The Foundations of Quantum Causal Inference: the Case of Machine Learning Methods

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilians-Universität
München

vorgelegt von

Omid Charrakh

aus

Mahshahr, Iran

2023

Referent/in: Prof. Dr. Stephan Hartmann

Korreferent/in: Prof. Dr. Thomas Augustin

Tag der mündlichen Prüfung: 03.07.2023

Acknowledgments

I would like to express my deepest gratitude to Prof. Dr. Stephan Hartmann, my advisor, for his guidance, support, and invaluable feedback throughout my graduate studies. His expertise, insights, and encouragement have been instrumental in shaping my research.

In the last two years of my doctoral studies, I worked as a research assistant in the chair of Statistical Learning and Data Science at the Department of Statistics at the Ludwig-Maximilians-University Munich, where I fundamentally got acquainted with the concepts of machine learning and statistical analysis. I am especially grateful to Prof. Dr. Bernd Bischl and Dr. Ludwig Bothmann for giving me such a golden opportunity. The knowledge I gained during this period had a special role in the innovations of this project.

I would also like to thank Jan Dziewior for his invaluable contributions to Chapters 4 and 5 of this dissertation. He brought a wealth of expertise and insights to the research, and his dedication and hard work were instrumental in the success of the project.

I want to express my heartfelt thanks to my dear friend Dr. Navid Shokouhi an expert in Machine Learning, for his invaluable support throughout the years of writing this dissertation. His depth of knowledge and willingness to answer my scientific questions have been influential in shaping the direction and content of this work.

I am deeply grateful to my partner Simin Hosseinzadeh for her invaluable assistance in translating Chapter 1 into German. She also helped me in creating some visualizations for my dissertation. I am truly fortunate to have had such a supportive and talented partner. Thank you, Simin, for your unwavering encouragement and dedication to this project.

I would also like to thank Dr. Michael Cuffaro and Dr. Jan Philipp Dapprich for their time, thoughtful feedback, and critical engagement with my work. Their suggestions and comments have helped me improve my research and write a better dissertation.

It is a pleasure to thank my colleagues, friends, and family, who have supported and inspired me in various ways throughout my graduate studies. Their encouragement, motivation, and camaraderie have made this journey more meaningful and enjoyable.

Contents

List of Figures

List of Tables

List of Algorithms

Einführung

Kausaler Eliminativismus, der von Bertrand Russell berühmt vertreten wurde, argumentiert, dass kausale Konzepte in der Wissenschaft keine Rolle spielen, weil die Gesetze in der fundamentalen Physik zeitlich symmetrisch sind, was bedeutet, dass die Zukunft die Vergangenheit genauso bestimmt, wie die Vergangenheit die Zukunft bestimmt (Russell 1912). Die Ansicht besagt, dass es keinen ontologischen Raum für kausale Beziehungen gibt, weil eine präzise Beschreibung der Anfangszustände in Verbindung mit dem zugrunde liegenden dynamischen Gesetz eine vollständige Beschreibung des Systemverhaltens in der fundamentalen Physik liefert.

Während Russell die Existenz einer objektiven, zugrunde liegenden Kausalbeziehung auf der mikroskopischen Ebene leugnete, gibt es mehrere Quellen von Asymmetrien, aus denen kausale Vorstellungen auf der makroskopischen Ebene entstehen könnten. Beispiele hierfür sind das Zweite Gesetz der Thermodynamik und der damit verbundene Zeitpfeil, die Möglichkeit von Interventionen sowie die unterschiedlichen Perspektiven verschiedener Akteure. Kausaler Eliminativismus wurde von Philosophen wie Nancy Cartwright moderiert, um kausalen Vorstellungen den Weg zurück in die Wissenschaft zu ermöglichen, insbesondere auf der makroskopischen Ebene. Für Cartwright besteht die Bedeutung kausaler Vorstellungen darin, dass sie die Mittel zur Unterscheidung zwischen effektiven und ineffektiven Strategien liefern: ”Kausale Gesetze können nicht beseitigt werden, denn sie sind nötig, um die Unterscheidung zwischen effektiven und ineffektiven Strategien zu begründen” (Cartwright 1979, S. 420).

Manipulierbarkeitstheorien der Kausalität sind eine etablierte Perspektive in der Philosophie der Wissenschaft, die Ursachen als Instrumente zur Manipulation von Effekten betrachten. Diese Theorien betonen, dass kausale Beziehungen durch die Fähigkeit zur Intervention und Manipulation von Variablen etabliert werden, wobei die Manipulation ein entscheidendes Kriterium zur Unterscheidung zwischen effektiven und ineffektiven Strategien darstellt. Obwohl Manipulierbarkeitstheorien der Kausalität historisch auf menschliche Interventionen ausgerichtet waren, haben zeitgenössische Versionen dieser Theorien ihren Fokus erweitert, um allgemeinere Formen der Intervention zu berücksichtigen und eine anthropozentrische Perspektive zu vermeiden.

Auf der anderen Seite konzentrieren sich probabilistische Theorien der Kausalität auf die Idee, dass kausale Beziehungen in der Regel von probabilistischen Abhängigkeiten begleitet werden. Normalerweise erhöht oder verringert ein Ereignis X die Wahrscheinlichkeit eines Ereignisses Y, wenn es dieses verursacht. Probabilistische Theorien der Kausalität beschreiben kausale Abhängigkeiten in Bezug auf probabilistische Abhängigkeiten. Mit anderen Worten versuchen sie, probabilistische Kriterien bereitzustellen, um zu entscheiden, ob X Y verursacht, und behaupten oft, dass Kausalität das entsprechende Muster probabilistischer Beziehungen ist. Die meisten probabilistischen Theorien der Kausalität werden von den folgenden zentralen Intuitionen motiviert: (i) Eine Veränderung der Ursache verändert ihre Effekte, und (ii) diese Veränderung zeigt sich in probabilistischen Abhängigkeiten zwischen Ursache und Wirkung (siehe Williamson 2009).

Manipulierbarkeits- und probabilistische Theorien der Kausalität werden oft als ergänzende Ansätze zur Erklärung von Kausalität angesehen. Manipulierbarkeitstheorien bieten eine Möglichkeit, darüber nachzudenken, wie wir intervenieren können, um bestimmte Effekte herbeizuführen oder zu verhindern, während probabilistische Theorien eine Möglichkeit bieten, darüber nachzudenken, wie wir Vorhersagen über die Wahrscheinlichkeit bestimmter Ereignisse basierend auf dem machen können, was wir über andere Ereignisse wissen. Die Kombination von Manipulierbarkeits- und probabilistischen Theorien der Kausalität hat zur Entwicklung von Berechnungsrahmen geführt, die probabilistische Theorien verwenden, um kausale Beziehungen zu identifizieren und die Auswirkungen bestimmter Interventionen vorherzusagen.

Die Theorie der kausalen Bayes-Netzwerke, die hauptsächlich von Judea Pearl und Clark Glymour entwickelt wurde (Pearl 2000, Spirtes et al. 2000), zeigt die erfolgreiche Integration von Manipulierbarkeits- und probabilistischen Theorien der Kausalität. Die Theorie bietet einen systematischen Ansatz zur Ableitung von kausalen Fakten aus statistischen Daten auf der Grundlage von einigen wohl begründeten Axiomen. Ein Modell, das von der Theorie der kausalen Bayes-Netzwerke generiert wird, wird oft als "graphisches kausales Modell" bezeichnet. Zwei zentrale Fragen, die ein graphisches kausales Modell zu beantworten versucht, sind:

- Gegeben die beobachteten statistischen Daten, wie kann man die zugrunde liegenden kausalen Beziehungen zwischen den Variablen entdecken?
- Gegeben die zugrunde liegenden kausalen Beziehungen, welche statistischen (Un-)Abhängigkeiten sollten erwartet werden?

Die grundlegende Idee besteht darin, die kausalen Beziehungen zwischen den Variablen eines Szenarios durch einen "kausalen Graphen" zu repräsentieren. In einem kausalen Graphen stellen die Knoten die Variablen des Szenarios dar und die Kanten repräsentieren die Richtung der Kausalität zwischen den Variablen. Wenn also die Variable X die Ursache der Variable Y ist, wird eine gerichtete Kante vom Knoten, der X entspricht, zum Knoten, der Y entspricht, gezogen. Die Anwesenheit einer solchen Kante bedeutet, dass, wenn man die Werte aller anderen Variablen außer X und Y in einem imaginären Experiment festlegt, durch Änderung des Wertes von X der Wert von Y geändert wird, während das Gegenteil nicht der Fall ist. In diesem Sinne ist die Theorie der kausalen Bayes-Netzwerke eine Anwendung von Manipulabilitätstheorien der Kausalität in der Praxis.

Um die zuvor genannten Fragen zu beantworten, gehen graphische Kausalmodelle oft davon aus, dass es einen zugrunde liegenden Daten-generierenden Prozess gibt, aus dem die statistischen Daten stammen. Die Struktur eines solchen zugrunde liegenden Prozesses wird dann durch Annahmen eingeschränkt, die durch unser Verständnis des Kausalitätsbegriffs motiviert sind. Die "kausale Markov-Bedingung" und die "kausale

Treuebedingung (causal faithfulness condition)" sind zwei bekannte Annahmen, die graphische Kausalmodelle in der Regel verwenden, um statistische (Un-)Abhängigkeiten mit Kausal (Un-)Abhängigkeiten in Verbindung zu bringen.

Während das Durchführen von Interventionen eine zuverlässige Möglichkeit ist, Kausalbeziehungen zu identifizieren, ist dies aufgrund praktischer Gründe wie Kosten und Ethik nicht immer möglich. In solchen Fällen muss die Identifizierung von Kausalbeziehungen ausschließlich auf Beobachtungsdaten basieren. Dieser Ansatz wird in der Wissenschaft seit mindestens dem siebzehnten Jahrhundert angewendet und ist beispielhaft für die Arbeit von Galileo, Pascal, Kepler und Newton (siehe Glymour et al. 2019). Ein wesentliches Merkmal der Theorie der kausalen Bayes-Netzwerke ist, dass sie Kausalinferenz auf der Grundlage von Beobachtungsdaten unterstützt. Obwohl die Theorie auf dem interventionistischen Kausalitätsverständnis beruht, ermöglicht sie unter bestimmten Umständen die Extraktion von Kausalwissen aus statistischen Daten ohne Durchführung von Interventionen. Es sei darauf hingewiesen, dass die Existenz von Methoden, die auf Beobachtungsdaten basieren, nicht im Widerspruch zum interventionistischen Konto stehen, da der Interventionist diese Ergebnisse als Zeigen interpretieren kann, *was passieren würde, wenn bestimmte Interventionen durchgeführt würden* (Woodward 2016*a*).

Observationsbasierte "kausale Entdeckungsalgorithmen" sind eine Familie von berechnungsmethoden, die darauf abzielen, kausale Beziehungen aus beobachtungsdaten aufzudecken. Diese Algorithmen verwenden statistische und berechnungstechniken, um kausale Strukturen zu identifizieren, die mit den Daten übereinstimmen, und stützen sich in der Regel auf Annahmen und Heuristiken, um die Suche nach dem plausibelsten kausalen Modell zu leiten. Der "Inductive Causation" (IC) (Verma 1993) und der Peter und Clark (PC) (Spirtes et al. 2000) sind bekannte Beispiele für Entdeckungsalgorithmen unter vielen anderen (weitere Übersichten siehe Guo et al. 2020, Glymour et al. 2019). Kausale Entdeckungsalgorithmen haben Anwendungen in verschiedenen Bereichen wie Wirtschaft, Epidemiologie, Genetik und Neurowissenschaften.

Trotz dieser Vielfalt von Anwendungen wurde festgestellt, dass kausale Entdeckungsalgorithmen bei der Anwendung auf bestimmte Quantenkorrelationen, die aus verschränkten

Quantensystemen entstehen, unbefriedigende Ergebnisse liefern. Insbesondere zeigten Wood & Spekkens (2015), dass die derzeitige Form von Entdeckungsalgorithmen Korrelationen, die Bells Ungleichung erfüllen, nicht von Korrelationen unterscheiden kann, die Bells Ungleichung verletzen. Folglich können sie den Herausforderungen der kausalen Erklärung von Quantenkorrelationen nicht gerecht werden. Das Problem zeigt daher einen schwerwiegenden Konflikt zwischen den zugrunde liegenden Axiomen der Theorie der kausalen Bayes'schen Netzwerke und bestimmten Vorhersagen der Quantentheorie auf. [1]

Zu beachten ist, dass das Problem, welches im letzten Absatz erwähnt wurde (eine Spannung zwischen dem nicht-lokalen Charakter der Quantenverschränkung (Quantum Entanglement) und der kausalen Entdeckungsalgorithmen), erst kürzlich entdeckt wurde. Zusätzlich zu diesem Problem hat die Quantenverschränkung eine weitere bekannte Schwierigkeit verursacht, die tatsächlich eine Spannung zwischen der Quantenmechanik und der Relativität darstellt. Gemäß Näger (2016) nenne ich diese Probleme "Das kausale Problem der Verschränkung" und "Das raumzeitliche Problem der Verschränkung" respektive. Obwohl sich diese Dissertation hauptsächlich auf das kausale Problem der Verschränkung konzentriert, ist es wichtig, das raumzeitliche Problem kurz zu erläutern, um zwischen den beiden Problemen zu unterscheiden.

Das raumzeitliche Problem der Verschränkung befasst sich mit der Frage, wie verschränkte Quantensysteme, die in räumlich getrennten Regionen der Raumzeit liegen, statistische Abhängigkeiten aufweisen können. Genauer gesagt kann das raumzeitliche Problem der Verschränkung als eine Inkonsistenz zwischen folgenden Aspekten angesehen werden:

1. Die Vorhersagen der Quantenmechanik

2. Eine bestimmte Vorstellung von Lokalität, die durch die Relativitätstheorie motiviert ist (d.h., dass kausale Einflüsse sich nicht schneller als das Licht ausbreiten können)

[1] Das Problem wurde von Philosophen bereits lange vor der Arbeit von Wood und Spekkens anerkannt. Wissenschaftler wie Butterfield (1989), Hausman (1999), Glymour (2006) haben alle erkannt, dass Quantenkorrelationen die Aufgabe erfordern, zumindest eine der Standardannahmen in der Theorie der kausalen Bayes'schen Netzwerke aufzugeben. Die Arbeit von Wood und Spekkens hat jedoch die Aufmerksamkeit vieler Forscher auf sich gezogen.

3. Prinzipien kausaler Erklärungen (z.B. die Annahme, dass alle Korrelationen kausal erklärt werden können)

Da die Vorhersagen der Quantenmechanik in verschiedenen Experimenten empirisch bestätigt wurden (beispielsweise Aspect et al. 1982), kann das spatiotemporale Problem als Dilemma zwischen der Aufgabe einer der beiden letztgenannten Annahmen betrachtet werden. Um das spatiotemporale Problem zu lösen, wurden in der Literatur eine erhebliche Anzahl von Lösungen vorgeschlagen. Während einige dieser Ansätze nicht-lokale Einflüsse postulieren (beispielsweise Egg & Esfeld 2014), tendieren die meisten Philosophen und Physiker dazu, die Lokalität aufrechtzuerhalten und zu dem Schluss zu kommen, dass zumindest eine der üblichen Prinzipien der kausalen Erklärung im quantenmechanischen Bereich versagt. Nicht-Screening-off gemeinsamer Ursachen (Butterfield 1989), nicht-kommutative gemeinsame Ursachen (Hofer-Szabó & Vecsernyés 2012), Retro-Kausalität (Price & Wharton 2015) und Super-Determinismus (Hooft 2009) sind Beispiele für solche Ansätze.

Das kausale Problem der Verschränkung ist eine neue Charakterisierung des Bell'schen Theorems, nach dem es selbst dann, wenn alle spatiotemporalen Einschränkungen ignoriert werden, immer noch einen Konflikt zwischen den Vorhersagen der Quantenmechanik und den üblichen Prinzipien der kausalen Erklärung gibt. Dieser Konflikt entsteht, weil die Verletzung der Bell-Ungleichung ohne die Verletzung mindestens eines Axioms der Theorie der kausalen Bayes-Netzwerke nicht kausal erklärt werden kann. Die beiden betroffenen Axiome sind die kausale Markov-Bedingung und die kausale Treuebedingung, die in der Theorie der kausalen Bayes-Netzwerke stark motiviert sind. Im quantenmechanischen Bereich führt das kausale Problem der Verschränkung zu einem Dilemma, dass entweder nicht-kausale Korrelationen vorliegen, die die kausale Markov-Bedingung verletzen, oder verursachte Unabhängigkeiten vorliegen, die die kausale Treuebedingung verletzen. Diese Inkonsistenz zwischen den beiden Axiomen kann physikalisch als Konflikt zwischen nicht-lokalen Verbindungen interpretiert werden, die notwendig sind, um Bell-Korrelationen kausal zu erklären, und den empirisch bestätigten no-signaling Unabhängigkeiten. Das kausale Problem der Verschränkung wirft daher die Frage auf, warum Bell'sche nicht-lokale Korrelationen nicht genutzt werden können, um superluminale Signale zu senden.

Eine mögliche Antwort auf das kausale Problem der Verschränkung besteht darin, zu argumentieren, dass es die Unzulänglichkeit der Interventionsansicht im quantenmechanischen Bereich aufdeckt. Daher müssen alternative Kausalitätskonten verwendet werden, um Quantenkorrelationen zu erklären. Zum Beispiel können die kausale Prozessansicht oder das kontrafaktische Kausalitätskonto verwendet werden, um Quantenkorrelationen zu erklären. [2]

Wenn man jedoch die Interventionsansicht im quantenmechanischen Bereich beibehalten möchte, vielleicht weil sie eine präzise Methode zur Bestimmung effektiver Strategien liefert, dann können zwei allgemeine Ansätze verfolgt werden, um das kausale Problem der Verschränkung anzugehen:

1. Die Beibehaltung des klassischen Rahmens der Kausalmodellierung und die Erklärung, warum bestimmte zugrunde liegende Annahmen des Rahmens im quantenmechanischen Bereich versagen.

2. Die Ablehnung des gesamten klassischen Rahmens der Kausalmodellierung und die Entwicklung einer explizit quantenmechanischen Verallgemeinerung des interventionistischen Kausalitätskonzepts.

Diese Arbeit diskutiert die beiden Ansätze zur kausalen Problematik der Verschränkung. Der erste Ansatz, wie er von Glymour (2006) und Näger (2016) exemplarisch dargelegt wird, lehnt grundlegende Prinzipien der kausalen Erklärung ab, wie etwa die kausale Markow-Bedingung und die kausale Treuebedingung. Der zweite Ansatz, wie er von Oreshkov et al. (2012), Barrett et al. (2019) und Shrapnel (2019) exemplarisch vertreten

[2] Zum Beispiel liefert Maudlins Untersuchung der kausalen Implikationen der Verletzung von Bells Ungleichung, die auf einer kontrafaktischen Analyse von Kausalität und einer ausreichenden Bedingung für eine kausale Verbindung zwischen raumartig getrennten Ereignissen beruht, Einblick in diese Frage. Maudlin (2011) betrachtet vier Möglichkeiten: (1) Kein superluminaler Materie- oder Energieaustausch, (2) Keine superluminalen Signale, (3) Keine superluminalen kausalen Einflüsse und (4) Kein superluminaler Informationsaustausch. Er zeigt, dass die Verletzung keine Verstöße gegen die ersten beiden Einschränkungen erfordert, sondern Verstöße gegen die letzten beiden.

wird, versucht, die Konzepte von Knoten, Mechanismen und Interventionen im quantenmechanischen Bereich neu zu definieren. Während die Komplexität der quantenmechanischen Verschränkung die Akzeptanz von gegenintuitiven Ergebnissen erfordert, kann die Ablehnung grundlegender Prinzipien unser Verständnis kausaler Beziehungen erheblich beeinträchtigen. In Kapitel 2 untersucht die Arbeit diese Ansätze kritisch, aber das Hauptmotiv ist die Einführung neuer Werkzeuge zur Untersuchung der kausalen Problematik der Verschränkung.

Diese Arbeit geht das kausale Problem durch eine intrinsisch empirische Methode an. Das Ziel besteht darin, die jüngsten Fortschritte im Maschinenlernen zu nutzen, um traditionelle Probleme in den Grundlagen der Quantenmechanik zu lösen. Insbesondere nutze ich Maschinelles Lernen, um neue Erkenntnisse über den Begriff der Kausalität im quantenmechanischen Bereich im Rahmen kausaler Bayes'scher Netzwerke zu gewinnen.

Maschinelles Lernen ist ein Studienbereich, der eine Vielzahl von Algorithmen und Modellierungstechniken umfasst, die es Systemen ermöglichen, aus Erfahrungen zu lernen und bestimmte Aufgaben ohne explizite Programmierung auszuführen. In der Ära von Big Data und hoch effizienten Algorithmen hat das Maschinenlernen weite Verbreitung gefunden und wird in verschiedenen Bereichen wie selbstfahrenden Autos, Wettervorhersage, Arzneimittelforschung, Wirtschaft und Sozialwissenschaften genutzt. Trotz dieser weit verbreiteten Nutzung in empirischen und sozialen Wissenschaften wurde das Maschinenlernen in grundlegenden Bereichen selten genutzt.

Die Hauptmotivation dieser Dissertation besteht darin, Machine Learning für grundlegende Debatten in der Quantenmechanik zu nutzen. Die in dieser Arbeit behandelten Themen bilden ein Dreieck, dessen drei Eckpunkte Machine Learning, Quantenmechanik und Kausales Modellieren sind. Die Arbeit umfasst zwei separate Projekte, die beide Machine Learning-Algorithmen zur Entdeckung kausaler Beziehungen in simulierten Quantenexperimenten nutzen. Das erste Projekt basiert auf dem berüchtigten EPR-Bell-Szenario. Ziel ist es, die Leistung verschiedener in der Literatur vorgeschlagener Modelle zur Lösung des kausalen Problems der Verschränkung zu vergleichen. Die Rolle von Machine Learning im ersten Projekt besteht darin, eine vereinheitlichende Arena für die empirische Bewertung zwischen verschiedenen Kandidatenmodellen zu schaffen. Das zweite Projekt betrachtet Szenarien, deren zugrunde liegende kausale Strukturen bereits bekannt sind. Das Ziel besteht darin, herauszufinden, inwieweit Machine Learning aus statistischen Daten ohne explizite Interventionen die wahren kausalen Strukturen identifizieren kann. Daher fungiert die Rolle von Machine Learning im zweiten Projekt eher als kausaler Entdeckungsalgorithmus, der an Quantensystemen arbeitet.

In den letzten Jahren haben sich kausale Inferenzmethoden auf der Basis von beobachteten Daten stark weiterentwickelt, hauptsächlich dank der schnellen Fortschritte im Bereich des Maschinellen Lernens und der Kombination mit kausalen Inferenztechniken. Mein Fokus liegt auf einigen der jüngsten Entwicklungen, die die Kraft des Maschinellen Lernens nutzen, um kausale Beziehungen in Situationen zu lernen, in denen herkömmliche Methoden nicht anwendbar sind. Genauer gesagt kann ein kausaler Entdeckungsalgorithmus bei Fehlen zusätzlicher Informationen zu einem Szenario, wie der zeitlichen Reihenfolge der Variablen oder der funktionalen Form von Mechanismen, kein eindeutiges kausales Graphen liefern, und seine Ausgabe enthält eine Reihe von ungerichteten Kanten. Um diese Einschränkung zu adressieren, wurde in den letzten Jahren eine Klasse von Entdeckungsalgorithmen entwickelt, die als "paarweise Algorithmen" bekannt sind, deren Schwerpunkt auf der Identifikation der kausalen Richtung in bivariaten Szenarien liegt, d.h. Szenarien, die nur aus zwei Variablen bestehen (für Überblicke siehe Mooij et al. 2016, Guyon et al. 2019).

