithmic recourse, refer to [76]. The algorithmic recourse setting aims at finding a counterfactual explanation[77], which would have led to a more desirable outcome for a particular individual represented by observations Y^F. This can be translated into an optimization problem in which the outcome is characterized by a given classifier $h : \mathcal{X} \to [0, 1]$ from which the outcome of an observation, e.g., getting a loan approved, can be predicted by thresholding $h(x) \geq 0.5$ or alternatively sampled according to the probability $h(x)$. In turn, the recourse problem can be formulated as a constrained optimisation problem which minimises the costs for performing an intervention under the constraint that it would have led to an alternative (more desirable) outcome. The costs are typically associated with the distance of the action of setting a particular value to the factum for which one would like to obtain a different outcome, as performing such action would require changing the individual or its properties. In [44], it is extended to also account for the uncertainty within the functional couplings resulting in the following algorithmic recourse formulation:

$$\min_{a=do(\mathbf{Y}_{\mathcal{I}})} \operatorname{cost}(a, Y^F)\; s.t. P_{\mathcal{M}[a]_{|Y^F}}(h(Y)) \geq 1-\delta \tag{1}$$

Eq. 1 minimises the cost of an action a (performing interventions on an intervention set $\mathcal{I}$) for an individual Y^F (one observation, "negatively" classified) such that the found counterfactual sample Y reaches the "positive" side after being applied to a classifier h under the counterfactual distribution ($P_{\mathcal{M}[a]|Y^F}$). Herein δ specifies the residual risk that one is willing to accept for not achieving the desired outcome. The constraint, therefore, measures the minimal probability, which can be stated as a threshold on the expectation of the classifier. Note that the constraint in the above optimisation problem is specified in terms of the counterfactual distribution. In this work, however, instead of requiring a high success rate under a single counterfactual SCM, we additionally average across possible SCMs, i.e. replacing the constraint in Eq. 1 by $P^j \left(P_{\mathcal{M}_\phi[a]_{|Y^F}}(h(Y))\right) \geq 1-\delta$. Here, P^j represents the distribution over possible ϕ-parametrized SCMs $\mathcal{M}_\phi$ that are all consistent with the observations. By introducing additional uncertainty, which only affects the counterfactual distribution, we expect a more uncertain classification outcome under the counterfactual distribution and hence also expect more robust recourse actions.

4.3 Counterfactual Inference with a Bayesian Warped Gaussian Process

The interventional or observational distribution of an SCM is determined by the conditional distributions $p(Y^j|Y_{\mathrm{pa}(j)})$. These distributions, however, can be realized with different combinations of functional coupling and exogenous noise influences. The chosen representation determines the counterfactual distribution in which the exogenous noise influence is kept fixed. To illustrate this effect of different parametrizations of the noise influence and functional coupling, consider the following adapted example from [8]. We construct a family of SCMs $\mathcal{M}_\phi$ with $\phi \in [0, 1)$, over two observational variables Y_1, Y_2. All members of the family give rise to the same observational and interventional distributions, but each leads to different counterfactual distributions. Specifically, these SCMs are constructed using the following relationship between Y_1, Y_2 and the corresponding noise influences N_1, N_2:

$$
\begin{aligned}
&Y_1 = N_1, \quad N_1 \sim \mathcal{U}[0,1]; \quad Y_2 = \mathbb{1}_{Y_1 < 0.5} N_2 Y_1 + \mathbb{1}_{Y_1 \geq 0.5} \zeta_\phi(N_2) \quad N_2 \sim \mathcal{U}[0,1] \\
&\zeta_\phi(n) = \mathbb{1}_{n+\phi \geq 1}(n + \phi - 1) + \mathbb{1}_{n+\phi<1}(n+\phi) \qquad (2)
\end{aligned}
$$

Here, ζ_ϕ modifies a uniform distribution $\mathcal{U}[0, 1]$ by shifting its support by ϕ and re-mapping it to $[0, 1]$ by cutting off all values larger than 1 and mapping them to $[0, \phi]$. Consequently, the resulting random variable shares the same cumulative distribution function as $\mathcal{U}[0, 1]$. However, solving for a particular realization n for a given factual observation (y_1, y_2) results in the following dependence on ϕ:

$$
n_1 = y_1; \qquad n_2 = \begin{cases} \frac{y_2}{y_1} & y_1 < 0.5 \\ \frac{y_2}{y_1} - \phi + \mathbb{1}_{y_2 < \phi y_1} & y_1 \geq 0.5 \end{cases}
$$

That is, depending on the value of y_1 we either observe a reparameterised version of n_2 or n_2 directly. In particular, if $y_1 < 0.5$ is observed, the noise posterior is independent of the parametrization, yet different parametrization will lead to different interventional predictions when interventions are applied in the $y_1 > 0.5$ regime. Due to this dependence, all these SCMs have different counterfactual distributions, as illustrated in Fig. 8. Here, the graphical causal model on the left contains a free parameter ϕ that characterises the way the influence of exogenous noise affects the SCM. The parameterisation is chosen such that each evaluation of such

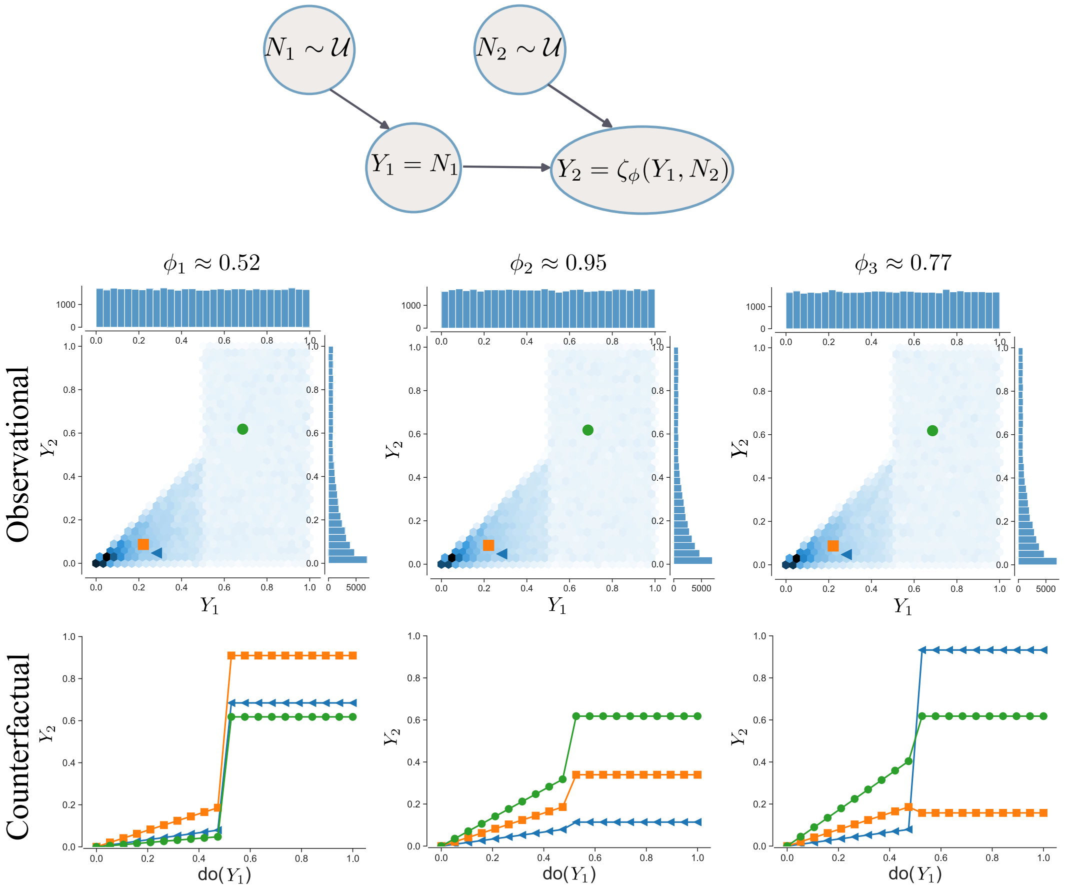


Figure 8: Illustration of the SCM in Eq.2

a representational parameter ϕ_1, ϕ_2, ϕ_3 leads to the same observational distribution of Y_1, Y_2 (top row in Fig. 8) when marginalising out N_1, N_2. The conditional of this observational distribution $p(Y_2|Y_1)$ coincides with the interventional distribution $p(y_2|do(Y_1) = y_1)$ due to the simple generating SCM in which Y_1 corresponds to the root-node. For the three different observations indicated by the markers in the upper row, we constructed the counterfactual distributions (three lines, lower row) for three different representational parameters ϕ_1, ϕ_2, ϕ_3. Although the observational distributions are identical, the bottom row shows different counterfactual distributions corresponding to the three SCMs. As the observational distributions are identical across all parameterisations, the exact SCM cannot be recovered even if an infinite number of data were available. In practice, this is further complicated by the limited amount of data that can be obtained, from which the functional couplings and interaction with the exogenous noise distributions would have to be estimated.

To this end, consider $i = 1, \ldots, N$ observations $(y^j)_i, Y^j = \{(y^j)_i\}_{i=1\ldots,N}$ from a node j as well as the corresponding observations $\mathbf{Y}_{\mathrm{pa}(j)} = \{(y_{\mathrm{pa}(j)})_i\}_{i=1,\ldots,N}, (y_{\mathrm{pa}(j)})_i \in \mathbb{R}^{\dim(\mathrm{pa}(j))}$ from the parent nodes $\mathrm{pa}(j)$. To model their relationship, we use the following generative model.

Definition 4.3.1 (Bayesian Warped GP (BW-GP))**.** *Given kernel parameter θ and a distribution over parametrizations p_ϕ, we refer to the following as a Bayesian Warped GP:*

$$Y^j = g_\phi^{-1}(f(Y_{pa(j)}) + N^j, Y_{pa(j)}), \quad f \sim \mathcal{GP}(\mu_{\mathcal{GP}}, k_{\theta^j}), \quad N^j \sim \mathcal{N}(0, \sigma^j), \quad \phi \sim p_\phi \tag{3}$$

Here, g_ϕ is a parametrized mapping, in this work modelled by a Normalizing flow, which is bijective w.r.t. Y^j for all $Y_{\mathrm{pa}(j)}$. This renders the model similar to the post-nonlinear causal model [78]. The possible parametrizations within the model are represented by the Bayesian belief p_ϕ.

Note that a BW-GP is equivalent to a transformed Gaussian Process with $\mathbb{G} = \mathbb{I}, \mathbb{T} = g_\phi$ within the notation of [30]. By inverting the bijective mapping g_ϕ w.r.t. its first argument, we transform the likelihood (not the prior) of a Gaussian process. As g_ϕ is non-linear $Y^j = g_\phi^{-1}(f(Y_{pa(j)}) + N^j, Y_{pa(j)})$ is *non-Gaussian* with *non-additive noise* [30]. By allowing for a non-linear warping using a normalising flow, this Gaussian distribution can be mapped to any other distribution of the same dimension arbitrarily well (under some mild regularity assumption, see [79]), provided that the neural network is sufficiently flexible. To learn such a model, we employ mean-field variational inference. More precisely, using $q_\phi = \mathcal{N}(m, \mathrm{diag}(s))$ as a variational approximation to the true posterior $p_\phi(\cdot|\mathbf{Y}_{\mathrm{pa}(j)}, Y^j, \theta)$, we optimise the following stochastic approximation (using S samples) to the evidence lower bound (ELBO) [30]:

$$\begin{aligned}\mathcal{L}(m, s, \theta) &= \mathbb{E}_{q_\phi}\left[\log\left(p(Y^j|\mathbf{Y}_{\mathrm{pa}(j)}, \phi, \theta)\right)\right] - \mathrm{KL}\left[q_\phi||p_\phi\right] \\ &\approx \frac{1}{S}\sum_{\phi_i \sim q_\phi} \log\left(p(Y^j|\mathbf{Y}_{\mathrm{pa}(j)}, \phi_i, \theta)\right) - \mathrm{KL}\left[q_\phi||p_\phi\right] \end{aligned} \tag{4}$$

Here, the marginal likelihood for a fixed transformation g_ϕ is given by (see also [29]):

$$\begin{aligned}\log\left(p(Y^j|\mathbf{Y}_{\mathrm{pa}(j)}, \phi, \theta)\right) &= \frac{1}{2}\log|\mathbf{K}_\theta| + \frac{1}{2}\mathbf{z}^\top \mathbf{K}_\theta{}^{-1}\mathbf{z} \\ &\quad - \sum_i \log\left|\frac{\partial g_\phi}{\partial x^j}\left(x_i^j, x_i^{\mathrm{pa}(j)}\right)\right| + \frac{N}{2}\log(2\pi), \\ &\quad \text{with} \quad \mathbf{K}_{\theta^j} = \left(k_\theta\left(\mathbf{Y}_{\mathrm{pa}(j)}, \mathbf{Y}_{\mathrm{pa}(j)}\right) + \sigma\mathbb{1}\right); \\ &\quad \mathbf{z} = \left(g_\phi\left(Y^j, \mathbf{Y}_{\mathrm{pa}(j)}\right) - \mu_{\mathcal{GP}}\left(\mathbf{Y}_{\mathrm{pa}(j)}\right)\right)\end{aligned}$$

The ELBO in Eq. 4 is a lower bound on the observational data distribution as a function of the parameters m, s, θ, where m and s are the mean and variance of the variational approximation q, ϕ while θ summarises parameters from the Gaussian process and therefore enters the first likelihood term only. Once we have obtained an approximate posterior distribution q_ϕ and kernel parameters θ by optimizing the ELBO Eq. 4, we can also perform predictions using the generative model Eq. 3. Specifically, as the generative model is a Gaussian process for any fixed transformation within the transformed space, we first sample parameters $\phi \sim q_\phi$. Using this fixed transformation, we can sample a function and noise values on any given test input and transform the sampled observation back into the original space [29].

The resulting process is a hierarchical Bayesian model in which the distribution q_ϕ determines the different noise parameterisations, and conditioned on this transformation, the residual uncertainty associated with a limited amount of data is captured by a Gaussian process. In [44] Gaussian Processes have also been used to model an SCM under imperfect knowledge. This allows for calculating counterfactual distributions and hence enables us to analyse the potential outcome of different decisions even when the functional couplings between the causal variables are not fully known. However, Gaussian processes fail to model non-Gaussian exogenous noise distributions for transitions between two causally linked variables $X \to Y$.

In contrast, Normalising Flows [80] offer an alternative that can model complex densities while maintaining analytical tractability for density evaluation and sampling. The combination of a Gaussian process with a normalising flow has already been pursued in [30]. However, they have not previously been used for the purpose of calculating counterfactual distributions. Exploiting the Gaussian process property for a fixed transformation in the hierarchical Bayesian model, we can use and extend the result Prop. 4.1.1 on calculating counterfactual SCMs for GPs to derive a sampling procedure for the counterfactual distribution of a Bayesian warped GP.

