D I S S.  E T H  N O.  2 6 0 8 3

PRAGMATISM
AND VARIABLE TRANSFORMATIONS
IN CAUSAL MODELLING

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
SEBASTIAN WEICHWALD
MSc Computational Statistics and Machine Learning
University College London
born on 3 January 1991
citizen of Germany

accepted on the recommendation of

Prof. Dr. Joachim M. Buhmann, examiner
Prof. Dr. Peter L. Bühlmann, co-examiner
Prof. Dr. Bernhard Schölkopf, co-examiner

2019
To do(ask = "Why?")
I am grateful to all the people who have had and still have a positive effect on my path of life. I am thankful for not having to live through the counterfactual world.

I thank Bernhard Schölkopf for continuous support, for responding to my application as a naïve first-year student that I failed to send to the dedicated application email address and that should later mark the beginning of my academic route, for being there and bolstering me up whenever things turned tough, for an excellent and supportive research environment, for taking everyone and everyone’s needs seriously, and for the trust and encouragement to pursue exciting research even beyond current hypers. I thank Joachim Buhmann for the opportunity to learn and spend a great time at ETH Zurich, for initiating and paving the way for a successful cross-disciplinary cardiology project, for being pragmatic and outcome-oriented in keeping a project with more than fifteen authors on track and a coffee machine running beyond warranty, and for his thoughtful advice and support. I thank Peter Bühlmann for contributing to and fostering an unique atmosphere within the Seminar for Statistics\(^1\) where I always felt very welcome and Gipfeli-blessed as a guest, for friendly and joking coffee break conversations that cheered us up whenever we were brooding over experimental results in the seminar room, and for valuable and constructive feedback throughout our joint project from which I learnt immensely.

Thank you Sabrina, Rita, and Susan for endless support in manoeuvring all administrative and formal hurdles.

I thank Tatiana Fomina and Timm Meyer for jump start, bolstering me up, and for our friendship ever since. I thank Vinay Jayaram, Matthias Hohmann, and Bernd Battes for “the good old BCI group days”, Moritz Grosse-Wentrup for early support and character-forming challenges, and Thorsten Zander and Sebastian Stark for Tichu evenings.

Thank you Niklas Pfister and Paul Rubenstein, two brilliant and curious collaborators together with whom it is stimulating to pursue research and fun to push projects to perfection. I thank Jonas Peters for always being passionate and enthusiastic about research and the joint projects to come. Thank you Arthur Gretton for making research more fun and for seeding

---

1 This thank you naturally extends to the SfS and warrants a footnote in the acknowledgements.
valuable research intuition in me. Thank you Frederick Eberhardt and Caroline Uhler for exquisite food for thought during our causality workshop. Thank you Russell Poldrack, Martin Lindquist, and Christoph Herrmann for the great opportunity and humbling experience as novice symposium speaker along beside you. I enjoyed working and connecting with many outstanding researchers and feel honoured to have been invited to meet many more during the Foundations and New Horizons for Causal Inference Workshop in the Mathematical Research Institute of Oberwolfach.

Thank you Rebekka Burkholz, Christian Matter, Alessandro Candreva, Dominik Janzing, Stefan Bauer, Konrad Kording, and many more for intellectual discourse from which I learnt a lot, Jonas Kübler for coffee breaks and thoughtful reminders of the causes outside of academia such as climate change and political discourse that deserve more attention, Viktor Wegmayr, Djordje Miladinovic, Luca Corinzia, et al. for making my ETH stay sociable and fun. Thank you James Townsend, James Owers, and Eric Schulz – I hope we will catch up soon. I thank my fellow colleagues at the MPI for making coffee breaks more and more fun, lunch discussions more intriguing, and for contributing to an atmosphere of mutual recognition and support, thanks Simon Bartels, Mateo Rojas-Carulla, Atalanti Mastakouri, Carl Johann Simon-Gabriel, Alessandro Ialongo, Niki Kilbertus, Krikamol Muandet, and everyone who is missing from this list completely at random.

I thank Kim and Steffi, who indirectly affected my life more than they would ever think, Tobias for making his way from Berlin when I pretended to be too busy for travelling, Anna for saving me from serious consequences of a burn and for dinner and movie nights so that home felt more like home, un to de Nörder Lüü segg ik eenfach Moin un piller nich rum!

Thank you Tobias and Markus for your friendship, that we maintain despite the distance and seeing us irregularly. Thank you for being there for me through thick and thin. Thank you Tobias for also being a patient and skilled explainer of physics and philosophy – some day you may perceive whatever holds the world together in its inmost folds.

A special thanks goes to my family for always believing in me, especially when I do not, and for being wonderful – Dankschee! Thank you Mama and Papa for always encouraging me to pursue what I wish to do and for helping me to see more clearly what it is that I want to do, without you I would have never sent that aforesaid email…Thank you Benjamin, often nothing could comfort me more but your reminder “Na und?”. Thank you Imke for knowing me better than I often do myself and for together growing our we despite a wee bit of PhD.
## 4. Introduction 77

- **4.1 Notation** 79
- **4.1.1 Assumptions & Constraint-Based Causal Discovery** 80
- **4.1.2 Revisit Causal Terminology in Conclusions of Neuroimaging Studies** 81

## 4.2 Causal and Anti-Causal Encoding and Decoding Models 82

- **4.2.1 Modelling in Stimulus- versus Response-Based Paradigms** 83
- **4.2.2 Relating Feature Relevance to Causal Inference** 84

## 4.3 Causal Interpretation of Either Encoding or Decoding 86

- **4.3.1 Causal Encoding Models** $P(X|S)$ 86
- **4.3.2 Anti-Causal Decoding Models** $P(S|X)$ 86
- **4.3.3 Anti-Causal Encoding Models** $P(X|R)$ 87
- **4.3.4 Causal Decoding Models** $P(R|X)$ 88
- **4.3.5 Subsumption** 88

## 4.4 Causal Interpretation of Both Encoding and Decoding 89

- **4.4.1 Stimulus-Based Setting** 89
- **4.4.2 Response-Based Setting** 91

## 4.5 Rules for Causal Interpretation of Feature Relevance 92

## 4.6 Exemplary Application 94

## 4.7 Causality and Variable Transformations in Neuroimaging 97

## 4.8 Reinterpretation of Previous Work in a Causal Framework 102

- **4.8.1 Potential Confounds** 102
- **4.8.2 Linear Encoding and Decoding Models** 103
- **4.8.3 Revisiting the Introductory Examples** 104

## 4.9 Discussion 104

- **4.9.1 The iid Assumption in Neuroimaging** 105
- **4.9.2 Finite Empirical Data** 105
- **4.9.3 Univariate versus Multivariate Analysis** 106
- **4.9.4 Whole Brain Analysis** 107
- **4.9.5 Untestable Assumptions** 107
- **4.9.6 Causal Inference in Neuroimaging** 107

## 5. Confounding-Robust Source Separation 109

- **5.1 Introduction** 109
  - **5.1.1 Relation to Existing Work** 111
  - **5.1.2 Our Contribution** 117

## 5.2 Methodology 118

- **5.2.1 Motivating Examples** 121
- **5.2.2 Identifiability** 125
- **5.2.3 Estimation** 128
5.2.4 Assessing the Quality of Recovered Sources 130
5.3 Causal Structure Learning via ICA 133
  5.3.1 Application to Climate Science 136
5.4 Measurement Transformations and ICA 140
  5.4.1 ICA for Blind Cortical Source Separation in EEG 141
5.5 Experiments 144
  5.5.1 Competing Methods 144
  5.5.2 Simulations 145
  5.5.3 EEG Experiments 151
5.6 Conclusion 161

6 CONCLUSION & OUTLOOK 163

A APPENDIX 167
  A.1 Supplement: Causal Consistency of Causal Models 167
     A.1.1 Useful Results regarding Contraction Mappings 167
     A.1.2 Convergence under any Intervention 169
  A.2 Supplement: Confounding-Robust Source Separation 171
     A.2.1 Details on the GARCH Simulation 171
     A.2.2 EEG Experiments on Data Set 4 172
     A.2.3 All Topographies and Activation Maps on Data Set 3 176

BIBLIOGRAPHY 179

PUBLICATIONS a
ABSTRACT

Causal reasoning is ubiquitous in everyday life. In the end, you supposedly can give a reason why you are reading this very question just now, can you? Yet, inferring answers to causal questions from observations and formalising how and under which assumptions one can do so is non-trivial. The statistical treatment of causal modelling contributes to this scientific endeavour: It lays out methodology that, under well specified assumptions, enables us to infer cause-effect relationships from observational data. The adoption and fruitful utilisation of such methods remains limited, however, despite the statistical foundations and numerous theoretical advances.

In this thesis, we present contributions towards closing the gap between statistical causal modelling and its successful application.

Foremost is the first characterisation of whether and how transformations of variables preserve or break the ability to causally reason about a system. We argue that this conceptual understanding is crucial for all practical uses and provides theoretical justification for the applicability of causal modelling tools. On the one hand, our ability to observe the world is limited and thus we can ever only observe a system through inevitable measurement transformations of some underlying variables. For example, when we reason about the effects of turning on the hob we can neither observe nor mentally represent the motions of all atoms separately. Instead, we are left with variables that we do sense or measure and that can be understood as transformations of the underlying particle motions such as the water temperature or the tenderness of pasta. On the other hand, we may wish to make the modelling choice of transforming the observed variables to arrive at a simpler, more interpretable, or more effective model for a given task. These ideas establish an underexplored connection between causal modelling and recent advances in representation learning.

We clarify the conditions under which transforming variables—due to either limited measurement abilities or modelling choice—preserves causal reasoning. Importantly, this involves a pragmatic choice of the interventions we wish to model the effects of, since the intervention set determines which transformations lead to causally consistent models. This choice ought to be tailored to the task we are trying to solve; there is no true all-purpose causal model but instead there are pragmatic causal models that are instrumental for a given purpose. Realising how variable transfor-
mations determine our ability to causally reason brings statistical causal modelling closer to application: Many existing algorithms start from observations of some measured variables or preprocessed features and aim to infer cause-effect relationships between these. In practice, however, the system may not admit a causal description in terms of these variables, which hinders the applicability of such methods. It is thus imperative that we consider the variable transformations at play when dealing with and proposing real-world applications of causal modelling.

Our conceptual contribution works from statistics towards application. Pursuing a pragmatic approach, we moreover devote our attention to the applied end to foster the acceptance of causal modelling. We do so by examining predominant practice in the analysis of neuroimaging studies, namely the interpretation of so-called encoding and decoding models. While the relevance of a brain feature in either model is routinely considered informative regarding its causal role, we provide the first exhaustive overview over which causal statements are warranted and which ones are not. We (ex post) undergird routine analyses by sound causal theory thereby clarifying to what extent a causal interpretation is justified. This complements traditional, per se non-causal, encoding and decoding analyses and settles debates in the community on the involved interpretational issues. Our approach received attention and we increased the readiness by the community to adopt causal methodology. Hence, we advocate that considering the causal underpinnings of prevailing methods in a given applied field is a promising way to bring causal modelling closer to application.

One must be beware of measurement transformations that break our ability to causally reason about a system. For example, in electroencephalography we do not measure cortical signals directly but instead each electrode signal is a linear superposition thereof. For meaningful analysis of such mixed signal we need to invert the measurement transformation, which is also referred to as solving the blind source separation problem. For electroencephalography this is commonly achieved by independent component analysis. Employing a causal perspective, we develop coroICA, a novel confounding-robust independent component analysis that is robust against group-wise stationary noise. coroICA is readily applicable to multi-subject studies where samples can be grouped by the subject they were recorded from. It is demonstrably more robust when transferred to new unseen subjects. We deploy our algorithm as open-source packages in Python, R, and Matlab and demonstrate a further relation to the identification of linear causal structure through application to Antarctic ice core data.
ZUSAMMENFASSUNG


In dieser Dissertation werden Forschungsbeiträge vorgestellt mit dem Ziel, die Lücke zwischen statistischer kausaler Modellierung und ihrer erfolgreichen Anwendung zu schließen.

Im Vordergrund steht die erste Charakterisierung davon, ob und wie Transformationen von Variablen es möglich oder unmöglich machen, kausal über ein System nachzudenken. Wir vertreten die Ansicht, dass dieses konzeptionelle Verständnis für jedwede praktische Anwendung notwendig ist und eine theoretische Begründung für die Anwendbarkeit kausaler Modellierungsinstrumente liefert. Einerseits ist unsere Fähigkeit, die Welt zu beobachten, begrenzt, sodass wir ein System stets nur mittels unvermeidbarer Messtransformationen einiger zugrundeliegender Variablen beobachten können. Wenn wir zum Beispiel über die Auswirkungen des Herdeinschaltens nachdenken, können wir die Bewegungen aller Atome weder separat beobachten noch mental abbilden. Stattdessen bleiben uns Variablen, die wir tatsächlich wahrnehmen oder messen, und die als Abbildung zugrundeliegender Partikelbewegungen verstanden werden können, so wie etwa die Wassertemperatur oder die Bissfestigkeit von Nudeln. Andererseits möchten wir uns bei der Modellierung möglicherweise dazu entscheiden, die beobachteten Variablen zu transformieren, um so ein einfacheres, interpretierbareres, oder effizienteres Modell für eine gegebene Aufgabe zu erhalten. Diese Ideen stellen eine nicht ausreichend untersuchte Verbindung her zwischen kausaler Modellierung und jüngsten Fortschritten im Lernen von Repräsentationen.


Vorsicht ist geboten, bei Messtransformationen, die es unmöglich machen, kausal über ein System nachzudenken. In der Elektroenzephalographie messen wir beispielsweise nicht direkt kortikale Signale, sondern
INTRODUCTION

In this thesis I argue for a pragmatic approach to statistical causal modelling. Such an approach needs to factor in a model’s purpose which in turn informs the choice of suitable variable transformations. The aim is to bring statistical causal modelling from pen and paper to fruitful application. Here we provide an intuitive introduction to pragmatic causal modelling and delineate variable transformations as one over-looked key prerequisite for a successful application and broader adoption of causal modelling techniques in the life sciences.

The contributions we present in this thesis are three-fold, all working towards closing the gap between statistical causal modelling and its application. First, we contribute conceptually by formalising how the ability to causally reason about a system and thus the applicability of causal modelling techniques is affected by variable transformations; these transformations are often imposed on us as measurement transformations that reflect our limited ability to observe all relevant quantities of a system at appropriate detail. Second, we complement routinely conducted neuroimaging analyses by providing the first exhaustive interpretation chart that reflects which causal statements are warranted when interpreting feature relevance in encoding and decoding models. Third, we develop a novel variant of independent component analysis that is robust to hidden confounding and can be understood as undoing a causality-breaking measurement transformation, e.g. in electroencephalography the electrode signals measured on the scalp capture a linear mixture of underlying cortical signals and thus do not lend themselves as is for a causal description of cortical activity.

1.1 A PLEA FOR PRAGMATIC CAUSAL MODELLING

We encounter causal questions every day and—oftentimes unknowingly—reason about them with ease. Should we put on a jumper or turn on the heater if we were feeling cold in a room? How should we decide between these two actions if another person with a generally higher sensation of warmth was in the room? Does it matter for our decision whether the door towards the cool hallway is briefly opened time and again when people
I N T R O D U C T I O N

enter or exit the room? Does the wall painting matter, whether the room is under the roof or not, or the music we are listening to?

We are remarkably proficient at identifying the relevant concepts and quantities that enable us to reason about which actions to take in a given situation. Importantly, we are able to do this even without an encompassing understanding of all the underlying forces and mechanisms at work. We do not and need not know exactly how a heater works or how its thermal capacity and conductivity determine the temporal temperature profile in the room when turning it on. Still, we have an idea about how different actions will shape the world around us that is effective and useful within a specific context. That is, we have a causal intuition that we can leverage pragmatically, that serves its purpose and helps us navigate the world.

In this thesis we examine how our ability to causally reason about a system depends on which variables and transformations thereof we include as descriptors in our model. The appropriate choice and construction of descriptor variables is an essential, yet largely overlooked, prerequisite to a pragmatic approach to statistical causal modelling.

What do we mean by pragmatism here? We appreciate that all models are wrong, yet some are useful. Thence, we do neither claim nor aim to identify from data the true causal mechanisms that exist in the world and govern its evolvement. It is not necessary that we settle on the physical reality behind all people, objects, atoms, and fields in a room before we can reason about whether we should put on a jumper. Instead, we strive to identify appropriate descriptors and a causal model that serves a specific purpose and that can be justified given explicit assumptions, observations, and prior knowledge. We require a causal model to be derived in a systematic way, that is, for the same ingredients (assumptions, observations, prior knowledge) people necessarily arrive at the same causal model and its predictions on the effects of interventions are testable. It reflects our current best guess on how to causally reason about a given system and forecasts how certain aspects of the world around us depend on the actions we take. As such, causal models are viewed as a pragmatic tool to make informed and principled decisions or to deliberate about what will happen if we took a certain route of actions.

In practice, the variables that qualify as causal variables are often unknown. In the following example we discuss how causal reasoning may be complicated by our chosen representations, concepts, and descriptors. First, it is often not sensible to consider certain interventions in isolation, just like it is not enough to fill the places on a sports team separately from one another.
(e.g. putting 11 skilled athletes on a football team and ending up with 4 goal keepers instead of one). Second, we may have no knowledge about, no access to, or no way to measure the descriptors that facilitate causal reasoning (e.g. experience of a player). Third, we may need to construct aggregate variables based on the observed quantities (e.g. the fraction of matches won reflecting success), but generally lack any prior knowledge to inform such construction.

Let us imagine a coach who is entrusted with the task of forming the first European Jugger\(^1\) team. Selecting players from the South League and the North League, she aims to build a strong team and win the next year’s world championship. If she were to stick with already existing teams, a classical prediction-based approach would be to pick the team of either league that was most successful lately and expect somewhat similar chances of winning in the near future. Our coach, however, is interested in forming a new team and is not restricted to draw from the pool of already existing Jugger teams in the South League or the North League. Her task is further complicated by the fact that not all players are allowed on a European team and, since it is the first championship ever, the competing teams are unknown which impedes targeted game play strategies. The causal question thus is “How should she fill the five places on her new team to maximise the chances of it being successful?” In contrast to sports betting, it is not enough to merely forecast the performance of existing teams for which historic observations are already available. Instead, she requires means to make an informed prediction about the chances of winning when putting together different players to form the new European team.

Her idea is intuitive: Pick the athletes which have been most successful in the past. An individual assessment of each player, however, is not enough to forecast the team’s performance. In particular, each Jugger team needs one runner, a role which requires elegant and swift movements and crouching down to sneak through the defence and score a goal, and four so-called enforcers. The fact that the team composition is important for predicting its success, suggests that she cannot decide on the five places on the team independently but instead needs to decide on the players jointly to ensure that at least one talented runner is among them. This illustrates our first point, that it may be problematic to consider interventions on separate variables in isolation. In our example we know that using the five places on a team as descriptors is inapt and we should rather think of a team as consisting of

\(^1\) Readers unfamiliar with this sports—be it fictitious or not—are encouraged to instead think of handball, association football, or a generic team sports with the respective number of athletes on a team.
one runner place and four enforcer places. Yet, in other realistic scenarios, when we are to infer a causal model from data, we often lack knowledge about such underlying structure and encounter problems in deriving meaningful causal descriptions as long as we consider interventions separately. If we did not know about the two roles in a Jugger team and scouted five Jugger talents independently, we would have a hard time explaining why one team may perform poorly while another may perform well.

Since the stakes are high, our coach consults a statistician to help her analyse historic data and inform the decision on which players to elect for her new team. She approaches all major Jugger clubs and requests information on all athletes who are entitled to play on the European team. To the statistician’s delight, she not only gathers all the data but also cleans and consolidates the data, resulting in a database as follows:

<table>
<thead>
<tr>
<th>athlete</th>
<th>body height</th>
<th>fraction of matches won</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carla</td>
<td>1.76m</td>
<td>0.84</td>
</tr>
<tr>
<td>Mariela</td>
<td>1.62m</td>
<td>0.91</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

To complicate matters for our coach and for the sake of this example, assume that the clubs do not provide information about the preferred roles of their players. Instead, our coach leverages her observation that smaller athletes tend to be skilled in the role of the runner. Based on the athlete database, the coach and statistician decide the best they can do is to select 1 shorter and 4 taller athletes. We emphasise that in our hypothetical scenario it is indeed instrumental for building a successful team to decide on the players based on body height. The reason is, that in this example being a runner type manifests itself in body height and so the coach’s decision to select a shorter athlete for the runner place on her team is likely to result in a reasonable team composition. This may be puzzling at first, since body height does not cause a player to be a talented runner. The following consideration is instructive. Assume we observe a tendency for runners to wear short hair (since it is less distracting during their swift movements) and hair length was reported in the athlete database. Then, just like body height, hair length may reflect being a runner type. Yet, if the coach were to select five skilled athletes and cut one’s hair, that would generally not result in a team with one runner and four enforcers. How pragmatically useful a variable is for causal reasoning depends on the context in which we are employing a certain descriptor and the (implicitly specified) way of intervening on it. In our example, having the runner place occupied by a
shorter athlete is implemented by selecting an athlete that happens to be not that tall; having it occupied by a short-haired athlete was implemented by cutting an athlete’s hair. The validity of the statement “putting one short-haired player on the team increases the chances of having one gifted runner on the team” depends on the context. It depends on whether a talented athlete is selected and then gets her hair cut or whether a talent is only selected among the short-haired athletes.

So far we have only talked about our coach selecting one gifted runner and four enforcers. We can use body height to capture information about a player’s role, but we have not yet discussed how our coach would select the skilled athletes. Provided a database as depicted above, one intuitive approach is to identify the five athletes with the highest fraction of matches won (within their respective role). High fractions correspond to athletes who often played on the victorious team of a match.

What could possibly go wrong when forming a team of five successful athletes? Imagine they query the database and obtain a list of athletes with the following highest fractions of matches won: 1.00, 1.00, 0.91, 0.90, 0.88. Impressed by these athletes’ track record, they invite them over for a trial training session. Soon the coach finds herself disappointed, though, and realises only then that all five are inexperienced Jugger athletes. What went wrong? The database lacks information that is important for reasoning about the athletes’ performance: Our coach only entered the fraction of matches won, effectively concealing the information about athletes’ experience as reflected in the total number of matches played. Indeed, in the raw data records she finds the fractions of above athletes to be made up as $\frac{1}{1}$, $\frac{1}{1}$, $\frac{10}{11}$, $\frac{9}{10}$, $\frac{7}{8}$. Having participated in 1, 1, 11, 10, 8 matches only, the players are evidently inexperienced. The corresponding high numbers of victorious matches can be explained by the common practice to rotate in inexperienced players only when the team has reached a clear point advantage early on in a match. As it stands, the data in the athlete database cannot inform an argument along the following lines: “A formation of five individually successful athletes can be expected to perform well as a team.”.

The problem is that the available variables are inapt to predict a newly formed team’s performance. This exemplifies how our ability to causally reason about a system crucially depends on which descriptors are available to us. In our hypothetical scenario we can explain and understand what went wrong. In practice, however, we cannot and it is unknown to us which descriptors we are possibly missing; and even if we knew, measurements of those variables may be impossible or expensive to obtain.
We carry on the example to demonstrate that sometimes we need to combine observed quantities to recover (proxies to) appropriate descriptor variables. Assume that Jugger clubs collect data inconsistently and that the total number of matches played per player is mostly missing. Instead, an information that all clubs list is the year a player joined the league. The athlete database is augmented as follows:

<table>
<thead>
<tr>
<th>athlete</th>
<th>body height</th>
<th>fraction of matches won</th>
<th>year joined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carla</td>
<td>1.76m</td>
<td>0.84</td>
<td>2015</td>
</tr>
<tr>
<td>Mariela</td>
<td>1.62m</td>
<td>0.90</td>
<td>2018</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Our coach assumes that on average a player participates in 20 matches per season. She thus suggests that for each athlete the number of matches played can be reasonably estimated by multiplying the difference between the current year and the year a player joined the league by 20. Multiplying the resulting number by the fraction of matches won, we obtain an estimate of the number of victorious matches an athlete took part in. Let us assume that during the first 10 matches new players are mostly only rotated in under clear point advantage. The first 10 matches are thus almost certainly victorious matches and are not informative about an athlete’s individual performance. The statistician proposes to construct an adjusted winning ratio from the available variables, subtracting 10 from the estimated numbers for both the victorious matches and total matches played, and dividing the former by the latter:

\[
\text{adjusted winning ratio} = \frac{\text{fraction of matches won} \times (\text{current year} - \text{year joined}) \times 20 - 10}{\text{current year} - \text{year joined} \times 20 - 10}
\]

The adjusted winning ratio is more suitable for capturing past and predicting future individual athlete’s success. Note that the specific construction is motivated and derived by reference to mechanisms that we assume to be in place in our exemplary scenario. When modelling a real-world system such knowledge is generally unavailable. It is a fundamental challenge to find the right transformations of the observed quantities to obtain descriptors that facilitate causal reasoning.

This closes our illustration of the problems we are facing in causal modelling of real-world systems. These are inspiration and motivation for the
work presented in this thesis. First, it is often not sensible to consider certain interventions in isolation, just like it is not enough to fill the five places on a Jugger team separately from one another; instead causal reasoning about the future prospects of a newly formed team requires us to understand and respect the concept of a 4-enforcers-plus-1-runner composition of a Jugger team. Second, we may have no knowledge about, no access to, or no way to measure the descriptors that facilitate causal reasoning (e.g. experience as reflected by the year a player joined the league). Third, we may need to construct aggregate variables based on the observed quantities (e.g. an adjusted winning ratio), but generally lack any prior knowledge to inform such construction.

In real-world scenarios, we have often implicitly solved the above issues of choosing the appropriate level of description and transformations of variables without even thinking about it. For example, we know which aspects to pay attention to and how to conceptualise the world around us when feeling cold and deliberating about whether to put on a jumper or turn on the heating; seeing that others in the lecture hall are feeling warm and that the door is being opened every now and then we reason that putting on a jumper is more likely to result in everyone feeling comfortable. The situation is more subtle when trying to infer pragmatically useful cause-effect relationships from data of new or less well understood systems, such as the brain or genetic pathways, for which we lack prior knowledge. Just like in our example, there are several complications before we can even begin to infer cause-effect relationships between the chosen descriptor variables, fit a model to our observations, and reason about the effect of interventions. Leveraging pragmatic causal modelling in practice requires an understanding about how to identify, characterise, and obtain the right causal variables as transformations of observed variables. Hoping to bring statistical causal modelling from pen and paper to fruitful application, in this thesis I advocate a pragmatic approach to causal reasoning and delineate one overlooked key prerequisite: causal variable definition and identifying the right descriptors as transformations of observed variables.

1.2 VARIABLE TRANSFORMATIONS MATTER FOR APPLICABILITY

We need to understand and characterise how variable transformations determine our ability to causally reason about a system in order to foster widespread adoption of causal modelling techniques in the life sciences. What concepts and aggregate descriptors are appropriate to causally rea-
INFORMATION

son about a system and how can we identify those? To date, most causal discovery and inference approaches (implicitly) presume that the measured variables and available features are the right players for a causal description and immediately proceed to estimating cause-effect relationships between them. It may be necessary, however, to transform the observed variables to arrive at meaningful representations that enable an effective causal description. Without knowing how higher-level causal concepts emerge from lower levels, for example, we could hardly make sense and use of a causal model of the 100 billion neurons in a brain. Instead, a model of averaged neuronal activity in distinct functional brain regions may be pragmatically useful to reason about the effect of different treatments, to decide how to proceed during a surgery, and to understand the brain. At other times, the observable variables may, due to limited measurement abilities, be mixtures of underlying variables and we need to extract those from the observables prior to inferring meaningful cause-effect relationships. For example, we need to separate individual sound sources from audio recordings and repetitive motion patterns from video recordings before reasoning about which aspect of a driver’s behaviour causes a rattling engine noise.

Determining the most instrumental concepts and suitable variables for a given situation is key to effective causal modelling. In everyday life, we effortlessly choose the description level and switch between different relata and representations. The choice depends on what we are reasoning about: are we deliberating whether or not to take an established pain killer or are we deciding on the chemical composition of a novel compounded medication; are we reasoning about the right water temperature for brewing our filter coffee or are we optimising the product of brewing time and water temperature to extract minimal amounts of bitter constituents with our espresso machine; are we discussing nationwide road policies to reduce emissions and fatalities or are we campaigning against a ring road around our hometown that invites more speeding and fatalities than the twisting main through-road; are we reasoning about global vaccination policies to hamper disease spread or are we explaining to our children why they ought to wash their hands? We argue that a pragmatic approach to statistical causal modelling needs to factor in a model’s use case, i.e. the question being asked and the problem being tackled, which in turn informs the choice and design of suitable variable transformations.
1.3 CONTRIBUTIONS AND PUBLICATIONS

To obtain pragmatic causal models in practice, we need not only to understand the context and purpose of the modelling exercise but we must also choose the right causal variables. Closing the gap between statistical causal modelling and its successful application requires collective efforts on several fronts and contributions from multiple perspectives. The statistical end needs to be moved closer to application with conceptual and methodological work being guided and inspired by questions that arise when modelling real-world systems. Conversely, it is crucial that we understand the questions asked in applied fields and the prevailing approaches to answer those. This allows us to identify situations in which existing causal inference tools may be more suitable to answer such questions or in which the used approaches can be undergirded (ex post) by formal causal theory. Furthermore, employing a causal perspective and interpretation may inspire extensions and improvements to prevailing methods that would have been difficult to come up with otherwise.

In this thesis we present the following contributions and respective publications towards a pragmatic approach to causal modelling and towards narrowing the theory–application gap on all of the aforementioned fronts:

- **Formalisation of causally consistent models under variable transformations.** We can often describe the same system with reference to different terminology, levels of detail, and concepts. We can, for example, reason about individual neurons’ firing rates, about average blood oxygen levels in different brain regions, or about electromagnetic activity of so-called cortical dipoles and about how any of those maintain faster reaction times or certain movements. We tackle the following conceptual challenge that is fundamental to causal modelling of real-world systems such as, for example, the brain: How can we formally characterise the relata, aggregate features, and representations that are suitable for a pragmatically useful causal model and how do different description levels relate to one another? The variables we can and do measure do not necessarily lend themselves as is for a causal description.

  In this thesis we present [R+W+17] in which we develop the first general framework to characterise when two causal models of the same system are causally consistent with one another and agree in their predictions of the effects of interventions. The link between two models is established by the variable transformation that maps
the relata of one model onto the relata of the other. Transformations here may correspond to some chosen preprocessing and feature extraction steps or reflect our limited ability to measure the underlying system. Instead of reasoning about how individual pixel colours in an image affect brain activity we may first segment it and identify the objects therein and then model the relationship between neuronal activity and the presence and position of objects in a visual scene. An example of an inevitable measurement transformation is electroencephalography (EEG) where we cannot measure the underlying cortical signals directly but only electrode signals that are a linear superposition thereof.

Our framework provides a formal account of how transformations of variables either break or preserve causal reasoning and how transformations may be even necessary to enable causal modelling of the underlying system in the first place. Importantly, this provides theoretical justification for the applicability of causal modelling tools in real-world situations where (a) we only measure and model a sub-system of the world, i.e. where variables ‘irrelevant’ to or outside of this sub-system are implicitly being marginalised out, (b) we seek a description based on macro-level features that are aggregates of underlying micro-level variables, or (c) we have only access to observations at particular points in time of an underlying time-evolving dynamical system.

Our take on the interplay between causal reasoning and variable transformations enables one to in principle consider and identify transformations that exhibit desired properties, e.g. that allow for ‘simpler’ (in terms of complexity), more ‘interpretable’ (in terms one would need to define precisely), or more ‘robust’ (against interventional regime changes) causal models as compared to using the plain observed variables. Building up on our work, Beckers and Halpern [BH19] consider the consistent abstraction of causal models via appropriate variable transformations. Robustness to domain shifts resulting from interventions is considered by Bengio et al. [Ben+19]: They argue in favour of a representation that is consistent with the underlying causal structure in order for a learner to adapt faster to new environments and to thus obtain good transfer.

Causal interpretation and extension of common analyses in neuroimaging.

Having acquainted ourselves with common analysis practice and the types of questions neuroimaging studies aim to answer [Wei+14b; FoW+17; MaW+17], we identified so-called encoding and decoding analyses as two of the prevailing approaches [Wei+14a]. In this thesis we present [Wei+15] in which we provide the first comprehensive set of causal interpretation rules for encoding and decoding models. We extend and support common analysis approaches by causal theory which allows us to delineate warranted from unwarranted conclusions. By making the assumptions and relation to causal inference explicit, we arrive at ad hoc solutions to enrich the causal interpretation of neuroimaging studies.

This work was perceived with great interest in the neuroimaging community. Among others, this contribution has informed the interpretation of semantic content decoding from brain activity [Hut+16], led to a refined understanding of whole-brain neural dynamics of probabilistic reward prediction [Bac+17], was used to derive an improved and better resolved brain-wide functional atlas [Var+18], and clarified long-lasting debates regarding the problem of potential confounds [TNC13; WGB14] as well as the interpretation of linear backward models [Hau+14].

In order to disseminate causal modelling tools and sustain the change it is important to understand the status quo and the current challenges in a given field. We must focus on pragmatic solutions that improve and complement existing practice. Our work serves as a starting point for future interdisciplinary efforts. These should focus on the development of new tools that are specifically tailored towards answering the causal questions eminent, for example, in neuroimaging [WG17].

• *A causality-inspired confounding-robust independent component analysis.*

Consider the case where—due to physical limitations—we can only observe a transformation of underlying causal sources with additional confounding terms, e.g. EEG where we observe a linear mixture of cortical signals. Undoing this causality-breaking measurement transformation is necessary in order to perform causal inference and causally reason about the brain features. Also, recovering causal variables is desirable since, as one can argue, the causal mechanisms between them remain invariant under transfer and thus the risk on new unseen environments is reduced [PBM16; Roj+18]. In EEG analyses, this blind signal separation is commonly achieved by independent component analysis (ICA). Existing methods are limited, however, by either assuming noise-free observations or only accounting for time-independent noise.

The following shortcomings provide a complementary motivation for and perspective on our ICA extension. ICA is closely linked to identifying linear causal structure [Shi+06]. Here, noise-free ICA procedures limit the causal identification to scenarios where no hidden confounding is allowed. Furthermore, existing ICA models are somewhat misspecified when analysing data from different interventional settings. This link to causal identification provides a natural starting point to consider an extension of ICA to noisy observations that can be grouped by the different settings under which the observations were obtained.

In this thesis we present [P+W+19] in which we introduce coroICA, confounding-robust independent component analysis. We extend the ordinary ICA model in a theoretically sound and explicit way to incorporate group-wise (or environment-wise) confounding noise, while existing methods only allow for time-independent noise, if at all. We show that the noise model has a natural relation to causality, prove identifiability under mild assumptions, and provide an efficient estimation procedure. The use for blind signal separation and thus undoing a causality-breaking measurement transformation is exemplified through application to EEG data. We demonstrate how coroICA is employed for causal identification and estimate climate sensitivity from Antarctic ice core data.

This contribution is an example of how a causal perspective can help to enhance existing computational methods such as ICA even if these methods have common applications that are per se not concerned
with causality. It has potential impact on standard procedures in the EEG community, which include other ICA routines as a common pre-processing step. We demonstrate that more robust source separation across subjects is achieved. coroICA can readily be applied as so-called group ICA to common EEG study scenarios that comprise multiple recordings on different patients.


○ Open-source, ready-to-use, and maintained software packages.

Striving for applicability, all algorithms developed as part of the author’s doctoral research are released and maintained as open-source ready-to-use software packages. For coroICA we provide Python, Matlab, and R implementations [P*W*+19].

We anticipated the need for rapid prototyping to optimise different variants of objective functions over orthogonal or fixed rank matrices; this is naturally the case when considering transformations of linearly mixed signals such as EEG. Therefore, we developed Pymanopt, a Python toolbox for optimisation on manifolds using automatic differentiation [TKW16]. Technicalities of differential geometry and the laborious calculation of derivatives usually pose a significant barrier for experimenting with Riemannian optimisation, while our toolbox makes it simple: it only requires the user to define the objective function, choose a manifold, choose a solver, while no laborious calculation of difficult derivatives is necessary since Pymanopt seamlessly integrates and makes use of automatic differentiation. To date, it is one of the few general toolboxes for optimisation on manifolds that is actively maintained and used [BAC18].

Porting the optimisation part to Pymanopt enabled a more efficient implementation of the MERLiN algorithm, which is an algorithm that recovers one causal variable from an observed linear mixture. The original version of this methodology was developed and assessed as part of the author’s master’s degree, while the necessary extensions and completion for publication as well as the re-implementation that rests upon Pymanopt [Wei+16; WGG16] were addressed during the doctoral studies.
Improved 1-year mortality prediction after acute coronary syndromes.

Despite impressive advances in the care of patients with acute coronary syndromes (ACS), the incidence of major adverse cardiovascular events remains high after ACS [Szu+17]. Improved risk prediction for ACS patients upon admission remains an unmet clinical need. The guidelines of the European Society of Cardiology recommend the Global Registry of Acute Coronary Events (GRACE) risk score which provides “the most accurate stratification of risk both on admission and at discharge” [Kas+17].

Yet, biomarkers and patient characteristics recently identified to be relevant for prognosis are not included in the GRACE risk score. In our interdisciplinary project, we conduct exhaustive feature importance analyses, considering traditional and novel baseline features for the prediction of 1-year mortality in ACS patients. In analogy to the Jugger team building example in Section 1.1, the clinically relevant question is “which variable combinations make a good choice for building a risk stratification score”. We develop robust model evaluation and feature selection procedures since many features but only few and class imbalanced samples are available.

We propose a novel risk score consisting of 8 features that are easily and readily available on patient admission. It outperforms the state-of-the-art GRACE risk score for 1-year all-cause mortality prediction on our derivation cohort, which we expect to hold up on a future validation cohort. Based on our exhaustive analysis of $1.4 \cdot 10^9$ variable selections, we highlight that for improving adverse event prediction we require features that capture information covering each of the following aspects: heart failure, age, stress, and inflammation. Similar to how 1 runner and 4 enforcers are required for a successful Jugger team, features for each pathophysiological pathway are needed for comprehensive risk stratification. Moving beyond the robust prognosis of mortality after ACS, the focus of future research is robust causal inference to establish how those parameters can inform personalised treatment plans that optimally affect health outcomes after ACS.
1.4 Thesis Outline

In Chapter 2 we introduce causal models viewed as posets of distributions that describe how distributions change under interventions (cf. Section 2.1). Structural Causal Models (SCMs) as convenient ways to describe such a poset of distributions, and their use for modelling real-world systems are
introduced in Sections 2.2 and 2.3. We outline the causal discovery problem within our proposed interpretation of viewing causal models as posets of distributions in Section 2.4.