Die Grundidee der paarweisen Algorithmen besteht darin, dass statistische Daten neben bedingten (Un)Abhängigkeiten verschiedene Arten von Asymmetrien enthalten, die zur Erkennung der kausalen Richtung verwendet werden können. Die Asymmetrien können eine Vielzahl von statistischen Eigenschaften umfassen, wie bedingte Entropie, gegenseitige Information, Momente, Reste von Regressionen und Standardabweichung bedingter Verteilungen. Paarweise Algorithmen können in zwei Kategorien eingeteilt werden: generative Algorithmen, die auf der Notion funktionaler kausaler Modelle (explizit oder implizit) basieren, und diskriminative Algorithmen, die die kausale Richtung direkt mit einem Klassifikationsalgorithmus bestimmen. Der in Kapitel 4 diskutierte Algorithmus ist generativ, während der in Kapitel 5 diskutierte Algorithmus diskriminativ ist.

Im Allgemeinen sind paarweisen Algorithmen hauptsächlich auf die Bemühungen von Machine-Learning-Experten zurückzuführen, die an Kausalitätsinferenz-Methoden interessiert sind, und haben zur Entdeckung neuer Aspekte der Theorie der kausalen Bayes-Netzwerke geführt. Obwohl solche Methoden als Fortsetzung des interventionistischen Kausalitätsansatzes verstanden werden können, haben sie weniger Aufmerksamkeit von Philosophen erhalten. Obwohl Climenhaga et al. (2021) kürzlich die Themen aus philosophischer Sicht untersucht hat. Die vorliegende Dissertation ist ein Versuch, diese jüngsten Fortschritte zur Bewältigung des kausalen Problems der Verschränkung in der Quantenmechanik zu nutzen.

Die Ergebnisse dieser Dissertation haben zwei wesentliche Beiträge zu den grundlegenden Diskussionen in der Quantenmechanik. Der erste Beitrag besteht darin, die Idee zu unterstützen, dass ein interventionistischer Kausalitätsansatz im quantenmechanischen Bereich möglich ist. Während viele glauben, dass der interventionistische Ansatz nicht auf den quantenmechanischen Bereich anwendbar ist (zum Beispiel Pearl 2009, Koller & Friedman 2009), haben einige Physiker auf der Grundlage des interventionistischen Ansatzes kausale Modelle für Quantensysteme vorgeschlagen (zum Beispiel Costa & Shrapnel 2016, Allen et al. 2017). Die Ergebnisse dieser Dissertation (hauptsächlich Kapitel 5) unterstützen letztere Idee und legen nahe, dass quantenmechanische Kausalbeziehungen durch generalisierte Begriffe des interventionistischen Ansatzes beschrieben werden können. Es ist jedoch notwendig zu betonen, dass diese Ergebnisse nicht bedeuten sollen, dass der interventionistische Ansatz der einzige Weg ist, um quantenmechanische Kausalbeziehungen zu charakterisieren. Andere Ansätze wie Prozesstheorien der Kausalität (Salmon 1984, Dowe 2000) könnten auch eine solche Charakterisierung bieten. Diese Ansätze werden jedoch in der vorliegenden Arbeit nicht untersucht.

Die zweite Beitrag dieser Dissertation besteht darin, die Idee zu unterstützen, dass Machine Learning ein zuverlässiges Instrument ist, um neue Perspektiven auf die Grundlagen der Quantenmechanik zu erforschen. Zum Beispiel zeigt das erste Projekt, wie traditionelle Diskussionen über das EPR-Bell-Szenario in Machine-Learning-Konzepte wie Verlustfunktionen umgewandelt und aus anderen Perspektiven betrachtet werden können. Darüber hinaus zeigt das zweite Projekt das Potenzial der Kombination von Kausalmodellierung mit Machine-Learning-Techniken, um einen quantenprozess kausal zu charakterisieren, auch wenn begrenzte Informationen über den Prozess verfügbar sind. Im Allgemeinen betonen die Ergebnisse dieser Arbeit das Potenzial von Machine-Learning-Ansätzen in den Grundlagen der Quantenmechanik und eröffnen neue Wege für weitere Erforschung und Untersuchung.

Inhalt

Diese Dissertation besteht aus sechs Kapiteln, einschließlich des aktuellen. Im Folgenden wird eine kurze Übersicht über den Inhalt jedes Kapitels gegeben.

Kapitel 2 bildet das Fundament der gesamten Dissertation und umfasst drei Hauptthemen: Machine Learning, Kausales Modellieren und Quantenmechanik. Der Abschnitt Machine Learning vermittelt ein Verständnis der grundlegenden Konzepte von überwachtem Lernen und dem Funktionsprinzip künstlicher neuronaler Netze. Der Abschnitt zum Kausalen Modellieren gibt einen kurzen Überblick über das interventionistische Kausalitätsmodell, gefolgt von einer Erklärung kausaler Bayes-Netzwerke, kausaler Entdeckungsalgorithmen und Paarweiser Methoden. Der Abschnitt zur Quantenmechanik geht tiefer auf das kausale Problem der Verschränkung und deren Beziehung zu den kausalen Markov- und Treueannahmen ein.

Kapitel 3 untersucht verschiedene Vorschläge zur Lösung des kausalen Problems der Verschränkung. Die zentrale Aussage des Kapitels ist, dass es zwar gemeinhin als Dilemma angesehen wird, im Quantenbereich entweder die kausale Markov-Bedingung oder die Treueannahmen aufzugeben, jedoch alternative Möglichkeiten aufgrund verschiedener physikalischer und philosophischer Überlegungen existieren. Das Kapitel zielt darauf ab, diese Möglichkeiten zu untersuchen, ihre Gültigkeit zu bewerten, die Konsequenzen zu untersuchen und die Zusammenhänge zwischen ihnen zu erforschen. Dieses Kapitel legt den philosophischen und grundlegenden Grundstein für die Machine Learning-Projekte in den folgenden Kapiteln.

Kapitel 4 bietet einen neuen Ansatz zur Lösung des kausalen Problems im Kontext einer kontinuierlichen Formulierung des EPR-Bell-Szenarios. Es wird ein Rahmen vorgestellt, um unterschiedliche Vorschläge für das kausale Problem empirisch zu bewerten. Der Rahmen basiert auf einem klassischen Algorithmus zur kausalen Entdeckung namens Causal Generative Neural Network (Goudet et al. 2018), der anhand philosophischer Überlegungen an das kausale Problem angepasst wurde. Die zentrale Idee besteht darin, die Vorhersagekraft verschiedener Kandidatenmodelle zu vergleichen und dasjenige zu finden, das Daten generiert, deren Verteilung der ursprünglichen Quantendaten am nächsten kommt. Dieser Rahmen ermöglicht den empirischen Vergleich einer breiten Palette von Modellen für das kausale Problem, ohne im Markov-Faithfulness-Dilemma stecken zu bleiben.

Kapitel 5 untersucht Quantenszenarien, deren kausale Strukturen bereits bekannt sind. Der Kern des Kapitels ist ein paarweiser Entdeckungsalgorithmus namens "Randomized Causation Coefficient" (Lopez-Paz et al. 2015). Obwohl der Algorithmus ursprünglich zur Unterscheidung von Ursache und Wirkung in klassischen Szenarien konzipiert wurde, wird er in mehreren Aspekten verallgemeinert, um einen Quantenkausalitätsentdeckungsalgorithmus zu erstellen, der bivariante und multivariate Quantenszenarien behandeln kann. Der resultierende Algorithmus funktioniert gut in verschiedenen simulierten Szenarien und hat Auswirkungen sowohl auf die Grundlagen der Quantenphysik als auch auf die Quanten-Engineering.

Kapitel 6 bietet eine Zusammenfassung der aus verschiedenen Richtungen gezogenen Schlussfolgerungen. Das Kapitel behandelt zwei Hauptpunkte: (1) die Auswirkungen der vorliegenden Ergebnisse auf den interventionistischen Kausalitätsbegriff im Quantendomäne und (2) die Relevanz von Machine Learning für Diskussionen in der Philosophie der Physik. Zusätzlich biete ich eine kurze Erklärung meiner Idee, wie die beiden Probleme der Verschränkung miteinander verbunden werden können.

Chapter 1

Introduction

Causal eliminativism, famously advocated by Bertrand Russell, argues that causal concepts have no role in science because the laws in fundamental physics are time-symmetric, meaning that the future determines the past in the same way the past determines the future (Russell 1912). The view contends that there is no ontological space for causal relations because a precise description of initial states, coupled with the underlying dynamical law, provides a complete description of system behavior in fundamental physics.

While Russell denied the existence of an objective, underlying relation of causation at the microscopic level, there are several sources of asymmetries from which causal notions might arise at the macroscopic level. Examples include the Second Law of Thermodynamics and the associated arrow of time, the possibility of performing interventions, and the different perspectives of different agents. More broadly, causal eliminativism was moderated by philosophers such as Nancy Cartwright to allow for causal notions to make their way back into science, particularly at the macroscopic level. For Cartwright, the importance of causal notions is that they provide the means to distinguish effective strategies from ineffective ones: "causal laws cannot be done away with, for they are needed to ground the distinction between effective strategies and ineffective ones" (Cartwright 1979, p.420).

Manipulability theories of causation are a well-established perspective in the philosophy of science that view causes as instruments for manipulating effects. These theories emphasize that causal relationships are established through the capacity to intervene and manipulate variables, making manipulation a key criterion to differentiate between effective and ineffective strategies. Although manipulability theories of causation have historically focused on human intervention, contemporary versions of these theories have expanded their scope to more general forms of intervention in order to avoid being anthropocentric.

On the other hand, probabilistic theories of causation focus on the idea that causal relationships are typically accompanied by probabilistic dependencies. Normally, when an event X causes an event Y, the former raises or lowers the probability of the latter. Probabilistic theories of causation characterize causal dependencies in terms of probabilistic dependencies. In other words, they attempt to provide probabilistic criteria to decide whether X causes Y and often maintain that causation is the corresponding pattern of probabilistic relationships. Most probabilistic theories of causation are motivated by the following central intuitions: (i) changing a cause alters its effects, and (ii) this alteration shows up in probabilistic dependencies between cause and effect (see Williamson 2009).

Manipulability and probabilistic theories of causation are often seen as complementary approaches to understanding causation. Manipulability theories provide a way of thinking about how we can intervene to bring about or prevent certain effects, while probabilistic theories provide a way of thinking about how we can make predictions about the probability of certain events occurring based on what we know about other events. The combination of manipulability and probabilistic theories of causation has led to the development of computational frameworks that use probabilistic theories to identify causal relationships and predict the effects of certain interventions.

The theory of causal Bayesian networks, primarily developed by Judea Pearl and Clark Glymour (Pearl 2000, Spirtes et al. 2000), exemplifies the successful integration of manipulability and probabilistic theories of causation. The theory offers a systematic approach to inferring causal facts from statistical data based on a few well-motivated axioms. A model generated by the theory of causal Bayesian networks is often referred to as a "graphical causal model." Two central questions that a graphical causal model aims to answer are:

- Given the observed statistical data, how can one discover the underlying causal relationships between the variables?

- Given the underlying causal relationships, what statistical (in)dependencies should be expected to be observed?

The basic idea is to represent the causal relationships among variables involved in a scenario by a "causal graph". In a causal graph, the nodes represent the variables of the scenario, and the edges represent the direction of causation between the variables. Thus, whenever variable X is the cause of variable Y, a directed edge is drawn from the node corresponding to X to the node corresponding to Y. The presence of such an edge makes the claim that if one fixes the values of all other variables except X and Y in an imaginary experiment, then by changing the value of X, the value of Y changes, while the converse is not true. In this sense, the theory of causal Bayesian networks is an instance of using manipulability theories of causation in practice.

To answer the questions mentioned earlier, graphical causal models often presume the existence of an underlying data-generating process from which the statistical data is aroused. The structure of such an underlying process is then constrained by assumptions motivated by our understanding of the notion of causation. The "causal Markov condition" and the "causal faithfulness condition" are two well-known assumptions that graphical causal models usually make to relate statistical (in)dependencies to causal (in)dependencies.

While performing interventions is a reliable way to identify causal relationships, it is not always feasible due to practical reasons such as cost and ethics. In such cases, identifying causal relationships must be based solely on observational data. This approach has been used in science since at least the seventeenth century, exemplified by the work of Galileo, Pascal, Kepler, and Newton (see Glymour et al. 2019). An essential feature of the theory of causal Bayesian networks is that it supports causal inference based on observational data. That is, although the theory is based on the interventionist account of causation, under certain circumstances, it allows for extracting causal knowledge from statistical data without performing interventions. Note that the existence of observational-based methods does not conflict with the interventionist account because the interventionist can interpret these results as showing *what would happen if certain interventions were be performed* (Woodward 2016*a*).

Observational-based "causal discovery algorithms" are a family of computational methods that aim to uncover causal relationships from observational data. These algorithms employ statistical and computational techniques to identify causal structures consistent with the data and typically rely on assumptions and heuristics to guide the search for the most plausible causal model. The Inductive Causation (IC) (Verma 1993) and the Peter and Clark (PC) (Spirtes et al. 2000) algorithms are well-known examples of discovery algorithms among others (for overviews see Guo et al. 2020, Glymour et al. 2019). Causal discovery algorithms have applications in various fields, such as economics, epidemiology, genetics, and neuroscience.

Despite such a diversity of applications, it has been noticed that causal discovery algorithms exhibit unsatisfactory results when applied to certain quantum correlations arising from entangled quantum systems. In particular, Wood & Spekkens (2015) showed that the current form of discovery algorithms could not distinguish correlations that satisfy Bell's inequality from correlations that violate Bell's inequality and, consequently, that they cannot do justice to the challenges of explaining quantum correlations causally. Thus, the problem demonstrates a serious conflict between the underlying axioms of the theory of causal Bayesian networks and certain predictions of quantum theory.[1]

Note that the problem mentioned in the last paragraph (which is a tension between the non-local character of quantum entanglement and causal discovery algorithms) has been found only recently. In addition to this problem, quantum entanglement has caused another well-known difficulty which, in fact, is a tension between Quantum Mechanics and Relativity. Following Näger (2016), I call these problems "the causal problem of

[1] The issue has been acknowledged by philosophers well long before the work of Wood and Spekkens. For instance, scholars such as Butterfield (1989), Hausman (1999), Glymour (2006) have all recognized that quantum correlations require the abandonment of at least one of the standard assumptions in the theory of causal Bayesian networks. However, the work by Wood and Spekkens has garnered the attention of many researchers.

entanglement" and "the spatiotemporal problem of entanglement," respectively. Although this dissertation primarily focuses on the causal problem of entanglement, it is important to briefly explain the spatiotemporal problem to differentiate between the two problems.

The spatiotemporal problem of entanglement deals with the question of how entangled quantum systems, being located in spacelike separated regions of spacetime, can exhibit statistical dependencies. More precisely, the spatiotemporal problem of entanglement can be regarded as an inconsistency between:

1. The predictions of Quantum Mechanics
2. A particular notion of locality motivated by Relativity (i.e., that causal influences cannot propagate faster than light)
3. Principles of causal explanation (e.g., the assumption that all correlations can be explained causally)

Since the predictions of Quantum Mechanics have been confirmed empirically in various experiments (for example Aspect et al. 1982), the spatiotemporal problem can be seen as a dilemma between abandoning either of the latter two assumptions. To deal with the spatiotemporal problem, a significant number of solutions have been proposed in the literature. While a few of these approaches postulate non-local influences (for example Egg & Esfeld 2014), the majority of philosophers and physicists tend to uphold locality and conclude that at least one of the usual principles of causal explanation fails in the quantum domain. Non-screening off common causes (Butterfield 1989), non-commutative common causes (Hofer-Szabó & Vecsernyés 2012), retro-causation (Price & Wharton 2015), and super-determinism (Hooft 2009) are the examples of such accounts.

The causal problem of entanglement is a new characterization of Bell's theorem, according to which even if one disregards all spatiotemporal constraints, there is still a conflict between the predictions of Quantum Mechanics and the usual principles of causal explanation. This conflict emerges because the violation of Bell's inequality cannot be explained causally without violating at least one axiom of the theory of causal Bayesian networks.

The two axioms in question are the causal Markov condition and the causal faithfulness condition, which are highly motivated in the theory of causal Bayesian networks. In the quantum domain, the causal problem of entanglement leads to a dilemma that either there are uncaused correlations, which violate the causal Markov condition, or there are caused independencies, which violate the causal faithfulness condition. This inconsistency between the two axioms can be viewed physically as a conflict between non-local connections necessary to explain Bell correlations causally and the empirically confirmed no-signaling independencies. Therefore, the causal problem of entanglement raises the question of why Bell's non-local correlations cannot be utilized to send superluminal signals.

One possible response to the causal problem of entanglement is to argue that it exposes the inadequacy of the interventionist view in the quantum domain. Therefore, alternative accounts of causation must be employed to account for quantum correlations. For example, the causal process view or the counterfactual account of causation may be used to explain quantum correlations.[2]

However, if one wishes to maintain the interventionist account in the quantum domain, perhaps because it provides a precise method for determining effective strategies, then two general approaches can be taken to address the causal problem of entanglement:

1. Maintaining the classical framework of causal modeling and explaining why certain underlying assumptions of the framework fail in the quantum domain.

2. Rejecting the entire classical causal modeling framework and developing an explicitly quantum generalization of the interventionist account of causation.

This thesis discusses the two approaches to the causal problem of entanglement. The first approach, exemplified by Glymour (2006) and Näger (2016), rejects fundamental principles

[2] For instance, Maudlin's investigation of the causal implications of the violation of Bell's inequality, based on a counterfactual analysis of causation and a sufficient condition for a causal connection between spacelike separated events, provides insight into this issue. Maudlin (2011) considers four possibilities: (1) no superluminal matter or energy transport, (2) no superluminal signals, (3) no superluminal causal influences, and (4) no superluminal informational exchange. He shows that the violation does not require violations of the first two constraints but requires violations of the second two.

of causal explanation, such as the causal Markov condition and causal faithfulness. The second approach, exemplified by Oreshkov et al. (2012), Barrett et al. (2019), and Shrapnel (2019), attempts to redefine the concepts of nodes, mechanisms, and interventions in the quantum realm. While the complexities of quantum entanglement require acceptance of counterintuitive results, rejecting fundamental principles can significantly affect our understanding of causal relationships. The thesis critically examines these approaches in Chapter 2, but the primary motivation is to introduce new tools for studying the causal problem of entanglement.

This thesis tackles the causal problem through an inherently empirical method. The goal is to exploit recent advances in Machine Learning to tackle traditional problems in the foundations of Quantum Mechanics. In particular, I exploit Machine Learning to shed new light on the notion of causation in the quantum domain within the framework of causal Bayesian networks.

Machine Learning is a field of study that encompasses a range of algorithms and modeling techniques that allow systems to learn from experience and perform specific tasks without explicit programming. In today's era of big data and highly efficient algorithms, Machine Learning has become widespread and is utilized in a variety of fields, such as self-driving cars, weather forecasting, drug discovery, economics, and social sciences. Despite this widespread usage in empirical and social sciences, Machine Learning has rarely been used in foundational areas.

It is the primary motivation of the present dissertation to use Machine Learning for foundational debates in Quantum Mechanics. The topics explored in this thesis form a triangle, with the three vertices being Machine Learning, Quantum Mechanics, and Causal Modeling. The thesis comprises two separate projects, both utilizing Machine Learning algorithms to discover causal relationships in simulated quantum experiments. The first project is based on the infamous EPR-Bell scenario. The aim is to compare the performance of different models proposed in the literature to solve the causal problem of entanglement. The role of Machine Learning in the first project is to build a unifying arena for empirical judgment between different candidate models. The second project considers scenarios whose underlying causal structures are already known. The goal is to find out the extent to which Machine Learning can identify the true causal structures from statistical data in the absence of explicit interventions. Therefore, the role of Machine Learning in the second project is more like a causal discovery algorithm functioning on quantum systems.

In recent years, causal inference methods based on observational data have developed a lot, mostly thanks to the rapid advances in Machine Learning and their combination with causal inference techniques. My attention is on some of the recent developments that use the power of Machine Learning to learn causal relationships in situations where traditional methods are inapplicable. To be precise, note that in the absence of additional information about a scenario, such as the temporal order of the variables or the functional form of mechanisms, a causal discovery algorithm cannot provide a unique causal graph, and its output contains a number of undirected edges. To address this shortcoming, a class of discovery algorithms, known as "pairwise algorithms", has been developed in recent years, whose focus is the identification of causal direction in bivariate scenarios, i.e., scenarios consisting of only two variables (for overviews see Mooij et al. 2016, Guyon et al. 2019).

The basic idea of pairwise algorithms is that statistical data, in addition to conditional (in)dependencies, contains various types of asymmetries that can be used to detect the causal direction. The asymmetries can include a wide range of statistical properties, such as conditional entropy, mutual information, moments, residuals of regressions, and standard deviation of conditional distributions. Pairwise algorithms can be divided into two categories: generative algorithms, which are based (explicitly or implicitly) on the notion of functional causal models, and discriminative algorithms, which directly determine the causal direction using a classification algorithm. The algorithm discussed in Chapter 4 is generative, while the algorithm discussed in Chapter 5 is discriminative.

In general, pairwise algorithms are primarily due to the efforts of Machine Learning experts interested in causal inference methods and have led to the discovery of new aspects of the theory of causal Bayesian networks. Although such methods can be understood as the continuation of the interventionist account of causation, they have received less attention from philosophers. Although Climenhaga et al. (2021) has recently examined the topics from a philosophical point of view. The current thesis is an attempt to use these recent advances in facing the causal problem of entanglement.

The results of this thesis have two primary contributions to foundational discussions in Quantum Mechanics. The first contribution is to support the idea that obtaining an interventionist account of causation in the quantum domain is possible. While many believe that the interventionist account does not apply to the quantum domain (for example Pearl 2009, Koller & Friedman 2009), some physicists have proposed causal models for quantum systems on the basis of the interventionist account (for example Costa & Shrapnel 2016, Allen et al. 2017). The results of this thesis (primarily Chapter 5) support the latter idea, suggesting that quantum causal relations can be described through generalized notions of the interventionist account. Nonetheless, as will be discussed later, to what extent such a description can be considered *causal* is a fundamental question that cannot be answered within the present results. It is also necessary to emphasize that these results are not meant to suggest that the interventionist account is the only way to characterize quantum causal relations. Other accounts, such as causal process theories of causation (Salmon 1984, Dowe 2000), might also provide such a characterization. However, the latter accounts are not explored in the current thesis.

The second contribution of this thesis is to provide support for the idea that Machine Learning is a reliable tool for exploring new perspectives on the foundations of Quantum Mechanics. For instance, the first project demonstrates how the traditional discussions about the EPR-Bell scenario can be transformed into Machine Learning concepts such as loss function and looked at from other perspectives. Moreover, the second project shows the potential of combining causal modeling with Machine Learning techniques to causally characterize a quantum process even when limited information is available about the process. In general, the results of this thesis highlight the potential of Machine Learning approaches in the foundations of Quantum Mechanics and open up new avenues for further exploration and investigation.

Outline

This dissertation consists of six chapters, including the current one. In what follows, a brief overview of the content of each chapter is provided.

Chapter 2 serves as a foundation for the entire dissertation, encompassing three main topics: Machine Learning, Causal Modeling, and Quantum Mechanics. The Machine Learning section provides an understanding of the basic concepts of supervised learning algorithms and the functioning of artificial neural networks. The Causal Modeling section gives a brief overview of the interventionist account of causation, followed by an explanation of causal Bayesian networks, causal discovery algorithms, and pairwise methods. The Quantum Mechanics section delves deeper into the causal problem of entanglement and its relation to the causal Markov and faithfulness assumptions.

Chapter 3 examines various proposals to address the causal problem of entanglement. The central claim of the chapter is that while it is commonly believed that the causal problem is a dilemma between abandoning either the causal Markov condition or the faithfulness assumptions in the quantum domain, alternative possibilities exist stemming from various physical and philosophical considerations. The chapter aims to study these possibilities, assess their validity, examine the consequences, and explore the interrelationships between them. This chapter lays the philosophical and foundational groundwork for the Machine Learning projects in subsequent chapters.