Proposition 4.3.1 (Noise posterior distribution of a BW-GP). *Let $\mathcal{M}$ be an uncertain*

SCM in which the functional couplings are distributed according to a BW-GP. For an observed factum Y^F with $y^{j,F}, y^F_{pa(j)}$ containing descendent and parent observations according to the graph $\mathcal{G}$, training data $\boldsymbol{Y}^j$, $\boldsymbol{Y}_{pa(j)}$, the noise posterior is given by:

$$
\begin{aligned}
&P(N^j|Y^F) = \int_\phi \mathcal{N}(\mu^j(\phi), s^j(\phi)) q_\phi(\phi) \mathrm{d}\phi, \text{ with} \\
&\mu^j(\phi) = \sigma_j^2 \left(\boldsymbol{K}^j \left(g_\phi(\boldsymbol{Y}, \boldsymbol{Y}) - \mu_{\mathcal{GP}}(\boldsymbol{Y})\right)\right)_{N+1}; \quad s^j(\phi) = \sigma_j^2 \left(\mathbb{1} - \sigma_j^2 \boldsymbol{K}^j\right)_{N+1,N+1} \\
&\boldsymbol{K}^j = \left(k_{\theta^j}(\boldsymbol{Y}, \boldsymbol{Y}) + \sigma_j^2 \mathbb{1}\right)^{-1}; \quad \boldsymbol{Y} = (\boldsymbol{Y}_{pa(j)}, y^F_{pa(j)}); \, \boldsymbol{Y} = (\boldsymbol{Y}_j, y_j^F)
\end{aligned}
\tag{5}
$$

where $N+1$ is the last entry, i.e., the index of the factum when concatenated with the training data $\boldsymbol{Y}_{pa(j)}$, $\boldsymbol{Y}_j$.

Proof The statement follows from the fact that for a given transformation, which is specified by ϕ, $g_\phi((\mathbf{Y}_j, y_j^F), (\mathbf{Y}_{\mathrm{pa(j)}}, y^F_{\mathrm{pa}(j)})) - \mu_{GP}((\mathbf{Y}_{\mathrm{pa(j)}}, y^F_{\mathrm{pa}(j)}))$ is distributed according to a zero-mean Gaussian with covariance given by $k_{\theta^r}((\mathbf{Y}_{\mathrm{pa(j)}}, y^F_{\mathrm{pa}(j)}), (\mathbf{Y}_{\mathrm{pa(j)}}, y^F_{\mathrm{pa}(j)}))$. The rest follows by applying Prop. 4.1.1. Equation (5) also directly gives us a way to approximate the noise posterior by first sampling ϕ from the variational approximation q_ϕ and subsequently sampling a latent function and corresponding observational noise. To sample from the counterfactual distribution, similarly to [44], we average across latent functions, but also across different parameterisations as modelled by $p(\phi)$. Specifically, by exploiting $p(f^j(y^*), N^j|\phi, y^*, Y^F) = p(f^j(y^*)|\phi, y^*, Y^F)p(N^j|\phi, Y^F)$, we can first sample from the predictive distribution of the BW-GP and add a sample from the noise distribution according to Equation (5) in order to get a sample from the counterfactual distribution in which we intervened on the parent node of j and estimate its effect for the observed factum Y^F. Note that the noise posterior depends on the transformation ϕ only via the transformed values for the descendant nodes. Consequently, the variance and especially the inverse of the kernel matrix can be computed beforehand and independently for all samples of the counterfactual distribution. Calculating the counterfactual distribution for the BW-GP only requires averaging across additional samples for the parameters of the normalising flow, for which also the training data have to be transformed. Consequently, although most causal reasoning methods, including algorithmic recourse, do not scale well due to the large number of possible intervention sets, the present method only adds linear computational effort compared to the GP-SCM due to the additional samples of different parameterisations.

4.4 Experiments

In the following, we evaluate our Bayesian Warped GP model (Eq .3) on the illustrative example (Eq. 2) as well as on an algorithmic recourse benchmark. In these experiments, we represent the bijective mapping g_ϕ by a neural spline flow with element-wise (referred to as bins) rational conditional spline functions [81, 82] and use an independent normal prior p_ϕ on the network weights.

4.4.1 Illustrative example

First, we analyse our proposed hierarchical Bayesian model w.r.t. its ability to cope with the inherent ambiguity of different parametrizations leading to the same interventional but different counterfactual distributions by learning a BW-GP on data arising from the SCM of Eq. 2 (see also Fig. 8). To also account for probing the learnt model in sparsely covered regimes of the training data, we selected 174 training points, all of which were within $[0, 0.6]$, but also tested the model in the regime $[0.6, 1]$. On these training datapoints, we fitted both a BW-GP as well as a Gaussian process. To assess the quality of the modelled SCM, we generated 1000 samples of X_1 uniformly across the range $[0, 1]$ and drew one sample from the modelled interventional distribution. The resulting predictive distribution of the BW-GP (left) and GP (right) is illustrated in Fig. 9. Both models are trained on points between 0 and 0.6, rendering the range between 0.6 and 1 as an extrapolation regime. The blue points in the

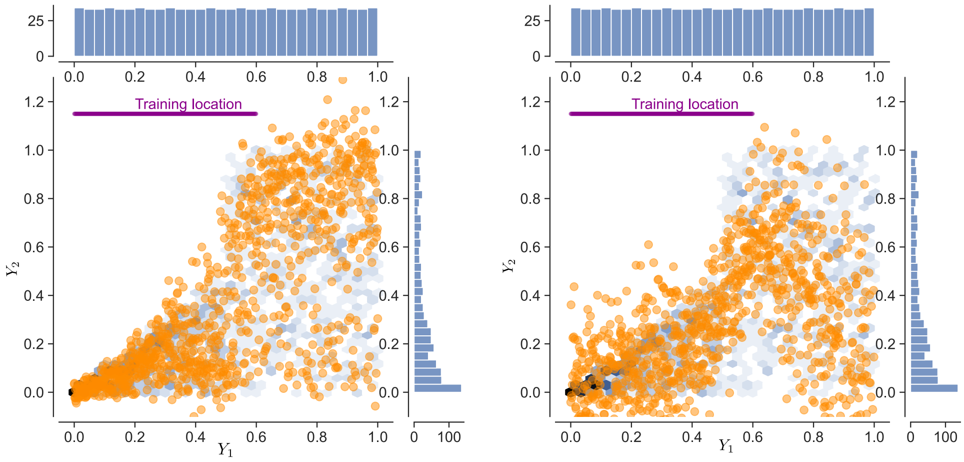


Figure 9: The predictive distribution of the BW-GP (left) and the GP (right) of the modelled interventional distribution.

background show the observational distribution of the SCM Fig .2, and the orange points correspond to samples of the interventional distribution of our BW-GP (left) and a GP (right). The blue points in the background indicate samples from the ground truth model of Eq. 2 (see Fig. 8). As can be seen in Fig. 9, the BW-GP provides a close fit to the ground-truth observational distribution, whereas a GP is not able to fit the observational data as accurately, due to the non-stationary noise distribution. This heteroscedasticity of the noise distribution also forces the plain GP to explain the data using non-zero functional coupling uncertainty. The BW-GP model, however, nicely adjusts for such uncertainty by allowing for non-stationary distributions over functional couplings.

Second, we also evaluate the counterfactual distribution for both a Gaussian process without parametrisation uncertainty and our BW-GP which includes such uncertainty. In Fig. 10, we plot the resulting counterfactual distribution estimates when intervening in Y_1 and using the noise posterior of the observation $Y_1^F = 0.22, Y_2^F = 0.08$ (marked by an orange square in Fig. 8) and compare them against counterfactual distributions arising from different parametrisations in Eq. 2. Here, the blue points in the background show samples of the true counterfactual distribution constructed

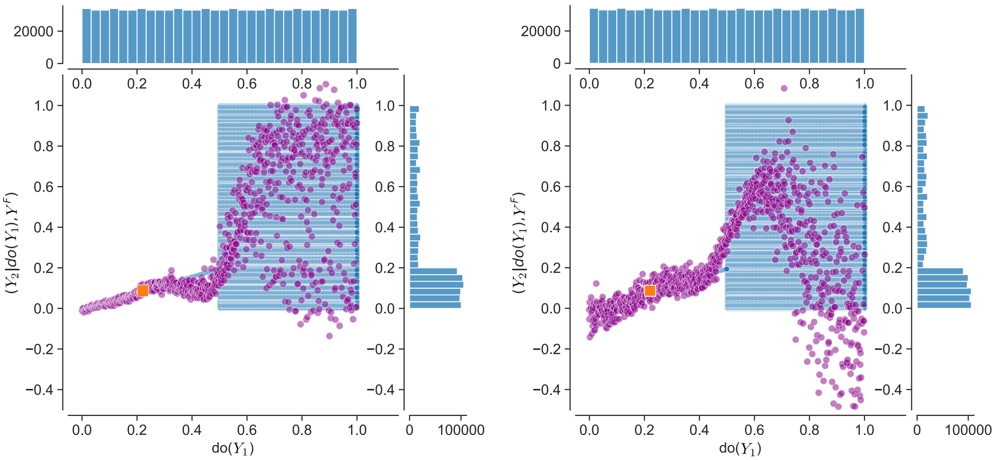


Figure 10: Illustration of the modelled counterfactual distribution of our BW-GP (left) and a GP (right).

from the factum (orange box) and varying parametrizations ϕ. The purple points represent a sample drawn of the counterfactual distribution of our BW-GP (left) and a GP (right). The interventional distribution of the counterfactual SCM (as shown in Fig. 10) is forced to recover the observation on which it is conditioned, if we would intervene in Y_1, forcing the variable to have the same value as observed (orange marker in Fig. 10). Although this property is recovered by both BW-GP and GP (by the construction of the counterfactual SCM), the stationarity assumption of the noise of the GP results in larger uncertainty around the observation in the counterfactual. Despite the non-stationarity of the noise of the BW-GP, it seems to also cover the uncertainty of counterfactual distribution in the out-of-training data regime. We focus on isolating the impact of uncertainties that arise from the inherent ambiguity of different parameterisations of the same observational and interventional distributions.

4.4.2 Benchmark Experiments

Besides the illustrative example, we evaluated the BW-GP on an important downstream task of a counterfactual distribution to assess the impact of the BW-GP on a more realistic decision-making process. To this end, we compare our model (BW-GP) against other baseline methods within the algorithmic recourse benchmark of [6], including a standard GP, a linear regressor and a conditional variational autoencoder (CVAE). For the CVAE, we use the implementation of [44], yet it can be regarded as a non-amortised version of the CVAE by [83]. Analogously to [6], we compared both the counterfactual model (denoted by $\mathcal{M}_{\texttt{<model>}}$) as well as the interventional variants of the different models (denoted by $\text{CATE}_{\texttt{<model>}}$).

Table 1: Experimental results of a three variable causal model in a recourse setting with 100 individuals. We compare our model ($\mathcal{M}_{\text{BW-GP}}$) against the reproduced baselines `LIN`, `GP`, `CVAE` of [6].

	Linear SCM			NON LINEAR SCM			NON ADDITIVE SCM		
	`Valid`(%)	Cost(%)	`MMD`	`Valid`(%)	Cost(%)	`MMD`	`Valid`(%)	Cost(%)	`MMD`
$\mathcal{M}_*$	100	11.2 ± 7.4	-	100	19.7 ± 12.3	-	100	10.3 ± 8.6	-
$\mathcal{M}_{\text{LIN}}$	100	12.0 ± 8.0	$0.019 \pm 2.37 \cdot 10^{-5}$	67	20.6 ± 10.8	0.202 ± 0.006	100	10.1 ± 8.3	0.383 ± 0.027
$\mathcal{M}_{\text{GP}}$	100	13.3 ± 9.7	0.043 ± 0.001	100	22.0 ± 13.5	0.036 ± 0.001	98	10.3 ± 8.5	0.369 ± 0.019
$\mathcal{M}_{\text{CVAE}}$	100	12.7 ± 8.2	0.031 ± 0.001	91	25.4 ± 14.3	0.139 ± 0.002	97	10.1 ± 8.1	0.146 ± 0.013
$\mathcal{M}_{\text{BW-GP}}$	100	13.0 ± 9.0	0.069 ± 0.002	99	22.3 ± 14.7	0.043 ± 0.001	99	10.2 ± 9.1	0.120 ± 0.009
CATE_*	88	12.4 ± 9.6	-	99	28.1 ± 28.9	-	100	10.1 ± 8.2	-
CATE_{GP}	90	12.6 ± 8.5	0.044	97	27.4 ± 17.8	0.043	94	9.6 ± 8.5	0.261
$\text{CATE}_{\text{CVAE}}$	87	12.8 ± 10.5	0.066	99	33.4 ± 25.0	0.069	100	10.1 ± 8.3	0.064
$\text{CATE}_{\text{BW-GP}}$	93	12.8 ± 9.0	0.073	98	29.8 ± 19.4	0.039	98	9.7 ± 7.9	0.089

Table 2: Experimental results of a seven variable semi-synthetic causal model on 100 facta in a recourse setting.

	LINEAR LOG. REGR.			NON-LINEAR LOG. REGR.			RANDOM FOREST		
	$\texttt{Valid}_*$(%)	Cost(%)	`MMD`	$\texttt{Valid}_*$(%)	Cost(%)	`MMD`	$\texttt{Valid}_*$(%)	Cost(%)	`MMD`
$\mathcal{M}_*$	100	17.4 ± 8.0	-	100	15.8 ± 9.3	-	100	19.3 ± 9.1	-
$\mathcal{M}_{\text{LIN}}$	100	18.0 ± 8.3	0.121 ± 0.007	96	16.2 ± 9.5	0.101 ± 0.009	94	19.5 ± 9.4	0.094 ± 0.007
$\mathcal{M}_{\text{GP}}$	100	22.0 ± 8.7	0.128 ± 0.004	100	18.6 ± 10.4	0.042 ± 0.001	100	21.2 ± 9.4	0.040 ± 0.001
$\mathcal{M}_{\text{BW-GP}}$	100	22.3 ± 9.3	0.050 ± 0.002	100	19.6 ± 12.1	0.053 ± 0.002	99	20.7 ± 9.2	0.049 ± 0.002
CATE_*	88	25.7 ± 9.3	-	89	21.4 ± 14.2	-	92	23.9 ± 9.0	-
CATE_{GP}	91	26.6 ± 9.5	0.082	93	22.3 ± 14.8	0.088	98	24.5 ± 9.5	0.086
$\text{CATE}_{\text{BW-GP}}$	95	28.1 ± 11.1	0.090	94	22.5 ± 14.4	0.087	98	24.6 ± 9.4	0.077

In this algorithmic recourse benchmark setting, the goal is to find both the optimal nodes for an intervention as well as the optimal intervention value in relation to the cost (Eq. 1). We report validity and cost of [6], where the validity defines the percentage of individuals with a beneficial outcome after a counterfactual sample is drawn. The cost is the L2-norm between the factum Y^F and the intervention, normalised by the range of each training variable. To assess the quality with which we represent the counterfactual distribution, not just the algorithmic recourse task, we additionally, evaluate the maximum mean discrepancy (MMD) [84] between the modelled counterfactual distribution and the counterfactual distribution of the ground truth model (denoted as $\mathcal{M}_*$). As both depend on the observed factum, we average the obtained MMD values across 100 facta. More precisely, to generate a sample of the modelled counterfactual distribution, we first calculate the posterior noise distribution and then perform a soft intervention on the root node by sampling values for the root node from the ground truth distribution. Using this sampling process, we obtain one counterfactual sample per factum. The same sampling process is used to evaluate the quality of the modelled interventional distribution in terms of MMD value, however, the noise prior is used instead of the noise posterior per factum to generate a sample. To generate samples from the counterfactual distribution of the ground truth SCM, we stored the noise variables $\mathbf{N}$ that generated a particular Y^F in the test data and substituted it into the structural equations of the SCM after performing an intervention. In order to use the same MMD metric across different models, we used a squared exponential kernel and used two independent samples of the ground truth distribution to estimate hyperparameters of the kernel according to the median heuristic [85].