Our work on causal consistency of SCMs in Chapter 3 is introduced and motivated by examples such as the historical debate within the cardiology community on the effect of cholesterol on the risk of heart disease (cf. Section 3.1.1). We lay out the notion of an exact transformation that preserves causal reasoning between two models and illustrate how otherwise our ability to causally reason about a system breaks (cf. Section 3.2). Exactness of transformations in three scenarios is proven, which justifies the use of causal models when (a) variables are marginalised out, (b) micro-variables are aggregated as macro-variables, or (c) we causally reason about equilibrium distributions of time-evolving dynamical systems (cf. Section 3.3). We provide an impossibility result in Section 3.4 that motivates a generalisation of our notion of an exact transformation: The notion of an approximate transformation enables a trade-off of model accuracy and complexity for causal models that motivates future directions of research (cf. Sections 3.5 and 3.6).

We devote Chapter 4 to the challenge of inferring causal statements from neuroimaging data. We build on constraint-based causal discovery and are motivated by so-called encoding and decoding models being commonly interpreted using causal terminology (cf. Section 4.1). We explicate why an interpretation requires to distinguish between stimulus- and response-based paradigms (cf. Section 4.2). Linking feature relevance in either model and for either experimental paradigm to marginal or conditional dependence with the experimental condition, we provide the first exhaustive set of interpretation rules for such analyses (cf. Sections 4.3–4.5). By means of an exemplary application of our interpretation rules, we further discuss how variable transformations, that naturally arise as measurement transformations in neuroimaging, affect our ability to draw causal conclusions and to sensibly talk about cause-effect relationships between brain areas (cf. Sections 4.6 and 4.7). The application of causal inference tools to neuroimaging data is hurdled by many intricacies which, bridging the gap from application towards methodology, should inform the development of methods that are tailored towards the application and which necessitate the adoption of causal inference tools that go beyond encoding and decoding models (cf. Section 4.9).

In Chapter 5 we introduce our novel confounding-robust independent component analysis, corolICA. To better position our contribution within the
vast ICA literature and clarify how coroICA improves and extends upon existing methods, we provide a thorough overview in Section 5.1 over noisy ICA, ICAs based on approximate joint matrix diagonalisation, and ICA for grouped data such as multi-subject EEG recordings or observations from multiple domains/environments. We prove an original identifiability result and develop an efficient estimation procedure that holds under group-wise stationary confounding, while previous results only allow for time-independent noise or no noise at all (cf. Section 5.2). The link to linear causal structure learning is established in Section 5.3 and exemplified through application to Antarctic ice core data for the estimation of equilibrium climate sensitivity. We draw the connection to causal modelling under variable transformations in Section 5.4 and show how both, the application to EEG data as well as the causal structure learning via ICA, are closely related to the idea of undoing a causality-breaking measurement transformation. By extensive experiments on simulated and EEG data, we illustrate that coroICA performs competitive to existing methods in the noise-free or time-independent noise cases, while it improves robustness in the more general case of group-wise stationary noise (cf. Section 5.5). Thence, we conclude that coroICA can be considered a conservative ICA variant that safeguards against a broader class of noise without significant performance loss if more restrictive assumptions on the noise were indeed met (cf. Section 5.6).

We conclude in Chapter 6 and outline future research directions that build upon the work presented in this thesis.
We begin this chapter by introducing causal models as mathematical objects, namely we view a causal model as a family of joint distributions as opposed to common statistical models that model only one such distribution. These distributions can be indexed by different regimes, environments, or interventional settings that they correspond to. Inheriting a partial ordering on the set of interventions, a causal model amounts to a partially ordered set of distributions. Structural causal models are a convenient way to describe such a poset avoiding a laborious description that merely lists all the interventional distributions. We provide examples of how these mathematical objects are used for causal modelling of real-world systems.

Disclaimer: Unless stated otherwise we do not consider causal models for counterfactual reasoning. We focus on interventional causation and reasoning about how distributions change under intervention.

2.1 CAUSAL MODELS AS POSETS OF DISTRIBUTIONS

Here we introduce causal models as posets of distributions and focus on providing an intuition. The mathematical definition of structural causal models, which we use throughout this thesis, is deferred to the following section. For the sake of illustration and instruction, in this section we specify the joint distributions for the considered example in terms of probability mass functions. This way we emphasise that a causal model is an abstract mathematical object that represents multiple distributions and admits a real-world interpretation only after explicitly identifying physical with model quantities. In Section 2.2 we revisit this example and provide a crisp and more intuitive way to specify the distribution poset by means of a structural causal model; we illustrate how it can be used to causally model a real-world system in Section 2.3. Taken together, both sections are a demonstration of why structural causal models are a convenient way to specify a causal model; we also provide a visualisation of how one can think about a causal model that is helpful to consult throughout Chapter 3.

Section 2.3 is adapted, Section 2.2 adapted and extended from [R* W*+17].
Let us first consider a “common” statistical model consisting of a number of random variables say $X = \{A, B, L\}$ and some parameters $\theta$ that parameterise the model. Once these parameters are fixed, the model implies a fixed distribution $P_X$ over the random variables. For example, we can have the following joint distribution $P_X$ over three binary variables $X = \{A, B, L\}$ as specified by the probability mass function

\[
f_{A,B,L}(a,b,l) = \begin{cases} 
p^2, & \text{if } a = b = l = 1, \\
p - p^2, & \text{if } b = l = 1 \text{ and } a = 0, \\
p - p^2, & \text{if } a = l = 1 \text{ and } b = 0, \\
(p - p^2)/2, & \text{if } a = b = 0, \\
0, & \text{otherwise}, \end{cases}
\]

and parameterised by $\theta \equiv p \in [0, 1]$.

Causal models are different in that, once the parameters are fixed, they imply a family of distributions $P^\text{do}(i)_X$ over the random variables, one for each so-called intervention $i$ in the intervention set $I_X$.\(^1\) The null-intervention corresponding to the so-called observational setting is denoted by $\emptyset$. For example, we can have the following four joint distributions, corresponding to different elements in an intervention set $I_X = \{\emptyset, i_1, i_2, i_3\}$,

\[
P^\emptyset_X, P^\text{do}(i_1)_X, P^\text{do}(i_2)_X, P^\text{do}(i_3)_X
\]

specified by the respective probability mass functions

\[
f^\emptyset_{A,B,L}(a,b,l) = f_{A,B,L}(a,b,l); \\
f^\text{do}(i_1)_{A,B,L}(a,b,l) = \begin{cases} 
(1 - p)/2, & \text{if } a = b = 0, \\
p, & \text{if } b = l = 1 \text{ and } a = 0, \\
0, & \text{otherwise}, \end{cases} \\
f^\text{do}(i_2)_{A,B,L}(a,b,l) = \begin{cases} 
(1 - p)/2, & \text{if } a = b = 0, \\
p, & \text{if } a = l = 1 \text{ and } b = 0, \\
0, & \text{otherwise}, \end{cases} \\
f^\text{do}(i_3)_{A,B,L}(a,b,l) = \begin{cases} 
1/2, & \text{if } a = b = 0, \\
0, & \text{otherwise}, \end{cases}
\]

and again parameterised by $\theta \equiv p \in [0, 1]$.

---

\(^1\) While for the abstract mathematical description of a causal model it bears no meaning, elements of $I_X$ connote interventions for reasons explicated in Section 2.3.
Below we provide an intuitive illustration of the difference between “common” statistical models and causal models; one can imagine how different values of \( \theta \) result in different distributions on the respective line segments and the red circles correspond to the so-called interventional distributions in a causal model:

\[
\begin{align*}
\text{“common” statistical model} & \quad \text{causal model} \\
\mathbb{P}_X & \quad \mathbb{P}_X^{\text{do}(i)} \\
\end{align*}
\]

The intervention set \( I_X \) can be equipped with a partial ordering where \( \emptyset \leq_X i \) for all \( i \in I_X \); say that in our example we additionally have \( i_1 \leq_X i_3 \) and \( i_2 \leq_X i_3 \) while all other pairs of elements are incomparable.\(^2\)

Below we present a visualisation of a causal model, i.e. a poset of distributions \( \{ \mathbb{P}_X^{\text{do}(i)} : i \in I_X \} \) that inherits the partial ordering from the intervention poset \( I_X \):

\[
\text{causal model as poset of distributions}
\]

---

\(^2\) Note that in our example the probability mass functions \( f_{A,B,L}^{\text{do}(i)} \) are more concentrated resulting in lower entropy the larger the \( i \in I_X \).
2.2 STRUCTURAL CAUSAL MODELS

SCMs are a widely used framework in causal modelling, with applications in neuroscience, economics and the social sciences [Bol89; Pea09]. In this section we introduce them as an abstract mathematical object and a convenient way to represent a poset of distributions; in Section 2.3 we describe their use as a causal modelling tool. Readers already familiar with SCMs should note that our definition is more general and deviates from the standard definition of SCMs in the following ways: we do not require that all possible perfect interventions be modelled; we do not assume independence of exogenous variables;\(^3\) and we do not require acyclicity.

**Definition 1** (Structural Causal Model (SCM)).

Let \( \mathbb{I}_X \) be an index set. An SCM \( M_X \) over variables \( X = (X_i : i \in \mathbb{I}_X) \) taking value in \( \mathcal{X} \) is a triple \( (S_X, \mathcal{I}_X, \mathbb{P}_E) \) where

- \( S_X \) is a set of structural equations, i.e. it is a set of equations \( X_i = f_i(X, E_i) \) for \( i \in \mathbb{I}_X \);
- \( (\mathcal{I}_X, \leq_X) \) is a subset of all perfect interventions equipped with a natural partial ordering (see below), i.e. it is an index set where each index corresponds to a particular perfect intervention on some of the \( X \) variables;
- \( \mathbb{P}_E \) is a distribution over the exogenous variables \( E = (E_i : i \in \mathbb{I}_X) \);
- with \( \mathbb{P}_E \)-probability one, under any intervention \( i \in \mathcal{I}_X \) there is a unique solution \( x \in \mathcal{X} \) to the intervened structural equations. This ensures that for any intervention \( i \in \mathcal{I}_X \), \( M_X \) induces a well-defined distribution over \( \mathcal{X} \).\(^4\)

In an SCM, each \( X_i \) is a function of the \( X \)-variables and the exogenous variable \( E_i \). In this mathematical model, a perfect intervention on a single variable \( \text{do}(X_i = x_i) \) is realised by replacing the structural equation for variable \( X_i \) in \( S_X \) with \( X_i = x_i \). Perfect interventions on multiple variables, e.g. \( \text{do}(X_i = x_i, X_j = x_j) \), are similarly realised by replacing the structural equations for each variable individually. Elements of \( \mathcal{I}_X \) correspond to perfectly intervening on a subset of the \( X \) variables, setting them to some particular combination of values. Unless stated otherwise, in this thesis we consider perfect interventions only.

---

3 Exogenous variables are also referred to as noise variables in the literature. Our relaxation of the assumption of independent exogenous variables means our models may be considered a type of semi-Markovian causal model.

4 That is, with probability one over the exogenous variables \( E \), for each draw \( E = e \) there exists a unique value \( x \in \mathcal{X} \) such that \( e \) and \( x \) satisfy the intervened structural equations. The distribution of \( E \) in conjunction with \( S_X \) then implies a distribution over \( \mathcal{X} \) for each intervention \( i \in \mathcal{I}_X \) via these unique solutions. If the SCM is acyclic, this is always satisfied; we impose this condition because we also consider cyclic SCMs [Bon+18].
\( \mathcal{I}_X \) has a natural partial ordering in which, for interventions \( i, j \in \mathcal{I}_X \), \( i \preceq_X j \) if and only if \( i \) intervenes on a subset of the variables that \( j \) intervenes on and sets them equal to the same values as \( j \). For example, \( \text{do}(X_i = x_i) \preceq_X \text{do}(X_i = x_i, X_j = x_j) \).\(^5\) The observation that this structure is important is one of our contributions in [R\( ^\ast W^\ast +17 \)]. We make crucial use of it in the next chapter.

The purpose of the following example is to illustrate how SCMs are written in our notation and to provide an example of a restricted set of interventions \( \mathcal{I}_X \).

**Example 2.** Consider the following SCM defined over the variables \( X = \{A, B, L\} \) and for \( p \in [0, 1] \)

\[
\begin{align*}
S_X &= \{A = E_1, \\
&\quad B = E_2, \\
&\quad L = \text{OR}(A, B, E_3)\} \\
\mathcal{I}_X &= \{\emptyset, \\
&\quad \text{do}(A = 0), \\
&\quad \text{do}(B = 0), \\
&\quad \text{do}(A = 0, B = 0)\}, \\
E_1, E_2 &\sim \text{Bernoulli}(p) \\
E_3 &\sim \text{Bernoulli}(\frac{1}{2})
\end{align*}
\]

where by the element \( \emptyset \in \mathcal{I} \) we denote the null-intervention corresponding to the unintervened SCM.

An SCM naturally gives rise to a graph with variables as nodes and edges pointing from variables on the right-hand side of the structural equations to the ones on the respective left-hand side (cf. [SGS01; Pea09]):

![Graph](image)

The reader is encouraged to contrast the definition of a poset of distributions in Example 2 with the one given in Section 2.1; both descriptions

\(^5\) Informally, this means that \( j \) can be performed after \( i \) without having to change or undo any of the changes to the structural equations made by \( i \). Not all pairs of elements must be comparable: for instance, if \( i = \text{do}(X_1 = x_1) \) and \( j = \text{do}(X_2 = x_2) \), then neither \( i \preceq_X j \) nor \( j \preceq_X i \).
specify the same four distributions over the three binary variables \( \{A, B, L\} \) while the SCM brings two advantages: it can be accompanied by a graph representation over the variables and admits the interpretation of interventions as setting certain variables to certain values as opposed to interventions merely being a partially ordered index set for distributions.

Below we provide, analogous to the illustrations in the preceding section, a graphical depiction of the causal model defined by the structural causal model in Example 2:

poset of distributions implied by the SCM in Example 2

Note that \( P_A = P_{do(B=0)}^A = P_B = P_{do(A=0)}^B = \text{Bernoulli}(p) \), i.e. the marginal distribution on \( A \) does not change upon intervention on \( B \) and vice versa, while the marginal distribution on \( L \) changes as

\[
\begin{align*}
    P_L &\sim \text{Bernoulli} \left( \frac{1}{2} + \frac{2p - p^2}{2} \right), \\
    P_{do(A=0)}^L &\sim \text{Bernoulli} \left( \frac{1}{2} + \frac{p}{2} \right), \\
    P_{do(B=0)}^L &\sim \text{Bernoulli} \left( \frac{1}{2} + \frac{p}{2} \right), \text{ and} \\
    P_{do(A=0,B=0)}^L &\sim \text{Bernoulli} \left( \frac{1}{2} \right);
\end{align*}
\]

the probability for \( L = 1 \) decreases with increasing intervention.

2.3 SCMs FOR CAUSAL MODELLING

In addition to being abstract mathematical objects, SCMs are used in causal modelling to describe distributions of variables and how they change un-
nder interventions [Pea09]. The do-interventions as abstract manipulations of SCMs are understood as corresponding to actual (or potentially only hypothetical) physical implementations in the real world, i.e. the model is ‘rooted in reality’. For instance, if a binary variable $A$ in an SCM reflects whether a light bulb is emitting light, then $\text{do}(A = 0)$ could be achieved by flipping the light switch or by removing the light bulb.

The SCM in Example 2 could be thought of as a simple causal model of two light bulbs $A$ and $B$ and the presence of light $L$ in a room with a window. Suppose that we have no access to the light switch and there are no curtains in the room but that we can intervene by removing the light bulbs. We can model this restricted set of interventions by $\mathcal{I}_X$, i.e. the do-intervention on the SCM side $\text{do}(A = 0)$ corresponds to removing the light bulb $A$.

The partial ordering of $\mathcal{I}_X$ corresponds to the ability to compose physical implementations of interventions. The fact that we can first remove light bulb $A$, i.e. $\text{do}(A = 0)$, and then afterwards remove light bulb $B$, resulting in the combined intervention $\text{do}(A = 0, B = 0)$, is reflected in the partial ordering via the relation $\text{do}(A = 0) \leq_X \text{do}(A = 0, B = 0)$.

2.4 CAUSAL DISCOVERY

In causal discovery we aim to infer a (structural) causal model from some observational data that correctly predicts the effects of interventions $i \in \mathcal{I}_X$. The problem can be pictured as follows:
Importantly, for causal discovery we oftentimes have access only to samples of the observational distribution $P^\emptyset_X$ or of only a few of the interventional distributions $P^{do(i)}_X$ where $i \in \mathcal{I}_{sub} \subseteq \mathcal{I}_X$.

It is a notoriously difficult task to not only model the distributions of which observations are available but to also infer a causal model that enables reasoning beyond the observed distributions and thus on the effects of any intervention $i \in \mathcal{I}_X$. Statistical causal discovery lays out different approaches that clarify under which additional assumptions causal structure can indeed be identified.

Let us briefly illustrate the intricacies of causal discovery. Assume that, unknown to us, the ground-truth interactions between relative daily variations of perceived work load ($A$) and blood pressure ($B$) are governed by

\[
A = E_1 \\
B = A + E_2 \\
E_1, E_2 \overset{iid}{\sim} \mathcal{N}(0, 1)
\]

such that perceived work load causes blood pressure. We wish to reason about how the distribution changes upon either intervening on $A$ or on $B$, while we only have access to observational data. That is, we have access to samples of

\[
P^\emptyset_X = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\right)
\]

while we wish to model e.g. the poset $\{P^\emptyset_X, P^{do(A=-1)}_X, P^{do(B=-1)}_X\}$ to answer the question: Does reducing the perceived work load $A$ by not forwarding emails to employees before lunch reduce blood pressure $B$, or should we rather advice employees to take blood pressure lowering medication to eventually also reduce their perceived work load?

The statistical modelling literature tells us how to learn about the observational distribution from samples of $P^\emptyset_X$. Yet, without further assumptions and consultation of the statistical causal modelling literature we cannot make reasonable predictions about the two interventional distributions
even if we obtained the exact $P^\emptyset_X$ in the infinite sample limit. It is instructive to note the difference between conditional and interventional distributions:

$$P_{X|A=-1} \sim \mathcal{N}\left(\begin{bmatrix} -1 \\ -1/2 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right) \quad \text{vs} \quad P^\text{do}(A=-1) \sim \mathcal{N}\left(\begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$$P_{X|B=-1} \sim \mathcal{N}\left(\begin{bmatrix} -1/2 \\ -1 \end{bmatrix}, \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \end{bmatrix}\right) \quad \text{vs} \quad P^\text{do}(B=-1) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right)$$

That is, the conditional distributions do not match the interventional distributions and the joint observational distribution does not readily allow us to make predictions about the effects of interventions. Also note that we can model the distribution $P^\emptyset_X$ equally well by considering $A$ and $B|A$, or $B$ and $A|B$, or even $A|H$ and $B|H$ for some so-called hidden confounder $H$, while only the first alternative would align with the posed ground-truth relationship $A \rightarrow B$. This example illustrates why causal discovery is a difficult endeavour and requires further assumptions.

The gold standard to infer the causal effect of a treatment is a randomised controlled trial, a scheme popularised by Sir Ronald A. Fisher [Con91]. Indeed, we may be able to disambiguate between $A \rightarrow B$, $A \leftarrow B$, or $A \leftarrow H \rightarrow B$ if we were to randomly put employees under intense work load and observe blood pressure or conversely have employees take random dosages of blood pressure medication and record their perceived work load. However, randomisation is often not feasible, expensive, or unethical (e.g. randomly assigning subjects to a cumbersome, expensive, and potentially risky treatment) or even technically impossible (e.g. randomising the blood oxygen level in a brain region to investigate its behavioural effect).

We find mathematical frameworks in the causal discovery literature that conceptualise the notion of causality and formally proof which assumptions make causal structure learning possible in a given uncontrolled non-randomised setting. Most notably are constraint-based methods based on conditional (in)dependences [Sch+98; SGS01; Pea09], methods that exploit assumptions on the function class or noise properties [Shi+06; Hoy+08b], information-geometric formalisations relying on a principle of “independence of cause and mechanism” [Jan+12], and methods harnessing the idea that the accuracy of predictions of a causal model (in contrast to a non-causal model) should be invariant to interventions [PBM16].

We revisit two of the aforementioned approaches in Chapters 4 and 5 and refer the reader to the respective references and [PJS17] for more details on causal structure discovery. Specifically, in Section 4.1.2 we introduce the necessary concepts and assumptions for constraint-based causal
discovery that we leverage for the interpretation of encoding and decoding analyses in neuroimaging. In Section 5.3 we discuss how our proposed confounding-robust independent component analysis is related to methods for linear causal structure learning that utilise assumptions on the noise distributions. In Chapter 3 we discuss the consistency of two causal models of the same system, a fundamental question of “causal modellability” that precedes and possibly limits the applicability of causal discovery routines.
CAUSAL CONSISTENCY OF CAUSAL MODELS

Complex systems can be modelled at various levels of detail. Ideally, causal models of the same system should be consistent with one another in the sense that they agree in their predictions of the effects of interventions. We formalise this notion of consistency in the case of Structural Causal Models (SCMs) by introducing exact transformations between SCMs. This provides a general language to consider, for instance, the different levels of description in the following three scenarios: (a) models with large numbers of variables versus models in which the ‘irrelevant’ or unobservable variables have been marginalised out; (b) micro-level models versus macro-level models in which the macro-variables are aggregate features of the micro-variables; (c) dynamical time series models versus models of their stationary behaviour. Our analysis stresses the importance of well specified interventions in the causal modelling process and sheds light on the interpretation of cyclic SCMs.

3.1 INTRODUCTION

Physical systems or processes in the real world are complex and can be understood at various levels of detail. For instance, a gas in a volume consists of a large number of molecules. But instead of modelling the motions of each particle individually (micro-level), we may choose to consider macroscopic properties of their motions such as temperature and pressure. Our decision to use such macroscopic properties is first necessitated by practical considerations. Indeed, for all but extremely simple cases, making a measurement of all the individual molecules is practically impossible and our resources insufficient for modelling the $\sim 10^{22}$ particles present per litre of ideal gas. Furthermore, the decision for a macroscopic description level is also a pragmatic one: if we only wish to reason about temperature and pressure, a model of $10^{22}$ particles is ill-suited.

Statistical physics explains how higher-level concepts such as temperature and pressure arise as statistical properties of a system of a large number

Sections 3.1–3.3 and 3.6 are adapted from [R*W*+17]; the other sections present further unpublished project contributions.
of particles, justifying the use of a macro-level model as a useful transformation of the micro-level model [Bal91]. However, in many cases aggregate or indirect measurements of a complex system form the basis of a macroscopic description of the system, with little theory to explain whether this is justified or how the micro- and macro-descriptions stand in relation to each other.

Due to deliberate modelling choice or the limited ability to observe a system, differing levels of model descriptions are ubiquitous and occur, amongst possibly others, in the following three settings:

(a) Models with large numbers of variables versus models in which the ‘irrelevant’ or unobservable variables have been marginalised out [Bon+18]; e.g. modelling blood cholesterol levels and risk of heart disease while ignoring other blood chemicals or external factors such as stress.

(b) Micro-level models versus macro-level models in which the macro-variables are aggregate features of the micro-variables [SA61; IS94; HAT13; CPE15; CEP16]; e.g. instead of modelling the brain as consisting of 100 billion neurons it can be modelled as averaged neuronal activity in distinct functional brain regions.

(c) Dynamical time series models versus models of their stationary behaviour [Fis70; IS94; DD01; Lac+08; MH13; MJS13]; e.g. modelling only the final ratios of reactants and products of a time evolving chemical reaction.

In the context of causal modelling, such differing model levels should be consistent with one another in the sense that they agree in their predictions of the effects of interventions. The particular causal models we focus on in this thesis are Structural Causal Models (SCMs, Section 2.2, Section 2.3) [SGS01; Pea09].

In Section 3.2, we introduce the notion of an exact transformation between two SCMs, providing us with a general framework to evaluate when two models can be thought of as causal descriptions of the same system. An important novel idea of [R⋆W⋆+17] is to explicitly make use of a natural ordering on the set of interventions. On a high level, if an SCM can be viewed as an exact transformation of another SCM, we are provided with an explicit correspondence between the two models in such a way that causal reasoning on both levels is consistent. We discuss this notion of consistency in detail in Sections 3.2.6 and 3.2.7.

In Section 3.3 we apply this mathematical framework and prove the exactness of transformations belonging to each of the three categories listed
In the following we give an example of the problems that can arise when there exists no consistent correspondence between two causal models, i.e. neither model can be viewed as an exact transformation of the other. This example falls into category (b) of the differing model levels listed above and was used by [SS04] to illustrate problems in the causal modelling process.

Historically, the level of total cholesterol in the blood (TC) was thought to be an important variable in determining risk of heart disease (HD). To investigate this, different experiments were carried out in which patients were assigned to different diets in order to raise or lower TC. Conflicting evidence was found by different experiments: some found that higher TC had the effect of lowering HD, while others found the opposite (cf. Figure 3.1b) [Ste07; Tru10].
From our point of view, this problem (seemingly conflicting studies) arose from trying to perform an ‘invalid’ transformation of the ‘true’ underlying model (cf. Figure 3.1a). According to the American Heart Association, the current scientific consensus is that the two types of blood cholesterol, low-density lipoprotein (LDL) and high-density lipoprotein (HDL), have a negative and positive effect on HD respectively. Assigning diets that raise LDL or HDL both raise TC but have different effects on HD. It is therefore not possible to transform the model in Figure 3.1a into the model in Figure 3.1b without leading to conflict: in order to reason about the causes of HD we need to consider the variables LDL and HDL separately.

3.2 CAUSALLY CONSISTENT TRANSFORMATIONS BETWEEN SCMS

We now work towards our definition of an exact transformation between SCMs. Our core idea is to analyse the correspondence between different levels of modelling by considering one model to be a transformation of the other. We discuss in Section 3.2.6 how causal reasoning in two SCMs relate when one SCM can be viewed as an exact transformation of the other and in Section 3.2.7 we illustrate what can go wrong when this is not the case.

3.2.1 Distributions implied by an SCM

Usually, a statistical model implies a single joint distribution over all variables once its parameters are fixed (cf. Section 2.1). SCMs are different in that, once the parameters are fixed, an SCM implies a family of joint distributions over the random variables, one for each intervention. That is, for each intervention \( i \in \mathcal{I}_X \), the SCM \( \mathcal{M}_X \) defines a distribution over \( \mathcal{X} \) which we denote by \( \mathbb{P}^{\text{do}(i)}_X \). Throughout, we will denote the null-intervention corresponding to the unintervened setting by \( \emptyset \in \mathcal{I}_X \). We can write the poset of all distributions implied by the SCM \( \mathcal{M}_X \) as

\[
\mathcal{P}_X := \left( \left\{ \mathbb{P}^{\text{do}(i)}_X : i \in \mathcal{I}_X \right\}, \leq_X \right)
\]

where \( \leq_X \) is the partial ordering inherited from \( \mathcal{I}_X \), i.e.

\[
\mathbb{P}^{\text{do}(i)}_X \leq_X \mathbb{P}^{\text{do}(j)}_X \iff i \leq_X j. \tag{1}
\]

More formally, one would need to define \( \mathcal{P}_X \) to be the poset of tuples \( (i, \mathbb{P}^{\text{do}(i)}_X) \) to avoid problems in the case that \( \mathbb{P}^{\text{do}(i)}_X = \mathbb{P}^{\text{do}(j)}_X \) for some \( i \neq X j \). Doing so would not require a
3.2 CAUSALLY CONSISTENT TRANSFORMATIONS BETWEEN SCMS

Note that $\mathcal{P}_X$ contains all of the information in $\mathcal{M}_X$ about the different distributions implied by the SCM and, importantly, how they are related via the interventions.\(^2\)

3.2.2 Transformations of Random Variables

Suppose we have a function $\tau: \mathcal{X} \to \mathcal{Y}$ which maps the variables of the SCM $\mathcal{M}_X$ to another space $\mathcal{Y}$. Observe that since $X$ is a random variable, $\tau(X)$ is also a random variable. For any distribution $\mathbb{P}_X$ on $\mathcal{X}$ we thus obtain the distribution of the variable $\tau(X)$ on $\mathcal{Y}$ as $\mathbb{P}_{\tau(X)} = \tau(\mathbb{P}_X)$ via the push-forward measure.

In particular, for each intervention $i \in \mathcal{I}_X$ we can define the induced distribution $\mathbb{P}_{\tau(X)}^i = \tau\left(\mathbb{P}_X^{\text{do}(i)}\right)$. We can write the poset of distributions on $\mathcal{Y}$ that are induced by the original SCM $\mathcal{M}_X$ and the transformation $\tau$ as

$$\mathcal{P}_{\tau(X)} := \left\{ \mathbb{P}_{\tau(X)}^i : i \in \mathcal{I}_X \right\}, \leq_X$$

where $\leq_X$ is the partial ordering inherited from $\mathcal{P}_X$ (and in turn from $\mathcal{I}_X$). The scenario can be graphically depicted as follows:

$\mathcal{P}_{\tau(X)}$ is just a structured collection of distributions over $\mathcal{Y}$, indexed by interventions $\mathcal{I}_X$ on the $\mathcal{X}$-level; importantly, the indices are not interventions on the $\mathcal{Y}$-level.

3.2.3 Exact Transformations between SCMs

Although $\mathcal{P}_{\tau(X)}$ is a poset of distributions over $\mathcal{Y}$, there does not necessarily exist an SCM $\mathcal{M}_Y$ over $\mathcal{Y}$ that implies it. For instance, if there is some change to Definition 3 or affect the further results in this chapter. To avoid notational burden in our exposition, we omit this treatment.

\(^2\) For example, the distribution over the variables $X$ in the observational setting, $\mathbb{P}_X^\emptyset$, changes to $\mathbb{P}_X^{\text{do}(i)}$ if we implement the intervention $\text{do}(i)$, and the partial ordering contains all information about which interventions can be composed.
intervention \( i \in \mathcal{I}_X \setminus \{\emptyset\} \) such that none of the variables \( Y_i \) is constant under the distribution \( \mathbb{P}^i_{\tau(X)} \), then \( \mathbb{P}^i_{\tau(X)} \) could not possibly be expressed as arising from a do-intervention \( j \in \mathcal{I}_Y \setminus \{\emptyset\} \) in any SCM over \( Y \).

The case in which there does exist an SCM \( M_Y \) that implies \( \mathbb{P}_{\tau(X)} \) is special, motivating our main definition.

**Definition 3 (Exact Transformations between SCMs).** Let \( M_X \) and \( M_Y \) be SCMs and \( \tau : \mathcal{X} \to \mathcal{Y} \) be a function. We say \( M_Y \) is an exact \( \tau \)-transformation of \( M_X \) if there exists a surjective order-preserving map \( \omega : \mathcal{I}_X \to \mathcal{I}_Y \) such that

\[
\mathbb{P}^i_{\tau(X)} = \mathbb{P}^{\text{do}(\omega(i))}_Y \quad \forall i \in \mathcal{I}_X
\]

where \( \mathbb{P}^i_{\tau(X)} \) denotes the distribution of the \( \mathcal{Y} \)-valued random variable \( \tau(X) \) with \( X \sim \mathbb{P}^{\text{do}(i)}_X \).

Order-preserving means that \( i \leq_X j \implies \omega(i) \leq_Y \omega(j) \). It is important that the converse need not in general hold as this would imply that \( \omega \) is injective, and hence also bijective. This would constrain the ways in which \( M_Y \) can be ‘simpler’ than \( M_X \). That \( \omega \) is surjective ensures that for any do-intervention \( j \in \mathcal{I}_Y \) on \( M_Y \) there is at least one corresponding intervention on the \( M_X \) level, namely an element of \( \omega^{-1}(\{j\}) \subseteq \mathcal{I}_X \). The two results presented in the next Section follow immediately from the definition.

### 3.2.4 Basic Properties of Exact Transformations

**Lemma 4.** The identity mapping and permuting the labels of variables are both exact transformations. That is, if \( M_X \) is an SCM and \( \pi : \mathbb{I}_X \to \mathbb{I}_X \) is a bijection then the transformation

\[
\tau : \mathcal{X} \to \mathcal{Y} \\
(x_i : i \in \mathbb{I}_X) \mapsto (x_{\pi(i)} : i \in \mathbb{I}_X)
\]

naturally gives rise to an SCM \( M_Y \) that is an exact \( \tau \)-transformation of \( M_X \), corresponding to relabelling the variables.

---

3 This problem is elaborated upon in [Ebe16]. As mentioned in Section 2.2, in this thesis we only consider perfect interventions unless stated otherwise.

4 Since \( \omega(i) = \omega(j) \iff (\omega(i) \leq_Y \omega(j)) \land (\omega(j) \leq_Y \omega(i)) \), which, if the converse held, would imply that \( (i \leq_X j) \land (j \leq_X i) \), which is equivalent to \( i = j \).

5 For instance, if it were necessary that \( \omega \) be bijective, Theorems 9 and 11 would not hold.
Proof. Consider the SCM $M_Y$ obtained from $M_X$ by replacing, for all $i \in I_X$, any occurrence of $X_i$ in the structural equations $S_X$ and interventions $I_X$ by $Y_{\pi(i)}$ and leaving the distribution over the exogenous variables unchanged. This is a good sanity check; it would be problematic if this were not the case and the labelling of our variables mattered. Similarly, compositions of exact transformations are also exact.

**Lemma 5** (Transitivity of exact transformations). If $M_Z$ is an exact $\tau_{ZY}$-transformation of $M_Y$ and $M_Y$ is an exact $\tau_{YX}$-transformation of $M_X$, then $M_Z$ is an exact $(\tau_{ZY} \circ \tau_{YX})$-transformation of $M_X$.

Proof. Let $\omega_{ZY} : I_Y \to I_Z$ and $\omega_{YX} : I_X \to I_Y$ be the mappings between interventions corresponding to the exact transformations $\tau_{ZY}$ and $\tau_{YX}$ respectively and define $\omega_{ZX} = \omega_{ZY} \circ \omega_{YX} : I_X \to I_Z$. Then $\omega_{ZX}$ is surjective and order-preserving since both $\omega_{ZY}$ and $\omega_{YX}$ are surjective and order-preserving. Since $\tau_{ZY}$ and $\tau_{YX}$ are exact it follows that for all $i \in I_X$

$$P^i_{\tau_{ZX}(X)} = P^{\omega_{ZY}(\omega_{YX}(i))}_{\tau_{ZY}(\tau_{YX}(X))} = P^Z_{\omega_{ZX}(i)}$$

i.e. $M_Z$ is an $\tau_{ZX}$-exact transformation of $M_X$.

3.2.5 **Exact Transformations ensure Causal Consistency**

The theorem below is a consequence of the fact that $\omega$ is order-preserving. This is a mathematical formalisation of the sense in which an exact transformation preserves causal reasoning, which will be elaborated upon in the next subsection.

**Theorem 6** (Causal consistency under exact transformations). Suppose that $M_Y$ is an exact $\tau$-transformation of $M_X$ and $\omega$ is a corresponding surjective order-preserving mapping between interventions. Let $i, j \in I_X$ be interventions such that $i \leq_X j$. Then the following diagram commutes:
Proof. Let \(i, j \in \mathcal{I}_X\) be interventions with \(i \leq_X j\). The commutativity of the left square of the diagram follows immediately from the definition of an exact transformation. It remains to be shown that the right square of the diagram commutes. By definition we have that \(\tau(P_{X}^{\text{do}(i)}) = P_{Y}^{\text{do}(\omega(i))}\) and \(\tau(P_{X}^{\text{do}(j)}) = P_{Y}^{\text{do}(\omega(j))}\).

Thus, we only have to show that \(P_{Y}^{\text{do}(\omega(i))} \leq_Y P_{Y}^{\text{do}(\omega(j))}\) as elements of \(P_Y\), i.e. that the arrow \(P_{Y}^{\text{do}(\omega(i))} \xrightarrow{\text{do}(\omega(j))} P_{Y}^{\text{do}(\omega(j))}\) exists. This follows from the order-preservingness of \(\omega\). \(\square\)

3.2.6 Causal Interpretation of Exact Transformations

The notion of an exact transformation between SCMs was motivated by the desire to analyse the correspondence between two causal models describing the same system at different levels of detail. The purpose of this section is to show that if one SCM can be viewed as an exact transformation of the other, then both can sensibly be thought of as causal models of the same system. In the following, we assume that \(\mathcal{M}_Y\) is an exact \(\tau\)-transformation of \(\mathcal{M}_X\) with \(\omega\) the corresponding map between interventions.

Surjectivity of \(\omega\) ensures that any intervention in \(\mathcal{I}_Y\) can be viewed as an \(\mathcal{M}_Y\)-level representative of some intervention on the \(\mathcal{M}_X\)-level. Consequently, if do-interventions on the \(\mathcal{M}_X\)-level are in correspondence with physical implementations, then surjectivity of \(\omega\) ensures that do-interventions on the \(\mathcal{M}_Y\)-level have at least one corresponding physical implementation, i.e. if \(\mathcal{M}_X\) is ‘rooted in reality’, then so is \(\mathcal{M}_Y\).

Commutativity of the left hand part of the diagram ensures that the effects of interventions are consistently modelled by \(\mathcal{M}_X\) and \(\mathcal{M}_Y\). Suppose we want to reason about the effects on the \(\mathcal{M}_Y\)-level caused by the inter-
vention \( j \in \mathcal{I}_Y \). For example, we may wish to reason about how the temperature and pressure of a volume of gaseous particles is affected by being heated. We could perform this reasoning by considering any corresponding \( \mathcal{M}_X \)-level intervention \( i \in \omega^{-1}([j]) \) and considering the distribution this implies over \( \mathcal{Y} \) via \( \tau \). In our example, this would correspond to considering how heating the volume of gas could be modelled by changing the motions of all the gaseous particles and then computing the temperature and pressure of the volume of particles. Commutativity of the left hand part of the diagram implies that \( \mathcal{M}_X \) and \( \mathcal{M}_Y \) are consistent in the sense that \( \mathcal{M}_Y \) allows us to immediately reason about the effect of the intervention \( j \in \mathcal{I}_Y \) while being equivalent to performing the steps above. That is, we can reason directly about temperature and pressure when heating a volume of gas without having to perform the intermediate steps that involve the microscopic description of the system.