Chapter 4 offers a novel approach to tackling the causal problem in the context of a continuous formulation of the EPR-Bell scenario. A framework is presented for empirically evaluating different proposals for the causal problem. The framework is based on a classical causal discovery algorithm called Causal Generative Neural Network (Goudet et al. 2018), adapted to the causal problem in light of philosophical considerations. The key idea is to compare the predictive power of various candidate models and find the one that generates data whose distribution is closest to the original quantum data. This framework allows for the empirical comparison of a wide range of proposed models for the causal problem without getting caught up in the Markov-faithfulness dilemma.

Chapter 5 studies quantum scenarios whose causal structures are already known. The cornerstone of the chapter is a pairwise discovery algorithm called Randomized Causation Coefficient (Lopez-Paz et al. 2015). While the said algorithm was originally designed to distinguish cause and effect in classical scenarios, it is generalized in several ways to create a quantum causal discovery algorithm that can handle bivariate and multivariate quantum scenarios. The resulting algorithm performs well in various simulated scenarios and has implications for both quantum foundations and quantum engineering.

Chapter 6 provides a summary of the conclusions drawn from different directions. The chapter covers two main points: (1) the implications of the present results for the interventionist account of causation in the quantum domain, and (2) the relevance of Machine Learning for discussions in the philosophy of physics. Additionally, I offer a brief explanation of my idea on how to link the two problems of entanglement.

Chapter 2

Foundations and Frameworks

This dissertation is an interdisciplinary work between the concepts of several research fields, including Machine Learning, Causal Modeling, and Quantum Mechanics. In order to create a consistent ground throughout the dissertation, I will introduce the definitions, concepts, and notations of the three fields in the present chapter. Nonetheless, each field encompasses a wide range of sub-fields that are not necessarily related to the topics of the present dissertation. Consequently, I will address only those concepts that are truly relevant to the topics of the current thesis.

2.1 Machine Learning

Machine Learning (ML) is a sub-field of Artificial Intelligence that aims to build models that "learn" through experience, where learning from experience is formalized as follows.

> A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E (Mitchell & Mitchell 1997, p. 2).

The term Artificial Intelligence has no strict definition and refers to a broad range of topics, including Machine Learning, natural language processing, computer vision, robotics,

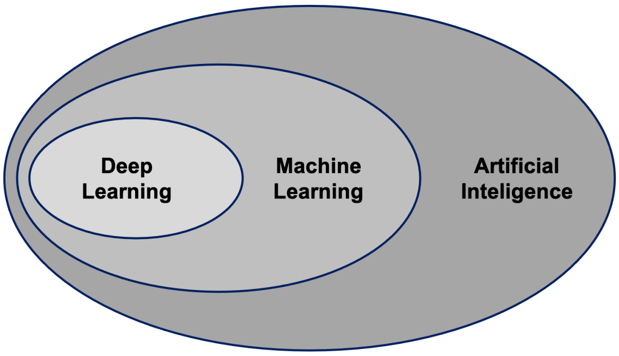


Figure 2.1: Deep Learning is a sub-field of Machine Learning which itself is a sub-field of Artificial Intelligence.

planning, search, game playing, and intelligent agents. In contrast to Artificial Intelligence, Machine Learning is a mathematically well-defined discipline and usually constructs predictive models from data instead of explicitly programming them. Deep Learning is part of a family of Machine Learning methods based on artificial neural networks. Figure 2.1 depicts the relation between Artificial Intelligence, Machine Learning, and Deep Learning.

Suppose we have a dataset $\mathscr{D}$ that contains n observations (or data points). Also, assume that a Machine Learning model is supposed to learn task T from dataset $\mathscr{D}$. Based on the nature of the data $\mathscr{D}$ and the task T, the learning problem can be divided into three categories: (1) supervised, (2) unsupervised, and (3) semi-supervised.

- In a **supervised** problem, $\mathscr{D}$ is a **labeled** dataset, i.e., it contains the values of both input variables $\mathbf{X}$ and output variables $\mathbf{Y}$. The goal is to learn the functional relation that maps variables $\mathbf{X}$ to variables $\mathbf{Y}$.

- In an **unsupervised** problem, $\mathscr{D}$ is **unlabeled**, i.e., it contains merely the values of $\mathbf{X}$. The goal is to learn the inherent structure of $\mathbf{X}$ without being concerned about $\mathbf{Y}$.

- In a **semi-supervised** problem, a (small) part of $\mathscr{D}$ is labeled while the rest is unlabeled. Similar to a supervised problem, the goal is to learn the functional relation that maps $\mathbf{X}$ to $\mathbf{Y}$. However, due to accessing a large number of unlabeled observations, a combination of supervised and unsupervised techniques is used.

Remark 2.1. *Different research fields use different names to refer to input and output variables. For instance, the input variables* $\mathbf{X}$ *are known as* ***features****, covariates, independent, exogenous, or explanatory variables. Similarly, the output variables* $\mathbf{Y}$ *are known as* ***targets****, outcomes, dependent, endogenous, or response variables.*

Depending on the type of output variable(s), a supervised learning problem is either a classification or regression problem.

- In a **classification** problem, the output(s) is a categorical variable whose values are expressed by a limited number of classes or labels. Classification problems with two and more than two classes are known as "binary" and "multiclass" classification problems, respectively. Identification of covid-19 from explanatory variables (e.g., chest X-Ray and CT scan images) is an example of a binary classification.

- In a **regression** problem, the output(s) is a numerical variable that takes a continuous range of values. Predicting the rental price of an apartment in Munich from explanatory variables (e.g., living area, year of construction, and location) is an example of a regression problem.

The present dissertation mostly deals with supervised learning tasks. Thus, in what follows, I assume that dataset $\mathscr{D}$ is labeled and is represented as

$$\mathscr{D} = \big((\mathbf{X}^{(1)}, \mathbf{Y}^{(1)}), \dots, (\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}) \big) \in (\mathcal{X} \times \mathcal{Y})^n, \tag{2.1}$$

where the tuple $(\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})$ denotes the i^{th} observation in the dataset, and $\mathcal{X}$ and $\mathcal{Y}$ respectively represent the input and output spaces. The goal is to learn a model that automatically maps the input space to the output space. To this end, a technique called "data splitting" is usually utilized. **Data splitting** is a well-known technique for training and evaluating Machine Learning models. The idea is to divide the original dataset into different parts and use each for a different purpose. Commonly, the dataset is divided into three disjoint subsets, i.e., training, validation, and test:

$$\mathscr{D} = (\mathscr{D}_{tr}, \mathscr{D}_{va}, \mathscr{D}_{te}). \tag{2.2}$$

The training data is used to fit the model, i.e., optimizing the model parameters. The validation data is not directly used for training but for monitoring the training quality to avoid unwanted issues such as underfitting and overfitting (explained below). The validation data is also used for tuning the model hyperparameters (explained below). The test data is used to evaluate the final performance of the model after being trained. It is also known as **unseen** data because the model cannot access it during training. The data checks how well the trained model performs or generalizes on unseen data.

To learn a model $f : \mathcal{X} \rightarrow \mathcal{Y}$, a learning algorithm usually makes the assumption that all observations in $\mathscr{D}$ are generated by an underlying **data-generating process** that is characterized by a joint probability distribution $\mathbb{P}_{\mathbf{XY}}$ on $\mathcal{X} \times \mathcal{Y}$. Because the true $\mathbb{P}_{\mathbf{XY}}$ is unknown to the algorithm, it attempts to approximate the distribution structure partially. While there are different approaches to achieving the said approximation, most supervised algorithms have three components based on which their procedure can be described: (1) hypothesis space, (2) risk, and (3) optimization.

Hypothesis Space

The hypothesis space is the space of all models a learning algorithm can use to predict the output(s) from the inputs. That is, the hypothesis space declares what kinds of models are allowed to be learned from the data. For instance, linear functions and neural networks lead to different hypothesis spaces based on which their corresponding algorithms solve a learning problem.

Note that a model $f : \mathcal{X} \rightarrow \mathcal{Y}$ induced by a learning algorithm includes a set of parameters that control the model's behavior. The said parameters are divided into learnable and non-learnable ones. The former refers to the parameters that can be inferred from the training data during the learning process. The latter refers to the parameters that cannot be directly inferred from the training data, and the user must usually determine their values before starting the learning process. Conventionally, the former is called model **parameters**, while the latter is called model **hyperparameters**. I denote these by $\mathbf{\Theta}$ and $\mathbf{\Omega}$, respectively.

The hypothesis space restricts the form of parameters and hyperparameters that models can exploit to map the input space to the output space. A too-simple hypothesis space induces weak models and can lead to **underfitting**: a situation where a model is so weak that it cannot learn the patterns of the training data and hence neither fits the training data nor generalizes to test data. On the contrary, a complex hypothesis space induces strong models and can lead to **overfitting**: a situation where a model is so strong that it learns not only the training data but also captures the statistical fluctuations as actual features. In such a case, the model achieves a very well performance on the seen data but a poor performance on the unseen data.

Therefore, the complexity of the hypothesis space must be determined proportionally to the complexity of the data under consideration. Moreover, to control the behavior of models, a tuning strategy is required, with which the values of model hyperparameters are determined. The strategy is called **hyperparameter tuning** and is usually based on the validation data.

Risk & Loss

The notion of risk is used to quantify the quality of a specific model on a given dataset. It is tied to the notion of the loss function.

Consider the model $f : \mathcal{X} \to \mathcal{Y}$ again. The model takes a true input $\mathbf{X}^{(i)}$ and returns an estimated output $\hat{\mathbf{Y}}^{(i)} = f(\mathbf{X}^{(i)})$. The loss function quantifies how well (more accurately, how badly) $\hat{\mathbf{Y}}^{(i)}$ approximates $\mathbf{Y}^{(i)}$. A **loss function** takes the true and estimated outputs and returns a real (typically, non-negative) number representing the quality of the estimation. The smaller the values of $L(\mathbf{Y}^{(i)}, \hat{\mathbf{Y}}^{(i)})$, the better $\hat{\mathbf{Y}}^{(i)}$ approximates $\mathbf{Y}^{(i)}$. The average of the loss values over all observations in the given dataset is known as the

empirical risk:

$$\mathscr{L}(f) = \sum_{i=1}^{n} L\left(\mathbf{Y}^{(i)}, f\left(\mathbf{X}^{(i)}\right)\right) \tag{2.3}$$

Depending on the type of supervised problem (i.e., classification or regression), different loss functions must be used. The cross-entropy loss is a standard loss for classification problems. Assuming that there is only one output variable, i.e., the output is a one-dimensional random variable, the cross-entropy loss is computed as:

$$L(\hat{\mathbf{Y}}^{(i)}, \mathbf{Y}^{(i)}) = -\sum_{j=1}^{m} \delta_{ij} \log\left(\widehat{p}_{ij}\right) \tag{2.4}$$

where m denotes the number of classification classes, and $\delta_{ij} = 1$ is if $\mathbf{Y}_i = j$ and 0 otherwise. $\widehat{p}_{ij}$ is the probability predicted by the model for the given observation, i.e., $\widehat{p}_{ij} = p(\mathbf{Y}^{(i)} = j|\mathbf{X}^{(i)})$. The L_p losses are standard losses for regression problems. These losses computed the p-norm of the difference between the true and estimated outputs:

$$L(\hat{\mathbf{Y}}^{(i)}, \mathbf{Y}^{(i)}) = \left\|\hat{\mathbf{Y}}^{(i)} - \mathbf{Y}^{(i)}\right\|_p. \tag{2.5}$$

Optimization

It is the goal of a supervised learning algorithm to find the model that minimizes the empirical risk defined in Equation 2.3. Put differently, the algorithm searches for a vector of optimal parameters $\mathbf{\Theta}$, leading to a model with the lowest empirical risk. This is accomplished in the training phase of the algorithm, wherein the model parameters $\mathbf{\Theta}$ are iteratively adjusted in the light of the training data.

A common optimization scheme used for this purpose is **gradient descent**, an iterative algorithm that finds local optima of differentiable functions. The said optimizer iteratively updates $\mathbf{\Theta}$ in the direction of the gradient of the empirical risk computed on the training data:

$$\mathbf{\Theta}^{t+1} = \mathbf{\Theta}^{t} - \eta \nabla_{\mathbf{\Theta}} \mathscr{L}_{tr}, \tag{2.6}$$

where t denotes the iteration index and η denotes the **learning rate** of the optimizer. In practice, more sophisticated methods such as **stochastic gradient descent (SGD)** and **Adaptive Moment Estimation (Adam)** are used. Moreover, instead of computing the full empirical risk, the contribution of a few observations is computed and optimized in each iteration.

The training phase can be repeated as many times as the user specifies, i.e., the number of training iterations is an example of a model hyperparameter. To avoid unwanted issues such as underfitting and overfitting, one should either (1) find a suitable value for this hyperparameter before starting the algorithm or (2) exploit techniques such as early stopping. **Early stopping** refers to a set of regularization techniques used in many ML algorithms to avoid overfitting (Prechelt 1998). In a nutshell, an early-stopping algorithm continuously monitors the quality of the Learning and terminates the training phase if the model stops being optimized. How and when to terminate the training phase depends on how the early-stopping algorithm is defined. It is common to check the model performance on the validation dataset and adapt the algorithm through adjustable hyperparameters such as **patience**, i.e., the number of iterations the algorithm waits before terminating the training phase.

2.1.1 Deep Learning

Deep Learning refers to a particular type of Machine Learning algorithm wherein neural networks are used as the building blocks of the hypothesis space. Deep Learning has progressed rapidly in recent years, primarily thanks to the high flexibility of neural networks. Neural networks are applied to a variety of data types such as image data (Shen et al. 2017), video data (Ngiam et al. 2011), text data (Kowsari et al. 2017), speech data (Deng et al. 2013), or tabular data (Gorishniy et al. 2021).

Neural networks can be used for both types of supervised learning problems, i.e., classification and regression. In this thesis, I use both forms. Depending on their architecture, neural networks are divided into **feed-forward** and **recurrent**. While in the former, there is no cycle in the network, the connections in the latter create cycles. In this thesis, I only use feed-forward networks. In particular, I only use **multilayer perceptrons**, a feed-forward neural network that is fully connected. In the following, whenever I use the term neural networks, I mean multilayer perceptrons.

A neural network consists of at least three layers: an input layer, an output layer, and one or more **hidden layers**. The input variables $\mathbf{X}$ are connected to the input layer, and the output variables $\mathbf{Y}$ are connected to the output layer. The whole network acts as a learning model $f : \mathcal{X} \rightarrow \mathcal{Y}$. The more the number of hidden layers, the more the complexity of the hypothesis space and the **deeper** the neural network.

Each hidden layer consists of a number of **neurons** or **hidden units**. Each neuron has a set of learnable parameters called weights and bias. The mathematical operations performed inside each neuron consist of two steps: an affine transformation followed by a non-linear transformation. Thus, when the values $x_1, \dots, x_p$ are given to the input of a neuron, the neuron returns the following output:

$$\hat{y} = \sigma\left(\sum_{i=1}^{p} w_i x_i + b\right) \tag{2.7}$$

where $w_i, b \in \mathbb{R}$ denote respectively the **weights** and the **bias** of the neuron, and σ denotes the **activation function**. The activation function is responsible for the non-linearity of neural networks. Two widely used activation functions are the ReLU function (Rectified Linear Unit: $\sigma(x) = \max(0, x)$) and the sigmoid function ($\sigma(x) = \frac{\exp(x)}{1+\exp(x)}$).

Figure 2.2 illustrates a neural network with several hidden layers. To calculate the outputs based on the network's inputs, the outputs of the first hidden layer are calculated based on the values of the input layer. Then, the outputs of the second hidden layer are calculated based on the outputs of the first hidden layer. This procedure is repeated until the last layer is reached. Therefore, the relation between the inputs and the outputs of the network in Figure 2.2 is

$$\begin{cases} f_i(\mathbf{Z}) = \left(\sigma_i\left(\mathbf{Z} \cdot \boldsymbol{W}^i + b_i\right)\right) \\ \widehat{\mathbf{Y}} = f_k \circ f_{k-1} \circ \dots \circ f_1(\mathbf{X}) \end{cases} \tag{2.8}$$

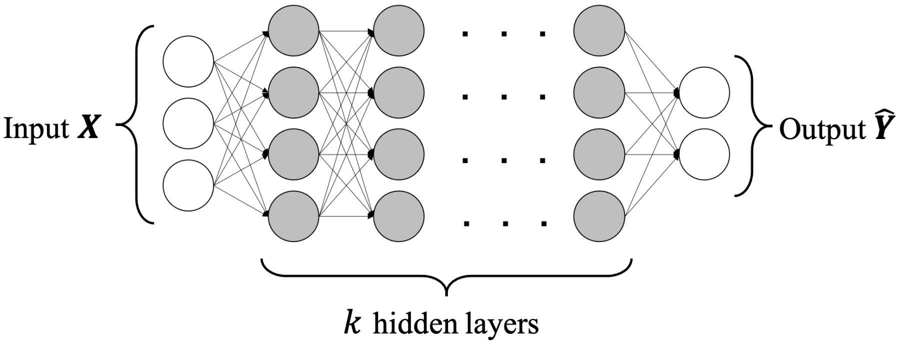


Figure 2.2: Illustration of a multilayer perceptron. Each edge represents a weight $\boldsymbol{W}^i[j, l]$ and each node of a hidden layer processes its inputs based on Equation 2.7.

where the $\boldsymbol{W}^i$ denotes the weight matrix corresponding to the neurons in the i^{th} hidden layer, i.e., $\boldsymbol{W}^i[j, :]$ represents the weight vector of the j^{th} neuron in the i^{th} layer.

To train a neural network is to optimize its weights and biases in light of the training data. Like other Machine Learning algorithms, the optimization is achieved iteratively. In each iteration or **training epoch**, the loss function is evaluated by comparing $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$. Then, the loss gradient is calculated with respect to the model parameters (i.e., weights and biases). Finally, the model parameters are updated in the direction of the loss gradient.

Optimizing the parameters in neural networks requires lots of differentiation calculations. Standard deep learning frameworks such as PyTorch and TensorFlow perform these calculations automatically through the **backpropagation** algorithm.

Another aspect of the training procedure in neural networks is when $\mathscr{D}_{tr}$ is very large (i.e., it contains a large number of observations) or when the computations of the loss function are very complicated. The **mini-batch training** strategy is used in such situations. In this strategy, in each training epoch, the $\mathscr{D}_{tr}$ is randomly broken into several smaller parts (or batches), and the optimization algorithm is performed on each batch. The strategy is used to avoid issues related to the lack of memory in computer systems. Meanwhile, the number of batches is a hyperparameter of the algorithm that the user must determine.

2.1.2 Kernel Embedding

Many Machine Learning algorithms are based on the analysis of probability distributions of the variables involved in a scenario. Since there is usually no access to original distributions, these algorithms have to estimate the distribution as an initial step and then start the analysis. In practice, such a strategy becomes almost impossible for high-dimensional distributions (i.e., when the number of variables in a scenario is large).

Methods based on the kernel embedding of distributions overcome this problem by mapping distributions into an intermediate space, known as a Reproducing Kernel Hilbert Space (RKHS), so that calculations on distributions can be performed on the image of the distributions in the RKHS. The relevant framework for doing so is known as the **kernel mean embedding** framework. Due to its unique features, the framework is widely used in causal inference methods based on Machine Learning. Both algorithms I investigate in my dissertation are based on the said framework; thus, I review the framework's principles in this section.

Let $\mathbf{X} \in \mathcal{X}$ be a possibly multivariate random variable defined on the space $\mathcal{X}$ with the probability distribution $\mathbb{P}$.

Definition 2.1 (**Kernel Function**). *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function, or kernel, if it is symmetric, i.e., $k(\mathbf{X}_i, \mathbf{X}_j) = k(\mathbf{X}_j, \mathbf{X}_i)$, and positive definite, i.e., the following holds for any $n \in \mathbb{N}$, any choice of $\mathbf{X}_1, \ldots \mathbf{X}_n \in \mathcal{X}$, and any $c_1, \ldots, c_n \in \mathbb{R}$:*

$$\sum_{i,j=1}^{n} c_i c_j k\left(\mathbf{X}_i, \mathbf{X}_j\right) \geq 0. \tag{2.9}$$

According to a theorem proved in Aronszajn (1950), associated with each kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ there exists a unique **feature space** $\mathcal{H}_k$ and a unique **feature map** $\phi_k : \mathcal{X} \to \mathcal{H}_k$ such that

$$k\left(\mathbf{X}, \mathbf{X}'\right) = \langle \phi_k(\mathbf{X}), \phi_k\left(\mathbf{X}'\right) \rangle_{\mathcal{H}_k} \tag{2.10}$$

The unique feature space is known as a **reproducing kernel Hilbert space (RKHS)**,

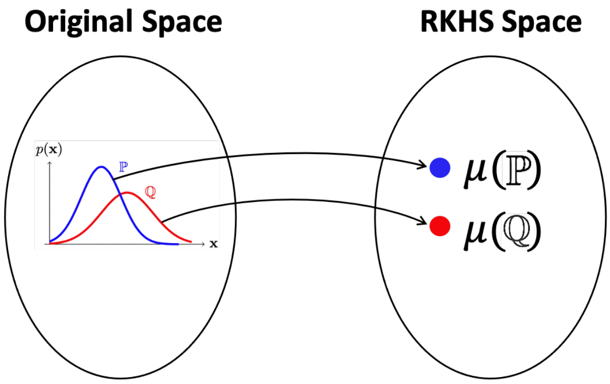


Figure 2.3: Embedding of marginal distributions: each distribution is mapped into a reproducing kernel Hilbert space via an expectation operation.

a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ equipped with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and norms $\|\cdot\|_{\mathcal{H}_k}$ in which the element $k(\mathbf{X}, \cdot)$ satisfies the **reproducing property** $\langle f, k(\mathbf{X}, \cdot) \rangle_{\mathcal{H}_k} = f(\mathbf{X})$ for any $f \in \mathcal{H}_k$ and any $\mathbf{X} \in \mathcal{X}$.

To see the importance of the above result, notice that an inner product between two vectors can be construed as a measure of similarity between the vectors. In the present context, a vector represents a data point in a high-dimensional space such as the original space $\mathcal{X}$ and the feature space $\mathcal{H}_k$. The right-hand side of Equation 2.10 asserts that computing the similarity measure between vectors in the feature space (i.e., $\phi_k(\mathbf{X})$ and $\phi_k(\mathbf{X}'))$ requires a linear calculation. However, the left-hand side of Equation 2.10 represents a non-linear calculation between the vectors in the original space. Therefore, Equation 2.10 implies that one can convert highly-complex non-linear calculations in the original space into simple inner product calculations in the feature space.

The theorem reveals its power when applied to probability distributions rather than single data points. In such a case, distributions are mapped into the RKHS, and complex calculations on the distributions, such as similarities between distributions, are performed efficiently on the image of the distributions in the RKHS (Muandet et al. 2017). The kernel mean embedding framework represents a distribution $\mathbb{P}$ as the following mean function (see Figure 2.3):

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} k(\mathbf{X}, \cdot) \mathrm{d}\mathbb{P}(\mathbf{X}). \tag{2.11}$$

There are two essential points about the above method. First, if the embedding $\mu_k : \mathbb{P} \rightarrow \mu_k(\mathbb{P})$ is an injective map, then no information is lost when mapping distributions into the RKHS. Such a property is realized in **characteristic** kernel functions. Put differently, if a characteristic kernel is used, the embedding preserves all information about a distribution while allowing the computations to be performed on the RKHS.

Second, for a **shift-invariant** kernel function, i.e., $k(\mathbf{X}, \mathbf{Y}) = \psi(\mathbf{X} - \mathbf{Y})$, it is possible to obtain a low-dimensional representation of distributions using the normalized Fourier transform of the kernel. Given a high-dimensional distribution, it is possible to extract a finite number of features from the distribution such that the extracted features represent the most important properties of the distribution. Thus, shift-invariant kernel functions allow one to represent a high-dimensional distribution through a 1-dimensional vector of features.