4.4.3 Interventional distribution

In this section, we provide additional plots illustrating the properties of the different models visually. Due to the complex yet low-dimensional setting, the non-additive SCM of the three-variable model (see Tab. 1)) is of particular interest. Since this was visually not the case for the other SCMs, we do not explicitly show them.

In Fig. 11, we plotted the ground truth distribution (see Fig. 11 (left)) and in (Fig. 11 (right)) the corresponding distribution as modelled by the BW-GP. To generate samples from the different models, we generated samples from the ground truth model for the variable of the root-node X_1. Using the different models for the conditional

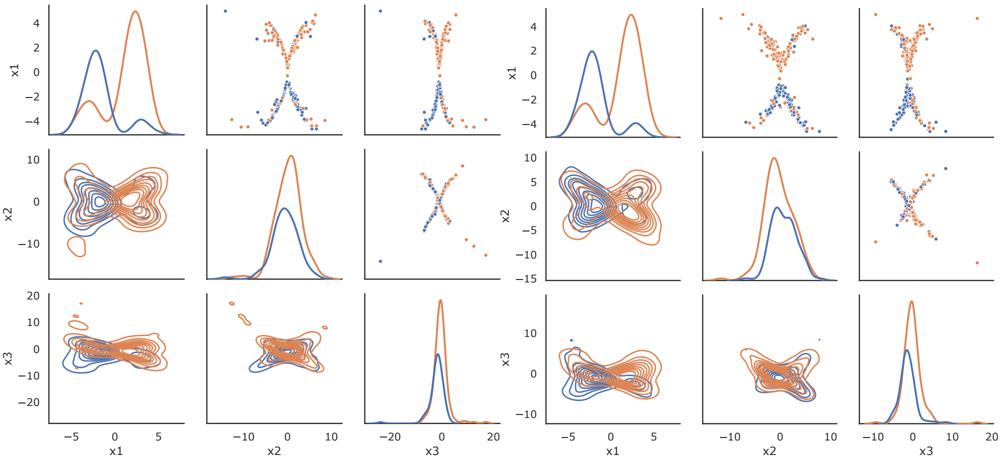


Figure 11: On the left, we plot the ground truth, and on the right, the Bayesian Warped Gaussian Process model. The colouring corresponds to the classes the classifier yields in the recourse task (blue are the negative and orange are the positive classified points).

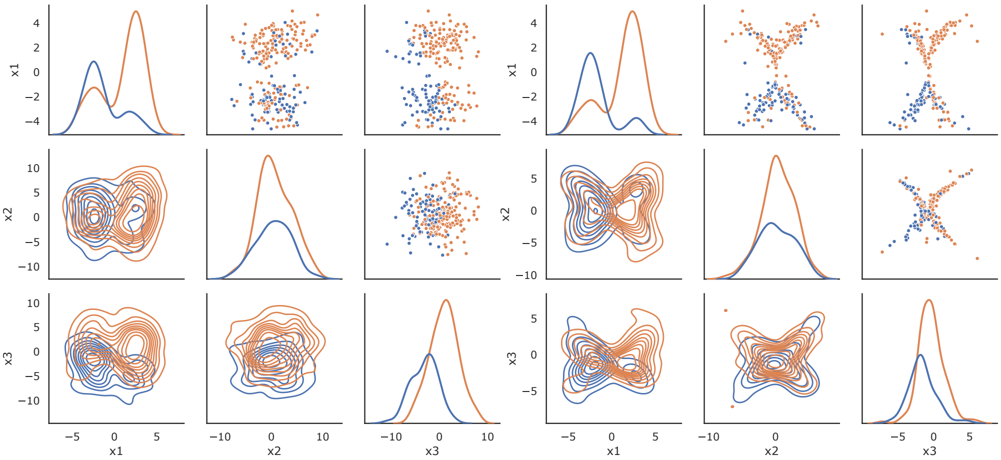


Figure 12: On the left, the Gaussian Process is plotted, and on the right, the CVAE model. The colouring corresponds to the classes the classifier yields in the recourse task (blue is the negative and orange are the positive classified points).

distributions of children given to the parents, we generated the remaining variables X_2, X_3 according to the causal graph. As also indicated by the small MMD-Values (Tab. 1), the BW-GP also visually fits the ground truth data much better than the GP (see Fig. 12 (left)), which learns two Gaussian distributions for the multimodal distribution. While the CVAE in (Fig. 12 (right)) fits the data also better than the

GP, it exhibits a higher variance than the BW-GP, which is also reflected by a slightly larger MMD-value.

Synthetic three variable causal model

First of all, we evaluate our model on three SCMs, a linear and a non-linear both with additive noise and a nonlinear with non-additive noise. Each SCM has the same underlying causal graph consisting of three variables, yet differs in the functional couplings being either linear, non-linear or exhibiting non-additive noise. Since the ground truth is known of this artificial, we can generate data from it. Analogously to [6], we trained each model on 250 such samples from the observational distribution and evaluated 100 facta sampled from the observational distribution that are found to be negatively classified according to a logistic regression, see Tab. 1. Here, $CATE_*$ refers to the optimisation process in which interventions are evaluated with respect to the interventional SCM rather than the counterfactual SCM within Eq. 1 (in the constraint set); see [6]. Therefore, interventions found by $CATE_*$ in Tab. 1,2 are not necessarily achieving 100 per cent validity when checked with counterfactual ground truth SCM. To set hyperparameters for our model (number of bins in the spline and size of the neural network), we performed a Bayesian optimisation on a validation set (details can be found in the Appendix; see Chapter 8). Although the BW-GP performs comparably in terms of costs and validity as the other best models, on the non-additive SCM we show a significantly smaller MMD than the GP in the counterfactual and interventional (CATE) task. This could be due to the fact that the normalising flow is able to learn multimodal distributions well. Nevertheless, the GP achieves high validity and comparable loss, which means that the learnt conditional distributions do not have a strong impact on the recourse task itself. The conditional variational auto-encoder (CVAE) performs similarly well on the non-additive SCM but operates considerably worse on the non-linear SCM counterfactual task. As noted in [6], samples of $\mathcal{M}_{\texttt{CVAE}}$ are "pseudo-counterfactual", possibly amounting to a reduced accuracy.

In the linear SCM experiment, we observe that the BW-GP performed slightly worse than the GP in terms of the MMD, yet without significant impact on validity or costs. Note, however, that the costs and validity are computed based on a counterfactual distribution which is constructed from a single ground truth SCM and hence does not include the additional uncertainty of potentially different parametrizations. We argue that the slight drop in performance of the MMD can, therefore, be attributed to the additional uncertainties accounted for by the BW-GP. Therefore, we additionally measured the variance of the counterfactual distribution samples over the different facta to assess a potential increase in the overall uncertainty of the counterfactual distribution modelled by the different methods. Indeed, we observed that our model has the highest variance (2.9907) across counterfactual distribution samples, followed by the GP (2.9531), the linear model (2.9271), the CVAE (2.9189).

Semi synthetic seven variable causal model

The semi-synthetic seven variable system is inspired by the German Credit UCI dataset as it features relevant variables such as age, savings, gender, etc., as well as a labelling mechanism representing the loan-approval. Based on data generated from this constructed SCM, different classifiers are trained: linear and non-linear logistic regression, and a random forest. Similarly to the three-variable model we used the same benchmark setting, models and computation as in [6] and performed a hyper-parameter optimization on a validation set. In this more realistic and higher-dimensional setting, we also observe a more accurate characterisation of the counterfactual distribution, as indicated by significantly lower MMD scores without sacrificing validity (see Tab. 2)[1]. Although the BW-GP performs slightly worse than the GP in terms of the accuracy of the interventional distribution for the

---

[1] Note that the results for the variational auto-encoder could not be reproduced with the provided source code.

logistic regression setting, it still achieves better validity. Similarly to the evaluation within the three-variable model, each method is only evaluated against a single SCM assumed to be the ground truth. However, the BW-GP additionally accounts for the uncertainty in the parameterisation, leading to a larger spread of counterfactual costs as indicated by the standard errors, yet without sacrificing validity.

4.5 Conclusion

In this chapter, we proposed a hierarchical Bayesian model to account for ambiguities in the underlying SCM as well as for the uncertainties arising from imperfect knowledge of functional couplings due to limited observational data. By using a Bayesian Warped GP, we were able to allow not only non-Gaussian distribution at descendent nodes but also non-stationary noise distributions. This seems to be particularly beneficial for counterfactual distributions (see Figure 10). Although we introduced an additional source of uncertainty about the parameterisation, this resulted in a more accurate fit of the counterfactual distribution also in more realistic settings (see Table 12).

In summary, we presented a method which allowed us to (i) capture uncertainty about the parametrisation of an SCM additionally to the uncertainty in the functional couplings and exogenous noise uncertainty about continuous variables; (ii) derive a counterfactual distribution in this extended setting; and (iii) investigate the impact of modelling additional uncertainties on an important downstream task of algorithmic recourse.

The gained expressiveness of the model also leads to robust recourse actions in terms of the achieved validity without an increase in costs due to the additional uncertainty within considered SCMs. In fact, our BW-GP (Eq. 3) theoretically provides a sufficiently flexible model to capture any conditional distribution $p(Y^j|Y_{\mathrm{pa}(j)})$. However, in practice, the flexibility of the neural network, as well as the amount of observational data, is limited. The proposed method could also be used in settings with unobserved confounders by introducing additional, yet unobserved, nodes within the SCM and integrating out their values during the training phase. However, when falsely assuming potential hidden confounders by introducing latent variables, each of which is associated with a flexible probability distribution, predictive power is likely to decline. Although we have shown that the proposed model can account for ambiguities to a certain degree, it still contains hard and soft assumptions, which could be relaxed. For example, in this research, we assumed that the graphical structure between the static-modelled variables is known. By imposing yet another probability distribution on the graphical structure, such a hard assumption could be relaxed with the downside of additional computational complexity to learn these models [86].

5 Counterfactual-based Root Cause Analysis in Dynamic Systems

Explaining unexpected behaviour in terms of underlying causes is a difficult challenge with a broad range of applications. Such applications range from identifying potential problems in industrial processes to understanding the factors influencing anomalous weather phenomena. For example, within an assembly line of an industrial manufacturing plant, faster identification of root causes of increased scrap rate (the rate at which assembled products fail quality assessment audits) can minimise cost, increase production yield, and increase overall efficiency. If one can observe sufficiently many instances of anomalous behaviour or faulty traces of a process, one option would be to perform a correlation-based analysis or causal discovery [87], thereby estimating the influencing factors to the variable "fault" [50, 88, 7]. Alternatively, causal inference can be used even if only one anomalous observation is available [49, 3]. Here, the identification of root causes is formulated in terms of a counterfactual query: "Would the observed failure also occur if we had replaced the behaviour of a sub-system at a certain point in time with its *normal* behaviour?".

To this end, a formal description of the behaviour of the full system is needed in which such counterfactual questions can be answered. However, existing causal methods for root cause identification are typically limited to static settings and focus on additive external influences causing failures rather than structural influences; see Section 3.2. In this chapter, we address these problems by modelling the dynamic causal system using a Residual Neural Network and deriving corresponding counterfactual distributions over trajectories. We show quantitatively that more root causes are identified when an intervention is performed on the structural equation and the external influence, compared to an intervention on the external influence only. By employing an efficient approximation to a corresponding Shapley value, we also obtain a ranking between the different subsystems at different points in time being responsible for an observed failure, which is applicable in settings with a large number of variables. We illustrate the effectiveness of the proposed method on a benchmark dynamic system as well as on a real-world river dataset.

We briefly restate and specify the notation used throughout this chapter. Following the notation of Peters et al. [8], we denote the sequence of observations of the system of interest by d-variate time series $(\mathbf{Y}_t)_{t\in\mathbb{Z}}$, where each $\mathbf{Y}_t$ for fixed t is the vector $(Y_t^1, ..., Y_t^d)$. Each $\mathbf{Y}_t^j$ represents the jth observable of a system at time t. By some abuse of notation, if we omit superscripts or subscripts, we refer to the full time series. That is, $\mathbf{Y} = (\mathbf{Y}_t)_{t\in\mathbb{Z}}$, $\mathbf{Y}^j = (Y_t^j)_{t\in\mathbb{Z}}$, and $\mathbf{Y}_t = (Y_t^j)_{j\in\{1,...,d\}}$. The full-time causal graph $\mathcal{G}_t$ with a node for each time point and the signal Y_t^j for $(j,t) \in 1, ..., d \times \mathbb{Z}$ theoretically has infinitely many nodes and is assumed to be acyclic, while the summary graph $\mathcal{G}$ with nodes $Y^1, ..., Y^d$ may be cyclic.

5.1 Interventional Dynamic SCM

As each SCM (interventional or not) defines structural equations and noise distributions, it can generate a trajectory of observations. We denote the distribution of the observations generated by the SCM as $P_{\mathcal{M}}$ and the distribution of the observations generated by the intervened SCM as $P_{\mathcal{M}_{\mathcal{J}}}$.

Definition 5.1.1. *(Interventional Dynamic SCM) Let $\mathcal{J}$ be a set of interventions in which each element ξ can be of the following form:*

$$\xi := do(P_{N_t^j}) = \tilde{P}_{N_t^j}, \quad (6) \qquad or \qquad \xi := do(S_t^j) = \tilde{S}_t^j \quad (7)$$

where $\tilde{P}_{N_t^j}$ *is a new noise distribution and* $\tilde{S}_t^j$ *is a new structural equation for the node* j *at time* t*. The interventional dynamic SCM is then defined by replacing the noise distribution or structural equation within a given dynamic SCM (see Definition 2.2.2)* $\mathcal{M}(\mathcal{S}_t, P_{N_t}, \mathcal{G}_t)$*. Here, Eq. 6 denotes a soft intervention on the noise distribution whereas Eq. 7 denotes an intervention on the structural equation. We then denote the intervened dynamic SCM resulting by* $M_{\mathcal{J}}(S_t, P_{N_t}, G_t)$*.*

5.2 Abducted and Counterfactual SCM

Given an observed trajectory, we can now also define the counterfactual distribution describing hypothetical trajectories that would have been observed if an (alternative) intervention had been performed. Let $\mathbf{Y}^F$ be an observed trajectory and $\mathcal{M}$ a given dynamic SCM (see Definition 2.2.2). In order to construct a counterfactual dynamic SCM, we define the noise posterior distribution $P_{N_t^j}(N_t^j|\mathbf{Y}^F) = \delta(N_t^j - N_t^{F,j})$ by:

$$N_t^{F,j} = -Y_{t-1}^{F,j} - f^j(Y_{t-1}^{F,PA(j)}, Y_{t-1}^{F,j}) + Y_t^{F,j} \tag{8}$$

where f^j is the structural equation of the node j and $PA(j)$ are the parents of the node j according to the summary graph $\mathcal{G}$. The resulting dynamic SCM, in which the noise distributions $P_{N_t^j}$ are replaced with the above defined noise posterior distributions, is then denoted as $\mathcal{M}^F$ indicating that the noise distributions are abducted from the observed trajectory $\mathbf{Y}^F$. In fact, when generating trajectories from this abducted SCM, it only generates the observed trajectory $\mathbf{Y}^F$ due to the above setting of the noise variables. In order to generate new counterfactual trajectories reflecting alternative outcomes, we need to perform an intervention on this abducted SCM, leading to the counterfactual SCM. That is, given an abducted SCM $\mathcal{M}^F$ and a set of interventions $\mathcal{J}$, we refer to the resulting interventional SCM $\mathcal{M}^F_{\mathcal{J}}$ as the counterfactual SCM. For example, when performing an intervention $do(P_{N^j_t}) = \tilde{P}_{N^j_t}$ on the noise distribution at a specific point in time t and a node j, the counterfactual SCM is defined by the following structural equations:

$$Y^e_s = Y^e_{s-1} + f^e(Y^{PA(e)}_{s-1}, Y^e_{s-1}) + N^e_s, \qquad \text{where} \tag{9}$$

$$N^e_s \sim \begin{cases} \tilde{P}_{N^j_t} & \text{if } s = t \text{ and } e = j \\ \delta(N^j_t - N^{F,j}_t) & \text{otherwise} \end{cases} \tag{10}$$

5.3 Root cause

As we are interested in identifying a root cause, we state here more precisely what we mean by this term. We define a root cause as an intervention according to $M_{\mathcal{J}}(S_t, P_{N_t}, G_t)$ leading to a faulty behaviour. Here, we assume that a faulty behaviour can be detected or defined using a known classifier ϕ. This classifier maps a time series to a binary value, indicating whether the time series is faulty. Such a classifier can either be given as a known test function (e.g. corresponding to an end-of-line test in an assembly line, an assertion in a software system, or a medical diagnosis) or can be learned from data (e.g. an outlier-score function learned on normal data).