Commutativity of the right hand side of the diagram ensures that once an intervention that fixes a subset of the variables has been performed, we can still consistently reason about the effects of further interventions on the remaining variables in \( \mathcal{M}_X \) and \( \mathcal{M}_Y \). Furthermore, it ensures that compositionality of do-interventions on the \( \mathcal{M}_X \)-level carries over to the \( \mathcal{M}_Y \)-level, i.e. if the intervention \( j \) on the \( \mathcal{M}_X \)-level carries over to the \( \mathcal{M}_Y \)-level, then the same is true of their representations in \( \mathcal{M}_Y \).

If \( \mathcal{M}_X \) and \( \mathcal{M}_Y \) are models of the same system and it has been established that \( \mathcal{M}_Y \) is an exact \( \tau \)-transformation of \( \mathcal{M}_X \) for some mapping \( \tau \), then the commutativity of the whole diagram in Theorem 6 ensures that they are causally consistent with one another in the sense described in the preceding paragraphs. If we wish to reason about the effects of interventions on the \( \mathcal{Y} \)-variables then it suffices to use the model \( \mathcal{M}_Y \), rather than the (possibly more complex) model \( \mathcal{M}_X \). In particular, this means that we can view the \( \mathcal{Y} \)-variables as causal entities, rather than only functions of underlying ‘truly’ causal entities. Only if this is the case, causal statements such as ‘raising temperature increases pressure’ or ‘LDL causes heart disease’ are meaningful.

### 3.2.7 Non-Exact Transformations break Causal Consistency

In the previous section we argued that our definition of exact transformations between SCMs is a sensible formalisation of causal consistency. In this section we will try to give the reader an intuition for why weakening the
conditions of our definition would be problematic. In particular we focus on the requirement that $\omega$ be order-preserving, which we view as one of the core ideas of our paper [R*W*+17].

The requirement that $\omega$ be surjective is, as discussed above, required so that all interventions on the $\mathcal{M}_Y$-level have a corresponding intervention on the $\mathcal{M}_X$-level. If we were to only require that $\omega$ be surjective (but not order-preserving), the observational distribution of $\mathcal{M}_X$ may be mapped to an interventional distribution of $\mathcal{M}_Y$, as illustrated by the following example (cf. Figure 3.2 for an illustration).

**Example 7.** Consider the SCM $\mathcal{M}_X = \{S_X, \mathcal{I}_X, \mathcal{P}_E\}$ over $\mathcal{X} = \mathbb{R}^3$ where

\[S_X = \{X_1 = E_1, X_2 = E_2, X_3 = X_1 + X_2 + E_3\}\]
\[\mathcal{I}_X = \{\emptyset, \text{do}(X_2 = 0), \text{do}(X_1 = 0, X_2 = 0)\},\]
\[E_1 \sim \mathbb{P}_{E_1}, \ E_2 = -E_1, \ E_3 \sim \mathbb{P}_{E_3}\]

where $\mathbb{P}_{E_1}$ and $\mathbb{P}_{E_3}$ are arbitrary distributions. Let $\tau : \mathcal{X} \to \mathcal{Y} = \mathbb{R}^2$ be the mapping such that

\[\tau(x_1, x_2, x_3) = (y_1, y_2) = (x_1 + x_2, x_3)\]

Let $\mathcal{M}_Y = \{S_Y, \mathcal{I}_Y, \mathbb{P}_F\}$ be an SCM over $\mathcal{Y}$ with

\[S_Y = \{Y_1 = F_1, Y_2 = Y_1 + F_2\}\]
\[\mathcal{I}_Y = \{\emptyset, \text{do}(Y_1 = 0)\},\]
\[F_1 \sim \mathbb{P}_{E_1}, \ F_2 \sim \mathbb{P}_{E_3}\]
Let \( \omega : \mathcal{I}_X \rightarrow \mathcal{I}_Y \) be defined by
\[
\omega : \begin{cases}
\emptyset & \mapsto \text{do}(Y_1 = 0) \\
d\text{o}(X_2 = 0) & \mapsto \emptyset \\
d\text{o}(X_1 = 0, X_2 = 0) & \mapsto \text{do}(Y_1 = 0)
\end{cases}
\]

Then it is true that \( P^i_{\tau(X)} = P^{\text{do}(\omega(i))}_Y \) for all \( i \in \mathcal{I}_X \), while \( \omega \) is not order-preserving and \( \omega(\emptyset) \neq \emptyset \).

If the SCMs in the above example were used to model the same system, it would be problematic that the observational setting of \( \mathcal{M}_X \)—a description of the system when not having physically performed any intervention—would correspond to an interventional setting in \( \mathcal{M}_Y \), conversely suggesting that the system had been intervened upon.

To avoid the above conflict, we could demand in addition to surjectivity that \( \omega \) map the null intervention of \( \mathcal{M}_X \) to the null intervention of \( \mathcal{M}_Y \). This additional assumption would ensure commutativity of the left-hand part of the diagram in Theorem 6. However, as the following example shows, this would not ensure that the right-hand part of the diagram commutes for all pairs of interventions \( i \leq_X j \), since in this case the arrow from \( P^{\text{do}(\omega(i))}_Y \) to \( P^{\text{do}(\omega(j))}_Y \) may not exist.  

**Example 8.** Let \( X, Y \) and \( \tau \) be as in Example 7. Consider the SCM \( \mathcal{M}_X = \{S_X, \mathcal{I}_X, \mathcal{P}_E\} \) where
\[
S_X = \{X_1 = E_1, X_2 = E_2, X_3 = X_1 + X_2 + E_3\}
\]
\[
\mathcal{I}_X = \{\emptyset, \text{do}(X_2 = 0), \text{do}(X_1 = 0, X_2 = 0)\},
\]
\[
E_1 = 1, \; E_2 \sim \mathcal{P}_E, \; E_3 \sim \mathcal{P}_E
\]

where \( \mathcal{P}_E \) and \( \mathcal{P}_E \) are arbitrary distributions. Let \( \mathcal{M}_Y = \{S_Y, \mathcal{I}_Y, \mathcal{P}_F\} \) be the SCM over \( Y \) with
\[
S_Y = \{Y_1 = 1 + F_1, Y_2 = Y_1 + F_2\}
\]
\[
\mathcal{I}_Y = \{\emptyset, \text{do}(Y_1 = 0), \text{do}(Y_1 = 1)\},
\]
\[
F_1 \sim \mathcal{P}_E, \; F_2 \sim \mathcal{P}_E
\]

---

6 By definition of the poset \( \mathcal{P}_Y \), this arrow exists if and only if \( \omega(i) \leq_Y \omega(j) \).
Let \( \omega : I_X \rightarrow I_Y \) be defined by
\[
\omega : \begin{cases} 
\emptyset \mapsto \emptyset \\
do(X_2 = 0) \mapsto \do(Y_1 = 1) \\
do(X_1 = 0, X_2 = 0) \mapsto \do(Y_1 = 0)
\end{cases}
\]
Then it is true that \( P_i^{\tau(X)} = P_Y^{\do(\omega(i))} \) for all \( i \in I_X \) and \( \omega(\emptyset) = \emptyset \), although \( \omega \) is not order-preserving.

The diagram for above example can be depicted as follows:

If the above SCMs were used as models of the same system, they would not suffer from the problem illustrated in Example 7. Suppose now, however, that we have performed the intervention \( \do(X_2 = 0) \) in \( M_X \), corresponding to the intervention \( \do(Y_1 = 1) \) in \( M_Y \). If we wish to reason about the effect of the intervention \( \do(X_1 = 0, X_2 = 0) \) in \( M_X \), we run into a problem. \( M_X \) suggests that \( \do(X_1 = 0, X_2 = 0) \) could be implemented by performing an additional action on top of \( \do(X_2 = 0) \). In contrast, \( M_Y \) suggests that implementing the corresponding intervention \( \do(Y_1 = 0) \) would conflict with the already performed intervention \( \do(Y_1 = 1) \).

3.3 Examples of Exact Transformations

In the introduction we motivated the problem considered in this chapter by listing three settings in which differing model levels naturally occur. Hav-
ing now introduced the notion of an exact transformation between SCMs, we provide in this section examples of exact transformations falling into each of these categories. The fact that a single framework can be used to draw an explicit correspondence between differing model levels in each of these settings demonstrates the generality of our framework.

Observe that in each of the following examples, the particular set of interventions considered is important. If we were to allow larger sets of interventions $\mathcal{I}_X$ in the SCM $\mathcal{M}_X$, the transformations given would not be exact. This highlights the importance to the causal modelling process of carefully considering the set of interventions.

3.3.1 Marginalisation of Variables

In the following two Theorems we consider two operations that can be performed on SCMs, namely marginalisation of childless or non-intervened variables, and prove that these are exact transformations. That is, an SCM can be simplified into an SCM with fewer variables by either of these operations without losing any causal content concerning the remaining variables.

Thus if the SCM $\mathcal{M}_Y$ can be obtained from another SCM $\mathcal{M}_X$ by successively performing the operations in the following theorems, then $\mathcal{M}_Y$ is an exact transformation of $\mathcal{M}_X$ and hence the two models are causally consistent. This formally explains why we can sensibly consider causal models that focus on a subsystem $\mathcal{M}_Y$ of a more complex system $\mathcal{M}_X$ (cf. Figure 3.3). For a measure-theoretic treatment of marginalisation in SCMs, see [Bon+18].

**Theorem 9** (Marginalisation of childless variables). Let $\mathcal{M}_X = (\mathcal{S}_X, \mathcal{I}_X, \mathcal{P}_E)$ be an SCM and suppose that $\mathbb{I}_Z \subset \mathbb{I}_X$ is a set of indices of variables with no children, i.e. if $i \in \mathbb{I}_Z$ then $X_i$ does not appear in the right-hand side of any structural equation in $\mathcal{S}_X$. Let $\mathcal{Y}$ be the set in which $Y = (X_i : i \in \mathbb{I}_X \setminus \mathbb{I}_Z)$ takes value. Then the transformation $\tau : \mathcal{X} \to \mathcal{Y}$ mapping

$$\tau : (x_i : i \in \mathbb{I}_X) = x \mapsto y = (x_i : i \in \mathbb{I}_X \setminus \mathbb{I}_Z)$$

naturally gives rise to an SCM $\mathcal{M}_Y$ that is an exact $\tau$-transformation of $\mathcal{M}_X$, corresponding to marginalising out the childless variables $X_i$ for $i \in \mathbb{I}_Z$.

**Proof.** By Lemma 5 it suffices to proof this for marginalisation of one childless variable. Without loss of generality, let $X_1$ be the childless variable to be marginalised out.

Let $\mathcal{M}_Y = (\mathcal{S}_Y, \mathcal{I}_Y, \mathcal{P}_F)$ be the SCM where
• the structural equations $S_Y$ are obtained from $S_X$ by removing the structural equation corresponding to the childless variable $X_1$;

• $I_Y$ is the image of the map $\omega : I_X \to I_Y$ that drops any reference to the variable $X_1$ (e.g. $\text{do}(X_1 = x_1, X_2 = x_2) \in I_X$ would be mapped to $\text{do}(X_2 = x_2) \in I_Y$);

• $F = (E_i : i \in I_X \setminus \{1\})$ are the remaining noise variables distributed according to their marginal distribution under $P_F$.

By construction, $\omega$ is surjective and order-preserving. Let $i \in I_X$ be any intervention. The variable $X_1$ being childless ensures that the law on the remaining variables $X_k, k \in I_X \setminus \{1\}$ that we obtain by marginalisation of the childless variable, i.e. $P_i^{\tau(X)}$, is equivalent to the law one obtains by simply dropping the childless variable, which is exactly what the law under $M_Y$ amounts to, i.e. $P_Y^{\omega(\text{do}(i))}$.

\begin{proof}
By Lemma 5 it suffices to proof this for marginalisation of one never-intervened-upon variable. Without loss of generality, let $X_1$ be the never-intervened-upon variable to be marginalised out. By acyclicity of the SCM $M_X$, the structural equation corresponding to variable $X_1$ is of the form $X_1 = f_1 \left( X_{\text{pa}(1)}, E_1 \right)$ and $X_1$ does not appear in the structural equation for any of its ancestors.

Now let $M_Y = (S_Y, I_Y, P_Y)$ be the SCM where

• $I_Y = I_X$;

• $F_i = (E_i, E_1) : i \in I_X \setminus \{1\}$ are the noise variables distributed as implied by $P_E$.

\end{proof}

\textbf{Theorem 10 (Marginalisation of non-intervened variables).}
Let $M_X = (S_X, I_X, P_E)$ be an acyclic SCM and suppose that $I_Z \subset I_X$ is a set of indices of variables that are not intervened upon by any intervention $i \in I_X$. Let $Y$ be the set in which $Y = (X_i : i \in I_X \setminus I_Z)$ takes value. Then the transformation $\tau : \mathcal{X} \to \mathcal{Y}$ mapping

$$\tau : (x_i : i \in I_X) = x \mapsto y = (x_i : i \in I_X \setminus I_Z)$$

naturally gives rise to an SCM $M_Y$ that is an exact $\tau$-transformation of the SCM $M_X$, corresponding to marginalising out the never-intervened-upon variables $X_i$ for $i \in I_Z$.

\textbf{Proof.} By Lemma 5 it suffices to proof this for marginalisation of one never-intervened-upon variable. Without loss of generality, let $X_1$ be the never-intervened-upon variable to be marginalised out. By acyclicity of the SCM $M_X$, the structural equation corresponding to variable $X_1$ is of the form $X_1 = f_1 \left( X_{\text{pa}(1)}, E_1 \right)$ and $X_1$ does not appear in the structural equation for any of its ancestors.

Now let $M_Y = (S_Y, I_Y, P_F)$ be the SCM where

• $I_Y = I_X$;

• $F_i = (E_i, E_1) : i \in I_X \setminus \{1\}$ are the noise variables distributed as implied by $P_E$.
3.3 Examples of exact transformations

Figure 3.3: Suppose that there is a complex model $\mathcal{M}_X$ but that we only wish to model the distribution over $X_1, X_2, X_3$ and how it changes under some interventions on $X_1, X_2, X_3$. By Theorem 9, we can ignore downstream effects (●) after grouping them together as one multivariate variable and by Theorem 10 we can ignore intermediate steps of complex mechanisms (●) and treat upstream causes as noise fluctuations (●). That is, we can exactly transform the complex SCM $\mathcal{M}_X$ into a simpler model $\mathcal{M}_Y$ by marginalisation.

- the structural equations $S_Y$ are obtained from $S_X$ by removing the structural equation of $X_1$ and replacing any occurrence of $X_1$ in the right-hand side of the structural equations of children of $X_1$ by $f_1 \left( X_{pa(1)}, E_1 \right)$, yielding $X_i = f_i \left( f_1 \left( X_{pa(1)}, E_1 \right), X_{pa(i)}, E_i \right)$.

Note that the structural equations of the resulting SCM are still acyclic and are all of the form $X_i = h_i \left( X_{\backslash i}, F_i \right)$.

Then $\mathcal{M}_Y$ is, by construction, an $\tau$-exact transformation of the SCM $\mathcal{M}_X$ for $\omega = id$. \qed

The assumption of acyclicity made in Theorem 10 can be relaxed to allow marginalisation of non-intervened variables in cyclic SCMs, at the expense of extra technical conditions (see Section 3 of [Bon+18]).

We remind the reader that our definition of an SCM does not require that the exogenous $E$-variables be independent. Theorem 10 would not hold if this restriction were made (which is usually the case in the literature); marginalising out a common parent node will in general result in its children having dependent exogenous variables.
3.3.2 Micro- to Macro-Level

Transformations from micro- to macro-levels may arise in situations in which the micro-level variables can be observed via a ‘coarse’ measurement device, represented by the function \( \tau \), e.g. we can use a thermometer to measure the temperature of a gas, but not the motions of the individual particles. They may also arise due to deliberate modelling choice when we wish to describe a system using higher level features, e.g. viewing the motor cortex as a single entity responsible for movements, rather than as a collection of individual neurons.

In such situations, our framework of exact transformations allows one to investigate whether such a macro-level model admits a causal interpretation. The following theorem provides an exact transformation between a micro-level model \( M_X \) and a macro-level model \( M_Y \) in which the variables are aggregate features of variables in \( M_X \) obtained by averaging (cf. Figure 3.4).

**Theorem 11** (Micro- to macro-level). Let \( M_X = (S_X, T_X, P_{E,F}) \) be a linear SCM over the variables \( W = (W_i : 1 \leq i \leq n) \) and \( Z = (Z_i : 1 \leq i \leq m) \) with

\[
S_X = \{W_i = E_i : 1 \leq i \leq n\} \\
\cup \left\{ Z_i = \sum_{j=1}^{n} A_{ij} W_j + F_i : 1 \leq i \leq m \right\}
\]

\( T_X = \{ \emptyset, \text{do}(W = w), \text{do}(Z = z), \text{do}(W = w, Z = z) : w \in \mathbb{R}^n, z \in \mathbb{R}^m \} \)

and \((E, F) \sim P\) where \( P\) is any distribution over \( \mathbb{R}^{n+m}\) and \( A\) is a matrix.

Assume that there exists an \( a \in \mathbb{R}\) such that each column of \( A\) sums to \( a\). Consider the following transformation that averages the \( W\) and \( Z\) variables:

\[
\tau : \mathcal{X} \to \mathcal{Y} = \mathbb{R}^2 \\
\begin{pmatrix} W \\ Z \end{pmatrix} \mapsto \begin{pmatrix} \hat{W} \\ \hat{Z} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} W_i \\ \frac{1}{m} \sum_{j=1}^{m} Z_j \end{pmatrix}
\]
Further, let $M_Y = (S_Y, I_Y, \mathbb{P}_{E, F})$ over the variables $\{\hat{W}, \hat{Z}\}$ be an SCM with

$$S_Y = \{\hat{W} = \hat{E}, \hat{Z} = \frac{a}{m} \hat{W} + \hat{F}\}$$

$$I_Y = \{\emptyset, \text{do}(\hat{W} = \hat{w}), \text{do}(\hat{Z} = \hat{z}), \text{do}(\hat{W} = \hat{w}, \hat{Z} = \hat{z}) : \hat{w} \in \mathbb{R}, \hat{z} \in \mathbb{R}\}$$

$$\hat{E} \sim \frac{1}{n} \sum_{i=1}^{n} E_i, \quad \hat{F} \sim \frac{1}{m} \sum_{i=1}^{m} F_i$$

Then $M_Y$ is an exact $\tau$-transformation of $M_X$.

**Proof.** We begin by defining a mapping between interventions

$$\omega : I_X \rightarrow I_Y$$

$$\emptyset \mapsto \emptyset$$

$$\text{do}(W = w) \mapsto \text{do} \left( \hat{W} = \frac{1}{n} \sum_{i=1}^{n} w_i \right)$$

$$\text{do}(Z = z) \mapsto \text{do} \left( \hat{Z} = \frac{1}{m} \sum_{i=1}^{m} z_i \right)$$

$$\text{do}(W = w, Z = z) \mapsto \text{do} \left( \hat{W} = \frac{1}{n} \sum_{i=1}^{n} w_i, \hat{Z} = \frac{1}{m} \sum_{i=1}^{m} z_i \right)$$

Note that $\omega$ is surjective and order-preserving (in fact, it is an order embedding). Therefore, it only remains to show that the distributions implied by $\tau(X)$ under any intervention $i \in I_X$ agree with the corresponding distributions implied by $M_Y$. That is, we have to show that

$$\mathbb{P}_{\tau(X)}^i = \mathbb{P}_{Y}^{\text{do}(\omega(i))} \quad \forall i \in I_X$$

In the observational setting, the distribution over $Y$ is implied by the following equations:

$$\hat{W} = \frac{1}{n} \sum_{i=1}^{n} W_i = \frac{1}{n} \sum_{i=1}^{n} E_i$$

$$\hat{Z} = \frac{1}{m} \sum_{i=1}^{m} Z_i = \frac{1}{m} \sum_{i=1}^{m} \left( \sum_{j=1}^{n} A_{ij} W_j + F_i \right) = \frac{a}{m} \hat{W} + \frac{1}{m} \sum_{i=1}^{m} F_i$$
Figure 3.4: An illustration of the setting considered in Theorem 11. The micro-variables $W_1, \ldots, W_n$ and $Z_1, \ldots, Z_m$ in the SCM $M_X$ can be averaged to derive macro-variables $\hat{W}$ and $\hat{Z}$ in such a way that the resulting macro-level SCM $M_Y$ is an exact transformation of the micro-level SCM $M_X$.

Since the distributions of the exogenous variables in the SCM $M_Y$ are given by $\hat{E} \sim \frac{1}{n} \sum_{i=1}^{n} E_i$ and $\hat{F} \sim \frac{1}{m} \sum_{i=1}^{m} F_i$, it follows that $P_{\text{do}(\emptyset)}^{\tau(X)}$ and $P_{\text{do}(\emptyset)}^{\tau(Y)}$ agree. Similarly, the push-forward measure on $Y$ induced by the intervention $\text{do}(W = w) \in \mathcal{I}_X$ is given by

$$
\hat{W} = \frac{1}{n} \sum_{i=1}^{n} W_i = \frac{1}{n} \sum_{i=1}^{n} w_i
$$

$$
\hat{Z} = \frac{1}{m} \sum_{i=1}^{m} Z_i = \frac{1}{m} \sum_{i=1}^{m} \left( \sum_{j=1}^{n} A_{ij} W_j + F_i \right) = \frac{a}{m} \hat{W} + \frac{1}{m} \sum_{i=1}^{m} F_i
$$

which is the same as the distribution induced by the $\omega$-corresponding intervention $\text{do}(\hat{W} = \frac{1}{n} \sum_{i=1}^{n} w_i)$ in $M_Y$.

Similar reasoning shows that this also holds for the other interventions $\text{do}(Z = z)$ and $\text{do}(W = w, Z = z)$.

3.3.3 Stationary Behaviour of Dynamical Processes

In this section we provide an example of an exact transformation between an SCM $M_X$ describing a time-evolving system and another SCM $M_Y$ describing the system after it has equilibrated. In this setting, $\tau$ could be thought of as representing our ability to only measure the time-evolving system at a single point in time, after the transient dynamics have taken place.
In particular, we consider a discrete-time linear dynamical system with identical noise and provide the explicit form of an SCM that models the distribution of the equilibria under each intervention (cf. Figure 3.5).

**Theorem 12** (Discrete-time linear dynamical process with identical noise). Let $\mathcal{M}_X = (S_X, \mathcal{I}_X, \mathbb{P}_E)$ over the variables $\{X^i_t : t \in \mathbb{Z}, i \in \{1, \ldots, n\}\}$ be a linear SCM with

$$S_X = \left\{ X^i_{t+1} = \sum_{j=1}^{n} A_{ij} X^j_t + E^i_t : i \in \{1, \ldots, n\}, t \in \mathbb{Z} \right\}$$

i.e. $X_{t+1} = AX_t + E_t$

$$\mathcal{I}_X = \left\{ \text{do}(X^i_t = x_j \ \forall t \in \mathbb{Z}, \forall j \in J) : x \in \mathbb{R}^{|J|}, J \subseteq \{1, \ldots, n\} \right\}$$

$E_t = E \ \forall t \in \mathbb{Z}$ where $E \sim \mathbb{P}$

where $\mathbb{P}$ is any distribution over $\mathbb{R}^n$ and $A$ is a matrix.

Assume that the linear mapping $v \mapsto Av$ is a contraction. Then the following transformation is well-defined under any intervention $i \in \mathcal{I}_X$:  

$$\tau : \mathcal{X} \rightarrow \mathcal{Y} \quad (x_t)_{t \in \mathbb{Z}} \mapsto y = \lim_{t \rightarrow \infty} x_t$$

Let $\mathcal{M}_Y = (S_Y, \mathcal{I}_Y, \mathbb{P}_F)$ be the (potentially cyclic) SCM over the $\mathcal{Y}$-level variables $\{Y^i : i \in \{1, \ldots, n\}\}$ with

$$S_Y = \left\{ Y^i = \frac{\sum_{j \neq i} A_{ij} Y^j}{1 - A_{ii}} + \frac{F^i}{1 - A_{ii}} : i \in \{1, \ldots, n\} \right\}$$

$$\mathcal{I}_Y = \left\{ \text{do}(Y^j = y_j \ \forall j \in J) : y \in \mathbb{R}^{|J|}, J \subseteq \{1, \ldots, n\} \right\}$$

$F \sim \mathbb{P}$

Then $\mathcal{M}_Y$ is an exact $\tau$-transformation of $\mathcal{M}_X$.  

---

7 Note that the assumption that the transition dynamics be linear can be relaxed to more general non-linear mappings. In this case, however, the structural equations of $\mathcal{M}_Y$ can only be written in terms of implicit solutions to the structural equations of $\mathcal{M}_X$. For purposes of exposition, we stick here to the simpler case of linear dynamics.

8 In Appendix A.1.2 we show that $A$ being a contraction mapping ensures that the sequence $(X_t)_{t \in \mathbb{Z}}$ defined by $\mathcal{M}_X$ converges everywhere under any intervention $i \in \mathcal{I}_X$. That is, for any realisation $(x_t)_{t \in \mathbb{Z}}$ of this sequence, its limit $\lim_{t \rightarrow \infty} x_t$ as a sequence of elements of $\mathbb{R}^n$ exists.
Proof. We begin by defining a mapping between interventions

\[ \omega : \mathcal{I}_X \to \mathcal{I}_Y \]

\[ \text{do}(X^i_t = x_j \ \forall t \in \mathbb{Z}, \forall j \in J) \mapsto \text{do}(Y^i = x_j \ \forall j \in J) \]

Note that \( \omega \) is surjective and order-preserving (in fact, it is an order embedding). Therefore, it only remains to show that the distributions implied by \( \tau(X) \) under any intervention \( i \in \mathcal{I}_X \) agree with the corresponding distributions implied by \( \mathcal{M}_Y \). That is, we have to show that

\[ \mathbb{P}_i^\tau(X) = \mathbb{P}_Y^{\text{do}(\omega(i))} \quad \forall i \in \mathcal{I}_X \]

For this we consider, without loss of generality, the distribution arising from performing the \( \mathcal{M}_X \)-level intervention

\[ i = \text{do}(X^i_t = x_j \ \forall t \in \mathbb{Z}, \forall j \leq m \leq n) \in \mathcal{I}_X \]

for \( m \in [n] \) (for \( m = 0 \) this amounts to the null-intervention).

Since \( A \) is a contraction mapping, it follows from Lemma 29 in Appendix A.1.2 that for any intervention in \( \mathcal{I}_X \), the sequence of random variables \( X_t \) defined by \( \mathcal{M}_X \) converges everywhere. That is, there exists a random variable \( X_* \) such that \( X_t \xrightarrow{\text{everywhere}} t \to \infty X_* \). In the case of the intervention \( i \) above, the random variable \( X_* \) satisfies:

\[
\begin{cases}
X_*^k = x_k & \text{if } k \leq m \\
X_*^k = \sum_j A_{kj} X_*^j + E_k & \text{if } m < k \leq n
\end{cases}
\tag{3.3.1}
\]

Since \( \tau(X) = \lim_{t \to \infty} X_t \), it follows from the definition of \( X_* \) that \( \tau(X) = X_* \), and hence \( \tau(X) \) also satisfies the equations above. It follows (rewriting the second line in Equation 3.3.1 above) that under the push-forward measure \( \mathbb{P}_i^\tau(X) = \tau \left( \mathbb{P}_X^{\text{do}(i)} \right) \) the distribution of the random variable \( \tau(X) = X_* \) is given by:

\[
\begin{cases}
X_*^k = x_k & \text{if } k \leq m \\
X_*^k = \sum_{j \neq k} A_{kj} X_*^j + \frac{E_k}{1 - A_{kk}} & \text{if } m < k \leq n
\end{cases}
\]

We need to compare this to the law of \( Y \) as implied by \( \mathcal{M}_Y \) under the intervention \( \omega(i) \), i.e. \( \mathbb{P}_Y^{\omega(i)} \). The \( \mathcal{M}_Y \)-level intervention \( \omega(i) \) corresponding to \( i \) is

\[ \omega(i) = \text{do}(Y^i = x_j \ \forall j \leq m \leq n) \in \mathcal{I}_Y \]
and so the structural equations of $\mathcal{M}_Y$ under the intervention $\omega(\text{do}(i))$ are

$$
\begin{cases}
Y^k = x_k & \text{if } k \leq m \\
Y^k = \frac{\sum_{j \neq k} A_{kj} Y^j}{1 - A_{kk}} + \frac{F^k}{1 - A_{kk}} & \text{if } m < k \leq n
\end{cases}
$$

Since $F \sim E$ it indeed follows that $\tau(X) \sim Y$, i.e. $\mathbb{P}_{\tau(X)}^i = \mathbb{P}_{Y}^{\text{do}(\omega(i))}$.

Thus $\mathcal{M}_Y$ is an exact $\tau$-transformation of $\mathcal{M}_X$. $\square$

The above theorem demonstrates how a linear additive SCM can arise as a result of making observations of a dynamical process. This supports one interpretation of SCMs as a description of a dynamical process that equilibrates quickly compared to its external environment. The framework of exact transformations allows us to explain in a precise way the sense in which such equilibrium models can be used as causal descriptions of an underlying dynamical process.

This result also sheds light on the interpretation of cyclic causal models. One interpretation of the structural equations of an acyclic SCM is that they

---

9 This interpretation corresponds to the assumption that the noise in the dynamical model is constant through time, and is used by e.g. [Lac+08; Moo+11; HEH12; MJS13] and [MH13] to meaningfully interpret cyclic SCMs.
represent a temporally ordered series of mechanisms by which data are generated. This is not possible in the case that the SCM exhibits cycles: there does not exist a partial ordering on the variables and hence one cannot think of each variable being generated temporally downstream of its parents. By showing that cyclic SCMs can arise as exact transformations of acyclic SCMs, we provide an interpretation of cyclic SCMs that does not suffer from the above problem.

3.4 Example of an Impossibility Result Precluding Exactness

In the previous sections we have seen how the ability to causally reason about a system breaks under inexact transformations and have proved exactness of several transformations. We provide an impossibility result in this section which motivates a generalisation of the notion of exact transformations to approximate transformations in the next section.

By the concept of an exact transformation we can link two causal models $M_X$ and $M_Y$ via a variable transformation $\tau$. There is many ways in which exactness of a transformation can fail; for example

- we may not be able to exactly transform a causal model if the model class that we consider for $M_Y$ is too restricted, e.g. if there are non-linear structural equations in $M_X$ but we only allow linear structural equations for $M_Y$, then even with $\tau$ being the identity mapping we may not obtain an exact transformation;

- we cannot marginalise out a variable if we wish to reason about the effects of intervening on that very variable again emphasising the importance of a restricted intervention set;

- we cannot map two variables $X_1$ and $X_2$ to the sum $Y_1 = X_1 + X_2$ if there is interventions on both of them in $I_X$ that result in the same total value, say $m$, but imply different distributions on the remaining variables, i.e. if $P_{X}^{do}(X_1=x_1, X_2=m-x_1) \neq P_{X}^{do}(X_1=x'_1, X_2=m-x'_1)$ for some $x_1, x'_1$, then we run into a problem since we necessarily need to identify both $X$-level interventions $do(X_1 = x_1, X_2 = m - x_1)$ and $do(X_1 = x'_1, X_2 = m - x'_1)$ with the same $Y$-level intervention, say $do(Y_1 = m)$, while we cannot disambiguate anymore between the corresponding conflicting distributions;
• the proof of Theorem 11 crucially depends on each column of $A$ summing to $a$, if that was not the case the micro- to macro-mapping discussed therein would not be exact.

In the following we show another such instance of the impossibility of an exact transformation: The stationary distributions of a discrete-time linear dynamical system with iid Gaussian noise cannot be modelled by a linear Gaussian SCM which is in contrast to Theorem 12 where exactness of this transformation was proven for identical noise. We provide an illustration of the considered scenario in Figure 3.6. A similar scenario is also presented in [JRS18] as a starting point for a more exhaustive analysis of the conditions for bivariate linear autoregressive models to be exactly transformable; in particular, the disturbing result is that exactness requires different coarse-grainings for $X$ and $Y$ instead of mapping both to their stationary distributions.
Theorem 13 (Impossibility of an exact transformation of a VAR process). Consider the following SCM $\mathcal{M}_X$ defined over the variables $\{A_t, B_t : t \in \mathbb{Z}\}$

$$S_X = \{A_{t+1} = \alpha A_t + E_t, \quad B_{t+1} = \beta A_t + \gamma B_t + F_t : t \in \mathbb{Z}\}$$

$$I_X = \{\emptyset\} \cup \{\text{do}(A_t = a \ \forall t \in \mathbb{Z}) : a \in \mathbb{R}\} \cup \{\text{do}(B_t = b \ \forall t \in \mathbb{Z}) : b \in \mathbb{R}\}$$

$E_t, F_t \overset{iid}{\sim} \mathcal{N}(0, 1) \ \forall t \in \mathbb{Z}$

Further assume that $|\alpha|, |\beta|, |\gamma| < 1$, which ensures that under any intervention the process is a stable stationary VAR process and the following transformation is well-defined:\footnote{Since $|\alpha|, |\beta|, |\gamma| < 1$, the map $v \mapsto (\alpha \ 0 \ \beta \ \gamma) v$ is a contraction mapping. As shown in Appendix A.1.1 and similar to Lemma 29 in Appendix A.1.2, this ensures that the sequence of $(A_t, B_t)$ is, under any intervention $i \in I_X$, transitioning according to a contraction and is, by Lemma 28 in Appendix A.1.1 of this thesis as well as Section 2.1.1 and Proposition 2.1 in [Lüt05], a stable and stationary first-order VAR process. The transformation $\tau$ maps the process to its stationary distribution.}

$$\tau : \mathcal{X} \to \mathcal{Y}$$

$$(x_t)_{t \in \mathbb{Z}} \mapsto y = \lim_{t \to \infty} x_t$$

If $\beta \neq 0$, then there exists no linear additive SCM with Gaussian noise $\mathcal{M}_Y$ that is an exact $\tau$-transformation of $\mathcal{M}_X$, i.e. no linear Gaussian model of the form

$$S_Y = \begin{cases} 
Y_1 = pY_2 + qE_1 \\
Y_2 = rY_1 + sE_2 
\end{cases}$$

$$I_Y$$

$\mathbb{P}_E \sim \mathcal{N}(\mu, \Lambda)$

can explain the stationary distributions under any intervention of the process defined by $\mathcal{M}_X$.

Proof. Note, that, without loss of generality, we can assume $q = s = 1$. 
We will first give the form of the distributions implied by $\mathcal{M}_X$ and $\tau$:\footnote{These can be computed as the stationary distributions of the corresponding stable first-order VAR processes using Equation (2.1.32) in [Lüt05].}

\[
\Pr_{\tau(X)}^{\text{do}(\emptyset)} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} S_A & S_A \frac{\alpha \beta}{1-\alpha \gamma} \\ \frac{\alpha \beta}{1-\alpha \gamma} S_A & S_B + S_A \frac{\beta^2 (1+\alpha \gamma)}{(1-\alpha \gamma)(1-\gamma^2)} \end{pmatrix}\right)
\]

\[
\Pr_{\tau(X)}^{\text{do}(A_t=a \ \forall t \in \mathbb{Z})} \sim \mathcal{N}\left(\begin{pmatrix} a \\ \frac{\beta}{1-\gamma} a \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & S_B \end{pmatrix}\right) \quad \text{for } a \in \mathbb{R}
\]

\[
\Pr_{\tau(X)}^{\text{do}(B_t=b \ \forall t \in \mathbb{Z})} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ b \end{pmatrix}, \begin{pmatrix} S_A & 0 \\ 0 & 0 \end{pmatrix}\right) \quad \text{for } b \in \mathbb{R}
\]

where $S_A = \frac{1}{1-\alpha^2}$ and $S_B = \frac{1}{1-\gamma^2}$.

Observe that the above distributions fully determine $\omega$, i.e.

\[
\omega : \mathcal{I}_X \rightarrow \mathcal{I}_Y
\]

\[
do(A_t = a \ \forall t \in \mathbb{Z}) \rightarrow \do(Y_1 = a) \quad \text{for } a \in \mathbb{R}
\]

\[
do(B_t = b \ \forall t \in \mathbb{Z}) \rightarrow \do(Y_2 = b) \quad \text{for } b \in \mathbb{R}
\]

\[
\emptyset \rightarrow \emptyset
\]

We require that $\Pr_{\tau(X)}^\tau = \Pr_{\tau(X)}^{\text{do}(\omega(i))}$ for $\mathcal{M}_Y$ to be an exact $\tau$-transformation of $\mathcal{M}_X$. We demonstrate, that this requirement puts constraints on the parameters of $\mathcal{M}_Y$ that cannot be satisfied if $\beta \neq 0$.

- Firstly, if $\Pr_{\tau(Y)}^{\text{do}(B=b)}$ should be equal to $\Pr_{\tau(X)}^{\text{do}(B_t=b \ \forall t \in \mathbb{Z})}$ for all $b \in \mathbb{R}$, then this implies that $p = 0$, $\mu_1 = 0$ and $\Lambda_{1,1} = S_A$.

- Secondly, if $\Pr_{\tau(Y)}^{\text{do}(A=a)}$ should be equal to $\Pr_{\tau(X)}^{\text{do}(A_t=a \ \forall t \in \mathbb{Z})}$ for all $a \in \mathbb{R}$, then this implies that $r = \frac{\beta}{1-\gamma}$, $\mu_2 = 0$ and $\Lambda_{2,2} = S_B$.