The **Gaussian** and **Laplace** kernels are examples of kernel functions that are characteristic and shift-invariant. Given a parameter $\sigma > 0$, which is known as bandwidth, the two kernels are defined as

$$k\left(\mathbf{X}, \mathbf{X}'\right) = \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}'\|_2^2}{2\sigma^2}\right), \quad k\left(\mathbf{X}, \mathbf{X}'\right) = \exp\left(-\frac{\|\mathbf{X} - \mathbf{X}'\|_1}{\sigma}\right). \tag{2.12}$$

The two points mentioned above are the cornerstones of the Machine Learning algorithms I investigate in Chapters 4 and 5. I utilize the Gaussian kernel function in both of the chapters.

2.2 Causal Modeling

This dissertation centers on the interventionist account of causation, specifically the theory of causal Bayesian networks and its compatibility with Quantum Mechanics. It explores recent developments in ML-based causal discovery algorithms to learn causal relationships in scenarios where traditional methods are insufficient. This section outlines the key elements relevant to the main topic of the thesis. The section begins with a discussion of the assumptions of causal Bayesian networks that are believed to be at the heart of the conflict with quantum correlations. This is followed by an examination of traditional causal discovery algorithms and inference techniques based on ML.

2.2.1 Framework

Let $\mathbf{V} = \{X_1, \ldots, X_p\}$ be a set of 1-dimensional random variables. Assume dataset $\mathscr{D}$ contains n observations from the variables' values in $\mathbf{V}$. The dataset takes the following form, wherein $\mathbf{V}^{(i)}$ represents the variables' values in the i^{th} observation:

$$\mathscr{D} = \left(\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(n)}\right) \quad \text{with} \quad \mathbf{V}^{(i)} = \left(x_1^{(i)}, \ldots x_p^{(i)}\right) \tag{2.13}$$

Associated with these variables, a joint probability distribution $\mathbb{P}$ with density p can be defined from which the observations in $\mathscr{D}$ are drawn randomly. Assume that some variables in $\mathbf{V}$ are statistically independent while others reveal statistical dependencies. Besides the statistical relationships, assume a set of causal relationships among the variables in $\mathbf{V}$ that might be exploited to *explain* the observed statistics causally. Two central questions arise here:

- Given the observed statistics, i.e., the dataset $\mathscr{D}$, how can one discover the underlying causal relationships?
- Given the causal relationships, what sorts of statistical (in)dependencies are expected

to be observed?

The causal modeling framework provides tools for addressing these questions and enables one to relate causal relations to statistical relations. In particular, the framework provides inference schemes known as **causal discovery algorithms**, which enable one to infer the causal relationships between variables from their statistical (in)dependencies.

The primary idea is to represent the causal relationships among the variables $\mathbf{V}$ through a graph $\mathcal{G}$. The nodes of such a graph denote the variables in $\mathbf{V}$, and the edges represent the causal directions among the variables, i.e., for a causal relation $X_i \rightarrow X_j$, an edge is sketched from X_i to X_j.

Remark 2.2. *Because of the correspondence between the "variables" and the "nodes," I use the two terms interchangeably. Moreover, I label each node with the name of the corresponding variable.*

In the presence of complete knowledge about the causal directions between the variables, the causal structure is represented by a directed acyclic graph (DAG). A **directed acyclic graph (DAG)** $\mathcal{G}$ is a graph whose all edges are directed, with no cycle among the nodes. $\mathcal{G}$ being acyclic means that no variable can (directly or indirectly) affect itself, while $\mathcal{G}$ being directed means that all causal relationships within $\mathcal{G}$ have a determined direction. Some basic terminology is needed to continue the discussion.

If there is a directed edge from X_i to X_j, the former is called the **parent** of the latter, and the latter is called the **child** of the former. Similarly, if there are edges like $X_i \rightarrow X_j \rightarrow X_k$, then X_i is a **ancestor** of X_k and X_k is a **descendant** of X_i. In the following, I denote the set of all parents and non-descendants for the variable X_i by $\mathbf{PA}_i^{\mathcal{G}}$ and $\mathbf{ND}_i^{\mathcal{G}}$, respectively. Moreover, the values taken by the former variables are noted by $\mathbf{pa}_i^{\mathcal{G}}$ and $\mathbf{nd}_i^{\mathcal{G}}$, respectively.

A node with no incoming edge is called **exogenous**, and a node with at least one incoming edge is called **endogenous**. Exogeneity of a node means that it is not under the causal influence of any other variable within that graph. A graphical structure $X_i \rightarrow X_k \leftarrow X_j$ is called **collider** (or v-structure), while graphical structures $X_i \rightarrow X_k \rightarrow X_j$ and

$X_i \leftarrow X_k \rightarrow X_j$ are called **chain** and **fork**, respectively. The **skeleton** of a DAG is the undirected graph obtained by substituting all the directed edges with undirected ones.

The next topic to be discussed is the notion of statistical (in)dependence. In statistics, two variables X and Y are said to be **marginally independent** if and only if $p(x,y) = p(x)p(y)$ for all x, y. That is, their joint distribution factorizes as the product of two marginal distributions. Moreover, two variables X and Y are said to be **conditionally independent** given Z if and only if $p(x,y|z) = p(x|z)p(y|z)$ for all x, y, z such that $p(z) > 0$. That is, their joint distribution factorizes as the product of two conditional distributions. I denote the sketched marginal and conditional independencies by $X \perp\!\!\!\perp_{\mathbf{s}} Y$ and $X \perp\!\!\!\perp_{\mathbf{s}} Y|Z$, where the subscript s is chosen to emphasize the "statistical" nature of these relations. Moreover, I use the notations $X \not\perp\!\!\!\perp_{\mathbf{s}} Y$ and $X \not\perp\!\!\!\perp_{\mathbf{s}} Y|Z$ to denote the dependencies corresponding to the above relations.

Remark 2.3. *Statistical (in)dependence can be generalized to random vectors (i.e., multivariate random variables). For instance, random vectors* $\mathbf{X} = [X_1, \ldots, X_n]^\top$ *and* $\mathbf{Y} = [Y_1, \ldots, Y_m]^\top$ *are marginally independent, denoted by* $\mathbf{X} \perp\!\!\!\perp_{\mathbf{s}} \mathbf{Y}$*, if and only if there is no dependence between the elements of* $\mathbf{X}$ *and the elements of* $\mathbf{Y}$*:*

$$p(x_1, \ldots, x_n, y_1, \ldots, y_m) = p(x_1, \ldots, x_n)p(y_1, \ldots, y_m) \quad \forall x_i, y_j$$

The notion of statistical (in)dependence can be reflected in causal graphs through the so-called "d-separation criterion" introduced by Pearl (1988). The d-separation criterion is a graphical tool for identifying causally independent nodes within a causal graph. By combining the criterion with the causal Markov condition (defined later), one can extract all the conditional independencies implied by the graph.

Definition 2.2 (d-separation). *In graph* $\mathcal{G}$*, two disjoint subsets of nodes* $\mathbf{X}$ *and* $\mathbf{Y}$ *are said to be* ***d-separated*** *by a third disjoint subset* $\mathbf{Z}$*, denoted* $\mathbf{X} \perp\!\!\!\perp_{\mathbf{c}} \mathbf{Y}|\mathbf{Z}$*, if for every path between nodes in* $\mathbf{X}$ *and* $\mathbf{Y}$ *there exists a node* Z *such that (1)* Z *is either a "chain"* $(\mathbf{X} \rightarrow Z \rightarrow \mathbf{Y})$ *or a "fork"* $(\mathbf{X} \leftarrow Z \rightarrow \mathbf{Y})$*, and* $Z \in \mathbf{Z}$ *, or (2)* Z *is a "collider"* $(\mathbf{X} \rightarrow Z \leftarrow \mathbf{Y})$*,*

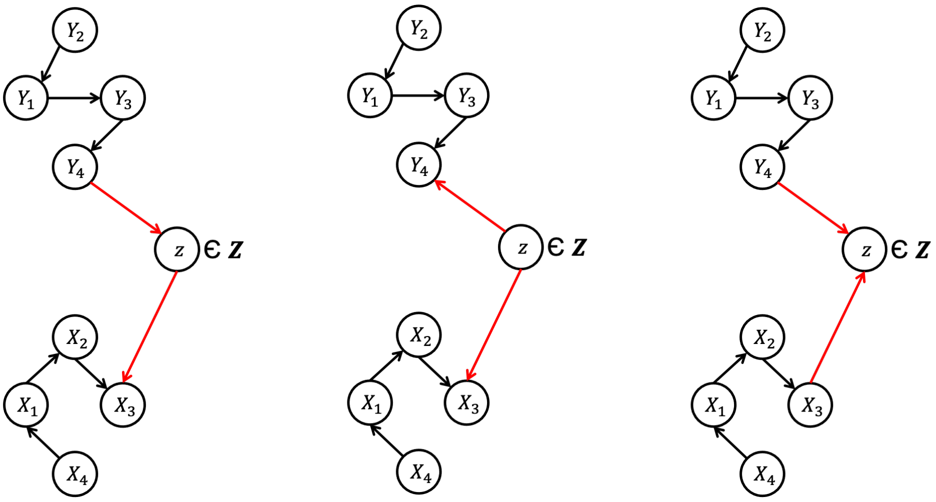


Figure 2.4: The d-separation criterion.

and neither Z nor any of its descendants are in $\mathbf{Z}$.

The basic idea behind the d-separation criterion is that a suitable set of variables $\mathbf{Z}$ can "block" the flow of causal influences between the variables within $\mathbf{X}$ and $\mathbf{Y}$ (see Figure 2.4). Therefore, the criterion provides a way for a causal graph to extract causal (in)dependencies in complex systems.

Definition 2.3 (**Causal Sufficiency**). *The set of variables $\mathbf{V}$ is said to be causally sufficient if it includes all of the common causes of pairs in $\mathbf{V}$. That is, for all pairs $X_i, X_j \in \mathbf{V}$, it holds that there no external common cause $X_k \notin \mathbf{V}$.*

With the basic definitions established, let us revisit the dataset $\mathscr{D}$ in Equation 2.2.1. As mentioned in the Machine Learning section, most supervised algorithms presume the existence of an underlying data-generating process, which is characterized by a distribution $\mathbb{P}$, and assume that data $\mathscr{D}$ is randomly sampled from $\mathbb{P}$. Graphical causal models adopt a similar strategy to supervised learning algorithms, but in addition, they impose further constraints on the underlying generating process, which arise from the causal relationships between the variables. This is a fundamental distinction between the two: while Machine Learning models may not necessarily enforce constraints on the form of the underlying generating process, causal models are designed to be sensitive to causal information. The causal Markov condition is an example of such additional constraints on the distribution $\mathbb{P}$.

Definition 2.4 (Causal Markov Condition (CMC)). *The causal Markov condition (CMC) can be expressed in various forms:*

- Factorization: *the joint distribution $\mathbb{P}$ factorizes as:*

$$p\left(x_1, \ldots, x_p\right) = \prod_{i=1}^{p} p(x_i | \mathbf{pa}_i^{\mathcal{G}}) \tag{2.14}$$

- Screen-off: *each variable is independent of its non-descendants given its parents:*

$$(X_i \perp\!\!\!\perp_{\mathbf{s}} \mathbf{ND}_i^{\mathcal{G}} \mid \mathbf{PA}_i^{\mathcal{G}}) \quad \forall i = 1, \ldots, p \tag{2.15}$$

- Explanation & Discovery: *causal independencies imply statistical independencies; equivalently, statistical dependencies imply causal dependencies:*

$$(\mathbf{X} \perp\!\!\!\perp_{\mathbf{c}} \mathbf{Y} | \mathbf{Z}) \Rightarrow (\mathbf{X} \perp\!\!\!\perp_{\mathbf{s}} \mathbf{Y} | \mathbf{Z}) \;\equiv\; (\mathbf{X} \not\perp\!\!\!\perp_{\mathbf{s}} \mathbf{Y} | \mathbf{Z}) \Rightarrow (\mathbf{X} \not\perp\!\!\!\perp_{\mathbf{c}} \mathbf{Y} | \mathbf{Z}) \tag{2.16}$$

There are several points about the CMC, which I will discuss below. First, if the distribution $\mathbb{P}$ has a density p, it can be proven that the three forms mentioned above are equivalent (see, e.g., Lauritzen 1996).

Second, each of the three mentioned forms has alternative names. The factorization form is sometimes expressed as the "Markov property" of the distribution $\mathbb{P}$ relative to the graph $\mathcal{G}$. Other names for the screen-off form are the "local Markov property" and the "parental Markov condition." The explanation form ($\perp\!\!\!\perp_{\mathbf{c}} \Rightarrow \perp\!\!\!\perp_{\mathbf{s}}$) is also called the "d-separation form." Note that the discovery form (i.e., $\not\perp\!\!\!\perp_{\mathbf{s}} \Rightarrow \not\perp\!\!\!\perp_{\mathbf{c}}$) is logically equivalent to the explanation form. Following Näger (2016), the explanation and discovery form more clearly demonstrate the content of the CMC in the context of causal discovery algorithms. In future chapters, I often mean the explanation and discovery form when I refer to the CMC.

Third, the screen-off form can be seen as an extrapolation of Reichenbach's principle common cause (Reichenbach 1956) to directed acyclic graphs. The latter states that if two variables X_1 and X_2 are statistically dependent but there is no direct causation between them, there must be a common cause or confounder named X_3 that, if taken into account, screens off the statistical dependence between them, i.e., $X_1 \perp\!\!\!\perp_{\mathbf{s}} X_2 | X_3$ or $p(x_1, x_2|x_3) = p(x_1|x_3)p(x_2|x_3)$. In other words, the third variable explains their statistical dependence. While Reichenbach's principle is restricted to a common cause scenario, the CMC generalizes this idea to more complex causal structures.

Fourth, one of the presumptions of the CMC is that the set $\mathbf{V}$ is causally sufficient, meaning it includes all the common causes relevant to the pairs in $\mathbf{V}$. If this assumption is not satisfied (i.e., some latent common causes exist), the distribution $\mathbb{P}$ does not satisfy the Markov condition relative to the graph $\mathcal{G}$. However, as soon as the role of these latent common causes is taken into account, the distribution $\mathbb{P}$ satisfies the CMC again.

As previously mentioned, the CMC is an additional constraint on the underlying data-generating process. The importance of the CMC is that it can connect causal statements to statistical ones. The CMC can be understood as the claim that the causal parent of each variable contains all the relevant information to describe the distribution of that variable. Therefore, by considering the role of causal parents, the joint distribution factorizes into distinct conditional terms.

In a broader sense, the CMC ensures the possibility of performing so-called surgical interventions on the variables of a scenario. In a graphical representation, surgical intervention on node X_i means that we cut all incoming edges of this variable and fix the value of X_i through an intervention. Such intervention only changes the values of variables that are causal descendants of X_i. Therefore, the CMC helps us find the causal structure between variables through the concept of intervention.

Pearl expresses the notion of intervention through the primitive notion of causal mechanisms. In this expression, the values of variables are determined through deterministic functions such as $x_i = f_i(\mathbf{pa}^{\mathcal{G}}_{,} E_i)$, where the function f_i represents the functional form of the causal mechanism associated with X_i, and E_i summarizes the impact of all excluded variables. For Pearl, deterministic relationships are behind the scenes of statistical relationships, and randomness in the data, expressed through the variables E_i, arises from our epistemic ignorance.

> I take causal relationships to be the fundamental building blocks both of physical reality and of human understanding of that reality, and I regard probabilistic relationships as but the surface phenomena of the causal machinery that underlies and propels our understanding of the world (Pearl 2000, xvi).

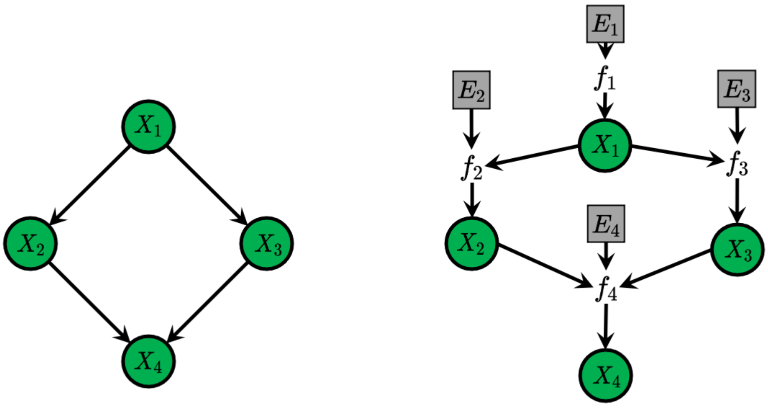
In this view, an intervention on a variable means changing the causal mechanism of that variable to a new causal mechanism without changing the other mechanisms. Therefore, an intervention on X_i can change the function f_i to a constant value or replace it with a new function such as g_i. Such an intervention requires an assumption of independence between the causal mechanisms of different variables, usually expressed under the name of **autonomy of causal mechanisms** or **independence of causal mechanisms**. In Chapter 3, I will talk more about this assumption.

Remark 2.4. *Woodward provides a more advanced characterization of the notion of intervention. Since the focus of this thesis is on discovery methods based on observational data, I skip this characterization and refer the reader to Woodward (2007, p. 75).*

Assuming that the underlying data-generating process arises from a set of autonomous mechanisms allows one to precisely define a functional causal model associated with the causal structure of a scenario.

Definition 2.5 (**Functional Causal Model (FCM)**). *A functional causal model, also known as a structural equation model, defined over a set of variables* $\mathbf{V}$ *is a triplet* $\mathscr{F} = (\mathcal{G}, \mathbf{f}, \mathbb{P}_{\mathbf{E}})$*, where* $\mathcal{G}$ *is a causal graph over* $\mathbf{V}$*,* $\mathbf{f} = \{f_1, \dots, f_p\}$ *is a set of functions representing the causal mechanisms among the variables,* $\mathbb{P}_{\mathbf{E}}$ *is a (product) joint distribution over independent noise terms* $\mathbf{E} = \{E_1 \dots E_n\}$*, such that:*

$$X_i := f_i\left(E_i, \mathbf{PA}_i^{\mathcal{G}}\right), \quad \forall i = 1, \dots, p \tag{2.17}$$



$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(E_2, X_1) \\ X_3 = f_3(E_3, X_1) \\ X_4 = f_4(E_4, X_2, X_3) \end{cases}$$

Figure 2.5: Example of a functional causal model: the graph nodes X_i represent the original observed variables, whereas f_i and E_i represent the causal mechanisms and noise terms postulated by the FCM. Left: original causal graph, middle: schematic representation of the corresponding FCM, and right: structural equations postulated by the FCM.

In an FCM, each variable is assumed to be a deterministic function of its causes and some noise terms, where the noises are assumed to be jointly independent. Figure 2.5 illustrates how an FCM models the underlying causal mechanisms in a scenario. The functions in this figure represent the causal mechanisms generating the variables. From Pearl's point of view, such functions are deterministic, while the noise terms are taken to be random, arising from our ignorance about other factors influencing the variables.

Remark 2.5. *While Pearl's view on causal mechanisms is deterministic, functional causal models can be used to model indeterministic scenarios, such as those found in Quantum Mechanics. In such cases, the randomness associated with noise terms should be interpreted as irreducible randomness inherent in the scenario rather than as epistemic randomness arising from the observer's lack of knowledge. This approach will be employed in Chapter 4.*

Another assumption commonly made within the framework of causal Bayesian networks is the causal Faithfulness condition.

Definition 2.6 (**Causal Faithfulness Condition (CFC)**)**.** *The causal faithfulness condition (CFC) can be expressed as either of the following equivalent forms:*

- No Fine-Tuning: *All statistical independencies induced by a causal model must originate from the d-separation relations holding in $\mathcal{G}$, and not from the particular values for the model parameters (Wood & Spekkens 2015).*

- Explanation and Discovery: *Causal dependencies imply statistical dependencies; equivalently, statistical independencies imply causal independencies (Näger 2016):*

$$(\mathbf{X} \not\perp\!\!\!\perp_{\mathbf{c}} \mathbf{Y}|\mathbf{Z}) \Rightarrow (\mathbf{X} \not\perp\!\!\!\perp_{\mathbf{s}} \mathbf{Y}|\mathbf{Z}) \;\equiv\; (\mathbf{X} \perp\!\!\!\perp_{\mathbf{s}} \mathbf{Y}|\mathbf{Z}) \Rightarrow (\mathbf{X} \perp\!\!\!\perp_{\mathbf{c}} \mathbf{Y}|\mathbf{Z}) \tag{2.18}$$

Like the CMC, the CFC is a constraint that causal models use to restrict the structure of the underlying data-generating process. However, the CFC's role is not as significant as the CMC. The CMC acts as a guideline principle, motivating one to find a causal explanation for statistical correlations. In contrast, the CFC restricts the space of causal graphs compatible with a probability distribution and reduces the degree of underdetermination of causal structure by a probability distribution (Zhang & Spirtes 2016).

The CFC is justified by the fact that the parameters of unfaithful models must be chosen from a zero-measure set, implying that the parameter values needed to violate the CFC are unstable. In other words, if these parameters have slightly different values, the fine-tuning would not occur, making the violation unlikely (Weinberger 2018). To illustrate this idea, Pearl gives an example of an observer trying to distinguish between a chair and two chairs positioned such that one hides the other. Pearl argues that it would be unlikely for the two objects to align in such a way that one perfectly hides the other because this alignment would be unstable relative to slight changes in environmental conditions or viewing angle (Pearl 2009, p. 49). Chapter 3 will investigate the controversies surrounding these types of arguments.

Definition 2.7 (CPDAG). *A completed partially directed acyclic graph (CPDAG) is a DAG wherein only edges belonging to v-structures, and those that would introduce additional v-structures or cycles are directed.*

Definition 2.8 (Markov Equivalent). *$\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be Markov equivalent if* *they induce the same set of d-separation relations. This is the case if and only if the two DAGs have the same skeleton and the same set of v-structures (Pearl & Verma 1991). A collection of all DAGs that are Markov equivalent is called a "Markov equivalence class" and is represented by a CPDAG.*

2.2.2 Causal Discovery Algorithms

Causal discovery algorithms refer to computational techniques for inferring causal facts from statistical data. Given a dataset, a discovery algorithm finds a causal model that might explain the underlying causal relationships among the variables in the scenario under consideration. Discovery algorithms usually make four assumptions to identify the causal structure of a scenario: (1) acyclicity, (2) CMC, (3) CFC, and (4) causal sufficiency.[1]

Under these assumptions, a discovery algorithm identifies the causal structure up to a Markov equivalent class and represents it graphically by a CPDAG. This strategy is followed in most "traditional" discovery algorithms. Discovery algorithms are traditionally divided into three classes:

1. **Constraint-based** algorithms perform conditional independence tests to identify independencies holding in a given dataset. The goal is to return a CPDAG compatible with the statistical independencies. Under the assumptions mentioned above, the CPDAG found in this way is unique. The Inductive Causation (IC) (Verma 1993) and the Peter and Clark (PC) (Spirtes et al. 2000) algorithms are well-known examples of this class. Both algorithms (i.e., IC and PC) have extended variants that relax the sufficiency assumption and perform the discovery task in the presence of latent confounders. The extended variant of the IC is the IC* algorithm (Verma 1993), while the Fast Causal Inference (FCI) algorithm (Spirtes et al. 2000) is the PC extension.

2. **Score-based** algorithms treat the discovery task as an optimization problem. The goal is to find a CPDAG that maximizes a pre-specified scoring function. The found

[1] The causal sufficiency assumption is relaxed in some algorithms.

CPDAG is unique if the score function satisfies desirable properties (e.g., decomposable, score-equivalent, and consistent). An example of this class is the Greedy Equivalent Search algorithm (Chickering 2002), which implements a greedy search strategy and uses the Bayesian Information Criterion as its scoring function.