Definition 5.3.1. *(Root cause) Given a classifier ϕ that determines whether an observed trajectory is faulty, we refer to a (set of) intervention(s) Ξ to be the root cause of a failure associated with the classifier ϕ, if observations $(\boldsymbol{Y}^F_{t,t=1,\ldots T})_j$ from the interventional SCM $\mathcal{M}_{\{\Xi\}}$ are leading to an increased failure rate:*

$$\mathbb{E}_{\boldsymbol{Y}^F_{t,t=1,\ldots T} \sim \mathcal{M}_{\Xi}}[\phi(\boldsymbol{Y}^F_{t,t=1,\ldots T})] - \mathbb{E}_{\boldsymbol{Y}_{t,t=1,\ldots T} \sim \mathcal{M}}[\phi(\boldsymbol{Y}_{t,t=1,\ldots T})] > 0$$

Note that this corresponds to the average treatment effect of an intervention on the external influence or structural intervention. If the probability of a failure for an external intervention on the noise or structure is higher than without any intervention, we assume that the failure has an underlying root cause.

5.4 Shapley Value

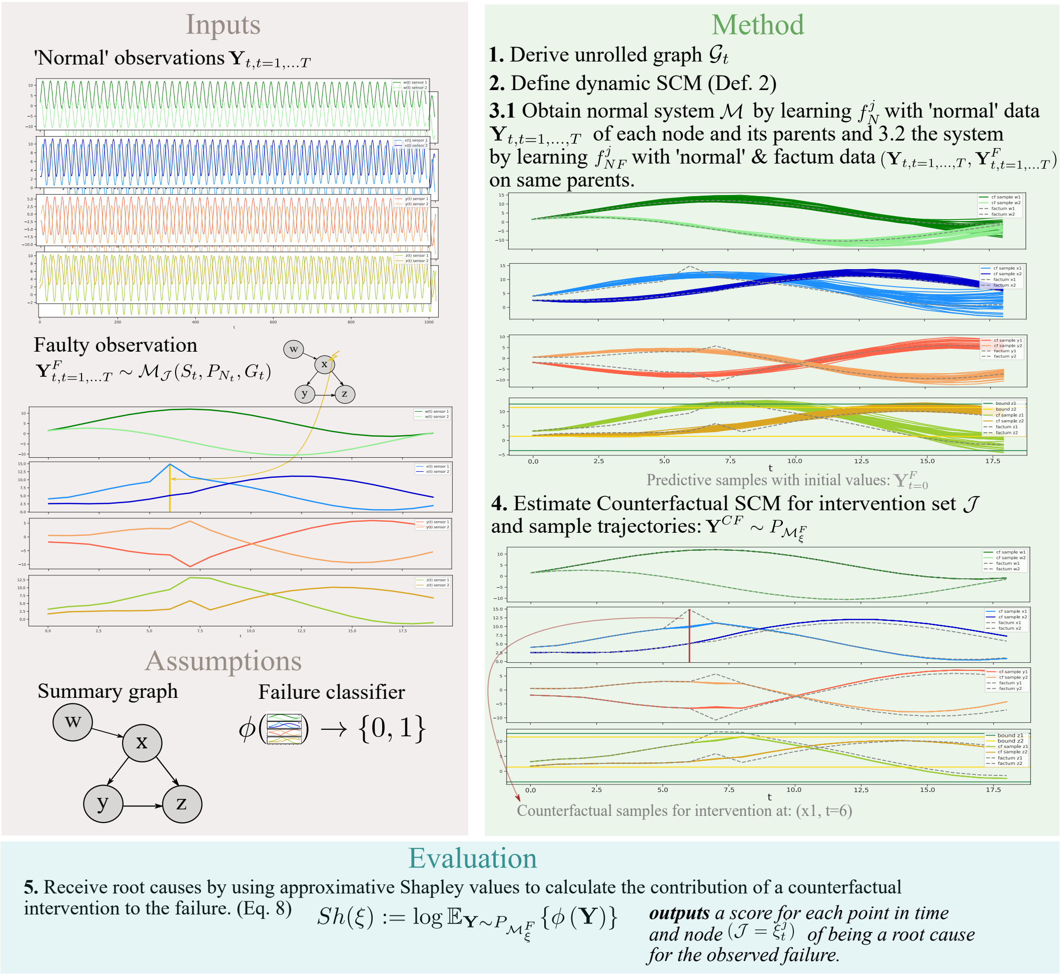
Shapley values, originally defined to quantify the contribution of individual players to the outcome of a game, have been used by Budhathoki et al. [3] in a static setup to define a score for nodes that are potential root causes of an observed fault. To this end, interventions (or possible root causes) are identified with players in a game whose outcome is determined by a value function that quantifies the degree to which a set of interventions can increase the likelihood of correcting a failure (defined below).

Definition 5.4.1. *(Shapley value) The Shapley value [53] of a player i out of a set N of possible players to the outcome of a game characterized by the outcome function v is defined by:*

$$Sh(i) := \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

Note that in order to calculate the Shapley value, one has to sum over exponentially many subsets of the set of possible players. This is feasible only for small sets of players. In the context of root cause analysis within a dynamic framework, the set of players represents the potential interventions, which span all possible times and nodes within the unrolled graph of a dynamic Structural Causal Model (SCM). Due to the exponential growth of the number of possible interventions, exact Shapley value estimation is computationally infeasible for dynamic SCMs, and we have to resort to an approximate version.

5.5 Method for identifying root causes


Inputs
'Normal' observations $\mathbf{Y}_{t,t=1,...T}$
Faulty observation
$\mathbf{Y}^F_{t,t=1,...T} \sim \mathcal{M}_\mathcal{J}(S_t, P_{N_t}, G_t)$
w
x
y
z
t
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Assumptions
Summary graph
Failure classifier
$\phi(\;) \to \{0, 1\}$
Method
Predictive samples with initial values: $\mathbf{Y}^F_{t=0}$
Counterfactual samples for intervention at: (x1, t=6)
Evaluation

1. Derive unrolled graph $\mathcal{G}_t$

2. Define dynamic SCM (Def. 2)

3.1 Obtain normal system $\mathcal{M}$ by learning f^j_N with 'normal' data $\mathbf{Y}_{t,t=1,...,T}$ of each node and its parents and 3.2 the system by learning f^j_{NF} with 'normal' & factum data $(\mathbf{Y}_{t,t=1,...,T}, \mathbf{Y}^F_{t,t=1,...T})$ on same parents.

4. Estimate Counterfactual SCM for intervention set $\mathcal{J}$ and sample trajectories: $\mathbf{Y}^{CF} \sim P_{\mathcal{M}^F_\xi}$

5. Receive root causes by using approximative Shapley values to calculate the contribution of a counterfactual intervention to the failure. (Eq. 8) $Sh(\xi) := \log \mathbb{E}_{\mathbf{Y} \sim P_{\mathcal{M}^F_\xi}} \{\phi(\mathbf{Y})\}$ ***outputs*** *a score for each point in time and node* $(\mathcal{J} = \xi^j_t)$ *of being a root cause for the observed failure.*

Figure 13: This figure shows an overview of the individual steps of our method.

Now that we have the necessary background, we can describe our method for identifying root causes in dynamic SCMs. The method is based on the following steps and is illustrated in Fig. 13. We want to identify the root cause that caused an observed failure in a system. To this end, we cast this problem in a counterfactual query: "Would the observed failure also occur if we had replaced the *faulty* behaviour of a sub-system at a certain point in time with its *normal* behaviour?". To answer this question after we observed a faulty observation $(\mathbf{Y}^F_{t,t=1,...T})_j$, as illustrated in *Inputs* in Fig. 13, we follow the steps of counterfactual distribution calculation: abduction, action and prediction [89]. However, in order to apply those steps, we need an SCM characterising the normal and potentially abnormal systems. To characterise the normal system, we assume that we have access to data representing the normal behaviour of the system, as shown in *Inputs* in Fig. 13. Additionally, we assume that we have at least a summary graph $\mathcal{G}$ of the system. This summary graph can be obtained from expert knowledge or from data. Furthermore, as shown in *Assumptions* in Fig. 13, we assume that we know a function ϕ that classifies an observation as faulty or normal.

Fitting the model: In step 3.1 of the *Method* part in Fig. 13 we obtain the normal behaviour system $\mathcal{M}$ by learning the functions f_N^j with the inputs being normal observations Y_t of each node and its parents of the summary graph $\mathcal{G}$. If for both normal and abnormal data, a node and hence its transition function is not anomalous, the transition function would be identical for both settings. Therefore, in 3.2 we additionally fit a transition function f_{NF}^j with normal and factual data as input to the same parents and children as in 3.1 of the known graph $\mathcal{G}$ and with that we define the SCM $\mathcal{FM}$. We show predictive samples of $\mathcal{M}$ in the graph; see Fig. 13 step 3.2.

Estimating the Counterfactual: In the abduction step, we first infer the noise distribution corresponding to the observed factum. We refer to the abducted SCMs $\mathcal{M}^F$ and $\mathcal{FM}^F$ by applying the factum as function input to f_N^j and f_{NF}^j and constructing the resulting noise posterior distributions as described in Eq. 8. We need to calculate the noise variables for both SCMs separately because function couplings and noise variables are coupled. In the action step, we perform an intervention in $\mathcal{M}$ by $\xi_{\mathcal{M}} := \{do(P_{N_t^j}) = \tilde{P}_{N_t^j}\}$ (see Eq. 6), where we use the prediction error of our model to estimate the Gaussian noise variance:

$$\tilde{P}_{N_t^j} = \mathcal{N}(0, \sigma_{val}^2), \qquad \sigma_{val}^2 = \frac{1}{V}\frac{1}{T}\sum_v \sum_t (Y_{t+1}^{j,v} - f_j(Y_t^{Pa(j),v}))^2 \tag{11}$$

with $Y^{j,v}$ being a validation trajectory of the normal data, and V the number of validation trajectories. For an intervention in $\mathcal{FM}$ we intervene on the noise as before and we additionally intervene on the structure by $do(S_t^j) = \tilde{S}_t^j$ (see Eq. 7), which replaces the previous transition function f_{FN}^j with a new structural equation $\tilde{S}_t^j$ consisting of the transition function f_N^j originating from the "normal" SCM, obtained purely from training data $\xi_{\mathcal{FM}} := \{do(P_{N_t^j}) = \tilde{P}_{N_t^j}, do(S_t^j) = \tilde{S}_t^j\}$. After the construction of the corresponding counterfactual SCM, we can then generate counterfactual trajectories under the different interventions $\mathbf{Y}^{CF} \sim P_{\mathcal{M}^F_{\xi_{\mathcal{M}}}}$, as illustrated in 4. in Fig. 13. If an external influence on node j at time t leads to an abnormal factum, an intervention of this type should remove the abnormal behaviour and, therefore, lead to a normal trajectory.

Evaluation: To quantify how close these counterfactual samples are to normal trajectories, the trajectories are processed using a classifier function ϕ (Eq. 5.5). In turn, we receive a score for each counterfactual sample indicating whether the failure was removed by counterfactual intervention ξ. We then averaged over multiple counterfactual samples. To rank interventions at different times and nodes, we can use Shapley values by identifying players with interventions and matching outcomes by the average normality of the counterfactual sample. Shapley values, however, scale exponentially, and therefore, we use the following simple approximation, which we obtain by ignoring interactions between different interventions, thereby only considering single-ton intervention sets. Although mainly motivated by pure computational tractability, we can alternatively assume that a perfectly synchronous occurrence of multiple root causes is very unlikely, thereby justifying the restriction to single-ton intervention sets. This simplification enables static causal reasoning in a dynamic setup but introduces two key challenges: (1) we cannot capture instantaneous interactions in physical models directly, and (2) multiple-step delays must be represented with intermediate states, leading to larger models. This adds complexity, as each intermediate state introduces additional exogenous variables and functional couplings. Thus, fitting a dynamic SCM becomes more computationally intensive, as each new intermediate state must maintain functional consistency with previous steps while ensuring that causal relationships are preserved across multiple time steps. However, in practice, although physical models may involve instantaneous interactions, causal dependencies typically unfold gradually over time. Therefore, with an appropriately high model resolution (sufficient small-time steps), the unit-level approximation can capture essential dynamic causal behaviour with sufficient accuracy. Secondly, although in this chapter we consider a model trained on one-time step (i.e. a mapping from t to $t+1$ is learned), further trainings could be performed considering larger time windows. This would allow multiple-step delays to be addressed. However, the ability to learn over multiple time steps and the increase in the number of time steps through very small increments have a complementary effect, and their actual compatibility in practice still needs to be studied.

Consequently, we arrive at the following simple expression of the contribution score of individual interventions ξ for each point in time and node:

$$Sh(\xi) := \log \mathbb{E}_{\mathbf{Y} \sim P_{\mathcal{M}_{\xi}^{F}}} \{\phi(\mathbf{Y})\} \tag{12}$$

5.6 Experiments

In the following experiments, we evaluate the effectiveness of the proposed method for different synthetic and real-world data sets. As for synthetic data sets, we consider both linear and non-linear dynamic systems with single-point external failure-causing disturbances as well as a benchmark data set for identifying structural causes for anomalies [7]. For the real-world data set, we are investigating the dynamic water flow rate in rivers [90]. For our synthetic experiments, we perform two meta-experiments, which analyse the influence on the model performance of varying root cause injections and how robust the model is against violating the assumption that the causal graph is known. We denote our models, a linear and a non-linear model both performing a counterfactual intervention on the external noise influence and on the structural equation with *Lin*(S_t^j, N_t^j) and *NLin*(S_t^j, N_t^j). We compare against a linear layer model with counterfactual noise-influence intervention *Lin*(N_t^j), similar to [3] and EasyRCA [7] in the benchmark experiment. For completeness, we additionally provide a non-linear model *NLin*(N_t^j) with counterfactual noise influence intervention. In order to model the nonlinear dynamic SCM, for *NLin* we use a simple three-layer residual neural network (ResNet) with hyperbolic tangent activation functions and 128 neurones as latent layer.