Up to here, the only remaining free parameter in $\mathcal{M}_Y$ is $\Lambda_{1,2}$ and we have

\[
\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \begin{pmatrix} E_1 \\ \frac{\beta}{1-\gamma} Y_1 + E_2 \end{pmatrix} \quad \text{where} \quad \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} S_A & \Lambda_{1,2} \\ \Lambda_{1,2} & S_B \end{pmatrix}\right)
\]
such that the $\mathcal{M}_Y$ model implies the following distribution over $(Y_1, Y_2)$

$$
\begin{pmatrix}
Y_1 \\
Y_2
\end{pmatrix} \sim \mathcal{N}
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
S_A & \Lambda_{1,2} + S_A \frac{\beta}{1-\gamma} \\
\Lambda_{1,2} + S_A \frac{\beta}{1-\gamma} & S_B + 2 \frac{\beta}{1-\gamma} \Lambda_{1,2} + \left( \frac{\beta}{1-\gamma} \right)^2 S_A
\end{pmatrix}
$$

That is, for the observational distributions to agree we require that the covariance of $(Y_1, Y_2)$ above agrees with the covariance of $\mathcal{P}^\text{do}(\emptyset)_{\tau(X)}$. In particular, $\Lambda_{1,2}$ needs to satisfy

$$
S_A \frac{\alpha \beta}{1-\alpha \gamma} = \Lambda_{1,2} + S_A \frac{\beta}{1-\gamma}
$$

$$
S_B + S_A \frac{\beta^2 (1+\alpha \gamma)}{(1-\alpha \gamma)(1-\gamma^2)} = S_B + 2 \frac{\beta}{1-\gamma} \Lambda_{1,2} + \left( \frac{\beta}{1-\gamma} \right)^2 S_A
$$

Rewriting the first and second equation we obtain

$$
\Lambda_{1,2} = \frac{-\beta}{(1-\alpha \gamma)(\alpha + 1)(1-\gamma)}
$$

$$
\Lambda_{1,2} = \frac{-\beta \gamma}{(1-\alpha \gamma)(\alpha + 1)(1-\gamma^2)}
$$

which has a solution if and only if

$$
0 = \frac{\beta}{(\gamma^2 - 1)(\alpha \gamma - 1)(\alpha + 1)}
$$

That is, if and only if $\beta = 0$, then there exists a linear additive SCM with Gaussian noise $\mathcal{M}_Y$ that is an exact $\tau$-transformation of $\mathcal{M}_X$.

3.5 APPROXIMATE TRANSFORMATIONS BETWEEN SCMS

Trade-offs between model complexity and model accuracy are commonly considered in statistical modelling; the following definition extends the notion of an exact transformation between SCMs and provides a way to quantify this trade-off when using SCMs for causal modelling.

**Definition 14** (Approximate Transformations of SCMs). Let $\mathcal{M}_X$ and $\mathcal{M}_Y$ be SCMs, $\tau: \mathcal{X} \to \mathcal{Y}$ and $\epsilon < \infty$. We say $\mathcal{M}_Y$ is an $\epsilon$-approximate $\tau$-transformation of $\mathcal{M}_X$ if there exists a surjective order-preserving map $\omega: \mathcal{I}_X \to \mathcal{I}_Y$ such that

$$
\text{KL} \left[ \mathcal{P}^i_{\tau(X)} \parallel \mathcal{P}^\text{do}(\omega(i)) \right] \leq \epsilon \quad \forall i \in \mathcal{I}_X
$$

where $\mathcal{P}^i_{\tau(X)}$ denotes the distribution of the $\mathcal{Y}$-valued random variable $\tau(X)$ with $X \sim \mathcal{P}^\text{do}(i)$.
For two distributions $P$ and $Q$, the KL divergence $KL[P\|Q]$ can be thought of as a measure of the error induced by trying to approximate $P$ by using $Q$. Thus if $M_Y$ is an $\epsilon$-approximate $\tau$-transformation of $M_X$, we can think of $M_Y$ as approximating the push-through measures $P_{\tau(X)}$ with an ‘approximation error’ that is uniformly bounded by $\epsilon$.

Restrictive conditions need to be met for a transformation to be exact. There may be cases in which, for a particular $\tau$ (corresponding to either deliberate model design or constrained measurement ability), the poset $P_{\tau(X)}$ is not implied by any SCM within a particular desired model class. For instance, as shown in Theorem 13 in the previous section, the stationary distributions of a discrete-time linear dynamical system $M_X$ with iid Gaussian noise cannot in general be modelled by a linear Gaussian SCM $M_Y$ that is an exact transformation of $M_X$. The following theorem demonstrates that we can, however, bound the error induced when trying to do so.

**Theorem 15** (Approximate transformation of a VAR process).

Let $M_X = (S_X, I_X, P_E)$ over the variables $\{X_i^t : t \in \mathbb{Z}, i \in \{1, \ldots, n\}\}$ be an SCM with

\[
S_X = \left\{ X_{t+1}^i = \sum_{j=1}^n A_{ij} X_t^j + E_t^i : i \in \{1, \ldots, n\}, t \in \mathbb{Z} \right\}
\]

i.e. $X_{t+1} = AX_t + E_t$

$I_X = \left\{ \text{do}(X_t^j = x_j \ \forall t \in \mathbb{Z}, \forall j \in J) : x \in \mathbb{R}^{|J|}, J \subseteq \{1, \ldots, n\} \right\}$,

$E_t \overset{iid}{\sim} \mathcal{N}(\mu, \Lambda) \ \forall t \in \mathbb{Z}$

Assume that the linear mapping $v \mapsto Av$ is a contraction, which ensures that under any intervention the process is a stable stationary VAR process and the following transformation is well-defined:\footnote{As shown in Appendix A.1.1 and similar to Lemma 29 in Appendix A.1.2, $v \mapsto Av$ being a contraction ensures that the sequence of $X_t$ is, under any intervention $i \in I_X$, transitioning according to a contraction and is, by Lemma 28 in Appendix A.1.1 of this thesis as well as Section 2.1.1 and Proposition 2.1. in [Lüt05], a stable and stationary first-order VAR process. The transformation $\tau$ maps the process to its stationary distribution.}

\[
\tau : \mathcal{X} \rightarrow \mathcal{Y}
\]

\[
(x_t)_{t \in \mathbb{Z}} \mapsto y = \lim_{t \rightarrow \infty} x_t
\]
Further, let $\mathcal{M}_Y = (S_Y, I_Y, P_F)$ over the variables $\{Y^i : i \in \{1, \ldots, n\}\}$ be an SCM with

$$S_Y = \left\{ Y^i = \frac{\sum_{j \neq i} A_{ij} Y^j}{1 - A_{ii}} + \frac{F^i}{1 - A_{ii}} : i \in \{1, \ldots, n\}\right\}$$

$$I_Y = \left\{ \text{do}(Y^j = y_j : \forall j \in J) : y \in \mathbb{R}^{|J|}, J \subseteq \{1, \ldots, n\}\right\}$$

$$F \sim \mathcal{N}(\mu, \Lambda)$$

Then $\mathcal{M}_Y$ is an $\epsilon$-approximate $\tau$-transformation of $\mathcal{M}_X$. In particular, $\epsilon$ is given by

$$\epsilon = \max_{J \subseteq \{1, \ldots, n\}} \frac{1}{2} \left( \text{tr} \left( \left( \Psi^I_{Y, \tau(X)} \right)^{-1} \Psi^I_{\tau(X)} \right) - (n - |J|) + \ln \left( \frac{\det \Psi^I_{Y, \tau(X)}}{\det \Psi^I_{\tau(X)}} \right) \right)$$

where

$$\Psi^I_Y = \left( I - A^I \right)^{-1} \Lambda^I \left( I - A^I \right)^{-\top}$$

$$\Psi^I_{\tau(X)} = \sum_{k=0}^{\infty} \left( A^I \right)^k \Lambda^I \left( \left( A^I \right)^\top \right)^k$$

where for any $n \times n$ matrix $W$ and any subset $J \subseteq \{1, \ldots, n\}$, $W^J$ denotes the submatrix of $W$ consisting of the rows and columns with indices in $J$.

**Proof.** We begin by defining a mapping between interventions

$$\omega : I_X \rightarrow I_Y$$

$$\text{do}(X^i_t = x_j \ \forall t \in \mathbb{Z}, \forall j \in J) \mapsto \text{do}(Y^i = x_j \ \forall j \in J)$$

Note that $\omega$ is surjective and order-preserving (in fact, it is an order embedding). Therefore, it only remains to show that the distributions implied by $\tau(X)$ under any intervention $i \in I_X$ are $\epsilon$-close to the corresponding distributions implied by $\mathcal{M}_Y$. That is, we have to show that

$$\text{KL} \left[ P^i_{\tau(X)} \left| \right| P^\text{do}(\omega(i))_Y \right] \leq \epsilon \ \forall i \in I_X$$

For this we consider, without loss of generality, the distribution arising from performing the $\mathcal{M}_X$-level intervention

$$i = \text{do}(X^i_t = x_j \ \forall t \in \mathbb{Z}, \forall j \leq m \leq n) \in I_X$$
for $m \in [n]$ (for $m = 0$ this amounts to the null-intervention). Under the intervention $i$ the structural equations defining the sequence $X_t$ can be written as

$$
\begin{pmatrix}
X_{t+1}^\bullet \\
X_t^\bullet
\end{pmatrix} = \begin{pmatrix}
I & 0 \\
A^\bullet & A_t^\bullet
\end{pmatrix} \begin{pmatrix}
X_{t+1}^\bullet \\
X_t^\bullet
\end{pmatrix} + \begin{pmatrix}
0 \\
E_t^\bullet
\end{pmatrix}
$$

In particular, the sequence of random variables $X_t^\bullet$ can be rewritten as

$$
X_{t+1}^\bullet = \left( A^\bullet X_t^\bullet + \mu_t^\bullet \right) + A^\bullet X_t^\bullet + \left( E_t^\bullet - \mu_t^\bullet \right)
$$

which is a first-order vector autoregressive (VAR) process. By Lemma 27 the mapping $\omega^\bullet \mapsto A^\bullet \omega^\bullet$ is a contraction mapping, and thus, by Lemma 28, all eigenvalues of $A^\bullet$ have modulus less than 1.\footnote{Lemma 27 and 28 can be found in Appendix A.1.1.} Therefore, $(X_t^\bullet)_{t \in \mathbb{Z}}$ is a stable VAR process with the following stationary distribution (refer to Section 2.1.1 and Proposition 2.1 in [Lüt05] for this result):

$$
X_\infty^\bullet \sim \mathcal{N} \left( \left( I - A^\bullet \right)^{-1} \nu, \sum_{k=0}^{\infty} \left( \left( A^\bullet \right)^k \Lambda^\bullet \left( \left( A^\bullet \right)^\top \right)^k \right) \Psi_{\tau(X)} \right)
$$

Since $\tau(X) = X_\infty$ maps the process to its stationary distribution, it follows that under the push-through measure $\mathbb{P}_{\tau(X)}^i = \tau \left( \mathbb{P}_X^{do(i)} \right)$ we have

$$
\mathbb{P}_{\tau(X)}^i \sim \begin{pmatrix} X_t^\bullet \\ X_\infty^\bullet \end{pmatrix} \quad \text{where} \quad X_\infty^\bullet \sim \mathcal{N} \left( z, \Psi_{\tau(X)} \right)
$$

We need to compare this to the law of $Y$ as implied by $\mathcal{M}_Y$ under the intervention $\omega(i)$, i.e. $\mathbb{P}_Y^{do(\omega(i))}$. The $\mathcal{M}_Y$-level intervention $\omega(i)$ corresponding to $i$ is

$$
\omega(i) = do(Y^j = x_j \forall j \leq m \leq n) \in \mathcal{I}_Y
$$

where e.g. $X_{t+1}^\bullet \in \mathbb{R}^m$, $X_t^\bullet \in \mathbb{R}^{n-m}$, $A^\bullet \in \mathbb{R}^{(n-m) \times m}$ and $A^\bullet \in \mathbb{R}^{(n-m) \times (n-m)}$.
and so the structural equations of $\mathcal{M}_Y$ under the intervention $\omega(\text{do}(i))$ are

$$
\begin{cases}
Y^k = x_k & \text{if } k \leq m \\
Y^k = \sum_{j \neq k} A_{kj} Y^j + \frac{F^k}{1-A_{kk}} & \text{if } m < k \leq n
\end{cases}
$$

which we can rewrite as

$$(Y^*)^T = (x^*, A^*x^* + A^*Y^* + F^*)$$

It follows, bringing $Y^*$ to the left-hand side and since $F \sim \mathcal{N}(\mu, \Lambda)$, that

$$Y^* = (I - A^*)^{-1} (A^*x^* + F^*)$$

is the distribution of $Y^*$ under $\mathbb{P}^{\text{do}(\omega(i))}$, while the remaining variables are constant $Y^* = x^*$.

For the KL divergence between the distributions we obtain

$$KL\left[\mathbb{P}^i_{\tau(X)} \mid \mid \mathbb{P}^{\text{do}(\omega(i))}_Y\right] = KL\left[\mathcal{N}(z, \Psi_{\tau(X)}) \mid \mid \mathcal{N}(z, \Psi_Y)\right]$$

$$= \frac{1}{2} \left( \text{tr} \left( (\Psi_{\tau(X)}^j)^{-1} \Psi_{\tau(X)}^j \right) - (n - m) + \ln \left( \frac{\det \Psi_{\tau(X)}^j}{\det \Psi_Y} \right) \right)$$

Observing that we could have chosen any other subset $J \subseteq [n]$ of variables to intervene on and we would have obtained the same calculation (but with different submatrices, e.g. $A^j$ instead of $A^*$), we see that $\epsilon$ in the statement of the theorem corresponds to taking the maximum of the above KL divergence over all possible subsets $J \subseteq [n]$.

This is guaranteed to be bounded as the specific values $x^*$ to which the intervened variables are set do not affect the above KL divergence. Thus, since this is finite for each intervention, the maximum is taken over a finite set of finite values and so is itself finite. Thus, it follows that

$$KL\left[\mathbb{P}^i_{\tau(X)} \mid \mid \mathbb{P}^{\text{do}(\omega(i))}_Y\right] \leq \max_{j \in \mathcal{I}_X} KL\left[\mathbb{P}^j_{\tau(X)} \mid \mid \mathbb{P}^{\text{do}(\omega(j))}_Y\right] = \epsilon \quad \forall i \in \mathcal{I}_X$$

which closes the proof.
3.6 DISCUSSION AND FUTURE WORK

It’s turtles all the way down! There is no such thing as a ‘correct’ model, but in \([R^W + 17]\) we introduced the notions of exact transformations between SCMs to evaluate when two SCMs can be viewed as causally consistent models of the same system. Illustrating how these notions can be used in order to relate differing model levels, we proved in Section 3.3 the exactness of transformations occurring in three different settings. These have implications for the following questions in causal modelling: When can we model only a subsystem of a more complex system? When does a micro-level system admit a causal description in terms of macro-level features? How do cyclic causal models arise?

Our work has implications for other problems in causal modelling. It suggests that ambiguous manipulations [SS04] may be thought of as arising due to the application of an inexact transformation to an SCM \(M_x\). This was illustrated in Section 3.1.1 in which LDL and HDL cholesterol were only measured via their sum TC, resulting in a model that suffered from the problem of ambiguous manipulations (cf. Figure 3.1a) since it was not an exact transformation of the underlying model (cf. Figure 3.1b). This is related to the problem of causal variable definition as studied by [Ebe16].

A future line of enquiry would be to further pursue generalised notions of an exact transformation in order to analyse the trade-off between model accuracy and model complexity for causal modelling using SCMs. For a transformation to be exact, we require that the posets \(\mathcal{P}_{\tau(x)}\) and \(\mathcal{P}_Y\) be equal. One could imagine and further analyse a ‘softening’ of this requirement such that the distributions in the posets are required to be only approximately equal similar to the preliminary results in Section 3.5. A slightly inaccurate model with a small number of variables may be preferable to an accurate but complex model.

We discussed the importance of an order-preserving \(\omega\) to ensure a notion of causal consistency between two SCMs. It would be interesting to better understand the conditions under which different properties of consistency between causal models hold – for instance, counterfactual reasoning, which we have not discussed.

While we have introduced the notion of an exact transformation, we have not provided any criterion to choose from amongst the set of all possible exact transformations of an SCM. Foundational work in a similar direction to ours has been done by [CPE15; CEP16], who consider a particular discrete setting. They provide algorithms to learn a transformation of a micro-level
model to a macro-level model with desirable information-theoretic properties. We conjecture that our framework may lead to extensions of their work, e.g. to the continuous setting.

Finally, suppose that we have made observations of an underlying system $M_X$ via a measurement device $\tau$, and that we want to fit an SCM $M_Y$ from a restricted model class to our data. By using our framework, asking whether or not $M_Y$ admits a causal interpretation consistent with $M_X$ reduces to asking whether the transformation is exact. More generally, by fixing any two of $M_X$, $\tau$ and $M_Y$, we can ask what properties must be fulfilled by the third in order for the two models to be causally consistent. We hope that this may lead to the practical use of SCMs being theoretically grounded.
Causal terminology is often introduced in the interpretation of encoding and decoding models trained on neuroimaging data. In this chapter, we investigate which causal statements are warranted and which ones are not supported by empirical evidence. We argue that the distinction between encoding and decoding models is not sufficient for this purpose: relevant features in encoding and decoding models carry a different meaning in stimulus- and in response-based experimental paradigms. We show that only encoding models in the stimulus-based setting support unambiguous causal interpretations. By combining encoding and decoding models trained on the same data, however, we obtain insights into causal relations beyond those that are implied by each individual model type. We provide an exemplary application of our interpretation rules, and discuss how measurement transformations in neuroimaging affect our ability to causally reason about relationships between brain areas.

4.1 INTRODUCTION

The question how neural activity gives rise to cognition is arguably one of the most interesting problems in neuroimaging [Ham01; War03; Atl+10]. Neuroimaging studies per se, however, only provide insights into neural correlates but not into neural causes of cognition [RKK02; War03]. Nevertheless, causal terminology is often introduced in the interpretation of neuroimaging data. For instance, Hamann writes in a review on the neural mechanisms of emotional memory that “Hippocampal activity in this study was correlated with amygdala activity, supporting the view that the amygdala enhances explicit memory by modulating activity in the hippocampus” [Ham01], and Myers et al. note in a study on working memory that “we tested [...] whether pre-stimulus alpha oscillations measured with electroencephalography (EEG) influence the encoding of items into working memory” [Mye+14] (our emphasis of causal terminology). The apparent contradiction between the prevalent use of causal terminology and the cor-

Sections 4.1–4.5, 4.8, and 4.9 are adapted from [Wei+15].
Causality and Common Neuroimaging Analyses

The relational nature of neuroimaging studies gives rise to the following question: Which causal statements are and which ones are not supported by empirical evidence?

We argue that the answer to this question depends on the experimental setting and on the type of model used in the analysis of neuroimaging data. Neuroimaging distinguishes between encoding and decoding models [Nas+11], known in machine learning as generative and discriminative models [NJ01]. Encoding models predict brain states, e.g. BOLD activity measured by fMRI or event-related potentials measured by EEG/MEG, from experimental conditions [Fri+94; FHP03; Dav+06]. Decoding models use machine learning algorithms to quantify the probability of an experimental condition given a brain state feature vector [Mit+04; PMB09]. Several recent publications have addressed the interpretation of encoding and decoding models in neuroimaging, discussing topics such as potential confounds [TNC13; WGB14], the dimensionality of the neural code [Dav+14], and the relation of linear encoding and decoding models [Hau+14]. We contribute to this discussion by investigating, for each type of model, which causal statements are warranted and which ones are not supported by empirical evidence. Our investigation is based on the seminal work by Pearl [Pea09] and Spirtes et al. [SGS01] on causal inference (cf. [Ram+10; GSH11; WCV11; MR14] for applications of this framework in neuroimaging). We find that the distinction between encoding and decoding models is not sufficient for this investigation. It is further necessary to consider whether models work in causal or anti-causal direction, i.e. whether they model the effect of a cause or the cause of an effect [Sch+12]. To accommodate this distinction, we distinguish between stimulus- and response-based paradigms. We then provide causal interpretation rules for each combination of experimental setting (stimulus- or response-based) and model type (encoding or decoding). We find that when considering one model at a time, only encoding models in stimulus-based experimental paradigms support unambiguous causal statements. Also, we demonstrate that by comparing encoding and decoding models trained on the same data, experimentally testable conditions can be identified that provide further insights into causal structure. These results enable us to reinterpret previous work on the relation of encoding and decoding models in a causal framework [TNC13; Hau+14; WGB14].

We note that the provided causal interpretation rules apply to any encoding and decoding model trained on experimental data. This provides a guideline to researchers on how (not) to interpret encoding and decod-
ing models when investigating the neural basis of cognition. A prequel to [Wei+15] was presented in [Wei+14a], where we argued for the first time that an analysis of the interpretation of encoding and decoding models requires us to distinguish between models that work in causal or anti-causal direction.

We begin in Sections 4.1.1 and 4.1.2 by introducing notation, assumptions and the basic concepts behind constraint-based causal discovery that we rely on throughout this chapter. Based on this, in Section 4.1.3 we revisit the causal statements cited above and demonstrate how they imply testable predictions.

We move forward and introduce the distinction between causal and anti-causal encoding and decoding models and establish a connection between these models and causal inference in Section 4.2. This connection enables us to derive the causal interpretation rules for separate encoding or decoding models in Section 4.3. In Section 4.4, we show that combining an encoding and a decoding model trained on the same data can provide further insights into causal structure. We provide an overview over interpretation rules in Section 4.5 and a reinterpretation of previous work on encoding and decoding models in a causal framework in Section 4.8. In Sections 4.6 and 4.7, we exemplify the application of our interpretation rules and discuss how variable transformations, that naturally arise as measurement transformations in neuroimaging, affect our ability to draw causal conclusions and to sensibly talk about cause-effect relationships between brain areas.

4.1.1 Notation

By \( X \) we denote the finite set of \( d \) random variables representing the brain state features, i.e. \( X = \{X_1, \ldots, X_d\} \). Samples of these variables may correspond to any type of independent and identically distributed (iid) samples of \( d \) brain state features. It is helpful to consider bandpower features of different EEG channels, trial-averaged BOLD activity at various cortical locations, or mean spike rates of multiple neurons as possible examples. By \( C \) we denote the random variable representing the (usually discrete) experimental condition. \( C \) stands for a stimulus (\( C \equiv S \)) or response (\( C \equiv R \)) variable and it will be made clear when \( C \) is restricted to either particular case. For convenience, we denote the set of all random variables by \( \hat{X} = \{C, X_1, \ldots, X_d\} \). Throughout this chapter, we denote marginal, conditional and joint distributions by \( P_X \), \( P_{X|C} \) and \( P_{X,C} \), respectively. For our theoretical investigations, we assume that the involved distributions have
probability mass or density functions (PMFs or PDFs) with values denoted by \( p(x) \), \( p(x|c) \) and \( p(x,c) \) respectively, overloading the notation of \( p \) while it is always clear from the argument which function is meant. We use the common notations for independence and conditional independence

\[
X \perp C \quad : \iff \quad P_{X|C} = P_X,
\]
\[
X \perp C|Y \quad : \iff \quad P_{X|C,Y} = P_{X|Y}.
\]

### 4.1.2 Assumptions & Constraint-Based Causal Discovery

In the framework of Causal Bayesian Networks (CBNs) [SGS01; Pea09], a variable \( X_i \) is said to be a cause of another variable \( X_j \) if the distributions \( P_{\text{do}(X_i=x_i)} \) are sensitive to \( x_i \) (cf. [Pea09], p. 24f.). In this notation, the intervention \( \text{do}(X_i=x_i) \) signifies that \( X_i \) is externally set to a constant \( x_i \), possibly resulting in a change of the distribution of \( X_j \). The framework of CBNs thus defines cause-effect relations in terms of the impact of external manipulations. This is in contrast to frameworks that define causality in terms of information transfer [Gra69; RFG05; LP10].

Causal relations between variables in CBNs are represented by directed acyclic graphs (DAGs). If we find a directed edge \( X_i \rightarrow X_j \), we call \( X_i \) a direct cause of \( X_j \) and \( X_j \) a direct effect of \( X_i \). In case there is no directed edge but at least one directed path \( X_i \rightarrow\rightarrow X_j \), we call \( X_i \) an indirect cause of \( X_j \) and \( X_j \) an indirect effect of \( X_i \). Note that the terms (in-)direct cause/effect depend on the set \( \hat{X} \) of observed variables: consider \( \hat{X} = \{C, X_1, X_2\} \) and the causal DAG \( C \rightarrow X_1 \rightarrow X_2 \). Then \( C \rightarrow\rightarrow X_2 \) and \( C \notall X_2 \) wrt. \( \hat{X} \), while \( C \rightarrow X_2 \) wrt. \( \hat{X}' = \{C, X_2\} \). That is, whether a cause or effect is direct or indirect depends on the set of observed brain state features. We omit the considered set of nodes if it is clear from the context.

To establish a link between conditional independences and DAGs the following concepts are required:

**d-separation** Disjoint sets of nodes \( A \) and \( B \) are d-separated by another disjoint set of nodes \( C \) if and only if all \( a \in A \) and \( b \in B \) are d-separated by \( C \). Two nodes \( a \neq b \) are d-separated by \( C \) if and only if every path between \( a \) and \( b \) is blocked by \( C \). A path between nodes \( a \) and \( b \) is blocked by \( C \) if and only if there is an intermediate node \( z \) on the path such that (i) \( z \in C \) and \( z \) is a tail-to-tail (\( \leftarrow z \rightarrow \)) or head-to-tail (\( \rightarrow z \rightarrow \)) or (ii) \( z \) is head-to-head (\( \rightarrow z \rightarrow \)) and neither \( z \) nor any of its descendants is in \( C \).
4.1 INTRODUCTION

**CAUSAL MARKOV CONDITION (CMC)** The CMC expresses the notion that each node in a causal DAG becomes independent of its non-descendants given its direct causes, i.e. that the causal structure implies certain (conditional) independences.

**FAITHFULNESS** The faithfulness assumption states that all (conditional) independences between the random variables of a DAG are implied by its causal structure, i.e. there are no more (conditional) independences than those implied by the CMC.

Throughout this chapter we assume faithfulness and the causal Markov condition. Under these assumptions d-separation is equivalent to conditional independence, i.e. $C$ d-separates $A$ and $B$ if and only if $A$ and $B$ are independent given $C$ [SGS01]. The following three examples are the most relevant instances of d-separation for our following arguments. Firstly, consider the chain $X_0 \rightarrow X_1 \rightarrow X_2$. Here, $X_1$ d-separates $X_0$ and $X_2$ by blocking the directed path from $X_0$ to $X_2$. This implies that $X_0 \perp \perp X_2|X_1$. Secondly, consider the fork $X_0 \leftarrow X_1 \rightarrow X_2$. Here, $X_1$ d-separates $X_0$ and $X_2$, as $X_1$ is a joint cause of $X_0$ and $X_2$. This again implies that $X_0 \perp \perp X_2|X_1$. Thirdly, consider the collider $X_0 \rightarrow X_1 \leftarrow X_2$. In this case, $X_1$ does not d-separate $X_0$ and $X_2$. As $X_1$ is a joint effect of $X_0$ and $X_2$, it unblocks the previously blocked path between $X_0$ and $X_2$, implying that $X_0 \not\perp \perp X_2|X_1$.

The equivalence between d-separation and conditional independence enables us to infer causal relations between variables in $\hat{X}$ from observational data. By identifying conditional independences that hold in our data, and mapping them onto the equivalent d-separations, we gain knowledge about the causal structures that can give rise to our data. This link forms the basis of the inference rules we describe in Sections 4.3, 4.4, and 4.5. We refer the interested reader to [MR14] for a more exhaustive introduction to this causal inference framework in the context of neuroimaging.

4.1.3 Revisit Causal Terminology in Conclusions of Neuroimaging Studies

We now demonstrate how the causal statements, that we cited in Section 4.1 to motivate our work, can be expressed in the framework of CBNs.

Firstly, consider the statement “Hippocampal activity in this study was correlated with amygdala activity, supporting the view that the amygdala enhances explicit memory by modulating activity in the hippocampus” [Ham01]. Here, it is implicitly assumed that hippocampal activity is a cause of explicit memory; that is, manipulating hippocampal activity
results in measurable changes in explicit memory. In the notation of CBNs, this is expressed as \( \text{hippocampal activity} \rightarrow \text{explicit memory} \). Further, it is implied that the amygdala enhances explicit memory via modulating activity in the hippocampus; that is, manipulating activity in the amygdala leads to measurable changes in hippocampal activity, which results in changes in explicit memory. This gives the causal hypothesis \( \text{amygdala activity} \rightarrow \text{hippocampal activity} \rightarrow \text{explicit memory} \). Assuming faithfulness and the CMC, this causal hypothesis makes the empirically testable predictions \( \text{amygdala activity} \not\perp \not\perp \text{explicit memory} \) and \( \text{amygdala activity} \perp \perp \text{explicit memory} \mid \text{hippocampal activity} \).

Secondly, consider the statement “we tested [...] whether pre-stimulus alpha oscillations measured with electroencephalography (EEG) influence the encoding of items into working memory” [Mye+14]. The authors thereby express the notion that pre-stimulus alpha oscillations are a cause of working memory; that is, manipulating pre-stimulus alpha oscillations has a measurable effect on working memory. This gives the causal hypothesis \( \text{pre-stimulus alpha oscillations} \rightarrow \text{working memory} \). Again assuming faithfulness and the CMC, this causal hypothesis makes the empirically testable prediction \( \text{pre-stimulus alpha oscillations} \not\perp \not\perp \text{working memory} \).

In the following, we investigate how these empirical predictions can be tested by encoding and decoding models, and under which conditions these predictions are sufficient to prove a causal hypothesis. In Section 4.8.3 we revisit those two examples illustrating our theoretical results.

4.2 CAUSAL AND ANTI-CAUSAL ENCODING AND DECODING MODELS

In Section 4.2.1 we clarify the sense in which encoding and decoding models either work in causal- or anti-causal direction depending on whether the study follows a stimulus- or response-based paradigm. We link the notion of feature relevance, commonly employed and interpreted in the analysis of neuroimaging data, to marginal and conditional dependence statements (cf. Section 4.2.2).
4.2 CAUSAL AND ANTI-CAUSAL ENCODING AND DECODING MODELS

4.2.1 Modelling in Stimulus- versus Response-Based Paradigms

An encoding model $P(X|C)^1$, as illustrated in Figure 4.1, represents how various experimental conditions are encoded in the brain state feature vector. We ask “How does the brain state look like given a certain experimental condition?”. Examples for encoding models are the general linear model [Fri+94] or the class-conditional mean.

A decoding model $P(C|X)$, as illustrated in Figure 4.1, represents how experimental conditions can be inferred from the brain state feature vector [Mit+04]. We ask “How likely is an experimental condition given a certain brain state feature vector?”. Decoding models are for example obtained using support vector machines (SVMs) or linear regression with $X$ as the independent and $C$ as the dependent variable.

A priori, encoding and decoding models are oblivious to the causal relation between experimental conditions and brain state features, i.e. they disregard whether they model the effect of a cause or the cause of an effect [Sch+12]. As we show in Section 4.2.2, however, this has implications for their interpretation. We hence introduce the distinction between stimulus- and response-based experimental paradigms.

We categorise an experimental setup as stimulus-based, if the experimental conditions precede the measured brain states and are randomised. An example of a stimulus-based setup is the investigation of the brain’s activity when exposed to auditory stimuli. As the auditory stimuli precede the mea-

---

1 In this chapter, we use notation such as $P(X|C)$ for models to indicate the corresponding distribution, here $P_{X|C}$, and modelling direction, here $X$ is modelled given $C$. 

Figure 4.1: Illustration of an encoding (ENC) and a decoding (DEC) model. The former models the brain state features as a function of the experimental condition, while the latter works in the opposite direction.
sured brain activity, the brain state features cannot be a cause of the stimuli. In the stimulus-based setting, an encoding model \( P(X|C) \equiv P(X|S) \) works in causal direction, as it models the effect of a cause. A decoding model \( P(C|X) \equiv P(S|X) \) works in anti-causal direction, as it models the cause of an effect. We note that in the stimulus-based setting we can control for and randomise the stimulus, i.e. we can externally set the stimulus to a desired value denoted by \( \text{do}(S = s) \).

We categorise an experimental setup as response-based, if the measured brain states precede the experimental conditions. An example of a response-based setup is the prediction of reaction times or prediction of the laterality of a movement from pre-movement brain state features. As the measured brain state features precede the actual reaction time or movement, the reaction time or movement cannot be a cause of this brain activity. In the response-based setting, an encoding model \( P(X|C) \equiv P(X|R) \) works in anti-causal direction, as it models the cause of an effect. A decoding model \( P(C|X) \equiv P(R|X) \) works in causal direction, as it models the effect of a cause. We note that in contrast to the stimulus-based setting, we cannot control for and randomise the response, i.e. we cannot set the response to a desired value by an external intervention.

In the following, we hence distinguish between four types of models:

1. Causal encoding models – \( P(X|S) \)
2. Anti-causal decoding models – \( P(S|X) \)
3. Anti-causal encoding models – \( P(X|R) \)
4. Causal decoding models – \( P(R|X) \)

### 4.2.2 Relating Feature Relevance to Causal Inference

In this section, we establish a link between causal inference and the identification of relevant features in encoding and decoding models. This link forms the basis for the causal interpretation rules in the next section.

When using an encoding model to analyse neuroimaging data, we wish to identify features that show a statistically significant variation across experimental conditions. In practice, this can be carried out by a variety of methods including but not limited to a general linear model [Fri+94], class-conditional differences in mean activation [DM04], and non-linear independence tests [Gre+07; GSH11]. Common to all these approaches is that they admit univariate statistical tests to quantify the likelihood of the data under
the null-hypothesis $X_i \perp \perp C$. Features, for which the null-hypothesis of independence is rejected, are considered relevant for the encoding model in this experimental paradigm. Features with insufficient evidence for rejection of the null-hypothesis are considered irrelevant for the present encoding model. The set of relevant and irrelevant features in an encoding model is subsequently denoted as $X^{+\text{enc}}$ and $X^{-\text{enc}}$, respectively. Assuming faithfulness, the relevance of features in an encoding model can be translated into d-separation properties that provide insights into the causal structure of the data-generating process. In particular, features in $X^{-\text{enc}}$ are d-separated from the experimental condition by the empty set while features in $X^{+\text{enc}}$ are not d-separated from the experimental condition by the empty set.

When using a decoding model to analyse neuroimaging data, we wish to identify features that matter for decoding the experimental condition; that is, we wish to determine for each feature if its removal increases the minimum Bayes error or changes our predicted class probabilities. If $X_1 \perp \perp C|X \setminus X_1$, then the predicted class probabilities remain unchanged upon dropping $X_1$, the Bayes classifier satisfies $\arg\max_c P(c|x_1,\ldots,x_d) = \arg\max_c P(c|x_2,\ldots,x_d)$, and the Bayes error rate is not impeded. This leads to the mathematical concept of conditional independence for a feature’s relevance in decoding [Str+08]: A feature $X_i$ is considered relevant for decoding if and only if $X_i \not\perp \perp C|X \setminus X_i$. In practice, one commonly used approach is to permute a feature with respect to the experimental conditions and check if this results in a statistically significant decrease in decoding accuracy or change in the predictions [Bre01]; even if conditional independence is often not explicitly tested for or the required conditional permutation scheme [Str+08] not strictly adhered to, it is important to note that conditional independence indeed captures the intended meaning of feature relevance in decoding. Features, for which the null-hypothesis of conditional independence is rejected, are considered relevant for the decoding model. Features with insufficient evidence for rejection of the null-hypothesis are considered irrelevant for decoding the experimental condition given the other features. The set of relevant and irrelevant features in a decoding model is subsequently denoted as $X^{+\text{dec}}$ and $X^{-\text{dec}}$, respectively. Assuming faithfulness again, every feature $X_i \in X^{-\text{dec}}$ is d-separated from the experimental condition by the set $X \setminus X_i$, while every feature $X_i \in X^{+\text{dec}}$ is not d-separated from the experimental condition by the set $X \setminus X_i$.

The link between the relevance of features in encoding and decoding models and d-separation properties allows us to provide the causal interpretation rules given in the next section. For our theoretical arguments, we
assume that we can identify all relevant features for each type of model, that is, we assume we can identify all (conditional) (in-)dependence relations. The aim is to demonstrate interpretational limits that persist even beyond statistical testing issues on a finite sample. We discuss the practical intricacies of identifying independence relations on finite empirical data in Section 4.9.

4.3 CAUSAL INTERPRETATION OF EITHER ENCODING or DECODING

4.3.1 Causal Encoding Models $P(X|S)$

According to Reichenbach’s principle [Rei56], the dependence between $S$ and $X_i \in X^{+\text{enc}}$ implies that $S \rightarrow X_i$, $S \leftarrow X_i$, or $S \leftarrow H \rightarrow X_i$ with $H$ a joint common cause of $S$ and $X_i$. In the stimulus-based setting, we control for and randomise the stimulus. This enables us to rule out the last two cases and conclude that $S \rightarrow X_i$, i.e. the features in $X^{+\text{enc}}$ are effects of $S$ [Hol86].

In contrast, for $X_i \in X^{-\text{enc}}$ we have $S \perp \perp X_i$, which allows us to conclude that features in $X^{-\text{enc}}$ are not effects of $S$.

As such, all relevant features in a causal encoding model are effects of $S$, while irrelevant features are not effects of $S$. We hence have the following two interpretation rules:

**Interpretation rule S1.** For $C \equiv S$:

\[ X_i \in X^{+\text{enc}} \iff X_i \text{ is an effect of } S, \text{ i.e. } S \rightarrow X_i \]

**Interpretation rule S2.** For $C \equiv S$:

\[ X_i \in X^{-\text{enc}} \iff X_i \text{ is not an effect of } S, \text{ i.e. } S \not\rightarrow X_i \]

4.3.2 Anti-Causal Decoding Models $P(S|X)$

We describe two counterexamples that show that features in $X^{+\text{dec}}$ are not necessarily effects of $S$ and that features in $X^{-\text{dec}}$ can be effects of $S$.

Firstly, assume $S \rightarrow X_1 \leftarrow X_2$. Since $X_1$ does not d-separate $S$ and $X_2$, i.e. $S \not\perp X_2|X_1$, we obtain that $X_2 \in X^{+\text{dec}}$ although $X_2$ is not an effect of $S$.