3. **Hybrid** algorithms combine techniques from the constraint-based and score-based methods to make the structural learning process computationally efficient. The goal is to find a CPDAG whose skeleton is compatible with the conditional independence relations, and at the same time, the score is maximum. The Max-Min Hill-Climbing (Tsamardinos et al. 2006) algorithm is a well-known example of a hybrid algorithm.

In the absence of additional information about a scenario, such as the temporal order of the variables or the functional form of mechanisms, a discovery algorithm cannot provide a unique DAG, and its output contains a number of undirected edges. To address this shortcoming, another class of discovery algorithms, known as **pairwise algorithms**, has been developed in recent years, whose focus is the identification of causal direction in bivariate scenarios, i.e., scenarios consisting of only two variables.

To make a start, let us first formalize the problem that a pairwise algorithm is supposed to solve. Consider a **bivariate scenario** composed of two variables X and Y in which it is always the case that either X is a cause for Y ($X \rightarrow Y$), or Y is a cause for X ($Y \rightarrow X$). Given a sample from the joint distribution of X and Y, the goal is to discover the true causal direction. Note that the two possible directions lead to causal models that are Markov-equivalent, and therefore traditional causal discoveries are unable to solve the problem. Moreover, no intervention is allowed to be performed on the variables, and determining the causal direction must only be based on the given observational data.

Pairwise algorithms are designed to deal with such bivariate scenarios. The basic idea of these algorithms is that statistical data, in addition to conditional independencies, contains various types of asymmetries that can be used to detect the causal direction. The asymmetries can include a wide range of statistical properties, such as conditional entropy, mutual information, moments, residuals of regressions, and standard deviation of

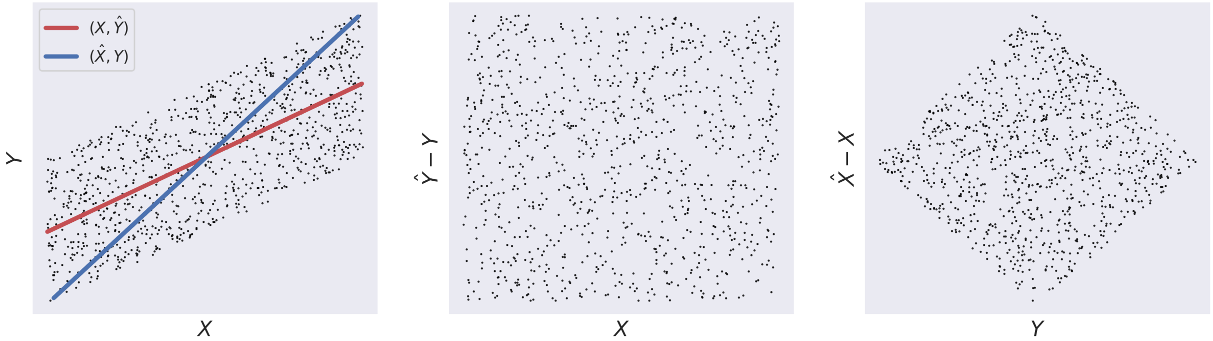


Figure 2.6: Statistical origin of asymmetries. Left: the scatter plot of X and Y; Middle: the estimated residual $\hat{E}_y = \hat{Y} - Y$ against the values of X; Right: the residual $\hat{E}_x = \hat{X} - X$ against Y.

conditional distributions.

Remark 2.6. *The Tübingen database*[2] *contains over one hundred hand-collected, real-world cause-effect samples (Mooij et al. 2016). These data are collected from different domains, and the causal direction in each dataset is determined with the help of experts in each domain. Pairwise algorithms are usually trained and tested on the said database.*

Figure 2.6 illustrates an example of a bivariate scenario in which asymmetries lead to the discovery of the causal direction. The underlying causal direction is $X \to Y$, and the causal mechanism is linear. To detect the causal direction in this scenario is to distinguish between model $Y = a_y X + b_y + E_y$ & $X \perp\!\!\!\perp_{\mathbf{s}} E_y$ and model $X = a_x Y + b_x + E_x$ & $Y \perp\!\!\!\perp_{\mathbf{s}} E_x$. To identify the causal direction, one needs to fit two linear regression models, once from X to Y and once from Y to X, and check which model leads to an error or residual that is independent of the input variable. As seen in Figure 2.6 (middle image), only the independence relation $\hat{E}_y \perp\!\!\!\perp_{\mathbf{s}} X$ holds, and therefore $X \to Y$ must be selected as the causal direction.

The previous example demonstrates a specific type of asymmetries arising when the causal mechanism is linear, and the noise is non-Gaussian. The literature on pairwise algorithms contains models encompassing more complex and non-linear scenarios. For instance, additive noise models postulate $Y = f(X) + E$ and post-nonlinear models postulate

[2] https://webdav.tuebingen.mpg.de/cause-effect/

$Y = f(g(X) + E)$. Nonetheless, pairwise algorithms are not restricted to models with predetermined functional forms.

In the two competitions (Guyon 2013, 2014), pairwise algorithms found a deeper relationship with Machine Learning approaches. In these competitions, participants were asked to solve the bivariate causal inference problem as a supervised Machine Learning problem. That is, they had to design and train a Machine Learning algorithm that could learn the causal direction in bivariate scenarios. The training and test data provided to the participants were a mixture of real and synthetic data from various domains, including chemistry, climatology, economy, epidemiology, medicine, physics, and sociology. Participants used innovative methods to solve this problem, which paved the way for more systematic approaches to pairwise discovery algorithms.

In general, pairwise algorithms can be divided into two categories: generative algorithms, which are based (explicitly or implicitly) on the notion of functional causal models, and discriminative algorithms, which directly determine the causal direction using a classification algorithm. The CGNN algorithm, discussed in Chapter 4, is a generative algorithm, while the RCC algorithm, discussed in Chapter 5, is a discriminative algorithm.

2.3 Quantum Mechanics

Quantum Mechanics is a physical theory that deals with the behavior of particles at the atomic and subatomic levels. Its unparalleled precision has led to significant advances in various fields, including quantum chemistry, quantum optics, semiconductors, and medical imaging. However, the foundations of Quantum Mechanics have been the subject of numerous debates due to its counter-intuitive properties, such as wave-particle duality, Schrödinger's cat paradox, Heisenberg's uncertainty principle, the measurement problem, the role of the observer, the reality of the wavefunction, quantum non-contextuality, and quantum non-locality.

Although the standard formalism of quantum theory is precise, it is silent about how to resolve these puzzles. Therefore, various interpretations of Quantum Mechanics have emerged, including the Copenhagen, many-worlds, pilot-wave, retrocausal, superdeterministic, and relational interpretations. These interpretations attempt to provide a perspective on the underlying nature of the theory to describe intuitively why standard formalism is the way it is.

This dissertation focuses on the conflict between the non-local character of quantum entanglement and the assumptions usually made in the causal modeling framework. Therefore, the other quantum puzzles and complexities will not be addressed. In the following sections, I will briefly introduce the mathematical formalism of Quantum Mechanics and explain how a famous Bell-type scenario is formulated. Then, I will discuss why quantum non-locality arises from the heart of Quantum Mechanics formalism and why it conflicts with the causal modeling framework.

2.3.1 Formalism

In Quantum Mechanics, a complex Hilbert space $\mathcal{H}$ is associated with each physical system, which serves as the basis for all mathematical calculations used to make predictions about the system. In this context, the term "system" refers to a physical object or a collection

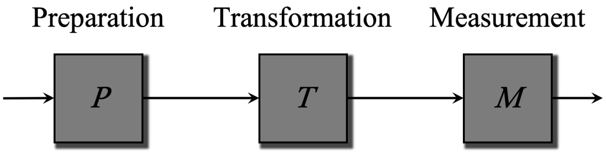


Figure 2.7: Preparation, Transformation, and Measurement procedures: a quantum system is firstly prepared in a quantum state through a preparation procedure P. The state is then transformed by procedure P. Finally, the system is measured by the measurement procedure M.

of objects under consideration. Examples of quantum systems include simple objects like electrons or complex collections like atoms or molecules.

The state of a quantum system is represented by a vector $|\psi\rangle \in \mathcal{H}$ or, more generally, an operator ρ. Each quantum system has a set of physical quantities, such as position, linear momentum, angular momentum, and energy, and these are represented in Quantum Mechanics by Hermitian operators such as A, where $A = A^{\dagger}$. The eigenvalues of the operator A correspond to the possible values that can be measured for the corresponding physical quantity. Unlike classical systems, where measurements are passive observations of physical properties, measurements affect systems in Quantum Mechanics. Thus, the physical quantities of a quantum system cannot be described independently of the measurement procedure.

Describing a system in Quantum Mechanics generally involves three procedures: preparation, transformation, and measurement. These procedures are illustrated in Figure 2.7. First, the system is prepared in a particular quantum state. Then, the system undergoes some interactions, which can lead to the transformation of the initial quantum state. Finally, a measurement is performed on the system to reveal the value of one of its physical properties.

Axiom 2.1 (**Preparation**). *A preparation P is associated with a trace-one positive operator ρ, known as the* ***density operator****, acting on the Hilbert space $\mathcal{H}$. The density operator represents the quantum state of a system.*

- *If the density matrix ρ has $\mathrm{Tr}(\rho^2) = 1$, the system is in a "pure state". Otherwise, it* *is in a "mixed state".*

- *To prepare a mixed state, we probabilistically combine several pure states $|\psi_i\rangle$ with probabilities p_i, giving us the density operator $\rho = \sum_i p_i|\psi_i\rangle\langle\psi_i|$.*

- *When two systems with states ρ_A and ρ_B are combined, the quantum state of the compound system is $\rho_{AB} = \rho_A \otimes \rho_B$.*

- *To obtain the state of one sub-system, we trace out the other sub-system: $\rho_A = \mathrm{Tr}_B(\rho_{AB})$.*

Axiom 2.2 (**Transformation**)**.** *A transformation T is associated with a completely-positive, trace-nonincreasing linear map $\mathcal{E} : \rho \to \mathcal{E}(\rho)$, transforming the density operator ρ to a new density operator $\mathcal{E}(\rho)$.*

- *The map $\mathcal{E}$ is trace-nonincreasing, meaning that $0 \leq \mathrm{Tr}(\mathcal{E}(\rho)) \leq 1$ for all density operators ρ.*

- *The map $\mathcal{E}$ is linear, meaning that $\mathcal{E}(\sum_i p_i\rho_i) = \sum_i p_i\mathcal{E}(\rho_i)$.*

- *The map $\mathcal{E}$ is completely positive, meaning that $\mathcal{E}(A)$ and $(I \otimes \mathcal{E})(A)$ are positive for any positive operator A.*

Axiom 2.3 (**Measurement**)**.** *A measurement M is a special transformation associated with a positive operator valued measure (POVM) M_m, where M_m is a positive operator corresponding to the outcome m. The POVM satisfies the constraint $\sum_m M_m = I$. The probability of obtaining outcome m from a measurement M for a system prepared using P and then transformed by T is calculated by taking the trace of the product of the POVM element M_m and the transformed density operator $\mathcal{E}(\rho)$:*

$$p(m \mid P, T, M) = \mathrm{Tr}(M_m\mathcal{E}(\rho)) \tag{2.19}$$

2.3.2 A Bell-Type Scenario

Linearity of transformations in Quantum Mechanics allows for situations where the state of a compound system can be non-factorizable, meaning it cannot be written as a tensor product of the states of its sub-systems, $\rho_{AB} \neq \rho_A \otimes \rho_B$. In such a situation, the sub-systems cannot be described by independent states but with the **entangled** state of their composition. Quantum entanglement is a phenomenon unique to quantum systems and has no counterpart in classical physics. A compound system in an entangled state can exhibit a particular type of correlation between the physical properties of its sub-systems, which remain even when the sub-systems are separated. Such statistical dependence between the statistics (i.e., outcomes of measurements) of entangled systems is called **non-local correlations**.

The non-local behavior of entangled particles has been a source of much debate in the foundations of Quantum Mechanics. Following Maudlin (2011), the quantum connection between entangled objects has at least three counter-intuitive features:

1. The quantum connection is **unattenuated**: no matter how far apart the two objects are, the statistical dependencies between the objects remain the same and are not affected by the spatial distance.

2. The quantum connection is **discriminating**: it is a private arrangement only retained by objects that have interacted in the past.

3. The quantum connection is **faster than light**: unlike other physical interactions, which are limited by the speed of light, the quantum connection appears to be instantaneous.

The non-local nature of quantum entanglement and the indeterministic nature of quantum theory leads to the idea that the description of Quantum Mechanics of systems may be incomplete. "Hidden variable theories" are proposals that attempt to explain quantum puzzles by postulating hypothetical entities. Einstein, Podolsky, and Rosen (EPR) famously formulated a paradoxical scenario in which (at least) one of the following conditions must be wrong:

1. A particular notion of locality motivated by Relativity ("...no real change can take place in the second system in consequence of anything that may be done to the first system.")

2. There exists a physical reality underlying Quantum Mechanics ("If, without in any way disturbing a system, we can predict with certainty (i.e., with probability equal to unity) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity.")

3. The description of quantum theory is complete.

By defending the first two conditions, the authors argued that the description of Quantum Mechanics is incomplete and supported the idea of hidden variables (Einstein et al. 1935).

Bell's theorem, proposed by John S. Bell in 1964, is a fundamental result in the foundations of Quantum Mechanics (Bell 1964). It is a test of the compatibility of hidden variable theories with Quantum Mechanics, designed to re-formulate the EPR idea in a more clear way. Bell's inequality is a mathematical expression that describes the expected correlations between the outcomes of measurements performed on entangled particles. According to Quantum Mechanics, this inequality can be violated, but according to hidden variable theories, it cannot.

Bell formulated his theorem in the context of a physical scenario whose rationale is very close to the EPR scenario; thus, I refer to it as the EPR-Bell scenario. The EPR-Bell scenario is a thought experiment that involves two particles in an entangled state. The experiment is designed to measure the correlation between the particles' properties and to compare it to the predictions of both Quantum Mechanics and hidden variable theories. Bell's inequality is derived from this scenario, and it states that the sum of the correlations between the particles' properties cannot exceed a certain value. If this inequality is violated, it is an indication that Quantum Mechanics is correct and that there

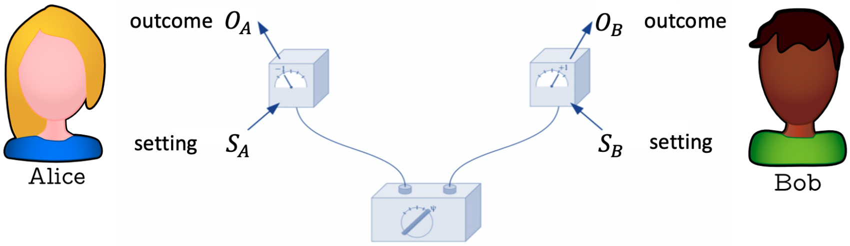


Figure 2.8: The CHSH scenario: two experimenters (Alice and Bob) locate in two spacelike separated regions. Each experimenter is allowed to set its measuring device freely.

are no hidden variables that can explain the correlations.

The Aspect experiment, conducted in the 1980s, confirmed the prediction of Quantum Mechanics that the Bell inequality is violated (Aspect et al. 1982). This experimental result ruled out a large class of hidden variable theories. The CHSH version of the EPR-Bell scenario, named after its inventors, Clauser, Horne, Shimony, and Holt, is a modification of the original scenario that simplifies the experimental setup while preserving the essential features of the experiment (Clauser et al. 1969).

Figure 2.8 provides a visual representation of the EPR-Bell scenario. In this scenario, two experimenters named Alice and Bob are located in distant regions that are spacelike-separated, which means that there is no causal influence between them. Entangled pairs of particles are emitted from a preparation device and sent to Alice and Bob, each of whom has a measurement device that measures the input particle. The measurement devices have several settings that determine the type of measurement to be performed on the particle, and experimenters can choose which measurement to apply to the input particle. In each run of the experiment, when an experimenter chooses a variable setting and performs a measurement, a measurement outcome appears. The experiment is repeated n times, with pairs of entangled particles sent to both experimenters and measurement outcomes appearing each time.

In the CHSH scenario, all variables in the experiment are binary, and their values are typically represented as $o_{A/B} \in \{-1, 1\}$ and $s_{A/B} \in \{0, 1\}$, respectively, denoting the outputs and settings of Alice and Bob. The dataset $\mathscr{D}$ corresponding to this experiment includes the results of all runs and is in the form of:

$$\mathscr{D} = \left(\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(n)}\right) \quad \text{with} \quad \mathbf{V}^{(i)} = \left(o_A^{(i)}, o_B^{(i)}, s_A^{(i)}, s_B^{(i)}\right). \tag{2.20}$$

Due to the quantum entanglement between the particles sent to the experimenters, quantum theory predicts that non-local correlations can be observed in the dataset $\mathscr{D}$. These correlations can be expressed in the form of the CHSH inequality, which takes the following form:

$$S = \mathrm{E}\left[O_A O_B|0,0\right] + \mathrm{E}\left[O_A O_B|0,1\right] + \mathrm{E}\left[O_A O_B|1,0\right] - \mathrm{E}\left[O_A O_B|1,1\right] \leq 2\sqrt{2} \tag{2.21}$$

where $\mathrm{E}\left[O_A O_B|i,j\right]$ represents the conditional expectation of the outcomes given the settings, i.e., the average of the product of the two outcomes in runs in which Alice and Bob have respectively chosen $S_A = i$ and $S_A = j$. The value $2\sqrt{2}$ is known as the Tsirelson bound and is an upper limit on the strength of quantum correlations between distant events.

In addition to the conditional dependence between the outcomes, the dataset $\mathscr{D}$ contains other statistical patterns from which causal information might be extracted. Remarkably, the dataset satisfies the **no-signaling** relations, which assert that the outcome at one wing of the experiment is independent of the setting at the opposite wing given the setting at the first wing:

$$O_A \perp\!\!\!\perp_{\mathbf{s}} S_B | S_A \quad \text{and} \quad O_B \perp\!\!\!\perp_{\mathbf{s}} S_A | S_B \tag{2.22}$$

The no-signaling constraint ensures that no communication can occur between the wings of the experiment. Another statistical relation that exists in $\mathscr{D}$ is that the two settings are marginally independent, i.e., $S_A \perp\!\!\!\perp_{\mathbf{s}} S_B$. The justification for the independence between the settings is straightforward: experimenters can choose their settings freely. If the shared entangled state is non-maximal, the outcome at one wing is statistically dependent on the setting at the same wing, i.e., $O_A \not\perp\!\!\!\perp_{\mathbf{s}} S_A$ and $O_B \not\perp\!\!\!\perp_{\mathbf{s}} S_B$.

The above inequality and the statistical (in)dependencies mentioned above have been confirmed empirically through experiments such as the Aspect experiment, indicating that the predictions of quantum theory are in perfect agreement with experimental results. The violation of the CHSH inequality is considered one of the strongest pieces of evidence for the non-locality of entanglement and has important implications for the foundations of Quantum Mechanics.

2.3.3 The Causal Problem of Entanglement

Having explained statistical (in)dependencies in the EPR-Bell scenario, we can now explore the question of what causal graph can explain these statistics. As we will see in the following paragraphs, none of the causal graphs that adhere to the CMC and CFC assumptions can reproduce the scenario's statistics. Consequently, the CMC and CFC assumptions contradict each other in the quantum domain, which leads to the so-called **causal problem of entanglement**.

John Bell formulated an initial version of this problem. Bell's original formulation was based on physical considerations (e.g., the concept of local causality) rather than the CMC and the CFC assumptions. Moreover, Bell's formulation did not account for all types of causal graphs that could be defined for the EPR-Bell scenario. A generalized version of Bell's argument is presented by Wood & Spekkens (2015), in which the authors show that, regardless of the physical considerations Bell employed, it is possible to demonstrate that no causal graph that complies with the acyclicity, CMC, and CFC assumptions can account for the scenario's statistics. This section reviews Bell's formulation of the problem and directs the reader to Wood & Spekkens (2015) for a more comprehensive argument.

The term "be-able" derives from the English verb "be" and refers to elements of a physical theory that correspond to real objects in the physical world. For Bell, beables are the primitives of a physical theory, and other concepts, such as observables and mathematical operators, must ultimately be reducible to beables. Local beables are a particular type of beables that can be assigned to a bounded region of spacetime. Suppose we have a complete specification of the local beables of a region. In that case, we should be able to express the probabilities the theory attaches to events occurring in that region in terms of

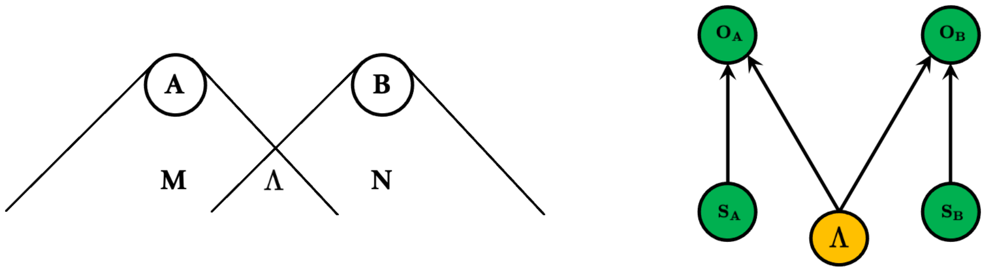


Figure 2.9: The Local Causality assumption. Left: local beables in the shared region (Λ) screen-off the probabilities in A and B. Right: The corresponding causal graph of the CHSH scenario.

local beables.

Figure 2.9 (left) depicts a scenario where two events A and B occur in two spacelike separated regions. The light cones associated with these regions have an overlapping region that characterizes the causal past of the two. According to Bell, a theory that is locally causal satisfies the following condition:

$$p(A, B|M, N, \Lambda) = p(A|M, \Lambda)\, p(B|N, \Lambda) \tag{2.23}$$

where M and N are the specifications of all local beables in the regions associated with A and B, respectively, and Λ is the specification of local beables in the overlapping region. The condition says that a complete specification of local beables in the overlapping region "screens off" the probabilities of events in A and B.

Figure 2.9 (right) depicts the causal graph of the EPR-Bell scenario based on the idea of local beables. Here, Λ represents a possible latent common cause for the outcomes in the two wings of the experiment and contains all the necessary information to characterize the entangled state shared by Alice and Bob. Since the entangled pair is prepared before the settings are chosen, it is natural to treat Λ as an exogenous variable. Similarly, since the value of the settings is determined before the outcomes, it is reasonable to represent the causal dependence between the setting and outcome of each wing through a directed edge from the setting to the outcome.

Although assuming the CMC and CFC, the causal graph in Figure 2.9 can explain some statistical (in)dependencies present in quantum statistics, it cannot reproduce the Tsirelson bound. Let us first see what relations the graph does reproduce. First, the no-signaling relations are satisfied due to the d-separation criterion (and thus the CMC assumption). Similarly, the setting independence condition is reproduced due to the d-separation criterion. Moreover, the statistical dependencies $S_A \not\perp\!\!\!\perp_{\mathbf{s}} O_A$ and $S_B \not\perp\!\!\!\perp_{\mathbf{s}} O_B$ are reproduced because the CFC would be violated otherwise.