5.6.1 Experimental datasets

Linear synthetic system: In our first data set, we consider a linear multivariate system with additive Gaussian noise consisting of four nodes (w, x, y, z), each having two dimensions. The summary graph of the system is shown in *Assumptions* in Fig. 13. The structural equations of the system are of the form:

$$Y_t^j := A^i \mathbf{Y}_{t-1}^j + \sum_{k \in PA(j)} B^k \mathbf{Y}_{t-1}^k + C^l N_t^j, \qquad (N_t^j)_d \sim \mathcal{N}(0,1) \quad \forall d$$

with N_t^j being zero mean standard Gaussian noise. In the following we show the coefficient matrices for the data generation of the linear synthetic system. The matrix of the root node w was chosen such that the eigenvalues are smaller than 1, which guarantees a stable system.

$$A^{ww} = \begin{bmatrix} 0.949 & 0.313 \\ -0.313 & 0.949 \end{bmatrix}, A^{xx} = \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.2 \end{bmatrix}, A^{yy} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}, A^{zz} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$$

$$B^{wx} = \begin{bmatrix} 0.5 & 0.2 \\ 0.2 & -0.5 \end{bmatrix}, B^{xy} = \begin{bmatrix} -0.9 & 0.7 \\ 0.7 & -0.9 \end{bmatrix}, B^{xz} = \begin{bmatrix} 0.4 & 0.9 \\ 0.9 & 0.4 \end{bmatrix}, B^{yz} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

$$C^{w}, C^{x}, C^{y}, C^{z} = \begin{bmatrix} 0.01 & 0.01 \\ 0.01 & 0.01 \end{bmatrix} \tag{13}$$

To simulate a root cause, we inject an additive constant term at a single dimension of a node j at time t to the equation above. Instead of a learned anomaly scoring function, in this experiment, we assume to have access to a function that checks the validity of a given observation, similarly, as it would be in a manufacturing scenario, in which an end-of-line test is performed [91]. Therefore, we examine if a failure on the "last" node in a manufacturing line (here "last" node in the summary graph is z) has occurred. To this end, we use a threshold function, fixed over time for each dimension of node z. More precisely, this classifier can be applied to any time-series observation $(\mathbf{Y}_t^j)_{t\in\{1,\ldots,T\},\, j\in\{w,x,y,z\}}$:

$$\phi(\mathbf{Y}) = 1 - \frac{1}{D_z}\sum_{k=1}^{D_z} \mathbb{1}_{[(\mu_z)_k-(\sigma_z)_k,(\mu_z)_k+(\sigma_z)_k]}(\mathbf{Y}_k^z)$$

Here, the dimension of node z is denoted with D_z. Note that this function provides gradual feedback on how many of the dimensions in node z are outside of the pre-specified corridor given by the threshold function.

FitzHugh–Nagumo system: Next, to allow for non-linear dynamic behaviour, we are generating data of the FitzHugh–Nagumo system (FHN), which is cyclic with regard to its summary graph, but acyclic in the unrolled graph $\mathcal{G}_t$. Although being a multivariate system, as the two dimensions interact, the corresponding dynamic SCM consists of one node x with two dimensions:

$$\dot{x_1} = 3(x_1 - x_1^3/3 + x_2), \qquad \dot{x_2} = (0.2 - 3x_1 - 0.2x_2)/3$$

We chose the initial values as in [92] but with slightly reduced additive Gaussian noise variance $\sigma^2 = 0.0025$. The root cause is simulated similarly to the linear system by adding a constant to the difference equation at one dimension and time point. We classify an observation as faulty if it deviates too much from a normal observation. As we have, in this setting, access to the ground truth, a normal observation is represented by a trajectory generated from the ground truth system. Consequently, the classifier consists of a time-varying threshold bound around each dimension of the normal observation without the injected root cause of node x. Denoting the expected trajectory from the system by $\mathbf{E}$ a given observation $\mathbf{Y}$ is then classified to be faulty if it does deviate more than 10 standard deviations at any point in time from the expected trajectory: $\phi(\mathbf{Y}) = 1 - \prod_t \mathbb{1}_{[\mathbf{E}_t^x - 10\sigma_x, \mathbf{E}_t^x + 10\sigma_x]}(\mathbf{Y}_t^x)$.

5.6.2 Evaluation

When we have drawn counterfactual samples from our model, we calculate the approximate Shapley values (see Eq. 5.5) and use the ϕ function to evaluate each intervention performed based on whether it corrected the failure. The root cause is the intervention of the node j at time t that has the highest influence on failure. If all counterfactual samples lead to the same ϕ evaluation for all interventions, then no unique root cause could be identified. However, due to random sampling of the counterfactual, this is an unlikely scenario (see, for example, Fig. 18). Nevertheless, for the evaluation, we only require that the ground-truth root cause is within the set of identified root causes. In Fig. 14 we show five counterfactual samples for the

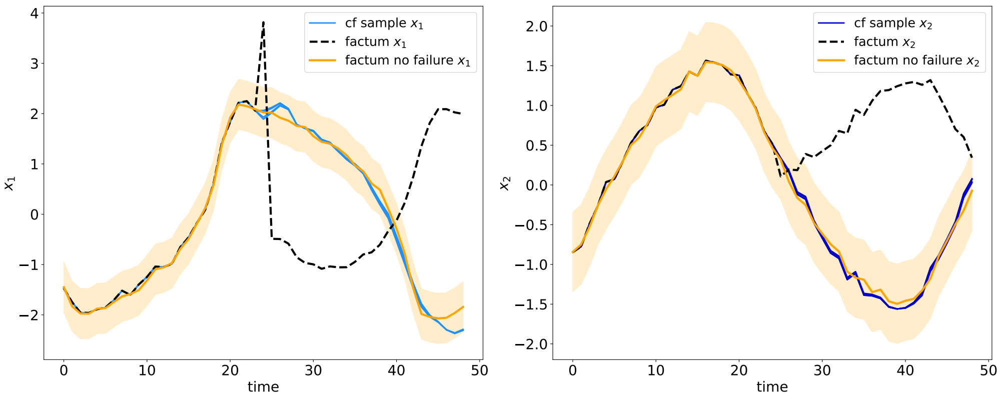


Figure 14: The figure shows the counterfactual samples for the FHN system with injected root cause at $(j = x_1, t = 24)$. The injected root cause disrupts the system observation heavily (black dashed line). However, the counterfactual intervention performed by our model $NLin(S_t^j, N_t^j)$ corrects the failure in both dimensions, such that it lies inside the threshold region (orange area).

nonlinear FHN system at the actual root cause injection point. Although the injected root cause is fairly large with regard to the interval of the normal observation without failure (drawn as an orange line), the counterfactual intervention performed by our model $NLin(S_t^j, N_t^j)$ corrects the failure for both dimensions of x. In order to analyse root cause injections and how the identification capabilities of our model behave under varying injections, we performed an *Injection experiment* for the synthetic linear and nonlinear FHN system. External disturbances in dynamic systems may be propagated and thereby increase their impact. Alternatively, if the system is robust against incremental noise (as is the case in the defined systems above due to the external noise influence even under the 'normal' conditions), it is not obvious how large an external influence at which point in time is noticeable. In Fig. 15, we show varying root cause injections for the linear synthetic system (varying constant added to the structural equation) over 20 randomly sampled facta with $T = 20$. It can be seen that the models that intervene in the structure and noise achieve a significantly higher identification score for large added constants. This could be due to a large root cause, in this setting leading to a factum with high distance to the normal data,

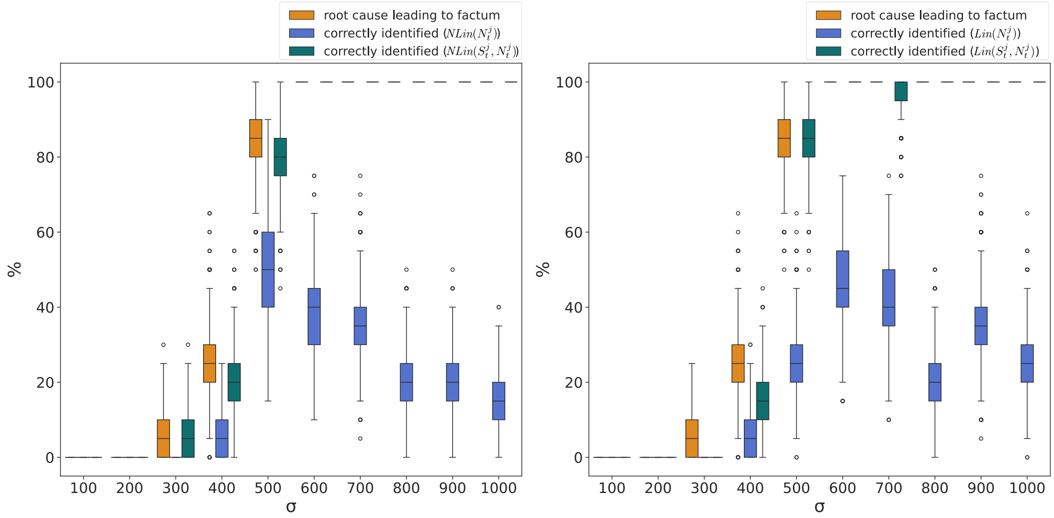


Figure 15: The root cause was injected at a random node $j = x_1$ at $t = 6$ with varying constants in $[1, 10]$. The horizontal axis shows the injected constant in relation to the noise standard deviation denoted by σ. We report how many root causes could be identified in %.

which can lead to a divergence over time of the normal behaviour system $\mathcal{M}$. We performed the same kind of injection experiment as for the linear system. In Fig. 16, we show varying root cause injections (varying constant added to the structural equation) over 20 randomly sampled facta with $T = 50$.

Assumption violation. We probe our models in violation of the causal graph assumption for the linear synthetic system. For this, we modify the causal graph used by the underlying model by adding or removing random edges, while keeping the original summary graph for data generation. We use the same facta generated as σ $= 500$ in Fig. 15. In Tab. 3 it can be seen that removing edges for all models has a stronger impact on predictive performance than adding. As expected, $Lin((S_t^j, N_t^j))$ performs best on this linear system, closely followed by $NLin((S_t^j, N_t^j))$. It must be mentioned that, in a graph with only four edges, removing an edge is a major incision in the model assumption.

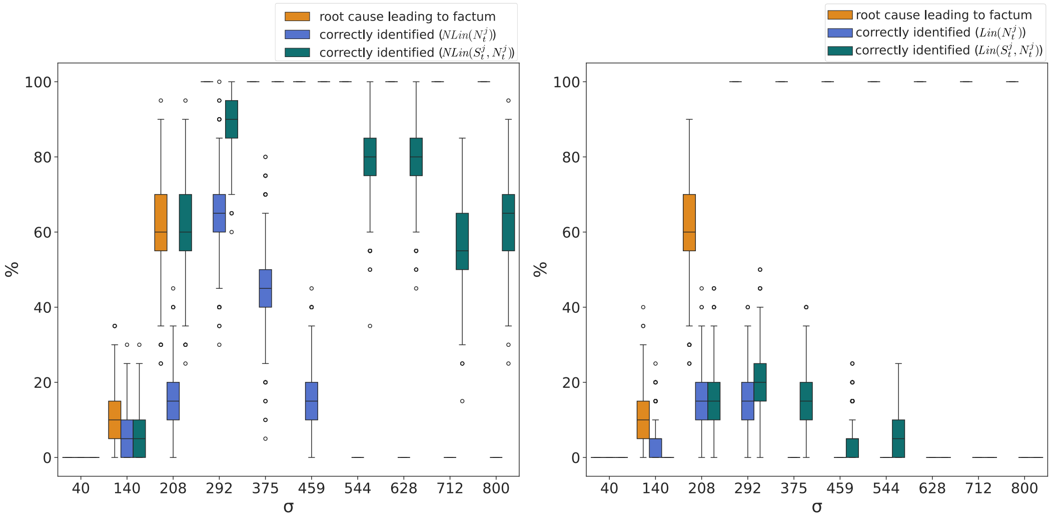


Figure 16: The root cause was injected at a random node $j = x_1$ at $t = 24$ with varying constants in $[0.1, 2.0]$. The x-axis shows the injected constant in relation to the noise standard deviation denoted by σ. We report how many root causes could be identified in %. On the left, for *NLin* it can be seen that the model intervening on the structure and the noise achieves a significantly higher identification score for larger added constants. This could be due to a large root cause, in this setting leading to a factum with high distance to the normal data, which may lead to a divergence over time of the normal behaviour system $\mathcal{M}$. On the right, we illustrated *Lin* and as expected, it is inadequate for addressing the complexities of the non-linear problem.

Table 3: We show the **Accuracy** for the setup $\sigma = 500$ of the linear synthetic system (see Fig. 15) with varying number of *removed or added edges* of the summary graph $\mathcal{G}$ used by the models.

	$NLin(S_t^j, N_t^j)$	$NLin(N_t^j)$	$Lin(S_t^j, N_t^j)$	$Lin(N_t^j)$
nr. of removed edges				
1	0.47 ± 0.25	0.23 ± 0.18	0.47 ± 0.24	0.18 ± 0.15
2	0.29 ± 0.20	0.06 ± 0.06	0.47 ± 0.24	0.12 ± 0.10
nr. of added edges				
1	0.82 ± 0.15	0.12 ± 0.10	1.0 ± 0.0	0.18 ± 0.15
2	0.88 ± 0.10	0.0 ± 0.0	0.88 ± 0.10	0.06 ± 0.06

Linear EasyRCA Benchmark. We compare against the linear univariate benchmark of [7] consisting of six nodes and two types of root causes. The parametric root cause means they change the coefficient of the parent nodes to a random uniform sampled value. As a special case of the parametric setting, they inject structural root causes, which set the coefficient of the parent nodes to zero. Since EasyRCA excludes single time point root causes, in order to do a fair comparison, we only rank sets of interventions, where we intervene on all times for a given node and evaluate it accordingly by $Sh(\xi_0^j, ...\xi_T^j)$. In their work, they inject on two nodes, where one is always the root node of the system and the other one a randomly chosen node. As in their benchmark comparison, the root cause of the root node is excluded, and we exclude it from the evaluation as well. In the evaluation, they distinguish for parametric and structural root causes, but because our model does not make a prediction about the type of root cause, it is sufficient if the EasyRCA predicted root causes contain the true root cause, regardless of the type. To rate the normality of a given trajectory $\mathbf{Y}$, we make use of the learnt dynamical SCM $\mathcal{M}$ that was fitted on normal observations of the system. More precisely, for the EasyRCA benchmark and the following River experiment, we used an outlier score similar to [3], based on the learnt dynamic SCM. That is, given a dynamic SCM $\mathcal{M}$ consisting of N nodes and providing the conditional distribution $p(\mathbf{Y}_t^j | \mathbf{Y}_t^{PA(j,t)})$ via the dynamics equation learned from normal observational data $(\mathbf{Y})_k$, we can define the following outlier score:

$$\phi(\mathbf{Y}) = \frac{1}{NT} \sum_{j,t} \log p(\mathbf{Y}_t^j | \mathbf{Y}_t^{PA(j,t)}) \tag{14}$$

In Tab. 4, it can be seen that, in general, the intervention (S_t^j, N_t^j) is preferable to an intervention only on (N_t^j). For linear systems, the accuracy of *NLin*(S_t^j, N_t^j) and *Lin*(S_t^j, N_t^j) are similarly good, while *EasyRCA* shows lower performance in the factum experiments $T = 100$. However, *Lin*(S_t^j, N_t^j) is inadequate to address the complexities of the non-linear problem.