Secondly, assume $S \rightarrow X_1 \rightarrow X_2$. Since $X_1$ d-separates $S$ and $X_2$, i.e. $S \perp X_2|X_1$, we have $X_2 \in X^{-\text{dec}}$ although $X_2$ is an effect of $S$. 
This establishes that interpreting relevant features in anti-causal decoding models has two drawbacks. Firstly, features in $X_{+\text{dec}}$ are only potentially effects of $S$. And secondly, effects of $S$ might be missed, since effects are not necessarily relevant for decoding the experimental condition. This yields the following two interpretation rules:

**Interpretation rule S3.** For $C \equiv S$:

$$X_i \in X_{+\text{dec}} \iff X_i \text{ is an effect of } S, \ i.e. \ S \rightarrow X_i$$

**Interpretation rule S4.** For $C \equiv S$:

$$X_i \in X_{-\text{dec}} \iff X_i \text{ is not an effect of } S, \ i.e. \ S \not\rightarrow X_i$$

### 4.3.3 Anti-Causal Encoding Models $P(X|R)$

According to Reichenbach’s principle, the dependence between $X_i \in X^{+\text{enc}}$ and $R$ implies that $X_i \rightarrow R$, $X_i \leftarrow R$, or $X_i \leftarrow H \rightarrow R$ with $H$ a joint common cause of $X_i$ and $R$. A priori, we know that brain activity $\rightarrow$ response. This enables us to rule out the case $X_i \leftarrow R$. We now show that one cannot uniquely determine which of the last two scenarios is the case, i.e. features in $X^{+\text{enc}}$ are only potentially causes of $R$. Consider $X_2 \leftarrow X_1 \rightarrow R$. In this case, we have $X_1 \perp \perp R$ and $X_2 \perp \perp R$ and therefore $X_1, X_2 \in X^{+\text{enc}}$. But note that $X_1 \rightarrow R$ while $X_2 \not\rightarrow R$, i.e. $X_2$ is not a cause of $R$. This shows that features in $X^{+\text{enc}}$ are not necessarily causes of $R$.

Features in $X^{-\text{enc}}$, on the other hand, are independent of $R$ and hence cannot be causes of $R$.

As such, not all relevant features in anti-causal encoding models are causes of $R$, while irrelevant features are indeed not causal for $R$. We hence have the following two interpretation rules:

**Interpretation rule R1.** For $C \equiv R$:

$$X_i \in X^{+\text{enc}} \iff X_i \text{ is only potentially a cause of } R, \ i.e. \ X_i \rightarrow R \text{ or } X_i \leftrightarrow H \rightarrow R$$

**Interpretation rule R2.** For $C \equiv R$:

$$X_i \in X^{-\text{enc}} \implies X_i \text{ is not a cause of } R, \ i.e. \ X_i \not\rightarrow R$$
4.3.4 Causal Decoding Models $P(R|X)$

We describe two counterexamples that show that one can neither conclude that features in $X_{+\text{dec}}$ are causes of $R$ nor that features in $X_{-\text{dec}}$ are not causes of $R$.

Firstly, consider $X_2 \rightarrow X_1 \leftarrow H \rightarrow R$ where $H$ is a hidden common cause of $X_1$ and $R$, yet $H$ is a non-observable brain state feature. We have $X_1, X_2 \in X_{+\text{dec}}$ in this example, as $X_1$ as well as $X_2$ are d-separated from $R$ only by the non-observable common cause $H$. But both $X_1$ and $X_2$ are not causes of $R$.

Secondly, assume $X_2 \rightarrow X_1 \rightarrow R$. Since $X_1$ d-separates $X_2$ and $R$, i.e. $X_2 \perp R|X_1$, we have $X_2 \in X_{-\text{dec}}$ although $X_2$ is a cause of $R$.

This establishes that interpreting relevant features in causal decoding models has two drawbacks. Firstly, features in $X_{+\text{dec}}$ are not necessarily causes of $R$. Secondly, causes of $R$ might be missed, since causes are not necessarily relevant for decoding the experimental condition. This yields the following two interpretation rules:

**Interpretation rule R3.** For $C \equiv R$:

$$X_i \in X_{+\text{dec}} \iff X_i \text{ is a cause of } R, \text{ i.e. } X_i \rightarrow R$$

**Interpretation rule R4.** For $C \equiv R$:

$$X_i \in X_{-\text{dec}} \iff X_i \text{ is not a cause of } R, \text{ i.e. } X_i \not\rightarrow R$$

4.3.5 Subsumption

In the previous subsections, we showed that the interpretation of relevant features in encoding and decoding models depends on the experimental setting. This justifies our argument that the distinction of encoding and decoding models is not sufficient to determine their causal interpretation. In particular we argued that, without employing further assumptions,

1. (cf. 4.3.1) causal encoding models $P(X|S)$ identify all effects of $S$ in $\hat{X}$.

2. (cf. 4.3.2) anti-causal decoding models $P(S|X)$ only identify some features being potentially effects of $S$ in $\hat{X}$.

3. (cf. 4.3.3) anti-causal encoding models $P(X|R)$ identify all features being potentially causes of $R$ in $\hat{X}$.

4. (cf. 4.3.4) causal decoding models $P(R|X)$ only identify some features being potentially causes of $R$ in $\hat{X}$. 
4.4 CAUSAL INTERPRETATION OF BOTH ENCODING AND DECODING

In Section 4.3, we showed that only encoding models in a stimulus-based setting support unambiguous causal statements. This result appears to imply that decoding models, despite their gaining popularity in neuroimaging, are of little value for investigating the neural causes of cognition. In the following, we argue that this is not the case. Specifically, we show that by combining encoding and decoding models, we gain insights into causal structure that are not possible by investigating each type of model individually. In analogy to Section 4.3, we again distinguish between stimulus- and response-based paradigms.

For a combined analysis of encoding and decoding models, we intuitively extend our notation and define the following four sets of brain state features, which yield a disjoint partition of $X$:

$$X^{+enc} = \{X_i \mid X_i \in X^{+enc} \cap X^{+dec}\}$$
$$X^{-enc} = \{X_i \mid X_i \in X^{-enc} \cap X^{-dec}\}$$
$$X^{+enc} = \{X_i \mid X_i \in X^{+enc} \cap X^{-dec}\}$$
$$X^{-enc} = \{X_i \mid X_i \in X^{-enc} \cap X^{-dec}\}$$

We now provide causal interpretation rules for each type of feature set and experimental setting.

4.4.1 Stimulus-Based Setting

4.4.1.1 Features relevant in Encoding and relevant in Decoding: $X^{+enc}_{+dec}$

As $X_i \in X^{+enc}$, it is an effect of $S$ (rule S1). Intuitively, the fact that furthermore $X_i \in X^{+dec}$ suggests that $X_i$ is in some sense closer to $S$. Figure 4.2, however, establishes that this intuition is not correct. Since $S \not\perp \perp X_2$ and $S \not\perp \perp X_2 \mid \{X_1, X_3\}$, we have $X_2 \in X^{+enc}_{+dec}$ even though $X_2$ is an indirect effect of $S$. Hence, features that are relevant in both encoding and decoding do not provide further insights into causal structure. This leads to the following interpretation rule (note the missing bi-implication compared to interpretation rule S1):

**Interpretation rule S5.** For $C \equiv S$:

$$X_i \in X^{+enc}_{+dec} \implies X_i \text{ is an effect of } S, \text{ i.e. } S \rightarrow X_i$$
4.4.1.2 Features relevant in Encoding and irrelevant in Decoding: $X^{+}\text{enc}$

As $X_i \in X^{+}\text{enc}$, it holds that $X_i$ is an effect of $S$ (rule S1). Only looking at the encoding side, however, does not determine whether $X_i$ is a direct ($S \rightarrow X_i$) or indirect effect ($S \dashrightarrow X_i$) of $S$ wrt. $\hat{X}$. As $X_i \in X^-\text{dec}$, however, $S$ and $X_i$ are d-separated by the set $X \setminus X_i$ (cf. Section 4.2.2). This rules out that $X_i$ is a direct effect of $S$ wrt. $\hat{X}$, leaving $S \dashrightarrow X_i$ as the only explanation. This leads to the following interpretation rule:

**Interpretation rule S6.** For $C \equiv S$:

$$X_i \in X^{+}\text{enc} \implies X_i \text{ is an indirect effect of } S \text{ wrt. } \hat{X}$$

4.4.1.3 Features irrelevant in Encoding and relevant in Decoding: $X^{-}\text{enc}$

As $X_i \in X^{-}\text{enc}$, it is not an effect of $S$ (rule S2). As further $X_i \in X^{+}\text{dec}$, $X_i$ and $S$ are not d-separated by $X \setminus X_i$. This implies that $X_i$ is either a cause of variables in $X \setminus X_i$ or that $X_i$ and $X \setminus X_i$ share at least one common cause. Examples of these scenarios are $S \rightarrow X_1 \leftarrow X_2$ and $S \rightarrow X_1 \leftarrow H \rightarrow X_2$. In both cases, knowledge about $X_2$ can be used to better decode $S$ from $X_1$ by removing variations in $X_1$ that are not due to $S$. This leads to the following interpretation rule:

**Interpretation rule S7.** For $C \equiv S$:

$$X_i \in X^{-}\text{enc} \implies X_i \text{ provides brain state context wrt. } S$$
4.4.1.4 Features irrelevant in Encoding and irrelevant in Decoding: $X^{-\text{enc}}_{-\text{dec}}$

Features in $X^{-\text{enc}}_{-\text{dec}}$ are neither effects of $S$ (rule S2) nor do they provide brain state context wrt. $S$. Hence, they can be considered irrelevant for the present experimental context. This is summarised in the following interpretation rule:

**Interpretation rule S8.** For $C \equiv S$:

\[ X_i \in X^{-\text{enc}}_{-\text{dec}} \implies X_i \text{ is neither an effect of } S \text{ nor provides brain state context} \]

4.4.2 Response-Based Setting

4.4.2.1 Features relevant in Encoding and relevant in Decoding: $X^{+\text{enc}}_{+\text{dec}}$

As $X_i \in X^{+\text{enc}}$, it is potentially a cause of $R$ (rule R1). Intuitively, the fact that furthermore $X_i \in X_{+\text{dec}}$ suggests that $X_i$ is in some sense closer to $R$, e.g. that $X_i$ is a cause of $R$. The DAG $X_1 \leftarrow H \rightarrow R$, however, establishes that this intuition is not correct. Since $R \not\perp\!\!\!\perp X_1$, we have $X_1 \in X^{+\text{enc}}_{+\text{dec}}$, even though $X_1$ is not a cause of $R$. Hence, features that are relevant in both encoding and decoding do not provide further insights into causal structure. This leads to the following interpretation rule (note the missing bi-implication compared to interpretation rule R1):

**Interpretation rule R5.** For $C \equiv R$:

\[ X_i \in X^{+\text{enc}}_{+\text{dec}} \implies X_i \text{ is only potentially a cause of } R, \]

i.e. $X_i \rightarrow R$ or $X_i \leftarrow H \rightarrow R$

4.4.2.2 Features relevant in Encoding and irrelevant in Decoding: $X^{+\text{enc}}_{-\text{dec}}$

As $X_i \in X^{+\text{enc}}$, it holds that $X_i \rightarrow R$ or $X_i \leftarrow H \rightarrow R$ (rule R1). Although we cannot distinguish between those two scenarios, we can learn more about $X_i$ from its irrelevance in decoding. $X_i \rightarrow R$ would imply that $X_i$ and $R$ are not d-separated by $X \setminus X_i$ and hence $X_i \in X^{+\text{enc}}_{+\text{dec}}$. Thus, $X_i$ cannot be a direct cause of $R$ wrt. $\hat{X}$. This leads to the following interpretation rule:

**Interpretation rule R6.** For $C \equiv R$:

\[ X_i \in X^{+\text{enc}}_{-\text{dec}} \implies X_i \text{ is no direct cause of } R \text{ wrt. } \hat{X} \]
4.4.2.3 *Features irrelevant in Encoding and relevant in Decoding: \( X^{-\text{enc}}_{+\text{dec}} \)

As \( X_i \notin X^{-\text{enc}} \), it is not a cause of \( R \) (rule R2). As further \( X_i \in X_{+\text{dec}} \), \( X_i \) and \( R \) are not d-separated by \( X \setminus X_i \). This implies that \( X_i \) is either a cause of variables in \( X \setminus X_i \) or that \( X_i \) and \( X \setminus X_i \) share at least one common cause. An example of this scenario is \( X_2 \rightarrow X_1 \leftarrow H \rightarrow R \). Here, knowledge about \( X_2 \) can be used to better decode \( R \) from \( X_1 \) by removing variations in \( X_1 \) that are not due to \( H \). In analogy to the stimulus-based setting, this leads to the following interpretation rule:

**Interpretation rule R7.** For \( C \equiv R \):

\[
X_i \in X^{-\text{enc}}_{+\text{dec}} \implies X_i \text{ provides brain state context wrt. } R
\]

4.4.2.4 *Features irrelevant in Encoding and irrelevant in Decoding: \( X^{-\text{enc}}_{-\text{dec}} \)

Features in \( X^{-\text{enc}}_{-\text{dec}} \) are neither causes of \( R \) (rule R2) nor do they provide brain state context wrt. \( R \). Hence, they can be considered irrelevant for the present experimental context. This leads to the following interpretation rule:

**Interpretation rule R8.** For \( C \equiv S \):

\[
X_i \in X^{-\text{enc}}_{-\text{dec}} \implies X_i \text{ is neither a cause of } R \text{ nor provides brain state context}
\]

4.5 **Rules for Causal Interpretation of Feature Relevance**

We summarise all causal interpretation rules in Table 4.1. Combining encoding and decoding models is particularly useful if features turn out to be relevant in only one type of model: features only relevant in encoding are not direct effects/causes, while features only relevant for decoding do provide brain state context wrt. \( S/R \) while not being effects/causes.

We further note that due to the possibility of hidden confounders, features potentially being a cause and genuine causes cannot be distinguished in the response-based setting without introducing further assumptions.
<table>
<thead>
<tr>
<th>relevance in encoding</th>
<th>relevance in decoding</th>
<th>causal interpretation</th>
<th>rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>X effect of S</td>
<td>S1</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>X no effect of S</td>
<td>S2</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>✓ inconclusive</td>
<td>S3</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>✓ inconclusive</td>
<td>S4</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>X effect of S</td>
<td>S5</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>X indirect effect of S</td>
<td>S6</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>provides brain state context</td>
<td>S7</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>neither effect nor provides brain state context</td>
<td>S8</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓ inconclusive</td>
<td>R1</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>X no cause of R</td>
<td>R2</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>✓ inconclusive</td>
<td>R3</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>✓ inconclusive</td>
<td>R4</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓ inconclusive</td>
<td>R5</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>X no direct cause of R</td>
<td>R6</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>provides brain state context</td>
<td>R7</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>neither cause nor provides brain state context</td>
<td>R8</td>
</tr>
</tbody>
</table>

Table 4.1: Causal interpretation rules for relevant (✓) and/or irrelevant (×) features $X_i$ in encoding and decoding models for stimulus- ($C \equiv S$) and response-based ($C \equiv R$) paradigms.
We present an exemplary application of our causal interpretation chart in Table 4.1. For this we simulate samples from a ground-truth model as depicted in Figure 4.3. In this Section we assume an idealised setting in which the underlying variables can be directly imaged. We consider in Section 4.7 how interpretability is hampered once observations are taken under an inevitable measurement transformation.

In our example scenario, $S$ can be understood as a randomised binary experimental condition such as presenting stimuli to either the left or right ear. The variables $X = \{X_1, \ldots, X_6\}$ represent neuronal activity that we assume can be imaged directly. Similar to voxels in functional magnetic resonance imaging of the human brain, the variables $X_1, \ldots, X_6$ can be identified with spatial locations on a toy 6-pixel “brain map”. This allows for abstract and simplified visualisations of feature importance which resemble brain maps that are commonly encountered in reports of neuroimaging studies, e.g. highlighting specific cortices that were deemed relevant for processing a certain stimulus. In our example, we consider a binary response variable $R$ which could for example reflect whether a subject demonstrated a reaction time below a certain threshold.
4.6 EXEMPLARY APPLICATION

Figure 4.4: A toy 6-pixel brain map analogous to brain maps commonly found in reports of neuroimaging study results. Visualisation of directly observed features $X_1, \ldots, X_6$ relevant/irrelevant for encoding or decoding of a stimulus ($S$) or response ($R$) variable as per Section 4.6.

We assume that we are imaging a subject’s brain activity in 100 trials. During each trial we present a stimulus to either the left or right ear with randomised laterality and observe whether the subject shows a reaction time below a certain threshold. As is common in neuroimaging studies, we assume trials are interleaved by sufficiently long breaks and samples are hence considered iid (cf. Section 4.9.1 for a discussion). That is, our data consists of 100 samples of $S$, $R$, and the 6 brain features $X$.

For our encoding and decoding analysis we employ linear ordinary least squares regression:

**Decoding Model $P(C|X)$**

regression model: $C = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6 + \epsilon$

**Encoding Model $P(X|C)$**

regression model: $X_i = \alpha_0 + \alpha_i C + \epsilon$ for each $i \in \{1, \cdots, 6\}$

where either $C \equiv S$ or $C \equiv R$ for the stimulus- or response-based setting and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Relevance of a feature $X_i$ in encoding or decoding is assessed by the p-value obtained in the respective t-test for testing whether the regression coefficients $\alpha_i$ or $\beta_i$ are non-zero.

Results of the encoding and decoding analyses in the stimulus- and response-based setting are depicted as toy 6-pixel brain maps in Figure 4.4.
where we colourcode the feature relevance, i. e. the respective coefficient’s p-value [0, 1]. Below we detail the different interpretations according to our Table 4.1 (employing a cutoff p-value of 0.05). It is instructive to compare this to the ground-truth as described in Figure 4.3.

**STIMULUS-BASED**

**S5**: **RELEVANT IN ENCODING, RELEVANT IN DECODING**  
$X_3, X_4, X_5$ are effects of $S$, yet, not necessarily direct effects which can be confirmed by considering the ground-truth role of $X_4$

**S6**: **RELEVANT IN ENCODING, IRRELEVANT IN DECODING**  
$X_6$ is an indirect effect of $S$

**S7**: **IRRELEVANT IN ENCODING, RELEVANT IN DECODING**  
$X_1$ provides brain state context; it explains some variance in $X_3$

**S8**: **IRRELEVANT IN ENCODING, IRRELEVANT IN DECODING**  
$X_2$ is neither and effect of $S$ nor does it provide brain state context for decoding $S$; the path between $X_2$ and $S$ is blocked by $X_1$; importantly, $X_2$ is irrelevant in both models while it is *not* independent of all the other brain state features

**RESPONSE-BASED**

**R5**: **RELEVANT IN ENCODING, RELEVANT IN DECODING**  
$X_4$ shows up as relevant in both encoding and decoding, yet its causal interpretation is inconclusive; in our example it is not a cause of $R$ in the ground-truth model, but as shown in Section 4.4.2.1 we cannot disambiguate its role from encoding and decoding analyses only

**R6**: **RELEVANT IN ENCODING, IRRELEVANT IN DECODING**  
$X_5, X_6$ are no direct causes of $R$; indeed, in our example they are not direct (neither indirect) causes of $R$ in the ground-truth model

**R7**: **IRRELEVANT IN ENCODING, RELEVANT IN DECODING**  
$X_3$ provides brain state context; it explains some variance in $X_4$

**R8**: **IRRELEVANT IN ENCODING, IRRELEVANT IN DECODING**  
$X_1, X_2$ are neither causes of $R$ nor do they provide brain state context for decoding $R$; the path between $X_1, X_2$ and $R$ is blocked by $X_3$; importantly, $X_1$ and $X_2$ are irrelevant in both models while they are *not* independent of all the other brain state features
We note that we can infer more about the effects of a randomised stimulus than about the causes of a response variable. Also, due to the conservative phrasing of the interpretations in Table 4.1 we obtain correct statements, yet sometimes of limited value; e.g. R6 only rules out that a variable is a direct cause, but does not give any further insight as to whether it is an indirect cause or not. This cautious approach safeguards against wrong conclusions.

In the preceding subsection we assumed that $X_1, \ldots, X_6$ can be directly observed. This assumption is arguably implausible in neuroimaging. Neuroimaging data is a prime example of data acquired under inevitable measurement transformations that, as discussed in Chapter 3, crucially affect our ability to causally reason about the imaged brain. The number of observed brain state features is orders of magnitude below the number of 100 billion neurons in a human brain; we observe a dimensionality reduced signal with drastic loss of information due to the omission of variables or more general non-injective transformations of the underlying activity. Functional magnetic resonance imaging can be considered a spatial smearing of local average blood-oxygen-levels and EEG recordings are considered a linear superposition of cortical electromagnetic activity [NS06]. Last but not least, the different neuroimaging modalities measure different aspects of “brain activity” such as blood-oxygen-levels or electrical activity; it is unclear how these different representations of brain activity relate to one another in terms of capturing an underlying causal signal and to what extent they capture something like a “true underlying causal neuronal level”.

Here we illustrate how spatial smearing hampers our causal interpretations in the example scenario of Section 4.6. In particular, we do not assume that the underlying $X$ variables are observed directly but instead we consider our imaging ability suffers from spatial smearing, i.e. our measurements $Y$ are a transformation of the $X$ variables. As we encounter interpretational issues even in our simplistic scenario, we emphasise the importance of the conceptual question and contribution in Chapter 3. Furthermore, this motivates preprocessing or blind source separation techniques that undo such measurement transformations and enable a causal description. To this end, we present a robustified independent component analysis (ICA) in Chapter 5. ICA is a blind source separation technique. It is a com-
commonly employed preprocessing step in the analysis of EEG data in order to unmix the observed mixture of cortical signals.

So how does spatial smearing affect our causal interpretations? Instead of observing $X_1, \ldots, X_6$ directly, we assume we can only measure $Y_1, \ldots, Y_6$. Reflecting our limited imaging ability, the $Y$ variables are assumed to be local weighted averages of the $X$ variables, that is the $Y$ variables are a (linear) transformation of the $X$ variables. We consider the following measurement transformation

$$
\begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
Y_4 \\
Y_5 \\
Y_6
\end{bmatrix} =
\begin{bmatrix}
1 & 1/2 & 1/2 & 0 & 0 & 0 \\
1/2 & 1 & 0 & 1/2 & 0 & 0 \\
1/3 & 0 & 1 & 1/3 & 1/3 & 0 \\
0 & 1/3 & 1/3 & 1 & 0 & 1/3 \\
0 & 0 & 1/2 & 0 & 1 & 1/2 \\
0 & 0 & 0 & 1/2 & 1/2 & 1
\end{bmatrix}
\begin{bmatrix}
X_1 \\
X_2 \\
X_3 \\
X_4 \\
X_5 \\
X_6
\end{bmatrix}
$$

such that each $Y_i$ is the sum of $X_i$ plus the average of the neighbouring $X_j$, excluding neighbours along the diagonals.

While we assume that the exact smearing transformation is unknown to us, we presume—in analogy to the analysis of neuroimaging data—that each $Y_i$ picks up mostly the signal of the corresponding underlying $X_i$. Doing so enables us to sensibly identify each $Y_i$ with a spatial location (in neuroimaging we say “cortical area”) and thus can again depict our analysis results on toy brain maps. The analogous analysis to the one in Section 4.6 leads to the maps shown in Figure 4.5.

The variables in the following cortical areas have a different interpretation between the transformation-free setting in Section 4.6 and the spatially smeared imaging considered in this section:

**STIMULUS-BASED**

*(s7 $\rightarrow$ s8)* Area 1 becomes irrelevant also in decoding when we consider spatial smearing. While area 1 was previously considered to provide brain state context, the analysis on the basis of the $Y$ variables suggests that area 1 is neither an effect of $S$ nor provides brain state context for decoding $S$. We may thus falsely conclude that area 1 plays no role in the context of processing the stimulus $S$.

*(s5 $\rightarrow$ s6)* Area 3 and 4 become irrelevant in decoding when we consider spatial smearing. While area 3 and 4 were previously con-
FIGURE 4.5: A toy 6-pixel brain map analogous to brain maps commonly found in reports of neuroimaging study results. Visualisation of features $Y_1, \ldots, Y_6$ relevant/irrelevant for encoding or decoding of a stimulus ($S$) or response ($R$) variable as per Section 4.7. The observed variables are a measurement transformation of the underlying $X$ variables that, due to limited measurement ability, cannot be observed directly.

Considered effects of $S$, leaving open the possibility of being direct effects, the analysis on the basis of the $Y$ variables suggests that both area 3 and 4 are only indirect effects of $S$. In the case of area 3 the latter interpretation conflicts with ground-truth as depicted in Figure 4.3.

RESPONSE-BASED

(R8 $\leadsto$ R6) Area 2 becomes relevant in encoding when we consider spatial smearing. While area 2 was previously considered irrelevant to the process under investigation, the analysis on the basis of the $Y$ variables suggests that we can only rule out area 2 being a direct cause of $R$. In general, this conclusion may be wrong or too conservative; while indeed still true, the interpretation of not being a direct cause is less informative than also concluding that it is neither an indirect cause nor a context variable.

(R7 $\leadsto$ R5) Area 3 becomes additionally relevant in encoding when we consider spatial smearing. While area 3 was previously considered to provide brain state context for decoding $R$, the analy-
sis on the basis of the Y variables suggests that area 3 may not only provide brain state context but also be related to R in some other way while its precise role is inconclusive. In general, this conclusion may be wrong or too conservative; since area 3 can be ruled out as cause of R the more informative interpretation as brain state context should reflect this, while R5 leaves its role open entirely.

(R5 $\rightarrow$ R6) Area 4 becomes irrelevant in decoding when we consider spatial smearing. While the role of area 4 was previously considered inconclusive, leaving open the possibility of being also a direct cause of R, the analysis on the basis of the Y variables suggests that area 4 is not a direct cause of R. In general, this conclusion may be wrong while it happens to be compatible with ground-truth in our considered example.

Measurement transformations may mislead to wrong causal interpretations. Even in our simplistic example, a mere linear transformation that spatially smears the signals alters or hinders causal interpretation. This problem is even more pressing in real neuroimaging scenarios where the observed variables are unknown transformations of an unknown underlying neuronal level, possibly further corrupted by noise stemming from the measurement device. It is pragmatic to undergird existing analysis practices—as we do by providing our interpretation rules—and also to restrict oneself to statements about the effects of a stimulus variable or causes of a response variable. By asking only those partial questions and the conservative phrasing of the causal interpretations in Table 4.1, however, we may underestimate how severely our ability to reason about cause-effect relationships between brain areas is affected by imaging measurement transformations.

In Chapter 3 we showed that exact transformations preserve causal reasoning. This puts restrictive constraints on the allowed transformations that are untestable in the case of neuroimaging where the transformation is not a modelling choice but a result of our limited ability to image the brain. We argue that it is thence important to understand the problem domain and imaging modality as well as possible. To make sense out of neuroimaging data and sensibly talk about cause-effect relationships in the brain we must investigate and identify the most plausible transformations of how the measured activity may relate to some underlying causal/neuronal level.

We emphasise that the interpretational issues only occur if we aim to arrive at statements such as “motor cortex activity drives movement speed” or “cat and dog images cause differentiating activity in the visual cortex”.
It is the spatial concepts and mental pictures of how the brain is organised, that we strive for but that at the same time complicate the matter. We could refrain from such aspirations and instead say “cat and dog images cause differentiating measurements as we image the brain with this specific device and compute that specific brain state feature from the observables”. Explanations like these are not satisfactory and almost as uninformative as the statement $S \rightarrow \{X_1, \ldots, X_n\} \rightarrow R$ or $S \rightarrow \tau(X)$; in a way it is a trivial statement that “the entirety of neuronal activity causes one to move the left arm” or “the stimuli cause the entirety of the brain activity measurements”. Instead, we wish to further dissect the cause-effect relationships among the $X$ or $Y$ variables. This way we hope to obtain an understanding of and enable reasoning about neuronal organisation, representation, and how to intervene in case of malfunction. If we assume there is some underlying $X$ level on which we can causally reason about the brain in a reasonable way, we need to understand (1) how our observables $Y = \tau(X)$ admit or facilitate a causal description, (2) how the $Y$ level can inform our understanding about the underlying $X$ level, and (3) how we can transform our $Y$ variables such that either $f(Y) = f(\tau(X)) \approx X$ or $Z = f(Y)$ where $Z$ enables a pragmatic causal description of the brain.

For reasoning about the neural causes of responses we may be fundamentally limited by the current inability to intervene on “blood-oxygen-level in voxel 42” or “bandpower of alpha-oscillations in the motor cortex”; without such facilities we may not be able to disambiguate between two causal descriptions, one reasoning about components of $f(\tau(X))$ as being causes of $R$, the other being an equally plausible “causal story” based on some other $g(\tau(X))$.\(^2\) Only with an objective, the causal modelling exercise becomes meaningful and decidable, e.g. when asking “Where and how should we apply transcranial magnetic stimulation, raising the bandpower of alpha-oscillations in a certain brain region, to cause improved memory or faster reaction times?”. Unless we can intervene on variables in either representation or do ask causal questions with reference to specific concepts, it is an arbitrary choice whether we want to reason about cause-effect relationships between blood-oxygen-levels or between oscillatory activity in different brain regions. Neither model can be refuted unless intervention facilities become available and neither can claim to capture some true ontology. Models that predict the effect of interventions are not falsifiable as long as

\(^2\) More formally, both $f \circ \tau$ and $g \circ \tau$ may correspond to exact transformations (cf. Definition 3) of the underlying $X$ level while components of the former representation correspond to blood-oxygen-levels in certain brain regions and components of the latter correspond to alpha-oscillation bandpower in different cortices.
those interventions are merely hypothetical. There can be other objectives that do not depend on the (hypothetical) ability to intervene, such as human cognitive constraints or parsimony of cause-effect relationships. For example, if we wish to reason about the causes of reaction times while keeping the number of nodes and edges in a manageable order, it may be possible to prefer one representation over the other depending on which of the two is the closer approximate transformation (cf. Section 3.5) of the underlying neuronal level. In that sense one causal model can be more pragmatic than another to guide further research on the causes of reaction times, while both provide “causal stories” that equally well describe the non-intervened system.

4.8 REINTERPRETATION OF PREVIOUS WORK IN A CAUSAL FRAMEWORK

In this section, we discuss previous work on the interpretation of encoding and decoding models in light of the causal interpretation rules that we introduced in the previous sections.

4.8.1 Potential Confounds

In [TNC13; WGB14] the problem of potential confounds in multi-voxel pattern analysis (MVPA) has been discussed. In particular, Todd, Nystrom and Cohen [TNC13] demonstrated that decoding models may determine brain state features as relevant that are statistically independent of the experimental condition. This finding is confirmed by interpretation rules S7 and R7 that we presented in Section 4.4. These rules reveal that what Todd et al. termed confounds are exactly those features that are irrelevant in encoding and relevant in decoding. Hence, our reinterpretation in a causal framework re-emphasises the potential problem highlighted in [TNC13] and additionally allows to specify the characteristics of such features, i.e. being irrelevant in encoding and relevant in decoding.

In contrast to Todd et al. we do not term those features confounds, as this terminology suggests that such features invalidate interpretation of other features and cannot be interpreted. Instead, we propose to use the terminology features that provide brain state context. Knowledge about the specific characteristics of features that provide brain state context wrt. S/R can indeed lead to interesting causal statements, as demonstrated in [GSH11]: under the additional assumption of causal sufficiency, the assumption that all
causally relevant features have been observed, a causal influence of gamma oscillations ($\gamma$) on the sensorimotor rhythm was concluded from the fact that $\gamma \in X^{enc} - dec$.

As we have shown in the previous section, encoding and decoding models provide complementary information. We hence argue that the problem discussed in [TNC13; WGB14] is not a shortcoming of decoding models, but rather a useful feature. Decoding models provide insights into causal structure that cannot be gained by only investigating encoding models. Pitfalls are not due to weaknesses of decoding models, but a result of negligent interpretation of relevant features. Our findings are thus in line with Todd et al. and Woolgar et al. and strengthen the point that MVPA results need to be interpreted carefully.

### 4.8.2 Linear Encoding and Decoding Models

Haufe et al. [Hau+14] demonstrated that linear backward models, i.e. linear decoding models, “cannot be interpreted in terms of the studied brain processes”, as a large weight does not necessarily correspond to a feature that picks up the signal and, vice-versa, a feature that picks up the signal does not necessarily have a large weight. The causal interpretation rules S3, S4 and R3, R4 extend this argument to non-linear decoding models and yield a reinterpretation of these findings in the framework of CBNs. What is more, our distinction of stimulus- and response-based experimental settings allows us to specify the finding that linear forward models, i.e. linear encoding models, are not affected by this problem: in accordance with Haufe et al. we showed that in the stimulus-based setting encoding models, both linear and non-linear, allow unambiguous causal statements. However, in the response-based setting only those features irrelevant in encoding are unproblematic to be interpreted as non-causes of $R$, while the meaning of relevant features remains ambiguous in this setting.

For the linear case Haufe et al. presented an intriguing way to obtain an encoding from a decoding model. Linear models only allow to test for correlation, and jointly Gaussian random variables are the only general case for which the lack of correlation implies independence. As such, if one is willing to assume that all variables are jointly Gaussian, this method together with our findings yields an easy way to enrich causal interpretation when using linear models. Firstly, in contrast to the decoding model one started with, the encoding model obtained by inversion allows for causal statements on its own (interpretation rules S1, S2 or R2). Secondly, since
then both an encoding and a decoding model are at hand, also interpretation rules S5-8 or R5-8 can be applied.

4.8.3 Revisiting the Introductory Examples

We revisit the introductory examples for which we derived testable predictions in Section 4.1.3.

Firstly, consider the causal hypothesis \( \text{amygdala activity} \rightarrow \text{hippocampal activity} \rightarrow \text{explicit memory} \). The two predictions derived in Section 4.1.3, i.e. that \( \text{amygdala activity} \) is marginal dependent on \( \text{explicit memory} \) and becomes conditionally independent given \( \text{hippocampal activity} \), can be tested by assessing the relevance of \( \text{amygdala activity} \) in a response-based encoding and decoding model, respectively. Without employing further assumptions these conditions are not sufficient to prove the hypothesised causal structure, though, as \( \text{amygdala activity} \leftarrow \text{hippocampal activity} \rightarrow \text{explicit memory} \) or \( \text{amygdala activity} \leftarrow h \rightarrow \text{hippocampal activity} \rightarrow \text{explicit memory} \), where \( h \) is hidden, are also consistent with the tested conditions. A statement that is warranted by interpretation rule R6 in case that both predictions hold true is the following: \( \text{amygdala activity} \) is not a direct cause of \( \text{explicit memory} \).

Secondly, consider the causal hypothesis \( \text{pre-stimulus alpha oscillations} \rightarrow \text{working memory} \). The prediction derived in Section 4.1.3, i.e. \( \text{pre-stimulus alpha oscillations} \not\perp \perp \text{working memory} \), can be tested by assessing the relevance of \( \text{pre-stimulus alpha oscillations} \) in a response-based encoding model. Testing this prediction is again not sufficient to conclude the hypothesised causal structure: in case the prediction holds true, \( \text{pre-stimulus alpha oscillations} \leftarrow h \rightarrow \text{working memory} \), where \( h \) is hidden, is also consistent with the tested condition. Under the rather strong assumption of causal sufficiency the existence of hidden confounders is denied, and the hypothesised causal structure may indeed be concluded.

Note that a hypothesised causal structure can be rejected whenever one of the conditional independences implied by the DAG is not present in the data.

4.9 DISCUSSION

The empirical relevance of our theoretical results was illustrated with a preliminary example in [Wei+15] through application to EEG data recorded during a visuo-motor learning task. We demonstrated that an encoding
model allows us to determine EEG features that are effects of the instruction to rest or to plan a reaching movement, but does not enable us to distinguish between direct and indirect effects. By comparing relevant features in an encoding and a decoding model, we provided empirical evidence that low-level sensorimotor $\mu$- and/or occipital $\alpha$-rhythms (8–14 Hz) are direct effects, while brain rhythms in higher cortical areas, including precuneus and anterior cingulate cortex, respond to the instruction to plan a reaching movement only as a result of the modulation by other cortical processes. These causal conclusions were in line with the roles commonly attributed to the corresponding cortical areas. Among others, our contribution has also informed the interpretation of semantic content decoding from brain activity [Hut+16], led to a refined understanding of whole-brain neural dynamics of probabilistic reward prediction [Bac+17], and was used to derive an improved and better resolved brain-wide functional atlas [Var+18].

The rules presented in [Wei+15] provide a guideline to researchers which causal statements are and which ones are not supported by empirical data when analysing encoding and decoding models. In particular, we argued that only encoding models in stimulus-based paradigms support unambiguous causal statements. We demonstrated that further causal insights can be derived by combining encoding and decoding models. Our causal interpretation rules apply to any type of brain state feature. In the following, we discuss limitations and potential extensions of this framework.

4.9.1 The iid Assumption in Neuroimaging

It is likely that neither of the iid assumptions is met by neuroimaging data. Consider for example subjects getting tired through the course of an experiment. In this case, the features for later trials may follow another distribution than those for earlier trials, violating the assumption of identical distributions. Also the brain’s state may depend on previous activations and hence later trials may be dependent on previous trials, violating the independence assumption. As statistical tests rest on the iid assumption, it is important to consider this limitation when interpreting test results.

4.9.2 Finite Empirical Data

Our theoretical arguments rest on the assumption that it is possible to distinguish between relevant and irrelevant features in encoding and decoding models, i.e. that we have access to an oracle for univariate independence
and multivariate conditional independence tests. In practice, we are faced with several interrelated problems. Firstly, the identification of irrelevant features rests on the readiness to interpret negative results. We need to interpret the lack of evidence against independence as evidence in favour of it. As such, we need to keep in mind that observing more data and/or using more powerful (conditional) independence tests may falsify previous statistical tests, thereby altering our causal conclusions. Secondly, we either need to employ non-linear encoding and decoding models to test for (conditional) independences, or we need to assume that all observed brain state features are jointly Gaussian. This follows from the fact that linear models only test for uncorrelatedness, and in general uncorrelatedness only implies independence if all variables are jointly Gaussian. Note that other frameworks enable causal discovery in linear models, e.g. by introducing the additional assumption of additive non-Gaussian noise [Shi+06; Hoy+08a]. Lastly, it is difficult to base conditional independence tests on permutation approaches, as these are biased towards conditional dependence [Str+08]. The development of unbiased conditional independence tests is an area of active research [Zha+11].

4.9.3 Univariate versus Multivariate Analysis

We based our causal analysis on commonly employed and intuitive notions of feature relevance (cf. Section 4.2.2). In encoding models, a feature is relevant if it varies with the experimental condition. This corresponds to an univariate independence test $X_i \perp \perp C$. In decoding models, a feature is relevant if it cannot be removed without increasing the minimum Bayes error. This corresponds to a multivariate conditional independence test $X_i \perp \perp C|X \setminus X_i$. The interpretation rules presented in this chapter apply whenever these tests are employed, independently of whether encoding and decoding models or direct statistical tests for (conditional) independence are being used [Gre+07; Zha+11].