Nonetheless, the graph predicts the classical bound 2 instead of the Tsirelson bound $2\sqrt{2}$. This discrepancy arises because, under the assumption of CMC, the joint distribution induced by the graph takes on a specific form:

$$p\left(o_A, o_B, s_A, s_B, \lambda\right) = p\left(o_A | s_A, \lambda\right) p\left(o_B | s_B, \lambda\right) p\left(s_A\right) p\left(s_B\right) p\left(\lambda\right) \quad \forall o_A, o_B, s_A, s_B, \lambda \tag{2.24}$$

From the above equation, the following condition is derived, which is known as **factorizability** condition:

$$p(o_A, o_B | i, j, \lambda) = p(o_A, | i, \lambda) p(o_B | j, \lambda) \tag{2.25}$$

Since the latent variable Λ can take probabilistic values, one can formulate the S parameter in Equation 2.21 by averaging over all the values taken by the latent variable:

$$S = \int d\lambda p(\lambda) S(\lambda) \quad \text{where} \quad S(\lambda) = \sum_{i,j} \mathrm{E}\left[O_A O_B | i, j, \lambda\right] (-1)^{ij}$$

Moreover, from the factorizability condition in Equation 2.25 it implies that the expectation values of the outcomes can be factorized into separate terms, i.e., $\mathrm{E}\left[O_A O_B | i, j, \lambda\right] =$

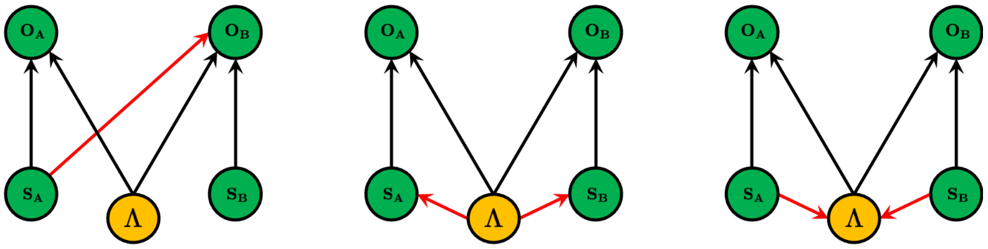


Figure 2.10: Three major interpretations of Quantum Mechanics violate the CFC assumption. Left: superluminal; Middle: superdeterministic; Right: retrocausal.

$\mathrm{E}\left[O_A|i,\lambda\right]\mathrm{E}\left[O_B|j,\lambda\right]$. Therefore, $S(\lambda)$ is expressed as

$$\begin{aligned} S(\lambda) = &\underbrace{\mathrm{E}\left[O_A|0,\lambda\right]}_{\leq 1}\left(\mathrm{E}\left[O_B|0,\lambda\right] + \mathrm{E}\left[O_B|1,\lambda\right]\right) + \\ &\underbrace{\mathrm{E}\left[O_A|1,\lambda\right]}_{\leq 1}\left(\mathrm{E}\left[O_B|0,\lambda\right] - \mathrm{E}\left[O_B|1,\lambda\right]\right) \end{aligned}$$

Finally, given that the only values the outcomes can take are either -1 or $+1$, it holds that $|S(\lambda)| \leq 2$ and hence $|S| \leq 2$. The inequality presented here is known as the **CHSH inequality**. The CHSH inequality provides the upper bound on the strength of correlations in classical systems. That is, the upper bound for classical correlations is 2, in contrast to the quantum correlations whose upper bound is given by the Tsirelson bound $2\sqrt{2}$. Therefore, the graph in Figure 2.9 cannot fully reproduce quantum statistics, mainly because it predicts a weak strength for the correlations between the outcomes.

As previously noted, not only is the graph in Figure 2.9 unable to reproduce quantum statistics faithfully, but any causal graph based on the assumptions of acyclicity, CMC, and CFC cannot reproduce quantum statistics either. In particular, many proposed models for the underlying causal structure in the EPR-Bell scenario, including those which exploit superluminal, superdeterministic, and retrocausal influences, violate the CFC assumption (see Figure 2.10).

2.4 Conclusion

In this chapter, we have covered the fundamental concepts required for understanding the subsequent chapters. A critical limitation of causal models in reproducing quantum phenomena was highlighted, known as the causal problem of entanglement. The CMC and CFC assumptions were also discussed, as they are fundamental concepts in causal modeling. Moving forward, it is important to clarify the approach taken in exploring interpretations of Quantum Mechanics in this thesis. Specifically, our arguments are based on the causal modeling framework, and as such, we only consider interpretations that can be expressed in the form of a causal graph. This includes interpretations such as superluminal, superdeterministic, and retrocausal models. Whether other interpretations of Quantum Mechanics, such as QBism and many-minds interpretations, can be examined within the causal modeling framework is an interesting open question that is not addressed in this thesis. Overall, this chapter has laid the foundation for the subsequent chapters, which will explore various interpretations of Quantum Mechanics through the lens of causal modeling.

Chapter 3

Quantum Conflicts

It is commonly believed that the causal problem of entanglement is a dilemma between abandoning either the CMC or the CFC assumptions in the quantum realm. While giving up either of these assumptions might be a valid response to the problem, this chapter argues that there are alternative possibilities stemming from various physical and philosophical considerations. The chapter aims to study these possibilities, evaluate their validities, examine their consequences, and explore their interrelationships. Moreover, the chapter builds the philosophical and foundational roots for the subsequent chapters on ML approaches.

3.1 Causal Markov Condition (CMC)

The CMC assumption has a central role in the foundations of the causal modeling framework, which has a twofold application. On the one hand, as a *methodological* principle, it asserts that correlations require causal explanation. Thus, it guides one in searching for causal explanations whenever logically independent variables exhibit correlations. On the other hand, it is a *bridging* principle between statistical and causal facts. Thus, it builds connections between statistical inference methods and causal inference methods.

Despite having such a central role, the causal problem of entanglement has led some researchers to doubt the CMC's validity (or at least its range of validity). Most notably, after examining a tripartite quantum scenario and failing to provide a causal explanation for quantum correlations, Clark Glymour concluded that some phenomena might admit no causal explanation or their causal explanation is not based on the CMC:

> ...there is no causal explanation of the phenomenon, or that there is a causal explanation but it doesn't satisfy the Markov assumption. ... It is not a truth of logic that all experimental associations have a causal explanation, and it is not a truth of logic that all causal relations satisfy the Markov Assumption. That's up to Nature (Glymour 2006, p. 124).

Although it may be tempting to abandon the CMC as a solution to the causal problem, we must approach this suggestion with caution, as it raises many difficult questions. For example, why does the CMC hold in classical scenarios but not in quantum scenarios? Are there specific conditions under which the CMC applies? When should we search for causation when some correlations can exist without causation? While Glymour acknowledges these problems and reminds the reader to think of "why, then, does the Markov Assumption work with our experiments on middle sized dry and wet goods, with climate, and rats and drugs, and so much else?" (Glymour 2006, p. 125), he does not provide clear answers.

To address these questions, we can reject the CMC based on its underlying assumptions. This rejection can be justified by either (1) rejecting Reichenbach's principle of common cause or by (2) rejecting Pearl's assumption of independence of noise terms in functional causal models. I will briefly discuss and evaluate both of these proposals.

The status of Reichenbach's principle has been widely debated in the philosophical literature, and there are many alleged counterexamples against its validity (see, e.g., Arntzenius 2019). As the CMC is a generalization of Reichenbach's principle, any counterexample against Reichenbach's principle can be seen as a counterexample against the CMC. While there are many proposals, I will focus on two subtle cases: interactive common causes (Cartwright 1988) and separate common causes (Hofer-Szabó & Vecsernyés 2012).

3.1.1 Interactive Common Causes

One approach to relinquishing Reichenbach's principle (and therefore the CMC) is to accept the essence of the principle but reject its screening-off condition. This means rejecting the idea that the conditional probabilities of effect variables factorize given their complete common cause. An example, originally from Cartwright, can help understand how this can happen.

Consider a molecule that is initially in a state, which after a while breaks down into two equally-sized fragments, each having its own final state. The initial state is the common cause for the final states; however, it cannot entirely determine its effects. The only thing that can be inferred from the common cause is that the fragments move with similar speeds in opposite directions due to the conservation of linear momentum. As a result, the initial state is a non-screening-off (or interactive) common cause for the final states. This example can be generalized to a broader context, where there is a global quantity $Q = f\left(q_1, \ldots, q_n\right)$ whose value is conserved due to some physical law. Whenever one of the local quantities q_i changes indeterministically, the other local quantities inevitably change so that the global quantity remains constant. Here, Q is a common cause for the local quantities q_i, but again it is an interactive common cause for them.

Applying the idea of interactive common causes to the EPR-Bell scenario means considering an interactive common cause for the two outcomes such that the factorizability of conditional probabilities does not hold. Figure 3.1 depicts this idea.

Interactive common causes can be a compelling solution to the causal problem, for they preserve the methodological spirit of the CMC (i.e., correlations have causal explanations) and resolve the conflict by slight modifications to the screening-off condition. However, a study by Näger (2020) demonstrated that the correlations produced by interactive common causes are not strong enough to violate Bell inequalities. This means that although interactive common causes do weaken the CMC, the weakening is insufficient to reconcile Quantum Mechanics with classical causal models. Thus, interactive common causes in their current form are not a solution to the causal problem, and alternative solutions

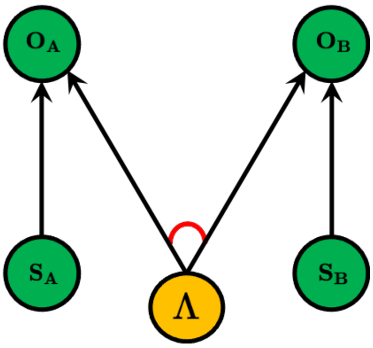


Figure 3.1: The causal graph of the EPR-Bell scenario in the presence of an interactive common cause (the arc between the edges symbolizes the idea of interactive common causes)

should be explored.

3.1.2 Separate Common Causes

Another approach for resolving the causal problem of entanglement based on Reichenbach's principle is to argue that Reichenbach's principle is about one pair of correlated events, whereas the EPR-Bell scenario concerns more than one pair. According to this argument, the issue lies not with Reichenbach's principle itself but with its application to the EPR-Bell scenario.

To simplify matters, let us assume that both Alice and Bob's setting variables are binary-valued. This means that there are four possible setting choices, such as $s_A, s_B \in \{0,1\} \Rightarrow (s_A, s_B) \in \{(0,0),(0,1),(1,0),(1,1)\}$. When we say that the outcomes of the two wings are correlated, we are actually referring to four pairs of correlated events (and hence four common causes), each corresponding to one possible choice of settings.

Nonetheless, Bell inequalities derivations are based on the tacit assumption that there exists a single common cause screening off all four correlations simultaneously. This assumption is known as the idea of having "common" common causes (Szabó 2000). Contrary to this idea is the notion of "separate" common causes, which asserts that corresponding to each correlated event, there should exist one hypothetical common cause. Hence, statistical dependencies of the two outcomes must be explained by a "system" of separate

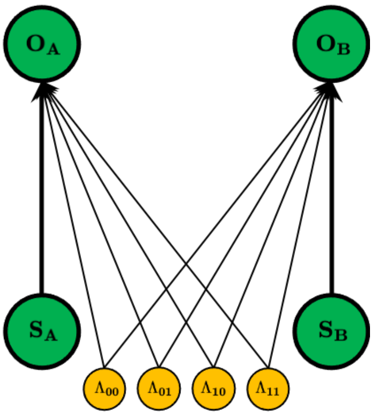


Figure 3.2: The causal graph of the EPR-Bell scenario in the presence of separate common causes.

common causes rather than a single common cause variable.

Figure 3.2 illustrates the causal structure of the EPR-Bell scenario in the presence of separate common causes. There are four common causes, each responsible for one correlation event, such as Λ_{00} being responsible for the outcomes' correlation when Alice and Bob choose $s_A = s_B = 0$.

According to Redei et al. (2013), an interesting feature of Figure 3.2 is that the graph does not automatically lead to problematic conditions such as "hidden locality".[1]

Thus, such a graph blocks the derivation of Bell inequalities. Moreover, it has been shown that a separate common cause system can generate the statistics of Quantum Mechanics in the EPR-Bell scenario with binary-valued settings (see proposition 9.10 Redei et al. 2013). Separate common causes might be a legitimate proposal for resolving the causal problem of entanglement. However, it is important to consider one caution.

The argument for separate common causes relies on the assumption that the settings are discrete-valued variables, which results in a finite number of common causes. This implies that the concept of separate common causes may not have a straightforward interpretation when applied to continuous scenarios, such as the one to be examined in the next chapter. Therefore, while separate common causes might block the derivation of Bell-

[1] The "hidden locality" condition says that given a pair of correlated outcomes and the hypothetical common cause of that correlation, the probability of the outcome in one wing is independent of the direction of the measurement in the other wing: $O_A^{(i)} \perp\!\!\!\perp S_B^{(j)} | (S_A^{(i)}, \Lambda_{ij})$ and $O_B^{(j)} \perp\!\!\!\perp S_A^{(i)} | (S_B^{(j)}, \Lambda_{ij})$.

type inequalities, further justification is needed to establish its validity as a solution to the causal problem. While there is one issue to consider, it may be possible to prove the same statements for continuous variables or to implement heuristic ideas that approximate continuous problems with discrete problems, such as by binning the continuous setting variable and converting it into a set of discrete partitions.

3.1.3 Correlated Noises

One alternative to abandoning the Causal Markov Condition (CMC) is to question the assumption that the noise terms in a functional causal model are statistically independent. Correlated noise terms can invalidate Pearl's derivation of the CMC and may necessitate a new formulation that is not affected by quantum correlations. As Steel (2020, p.229) suggests, it is necessary to provide a justification for assuming independent exogenous error terms.

In my opinion, the straightforward argument for assuming independent noise terms is that correlated noise terms can introduce statistical dependencies between variables without acknowledging causal connections, which can undermine the validity of the whole causal modeling approach. Therefore, postulating independent noise terms is a necessary assumption for most causal modeling tasks. However, there are situations where the assumption of independent noise terms may not hold. For example, in some cases, the noise terms may be correlated due to unmeasured common causes or feedback loops in the causal system. In such cases, the causal Markov condition may not hold, and other techniques, such as instrumental variables, may be necessary to recover the causal relationships between variables.

Overall, while the assumption of independent noise terms is a powerful simplification in causal modeling, it is important to be aware of situations where this assumption may not hold and to use appropriate techniques to address these challenges. I see two possibilities for postulating correlated noises depending on whether or not the causal sufficiency condition is satisfied.[2]

Firstly, if the causal sufficiency condition is not met and the system is non-Markovian, there may be correlated noise terms between variables whose common causes have not been measured. However, these correlated noise terms will disappear as soon as the relevant unmeasured common causes are added to the model. Pearl's position on non-Markovian models is as follows:

> ... such models - even if any exist in the macroscopic world - would have limited utility as guides to decisions. For example, it is not clear how one would predict the effects of interventions, save for explicitly listing the effect of every conceivable intervention in advance. (Pearl 2009, p. 61)

By the way, correlated noises stemming from the first possibility (i.e., the existence of common causes have not been detected) are less relevant to the EPR-Bell scenario since the hidden variable Λ is a complete specification of *all* local beables of the overlapping region of the two backward light cones, leaving no room for unmeasured common causes to bring about correlated noises.

The second possibility arises when the causal sufficiency condition is satisfied from the beginning, and correlated noise terms can arise due to non-separability caused by a meta-mechanism influencing the causal mechanisms of the scenario. The origin of this meta-mechanism and its mathematical modeling are discussed in detail in Section 3.3.2, but it is worth noting that this possibility has exciting implications for the causal problem.

3.2 Causal Faithfulness Condition (CFC)

The CFC plays a critical role in the causal modeling framework, serving a dual purpose. As a *methodological* principle, the CFC states that there is no reason to believe that causal connections are hidden from observers, and hence, causally related variables are expected

[2] Recall that causal sufficiency is met when all the common causes of the measured variables have been observed.

to display statistical dependencies. As a *bridging* principle, the CFC connects causal facts to statistical facts.

In practice, the CFC is a crucial assumption for many constraint-based causal discovery algorithms. These algorithms involve performing conditional independence tests, which result in a set of Markov-equivalent models. The CFC is then employed to identify the models whose conditional independence relations arise due to their graphical structure rather than the fine-tuning of their parameters. This is the approach taken by Wood & Spekkens (2015) to formulate the causal problem of entanglement: they assumed certain independence relations (i.e., independence of setting variables and no-signaling conditions) and searched for causal graphs that can reproduce these independencies under the CFC assumption. The authors demonstrated that there is no causal graph that can faithfully reproduce the statistics of Quantum Mechanics while respecting the CFC, thereby concluding that three major interpretations of Quantum Mechanics are unsatisfactory since they violate the CFC:

> It follows that all three of the main approaches for providing a causal explanation of Bell correlations, superluminal causes, superdeterminism and retrocausal influences, are unsatisfactory, and they are all unsatisfactory for the same reason (Wood & Spekkens 2015, p. 3).

Should we question the conclusion that the CFC confidently rules out three major interpretations of Quantum Mechanics? The philosophical literature offers various strategies to challenge this claim, one of the most notable being presented by Nancy Cartwright. In her work, Cartwright (2001) focuses on the famous canceling path scenario, in which the causal effect of one variable on another fades because the influence is carried by two paths that cancel each other out. Cartwright raises the question of how to distinguish between causally independent variables and variables subject to path cancellation. Her proposal is that we should incorporate independent evidence about the relevant causal mechanisms in our causal model alongside statistical observations. However, in the absence of such independent evidence, we must exercise caution:

> Where we don't know, we don't know. When we have to proceed with little information we should make the best evaluation we can for the case at hand—and hedge our bets heavily; we should not proceed with false confidence having plumped either for or against some specific hypothesis—like faithfulness—for how the given system works when we really have no idea (Cartwright 2001, p. 254).

What stance should we adopt towards the CFC in the causal problem of entanglement? It is worth noting that we lack independent evidence regarding the underlying causal mechanisms in the EPR-Bell scenario, as our knowledge is limited to a set of conditional independence relations and a strong correlation between the two outcomes. In fact, one of the primary motivations for formulating different interpretations of Quantum Mechanics is our lack of understanding of the underlying mechanisms of Quantum Mechanics. Therefore, following Cartwright's argument, we should not make any conclusive decisions either for or against the CFC based on our limited knowledge.

However, how should we proceed with causal discovery in the EPR-Bell scenario if the CFC fails? Should we accept the failure of the CFC and embrace the Quantum Mechanics interpretations, or should we set aside the Quantum Mechanics interpretations and accept the CFC? In the rest of this section, we will explore two general solutions to this problem: (1) utilizing "natural fine-tuning" and (2) utilizing alternative tools for causal discovery that go beyond the CFC.

3.2.1 Natural Fine-Tuning

As discussed in the previous chapter, the CFC assumption is motivated by the stability demand, which suggests that causal models should be stable against external perturbations. However, this demand can be challenged in cases where the counterbalance between canceling paths is always stable. In the context of the causal problem of entanglement, this is the approach proposed by Näger (2016).

What if the counterbalance between the two paths always remains stable? If the fine-tuning of causal parameters can be achieved in such a way that it always remains stable, then the stability demand for postulating the CFC loses its functionality. This is precisely the strategy that Näger (2016) suggests to deal with the causal problem of entanglement.

Näger argues that the no-signaling independencies observed in the EPR-Bell scenario may be the result of internal canceling paths that are finely tuned by the laws of nature in a nomological manner consistent with Quantum Mechanics. If the counterbalance between different paths is naturally fine-tuned, then the causal model does not need to rely on unstable canceling mechanisms. Näger explains:

> The most convincing answer is provided by the quantum mechanical formalism which assumes the laws of nature to have a specific form that guarantees that causal parameters, which change due to disturbances, only change in a balanced way (Näger 2016, p. 49).

In my view, Näger's approach to the EPR-Bell scenario is both philosophically and physically compelling. Philosophically, the approach is consistent with Cartwright's argument and does not take a firm stance either for or against the CFC. Moreover, the approach preserves the key concepts of causal explanations by introducing minor adjustments to the implementations of stable fine-tuning. From a physical standpoint, the approach is supported by certain interpretations of Quantum Mechanics. There are at least two physical models that explain how stable fine-tuning can be achieved in the EPR-Bell scenario. The first model is based on the "quantum equilibrium hypothesis" and lends credence to superluminal causal models. The second model is based on "internal symmetries" and supports retrocausal models.

Quantum Equilibrium Hypothesis

In Bohmian mechanics, the possibility of superluminal signaling is prevented by relying on the quantum equilibrium hypothesis. According to this hypothesis, in a Bohmian world, the initial configuration of particles is determined by the Born rule. Without this hypothesis, Bohmian mechanics could potentially violate the no-signaling independencies, allowing manipulation of the outcome on one wing by changing the setting on the other wing. However, the existence of the quantum equilibrium hypothesis ensures that this is not the case in Bohmian mechanics.

But can we reject Bohmian mechanics based on the violation of the CFC? I do not think so because this type of fine-tuning is law-like and stable. A physical law, namely the Born rule, is responsible for adjusting the initial configuration of particles, and this leads to stable fine-tuning. Therefore, if there are independent justifications in favor of the equilibrium hypothesis, fine-tuning in Bohmian mechanics is not problematic. In fact, some proponents of Bohmian mechanics use this argument to defend their theory:

> ...Wood and Spekkens take this to be a case of fine-tuning. If, however, the quantum equilibrium hypothesis can be justified in terms of the statistical behaviour arising from a typical initial configuration, as Dürr et al. (2013, ch. 2) argue, there is nothing problematic about this fine-tuning (Egg & Esfeld 2014, p. 9).

Internal Symmetries

The second example of stable fine-tuning is proposed by Almada et al. (2016), where a retrocausal model is presented that satisfies the no-signaling condition through the internal symmetries of the model structure. The author correctly recognizes that internal symmetries can naturally lead to stable fine-tuning:

> ...if one hopes to naturally restrict signalling without unnatural fine-tuning, one should look for models with an abundance of symmetries (Almada et al. 2016, p. 13).

However, Almada's model is not the only one that exploits stable fine-tuning. For example, Weinstein (2018) presents a machine-learning model for reproducing the statistics of the EPR-Bell scenario, where superluminal signals are blocked through the symmetries of weights and biases within the hidden layers of the model. Another example is the retrocausal model proposed by Argaman (2010).

3.2.2 Beyond the CFC

If no independent evidence about the causal mechanisms of the EPR-Bell scenario is available, why not go beyond the CFC and use alternative criteria? In this way, we no longer "proceed with false confidence having plumped either for or against" the CFC. Wood and Spekkens also suggest this remedy:

> CI-based (conditional-independence-based) causal discovery algorithms do not do justice to Bell's theorem. [Conditional] independencies simply do not provide enough information. One needs a causal discovery algorithm that looks at the strength of correlations to reproduce Bell's conclusion (Wood & Spekkens 2015, p. 17).

There are two systematic ways to extract causal information from statistical data: (1) conditional independence relations and (2) structural and distributional details of variables as an indication of the underlying functional forms of the causal mechanisms. While constraint-based algorithms only use the first source, pairwise and score-based algorithms are capable of using the second or even both sources. Accordingly, pairwise and score-based algorithms do not usually require the CFC assumption for causal discovery. Many algorithms fall into this type, and discussing all of them is beyond the scope of this section. However, the basic idea is that the direction of causal mechanisms leaves some footprints on the statistical data that differ from the conditional independence relations, and these footprints can be detected by certain methods. Importantly, Machine Learning approaches can be used to determine the direction of causality from these asymmetries. Chapter 4 and Chapter 5 will demonstrate how this idea can provide new tools to study the causal problem of entanglement.

It is worth mentioning that, besides Machine Learning approaches, there are proposals to weaken or generalize the notion of the CFC. For instance, Uhler et al. (2013) introduces a λ-strong-faithfulness criterion based on the strength of correlations. Alternatively, Cavalcanti (2018) suggests that a generalized CFC can be defined as follows: a causal model should not allow causal connections stronger than needed to explain the observed deviations from the observed conditional independence relations in the EPR-Bell scenario.

In conclusion, the causal problem of entanglement does not necessarily lead to the rejection of fine-tuned interpretations of Quantum Mechanics. On the one hand, there are good reasons to doubt the validity of the CFC in the quantum realm. On the other hand, there are interesting alternative criteria of the CFC that can be exploited for causal discovery.