Table 4: We report the **Accuracy** over 20 facta of the summary graph on a linear system and the FHN oscillator. In the lower part of the table we present the experimental results of the EasyRCA benchmark [7] comparing the accuracy for one factum over 30 graphs for different factum lengths T (here, normal data has the same size T). [1]

	NLin		Lin		EasyRCA
	(S_t^j, N_t^j)	(N_t^j)	(S_t^j, N_t^j)	(N_t^j)	
Lin. system	0.94 ± 0.05	0.59 ± 0.24	1.0 ± 0.0	0.29 ± 0.20	-
FHN oscillator	0.90 ± 0.09	0.65 ± 0.23	0.20 ± 0.16	0.15 ± 0.13	-
Lin. Parametric					
`Factum-100`	1.0 ± 0.0	0.97 ± 0.03	1.0 ± 0.0	0.97 ± 0.03	0.87 ± 0.12
`Factum-200`	0.97 ± 0.03	0.93 ± 0.06	1.0 ± 0.0	0.93 ± 0.06	0.93 ± 0.06
`Factum-500`	1.0 ± 0.0	0.97 ± 0.03	1.0 ± 0.0	0.97 ± 0.03	1.0 ± 0.0
`Factum-1000`	1.0 ± 0.0	1.0 ± 0.0	0.97 ± 0.03	0.83 ± 0.14	0.97 ± 0.03
Lin. Structural					
`Factum-100`	1.0 ± 0.0	0.87 ± 0.12	1.0 ± 0.0	0.83 ± 0.14	0.8 ± 0.16
`Factum-200`	0.90 ± 0.09	0.27 ± 0.20	0.70 ± 0.21	0.53 ± 0.25	0.90 ± 0.09
`Factum-500`	1.0 ± 0.0	1.0 ± 0.0	0.87 ± 0.12	1.0 ± 0.0	1.0 ± 0.0
`Factum-1000`	1.0 ± 0.0	0.8 ± 0.16	1.0 ± 0.0	1.0 ± 0.0	0.97 ± 0.03

Real World River Experiment. We analyse our method on real-world data considering a univariate river experiment consisting of four nodes. The nodes represent measuring stations of the Ribble River in England (data is from [90]). These mea-

---

[1] We excluded the 2000 factum length experiment of the EasyRCA benchmark for computational reasons. Additionally, note that since EasyRCA is univariate, it can not be applied to our synthetic systems.

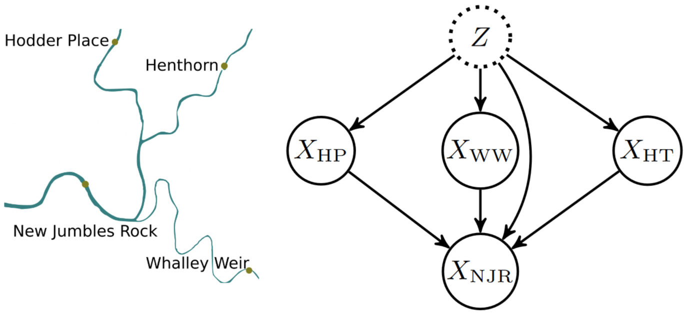


Figure 17: With the geographical knowledge of the river flow, a summary graph can be inferred (Figure taken from [3]).

suring stations are influenced by unknown external influences. For this reason, the summary graph includes a node Z representing unobserved common causes like weather conditions (e.g. rain, temperature) influencing all nodes. These unobserved confounders affect the accuracy of our model when learning the normal system $\mathcal{M}$ from observational data. The nodes represent the stations of the Ribble River that measure the flow rate. Although this data set has been investigated in [3], as a result of our dynamic viewpoint, we consider a slightly different factum. They consider four time points as static facta and infer the root causes of these. In contrast, we consider an entire time series as factum and infer the root cause. In addition, we use a finer time resolution of 15-minute intervals instead of averaged daily values, which has the advantage that the resulting SCM is less prone to instantaneous effects due to aggregation within a time-window. The finer resolution means that we consider a shorter period of time, namely the three days from 16.03.2019 to 19.03.2019 in which the flow rate is particularly high. As training data, we used the same time span as [3] from 01.01.2010 to 31.12.2018. They provide a z-score threshold for the New Jumbles Rock station, which we use as ϕ in 14, see Fig.18. We find the Shapley values with the highest scores at the Henthorn station, which is an upstream station of the New Jumbles Rock station. Although no ground-truth root cause exists for this experiment since it is a real-world example, the result is plausible both geographically and with regard to the time point. However, counterfactual intervention cannot correct the

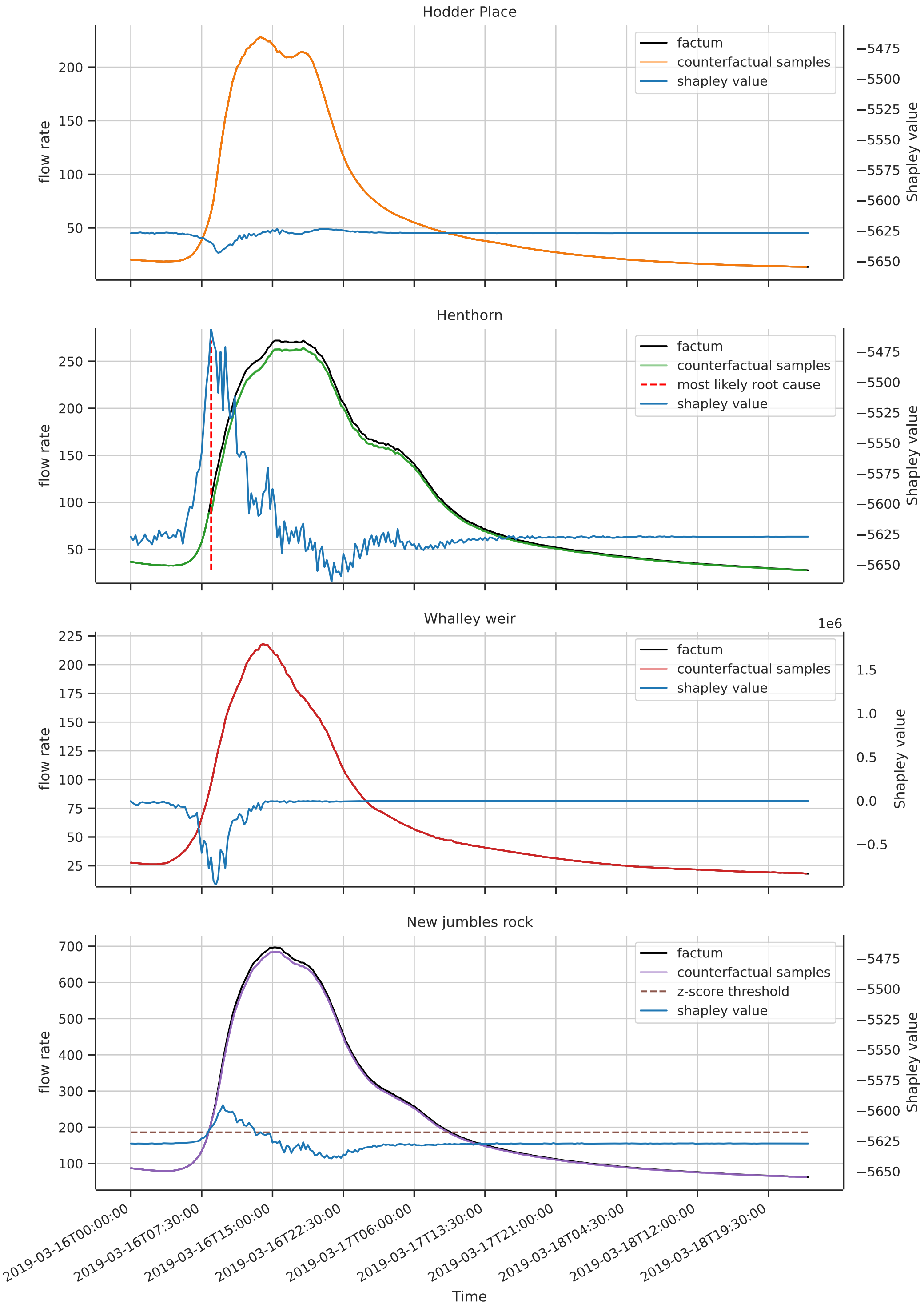


Figure 18: We show five counterfactual samples (for each station) of our model $NLin(S_t^j, N_t^j)$ with the intervention at the predicted root cause at 08:30 on 16.03.2019. Additionally, we illustrate the resulting Shapley values for each time point, showing that right before the failure occurs the Shapley values increase.

failure, as the counterfactual sample is not below the z-score threshold. This could be due to the fact that the influence of unobserved confounders is particularly high.

5.7 Conclusion

In this chapter, we have presented a method for identifying root causes in dynamic systems based on counterfactual reasoning. As the proposed method ranks individual interventions corresponding to individual nodes or sensors at particular times within a trajectory, our method is capable of exploiting not only the causal structure but also the natural direction of causality over time. By modelling temporal transitions with a non-linear neural network and a Shapley-value approximation, we are able to remove important limitations of current counterfactual root cause analysis methods. While we demonstrated both on synthetic as well as on real data the effectiveness of our method in identifying root causes in dynamic systems, there are several directions for further improvement. For example, our method is currently limited to the assumption that the root cause consists of a single intervention and that the causal graphical structure is known as well as the absence of latent confounders. In the next chapter, we sketch the extension of our method to account for the influence of latent confounders.

6 Confounding

To address the limitations of previous approaches, in this chapter, we analyse an existing approach that considers hidden confounding factors in a dynamic context. On the basis of this approach, we sketch ideas on how to perform root cause analysis with present hidden confounding factors. Up to this point, our analysis has relied on the assumption of strong ignorability (see 2.6), which presumes the absence of unobserved confounders. However, hidden confounders are prevalent in real-world systems. For instance, in the river experiment 5.6.2, rainfall could act as a hidden confounder, potentially influencing the observed flowrates both temporally and geographically. Additional complexities may arise in practice, such as multiple hidden confounders exerting simultaneous effects, large measurement errors obscuring the influence of hidden confounders or temporal shifts involving substantial time lags.

Given the prevalence of unobserved confounders, their presence introduces significant uncertainty in the causal inference process, as these hidden variables simultaneously affect both the treatment and the outcome. Consequently, assuming strong ignorability may lead to an inaccurate representation of the system and result in biased estimates of causal effects. This challenge arises because we lack direct knowledge of the underlying structural equations and must infer them from an (incomplete) causal graph with observational data which does not include information about the confounder. Even when we include the confounding variable in the causal graph and develop a latent variable model, determining the hidden confounder is not possible without imposing additional assumptions. For instance, we might assume that the latent variable follows a particular distribution, a temporal ordering or we need prior knowledge about the nature of confounding. Without such assumptions, there is no guarantee that the latent confounder is estimated accurately, leading to potentially biased causal estimates.

Extending the summary causal graph of the previous chapter to hidden confounding opens up new challenges in the counterfactual. Specifically, to the best of our knowledge, limited work exists on the separate inference of the exogenous variables and hidden confounders in the notion of SCMs during the counterfactual abduction step, see Sec. 3.3. Given these challenges, some methodologies employ the potential outcome framework rather than relying exclusively on a SCM.
The potential outcome framework provides a different perspective on causal inference by focusing on the comparison of potential outcomes under different treatment conditions. This framework is often more flexible when dealing with unobserved confounding because it does not require a fully specified causal graph to perform causal inference. One of the key advantages of the potential outcome framework in this setting is that it encapsulates the assumption of independent noise through the Stable Unit Treatment Value Assumption (SUTVA) (see Background Chapter 2.6).

Besides the theoretical causal framework, various inference methods exist to address a complex multivariate latent space. Here, we aim to outline several potential approaches, acknowledging that a comprehensive discussion of each is beyond the scope of this section. First, a hierarchical Bayesian model could be used, where latent noise and hidden confounders could be treated as distinct latent variables, each with its own prior distribution. This approach enables the separate modelling of these variables, with Bayesian inference used to estimate their posterior distributions based on observed data. For a recent overview of causal hierarchical modelling and the potential applications of Bayesian hierarchical modelling, see [93]. In addition, some work focuses on the latent space of dynamical systems [94], [95], also in a causal context [72]. Alternatively, Independent Component Analysis (ICA) could be employed to decompose a multivariate signal into additive, independent components. ICA leverages the statistical properties of the data to separate these components, but it requires the assumption that the latent noise and hidden confounder are statistically independent, which might be difficult to guarantee [96]. Another approach involves using proxy variables, which are observed variables correlated with the hidden confounders [97]. During the abduction step, these proxy variables could be used to approximate the influence of the hidden confounders, thereby facilitating more accurate causal estimation.

Finally, extending the causal model to include latent variables representing hidden confounders offers another strategy. These additional unobserved variables could be inferred from the observed data before estimating causal effects. But as mentioned before, when exogenous noise influencing each covariate is present, as typically in a causal model, it is quite difficult to derive the two latent variables separately. A specific latent variable model, known as the *Deconfounder* [4], will be explored in the subsequent section.

6.1 Static Deconfounder

In their approach, Bica et al. [5] estimate the treatment effect over time while addressing hidden confounding factors. Since their methodology builds upon the static deconfounding approach of [4], we will illustrate first the static concept before transitioning to the *Time Series Deconfounder* of Bica et al. The main idea involves a two-step process: first, estimating substitutes for the latent confounding variables and then using these substitutes for treatment effect estimation. Particularly the static deconfounder first fits a probabilistic factor model and outputs substitute confounders $Z_i \sim p(z_i|a_i)$. It then uses the individual factor weights $\hat{z}_i = \mathbb{E}_M[\mathbf{Z}_i|\mathbf{A}_i = \mathbf{a}_i]$ of the fitted model M to fit an outcome model $p(y_i|\mathbf{a}_i, \hat{z}_i)$. With that the average potential outcome estimate $\sum_{i=1}^{n} \mathbb{E}_Y[Y_i(\mathbf{A}_i)|\mathbf{A}_i = \mathbf{a}, Z_i = \hat{z}_i]$ can be computed, for more details see [4].

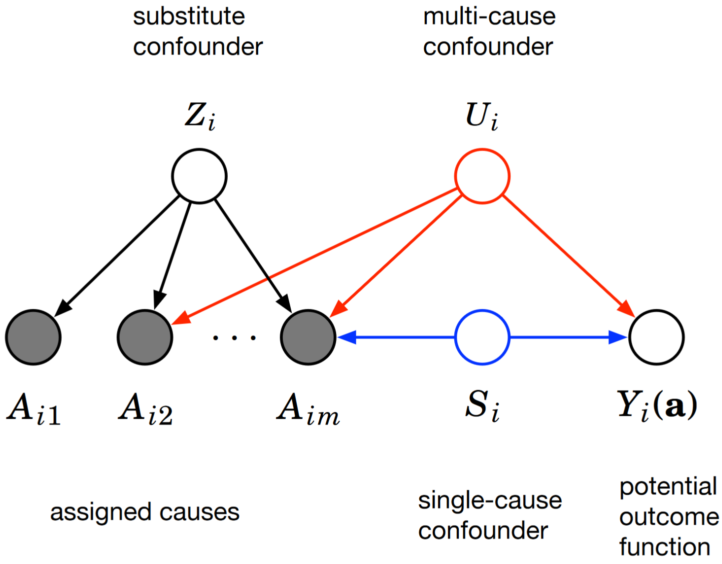


Figure 19: *Static multi-cause confounder* (Figure taken from [4]). Note that the substitute confounders Z_i are pre-treatment variables and should not contain any mediators along the assigned causes and the outcome because mediators cannot be identified by looking only at the assigned causes [4].