We note that there are instances in which decoding models are used to carry out an encoding analysis. Consider the searchlight technique [KGB06]. Here, it is tested whether a set of $k$ voxels as a whole contains information about the experimental condition. In this case, the decoding model is used for a marginal independence test $C \perp \perp \{X_{i_1}, \ldots, X_{i_k}\}$. This approach is oblivious to the causal structure within the set $\{X_{i_1}, \ldots, X_{i_k}\}$. As such, the searchlight technique does not provide causal insights beyond those implied by
an encoding model. The additional insights offered by decoding models rest on multivariate conditional independence tests.

4.9.4 Whole Brain Analysis

In case of a high-dimensional feature space, e.g. voxels in fMRI or band-power features in high-density EEG recordings, training a decoding model on the whole feature set may be infeasible. To harness the additional insights provided by a decoding model, it may be necessary to reduce the feature space dimensionality before training a decoding model, e.g. by clustering for fMRI or ICA for EEG recordings. We note that it is not trivial to reduce the dimensionality of the feature space without discarding causally relevant information, i.e. dimension reduction by an exact transformation in the sense discussed in Chapter 3.

4.9.5 Untestable Assumptions

As it is the case for any type of empirical inference, causal inference also rests on a set of untestable assumptions. In particular, causal inference in the framework of CBNs rests on the CMC and the assumption of faithfulness. Theoretical results show that the set of unfaithful distributions has measure zero relative to all discrete or Gaussian probability distributions that can be generated by a given DAG [Mee95]. As long as nature has no particular reason to favour unfaithful distributions, we are thus unlikely to encounter them in practice. We note that further assumptions might allow stronger causal statements, e.g. the assumption of causal sufficiency rules out the existence of hidden common causes.

4.9.6 Causal Inference in Neuroimaging

We note that the rules presented here have, to a certain extent, already been applied in the context of neuroimaging [Ram+10; GSH11; WCV11; MR14]. The primary contribution of our work is to point out their relation to widely used methods for analysing neuroimaging data. In particular, we show that interpreting encoding and decoding models is a form of causal inference. Given the prevalent use of causal terminology in the interpretation of neuroimaging studies, we believe it is essential to make the inherent assumptions and limitations explicit. We further note that combining encoding and decoding models is only a first step towards a causal analysis of empirical
data. More detailed insights can be obtained by additional conditional independence tests, e.g. by training decoding models on subsets of variables and/or permuting subsets of variables. Causal inference algorithms like the PC or FCI algorithm are designed for tackling such questions and hence can yield more detailed causal insights [SGS01; Pea09]. In contrast, encoding and decoding models are not especially designed for these purposes but, as shown, might still warrant some causal interpretation.

We emphasise that, if relevant features in encoding and decoding models are interpreted in a causal sense, one inevitably accepts the untestable assumptions and limitations expatiated in this chapter. The only way to elude this situation, in case one is not willing to make those assumptions, is to resign from causal interpretations.

If not only correlational statements but ultimately neural causes of cognition are of interest, further assumptions and causal inference algorithms, which go beyond encoding and decoding models, should be considered. Future research may investigate how causal inference methods can be facilitated in neuroimaging, discuss the appropriateness of different assumptions, and explore possible ways to weaken or refine those assumptions in case of neuroimaging data.
We introduce coroICA, confounding-robust independent component analysis, a novel ICA algorithm which decomposes linearly mixed multivariate observations into independent components that are corrupted (and rendered dependent) by hidden group-wise stationary confounding. It extends the ordinary ICA model in a theoretically sound and explicit way to incorporate group-wise (or environment-wise) confounding. We show that our proposed general noise model allows to perform ICA in settings where other noisy ICA procedures fail. Additionally, it can be used for applications with grouped data by adjusting for different stationary noise within each group. Our proposed noise model has a natural relation to causality. We explain how it can be applied in the context of causal inference and how it can be viewed as means to undoing a causality-breaking measurement transformation. In addition to our theoretical framework, we provide an efficient estimation procedure and prove identifiability of the unmixing matrix under mild assumptions. Finally, we illustrate the performance and robustness of our method on simulated data, provide audible and visual examples, and demonstrate the applicability to real-world scenarios by experiments on publicly available Antarctic ice core data as well as two EEG data sets. We provide a scikit-learn compatible pip-installable Python package coroICA as well as R and Matlab implementations accompanied by a documentation at https://sweichwald.de/coroICA/.

5.1 INTRODUCTION

The analysis of multivariate data is often complicated by high dimensionality and complex inter-dependences between the observed variables. In order to identify patterns in such data it is therefore desirable and often necessary to separate different aspects of the data. In multivariate statistics, for example, principal component analysis (PCA) is a common preprocessing step that decomposes the data into orthogonal principle components which are sorted according to how much variance of the original data each com-
ponent explains. There are two important applications of this. Firstly, one can reduce the dimensionality of the data by projecting it onto the lower dimensional space spanned by the leading principal components which maximise the explained variance. Secondly, since the principle components are orthogonal, they separate in some sense different (uncorrelated) aspects of the data. In many situations this enables a better interpretation and representation.

Often, however, PCA may not be sufficient to separate the data in a desirable way due to more complex inter-dependences in the multivariate data (see e.g., Section 1.3.3 in Hyvärinen, Karhunen and Oja [HKO02] for an instructive example). This observation motivates the development of independent component analysis (ICA), formally introduced in its current form by Cardoso [Car89a] and Comon [Com94]. ICA is a widely used unsupervised blind source separation technique that aims at decomposing an observed mixture of independent source signals. More precisely, assuming that the observed data is a linear mixture of underlying independent variables, one seeks the unmixing matrix that maximises the independence between the signals it extracts. There has been a large amount of research on different types of ICA procedures and their interpretations, e.g., Bell and Sejnowski [BS95, Infomax] who maximise the entropy, Hyvärinen [Hyv99b, fastICA] maximising the kurtosis or Belouchrani et al. [Bel+97, SOBI] who propose to minimise time-lagged dependences, to name only some of the widespread examples.

ICA has applications in many fields, for example in finance [e.g., BW97], the study of functional magnetic resonance imaging (fMRI) data [e.g., McK+98a; McK+98b; Cal+03], and notably in the analysis of electroencephalography (EEG) data [e.g., Mak+95; Mak+97; DM04]. The latter is motivated by the common assumption that the signals recorded at EEG electrodes are a (linear) superposition of cortical dipole signals [NS06]. Indeed, ICA-based preprocessing has become the de facto standard for the analysis of EEG data. The extracted components are interpreted as corresponding to cortical sources [e.g., Gha+96; ZWJ00; Mak+02] or used for artefact removal by dropping components that are dominated by ocular or muscular activity [e.g., Jun+00; DSM07].

In many applications, the data at hand is heterogeneous and parts of the samples can be grouped by the different settings (or environments) under which the observations were taken. For example, we can group those samples of a multi-subject EEG recording that belong to the same subject. For the analysis and interpretation of such data across different groups, it is de-
sirable to extract one set of common features or signals instead of obtaining individual ICA decompositions for each group of samples separately. Here, we present a novel, methodologically sound framework that extends the ordinary ICA model, respects the group structure and is robust by explicitly accounting for group-wise stationary confounding. More precisely, we consider a model of the form

$$X_i = A \cdot S_i + H_i,$$  \hspace{1cm} (5.1.1)

where $i$ denotes the sample index, $A$ remains fixed across different groups, $S_i$ is a vector of independent source signals and $H_i$ is a vector of stationary confounding noise variables with fixed covariance within each group (an intuitive example where such a scenario may be encountered in practice is illustrated in Figure 5.9). Based on this extension to ordinary ICA, we construct a method and an easy to implement algorithm to extract one common set of sources that are robust against confounding within each group and can be used for across-group analyses. The unmixing also generalises to previously unseen groups.

5.1.1 \textit{Relation to Existing Work}

ICA is well-studied with a tremendous amount of research related to various types of extensions and relaxations of the ordinary ICA model. In light of this, it is important to understand where our proposed procedure is positioned and why it is an interesting and useful extension. Here, we look at ICA research from three perspectives and illustrate how our proposed coroICA methodology relates to existing work. First off, in Section 5.1.1.1 we compare our proposed methodology with other noisy ICA models. In Section 5.1.1.2, we review ICA procedures based on approximate joint matrix diagonalisation. Finally, in Section 5.1.1.3 we summarise the existing literature on ICA procedures for grouped data and highlight the differences to coroICA.

5.1.1.1 Noisy ICA Models

The ordinary ICA model assumes that the observed process $X$ is a linear mixture of independent source signals $S$ without a confounding term $H$. Identifiability of the source signals $S$ is guaranteed by assumptions on $S$ such as non-Gaussianity or specific time structures. For coroICA we require—similar to other second-order based methods (cf. Section 5.1.1.2)—that the source process $S$ is non-stationary. More precisely, we require that
either the variance or the auto-covariance of \( S \) changes across time. An important extension of the ordinary ICA model is known as noisy ICA [e.g., MCG97] in which the data generating process is assumed to be an ordinary ICA model with additional additive noise. In general, this leads to further identifiability issues. These can be resolved by assuming that the additive noise is Gaussian and the signal sources non-Gaussian [e.g., Hyv99b], which enables correct identification of the mixing matrix. Another possibility is to assume that the noise is independent over time, while the source signals are time-dependent\(^1\) [e.g., CC00b]. In contrast, our assumption on the noise term \( H \) is much weaker, since we only require it to be stationary and hence in particular allow for time-dependent noise in coroICA. As we show in our simulations in Section 5.5.2.3 this renders our method robust with respect to confounding noise: coroICA is more robust against time-dependent noise while remaining competitive in the setting of time-independent noise. We refer to the book by Hyvärinen, Karhunen and Oja [HKO02] for a review of most of the existing ICA models and the assumptions required for identifiability.

5.1.1.2 ICA based on Approximate Joint Diagonalisation

As an extension of PCA, the concept of ICA is naturally connected to the notion of joint diagonalisation of covariance-type matrices. One of the first procedures for ICA was FOBI introduced by Cardoso [Car89b], which aims to jointly diagonalise the covariance matrix and a fourth order cumulant matrix. Extending on this idea Cardoso and Souloumiac [CS93] introduced the method JADE which improves on FOBI by diagonalising several different fourth order cumulant matrices. Unlike FOBI, JADE uses a general joint matrix diagonalisation algorithm which is the de facto standard for all modern approaches. In fact, there is a still-active field that focuses on approximate joint matrix diagonalisation, commonly restricted to positive semi-definite matrices, and often with the purpose of improving ICA procedures [e.g., CS96; Zie+04; TY09; ACG18].

Both JADE and FOBI are based on the assumption that the signals are non-Gaussian. This ensures that the sources are identifiable given independent and identically distributed observations. A different stream of ICA research departs from this assumption and instead assumes that the data are a linear mixture of independent weakly stationary time-series. This model

\(^1\) Autocorrelated signals are time-dependent, while the absence of autocorrelation does not necessarily imply time-independence of the signal. We thus use the terms time-dependence and time-independence throughout this chapter.
is often referred to as a second-order source-separation model (SOS). The
time structure in these models allows to identify the sources by jointly diag-
onalising the covariance and auto-covariance. The first method developed
for this setting is AMUSE by Tong et al. [Ton+90] who diagonalise the co-
variance matrix and the auto-covariance matrix for one fixed lag. The per-
formance of AMUSE is, however, fragile with respect to the exact choice of
the lag, which complicates practical application [Mie+12]. Instead of only
using a single lag, Belouchrani et al. [Bel+97] proposed the method SOBI
which uses all lags up to a certain order and jointly diagonalises all the
resulting auto-covariance matrices. SOBI is to date still one of the most
commonly employed ICA methods, in particular in EEG analysis.

The SOS model is based on the assumption of weak stationarity of
the sources which implies that the signals have fixed variance and auto-
covariance structure across time. This assumption can be dropped and
the resulting models are often termed non-stationary source separation
models (NSS). The non-stationarity can be leveraged to boost the per-
formance of ICA methods in various ways [see MOK95; CC00a; CC00b;
CC01; CCB01; Hyv01; PC01]. All aforementioned methods make use of
the non-stationarity by jointly diagonalising different sets of covariance
or auto-covariance matrices and mainly differ by how they perform the
approximate joint matrix diagonalisation. For example, the methods in-
troduced by Choi and Cichocki [CC00a; CC00b] and Choi, Cichocki and
Belouchrani [CCB01] make use of non-stationarity across sources by sepa-
rating the data into blocks and jointly diagonalising either the covariance
matrices, the auto-covariances or both across all blocks. For our experimen-
tal comparisons, we implemented all three of these methods with the slight
modification that we use the recent uwedge approximate joint matrix diag-
onalisation procedure due to Tichavsky and Yeredor [TY09]. We denote the
resulting three ICA variants as

- choiICA (var): jointly diagonalise blocks of covariances,
- choiICA (TD): jointly diagonalise blocks of auto-covariances,
- choiICA (var & TD): jointly diagonalise blocks of covariances and
  auto-covariances.

Depending on the type of matrix which is diagonalised, each procedure de-
tects different types of signals and behaves differently with respect to noise.
Choi and Cichocki [CC01] suggest a modification of choiICA (TD) in which
instead of auto-covariance matrices, differences of auto-correlation matrices
are diagonalised. The advantage being that it captures the non-stationarity of a signal more explicitly. Our proposed method similarly aims to use this type of signal but instead of considering the noise-free case, we explicitly formalise a model class that generalises to noisy settings. Furthermore, we provide an identifiability theorem allowing for group-wise stationary confounding. Such a result has not been proven for the aforementioned method in the noise-free case. For a detailed description of both SOS- and NSS-based methods we refer the reader to the review by Nordhausen [Nor14] and for recent developments on leveraging non-stationarity for identifiability in non-linear ICA see [HM16].

An exhaustive comparison of all methods is infeasible on the one hand due to the sheer amount of different models and methods and on the other hand due to the fact that appropriately maintained and easy adaptable code—for most methods—simply does not exist. Therefore, we focus our comparison on the following representative, modern methods that are most closely related to coroICA: fastICA, SOBI, choiICA (TD), choiICA (var), choiICA (TD & var). The methods and their respective assumptions on the source and noise characteristics are summarised in Table 5.2.

5.1.1.3 ICA Procedures for Grouped Data

Applications in EEG and fMRI have motivated the development of a wide variety of blind source separation techniques which are capable of dealing with grouped data, e.g., where groups correspond to different subjects or recording sessions. A short review is given in Hyvärinen [Hyv13] and a detailed exposition in the context of fMRI data is due to Calhoun et al. [Cal+03].

Consider we are given $m$ groups $\{g_1, \ldots, g_m\}$ and observe a corresponding data matrix $X_{g_i} \in \mathbb{R}^{d \times n_i}$ for each group, where $d$ is the number of observed signals and $n_i$ the number of observations. Using this notation, all existing ICA procedures for grouped data can be related to one of three underlying models extending the classical mixing model $X = A \cdot S$. The first, often also referred to as “temporal concatenation”, assumes that the mixing remains equal while the sources are allowed to change across groups leading to data of the form

\begin{equation}
(X_{g_1}, \ldots, X_{g_m}) = A \cdot (S_{g_1}, \ldots, S_{g_m}).
\end{equation} (5.1.2)
<table>
<thead>
<tr>
<th>method</th>
<th>type of signal</th>
<th>allowed noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>choiICA (TD)</td>
<td>varying time-dependence signal</td>
<td>time-independent noise</td>
</tr>
<tr>
<td>choiICA (var)</td>
<td>varying variance signal</td>
<td>no noise</td>
</tr>
<tr>
<td>choiICA (var &amp; TD)</td>
<td>varying variance and time-dependence signal</td>
<td>no noise</td>
</tr>
<tr>
<td>SOBI</td>
<td>fixed time-dependence signal</td>
<td>time-independent noise</td>
</tr>
<tr>
<td>fastICA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>non-Gaussian signal</td>
<td>no noise</td>
</tr>
<tr>
<td>coroICA</td>
<td>varying time-dependence and/or variance signal</td>
<td>group-wise stationary noise</td>
</tr>
</tbody>
</table>

Table 5.2: Important ICA procedures and the signal types they require as well as the noise they can deal with. coroICA is a confounding-robust ICA variant and is the only method for which an identifiability result under time-dependent noise is available.

<sup>a</sup> The fastICA method can be extended to include Gaussian noise [see Hyv99a].
The second model, often also referred to as “spatial concatenation”, as-
sumes the sources remain fixed \((n_1 = \cdots = n_m)\) while the mixing matrices
are allowed to change, i.e.,

\[
\begin{pmatrix}
X_{g_1} \\
\vdots \\
X_{g_m}
\end{pmatrix} =
\begin{pmatrix}
A_{g_1} \\
\vdots \\
A_{g_m}
\end{pmatrix} \cdot S. \tag{5.1.3}
\]

Finally, the third model assumes that both the sources and the mixing re-
mains fixed across groups which implies that for all \(k \in \{1, \ldots, m\}\) it holds
that

\[X_{g_k} = A \cdot S. \tag{5.1.4}\]

In all three settings the baseline approach to ICA is to simply apply a clas-
sical ICA to the corresponding concatenated or averaged data, i.e., to apply
the algorithm to the temporally/spatially concatenated data matrices on the
left-hand side of above equations or the average over groups. These ad-hoc
approaches are appealing, since they postulate straightforward procedures
to solving the problem on grouped data and facilitate interpretability of
the resulting estimates. It is these ad-hoc approaches that are implemented
as the default behaviour in toolboxes like the widely used \texttt{eeglab} for EEG
analyses [DM04].

Several procedures have been proposed tailored to specific applications
that extend on these baselines by employing additional assumptions. The
most prominent such extensions are tensorial methods that have found
popularity in fMRI analysis. They express the group index as an addi-
tional dimension (the data is thus viewed as a \(\mathbb{R}^{d \times n \times m}\) tensor) and construct
an estimate factorisation of the tensor representation. Many of these pro-
cedures build on the so called PARAFAC (parallel factor analysis) model
[Har70]. Recasting the tensor notation, this model is of the form (5.1.3)
with \(A_{g_k} = A \cdot D_{g_k}\) for all groups and for diagonal matrices \(D_{g_1}, \ldots, D_{g_m}\).
As can be seen from this representation, the PARAFAC model allows the
mixing matrices to change across groups while they are constrained to be
the same up to different scaling of the mixing matrix columns (intuitively,
across groups the source dimensions are allowed to project with different
strengths onto the observed signal dimensions). Given that the matrices
\(D_{g_1}, \ldots, D_{g_m}\) are sufficiently different it is possible to estimate this model
uniquely without further assumptions. However, in the case that some of
these diagonal matrices are equal identifiability is lost. In such cases Beck-
mann and Smith [BS05] suggest to additionally require that the individual
components of the sources be independent. This is comparable to the case where uncorrelatedness may not be sufficient for the separation of sources while independence is.

The coroICA procedure also allows for grouped-data but aims at inferring a fixed mixing matrix $A$, i.e., a model as given in (5.1.2) is considered. In contrast to vanilla concatenation procedures, our methodology naturally incorporates changes across groups by allowing and adjusting for different stationary confounding noise in each group. We argue why this leads to a more robust procedure and also illustrate this in our simulations and real data experiments. More generally, our goal is to learn an unmixing which allows to generalise to new and previously unseen groups; think for example about learning an unmixing based on several different training subjects and extending it to new so far unseen subjects. Such tasks can appear in brain-computer interfacing applications and can also be of relevance more broadly in feature learning for classification tasks where classification models are to be transferred from one group/domain to another. Since our aim is to learn a fixed mixing matrix $A$ that is confounding-robust and readily applicable to new groups, coroICA cannot naturally be compared to models that are based on spatial concatenation (5.1.3) or fixed sources and mixings (5.1.4); these methods employ fundamentally different assumptions on the model underlying the data generating process, the crucial difference being that we allow the sources and their time courses to change between groups.

5.1.2  Our Contribution

One strength of our methodology is that it explicates a statistical model that is sensible for data with group structure and can be estimated efficiently, while being supported by provable identification results. Furthermore, providing an explicit model with all required assumptions enables a constructive discussion about the appropriateness of such modelling decisions in specific application scenarios. The model itself is based on a notion of invariance against confounding structures from groups, an idea that is also related to invariance principles in causality [Haa44; PBM16]; see also Sections 5.3 and 5.4 for a discussion on the relation to causality and the interplay between causal reasoning and undoing measurement variable transformations.

We believe that coroICA is a valuable contribution to the ICA literature on the following grounds:
• We introduce a methodologically sound framework which extends ordinary ICA to settings with grouped data and confounding noise.
• We prove identifiability of the unmixing matrix under mild assumptions, importantly, we explicitly allow for time-dependent noise thereby lessening the assumptions required by existing noisy ICA methods.
• We provide an easy to implement estimation procedure.
• We illustrate the usefulness, robustness, applicability, and limitations of our newly introduced coroICA algorithm as well as characterise the advantage of coroICA over existing ICAs: The source separation by coroICA is more stable across groups since it explicitly accounts for group-wise stationary confounding.
• We provide an open-source scikit-learn compatible ready-to-use Python implementation available as coroICA from the Python Package Index repository as well as R and Matlab implementations and an intuitive audible example which is available at https://sweichwald.de/coroICA/.

5.2 METHODOLOGY

We consider a general noisy ICA model inspired by ideas employed in causality research (see Sections 5.3 and 5.4). We argue below that it allows to incorporate group structure and enables joint inference on multi-group data in a natural way. For the model description, let \( S_i = (S_1^i, \ldots, S_d^i)^\top \in \mathbb{R}^{d \times 1} \) and \( H_i = (H_1^i, \ldots, H_d^i)^\top \in \mathbb{R}^{d \times 1} \) be two independent vector-valued sequences of random variables where \( i \in \{1, \ldots, n\} \). The components \( S_1^i, \ldots, S_d^i \) are assumed to be mutually independent for each \( i \) while, importantly, we allow for any weakly stationary noise \( H \). Let \( A \in \mathbb{R}^{d \times d} \) be an invertible matrix. The \( d \)-dimensional data process \((X_i)_{i \in \{1, \ldots, n\}}\) is generated by the following noisy linear mixing model

\[
X_i = A \cdot S_i + H_i, \quad \text{for all } i \in \{1, \ldots, n\}. \tag{5.2.1}
\]

\( X \) is a linear combination of source signals \( S \) and confounding variables \( H \). In this model, both \( S \) and \( H \) are unobserved. One aims at recovering the mixing matrix \( A \) as well as true source signals \( S \) from observations of \( X \). Without additional assumptions, the confounding \( H \) makes it impossible to

---

2 Throughout this chapter superscripts denote vector components and subscripts denote sequence indices.
identify the mixing matrix $A$. Even with additional assumptions it remains a difficult task (see Section 5.1.1.1 for an overview of related ICA models). Given the mixing matrix $A$ it is straightforward to recover the confounded source signals $\tilde{S}_i = S_i + A^{-1} \cdot H_i$.

Throughout this chapter, we denote by $X = (X_1, \ldots, X_n) \in \mathbb{R}^{d \times n}$ the observed data matrix and similarly by $S$ and $H$ the corresponding (unobserved) source and confounding data matrices. For a finite data sample generated by this model we hence have

$$X = A \cdot S + H.$$ 

In order to distinguish between the confounding $H$ and the source signals $S$ we assume that the two processes are sufficiently different. This can be achieved by assuming the existence of a group structure such that the covariance of the confounding $H$ remains stationary within a group and only changes across groups.

**Assumption 16 (group-wise stationary confounding).** There exists a collection of $m$ disjoint groups $G = \{g_1, \ldots, g_m\}$ with $g_k \subseteq \{1, \ldots, n\}$ and $\bigcup_{k=1}^{m} g_k = \{1, \ldots, n\}$ such that for all $g \in G$ the process $(H_i)_{i \in g}$ is weakly stationary.

Under this assumption and given that the source signals change enough within groups, the mixing matrix $A$ is identifiable (see Section 5.2.2). Similar to existing ICA methods discussed in Section 5.1.1.2, we propose to estimate the mixing matrix $A$ by jointly diagonalising empirical estimates of dependence matrices. In contrast to existing methods, we explicitly allow and adjust for the confounding $H$. The process of finding a matrix $V$ that simultaneously diagonalises a set of matrices is known as joint matrix diagonalisation and has been studied extensively [e.g., Zie+04; TY09]. In Section 5.2.3, we show how to construct an estimator for $V$ based on approximate joint matrix diagonalisation.

The key step in adjusting for the confounding is to make use of the assumption that in contrast to the signals $S$ the confounding $H$ remains stationary within groups. Depending on the type of signal in the sources one can consider different sets of matrices. Here, we distinguish between two types of signals.

**Variance Signal** In case of a variance signal, the variance process of each signal source $\text{Var}(S^i_l)$ changes over time. These changes can be de-
ected by examining the covariance matrix \( \text{Cov}(X_i) \) over time. For \( V = A^{-1} \) and using (5.2.1) it holds for all \( i \in \{1, \ldots, n\} \) that

\[
V \text{Cov}(X_i)V^\top = \text{Cov}(S_i) + V \text{Cov}(H_i)V^\top.
\]

Since the source signal components \( S^j_i \) are mutually independent, the covariance matrix \( \text{Cov}(S_i) \) is diagonal. Moreover, due to Assumption 16 the covariance matrix of the confounding \( H \) is constant, though not necessarily diagonal, within each group. This implies for all groups \( g \in G \) and for all \( k, l \in g \) that

\[
V (\text{Cov}(X_k) - \text{Cov}(X_l)) V^\top = \text{Cov}(S_k) - \text{Cov}(S_l) \tag{5.2.2}
\]

is a diagonal matrix.

**TIME-DEPENDENCE SIGNAL**

In case of a time-dependence signal, the time-dependence of each signal source \( S^j_i \) changes over time, i.e., for fixed \( \tau \), \( \text{Cov}(S^j_i, S^j_{i-\tau}) \) changes over time. These changes lead to changes in the auto-covariance matrices \( \text{Cov}(X_i, X_{i-\tau}) \). Analogous to the variance signal it holds for all \( i \in \{\tau + 1, \ldots, n\} \) that

\[
V \text{Cov}(X_i, X_{i-\tau}) V^\top = \text{Cov}(S_i, S_{i-\tau}) + V \text{Cov}(H_i, H_{i-\tau}) V^\top.
\]

Since the source signal components \( S^j_i \) are mutually independent, the auto-covariance matrix \( \text{Cov}(S_i, S_{i-\tau}) \) is diagonal and due the stationarity of \( H \) (see Assumption 16) the auto-covariance \( \text{Cov}(H_i, H_{i-\tau}) \) is constant within each group. This implies for all groups \( g \in G \), for all \( k, l \in g \) and for all \( \tau \) that

\[
V (\text{Cov}(X_k, X_{k-\tau}) - \text{Cov}(X_l, X_{l-\tau})) V^\top = \text{Cov}(S_k, S_{k-\tau}) - \text{Cov}(S_l, S_{l-\tau}) \tag{5.2.3}
\]

is a diagonal matrix.

For both signal types, we can identify \( V \) by simultaneously diagonalising differences of (auto-)covariance matrices. Details and identifiability results are given in Section 5.2.3. The two signal types considered differ from both, the more classical settings of non-Gaussian time-independent signals as considered for example by fastICA, and the stationary signals with fixed time-dependence assumed for SOBI (cf. Table 5.2). Owing to the non-stationarity of the signal we can allow for more general forms of noise.
5.2 Methodology

5.2.1 Motivating Examples

To get a better understanding of our proposed ICA model in (5.2.1), we illustrate two different aspects: the group structure and the noise model.

**Noise Model** coroICA can be viewed as a noisy ICA, where noise is allowed to be group-wise non-stationary. This generalises existing noisy ICA methods, which, to the best of our knowledge, all assume that the noise is time-independent with various further assumptions. Example 17 illustrates the intuition behind our model via a toy-application to natural images.

**Example 17** (unmixing noisy images). We provide an illustration of how our proposed method compares to other ICA approaches under the presence of noise. Four images, each $450 \times 300$ pixels and with three RGB colour channels, are used to construct four sources $S_1, S_2, S_3, S_4$ as follows.\(^3\) Every colour channel is converted to a one dimensional vector by cutting each image into $15 \times 10$ equally sized patches (i.e., each patch consists of $30 \times 30$ pixels) and concatenating the row-wise vectorised patches. This procedure preserves the local structure of the image. We concatenate the three colour channels and consider them as separate groups for our model. Thus, each of the four sources $S_1, \ldots, S_4$ consists of $n = 3 \cdot 450 \cdot 300 = 405,000$ observations, that is, three groups of $135,000$ observations corresponding to the RGB colour channels. Next, we construct locally dependent noise that differs across colour channels. Here, locally dependent means that the added noise is similar (and dependent) for pixels which are close to each other. This results in four noise processes $H_1, \ldots, H_4$. We combine the sources with the noise and apply a random mixing matrix $A$ to obtain the following observed data

$$X = A \cdot S + H.$$  

The recast noisy images $\tilde{S} = S + A^{-1}H$ are illustrated in the first row and the recast observed mixtures $X$ in the second row of Figure 5.1. The last three rows are the resulting reconstructions of three different ICA procedures, coroICA (TD), fastICA and choiICA (TD). As expected, fastICA as a noise-free ICA method, appears frail to the noise in the images. While choiICA (TD) is able to adjust for independent noise, it is unable to properly adjust for the spatial dependence of the noise process and thus leads to undesired reconstruction results. In contrast, coroICA (TD) is able to recover the noisy images. It is the noise and its characteristics that break the two competing ICA methods, since all three methods are able to unmix the images in the noise-free case (not shown here).

---

\(^3\) The images are freely available from Pexels GmbH [Pex18].
**Figure 5.1:** Images accompanying Example 17. The top row shows noisy unmixed images, the second row shows mixed images, and the last three rows show unmixed and rescaled images resulting from an application of coroICA (TD), choiICA (TD) and fastICA (cf. Table 5.2). Here, only coroICA (TD) is able to correctly unmix the images and recover the original (noise-corrupted) images.
The noise model we employ is motivated by recent advances in causality research where the group-wise stationary noise can be interpreted as unobserved confounding factors in linear causal feedback models. We describe this in more detail with an explicit example application to Antarctic ice core data in Section 5.3.

**GROUP STRUCTURE** A key aspect of our model is that it aims to leverage group-structure to improve the stability of the unmixing under the presence of group-wise confounding. Here we refer to the following notion of stability: A stable unmixing matrix extracts the same set of independent sources when applied to the different groups; it is robust against the confounding that varies across groups and introduces dependences. A standard ICA method is not able to estimate the correct unmixing $V = A^{-1}$, if the data generating process follows our confounded ICA model in (5.2.1). These methods extract signals that are not only corrupted by the group-wise confounding but also are mixtures of the independent sources and are thus not stable in the aforementioned sense. This is illustrated by the “America’s Got Talent Duet Problem” (cf. Example 18), an extension and alteration of the classical “cocktail party problem”.

**Example 18 (America’s Got Talent Duet Problem).** Consider the problem of evaluating two singers at a duet audition individually. This requires to listen to the two voices separately, while the singers perform simultaneously. There are two sound sources in the audition room (the two singers) and additionally several noise sources which corrupt the recordings at the two microphones (or the jury member’s two ears). A schematic of such a setting is illustrated in Figure 5.2. The additional noise comes from an audience and two open windows. One can assume that this noise satisfies our Assumption 16 on a single group. The sound stemming from the audience can be seen as an average of many sounds, hence remaining approximately stationary over time. Typical sounds from an open window also satisfy this assumption, for example sound from a river or a busy road. Our methodology, however, also allows for more complicated settings in which the noise shifts at known points in times, for example if someone opens or closes a window or starts mowing the lawn outside. In such cases we use the known time blocks of stationary noise as groups and apply coroICA (var) on this grouped data. An example with artificial sound data related to this setting is available at https://sweichwald.de/coroICA/. We show that coroICA (var) is able to recover useful sound signals with the two voices being separated into different dimensions and thus allows to listen to them individually. In contrast, existing ICAs applied to the time concatenated data fail to unmix the two singers.
Figure 5.2: Schematic of the “America’s Got Talent Duet Problem” described in Example 18. The sound from the windows and audience is taken to be confounding noise which has fixed covariance structure over given time blocks. The challenge is to recover the sound signals from the individual singers given the recordings of the two microphones.
5.2.2  Identifiability

Identifiability requires that the source signals $S$ change sufficiently strong within groups. The precise notion of a strong signal depends on the type of signal. As discussed previously, we consider two types of non-stationary signals (i) variance signals and (ii) time-dependence signals. Depending on the signal type we formalise two slightly different assumptions that characterise source signals that ensure identifiability. Firstly, in the case of a variance signal, we have the following assumption.

**Assumption 19** (signals with independently changing variance). For each pair of components $p, q \in \{1, \ldots, d\}$ we require the existence of three (not necessarily unique) groups $g_1, g_2, g_3 \in \mathcal{G}$ and three corresponding pairs $l_1, k_1 \in g_1, l_2, k_2 \in g_2$ and $l_3, k_3 \in g_3$ such that the two vectors

$$\begin{pmatrix}
    \text{Var}(S_{l_1}^p) - \text{Var}(S_{k_1}^p) \\
    \text{Var}(S_{l_2}^p) - \text{Var}(S_{k_2}^p) \\
    \text{Var}(S_{l_3}^p) - \text{Var}(S_{k_3}^p)
\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}
    \text{Var}(S_{l_1}^q) - \text{Var}(S_{k_1}^q) \\
    \text{Var}(S_{l_2}^q) - \text{Var}(S_{k_2}^q) \\
    \text{Var}(S_{l_3}^q) - \text{Var}(S_{k_3}^q)
\end{pmatrix}$$

are neither collinear nor equal to zero.

In case of time-dependence signals we have the analogous assumption.

**Assumption 20** (signals with independently changing time-dependence). For each pair of components $p, q \in \{1, \ldots, d\}$ we require the existence of three (not necessarily unique) groups $g_1, g_2, g_3 \in \mathcal{G}$ and three corresponding pairs $l_1, k_1 \in g_1, l_2, k_2 \in g_2$ and $l_3, k_3 \in g_3$ for which there exists $\tau \in \{1, \ldots, n\}$ such that the two vectors

$$\begin{pmatrix}
    \text{Cov}(S_{l_1}^p, S_{l_1}^{p,-\tau}) - \text{Cov}(S_{k_1}^p, S_{k_1}^{p,-\tau}) \\
    \text{Cov}(S_{l_2}^p, S_{l_2}^{p,-\tau}) - \text{Cov}(S_{k_2}^p, S_{k_2}^{p,-\tau}) \\
    \text{Cov}(S_{l_3}^p, S_{l_3}^{p,-\tau}) - \text{Cov}(S_{k_3}^p, S_{k_3}^{p,-\tau})
\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}
    \text{Cov}(S_{l_1}^q, S_{l_1}^{q,-\tau}) - \text{Cov}(S_{k_1}^q, S_{k_1}^{q,-\tau}) \\
    \text{Cov}(S_{l_2}^q, S_{l_2}^{q,-\tau}) - \text{Cov}(S_{k_2}^q, S_{k_2}^{q,-\tau}) \\
    \text{Cov}(S_{l_3}^q, S_{l_3}^{q,-\tau}) - \text{Cov}(S_{k_3}^q, S_{k_3}^{q,-\tau})
\end{pmatrix}$$

are neither collinear nor equal to zero.

Intuitively, these assumptions ensure that the signals are not changing in exact synchrony across components, which removes degenerate types of signals. In particular, they are satisfied in the case that the variance or auto-covariance processes change pair-wise independently over time. Whenever one of these assumptions is satisfied, the mixing matrix $A$ is uniquely identifiable.
**Theorem 21** (identifiability of the mixing matrix).

Assume the data process \( (X_i)_{i \in \{1, \ldots, n\}} \) satisfies the model in (5.2.1) and Assumption 16 holds. If additionally either Assumption 19 or Assumption 20 is satisfied, then \( A \) is unique up to permutation and rescaling of its columns.

The proof is based on Theorem 1 from Kleinsteuber and Shen [KS13]. For completeness, we introduce some of the notation therein and state their result with adapted notation to ease following our proof of Theorem 21. We begin by defining the empirical correlation between two vectors \( \mathbf{v}, \mathbf{w} \in \mathbb{R}^d \) as

\[
\hat{\text{Corr}}(\mathbf{v}, \mathbf{w}) := \begin{cases} 
\frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|}, & \text{if } \mathbf{v} \neq 0 \text{ and } \mathbf{w} \neq 0, \\
1, & \text{otherwise.}
\end{cases}
\]

Moreover, for a collection of \((d \times d)\)-real diagonal matrices \( \{Z_1, \ldots, Z_m\} \), we define the following collinearity measure

\[
\rho(Z_1, \ldots, Z_m) := \max_{1 \leq k < l \leq d} |\hat{\text{Corr}}(z_k, z_l)|,
\]  

(5.2.4)

where \( z_j := (z_1(j) \ldots, z_m(j)) \) and \( z_i(j) \) is the \( j \)-th diagonal element of the matrix \( Z_i \). Using this notation we can state the uniqueness result due to Kleinsteuber and Shen [KS13, Theorem 1] as follows.

**Theorem 22** (Kleinsteuber and Shen [KS13, Theorem 1]). Let \( D_i \in \mathbb{R}^{d \times d} \), for \( i \in \{1, \ldots, m\} \) be diagonal, and let \( M \in \mathbb{R}^{d \times d} \) be an invertible matrix so that \( M^\top D_i M \) is diagonal as well. Then \( M \) is essentially, up to scaling and permutation of its columns, unique if and only if \( \rho(D_1, \ldots, D_m) < 1. \)

Using this result we prove Theorem 21.