3.3 Non-separability Arguments

Quantum entanglement has another puzzling feature that takes the causal problem beyond the debates surrounding CMC and CFC. This feature stems from the tension between the non-local character of quantum entanglement and Relativity and leads to doubts about the localizability of entangled objects. The central issue can be framed as follows:

> ... whether one can imagine that entangled objects are embedded in a relativistic spacetime, is the central problem of entangled systems (Friebe et al. 2018, p. 135).

It should be noted that the standard interpretation of Relativity requires that causal processes propagate at most with the speed of light. That is, there are no causal processes that propagate faster than light. A "causal process" between two variables refers to a chain of causes and effects that connect the states of the two variables. A causal process has two components: the relata (i.e., the states of the two variables) and the relation itself. To say that a process is local is to say that the relata and the relation are both local. Hence, the non-locality of a process can arise either from the non-locality of states or the non-locality of relations.

In the context of the EPR-Bell scenario, quantum non-locality can thus arise from two sources: (1) the non-locality of the ontic state of entangled objects or (2) the non-locality of dynamics of the entangled objects. While many proposals in the physics literature tend to accept the second option to address the EPR-Bell scenario, some arguments favor the first option. I refer to these arguments as "non-separability arguments" and discuss two types of such proposals in the following sections: (1) non-localized states and (2) dependent causal mechanisms.

3.3.1 Non-localized States

The essence of the current view is that the non-local behavior of quantum objects originates from the fact that a pair of entangled particles should not be seen as a pair of distinct objects located in different regions. Instead, the pair should be constructed as one entity whose state is spread in the whole spacetime. Woodward has an obvious comment in this direction:

> One might go on to argue that the interactions of the electrons at the different measuring devices are in fact not distinct events. One has instead a single wave state extended over a large distance. So there is not only no causal explanation of separate events, but no separate events to be [non-accidentally] connected (Hausman & Woodward 1999, p. 657).

He makes a similar statement elsewhere to defend the reliability of the causal modeling framework in dealing with entangled particles:

> ... there is no well defined notion of intervention on the spin state of one of the separated particle pairs with respect to the other in EPR type experiments. This would represent a limitation on the application of an interventionist account of causation only if there was reason to suppose that there is a direct causal connection between these states. Hausman and Woodward argue there is no such reason, hence that it is a virtue...that the interventionist account does not commit us to such a connection (Woodward 2007, p. 70).

Figure 3.3 depicts a non-separable causal model that employs the concept of non-localized states. The variable Λ denotes the non-localized state that extends over the whole spacetime due to the presence of entanglement. Being non-localized, Λ can be influenced by

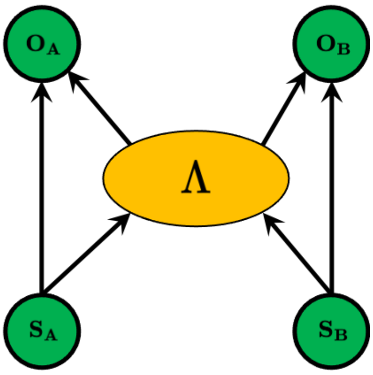


Figure 3.3: A non-separable causal model: Λ spreads throughout a whole spacetime region

both setting variables and can, in turn, influence both outcomes. This model adopts a holistic view regarding quantum entanglement, as the states of all sub-regions depend on the state of the whole region.

While this may seem like an unusual physical picture, there are observations that support it. For example, entanglement connections are insensitive to spatial and temporal coordinates of spacetime. They are "spatially insensitive" because the strength of entanglement correlations does not change with increasing or decreasing the spatial distance between the entangled particles. Similarly, they are "temporally insensitive" because the destruction of entanglement (i.e., the collapse of the wavefunction) happens simultaneously.

Non-separable causal models may be attractive because they can avoid tensions with the special theory of Relativity. Note that no definite temporal order can be assigned between the two wings of the EPR-Bell experiment because they are spacelike-separated. However, any causal model that postulates direct causation between the wings faces the problem of frame dependency for the causal direction. Non-separable models do not face this problem because they do not posit direct causation between the two wings. Despite this argument, recent works in quantum causal models have criticized the non-separable model. For instance, Shrapnel (2019) argues that by generalizing the notion of intervention for quantum systems, one can break the correlation between the two outcomes and undermine the view that the two objects are a single, indivisible event.

3.3.2 Independence of Mechanisms

The concept of non-separability can also be understood in terms of the independence of causal mechanisms. A causal mechanism is a function that maps the values of parent nodes to a child node, and the number of causal mechanisms in a graph is equal to the number of nodes. The independence of causal mechanisms assumes that each mechanism is autonomous and operates independently of all others. This assumption is crucial to the possibility of local interventions, i.e., changing one part of a system without affecting other parts. As Pearl puts it:

> The world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behaviour of a relatively small number of variables. Actions are local in the space of mechanisms (Pearl 2009, p. 229).

Pearl does not explain why the world is decomposable in this way, nor does he sketch how the world looks when the assumption is violated. He accepts the independence of causal mechanisms as a primitive ontological fact and uses the assumption to guarantee the possibility of local interventions, i.e., to alter one part of a system without altering the other parts.

Although I accept that Pearl's picture accords with everyday life experience and it might be convincing to expect a classical world to behave in this way, I argue that the validity of this picture in the quantum domain is not as straightforward as Pearl puts. First, quantum systems exhibit puzzling behaviors that make it hard to believe that how these systems behave always accords with everyday life experience. For instance, the non-local character of quantum entangled particles should make us cautious about statements such as "actions are [always] local". Or, the Kochen–Specker theorem, which shows the impossibility of non-contextually assigning properties to quantum systems (Kochen & Specker 1975), suggests that, unlike any known classical system, quantum systems do not admit preexisting properties. Second, we do not access a lower-level description of what happens in quantum mechanisms, so we do not know what those "small number of variables" are in a quantum process. This would be a caveat for Pearl's assertion since he utilizes these variables to guarantee the independence of mechanisms.

For these reasons, I think it is essential to examine the validity of the independence of causal mechanisms in the quantum realm instead of postulating it in the first place. To do so, we must provide a precise definition of the independence of causal mechanisms. One such definition is provided by Peters et al. (2017), a textbook that uses Machine Learning approaches to causal modeling.

Definition 3.1 (Independence of Causal Mechanisms). *The causal generative process of a system is composed of autonomous modules that do not "influence" or "inform" each other. In the probabilistic case, the conditional distribution of each variable given its parents does not inform/influence the other conditional distributions.*

Peters et al. (2017) point out that the above definition can be construed from two perspectives: physical and information-theoretic. From a physical point of view, the independence of causal mechanisms postulates that mechanisms do not "influence" each other. So, there is no physical pathway or meta-mechanism that connects different mechanisms. From this perspective, the independence of causal mechanisms guarantees the possibility of local interventions: one part of a system can be changed without affecting the other parts because mechanisms do not influence each other. From an information-theoretic point of view, the independence of causal mechanisms postulates that mechanisms do not carry information about each other; one cannot gain information about one mechanism by looking at the informational content of other mechanisms.

It is essential to differentiate between two levels of information in this context. While the effect variable is statistically dependent on the cause variable, the mechanism generating the effect is informationally independent of the mechanism generating the cause. For example, consider a scenario where X is the cause, and Y is the effect, i.e., $X \to Y$. Due to the causal connection between the two variables, X and Y are statistically dependent. However, the independence of causal mechanisms asserts that the conditional distributions

P_X and $P_{Y|X}$ are informationally independent because each distribution represents one causal mechanism, and causal mechanisms are independent of each other.

Therefore, to precisely define the independence of causal mechanisms, one needs to establish a formal notion of "informational independence" between mechanisms. Janzing & Schölkopf (2010) formalized this concept using Kolmogorov complexity and algorithmic mutual information. The idea is to represent each mechanism with a bit-string and define the Kolmogorov complexity of a mechanism as the length of the shortest program that generates the corresponding bit-string on a universal Turing machine. Two mechanisms are considered algorithmically or "informationally independent" if the length of the shortest description of the two bit-strings together is not shorter than the sum of the shortest individual descriptions. There are lots of technical details here, which I do not wish to review here. Instead, I draw some concluding remarks about the connections between the present discussion on the independence of causal mechanisms and the EPR-Bell scenario.

Firstly, the above discussion highlights that the independence of causal mechanisms can be transformed from a primitive ontological fact to an assumption that is quantifiable in different scenarios. For instance, in the EPR-Bell scenario, it is not necessary to assume that the causal mechanisms of the outcomes are necessarily independent; the validity of such independence can be quantified thanks to the tools provided by Machine Learning and information theory. In the next chapter, I propose a functional causal model that violates the independence of causal mechanisms and examine its predictive power in reproducing the statistics of the EPR-Bell scenario.

Secondly, similar to the strategy adopted by Cavalcanti (2018) for generalizing the CFC (see Section 3.2), a weakened version of the independence of causal mechanisms can be proposed for the EPR-Bell scenario: a causal model should not allow dependent mechanisms stronger than needed for reproducing the statistics of the EPR-Bell scenario. One can thus study the causal problem of entanglement beyond the dichotomy of relinquishing the CMC or the CFC.

Thirdly, the causal non-separability of entangled quantum states does not arise only through the non-localizability of states; the violation of the independence of causal mechanisms and hence the dependence between causal mechanisms can also bring about such non-separability.

Fourthly, further research is required to understand how to achieve the violation of the independence of causal mechanisms physically. One possibility is to assume the presence of a meta-mechanism that affects the causal mechanisms. The former meta-mechanism can be realized using correlated noises discussed in Section 3.1. Therefore, the weakening of the independence of causal mechanisms can be interpreted as a relaxation of the CMC.

3.4 The Intervention Assumption

An assumption often made in causal models proposed for the EPR-Bell scenario is that the settings are not affected by other variables, i.e., the settings are exogenous variables. This is known as the "intervention assumption" and is justified by the assertion that the experimenters have free will to select what to prepare and what to measure.

Rejecting the intervention assumption leads to models in which one or both settings are causally influenced by other variables, e.g., by the device preparing the entangled particles. This strategy is adopted by superdeterministic models. Figure 3.4 shows an example of a superdeterministic model in which the hidden variable μ affects not only the settings but also Λ, which itself is a hidden variable supposed to model the prepared state.

A well-known criticism of superdeterministic models is that they are often considered conspiratorial, as they propose mysterious mechanisms that restrict the free will of experimenters.[3]

However, I contend that the rejection of the intervention assumption does not necessarily entail a conspiracy, especially if the responsible mechanism is mandated by natural laws (even if these laws are yet to be discovered). The Pauli exclusion principle provides an example of how a physical law can produce correlations without any conspiracy. Specifically, two fermions can be correlated even if they have not interacted previously:

[3] There are more complaints against superdeterministic models; check Hossenfelder & Palmer (2020, sec. 4) for some of the most popular objections and responses.

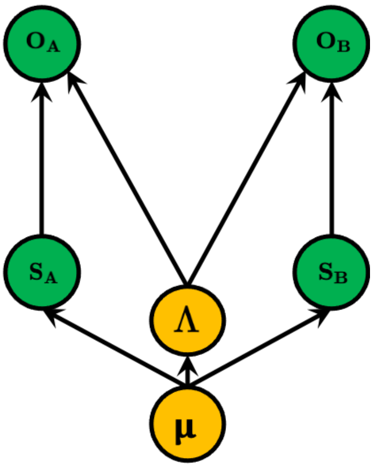


Figure 3.4: Violation of the intervention assumption: the setting variables are causally affected by other variables.

> ... there is a universal correlation of the EPR type which we do not have to cleverly set up, it is simply the total antisymmetrization of a many fermion state, which does correlate the electrons of my body with those of any inhabitant of the Andromeda galaxy (Lévy–Leblond 1985).

I conclude this section by stating that rejecting the intervention assumption is another valid option in the debate surrounding the causal problem of entanglement. The option would be more compelling when the laws of nature are responsible for violating the intervention assumption. In such a case, there is a stable fine-tuning whose viability can be justified by the arguments discussed in Section 3.2.

3.5 Causal Perspectivalism

This section points out that the standard causal modeling framework is incapable of addressing a set of causal models proposed for the EPR-Bell scenario. Such models can neither be supported nor rejected in light of the causal problem.

To begin, note that Evans (2018) proposed a retrocausal model reproducing the EPR-Bell statistics. At first, it seems that the model must be ruled out by the argument of Wood & Spekkens (2015) because of being retrocausal and hence violating the CFC. However,

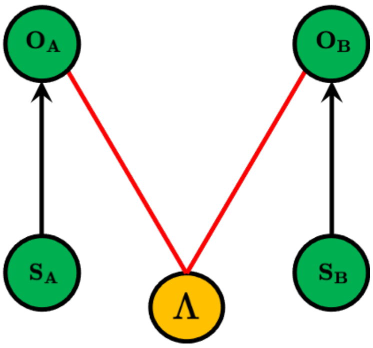


Figure 3.5: A model with undirected edges: dependency of the outcomes is nomological and not causal.

this is not a correct conclusion because the model cannot be characterized by the standard interventionist account of causation.

Figure 3.5 depicts a graph corresponding to Evans' retrocausal model. The graph represents a combination of causal and nomological dependencies, which is why some edges are directed, and others are undirected. Particularly, dependencies of the outcomes with the hidden variable are nomological, while dependencies of outcomes with the settings are causal. The undirected edges do not support surgical interventions, so the model cannot properly be described within the standard interventionist account underlying the causal modeling framework.

Nevertheless, Evans argues that it is possible to accommodate the above graph within the causal modeling framework if the interventionist account of causation is augmented by causal perspectivalism. By doing so, causation becomes an agent-dependent notion realized within a "block universe." Here, causation has nothing to do with changing certain aspects of the future or the past because all are equally fixed from an external perspective. Instead, because agents' epistemic constraints lead them to disagree about what should be taken as fixed and controllable, causation might be defined for each agent differently.

In such a proposal, observers in different frames of reference would assign different directions to the undirected edges in the above graph. But no one is wrong because there is no objective direction for nomologically related variables. For instance, Alice might consider a causal model in which her measurement outcomes influence Bob's outcomes, but at the same time, Bob considers a causal model in which his outcomes influence Alice's outcomes. Neither of the models is incorrect; the models are observer-dependent.

I conclude this section by saying that the standard causal modeling framework cannot accommodate all the proposed causal models for the EPR-Bell. As a result, no-go theorems derived within the standard framework have a limited scope in their conclusions. Of particular importance are models that postulate a nomological dependence between the two outcomes. These models evade the dichotomy of the CMC and CFC because the standard interventionist account of causation does not provide a natural framework for sketching such models.

3.6 Quantum Causal Modeling

This section investigates a fundamentally different approach to the causal problem of entanglement that is arisen from the literature on process matrix formalism and quantum causal models. According to this approach, instead of modifying one of the underlying assumptions in the framework of classical causal models, one needs to generalize the whole framework and equip it enough to deal with quantum systems. According to the present view, the take-home message of the EPR-Bell scenario is to redefine fundamental notions of the causal modeling framework, such as causal variables, mechanisms, and the notion of intervention.

> ... to establish an account of causality in quantum theory's own terms, without assuming a separate realm of classical systems or measurement outcomes. The account provides its own answer to many of the basic questions concerning causality in a quantum universe (Barrett et al. 2019).

The literature on quantum foundations contains various proposed methods for generalizing causal concepts to the quantum domain, and there is still no consensus on a single formulation of quantum causal models. Nevertheless, there is a common ingredient among all the proposed formulations of quantum causal models focusing on which facilitates understanding the essence of quantum causal models.

To make a start, note that a fundamental assumption in classical causal models is that the causal knowledge that can be inferred from a variable A using interventions on its causal parents is expressible in terms of a conditional probability distribution $p_{\mathrm{A|pa(A)}}$.

Such an assumption is challenged in quantum causal models by arguing that conditional probabilities are too restrictive to describe the informational content of quantum systems. In a quantum causal model, a quantum system is represented by two Hilbert spaces, one to represent the causal past of the system and the other to represent the causal future of the same system, such that both spaces together represent the set of all possible interventions that can be performed on the system. Under this assumption, the knowledge about a system given its causal parent is represented by an exotic mathematical object called "conditional density operator" $\rho_{\mathrm{A|pa(A)}}$ (Leifer 2006) or equivalently by a "completely positive map" (Costa & Shrapnel 2016).

From a mathematical point of view, conditional density operators provide richer algebraic structures compared to conditional probability distributions. For, the former allows the transition from commutative algebras to non-commutative algebras. In such a new framework, the concept of the joint probability distribution is replaced with a less intuitive notion called process matrix (see, e.g. Oreshkov et al. 2012), the concept of intervention takes a complex representation (see, e.g. Shrapnel 2019), the causal mechanism is represented by a unitary operator instead of a deterministic function (see, e.g. Barrett et al. 2019). More relevant to the present discussion is the screen-off condition in Reichenbach's principle, which finds its own translation in a quantum causal model. For example, Leifer & Spekkens (2013) showed that Reichenbach's principle holds in the quantum realm if the screen-off condition is about the factorizability of conditional density operators rather than conditional probability distributions.

Under these considerations, physicists have succeeded in formulating *causal* models that are applicable to large quantum networks and can be used for quantum engineering purposes. For instance, in a network consisting of a large number of interacting quantum

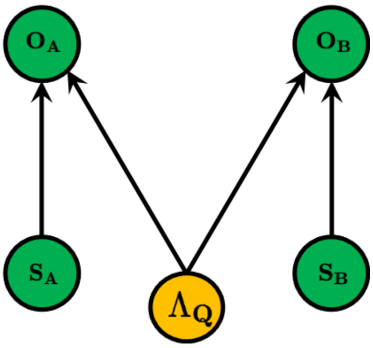


Figure 3.6: The causal graph of the EPR-Bell scenario proposed by a quantum common cause.

systems, a quantum causal model provides instructions on how to manipulate a set of systems such that a preferred effect can be brought about in a specific system. In other words, quantum causal models enable one to distinguish between effective and ineffective strategies. Furthermore, the process matrix formalism can delineate scenarios beyond non-relativistic Quantum Mechanics, such as those in which the global causal structure of a scenario is affected by laws of physics. Examples of this category are process matrices that are in a superposition of two or more certain causal structures or non-causal processes that are incompatible with any combination of causal structures and hence do not admits any causal explanation (see Branciard et al. 2015, Araújo et al. 2015). The process matrix has led to the experimental implementation of exciting examples such as the quantum switch, in which the causal order of a cause-effect scenario is in a superposition, i.e., a single event can be a cause and an effect of another event at the same time (see Rubino et al. 2017, Goswami et al. 2018).

3.6.1 Quantum Causal Models on the EPR-Bell Scenario

Quantum causal models have a straightforward causal story about the EPR-Bell scenario: there is no direct causation between the two wings of the experiment, but the two wings are under the influence of a common cause.

Figure 3.6 depicts the causal graph proposed by quantum causal models. In this graph, a “quantum” common cause brings about the observed correlations between the two outcomes. This implies that the peculiarity of the EPR-Bell scenario does not originate from the graphical structure (i.e., DAG) of the scenario. Rather, the peculiarity arises from the nature of the common cause, which is something more than a classical random variable. According to this view, the main lesson from the EPR-Bell scenario is admitting that quantum common causes require more complex mathematical machinery: conditional probabilities should be replaced by conditional operators to reproduce the statistics of Quantum Mechanics. It appears that two general types of models can reproduce the statistics of the EPR-Bell scenario: [4]

- models that are parametrically conservative but structurally radical
- models that are parametrically radical but structurally conservative

The first type includes classical causal models proposed usually in the quantum foundations literature. These models are "parametrically conservative" as they exploit classical random variables and conditional probability distributions to represent variables and mechanisms in a causal model. These models are "structurally radical" because they are willing to postulate weird causal structures such as superdeterministic and superluminal DAGs. The second type includes quantum causal models. These are "parametrically radical" as they use more complex objects to represent variables and mechanisms but are "structurally conservative" because their corresponding DAG is a simple common cause, depicted in Figure 3.6.

After explaining the fundamental concept of quantum causal models, the question arises: *to what extent can we consider the story presented by such models as "causal"?*

If we limit the concept of causation to the interventionist account of causation and argue that being causal means providing guidelines for distinguishing between effective and ineffective strategies, then the story can be considered causal. However, is it sufficient to reduce the notion of causation to a set of operational instructions about effective and ineffective strategies? Should not a causal story provide additional details about the phenomenon under investigation?

[4] This terminology is borrowed from Daley et al. (2022).

It is important to note that the quantum causal model's story about Bell's correlations may not provide new insights into the nature of entanglement, but it still contributes to our understanding of the phenomenon. Although the story does not address the causal ordering among entangled particles or why entanglement connections are attenuation-free, it is a step forward in our understanding of Quantum Mechanics and the causal relationships within it.

> ... what the quantum causal modelling framework tells us about the Bell correlations is simply what we already knew, i.e. that these correlations can be produced in the lab by an agent who prepares a joint entangled state in the causal past of both measurements (Adlam 2022, p. 18).

The above discussion can be investigated within a broader context by noting that the common cause explanation provided by quantum causal models does not naturally fit with the idea of local beables defended in Bell (1976). As mentioned in the previous chapter, the term beable can be construed as elements of a physical theory that refer to real objects in the world, in opposition to what is only observable. Local beables are those that can be assigned to some bounded region of spacetime, i.e., a local beable lives in a three-dimensional space or four-dimensional spacetime. For instance, the electric field in Maxwell's electromagnetic theory is an eligible candidate for being a local beable as it is a quantity whose values can be expressed by assigning to the spacetime coordinates of the region in which the field is defined. For Bell, the ultimate goal of local beables is to recast all local observables of a theory in terms of local beables:

> beables must include the settings of switches and knobs on experimental equipment, the currents in coils, and the reading of instruments. "Observables" must be made, somehow, out of beables. The theory of local beables should contain, and give precise physical meaning to, the algebra of local observables. (Bell 1976, p. 52)

Suppose the common cause postulated in the EPR-Bell scenario is to be a local beable. In that case, it must have a corresponding physical quantity living in a (3-dimensional or 4-dimensional) physical space such that its mathematical representation is expressible by assigning to spacetime coordinates of its defining region. Such a demand is respected by a *classical* common cause represented by a classical random variable and assigned to a particular region of spacetime. A quantum causal model seems incompatible with the local beable demand, for the "quantum" common cause has no corresponding physical quantity living in an ordinary physical space. In Shrapnel's words:

> ... if Bell's aim was to use beables in order to expunge the use of any mathematical representational devices that move beyond classical random variables, then clearly the quantum causal models fail in this respect (Shrapnel 2019, p. 20).

With this description, our previous question converts into the following question: *To what extent can a story whose constituent elements are not local beables be called "causal"?*

To answer the present question, we must ask what reasons we have to postulate the local beable condition in the first place. This is an important question, mainly because we already know that quantum systems exhibit puzzling behaviors that make it hard to believe that these systems can be represented by classical random variables. For example, the Kochen–Specker theorem suggests that, unlike classical systems, quantum systems admit no preexisting properties to be represented by classical random variables. It appears that Bell's motivation for demanding local beables roots back to the notion of "classical terms" in Niels Bohr's philosophy. In Bohr's view, classical terms are essential because they provide unambiguous language to communicate the results of experiments and observations:

> It is decisive to recognize that, however far the phenomena transcend the scope of classical physical explanation, the account of all evidence must be expressed in classical terms. The argument is simply that by the word "experiment" we refer to a situation where we can tell others what we have done and what we have learned and that, therefore, the account of the experimental arrangement and of the results of the observations must be expressed in unambiguous lan-

> guage with suitable application of the terminology of classical physics (Bohr 1949, p. 209).

Quoting the first sentence of the above paragraph, Bell explains his motivation for postulating local beables:

> It is the ambition of the theory of local beables to bring these “classical terms” into the equations, and not relegate them entirely to the surrounding talk (Bell 1976).