In their latent variable model, they assume that confounding variables are present as multi-cause confounders. This implies that a confounder must influence multiple causes to be inferable. For instance, in an economic context, the causes could be the prices of various items, and the effect would be the total consumer expenditure. Seasonality could act as a multi-cause confounder, affecting both prices and demand for multiple items.

In Fig.19, for each individual i, the latent substitute confounder Z_i renders the assigned causes A_{ij} (with j for each cause) conditionally independent $p(a_{i1}, ..., a_{ik}|z_i) = \prod_j p(a_{ij}|z_i)$. This implies that by contradiction, all confounding effects are captured by the substitute Z_i, and it can't exist another multi-cause confounder U_i, which has not been accounted for. However, this argument does not apply to single-cause confounders S_i [4].

6.2 Time Series Deconfounder

In this section we closely look at the work of [5] describing their Time Series Deconfounder. We introduce their approach including the latent model to infer hidden confounder variables and the outcome model to estimate the treatment effect over time.[1] When extending to time series, Bica et al [5] account for hidden confounders that may change over time and are affected by past treatments and covariates. To infer a latent confounder in the following, the probabilistic factor model of [5] is described, extending the work of [4] for time-varying treatments. They use multiple treatment assignments $\mathbf{A}_{tk}$ with k treatments (causes) at each timestep t to infer a sequence of latent variables $\overline{\mathbf{Z}}_t = (\mathbf{Z}_1, ..., \mathbf{Z}_t) \in \overline{\mathcal{Z}}_t$. These treatments could be binary and/or continuous. The time-dependent covariates are denoted by $\mathbf{X}_t$. The observational data of an individual i consists of realizations $\{\mathbf{x}_t^{(i)}, \mathbf{a}_t^{(i)}, \mathbf{y}_{t+1}^{(i)}\}_{t=1}^{T^{(i)}}$ with samples collected for $T^{(i)}$ discrete and regular timesteps.[2] Like in [4] the treatment is conditionally independent of the latent substitute confounder and the covariates:

$$p(a_{t1}, ..., a_{tk} | \mathbf{z}_t, \mathbf{x}_t) = \prod_{j=1}^{k} p(a_{tj} | \mathbf{z}_t, \mathbf{x}_t) \tag{15}$$

where $z_t = g(\overline{\mathbf{h}}_{t-1})$ being a function of the history and $\overline{\mathbf{h}}_{t-1} = (\overline{\mathbf{a}}_{t-1}, \overline{\mathbf{x}}_{t-1}, \overline{\mathbf{z}}_{t-1})$ being the realization of history $\overline{\mathbf{H}}_{t-1}$. To infer the latent confounding variables they model the marginal distribution of the assigned treatments $p(\overline{\mathbf{a}}_T)$ as a probabilistic factor model.

6.2.1 Factor model

Definition 6.2.1. *The factor model with joint distribution $p(\theta_{1:k}, \overline{\boldsymbol{x}}_T, \overline{\boldsymbol{z}}_T, \overline{\boldsymbol{a}}_T)$ is defined with a prior over the parameters $p(\theta_{1:k})$, the distribution of the observed*

---

[1] Note that not every detail of their work is discussed here, but only the main characteristics in order to describe the application for root cause analysis.

[2] We omit the individual superscript (i) in the following.

covariates $p(\overline{\boldsymbol{x}}_T)$ (over all timepoints T) capturing the dynamics of the observed variables by:

$$p(\theta_{1:k})p(\overline{\boldsymbol{x}}_T) \cdot \prod_{t=1}^{T}(p(\boldsymbol{z}_t|\overline{\boldsymbol{h}}_{t-1})\prod_{j=1}^{k}p(a_{tj}|\boldsymbol{z}_t, \boldsymbol{x}_t, \theta_j)) \tag{16}$$

They assume that at each timestep t, the treatment history, the covariates and the latent variable contain all dependencies between the assigned causes $(A_{t1}, ..., A_{tk})$ and the potential outcomes by defining the *sequential ignorable treatment assignment*:

$$\mathbf{Y}(\overline{\mathbf{a}}_{\geq t}) \perp\!\!\!\perp (A_{t1}, ..., A_{tk})|\overline{\mathbf{A}}_{t-1}, \overline{\mathbf{X}}_t, \overline{\mathbf{Z}}_t \tag{17}$$

for all $\overline{\mathbf{a}}_{\geq t}$ and for all $t \in 0, ..., T$. The potential outcome $\mathbf{Y}(\overline{\mathbf{a}}_{\geq t})$ is denoted for each possible course of treatment $\overline{\mathbf{a}}$ that starts at time t and consists of a sequence of treatments that end just before the patient outcome $\mathbf{Y}$ is observed. And $\overline{\mathbf{X}}_t = (\mathbf{X}_1, ..., \mathbf{X}_t) \in \overline{\mathcal{X}}_t$ is the history of covariates until timestep t. Besides sequential ignorable treatment assignment, they assume *positivity* and *consistency* (see 2.6) which are usually used among the existing methods.

The factor model in Fig 20 is showing the relationship between the $\mathbf{Z}_t$, $\mathbf{X}_t$ and the assigned treatments $\mathbf{A}_{tk}$. We want to emphasise that with their framework simultaneous treatments at each timepoint are possible. It is implemented by a Recurrent Neural Network (RNN), which is typically used for time series data. Note that this is not a causal graph since, for the potential outcomes framework, no causal graphical model is necessary. In the following, in the second stage, the inferred variables $\overline{Z}_t$ can be used to estimate the treatment effects over time by using an outcome model.

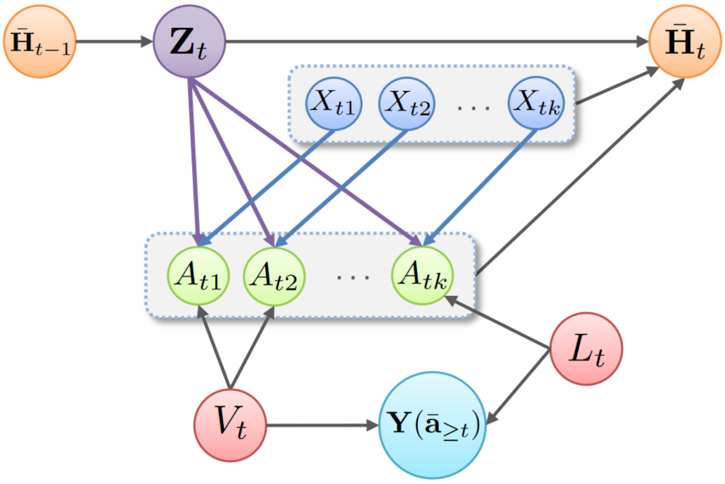


Figure 20: *Time Series Deconfounder* (Figure taken from [5]). The Time Series Deconfounder rules out the existence of other multi-cause confounders like V_t which are not captured by Z_t by assuming conditional independence between A_{tk} given Z_t and X_t. As in the static approach of [4], they assume no hidden single cause confounders L_t.

6.2.2 Outcome model

After the factor model has been fitted to observational data and captures the distribution of the assigned treatments, Bica et al [5] fit an outcome model to predict potential outcomes. They sample latent variables $\overline{\mathbf{Z}}_t$ from the factor model and augment it with the observational dataset (meaning $\overline{\mathbf{Z}}_t$ and the observational data is joined). Secondly, they fit a (Recurrent) Marginal Structural Model (R-MSN) to estimate the treatment effect (for each individual i) at each timestep (meaning they make one-step-ahead predictions) as $\mathbb{E}[\mathbf{Y}_{t+1}(\mathbf{a}_t)|\overline{\mathbf{A}}_{t-1}, \overline{\mathbf{X}}_t, \overline{\mathbf{Z}}_t]$. Using RNNs as R-MSN [98] to estimate the treatment effect over time, they use the *Inverse Probability of Treatment Weighting* to remove the bias from time-dependent confounders. They hereby present unbiased estimates in their work as long as the factor model captures the distribution of the assigned causes well[3] (see Eq. 15) and the assumptions hold (positivity, consistency and no hidden single cause confounders). However, the authors of [5] state that the

---

[3] To ensure this, they perform predictive checks on the factor model, but this is out of the scope of this section.

estimated hidden confounders from the factor model result in unbiased but higher variance estimates of the treatment effects. Therefore they trade off the confounding bias for estimation variance, which was also pointed out already in the static version [4]. Nevertheless, they argue that the treatment effects computed without considering the hidden confounders will inevitably be biased. After presenting the related work, we will now sketch how this approach can be applied to root cause analysis.

6.3 Deconfounding Dynamic Root Cause Analysis

In principle, the Time Series Deconfounder can be applied to root cause analysis. For this we are primarily interested in the counterfactual potential outcome of the outcome model. Probably the biggest distinction between the Time Series Deconfounder and our root cause approach is the way treatments, or called interventions in our approach, are carried out. They assume to have observed the covariates and, in addition, the assigned treatments (over the same timespan as the covariates) with the corresponding observed outcome $\mathbf{Y}_{t+1}$ of a diseased patient. In our approach, we observe the covariates $\mathbf{X}_t$ and the outcome $\mathbf{Y}_{t+1}$ of "healthy" patients and diseased patient samples in addition (we called this *normal* and factum data).

Since in the root cause analysis setup, we observed no separate historical treatment samples, we would assume that all covariates are potential intervenable variables, i.e., $\mathbf{A}_t = \mathbf{X}_t$.[4] An intervention could then be described by a change in a covariate starting at time t: $\mathbf{x}_{\geq t}$. In order to follow the assumption of a multi-cause confounder, since we utilized data from multivariate systems, we assume the confounder to affect multiple dimensions of a covariate (and the outcome) at time t. The dimensions of the influenced covariate are then assumed to be conditionally independent of $\mathbf{Z}_t$. The adapted factor model $p(\theta_{1:k}, \overline{\mathbf{x}}_T, \overline{\mathbf{z}}_T)$ is then described without explicit treatment

---

[4] We maintain this assumption here, although, in real-world applications, not all covariates are intervenable variables.

variables:

$$p(\theta_{1:k}, \overline{\mathbf{x}}_T, \overline{\mathbf{z}}_T) = p(\theta_{1:k}) \cdot \prod_{j=1}^{T} (p(\mathbf{z}_t | (\overline{\mathbf{x}}_{t-1}, \overline{\mathbf{z}}_{t-1}) \prod_{j=1}^{k} p(\mathbf{x}_{tj} | \mathbf{z}_t, \theta_j)) \tag{18}$$

After sampling substitute confounders $\mathbf{Z}_t$ for a factum sample i, we could then as in [5] estimate the potential outcomes for an performed treatment (intervention) $\mathbf{x}_{\geq t}$. With the (given) classifier ϕ applied to the counterfactual (potential) outcome, we receive the information if the intervention removed the failure or not. Afterwards, we could calculate the Shapley value over the performed intervention $\mathbf{x}_{\geq t}$ (as discussed in 5.4). As we do not have any *normal*, failure-free data in this setup, we cannot provide information on how close the counterfactual outcome has come to a *normal* state. In principle, the Time Series Deconfounder methodology enables the possibility of simultaneous interventions, thereby necessitating an extended Shapley value estimation, like [99].

6.3.1 Deconfounding with a Counterfactual SCM

If we stay in the counterfactual SCM framework instead of using potential outcomes as [5], we first need to define an extended causal graphical model. In the following, we demonstrate a potential extension for the linear synthetic system 5.6.1. We could, of course, extend the other presented dynamic systems as well. For the linear synthetic problem, an extended summary causal graph, as depicted in Fig. 21, is conceivable. In this setup, we define time dependent covariates as $X_t = \{\mathbf{w}_t, \mathbf{b}_t, \mathbf{d}_t, \mathbf{y}_{t+1}\}_{t=1}^{T}$. If we implement a latent confounder as a multi-cause confounder that influences all covariates (and in our case all treatment variables, $\mathbf{A}_t = \mathbf{X}_t$) like in the river example, see Fig.17, we have to assume that all treatment variables are conditionally independent. However, this breaks the dependencies in the causal graph.

Instead, we could imagine a simpler model where the latent confounder $\mathbf{v}$ affects the multiple dimensions of *one* variable $\mathbf{b}$ and the outcome value $\mathbf{y}$. Since we aim to use

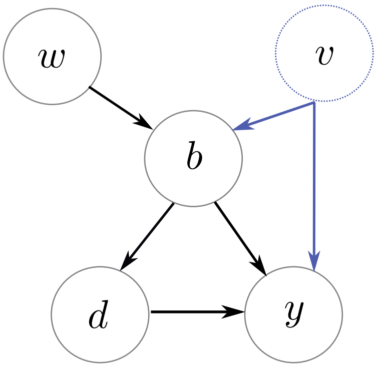


Figure 21: Extended causal graph with hidden confounder v. Note that variable names in the summary graph have been changed to prevent notation misuse.

the factor model of [5], the dimensions of (the multivariate treatment variable) $\mathbf{b}$ need to be conditionally independent given $\mathbf{v}$. After the latent confounder is inferred by the factor model and present in the observational data, the counterfactual intervention could, in principle, be performed on any covariate regardless of the treatment variable we used for estimating the confounder v. However, if the root cause is in b and we intervene on b, it is unclear if information about the root cause is still in b. We expect that the information about the root cause b could then be represented in v.

In summary, we believe that it is possible to use the *Time Series Deconfounder* of [5], with an attached evaluation of Shapley Values for root cause analysis. Using exclusively the factor model of the Time series deconfounder to derive the latent confounders and then using the (counterfactual) SCM framework with underlying causal graphs as in Chapt. 5 is not easily feasible. This is due to the assumption that the latent multi-cause confounder renders the treatments conditionally independent, which conflicts with the properties of a directed causal graph.
However, in principle, the counterfactuals in both frameworks are equivalent, as the potential outcomes framework and the counterfactual SCM framework are logically equivalent [20]. This may be only valid under the condition that the same $\mathbf{Z}_t$ values are sampled and the same interventions/treatments could be applied [100].

7 Conclusion

In this dissertation, we focused on the application of counterfactual reasoning within realistic and dynamic environments. First, we addressed the open research question of incorporating uncertainty into counterfactual reasoning. For this purpose, we introduced a hierarchical Bayesian model designed to handle ambiguities in the underlying SCM and uncertainties arising from limited observational data. By incorporating a Bayesian Warped Gaussian Process, we allowed for non-Gaussian and non-stationary noise distributions, which proved particularly effective for counterfactual distributions. Although we introduced additional uncertainty about the model's parameters, our approach led to a more accurate fit of counterfactual distributions in realistic scenarios. However, Gaussian Processes, along with Bayesian Warped Gaussian Processes, are computationally expensive, especially as the size of the training data increases. The added complexity of learning the warping function further intensifies the computational cost and reduces the model's interpretability. To improve the efficiency of Gaussian Processes, several extensions, such as sparse Gaussian Processes, have been proposed and could potentially be applied to Bayesian Warped Gaussian Processes. We leave this exploration as future work.

To address the second research question regarding the automatic identification of root causes in dynamic systems, we introduced a counterfactual-based approach. We demonstrated that our method is capable of ranking individual interventions as potential root causes, where these interventions correspond to failure candidates at specific points in time within a given trajectory. By utilizing a non-linear neural network and a Shapley value approximation for temporal transitions, our approach overcame the scalability and linearity limitations of current counterfactual root cause analysis techniques. However, given that the proposed method is limited to considering a single intervention, future research could explore extending this approach to handle multiple interventions simultaneously. While we discussed the effect of relaxing the assumption of a known causal graphical structure, extending the approach to an upstream causal discovery task would represent an intriguing direction for future work.