**Proof of Theorem 21.** The theorem is proven by the correct invocation of Theorem 22. We first define the unmixing matrix \( V := A^{-1} \) and introduce the sets of matrices

\[
\mathcal{D}_{\text{var}} := \{V(\text{Cov}(X_k) - \text{Cov}(X_l))V^\top | g \in \mathcal{G} \text{ and } k, l \in g\}.
\]

and

\[
\mathcal{D}_{\text{TD}} := \{V(\text{Cov}(X_k, X_{k-\tau}) - \text{Cov}(X_l, X_{l-\tau}))V^\top | g \in \mathcal{G} \text{ and } k, l \in g\}.
\]
Due to the assumed ICA model and Assumption 16, all matrices in the sets \( \mathcal{D}_{\text{var}} \) and \( \mathcal{D}_{\text{TD}} \) are diagonal (cf. (5.2.2) and (5.2.3)). Moreover, for \( g \in \mathcal{G} \) and \( k, l \in g \) it holds that

\[
V(\text{Cov}(X_k) - \text{Cov}(X_l)) = \text{Cov}(S_k) - \text{Cov}(S_l) = \text{diag} \left( \text{Var}(S^1_k) - \text{Var}(S^1_l), \ldots, \text{Var}(S^d_k) - \text{Var}(S^d_l) \right)
\]

and

\[
V(\text{Cov}(X_k, X_{k-\tau}) - \text{Cov}(X_l, X_{l-\tau})) = \text{Cov}(S_k, S_{k-\tau}) - \text{Cov}(S_l, S_{l-\tau}) = \text{diag} \left( \text{Cov}(S^1_k, S^1_{k-\tau}) - \text{Cov}(S^1_l, S^1_{l-\tau}), \ldots, \text{Cov}(S^d_k, S^d_{k-\tau}) - \text{Cov}(S^d_l, S^d_{l-\tau}) \right).
\]

Next, we define for all \( j \in \{1, \ldots, d\} \) the vectors

\[
z_j = \left( \left( \text{Var}(S^j_k) - \text{Var}(S^j_l) \right)_{k, l \in g} \right)_{g \in \mathcal{G}}
\]

or

\[
z_j = \left( \left( \text{Cov}(S^j_k, S^j_{k-\tau}) - \text{Cov}(S^j_l, S^j_{l-\tau}) \right)_{k, l \in g} \right)_{g \in \mathcal{G}},
\]

depending on whether a variance signal or time-dependence signal is being considered, respectively. Then, Assumption 19 or Assumption 20 implies for all distinct pairs \( p, q \in \{1, \ldots, d\} \) that

\[
|\hat{\text{Corr}}(z_p, z_q)| = \frac{|z_p \cdot z_q|}{\|z_p\| \|z_q\|} < 1.
\]

Hence, for either \( \mathcal{D} = \mathcal{D}_{\text{var}} \) or \( \mathcal{D} = \mathcal{D}_{\text{TD}} \) it holds that \( \rho(\mathcal{D}) < 1 \), where \( \rho \) is defined in (5.2.4). Since the identity matrix satisfies that \( \text{Id} D \text{Id}^\top \) is diagonal for all \( D \in \mathcal{D} \), we can invoke Theorem 22 to conclude that any matrix \( M \in \mathbb{R}^{d \times d} \) for which \( MDM^\top \) is diagonal for all \( D \in \mathcal{D} \), is equal to the identity matrix up to scaling and permutation of its columns. Next, we consider the two signal types separately.

- **variance signal:** If there is a variance signal that satisfies Assumption 19, assume there exists an invertible matrix \( \tilde{A} \) such that for all \( g \in \mathcal{G} \) and all \( k, l \in g \) it holds that

\[
\tilde{A}^{-1}(\text{Cov}(X_k) - \text{Cov}(X_l))(\tilde{A}^{-1})^\top = \text{Cov}(S_k) - \text{Cov}(S_l).
\]
Then, it also holds that
\[
(V\tilde{A})(\text{Cov}(S_k) - \text{Cov}(S_l))(V\tilde{A})^\top \in \mathcal{D}_{\text{var}} = V(\text{Cov}(X_k) - \text{Cov}(X_l))V^\top,
\]
which is diagonal.

• **time-dependence signal:** If there is a time-dependence signal that satisfies Assumption 20, assume there exists an invertible matrix \(\tilde{A}\) such that for all \(g \in \mathcal{G}\) and all \(k, l \in g\) it holds that
\[
\tilde{A}^{-1}(\text{Cov}(X_k, X_{k-\tau})) - \text{Cov}(X_l, X_{l-\tau}))(\tilde{A}^{-1})^\top
= \text{Cov}(S_k, S_{k-\tau}) - \text{Cov}(S_l, S_{l-\tau}).
\]
Then, it also holds that
\[
(V\tilde{A})(\text{Cov}(S_k, S_{k-\tau}) - \text{Cov}(S_l, S_{l-\tau}))(V\tilde{A})^\top \in \mathcal{D}_{\text{TD}} = V(\text{Cov}(X_k, X_{k-\tau})) - \text{Cov}(X_l, X_{l-\tau}))V^\top,
\]
which is diagonal.

Using the above reasoning, either of the two cases—depending on whether Assumption 19 or 20 holds—shows that \(V\tilde{A}\) is equal to the identity matrix up to permutation and rescaling of its columns. Moreover, this implies that \(\tilde{A}\) is equal to \(A\) up to scaling and permutation of its columns. This completes the proof of Theorem 21.

### 5.2.3 Estimation

In order to estimate \(V\) from a finite observed sample \(X \in \mathbb{R}^{d \times n}\), we first partition each group into subgroups. We then compute the empirical (auto-)covariance matrices on each subgroup. Finally, we estimate a matrix that simultaneously diagonalises the differences of these empirical (auto-)covariance matrices using an approximate joint matrix diagonalisation technique. This procedure results in three methods depending on which type of matrices we diagonalise. Similar to our notation for the different versions of choiICAs we denote these methods by coroICA(var) if we diagonalise differences of covariances, coroICA(TD) if we diagonalise differences of auto-covariances, and coroICA(var & TD) if we diagonalise both differences of covariance and auto-covariances.
More precisely, for each group \( g \in G \), we first construct a partition \( \mathcal{P}_g \) consisting of subsets of \( g \) such that each \( e \in \mathcal{P}_g \) satisfies that \( e \subseteq g \) and \( \bigcup_{e \in \mathcal{P}_g} e = g \). This partition \( \mathcal{P}_g \) should be granular enough to capture the changes in the signals described in Assumption 19 or 20. We propose partitioning each group based on a grid such that the separation between grid points is large enough for a reasonable estimation of the covariance matrix and at the same time small enough to capture variations in the signals. In our experiments, we observed robustness with respect to the exact choice; only too small partitions should be avoided since otherwise the procedure is fragile due to poorly estimated covariance matrices. More details on the choice of the partition size are given in Remark 23. Depending on whether a variance or time-dependence signal or a hybrid thereof is considered, we fix time lags \( T \subset \mathbb{N}_0 \).

Next, for each group \( g \in G \), each distinct pair \( e, f \in \mathcal{P}_g \), and each \( \tau \in T \) we define the matrix

\[
M_{e,f}^{g,\tau} := \text{Cov}_\tau(X_e) - \text{Cov}_\tau(X_f),
\]

where \( \text{Cov}_\tau(\cdot) \) denotes the empirical (auto-)covariance matrix for lag \( \tau \) and \( X_e \) is the data matrix restricted to the columns corresponding to the subgroup \( e \). Assumption 16 ensures that \( VM_{e,f}^{g,\tau} V^\top \) is approximately diagonal. We are therefore interested in finding an invertible matrix \( V \) which approximately jointly diagonalises the matrices in the set

\[
\mathcal{M}^{\text{all}} := \{ M_{e,f}^{g,\tau} \mid g \in G \text{ and } e, f \in \mathcal{P}_g \text{ and } \tau \in T \}. \tag{5.2.5}
\]

The number of matrices in this set grows quadratically in the number of partitions. This can lead to large numbers of matrices to be diagonalised. Another option that reduces the computational load is to compare each partition to its complement, which leads to the following set of matrices

\[
\mathcal{M}^{\text{comp}} := \{ M_{e,\bar{e}}^{g,\tau} \mid g \in G \text{ and } e \in \mathcal{P}_g \text{ (with } \bar{e} := g \setminus e \text{) and } \tau \in T \} \tag{5.2.6}
\]

or to compare only neighbouring partitions as in

\[
\mathcal{M}^{\text{neighbor}} := \{ M_{e,\text{neighbor}(e)}^{g,\tau} \mid g \in G \text{ and } e \in \mathcal{P}_g \text{ and } \tau \in T \}, \tag{5.2.7}
\]

where \( \text{neighbor}(e) \) is the partition to the right of \( e \).

The task of jointly diagonalising a set of matrices is a well-studied topic in the literature and is referred to as approximate joint matrix diagonalisation. Many solutions have been proposed for different assumptions made on the
matrices to be diagonalised. We use the uwedge algorithm\(^4\) introduced by Tichavsky and Yeredor \([TY09]\). The basic idea behind uwedge is to find a minimiser of a proxy for the loss function

\[
\ell(V) = \sum_{M \in M^*} \left( \sum_{k \neq l} \left[ VMV^\top \right]_{k,l}^2 \right),
\]

over the set of invertible matrices, where in our case \(M^*\) is either of \(M^{\text{all}}, M^{\text{comp}},\) or \(M^{\text{neighboring}}\).

The full estimation procedure based on the set \(M^{\text{neighboring}}\) defined in (5.2.6) is made explicit in the pseudo code in Algorithm 1 (where ApproximateJointDiagonalizer stands for a general approximate joint diagonaliser; here we use uwedge).

**Remark 23** (choosing the partition and the lags). *Whenever there is no obvious partition of the data, we propose to partition the data into equally sized blocks with a fixed partition size. The decision on how to choose a partition size should be driven by type of non-stationary signal one expects and the dimensionality of the data. For example, in the case of a variance signal the partition should be fine enough to capture areas of high and low variance, while at the same time being coarse enough to allow for sufficiently good estimates of the covariance matrices. That said, for applications to real data sets the signals are often of various length implying that there is a whole range of partition sizes which all work well. In cases with few data points, it can then be useful to consider several grids with different partition sizes and diagonalise across all resulting differences simultaneously. This somewhat removes the dependence of the results on the exact choice of a partition size and increases the power of the procedure. We employ this approach in Section 5.3.1. In general, the lags \(T\) should be chosen as \(T = \{0\}, T \subset \mathbb{N},\) or \(T \subset \mathbb{N}_0,\) depending on whether a variance signal, time-dependence signal, or a hybrid thereof is considered. For time-dependence signal, we recommend to determine up to which time-lag the autocorrelation of the observed signals has sufficiently decayed, and use all lags up to that point.*

5.2.4 **Assessing the Quality of Recovered Sources**

Assessing the quality of the recovered sources in an ICA setting is an inherently difficult task, as is typical for unsupervised learning procedures.

---

\(^4\) As a byproduct of our work, we are able to provide a new stable open-source Python/R/Matlab implementation of the uwedge algorithm which is also included in our respective coroICA packages.
Algorithm 1: coroICA

**input**: data matrix $X$
- group index $G$ (user selected)
- group-wise partition $(P_g)_{g \in G}$ (user selected)
- lags $T \subset \mathbb{N}_0$ (user selected)

 initialise empty list $M$

 for $g \in G$
   for $e \in P_g$
     for $\tau \in T$
       append $\hat{\text{Cov}}_{\tau}(X_e) - \hat{\text{Cov}}_{\tau}(X_{\text{neighbour}(e)})$ to list $M$
     end
   end

 $\hat{V} \leftarrow \text{ApproximateJointDiagonalizer}(M)$
 $\hat{S} \leftarrow \hat{V}X$

**output**: unmixing matrix $\hat{V}$
- sources $\hat{S}$

The unidentifiable scale and ordering of the sources as well as the unclear choice of a performance measure render this task difficult. Provided that ground truth is known, several scores have been proposed, most notably the Amari measure introduced by Amari, Cichocki and Yang [ACY95] and the minimum distance (MD) index due to Ilmonen et al. [Ilm+10]. Here, we use the MD index, which is defined as

$$\text{MD}(\hat{V}, A) = \frac{1}{\sqrt{p-1}} \inf_{C \in \mathcal{C}} \| C\hat{V}A - \text{Id} \|,$$

where the set $\mathcal{C}$ consists of matrices for which each row and column has exactly one nonzero element. Intuitively, this score measures how close $\hat{V}A$ is to a rescaled and permuted version of the identity matrix. One appealing property of this score is that it can be computed efficiently by solving a linear sum assignment problem. In contrast to the Amari measure, the MD index is affine invariant and has desirable theoretical properties [see Ilm+10].

We require a different performance measure for our real data experiments where the true unmixing matrix is unknown. Here, we check whether the desired independence (after adjustment for the constant confounding) is achieved by computing the following covariance instability score (CIS) ma-
trix. It measures the instability of the covariance structure of the unmixed sources $\hat{S}$ and is defined for a each groups $g \in G$ and a corresponding partition $\mathcal{P}_g$ (see Section 5.2.3) by

$$CIS(\hat{S}, \mathcal{P}_g) := \frac{1}{|\mathcal{P}_g|} \sum_{e \in \mathcal{P}_g} \left( \frac{\text{Cov}(\hat{S}_e) - \text{Cov}(\hat{S}_{\text{neighbour}(e)})}{\hat{\sigma}_{\hat{S}_g} \cdot \hat{\sigma}_{\hat{S}_g}^\top} \right)^2,$$

where $\hat{\sigma}_{\hat{S}} \in \mathbb{R}^{d \times 1}$ is the empirical standard deviation of $\hat{S}$ and the fraction is taken element-wise. The CIS matrix is approximately diagonal whenever $\hat{S}$ can be written as the sum of independent source signals $S$ and confounding $H$ with fixed covariance. This is condensed into one scalar that reflects how stable the sources’ covariance structure is by averaging the off-diagonals of the CIS matrix

$$\text{MCIS}(\hat{S}, \mathcal{P}_g)^2 := \frac{1}{d(d-1)} \sum_{i,j=1, i \neq j}^d \left[ \text{CIS}(\hat{S}, \mathcal{P}_g) \right]_{i,j}.$$

The differences taken in the CIS score extract the variance signals such that the mean covariance instability score (MCIS) can be understood as a measure of independence between the recovered variance signal processes. High values of MCIS imply strong dependences beyond stationary confounding between the signals. Low values imply weak dependences. MCIS is a reasonable score whenever there is a variance signal (as described in Section 5.2) in sources and is a sensible evaluation metric of ICA procedures in such cases. In case of time-dependence signal (as described in Section 5.2), one can define an analogous score based on the auto-covariances. Here, we restrict ourselves to the variance signal case as for all our applications this appeared to constitute the dominant part of the signal.

In case of variance signals the MCIS appears natural and appropriate as independence measure: It measures how well the individual variance signals (and hence the relevant information) are separated. To get a better intuition, let $A = (a_1, \ldots, a_d) \in \mathbb{R}^{d \times d}$ denote the mixing and $V = (v_1, \ldots, v_d)^\top \in \mathbb{R}^{d \times d}$ the corresponding unmixing matrix (i.e., $V = A^{-1}$, $a_i$ are columns of $A$ and $v_i$ are rows of $V$). Then it holds that,

$$\text{Cov}(X_i)v_j^\top = A \text{Cov}(S_i)A^\top v_j^\top + \text{Cov}(H_i)v_j^\top$$

$$= A \text{Cov}(S_i)e_j^\top + A \text{Cov}(H_i)e_j^\top$$

$$= a_j \text{Var}(S_i^j) + A \text{Cov}(H_i)e_j^\top \quad (5.2.8)$$
Under our group-wise stationary confounding assumption (Assumption 16) this implies that within all groups \( g \in \mathcal{G} \), it holds for all \( l, k \in g \) that
\[
(Cov(X_l) - Cov(X_k)) v_j^\top = a_j \left( Var(S^j_l) - Var(S^j_k) \right).
\]
This equation holds also in the confounding-free case and it reflects the contribution of the signal (in terms of variance signal) of the \( j \)-th recovered source \( S_j \) to the the variance signal in all components of the observed multivariate data \( X \).

While in the population case the equality in (5.2.9) is satisfied exactly, this is no longer the case when the (un-)mixing matrix is estimated on finite data. Consider two subsets \( e, f \in g \) for some group \( g \in \mathcal{G} \), then using the notation from Section 5.2.3 and denoting by \( \hat{v}_j \) and \( \hat{a}_j \) the estimates of \( v_j \) and \( a_j \), respectively, it holds that
\[
M_{e,f}^g \hat{v}_j^\top = [\widehat{Cov}(X_e) - \widehat{Cov}(X_f)] \hat{v}_j^\top
= \hat{A} [\widehat{Cov}(\hat{S}_e) - \widehat{Cov}(\hat{S}_f)] \hat{A}^\top \hat{v}_j^\top
= \hat{A} \left[ \widehat{Cov}(\hat{S}_e) - \widehat{Cov}(\hat{S}_f) \right] e_j^\top
\approx \hat{a}_j (Var(S^j_e) - Var(S^j_f)). \tag{5.2.10}
\]
The approximation is close only if the empirical estimate \( \hat{V} \) correctly unmixes the \( j \)-th source. Essentially, MCIS measures the extent to which this approximation holds true for all components simultaneously across the subsets specified by the partition \( \mathcal{P}_g \). It is also possible to consider individual components by assessing how closely the following proportionality is satisfied
\[
\sum_{M \in \mathcal{M}^*} \text{sign}(\hat{v}_j M \hat{v}_j^\top) M \hat{v}_j^\top \propto \hat{a}_j. \tag{5.2.11}
\]
In EEG experiments, this can also be assessed visually by comparing the topographic maps corresponding to columns of \( A \) with so-called activation maps corresponding to the left-hand side in (5.2.11). More details on this are provided in Section 5.5.3.3.

\section*{5.3 Causal Structure Learning via ICA}

Our underlying noisy ICA model (5.2.1) and the assumption on the noise (Assumption 16) are motivated by causal structure learning scenarios. ICA is closely linked to the problem of identifying structural causal models
(SCMs) [see Pea09; IR15; PJS17]. Shimizu et al. [Shi+06] were the first to make this connection explicit and used ICA to infer causal structures. To make this more precise consider the following linear SCM

\[ X_i = B \cdot X_i + \tilde{S}_i, \]  

(5.3.1)

where \( X_i \) are observed covariates and \( \tilde{S}_i \) are noise terms. An SCM induces a corresponding causal graph over the involved variables by drawing an edge from variables on the right-hand side to the one on the left-hand side of (5.3.1). Moreover, we can define noise interventions [Pea09] by allowing the distributions of the noise terms \( \tilde{S}_i \) to change for different \( i \). In the language of ICA, this means that the signals \( \tilde{S}_i \) encode the different interventions (over time) on the noise variables. Assuming that the matrix \( \text{Id} - B \) is invertible, we can rewrite (5.3.1) as

\[ X_i = (\text{Id} - B)^{-1}\tilde{S}_i, \]

which can be viewed as an ICA model with mixing matrix \( A = (\text{Id} - B)^{-1} \). Instead of taking the noise term \( \tilde{S}_i \) as independent noise sources one can also consider \( \tilde{S}_i = S_i + H_i \). In that case the linear SCM in (5.3.1) describes a causal model between the observed variables \( X_i \) in which hidden confounding is allowed. This is illustrated in Figure 5.3, which depicts a 3 variable SCM with feedback loops and confounding. Learning a causal model as in (5.3.1) with ICA is generally done by performing the following two steps.

(i) ICA

The matrix \( (\text{Id} - B) \) is inferred by ICA up to an undefined scale and permutation of its rows by using an appropriate ICA procedure. This step is often infeasible in the presence of confounding \( H \) since existing ICA methods only allow noise under restrictive assumptions (cf. Table 5.2).

(ii) IDENTIFY \( B \)

There are essentially two assumptions that one can make in order for this to work. The first is to assume the underlying causal model has an acyclic structure as in Shimizu et al. [Shi+06]. In such cases the matrix \( B \) needs to be permuted to an upper triangular matrix. The second option is to allow for feedback loops in the causal model but restrict the types of feedback to exclude infinite loops as in Hoyer et al. [Hoy+08a] and Rothenhäusler et al. [Rot+15].

When performing step (i) there are two important modelling assumptions that are made when selecting the ICA procedure: (a) the type of al-
allowed signals (types of interventions) and (b) the type of allowed confounding. For the classic ICA setting with non-Gaussian source signals and no noise this translates to the class of linear non-Gaussian models, such as Linear Non-Gaussian Acyclic Models (LiNGAMs) introduced by Shimizu et al. [Shi+06]. While such models are a sensible choice in a purely observational setting (i.e., no samples from interventional settings) they are somewhat misspecified in terms of (a) when data from different interventional settings or time-continuous intervention shifts are observed (see Remark 24). In those settings, it is more natural to use ICA methods that are tailored to sequential shifts as for example choiICA or coroICA. Moreover, most common ICA methods consider noise-free mixing, which from a causal perspective implies that no hidden confounding is allowed. While noisy ICA weakens this assumption, existing methods only allow for time-independent or even iid noise, which again greatly restricts the type of confounding. In contrast, our proposed coroICA allows for any type of block-wise stationary confounding, hence greatly increasing the class of causal models which can be inferred. This is attractive for causal modelling as it is a priori unknown whether hidden confounding exists. Therefore, our proposed procedure allows for robust causal inference under general confounding settings. In Section 5.3.1, we illustrate a potential application to climate science and how the choice of ICA can have a strong impact on the estimates of the causal parameters.

Remark 24 (relation between interventions and non-stationarity). A causal model does not only describe the observational distribution but also the behaviour of the data generating model under all of the allowed interventions. Here, we restrict the allowed interventions to distribution shifts in the source signals, that either change the distribution block-wise (e.g., abruptly changing environmental conditions) or continuously (e.g., continuous shifts in the environmental conditions). Any such shifts are by definition synonymous with the process \( S_i \) being non-stationary. In our proposed causal model (5.3.1) the non-stationarity of the signal therefore corresponds to shifts in the environmental conditions which can be utilised, using coroICA, to infer the underlying causal structure. From this perspective, the causal inference procedure we propose here is a method based on interventional data rather than plainly observational data, while the interventions are not exactly known.
5.3.1 Application to Climate Science

To motivate the foregoing causal model we consider a prominent example from climate science: the causal relationship between carbon dioxide concentration (CO\textsubscript{2}) and temperature (T). More precisely, we consider Antarctic ice core data that consists of temperature and carbon dioxide measurements of the past 800'000 years due to Bereiter et al. [Ber+15, carbon dioxide] and Jouzel et al. [Jou+07, temperature]. We combined both temperature and carbon dioxide data and recorded measurements every 500 years by a cubic interpolation of the raw data. The data is shown in Figure 5.5 (right). Oversimplifying, one can model this data as an SCM with time-lags as follows

\[
\begin{align*}
\begin{pmatrix}
\log(\text{CO}_2)_t \\
T_t
\end{pmatrix} &= \begin{pmatrix} 0 & \beta \\ \alpha & 0 \end{pmatrix} \begin{pmatrix}
\log(\text{CO}_2)_t \\
T_t
\end{pmatrix} + \sum_{k=1}^{p} B_k \begin{pmatrix}
\log(\text{CO}_2)_{t-k} \\
T_{t-k}
\end{pmatrix} + \tilde{S}_t,
\end{align*}
\]

(5.3.2)

where \(\tilde{S}_t = S_t + H_t\) with \(S_t\) component-wise independent non-stationary source signals and \(H_t\) a stationary confounding process. Vector-valued linear time-series models of this type are referred to as structural auto re-
gressive models (SVARs) [see e.g., Lüt05]. They have been previously analysed in the confounding free-case by Hyvärinen et al. [Hyv+10], using an ICA based causal inference approach. A graphical representation of such a model is shown in Figure 5.4. In this example, we can think of the source signals $S_t$ as being two independent summaries of important factors that affect both temperature and carbon dioxide and vary over time, e.g., environmental catastrophes like volcano eruptions and large wildfires, sunspot activity or ice-coverage. These variations can be considered as changing environmental conditions or interventions (see Remark 24). On the other hand the stationary confounding process $H_t$ can be thought of as factors which affect both temperature and carbon dioxide in a constant fashion over time, for example this could be effects due the shifts in the earth’s rotation axis.

Assuming that this was the true underlying causal model, we could use it to predict what happens under interventions. From a climate science perspective an interesting intervention is given by doubling the concentration of CO$_2$ and determining the resulting instantaneous (faster than 1000 years) effect on the temperature. This effect is commonly referred to as equilibrium climate sensitivity (ECS) due to CO$_2$ which is loosely defined as the change in degrees temperature associated with a doubling of the concen-
tation of carbon dioxide in the earth’s atmosphere. In the fifth assessment report of the United Nations Intergovernmental Panel on Climate Change it has been stated that "there is high confidence that ECS is extremely unlikely less than 1 °C and medium confidence that the ECS is likely between 1.5 °C and 4.5 °C and very unlikely greater than 6 °C" [Int14, Chapter 10]. Since the measurement frequency in our model is quite low (500 years) and we model the logarithm of carbon dioxide the ECS corresponds to

$$\text{ECS} = \log(2)\alpha.$$ 

Estimating the model in (5.3.2) can be done by first fitting a vector autoregressive model of the time lags using OLS resulting in a vector of residuals

$$R_t = \left(\log(\text{CO}_2)_t \right) - \left(\hat{\log}(\text{CO}_2)_t \right) - T_t.$$ 

Then, one can apply the two-step causal inference procedure described in Section 5.3 to

$$R_t = B_0 R_t + \hat{S}_t.$$ 

Since we are in a two-dimensional setting, step (ii) (i.e., identifying the causal parameters $\alpha$ and $\beta$ from the estimated mixing matrix) only requires to assume that feedback loops do not blow-up, which translates into $B_0$ having spectral norm less than one. Given that the signal is sufficiently strong (i.e., there are sufficient interventions on both $\text{CO}_2$ and $T$), it is possible to recover the causal parameters by trying both potential permutations of the sources with subsequent scaling and assessing whether the aforementioned condition is satisfied.

We applied this procedure based on coroICA (var) to the data in order to estimate climate sensitivity and compared it with results obtained when using fastICA or choiICA (var). The results are given in Figure 5.5. We believe the results illustrate two important aspects. Firstly, the choice of the lags has a strong effect on the estimation of the causal effect parameters, particularly for boundary cases. If it is chosen too small the remaining time-dependence in the data can obscure the signal. If it is chosen too big part of the signal starts being removed. Choosing an appropriate number of lags is therefore crucial. One option would be to apply an information criterion (AIC or BIC) for this. Secondly, the results illustrate that the choice of ICA has a large impact on the estimated causal effect parameters. More specifically, both the assumed signal as well as the assumed confounding have an impact on the estimation. Compare the results between fastICA (non-Gaussian signal)
FIGURE 5.5: (left) Estimated equilibrium climate sensitivity (ECS) for different ICAs depending on the number of lags included into the SVAR model. The light gray and dark gray overlay indicate likely and very likely value ranges, respectively, for the true value of climate sensitivity as per the fifth assessment report of the United Nations Intergovernmental Panel on Climate Change (cf. Section 5.3.1). The differences across procedures illustrate that the choice of ICA has a large effect on the estimation. (right) Interpolated time-series data, which we model with an SVAR model.
and choiICA/coroICA (variance signal) for the former and observe the differences between fastICA/choiICA (no confounding) and coroICA (adjusted for stationary confounding) for the latter. The choice of the ICA algorithm should therefore be driven by the assumptions (both on signal type and confounding) one is willing to employ on the underlying model. Considering a variance signal and adjusting for confounding, coroICA appears to lead to estimates of equilibrium climate sensitivity that are more closely in line with the highly likely bands previously identified by the United Nations Intergovernmental Panel on Climate Change. This observation is only indicative as all three methods yield highly variable results and also the panel’s highly likely band rests on certain assumptions that may become refuted at some later point. coroICA can be considered a conservative choice if no assumptions on confounding can be made, while noise-free methods may outperform if indeed there were no confounding factors.

5.4 MEASUREMENT TRANSFORMATIONS AND ICA

In this section we make explicit how ICA accounts for a causality-breaking transformation (cf. Chapter 3). First, we explicate how the application of ICA for causal structure learning (cf. Section 5.3) accomplishes two steps at once: ICA (i) separates shift interventions and stationary confounding, both observed only via a measurement transformation, and (ii) leverages this for causal structure identification between the observables. Second, we discuss the use of ICA for demixing recorded EEG signals that are commonly assumed a linear superposition of underlying cortical signals, thereby aiming to separate causal signals that can only be observed via a causality-breaking measurement transformation.

In the previous section we discussed the application of coroICA to identify $B$ in models such as

$$X_b = B \cdot X_b + S_b + H$$

where $X_b$ are observed covariates (e.g. temperature and carbon dioxide concentration), $S_b$ are component-wise independent factors (e.g. block-wise time-varying sunspot activity or volcano eruptions), and $H$ represents (group-wise) stationary confounding noise that our novel coroICA procedure can account for (e.g. regular shifts in the earth’s rotation axis). We rephrase this scenario to draw closer connection to causal modelling under
variable transformations. To this end, assume an underlying SCM with structural equations as follows

\[ S = E_S \]
\[ X = (\text{Id} - B)^{-1} \cdot (S + E_H), \]

and noise distributions \( E_H \sim \mathbb{P}_H \) and \( E_S \sim \mathbb{P}_S \). The components of \( E_S \) (and all \( S_b \) below) are assumed mutually independent, while the components of \( E_H \) can in general be dependent. Instead of perfect interventions as considered in Chapter 3, we consider so-called shift or soft interventions that reflect how the distribution of \( S \) changes block-wise over time. That is, we drop the time/block index in our SCM and instead relate the non-stationarity of \( S \) to interventions (cf. Remark 24) in the intervention set

\[ \mathcal{I}_{S,X} = \{ \text{do}(S = S_b) : b \} \cup \{ \emptyset \}. \]

Replacing the equation of \( S \) by \( S = S_b \), the structural equations induce an interventional distribution denoted by \( \mathbb{P}_{S,X}^{\text{do}(S=S_b)} \) whose marginal distribution on \( S \) agrees with \( S_b \), i.e. \( \mathbb{P}_S^{\text{do}(S=S_b)} \sim S_b \). Considering this as the underlying SCM—what we called the \( X \)-level in Chapter 3—the problem can be phrased as aiming to infer the parameter \( B \) of the underlying model from observations of \( Y \equiv X = \tau(S, X) \) under unknown block-wise shift interventions. The measurement transformation \( \tau \) conceals \( S \) (and \( E_H \)) from us. Causal inference is hurdled in this scenario as the observations correspond to some interventional distributions

\[ \mathbb{P}_{Y}^{\text{do}(b_1)}, \ldots, \mathbb{P}_{Y}^{\text{do}(b_k)} \]

which correspond to unknown interventions \( b_1 \in \mathcal{I}_{S,X} \). corolICA adjusts for the stationary confounding thereby implicitly separating stationary confounding \( E_H \) from shift interventions \( S_b \), which enables estimation of \( B \) despite the measurement transformation \( \tau \). If one were to naively infer causal structure, say by directly interpreting the dependence structure between the components of \( X \), the results would in general be distorted by the dependences introduced by the confounding term \( H \).

### 5.4.1 ICA for Blind Cortical Source Separation in EEG

An EEG is not a direct measurement of cortical signals. Instead, the recorded electrode signals are an inevitable measurement transformation
of the underlying cortical activity. They are commonly assumed to be a (linear) superposition of cortical dipole signals [NS06]:

\[
\begin{align*}
X^1 & \quad X^2 & \quad X^3 \\
\text{observed linear mixture} & \\
\text{linear mixing} & \\
\text{causal variables} & 
\end{align*}
\]

In general, this transformation does not satisfy the conditions of an exact transformation in Definition 3. Additionally, the ground-truth model on the cortical dipole level as well as the precise linear transformation are unknown and causal reasoning on the level of observed EEG signals is unreasonable. Statements such as “manipulating the electrode’s signal affects the subject’s attentional state” are nonsensical. Yet, variables such as the activity in the parietal cortex, extracted as a linear combination of electrode signals, may admit meaningful causal statements such as “manipulating the activity in the parietal cortex affects the subject’s attentional state”. We therefore need to solve the blind source separation problem, otherwise we cannot sensibly talk about causal structure between cortical dipoles (see Section 4.7 for more on causality and variable transformations in neuroimaging).

Since [Mak+95], ICA is the routine preprocessing step to accomplish blind source separation of EEG recordings. It is unknown how exactly the cortical dipoles project onto the scalp level as well as how many sources there are and where they are located. In ICA one aims to invert the measurement transformation by assuming independent cortical source components. This assumption may seem counterintuitive at first especially since it is common practice to investigate the dependence structure and information flow between those cortical components as extracted by ICA. Yet, the practical success of ICA is undeniable and its routine application to EEG data yields remarkably consistent and neurophysiologically plausible results across studies and subjects [OM06; Ont+06].
Several potential ways of explanation can be found in the literature. First, it may be assumed that muscular activity or eye movements projecting onto electrode level are indeed independent of the cortical components and ICA thus enables the removal of artefactual components. Indeed, this warrants the use of ICA for removing artefactual non-brain components from EEG data [OM06], while the interpretation of the other components as cortical components remains unclear. Second, since ICA algorithms optimise empirical proxy measures of independence they capture only certain (instantaneous) dependences while still allowing for higher-order time-delayed dependences between sources such as between commonly considered time-windowed bandpower features [Mul+11; Hus+14]. The identifiability result in Theorem 21, for example, relies on the covariance matrices $\text{Cov}(S_k) - \text{Cov}(S_l)$ being diagonal which does not imply that the variance processes $S^i$ be necessarily independent. Third, the components extracted by ICA can be understood as being maximally independent which may be appropriate for separating near independent cortical sources and cancelling the volume conduction [Mak+04; OM06; Ont+06]. For example, let us consider cortical sources $S_1$ and $S_2$ that are weakly coupled and exert time-delayed influence onto each other. Extracting $X_1 \approx S_1$ and $X_2 \approx S_2$ may arguably still result in less dependent sources than any other mixing such as $\tilde{X}_1 \approx aS_1 + bS_2$ and $\tilde{X}_2 \approx cS_1 + dS_2$, which has the same source component represented in several extracted components.

Beyond the aforementioned arguments and empirical success, no theoretical justification for how ICA indeed recovers sources that admit an interpretation as cortical sources and no explicit plausible generative model with a more nuanced assumption than independent sources is known. So far, the available explanations rely in some way or another on empirical failure modes of ICA routines, e.g. by assuming that only maximally independent sources are recovered or the empirical dependence measure misses certain higher-order interactions. Our framework presented in Chapter 3 enables future research into how observations of linear mixtures can be leveraged for inference about the nature and possible causal interactions between the unknown underlying sources.

Considering the widespread use and adoption of ICA, our corolICA as introduced in this chapter is an important first improvement upon the state-of-the-art ICA methods in EEG analyses. Existing and most commonly employed ICA methods are limited by assuming noise-free observations or only accounting for time-independent noise. In these restrictive noise scenarios, corolICA performs competitive to existing methods while it out-
performs under group-wise stationary noise (cf. Section 5.5). Furthermore, EEG recordings of multiple subjects are routinely pooled together for a group analysis; here coroICA is the more appropriate and principled approach than other ad hoc approaches in order to obtain common sets of source components from multiple subject recordings (cf. Section 5.1.1.3 and Figure 5.9). In a way, the methodology behind coroICA can also be viewed as accounting, in first-order approximation, for small subject-specific variations in the cortical- to scalp-level projection.

5.5 EXPERIMENTS

In this section, we analyse empirical properties of coroICA. To this end, we first illustrate the performance of coroICA as compared to time-concatenated versions of (noisy) ICA variants on simulated data with and without confounding. We also compare on real data and outline potential benefits of using our method when analysing multi-subject EEG data.

5.5.1 Competing Methods

In all of our numerical experiments, we apply coroICA as outlined in Algorithm 1, where we partition each group based on equally spaced grids and run a fixed number of $10 \cdot 10^3$ iterations of the uwedge approximate joint diagonaliser. Unless specified otherwise, coroICA refers to coroICA (var) (i.e., the variance signal based version) and we explicitly write coroICA (var), coroICA (TD) and coroICA (var & TD) whenever appropriate to avoid confusion. We compare with all of the methods in Table 5.2. Since no Python implementation was publicly available, we implemented the choiICAs and SOBI methods ourselves also based on a fixed number of $10 \cdot 10^3$ iterations of the uwedge approximate joint diagonaliser. For fastICA we use the implementation from the scikit-learn Python library due to Pedregosa et al. [Ped+11] and use the default parameters.

For the simulation experiments in Section 5.5.2, we also compare to random projections of the sources, where the unmixing matrix is simply sampled with iid standard normal entries. The idea of this comparison is to give a baseline of the unmixing problem and enhance intuition about the scores’ behaviour. In order to illustrate the variance in this method, we generally sample 100 random projections and show the results for each of them. A random mixing does not lead to interpretable sources, thus we do not compare with random projections in the EEG experiments in Section 5.5.3.
5.5.2 Simulations

In this section, we investigate empirical properties of coroICA in well-controlled simulated scenarios. First off, we show that we can recover the correct mixing matrix given that the data is generated according to our model (5.2.1) and Assumptions 16 and 19 hold, while the other ICAs necessarily fall short in this setting (cf. Section 5.5.2.1). Moreover, in Section 5.5.2.2 we show that even in the absence of any confounding (i.e., when the data follows the ordinary ICA model and \( H \equiv 0 \) in our model) we remain competitive with all competing ICAs. Finally, in Section 5.5.2.3 we analyse the performance of coroICA for various types of signals and noise settings. Our first two simulation experiments are based on block-wise shifting variance signals, which we describe in Data Set 1 and our third simulation experiment is based on GARCH type models described in Data Set 2.

5.5.2.1 Dependence on Confounding Strength

For this simulation experiment, we sample data according to Data Set 1 and choose to simulate \( n = 10^5 \) (dimension \( d = 22 \)) samples from \( m = 10 \) groups where each group contains \( n/m = 10^4 \) observations. Within each group, we select a random partition consisting of \( |P^g| = 10 \) subsets while ensuring that these have the same size on average. We fix the signal strength to \( c_1 = 1 \) and consider the behaviour of coroICA (trained on half of the groups with an equally spaced grid of 10 partitions per group) for different confounding strengths \( c_1 = \{0.125, 0.25, 0.5, 1, 1.5, 2, 2.5, 3\} \). The results for 1000 repetitions are shown in Figure 5.6. To allow for a fair comparison we take the same partition size for choiICA (var).

Data Set 1: Block-wise shifting variance signals

For our simulations we select \( m \) equally sized groups \( \mathcal{G} := \{g_1, \ldots, g_m\} \) of the data points \( \{1, \ldots, n\} \) and for each group \( g \in \mathcal{G} \) construct a partition \( \mathcal{P}_g \). Then, we sample a model of the form

\[
X_i = A \cdot (S_i + C \cdot H_i),
\]

where the values on the right-hand side are sampled as follows:
• $A, C \in \mathbb{R}^{d \times d}$ are sampled with iid entries from $\mathcal{N}(0,1)$ and $\mathcal{N}(0, \frac{1}{d})$, respectively.