Although this thesis focused on the uncertainties associated with counterfactual reasoning, other unobserved factors, such as latent confounders, could also be considered. We outlined potential extensions for addressing latent confounding within a dynamic counterfactual root cause framework. However, this outline should be regarded as a preliminary suggestion, and we leave the development of these extensions open for future research. In conclusion, this work provides a significant advancement in both causal inference and dynamic systems analysis, providing robust tools for counterfactual reasoning, dynamic root cause analysis, and decision-making under uncertainty.

8 Appendix

8.1 Estimation of Counterfactual Interventions under Uncertainties

8.1.1 Hyperparameter Setup

Since our Bayesian Warped GP consists of various components, there are a couple of hyperparameters that can be optimized, see Table (5). We represent the bijective mapping g_ϕ of the Normalizing flow by a neural spline flow transform with element-wise (referred to as bins) rational conditional spline functions, to represent the conditional distributions in the causal model. Each conditional spline transform consists of a dense neural network with three Bayesian linear layers and a RELU activation function. In this neural network the hidden dimensions (`hidden dims`) need to be set (implemented with Pyro [101]). Furthermore, the spline is defined in a bounding box (`bounds`), which should cover the range of input data, for details see [81]. To relax this requirement, we normalize the input data.

According to our variational inference scheme, we can optimize further parameters affecting the training: the number of Monte Carlo samples (`S`) to be drawn, the prior variance (`prior var`), the learning rate (`lr`), and the training steps (`steps`). We optimize these variables in the seven variable setup to minimize the MMD on a held-out validation dataset of size 250 (generated from the ground truth SCM). For the three-variable setup, we optimized the hyperparameters w.r.t. the cost due to time constraints. In both cases, we used the BOHB (Bayesian Optimization algorithm using Hyperband) algorithm [102] for optimization. More specifically, we used the python ray-tune package of [103] as the implementation of BOHB.

Table 5: Optimal Hyperparameters found with BOHB on a validation set for each SCM and classifier setting.

	LINEAR SCM	NON-LINEAR SCM	NON-ADDITIVE	LINEAR LOG. REGR.	NON-LINEAR LOG. REGR.	RANDOM FOREST
`bounds`	6	1	10	27	3	21
`hidden dims`	10	13	40	2	6	27
`lr`	0.03	0.03	0.01	0.04	0.008	0.05
`steps`	5719	5719	4501	6982	6198	4956
`S`	15	21	20	31	24	21
`prior var`	0.1	0.1	0.05	0.03	0.01	0.02

8.2 Counterfactual-based Root Cause Analysis in Dynamic Systems

8.2.1 Hyperparameter Setup

We report the hyperparameters of our Residual neural network in Table (6). It consists of three layers with hyperbolic tangent activation functions and 128 neurons as latent layer. Note that we chose $\Delta_t = 1.0$ for all experiments.

Table 6: Hyperparameter setup of *NLin* for the performed experiments.

	Lin. system	FHN oscillator	EasyRCA benchmarking	River
splits	4	4	6	4
T_{train}	1000	1000	-[1]	300.000
T_{factum}	20	50	-	90
lr	0.01	0.01	0.1	0.01
epochs	50	100	200	50
dim	2	2	1	1

---

[1] In the EasyRCA benchmark the normal data and the factum have for each experiment the same length, see Table 2 in the main paper for corresponding T.

Bibliography

[1] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] K. Budhathoki, L. Minorics, P. Bloebaum, and D. Janzing, "Causal structure-based root cause analysis of outliers," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 2022.

[4] Y. Wang and D. M. Blei, "The blessings of multiple causes," 2019.

[5] I. Bica, A. M. Alaa, and M. van der Schaar, "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders," 2020.

[6] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera, "Algorithmic recourse under imperfect causal knowledge: a probabilistic approach," in *Advances in Neural Information Processing Systems*, 2020.

[7] C. K. Assaad, I. Ez-Zejjari, and L. Zan, "Root cause identification for collective anomalies in time series given an acyclic summary causal graph with loops," in

Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, PMLR, 2023.

[8] J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press, 2017.

[9] J. Weilbach, S. Gerwinn, M. Kandemir, and M. Fränzle, "Estimation of counterfactual interventions under uncertainties," in *Proceedings of the 15th Asian Conference on Machine Learning*, PMLR, 2024.

[10] J. Weilbach, S. Gerwinn, K. Barsim, and M. Fränzle, "Counterfactual-based root cause analysis for dynamical systems," in *Machine Learning and Knowledge Discovery in Databases. Research Track*, 2024.

[11] J. Pearl, *Causality.* Cambridge university press, 2009.

[12] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, 2009.

[13] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis.* John Wiley & Sons, 2011.

[14] A. Balke and J. Pearl, "Probabilistic evaluation of counterfactual queries," in *Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1)*, 1994.

[15] C. Avin, I. Shpitser, and J. Pearl, "Identifiability of path-specific effects," in *International Joint Conference on Artificial Intelligence*, 2005.

[16] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, 1986.

[17] J. Neyman, "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 1923.

[18] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, 2005.

[19] K. A. Bollen and J. Pearl, *Eight Myths About Causality and Structural Equation Models*. Springer Netherlands, 2013.

[20] D. Galles and J. Pearl, "An axiomatic characterization of causal counterfactuals," *Foundations of Science*, 1998.

[21] P. Rosenbaum and D. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 1983.

[22] B. G. Vegetabile, "On the distinction between "conditional average treatment effects" (cate) and "individual treatment effects" (ite) under ignorability assumptions," 2021.

[23] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, 1974.

[24] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Adaptive computation and machine learning, MIT Press, 2000.

[25] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*. Harvard University Press, 2001.

[26] A. Shojaie and E. B. Fox, "Granger causality: A review and recent advances," 2021.

[27] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," 2023.

[28] D. J. MacKay *et al.*, "Introduction to gaussian processes," *NATO ASI series F computer and systems sciences*, 1998.

[29] E. Snelson, Z. Ghahramani, and C. Rasmussen, "Warped gaussian processes," in *Advances in Neural Information Processing Systems*, MIT Press, 2004.

[30] J. Maroñas, O. Hamelijnck, J. Knoblauch, and T. Damoulas, "Transforming gaussian processes with normalizing flows," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021.

[31] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*, PMLR, 2015.

[32] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," *Advances in neural information processing systems*, 2005.

[33] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, 2015.

[34] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural computation*, 1989.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," 2016.

[36] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, 2019.

[37] I. Shpitser and J. Pearl, "What counterfactuals can be tested," in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 2007.

[38] M. Oberst and D. Sontag, "Counterfactual off-policy evaluation with Gumbel-max structural causal models," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, 2019.

[39] D. M. Chickering and J. Pearl, “A clinician’s tool for analyzing non-compliance,” in *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI Press / The MIT Press, 1996.

[40] G. W. Imbens and D. B. Rubin, “Bayesian inference for causal effects in randomized experiments with noncompliance,” *The Annals of Statistics*, 1997.

[41] T. S. Richardson, R. J. Evans, and J. M. Robins, “Transparent Parametrizations of Models for Potential Outcomes,” in *Bayesian Statistics 9*, Oxford University Press, 2011.

[42] S. Tsirtsis, A. De, and M. Rodriguez, “Counterfactual explanations in sequential decision making under uncertainty,” *Advances in Neural Information Processing Systems*, 2021.

[43] J. Zhang, J. Tian, and E. Bareinboim, “Partial counterfactual identification from observational and experimental data,” in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[44] A.-H. Karimi, B. Schölkopf, and I. Valera, “Algorithmic recourse: From counterfactual explanations to interventions,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2021.

[45] V. Guyomard, F. Fessant, T. Guyet, T. Bouadi, and A. Termier, “Generating robust counterfactual explanations,” 2023.

[46] J. Jiang, F. Leofante, A. Rago, and F. Toni, “Formalising the robustness of counterfactual explanations for neural networks,” 2022.

[47] S. Dutta, J. Long, S. Mishra, C. Tilli, and D. Magazzeni, “Robust counterfactual explanations for tree-based ensembles,” in *International Conference on Machine Learning*, PMLR, 2022.

[48] K. Budhathoki, D. Janzing, P. Bloebaum, and H. Ng, "Why did the distribution change?," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, PMLR, 2021.

[49] E. V. Strobl and T. A. Lasko, "Identifying patient-specific root causes with the heteroscedastic noise model," 2022.

[50] S. Tonekaboni, S. Joshi, K. R. Campbell, D. Duvenaud, and A. Goldenberg, "What went wrong and when? instance-wise feature importance for time-series black-box models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[51] J. Y. Halpern and C. Hitchcock, "Graded causation and defaults," *The British Journal for the Philosophy of Science*, 2015.

[52] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[53] L. S. Shapley, *A Value for N-Person Games.* Santa Monica, CA: Defense Technical Information Center, 1952.

[54] J. Pearl and J. M. Robins, "Probabilistic evaluation of sequential plans from causal models with hidden variables.," in *UAI*, 1995.

[55] G. Elidan, N. Lotner, N. Friedman, and D. Koller, "Discovering hidden variables: A structure-based approach," in *Advances in Neural Information Processing Systems*, MIT Press, 2000.

[56] P. R. Rosenbaum and P. R. Rosenbaum, "Sensitivity to hidden bias," *Observational studies*, 2002.

[57] A. M. Alaa and M. Van Der Schaar, "Bayesian inference of individualized treatment effects using multi-task gaussian processes," *Advances in neural information processing systems*, 2017.

[58] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, PMLR, 2016.

[59] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International conference on machine learning*, PMLR, 2017.

[60] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2018.

[61] S. Bongers, P. Forré, J. Peters, and J. M. Mooij, "Foundations of structural causal models with cycles and latent variables," *The Annals of Statistics*, no. 5, 2021.

[62] A. Hyttinen, F. Eberhardt, and P. O. Hoyer, "Learning linear cyclic causal models with latent variables," *J. Mach. Learn. Res.*, 2012.

[63] J. M. Robins, "Association, causation, and marginal structural models," *Synthese*, 1999.

[64] D. Cheng, Z. Xu, J. Li, L. Liu, J. Liu, W. Gao, and T. D. Le, "Instrumental variable estimation for causal inference in longitudinal data with time-dependent latent confounders," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[65] T. Hatt and S. Feuerriegel, "Sequential deconfounding for causal inference with unobserved confounders," in *Causal Learning and Reasoning*, PMLR, 2024.

[66] M. Haussmann, T. M. S. Le, V. Halla-aho, S. Kurki, J. V. Leinonen, M. Koskinen, S. Kaski, and H. Lähdesmäki, “Estimating treatment effects from single-arm trials via latent-variable modeling,” 2024.

[67] Y. Zhu, Y. He, J. Ma, M. Hu, S. Li, and J. Li, “Causal inference with latent variables: Recent advances and future prospectives,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

[68] J. Peters, S. Bauer, and N. Pfister, “Causal models for dynamical systems,” 2020.

[69] V. Didelez, “Causal reasoning for events in continuous time: A decision-theoretic approach.,” in *ACI@ UAI*, 2015.

[70] G. Blondel, M. Arias, and R. Gavaldà, “Identifiability and transportability in dynamic causal networks,” 2016.

[71] E. De Brouwer, J. Gonzalez, and S. Hyland, “Predicting the impact of treatments over time with uncertainty aware neural differential equations.,” in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, PMLR, 2022.

[72] N. Seedat, F. Imrie, A. Bellot, Z. Qian, and M. van der Schaar, “Continuous-time modeling of counterfactual outcomes using neural controlled differential equations,” *arXiv:2206.08311*, 2022.

[73] P. Sanchez and S. A. Tsaftaris, “Diffusion causal models for counterfactual estimation,” *arXiv:2202.10166*, 2022.

[74] B. Ustun, A. Spangher, and Y. Liu, “Actionable recourse in linear classification,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019.

[75] G. König, T. Freiesleben, and M. Grosse-Wentrup, “A causal perspective on meaningful and robust algorithmic recourse,” *arXiv:2107.07853*, 2021.

[76] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, “A survey of algorithmic recourse: contrastive explanations and consequential recommendations,” *ACM Computing Surveys (CSUR)*, 2021.

[77] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” *Harv. JL & Tech.*, 2017.

[78] K. Zhang and A. Hyvärinen, “On the identifiability of the post-nonlinear causal model,” in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009.

[79] F. Koehler, V. Mehta, and A. Risteski, “Representational aspects of depth and conditioning in normalizing flows,” *CoRR*, 2020.

[80] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, “Normalizing flows for probabilistic modeling and inference,” *Journal of Machine Learning Research*, 2021.

[81] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, “Neural spline flows,” *Advances in neural information processing systems*, 2019.

[82] H. M. Dolatabadi, S. Erfani, and C. Leckie, “Invertible generative modeling using linear rational splines,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR, 2020.

[83] N. Pawlowski, D. Coelho de Castro, and B. Glocker, “Deep structural causal models for tractable counterfactual inference,” in *Advances in Neural Information Processing Systems*, 2020.

[84] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *JMLR*, 2012.

[85] D. Garreau, W. Jitkrittum, and M. Kanagawa, “Large sample analysis of the median heuristic,” *arXiv:1707.07269*, 2017.

[86] J. von Kügelgen, P. K. Rubenstein, B. Schölkopf, and A. Weller, “Optimal experimental design via bayesian optimization: active causal structure learning for gaussian process networks,” *arXiv:1910.03962*, 2019.

[87] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search.* MIT press, 2001.

[88] A. Ikram, S. Chakraborty, S. Mitra, S. Saini, S. Bagchi, and M. Kocaoglu, “Root cause analysis of failures in microservices through causal discovery,” in *Advances in Neural Information Processing Systems*, 2022.

[89] L. G. Neuberg, “Causality: Models, reasoning, and inference, by judea pearl, cambridge university press, 2000,” *Econometric Theory*, 2003.

[90] “Hydrology data explorer.” `https://environment.data.gov.uk/hydrology/explore`. Accessed: 2024-03-18.

[91] K. Göbler, T. Windisch, M. Drton, T. Pychynski, S. Sonntag, and M. Roth, “`causalAssembly`: Generating realistic production data for benchmarking causal discovery,” 2024.

[92] P. Hegde, Çağatay Yıldız, H. Lähdesmäki, S. Kaski, and M. Heinonen, “Variational multiple shooting for bayesian odes with gaussian processes,” 2022.

[93] E. N. Weinstein and D. M. Blei, “Hierarchical causal models,” 2024.

[94] P. Kidger, J. Foster, X. C. Li, and T. Lyons, “Efficient and accurate gradients for neural sdes,” *Advances in Neural Information Processing Systems*, 2021.

[95] P. Kidger, J. Foster, X. Li, and T. J. Lyons, "Neural sdes as infinite-dimensional gans," in *International conference on machine learning*, PMLR, 2021.

[96] P. Wu and K. Fukumizu, "Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020.

[97] Y. Wang and D. Blei, "A proxy variable view of shared confounding," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021.

[98] B. Lim, A. Alaa, and M. v. d. Schaar, "Forecasting treatment responses over time using recurrent marginal structural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[99] N. Jethani, M. Sudarshan, I. Covert, S.-I. Lee, and R. Ranganath, "Fastshap: Real-time shapley value estimation," 2022.

[100] J. Y. Halpern, "Axiomatizing causal reasoning," 2014.

[101] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, "Pyro: Deep universal probabilistic programming," *J. Mach. Learn. Res.*, 2019.

[102] S. Falkner, A. Klein, and F. Hutter, "BOHB: robust and efficient hyperparameter optimization at scale," *CoRR*, 2018.

[103] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv:1807.05118*, 2018.