• For each $g \in \mathcal{G}$ the $H_i \in \mathbb{R}^d$ are sampled from $\mathcal{N}(0, \sigma_g^2 \text{Id}_d)$, where the $\sigma_g^2$ are sampled iid from $\text{Unif}(0.1, b_1)$.

• For each $g \in \mathcal{G}$ and $e \in \mathcal{P}_g$ the variables $S_i \in \mathbb{R}^d$ are sampled from $\mathcal{N}(0, \eta_e^2 \text{Id}_d)$, where the $\eta_e^2$ are sampled iid from $\text{Unif}(0.1, b_2)$.

The parameters $b_1$ and $b_2$ are selected in such a way that the expected confounding strength $c_1 = \mathbb{E}(\sigma_g^2)$ and variance signal strength $c_2 := \mathbb{E}(|\eta_e^2 - \eta_f^2|)$ are as dictated by the respective experiment. Due to the uniform distribution this reduces to

$$b_1 = 2c_1 - 0.1 \quad \text{and} \quad b_2 = 3c_2 + 0.1.$$

The results indicate that in terms of the MD index the competitors all become worse as the confounding strength increases. All competing ICAs systematically estimate an incorrect unmixing matrix. coroICA on the other hand only shows a very small loss in precision as confounding increases; the small loss is expected due to the decreasing signal to noise ratio. In terms of MCIS, the behaviour is analogous but slightly less well resolved; with increasing confounding strength the unmixing estimation of all competing ICAs is systematically biased resulting in bad separation of sources and high MCIS scores both out-of-sample and in-sample.

### 5.5.2.2 Efficiency in the Absence of Group Confounding

For this simulation experiment, we sample data according to Data Set 1 and choose to simulate $n = 2 \cdot 10^4$ (dimension $d = 22$) samples from $m = 10$ groups where each group contains $n/m = 2 \cdot 10^3$ observations. Within each group, we then select a random partition consisting of $|\mathcal{P}_g| = 10$ subsets while ensuring that these have the same size on average. This time, to illustrate performance in the absence of confounding, we fix the confounding strengths $c_1 = 0$ and consider the behaviour of coroICA (applied to half of the groups with an equally spaced grid of 10 partitions per group) for different signal strengths $c_2 = \{0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$. The results for 1000 repetitions are shown in Figure 5.7. Again, choiICA (var) is applied with the same partition size.

The results indicate that overall coroICA performs competitive in the confounding-free case. In particular, there is no drastic negative hit on the
5.5 Experiments

FIGURE 5.6: Results of the simulation experiment described in Section 5.5.2.1. Plot shows performance measures (MD: small implies close to truth; MCIS: small implies stable) for fixed signal strength and various confounding strengths. The difference between the competing ICAs and coroICA is more prominent for higher confounding strengths where the estimates of the competing ICAs are increasingly different from the true unmixing matrix and the sources become increasingly unstable.
FIGURE 5.7: Results of the simulation experiment described in Section 5.5.2.2. Plot shows performance measures (MD: small implies close to truth; MCIS: small implies stability) for data generated without confounding and for various signal strengths. These results are reassuring, as they indicate that when applied to data that follows the ordinary ICA model, coroICA still performs competitive to competing ICAs even though it allows for a richer model class.

performance of coroICA as compared to choiICA (var) in settings where the data follows the ordinary ICA model. The slight advantage compared to fastICA in this setting is due to the signal type which favours ICA methods that focus on variance signals.

5.5.2.3 Comparison with other Noisy ICA Procedures

To get a better understanding of how our proposed ICA performs for different signal and noise types, we compare it on simulated data as described in Data Set 2. We illustrate the different behaviour with respect to the different types of signal by applying all three of our proposed coroICA procedures (coroICA (var), coroICA (TD) and coroICA (var & TD)) and compare them to the
corresponding choiICA variants which do not adjust for confounding (choi-ICA (var), choiICA (TD) and choiICA (var & TD)). While all corolICA procedures can deal with any type of stationary noise, choiICA (TD) only works for time-independent noise and choiICA (var) and choiICA (var & TD) cannot handle any type of noise at all (see Table 5.2). Additionally, we also compare with fastICA to assess its performance in the various noise settings. The results are depicted in Figure 5.8.

**Data Set 2: GARCH simulation**

For this simulation we consider different settings of the confounded mixing model

\[ X_t = AS_t + H_t. \]

More precisely, we consider the following three different GARCH type signals: (i) changing variance, (ii) changing time-dependence, and (iii) both changing variance and changing time-dependence. For each of these signal types we consider two types of confounding (noise) terms: (a) time-independent and (b) time-dependent auto-regressive noise. For both we construct \( d \) independent processes \( \tilde{H}_1, \ldots, \tilde{H}_d \) and then combine them with a random mixing matrix \( C \) as follows

\[ H_t = C \cdot \tilde{H}_t. \]

Full details are given in Appendix A.2.1.

In all settings the most general method corolICA (var & TD) is able to estimate the correct mixing. The two signal specific methods corolICA (TD) and corolICA (var) are also able to accurately estimate the mixing in settings where a corresponding signal exists. It is also worth noting that they slightly outperform corolICA (var & TD) in these settings. In contrast, when comparing with the choiICA variants, corolICA is in general able to outperform the corresponding method. Only in the setting of a changing time-dependence with time-independent noise, choiICA (TD) is able to slightly outperform corolICA (TD).

### 5.5.2.4 Summary of the Performance of corolICA

In summary, corolICA performs well on a larger model class consisting of both the group-wise confounded as well as the confounding-free case. An advantage over all competing ICAs is gained in confounded settings (as shown in Section 5.5.2.1) while there is at most a small disadvantage in the
Figure 5.8: Results of the simulation experiment described in Section 5.5.2.3 and Data Set 2. Plots show performance (MD: small implies close to truth) for data generated with auto-regressive (AR) or iid noise and for var, TD, and var & TD signal as described in Data Set 2. coroICA (var & TD) is able to estimate the correct mixing in all of the considered settings, while others break whenever the more restrictive signal/noise assumptions are not met.
unconfounded case (cf. Section 5.5.2.2). This suggests that whenever the data is expected to contain at least small amounts of stationary noise or confounding, one may be better off using coroICA as the richer model class will guard against wrong results. The results in Section 5.5.2.3 further underline the robustness of our proposed method to various types of noise (and signals) for which other methods break. Again, even in settings that satisfy the assumptions of the more tailored methods coroICA remains competitive.

5.5.3 EEG Experiments

ICA is often applied in the analysis of EEG data. Here, we illustrate the potential benefit and use of coroICA for this. Specifically, we consider a multi-subject EEG experiment as depicted in Figure 5.9. The goal is to find a single mixing matrix that separates the sources simultaneously on all subjects. Our proposed model allows that the EEG recordings for each subject have a different but stationary noise term $H$.

We illustrate the applicability of our method to this setting based on two publicly available EEG data sets.

**Data Set 3: CovertAttention data**

This data set is due to Treder et al. [Tre+11] and consists of EEG recordings of 8 subjects performing multiple trials of covertly shifting visual attention to one out of 6 cued directions. The data set contains recordings of

- 8 subjects,
- for each subject there exist 6 runs with 100 trials,
- each recording consists of 60 EEG channels recorded at 1000 Hz sampling frequency, while we work with the publicly available data that is downsampled to 200 Hz.

Since visual inspection of the data revealed data segments with huge artefacts and details about how the publicly available data was preprocessed was unavailable to us, we removed outliers and high-pass filtered the data at 0.5 Hz. In particular, along each dimension we set those values to the median along its dimension that deviate more than 10 times the median absolute distance from this median. We further preprocess the data by re-referencing to common average refer-
subject a

\[ X_a = AS_a + H_a \]

subject b

\[ X_b = AS_b + H_b \]

**Figure 5.9:** Illustration of a multi-subject EEG recording. For each subject, EEG signals \( X \) are recorded which are assumed to be corrupted by subject-specific (but stationary) noise terms \( H \). The goal is to recover a single mixing matrix \( A \) that separates signals well across all subjects.
ence (car) and projecting onto the orthogonal complement of the null component. For our unmixing estimations, we use the entire data, i.e., including intertrial breaks.

For classification experiments (cf. Section 5.5.3.2) we use, in line with Treder et al. [Tre+11], the 8–12 Hz bandpass-filtered data during the 500–2000 ms window of each trial, and use the log-variance as bandpower feature [Lot+18]. The classification analysis is restricted to valid trials (approximately 311 per subject) with the desired target latency as described in Treder et al. [Tre+11].

Results on the CovertAttention Data Set 3 are presented here, while the results of the analogous experiments on the BCICompIV2a Data Set 4 are deferred to Appendix A.2.2. For both data sets, we compare the recovered sources of coroICA with those recovered by competing ICA methods. Since ground truth is unknown we report comparisons based on the following three criteria:

STABILITY AND INDEPENDENCE
We use MCIS (cf. Section 5.2.4) to assess the stability and independence of the recovered sources both in- and out-of-sample.

CLASSIFICATION ACCURACY
For both data sets there is label information available that associates certain time windows of the EEG recordings with the task the subjects were performing at that time. Based on the recovered sources, we build a classification pipeline relying on feature extraction and classification techniques that are common in the field [Lot+18]. The achieved classification accuracy serves as a proxy of how informative and suitable the extracted signals are.

TOPOGRAPHIES
For a qualitative assessment, we inspect the topographic maps of the extracted sources, as well as the corresponding power spectra and a raw time-series chunk. This is used to illustrate that the sources recovered by coroICA do not appear random or implausible for EEG recordings and are qualitatively similar to what is expected from other ICAs. Furthermore, we provide an overview over all components achieved on Data Set 3 by SOBI, fastICA, and coroICA in Appendix A.2.3, where components are well resolved when the corresponding topographic map and activation map are close to each other (cf. Section 5.2.4).
5.5.3.1 Stability and Independence

We aim to probe stability not only in-sample but also verify the expected increase in stability when applying the unmixing matrix to data of new unseen subjects, i.e., to new groups of samples with different confounding specific to that subject. In order to assess stability and independence of the recovered sources in terms of the MCIS both in- and out-of-sample and for different amounts of training samples, we proceed by repeatedly splitting the data into a training and a test data set. More precisely, we construct all possible splits into training and test subjects for any given number of training subjects. For each pair of training and test set, we fit an unmixing matrix using coroICA and all competing methods described in Section 5.5.1. We then compute the MCIS on the training and test data for each method separately and collect the results of each training-test split for each number of training subjects.

Results obtained on the CovertAttention data set (with equally spaced partitions of \( \approx 15 \) seconds length) are given in Figure 5.10 and the results for the BCICompIV2a data set (with equally spaced partitions of \( \approx 15 \) seconds length) are shown in Appendix A.2.2.1, Figure A.1. For both data sets the results are qualitatively similar and support the claim that the unmixing obtained by coroICA is more stable when transferred to new unseen subjects. While for the competing ICAs the instability on held-out subjects does not follow a clear decreasing trend with increasing number of training subjects, coroICA can successfully make use of additional training subjects to learn a more stable unmixing matrix.

Due to the characteristics and low signal-to-noise ratio in EEG recordings, the evaluation based on the absolute MCIS score is less well resolved than what we have seen in the simulations before. For this reason we additionally provide a more focused evaluation by considering the MCIS fraction: the fraction of the MCIS achieved on a subject by the respective competitor method divided by the MCIS achieved on that subject by coroICA when trained on the same subjects. Thus, this score compares MCIS on a per subject basis, where values greater than 1 indicate that the respective competing ICA method performed worse than coroICA. Figure 5.11 shows the results on the CovertAttention Data Set 3 confirming that coroICA can successfully incorporate more training subjects to derive a better unmixing of signals.
Figure 5.10: Experimental results for comparing the stability of sources (MCIS: small implies stable) trained on different numbers of training subjects (cf. Section 5.5.3.1), here on the CovertAttention Data Set 3, demonstrating that corolICA, in contrast to the competing ICA methods, can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.
Figure 5.11: Experimental results for comparing the stability of sources of the competing methods relative to the stability obtained by coroICA (MCIS fraction: above 1 implies less stable than coroICA) trained on different numbers of training subjects (cf. Section 5.5.3.1), here on the CovertAttention Data Set 3, demonstrating that coroICA can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.
5.5.3.2 Classification based on Recovered Sources

While the results in the previous section indicate that coroICA can lead to more stable separations of sources in EEG than the competing methods, in scenarios with an unknown ground truth the stability of the recovered sources cannot serve as the sole determining criterion for assessing the quality of recovered sources. In addition to asking whether the recovered sources are stable and independent variance signals, we hence also need to investigate whether the sources extracted by coroICA are in fact reasonable or meaningful. In the “America’s Got Talent Duet Problem” (cf. Example 18) this means that each of the recovered sources should only contain the voice of one (independent) singer (plus some confounding noise that is not the other singer). For EEG data, this assessment is not as easy. Here, we approach this problem from two angles: (a) in this section we show that the recovered sources are informative and suitable for common EEG classification pipelines, (b) in Section 5.5.3.3 we qualitatively assess the extracted sources based on their power spectra and topographic maps.

In both data sets there are labelled trials, i.e., segments of data during which the subject covertly shifts attention to one of six cues (cf. Data Set 3) or performs one of four motor imagery tasks (cf. Data Set 4). Based on these, one can try to predict the trial label given the trial EEG data. To mimic a situation where the sources are transferred from other subjects, we assess the informativeness of the extracted sources in a leave-k-subjects-out fashion as follows. We estimate an unmixing matrix on data from all but \( k \) subjects, compute bandpower features for each extracted signal and for each trial (as described in Data Set 3 and Data Set 4), and on top of those we train an ensemble of 200 bootstrapped shrinkage linear discriminant analysis classifiers where each is boosted by a random forest classifier on the wrongly classified trials. This pipeline (signal unmixing, bandpower-feature computation, trained ensemble classifier), is then used to predict the trials on the \( k \) held-out subjects.

The results are reported in Figure 5.12 and Appendix 5.5.3.2, Figure A.3 which show for each number of training subjects, the accuracies achieved on the respective held-out subjects when using the unmixing obtained on the remaining subjects by either coroICA or one of the competitor methods. The results on both data sets support the claim that the sources recovered by coroICA are not only stable but in addition also capture meaningful aspects of the data that enable competitive classification accuracies in fully-out-of-sample classification. The mean improvement in classification accuracy of coroICA over the other methods increases with increasing number of train-
Figure 5.12: Classification accuracies on held-out subjects (cf. Section 5.5.3.2), here on the CovertAttention Data Set 3. Gray regions indicate a 95% confidence interval of random guessing accuracies.

It is worth noting that these classification results depend heavily on the employed classification pipeline subsequent to the source separation. Here, our goal is only to show that coroICA does indeed separate the data into informative sources. In practice, and when only classification accuracy matters, one might also consider using a label-informed source separation [Däh+14], employ common spatial patterns [KLZ90] or use decoding techniques based on Riemannian geometry [Bar+12].

5.5.3.3 Topographic Maps

The components that coroICA extracts from EEG signals are stable (cf. Section 5.5.3.1) and meaningful in the sense that they contain information that enables classification of trial labels, which is a common task in EEG studies (cf. Section 5.5.3.2). In this section, we complement the assessment of the recovered sources by demonstrating that the results obtained by coroICA lead to topographies, activation maps, power spectra and raw time-series that are similar to what is commonly obtained during routine ICA analyses of EEG data when the plausibility and nature of ICA components is to be judged.
Topographies are common in the EEG literature to depict the relative projection strength of extracted sources to the scalp sensors. More precisely, the column-vector \( a_j \) of \( A = V^{-1} \) that specifies the mixing of the \( j \)-th source component is visualised as follows. A sketched top view of the head is overlayed with a heatmap where the value at each electrodes’ position is given by the corresponding entry in \( a_j \). These topographies are indicative of the nature of the extracted sources, for example the dipolarity of source topographies is a criterion invoked to identify cortical sources [Del+12] or the topographies reveal that the source mainly picks up changes in the electromagnetic field induced by eye movements. Another way to visualise an extracted source is an activation map, which is commonly obtained by depicting the vector \( \hat{\text{Cov}}(X) v_j^\top \) (where \( v_j \) is \( j \)-th row of unmixing matrix \( V \)) and shows for each electrode how the signal observed at that electrode covaries with the signal extracted by \( v_j \) [Hau+14]. Besides inspecting the raw time-series data, another criterion invoked to separate cortical from muscular components is the log power spectrum. For example, a monotonic increase in spectral power starting at around 20 Hz is understood to indicate muscular activity [Gon+03] and peaks in typical EEG frequency ranges are used to identify brain-related components.

In Figure 5.13, we depict the aforementioned criteria for three exemplary components extracted by coroICA on the CovertAttention Data Set 3. Following the discussion in Section 5.2.4 we show the activation maps as

\[
\text{DiffX}(v_j^\top) = \sum_{\mathcal{M} \in \mathcal{M}^*} \text{sign}(v_j^\top M v_j^\top) M v_j^\top,
\]

which captures variance changing signal and allows to assess the quality of a recovered source by comparison to the topographic map \( a_j \) (cf. Equation 5.2.4). Here, the idea is to demonstrate that coroICA components are qualitatively similar to components extracted by commonly employed SOBI-ICA or fastICA. Therefore, we choose to display one example of an ocular component (2\(^{nd}\) where the topography is indicative of eye movement), a cortical component (7\(^{th}\) where the dipolar topography, the typical frequency peak at around 8–12 Hz, and the amplitude modulation visible in the raw time-series are indicative of the cortical nature), and an artefactual component (51\(^{st}\) where the irregular topography and the high frequency components indicate an artefact). For comparison, we additionally show

---

5 These are commonly employed criteria which are also advised in the eeglab tutorial [DM04, https://sccn.ucsd.edu/wiki/Chapter_09:_Decomposing_Data_Using_ICA] and the neurophysiological biomarker toolbox wiki [Har+12, https://www.nbtwiki.net/doku.php?id=tutorial:how_to_use_ica_to_remove_artifacts].
for each component the topographies of the components extracted by SOBI-ICA or fastICA by matching the recovered source which most strongly correlates with the one extracted by coroICA. The components extracted by coroICA closely resemble the results one would obtain from a commonly employed ICA analysis on EEG data.

For completeness, we provide an overview over all components extracted on Data Set 3 by SOBI, fastICA, and coroICA (var) in the Supplementary Section A.2.3. Components are well resolved when the corresponding topographic map and activation map are close to each other (cf. Section 5.2.4), which, by visual inspection, appears to be more often the case for coroICA than for the competing methods.
5.6 CONCLUSION

In [P*W*+19], we propose a method for recovering independent sources corrupted by group-wise stationary confounding. It extends ordinary ICA to an easily interpretable model, which we believe is relevant for many practical problems as is demonstrated in Section 5.3.1 for climate data and Section 5.5.3 for EEG data. We give explicit assumptions under which the sources are identifiable in the population case (cf. Section 5.2.2). Moreover, we introduce a straightforward algorithm for estimating the sources based on the well-understood concept of approximate joint matrix diagonalisation. As illustrated in the simulations in Section 5.5.2, this estimation procedure performs competitive even for data from an ordinary ICA model, while additionally being robust and able to adjust for group-wise stationary confounding. For real data, we show that the corolICA model indeed performs reasonably on EEG data and leads to improvements in comparison to commonly employed approaches, while at the same time preserving an enhanced interpretation of the recovered sources.
The gold standard to infer the causal effect of a treatment is a randomized controlled trial, a scheme popularized by Sir Ronald A. Fisher [Con91]. Randomisation, however, is often not feasible, expensive, or unethical (e.g. randomly assigning subjects to a cumbersome, expensive, and potentially risky treatment) or even technically impossible (e.g. randomising the blood oxygen level in a brain region to investigate its behavioural effect).

Causal inference research has led to mathematical frameworks that allow to conceptualise the notion of causality and to formally proof which assumptions make causal inference possible when working with observational data from potentially non-randomised uncontrolled settings. Examples include constraint-based methods that leverage the Markov and faithfulness assumptions to link conditional (in)dependences to causal structure [Sch+98; SGS01; Pea09], methods that exploit assumptions on the function class or noise properties [Shi+06; Hoy+08b], information-geometric formalisations relying on a principle of “independence of cause and mechanism” [Jan+12], and methods harnessing the idea that the accuracy of predictions of a causal model (in contrast to a non-causal model) should be invariant to interventions [PBM16].

To date, it is mostly (implicitly) assumed that the observable and actually measured variables are the right players that allow for a causal description of the system under investigation. Causal variables and observed variables do not need to coincide, however, which hinders the applicability of these methods. For example, when historically LDL and HDL cholesterol were only measured via their sum, reasoning about the causes of heart disease based on total cholesterol lead to conflicting study findings, problems of ambiguous manipulations, and The Cholesterol Wars [Ste07] (cf. Section 3.1.1). Viewed within our framework, this mishap can be explained by the measurement transformation (here, total cholesterol = LDL + HDL) not being an exact transformation of the underlying model.

We argue that the applicability and success of causal modelling tools for real-world problems depends crucially on a better understanding of how transformations (and preprocessing) of observed variables break or preserve causal reasoning or are even a necessary step to enable causal modelling in the first place. Transformations of the observed variables may be
required to isolate causal signals thus enabling causal inference algorithms to perform on those transformed variables. Since we only ever observe any system via inevitable measurement transformations, causal variable construction is an essential—but mostly overlooked—step when causal inference algorithms are eventually and fruitfully to be applied to real-world problems.

In this thesis we present three approaches to bring causal modelling from pen and paper closer to a pragmatic use (the contributions and their impact are also summarised in Section 1.3):

• We contribute conceptually by introducing the notion of exact transformations between SCMs in Chapter 3. The idea is that causal models of the same system should be consistent with one another in the sense that they agree in their predictions of the effects of interventions. This provides a framework to tackle the conceptual problem that we run into when applying causal modelling techniques in practice: We wish to causally reason about relationships between observables or other engineered features while it is unclear whether these variables admit a consistent causal description of the system at hand. We provide theoretical justification for modelling only a subsystem of a more complex system and explicate under which precise conditions we may causally reason about macroscopic properties of a system.

• Causal terminology is prevalent and often introduced in the interpretation of neuroimaging studies, especially of so-called encoding and decoding models. Appreciating the urge for causal explanation and aiming to foster the adoption of causal inference tools that go beyond the established associational analyses, we advocate a pragmatic first-step which is to explicate causal underpinnings and warranted inferences of techniques that are already commonplace in a given application field. To this end, we provide the first exhaustive interpretation chart that theoretically undergirds the causal interpretation of feature relevance in encoding and decoding models. It has informed neuroimaging studies that build upon our work and further clarified ongoing debates about interpretational issues in the community (cf. Section 1.3 and Chapter 4).

• Improving upon existing ICA variants that are routinely used in the analysis of EEG data, we present a causality-inspired confounding-robust ICA in Chapter 5. This contribution is an example of how a causal perspective can be pragmatic in improving computational
methods even if per se they are not only used for causal analyses. Our novel coroICA has a natural motivation in linear causal structure learning, where our extension allows existing approaches to be extended to confounded causal structures with observations from different environments. ICA can be understood as inverting a causality-breaking measurement transformation in EEG. Our novel ICA routine allows for group-wise stationary noise, while previous methods work under no or only iid noise. Since it safeguards against a broader class of noise and enables the principled application to and interpretation of multi-subject recordings, coroICA qualifies for routine application in EEG analyses.

Causality in neuroscience and neuroimaging remains an open and challenging task for multiple reasons. While it is arguably obvious that the observables in neuroimaging do not admit an immediate causal description of brain activity and how neural activity gives rise to cognition, a remedy is less obvious. Even a mere spatial smearing can break our ability to obtain meaningful causal insights (cf. Section 4.7). Furthermore, confounding and the massive dimensionality reduction from a 100 billion neurons to a couple of thousand voxel blood-oxygen levels or oscillatory activity in a dozen of EEG signals pose a major challenge for causal inference [MK18].

The problem of causal feature learning, viewed as recovering an causality-preserving exact transformation of some underlying SCM, can guide us towards new meaningful concepts in and representations of the data. Furthermore, while interpreting ICA components of EEG data as corresponding to cortical dipoles that exert a causal effect onto each other is problematic (cf. Section 5.4.1), noise injections into a system may help in identifying a suitable transformation that partially inverts a causality-breaking measurement transformation.

Another concrete future line of enquiry is to generalise the notion of an exact transformation between SCMs in order to analyse the trade-off between model accuracy and model simplicity for causal modelling using structural causal models. For a transformation between two SCMs to be exact (and hence to consistently preserve causal reasoning), we required the posets \( P_{\tau(X)} \) and \( P_Y \) be equal which is a restrictive constraint (cf. Section 3.4). One could imagine and further analyse a ‘softening’ of this requirement such that the distributions in the posets are required to only be approximately equal (cf. Section 3.5). A slightly inaccurate model with a small number of variables may be preferable to an accurate but complex model. This also naturally leads to the idea of causal abstraction, that is,
how a complex causal model on a fine-grained level of detail can be simplified and abstracted while preserving ‘as much causal information as possible’ and still be a pragmatically useful model. For example, when deciding patient treatment plans we may wish to systematically abstract away from a causal model over several billions of neurons and instead work with a simpler abstraction thereof that is pragmatically useful, humanly tractable, and preserves the essential cause-effect information and predictions about distributional changes under interventions that are required by clinicians. High-level causal descriptions may further be of interest, especially ‘When the Map Is Better Than the Territory’ [Hoe17], and the effective information may be more suitably represented at a macro- than a micro-level [HAT13].

While the observational and interventional distributions can sensibly be compared by symmetric versions of KL-divergence or maximum mean discrepancy, a sensible definition of SCM abstractions not only needs to bound the distributional distances but will also require a conceptualisation of ‘causal expressiveness’ of a model. This will prompt the development of a characterisation (and possibly quantification) of how much and which exact ‘causal content’ a transformed/simplified causal model still preserves about the system under investigation.

We discussed the importance of an order-preserving $\omega$ to ensure a notion of causal consistency between two SCMs. Future research on characterising the conditions under which different properties of consistency between causal models hold is important – for instance, under which conditions we can preserve counterfactual reasoning, which we have not discussed in this thesis. Also, we only introduced the notion of an exact transformation while it remains an open question how to choose among the set of all possible exact transformations of an SCM subject to some desirable model properties.

Furthermore, a better characterisation of the relation between observed variables and (latent) causal (proxy) variables may prove useful for real-world scenarios considered in fairness in machine learning research. As with the discussed EEG example, the (re)construction of causal variables may be a necessary prerequisite to identify and meaningfully reason about the causes of bias in datasets or machine learning models trained on those datasets. A protected feature (e.g. race) may not be observed while it still affects the observables (e.g. name, place of birth) in a specific way; recovering the relevant variables and the causal pathways between them provides a useful tool to systematically account and correct for undesired bias in a given scenario (e.g. credit scoring).
A

APPENDIX

A.1 SUPPLEMENT: CAUSAL CONSISTENCY OF CAUSAL MODELS

Lemmata 25, 26, and 29 are adapted from \([RW^*+17]\); Lemmata 27 and 28 present further unpublished project contributions.

A.1.1 Useful Results regarding Contraction Mappings

Lemma 25. Suppose that the function

\[ f : \mathbb{R}^n \to \mathbb{R}^m \]
\[ x \mapsto f(x) \]

is a contraction mapping. Then, for any \( e \in \mathbb{R}^m \), so is the function

\[ f^* : \mathbb{R}^n \to \mathbb{R}^m \]
\[ x \mapsto f(x) + e \]

Proof. By definition, there exists \( c < 1 \) such that for any \( x, y \in \mathbb{R}^n \),

\[ \|f^*(x) - f^*(y)\| = \|(f(x) + e) - (f(y) + e)\| = \|f(x) - f(y)\| \leq c\|x - y\| \]

and hence \( f^* \) is a contraction mapping. \( \square \)

Lemma 26. Suppose that the function

\[ f : \mathbb{R}^n \to \mathbb{R}^n \]
\[ x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix} \]
is a contraction mapping. Then for any $m \leq n$, and $x_i^* \in \mathbb{R}$, $i \in [m]$, so is the function

$$f^* : \mathbb{R}^n \to \mathbb{R}^n$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} x_1^* \\ \vdots \\ x_m^* \\ f_{m+1}(x) \\ \vdots \\ f_n(x) \end{pmatrix}$$

**Proof.** By definition, there exists $c < 1$ such that for any $x, y \in \mathbb{R}^n$,

$$\|f^*(x) - f^*(y)\| = \left\| \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f_{m+1}(x) - f_{m+1}(y) \\ \vdots \\ f_n(x) - f_n(y) \end{pmatrix} \right\|$$

$$\leq \left\| \begin{pmatrix} f_1(x) - f_1(y) \\ \vdots \\ f_n(x) - f_n(y) \end{pmatrix} \right\|$$

$$= \|f(x) - f(y)\|$$

$$\leq c\|x - y\|$$

and hence $f^*$ is a contraction mapping. \qed

**Lemma 27.** Suppose that the linear map $A : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction mapping. Let $A^\square : \mathbb{R}^m \to \mathbb{R}^m$ denote the linear map induced by the bottom-right $m \times m$ submatrix of $A$.

Then $A^\square$ is also a contraction mapping.
**Proof.** Consider the following map:

\[
A^* : \mathbb{R}^n \rightarrow \mathbb{R}^n
\]

\[
\begin{pmatrix}
  x^* \\
  x^\square
\end{pmatrix}
\mapsto
\begin{pmatrix}
  0 \\
  A^\square x^* + A^\square x^\square
\end{pmatrix}
\]

By Lemma 26 this is also a contraction mapping. Hence there exists a \( c < 1 \) such that

\[
\|A^*v - A^*w\| \leq c\|v - w\|
\]

for any \( v, w \in \mathbb{R}^n \).

Now, let \( v^\square, w^\square \in \mathbb{R}^m \) and define

\[
v = \begin{pmatrix}
  0_{n-m} \\
  v^\square
\end{pmatrix} \in \mathbb{R}^n, \quad w = \begin{pmatrix}
  0_{n-m} \\
  w^\square
\end{pmatrix} \in \mathbb{R}^n
\]

It follows that:

\[
\|A^\square v^\square - A^\square w^\square\| = \|A^*v - A^*w\| \leq c\|v - w\| = c\|v^\square - w^\square\|
\]

and hence \( A^\square \) is a contraction mapping. \( \square \)

**Lemma 28.** Suppose that the linear map \( A : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a contraction mapping.

Then the eigenvalues of \( A \) are all strictly less than 1 in absolute value.

**Proof.** Suppose for contradiction that \( A \) is a contraction mapping with an eigenvalue \( |\lambda| \geq 1 \). Then there exists a corresponding eigenvector \( v \) such that \( Av = \lambda v \).

It follows that:

\[
\|Av - A0\| = \|\lambda v\| = |\lambda|\|v - 0\| \geq \|v - 0\|
\]

which contradicts the fact that \( A \) is a contraction mapping. \( \square \)

### A.1.2 Convergence under any Intervention

The following Lemma shows that \( A \) being a contraction mapping ensures that the sequence \( (X_t)_{t \in \mathbb{Z}} \) defined by \( M_X \) in Theorem 12 converges everywhere under any intervention \( i \in I_X \). That is, for any realisation \( (x_t)_{t \in \mathbb{Z}} \) of this sequence, its limit \( \lim_{t \to \infty} x_t \) as a sequence of elements of \( \mathbb{R}^n \) exists.
Lemma 29. Consider the SEM $\mathcal{M}_X$ in Theorem 12, and suppose that the linear map $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a contraction mapping. Then, for any intervention $i \in \mathcal{I}_X$, the sequence of $X_t$ converges everywhere.

Proof. Consider, without loss of generality, the intervention

$$\text{do}(X_t^j = x_j \ \forall t \in \mathbb{Z}, \forall j \leq m \leq n) \in \mathcal{I}_X$$

for $m \in \{n\}$ (for $m = 0$ this amounts to the null-intervention). The structural equations under this intervention are

$$\begin{cases}
X_{t+1}^k = x_k & \text{if } k \leq m \\
X_{t+1}^k = \sum_j A_{kj} X_t^j + E^k & \text{if } m < k \leq n
\end{cases}$$

and thus the sequence $X_t$ can be seen to transition according to the function $f = g \circ h$, where

$$h : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$v \mapsto w = Av + E$$

$$g : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ \vdots \\ x_m \\ w_{m+1} \\ \vdots \\ w_n \end{pmatrix}$$

By Lemma 25 and Lemma 26 (both in Appendix A.1.1), $f$ is a contraction mapping for any fixed $E$. Thus, by the contraction mapping theorem, the sequence of $X_t$ converges everywhere to a unique fixed point. \hfill \Box
A.2 SUPPLEMENT: CONFOUNDING-ROBUST SOURCE SEPARATION

A.2.1 Details on the GARCH Simulation

The GARCH model that we simulate from in Section 5.5.2.3 is specified as follows. We simulate sources \( S^1, \ldots, S^d \) from the following GARCH-type model

\[
\begin{align*}
\sigma_i^2 &= a_1 + a_2 \cdot (S_{i-1}^j)^2 + a_3 \cdot \sigma_{i-1}^2, \\
S_i^j &= b_1 S_{i-1}^j + \cdots + b_p S_{i-p}^j + \sigma_i \epsilon_i,
\end{align*}
\]

where the \( \epsilon_i \) are independent and standard normal. Moreover, the noise terms \( H^1, \ldots, H^d \) are assumed to be either given by the following AR-process

\[
H_i^j = c_1 H_{i-1}^j + \cdots + c_q H_{i-q}^j + \nu_i,
\]

where \( \nu_i \) are independent standard normal, \( q \) is uniformly distributed on \{1, \ldots, 10\} and \( c_i \) independent \( \mathcal{N}(0, 1/(i + 1)^2) \) or simply as iid \( \mathcal{N}(0, 1) \) random variables. The final data is then constructed according to the following equation

\[
X_i = A \cdot S_i + \tilde{H}_i,
\]

where \( \tilde{H}_i = ACH_i \) and \( A, C \in \mathbb{R}^{d \times d} \) are sampled with iid entries from \( \mathcal{N}(0, 1) \) and \( \mathcal{N}(0, \frac{1}{q}) \), respectively. To illustrate, the effect of the signal type we consider the following three settings.

- **Setting 1 (time-independent with changing variance)**
  Set \( a = (0.005, 0.026, 0.97) \) such that the variance changes over time and \( p = 0 \) to ensure time-independent signals. Based on these settings we sample \( n = 200000 \) observations.

- **Setting 2 (varying time-dependence structure with constant variance)**
  Set \( a = (1, 0, 0) \) such that the variance is fixed to 1. Then, sample \( p \) 100 times uniformly from \{1, \ldots, 10\} and \( b_i \) independent from \( \mathcal{N}(0, 1/(i + 1)^2) \) and simulate 2000 observations for each of the 100 parameter settings, leading to a total of \( n = 200000 \) observations.

- **Setting 3 (varying time-dependence structure with changing variance)**
  Set \( a = (0.005, 0.026, 0.97) \) such that the variance changes over time.
Then, we sample $p$ 100 times uniformly from \{1, \ldots, 10\} and $b_i$ independent from $\mathcal{N}(0, 1/(i + 1)^2)$ and simulate 2000 observations for each of the 100 parameter settings, leading to a total of $n = 200000$ observations.

A.2.2 EEG Experiments on Data Set 4

Analogous to Sections 5.5.3.1 and 5.5.3.2 we conducted experiments on the BCICompIV2a Data Set 4, the results of which are presented in the subsequent sections.

**Data Set 4: BCICompIV2a data**

This data set is due to Tangermann et al. [Tan+12, Section 5] and consists of EEG recordings of 9 subjects performing multiple trials of 4 different motor imagery tasks. The data set contains recordings of

- 9 subjects, each recorded on 2 different days,
- for each subject and day there exist 6 runs with 48 trials,
- each recording consists of 22 EEG channels recorded at 250 Hz sampling frequency,
- and is bandpass filtered between 0.5 and 100 Hz and is 50 Hz notch filtered.

For our analysis we only use the trial-data, i.e., the concatenated segments of seconds 3–6 of each trial (corresponding to the motor imagery part of the trials [Tan+12]). We further preprocess the data by re-referencing to common average reference (car) and projecting onto the orthogonal complement of the null component.

As features for classification experiments (cf. Section 5.5.3.2) on this data set we use bandpower in the 8–30 Hz band as measured by the log-variance of the 8–30 Hz bandpass-filtered trial data [Lot+18].
A.2.2.1 Stability and Independence

![Graph showing the stability of sources trained on different numbers of training subjects](image)

**Figure A.1:** Experimental results for comparing the stability of sources (MCIS: small implies stable) trained on different numbers of training subjects (cf. Section 5.5.3.1), here on the BCICompIV2a Data Set 4, demonstrating that coroICA, in contrast to the competing ICA methods, can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.
**Figure A.2**: Experimental results for comparing the stability of sources of competitors relative to the stability obtained by coroICA (MCIS fraction: > 1 implies less stable than coroICA) trained on different numbers of training subjects (cf. Section 5.5.3.1), here on the BCICompIV2a Data Set 4, demonstrating that coroICA can successfully incorporate more training subjects to learn more stable unmixing matrices when applied to new unseen subjects.
A.2.2.2 Classification based on Recovered Sources

![Classification accuracies on held-out subjects](image)

**Figure A.3:** Classification accuracies on held-out subjects (cf. Section 5.5.3.2), here on the BCICompIV2a Data Set 4. Gray regions indicate a 95% confidence interval of random guessing accuracies.
A.2.3  All Topographies and Activation Maps on Data Set 3

Figure A.4: Activation maps (left of each pair of columns) and topographies (right of each pair of columns) of 59 components extracted by SOBI on the CovertAttention Data Set 3. For components that are well resolved, both should look similar (cf. Section 5.2.4 and 5.5.3.3).
**Figure A.5:** Activation maps (left of each pair of columns) and topographies (right of each pair of columns) of 59 components extracted by fastICA on the CovertAttention Data Set 3. For components that are well resolved, both should look similar (cf. Section 5.2.4 and 5.5.3.3).
**Figure A.6:** Activation maps (left of each pair of columns) and topographies (right of each pair of columns) of 59 components extracted by coroICA (var) on the CovertAttention Data Set 3. For components that are well resolved, both should look similar (cf. Section 5.2.4 and 5.5.3.3).
<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Title</td>
<td>Details</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
</tr>
</thead>
</table>


Peer-Reviewed Journal Articles


Peer-Reviewed Conference Articles


Preprints and Working Papers