Hence, we cognitive agents require local beables to grasp the content of physical theories in a way that our brain is adapted, i.e., in terms of classical notions. If this is a correct characterization of the problem, then judging whether a quantum causal model provides a causal story highly depends on the theoretical grounds of that particular quantum causal model. If the quantum causal model in question provides a “secure transition” from the quantum domain to the classical domain, its story might be considered causal. Here, a secure transition is meant to be an explanation of the classical limit and the origin of quantumness in causal models. Such a strategy is adopted in some proposed quantum causal models (see, e.g. Costa & Shrapnel 2016); however, they are still far from a principled transition to be accepted as an “unambiguous language.”

I conclude this section by asserting that quantum causal models offer a novel tool for dealing with the causal problem of entanglement that inherently differs from traditional approaches. The extent to which quantum causal models succeed in providing causal stories about the quantum entanglement phenomenon depends highly on how they transit from classical causal models to quantum causal models.

Quantum causal models are compelling from the perspective of locality considerations, for they are structurally conservative; they can be, however, uncompelling from the perspective of scientific realism and classical terms, for they are parametrically radical. Opposite statements can be made about classical causal models. How to compare the two general approaches (i.e., classical and quantum causal models) is an interesting question that has attracted less attention. Nonetheless, Daley et al. (2022) has recently proposed a numerical technique in this direction.

3.7 Conclusion

The aim of this chapter was to shed new light on the causal problem of entanglement by examining the sources of conflict between the axioms of the theory of causal Bayesian networks and quantum correlations. To achieve this goal, the chapter explored various possibilities for resolving the causal problem and analyzed the arguments for and against each option, as well as their interrelationships.

The analysis presented in this chapter leads to several conclusions. First, resolving the causal problem of entanglement does not necessarily require abandoning either the CMC or the CFC assumptions. Alternative possibilities emerge when the underlying presumptions are precisely addressed. Second, each possibility can be implemented in various ways, leading to a new physical picture. For example, violating the CMC assumption can be achieved by adopting interactive common causes, separate common causes, or correlated noises. Third, not all possibilities presented in this chapter can adequately address the causal problem of entanglement. For instance, interactive common causes cannot reproduce quantum statistics, making them insufficient for resolving the causal problem. Fourth, the vast potential of data-driven numerical methods has not been adequately explored in the context of the causal problem, with only a few exceptions.

The next chapter will focus on an ML-based framework for adjudicating between different possibilities for resolving the causal problem, as mentioned in the first conclusion. Specifically, the chapter will explore models that violate the CFC, such as retrocausal, superdeterministic, and superluminal models, as well as models that violate the independence of causal mechanisms. Although evaluating all the possibilities sketched in this chapter is impossible due to the lack of translation for some options (e.g., causal perspectivalism), the methodological limitation of the framework should not reflect on the possibilities themselves.

Chapter 4

Machine Learning Analysis of EPR-Bell Correlations[1]

In the previous chapters, we noticed that standard causal modeling techniques cannot provide a causal explanation for quantum correlations. In particular, we noticed that the statistical pattern of dependencies and independencies in the EPR-Bell scenario is so weird that any discovery algorithm that assumes the CMC and CFC assumptions cannot provide a causal graph for the scenario. We referred to this puzzle as the causal problem of entanglement and examined various proposals for overcoming this puzzle in Chapter 3.

The present chapter offers an alternative way of tackling the causal problem, which is empirical in spirit. I exploit recent advances in Machine Learning (ML) to build a framework to adjudicate between various proposals for the causal problem empirically. To this end, I take a classical causal discovery algorithm called Causal Generative Neural Network (CGNN) and adapt it to the causal problem in light of philosophical considerations discussed previously. The key idea is to compare the *predictive power* of different candidate models and find the candidate that can generate data whose distribution is "closer" to the

[1] The present project is the result of my collaboration with **Jan Dziewior** from the Max-Planck-Institute for Quantum Optics, 85748 Garching, Germany; Faculty of Physics, LMU Munich, 80797 Munich, Germany. Jan Dziewior has two main contributions to the project: (1) simulating the quantum experiments, and (2) resolving the problem explained in Section 4.4. He has also written an early version of Section 4.2.1.
The source code of this project can be found here: `https://github.com/OmidCharrakh/Quantum_CGNN/`

original quantum data. How to generate data using a candidate model and how to quantify the distance between two distributions are details that will be discussed during the chapter. The essential point, however, is that the present framework can empirically compare a wide range of proposed models for the causal problem without getting involved in the CMC-CFC dilemma.

4.1 Introduction

The causal modeling literature contains a multitude of discovery algorithms (see Glymour et al. 2019, Guo et al. 2020, for overviews). However, the CGNN benefits from a set of features making it exceptionally suitable for the present work. In what follows, I point out these features and explain why they fit nicely with the philosophical considerations addressed in Chapter 3.

First, the CGNN does not require the CFC assumption for the task of causal discovery. Apart from the technical flexibility it brings, this feature has two foundational implications for the application of the CGNN to study the causal problem. On the one hand, it fulfills Cartwright's concern regarding the CFC (Cartwright 2001, p. 254). That is, "we should not proceed with false confidence having plumped either for or against" the CFC, especially when we are not certain about its validity in the context of quantum correlations. On the other hand, it enables the CGNN to evaluate causal models originating from interpretations of Quantum Mechanics that violate the CFC (e.g., superluminal, superdeterministic, and retrocausal).

Second, the evaluation of the CGNN is based on the entire (i.e., joint) distribution of a scenario, unlike conventional approaches to Bell-type inequalities where the focus is merely on the conditional distribution of the outcomes given the settings. From a statistical perspective, the joint distribution over the four variables contains much more information than the conditional distribution of the outcomes, and therefore, judgments based on the former are more reliable than judgments based on the latter. This feature additionally satisfies Wood and Spekkens' concern regarding the conditional-independence-based discovery algorithms. Recall that algorithms based on conditional independencies "do not do justice to Bell's theorem" because conditional "independencies simply do not provide enough information," and thus, one "needs a causal discovery algorithm that looks at the strength of correlations to reproduce Bell's conclusion" (Wood & Spekkens 2015, p. 17). The CGNN is an example of such an algorithm that looks not only at the strength of correlations but also at the entire distribution of the scenario.

Third, the outputs of the CGNN go beyond a binary decision of whether or not a particular graph is the causal graph of a probability distribution. The algorithm assigns a continuous score to a given candidate, where the score indicates the extent to which the candidate can reproduce the original statistics of a scenario. It is possible to obtain even further insights into a candidate by shedding light on the algorithm's black box and scrutinizing its outputs. One can answer questions concerning the extent to which a candidate can reproduce a particular marginal or conditional relation.

Fourth, the CGNN accomplishes the discovery task by relying on merely observational data, i.e., it does not require any interventional scheme. This feature is necessary for any discovery algorithm that is supposed to be applied to the causal problem, for the outcomes are "uncontrollable" in the EPR-Bell scenario. In addition, the algorithm has strategies for handling latent confounders (i.e., hidden variables), enabling one to evaluate candidates exploiting hidden variables.

Fifth, the CGNN offers significant flexibility in the functional form of mechanisms and the distribution of variables and noises, which sets it apart from many other discovery algorithms that rely on functional causal models (henceforth, FCMs) like LiNGAM, ANM, and PNL. This flexibility is due, in part, to the use of artificial neural networks to represent causal mechanisms. Neural networks are known for their ability to uncover complex patterns in statistical data, allowing the CGNN to learn intricate functions for causal mechanisms.

Despite all the enumerated capabilities of the CGNN, there is a challenge in utilizing the algorithm for quantum scenarios. The CGNN is designed to work with continuous probability distributions and does not apply to typical quantum scenarios wherein the settings and the outcomes are discrete variables. To apply the CGNN to the causal problem, we must go beyond the usual scenarios and simulate a Bell-type scenario wherein an entangled state is shared between continuous quantum systems such that both outcomes and settings are continuous.

The remainder of the chapter is organized as follows. Section 4.2 explains how to formulate and simulate a continuous EPR-Bell scenario as well as the statistical relations holding in such a scenario. Section 4.3 describes the internal structure of the CGNN algorithm and why it is necessary to insert specific changes into it to be applicable to the causal problem of entanglement. Section 4.4 describes different criteria we use for distance quantification between two probability distributions and how to relate these criteria to the notion of the loss function. Section 4.5 discusses the candidates evaluated in this chapter and how to tune their hyperparameters. Section 4.6 presents the adjudication results by sketching several visualizations. Section 4.7 concludes.

4.2 Physical Scenario

To utilize the CGNN for studying the causal problem, it is necessary to go beyond the conventional formulation of the EPR-Bell scenario and find a continuous variant. In what follows, I briefly review the conventional formulation (i.e., the CHSH scenario) and then explain how to generalize it.

4.2.1 Formulation and Simulation

Conventional

A conventional Bell-type scenario is a two-wing scenario (wings A and B) wherein the wings are spacelike-separated, each containing a physical system. The statistics of such a scenario is a joint distribution over outcomes of measurements on the two systems, conditioned on the measurement settings applied at each wing. In the most basic setup, the two outcomes O_A and O_B, and the two settings S_A and S_B are binary-valued variables.

The physical realization of the said scenario consists of two spatially separated qubits, the most simple form of a system capable of having quantum characteristics. To generate the sought-after non-classical probability distributions, the two systems are taken to be in an entangled state $|\Psi\rangle$ with

$$|\Psi\rangle = \frac{1}{\sqrt{2}} \left(|00\rangle + |11\rangle \right), \tag{4.1}$$

where $|jk\rangle := |j\rangle \otimes |k\rangle$, with the two kets on the right-hand side belonging to the Hilbert spaces of the first and second qubits, respectively, and $\{|0\rangle, |1\rangle\}$ denoting bases of both Hilbert spaces.

Switching between non-commuting observables for measurements at the two wings is necessary to obtain the desired distributions. A possible representation of the measured observable σ_k in accordance with the setting parameter s_k is given by

$$\sigma_k = \cos s_k \, \sigma_x + \sin s_k \, \sigma_z, \tag{4.2}$$

where σ_x and σ_z are two Pauli-matrices corresponding to canonical observables on qubits. Following the proposal by Clauser et al. (1969), recent Bell-tests choose two values each for s_A and s_B to achieve a maximal violation of the CHSH inequality.

Continuous Variant

The scenario described above is suited for theoretical analysis and experimental implementation due to its clarity and simplicity. For our causal analysis, we require a scenario wherein the outcomes and the settings have continuous probability distributions. Such a generalization must contain the necessary features distinguishing quantum mechanical phenomena from classical ones. To this end, turning towards physical systems described by continuous Hilbert spaces is necessary.

In current experiments, the most common approach to Bell-type scenarios with continuous-variable systems involves optical modes, where the infinite-dimensional discrete photon number space is parametrized via continuous quadrature variables Wenger et al. (2003), Brask et al. (2012), Thearle et al. (2018). As the primary goal of these experiments is to detect the violation of Bell inequalities, the measurements realize a binning resulting in probability distributions over binary-valued outcomes.

We follow an approach similar to an experiment by Howell et al. (2004), where the transversal position and momentum distributions of entangled photons are observed. We pick up this scenario in our simulation by considering a pair of continuous systems and allowing for a gradual transition between measuring a position ($\mathbf{x}$) or a complementary observable ($\mathbf{k}$) on both systems, where $[\mathbf{x}, \mathbf{k}] = i$. Thus, the effective observables $\mathbf{o_A}$ and $\mathbf{o_B}$ measured at the two wings can be represented by the following operator, in which s_j denotes values for the settings chosen from the interval $[0, 1]$:

$$\mathbf{o_j} = \cos(s_j \pi/2)\mathbf{x} + \sin(s_j \pi/2)\mathbf{k}. \tag{4.3}$$

Equation 4.3 allows the experimenters to choose their setting variables from a continuous range of values. Here, $s = 0$ and $s = 1$ correspond to applying position and momentum measurements. In addition, there is a continuous range for the outcome values associated with each measurement choice.

In the conventional Bell-type scenario, the outcomes can be marginally independent of the settings when the reduced state at each wing is maximally mixed:

$$O_A \perp\!\!\!\perp_{\mathbf{s}} (S_A, S_B) \quad \& \quad O_B \perp\!\!\!\perp_{\mathbf{s}} (S_A, S_B) \tag{4.4}$$

We want to preserve these independencies in our simulation. However, since there is no analog to a maximally mixed state in continuous-variable systems, the choice of the quantum state shared by the two wings is non-trivial. We adopt a joint state that yields Gaussian marginal distributions such that the width of the marginals does not depend on the settings. This leads to the entangled state $|\Phi\rangle$ represented by the wavefunction

$\Phi(x_A, x_B)$ as

$$\Phi(x_A, x_B) := \langle x_A x_B | \Phi \rangle = \frac{1}{\sqrt{\pi}} \exp\left(-\frac{x_A^2 + x_B^2 - 2c x_A x_B}{2\sqrt{1-c^2}} \right) \tag{4.5}$$

with the position bases $\{|x_A\rangle\}_{x_A}$ and $\{|x_B\rangle\}_{x_B}$ of the two wings and the free parameter $c \in [0, 1[$. The parameter c corresponds to the correlation between the spatial observables $\mathbf{x_A}$ and $\mathbf{x_B}$ and satisfies the relation

$$c = \frac{\langle \mathbf{x_A} \mathbf{x_B} \rangle}{\Delta \mathbf{x_A} \Delta \mathbf{x_B}}, \tag{4.6}$$

which yields the standard definition of statistical correlation, written in Dirac notation. Also, $\Delta \mathbf{x_A}$ and $\Delta \mathbf{x_B}$ denote the standard deviations of the respective marginal distributions.

The parameter c controls the degree of entanglement between the two systems. For the case of $c = 0$, the two systems are entirely disentangled, while they have nonvanishing entanglement otherwise. For the ideal case of $c = 1$, a strong correlation is observed when measuring $\mathbf{x}$ at both wings, whereas a strong anti-correlation appears when measuring $\mathbf{k}$ at both wings, and the correlation disappears when measuring $\mathbf{x}$ at one wing and $\mathbf{k}$ at the other. For our simulation, we use a c close to but smaller than 1 because a perfectly correlated state (i.e., $c = 1$) might be unphysical.

The physical model described above yields a conditional distribution over the outcomes O_A and O_B given the settings S_A and S_B, i.e., $p(O_A, O_B | S_A, S_B)$. It is possible to calculate a closed mathematical formula for this conditional distribution in the eigenbases of the observables in 4.3 by taking the partial Fourier transform of the state $|\Phi\rangle$ in 4.5. This leads to the following equation:

$$p(O_A = o_A, O_B = o_B | S_A = s_A, S_B = s_B) = \frac{k_1}{\pi k_2} \exp\left(-\frac{k_1}{{k_2}^2} (o_A^2 + o_B^2 - 2c_{\text{eff}} o_A o_B) \right), \tag{4.7}$$

where $k_1 := \sqrt{1 - c^2}$ and $k_2 := \sqrt{\frac{2-c^2(1+\cos(\pi(s_A+s_B)))}{2}}$ are real parameters and $c_{\text{eff}} = c\cos\big((s_1 + s_2)\pi/2\big)$ is the effective correlation value.

Remark 4.1. *Equation 4.7 implies that if the settings S_A and S_B take uniform distributions, then the outcomes O_A and O_B take Gaussian distributions.*

We use Equation 4.7 to simulate the joint distribution of the scenario, i.e., $p(O_A, O_B, S_A, S_B)$. To this end, we posit uniform distributions over the setting variables and obtain the outcomes. A sample data from the joint distribution is then the concatenation of these four variables' values.

4.2.2 Statistical Properties

Having explained the physical grounds of our simulation, I now focus on some statistical properties of our simulated data that are important from a causal discovery perspective. To this end, I first provide a characterization of different types of relations holding in our scenario and then present a set of visualizations representing the statistical properties at different levels. Aside from providing deeper insights into our data, these visualizations shall be construed as references to evaluate the predictive power of different candidates.

Dependencies and Independencies

Discovery algorithms check the statistical pattern of dependencies and independencies to estimate the causal structure of a scenario. Thus, it is important to explore the (un)conditional (in)dependence relations holding in a scenario. As the number of these relations can be very large in general, discovery algorithms usually restrict their attention to the conditional independence relations required for checking the d-separation criterion. We know conditional independencies alone do not suffice for causal discovery in the EPR-Bell scenario. So, I go beyond the d-separation criterion and present characterization of all types of relations holding in our scenario. Such a characterization will be used in the coming sections to build a heuristic loss function and evaluate the learning quality of the candidates.

Type	Form	Number	Example
$(1,0)$	X_i	4	O_A
$(1,1)$	$X_i\|X_j$	12	$O_A\|O_B$
$(1,2)$	$X_i\|X_jX_k$	12	$O_A\|(O_B,S_A)$
$(1,3)$	$X_i\|X_jX_kX_l$	4	$O_A\|(O_B,S_A,S_B)$
$(2,0)$	X_iX_j	6	(O_A,O_B)
$(2,1)$	$X_iX_j\|X_k$	12	$(O_A,O_B)\|S_A$
$(2,2)$	$X_iX_j\|X_kX_l$	6	$(O_A,O_B)\|(S_A,S_B)$
$(3,0)$	$X_iX_jX_k$	4	(O_A,O_B,S_A)
$(3,1)$	$X_iX_jX_k\|X_l$	4	$(O_A,O_B,S_A)\|S_B$
$(4,0)$	$X_iX_jX_kX_l$	1	(O_A,O_B,S_A,S_B)

Table 4.1: Different types of conditional relations in a four-variable scenario. In sum, there are 65 relations.

Definition 4.1. *Let* $\mathbf{V} = \{V_1, \ldots, V_p\}$ *be a set of variables, and* $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ *be two disjoint subsets of* $\mathbf{V}$. *A* ***conditional relation*** *on* $\mathbf{X} = \{X_1, \ldots X_{p_x}\}$ *given* $\mathbf{Y} = \{Y_1, \ldots Y_{p_y}\}$ *is a statistical relation between the variables in* $\mathbf{X}$ *given the variables in* $\mathbf{Y}$ *and is denoted by*

$$\mathrm{R}_{\mathbf{X}|\mathbf{Y}} := (X_1 ... X_{p_x})|(Y_1 ... Y_{p_y}). \tag{4.8}$$

In the above definition, p_x and p_y can be used to characterize $\mathrm{R}_{\mathbf{X}|\mathbf{Y}}$ and assign a “type” to such a relation, where p_x and p_y are the numbers of conditioned and conditioning variables respectively. I will henceforth use (p_x, p_y) to refer to the type of the conditional relation $\mathrm{R}_{\mathbf{X}|\mathbf{Y}}$. The total number of conditional relations that can be defined on a set of p variables is

$$\sum_{p_y=0}^{p-p_x} \sum_{p_x=1}^{p} \binom{p}{p_x}\binom{p-p_x}{p_y}. \tag{4.9}$$

In the EPR-Bell scenario, four observed (non-hidden) variables exist, and the total number of relations to be investigated is equal to 65. Table 4.1 summarizes the type of these relations. Of particular importance are the six relations of type $(2,2)$, which describe the conditional behavior of each pair of variables given another pair. These relations are as

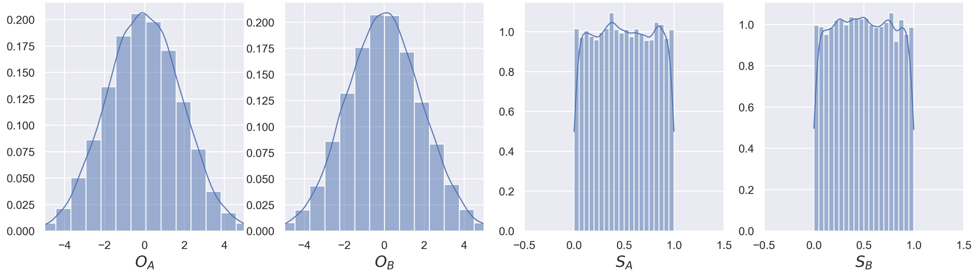


Figure 4.1: Marginal distributions of the four variables: the outcomes have Gaussian distributions, while the settings have uniform distributions.

follows:

$$\begin{aligned}&\{(O_A,O_B)|(S_A,S_B),\;\;(O_A,S_A)|(O_B,S_B),\;\;(O_A,S_B)|(O_B,S_A),\\&(O_B,S_A)|(O_A,S_B),\;\;(O_B,S_B)|(O_A,S_A),\;\;(S_A,S_B)|(O_A,O_B)\}.\end{aligned}\tag{4.10}$$

Marginal Distributions

In Table 4.1, type $(1,0)$ refers to univariate relations, i.e., how each variable behaves irrespective of other variables. The marginal distribution of one of the outcomes is an example of such behavior. Figure 4.1 depicts this behavior for our simulated data. The figure confirms that the settings S_A and S_B have uniform distributions, while the outcomes O_A and O_B have normal distributions.

Marginal Dependencies

In Table 4.1, type $(2,0)$ refers to bivariate relations among each pair of variables irrespective of other variables. The marginal dependence between the two outcomes is an instance of such a relation. Figure 4.2 evaluates such relations for our simulated data by exploiting four dependency criteria: Pearson correlation coefficient, Spearman's rank correlation coefficient, Kendall rank correlation coefficient Abdi (2007), and adjusted mutual information Vinh et al. (2010). The figure confirms that the simulated data exhibit no dependency of type $(2,0)$, i.e., all pairs of variables are marginally independent.

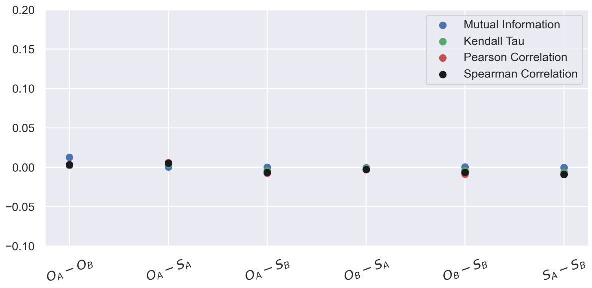


Figure 4.2: Marginal independencies

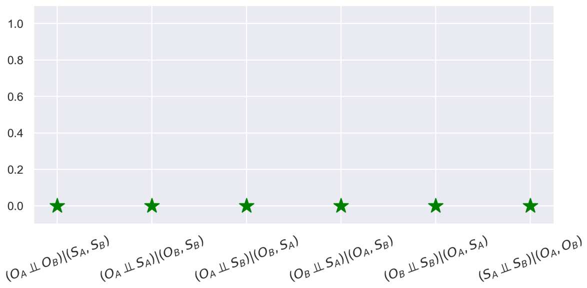


Figure 4.3: Conditional independence tests

Conditional Dependencies

In Table 4.1, type $(2, 2)$ refers to bivariate relations among each pair of variables conditioned on the remaining pair. The conditional dependence of the two outcomes given the two settings is an instance of such a relation. There are various approaches to investigating such relations. For example, Figure 4.3 evaluates these relations within the framework of statistical hypothesis testing, while Figure 4.4 pursues the same goal through the heuristic notion of "correlation matrix." I discuss both approaches in the following.

Conditional Independence Tests

In the framework of statistical hypothesis testing, where one pits a null hypothesis H_0 against an alternative hypothesis H_1 and exploits the notion of the p-value. A suitable pair of hypotheses for testing the conditional independence between two variables X and Y given a third Z is $\mathrm{H}_0 : X \perp\!\!\!\perp_{\mathbf{s}} Y|Z$ versus $\mathrm{H}_1 : X \not\perp\!\!\!\perp_{\mathbf{s}} Y|Z$. A sufficiently small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis and implies the conditional dependence $X \not\perp\!\!\!\perp_{\mathbf{s}} Y|Z$, while a large p-value fails to reject the hypothesized independence $X \perp\!\!\!\perp_{\mathbf{s}} Y|Z$. In practice, there are different conditional independence tests; the results presented here are based on the "Fast Independence Test" method proposed in Chalupka et al. (2018). Figure 4.3 depicts the p-values of the six conditional independence tests performed on the simulated data. The figure confirms that all pairs of variables are conditionally dependent, given the remaining pair.

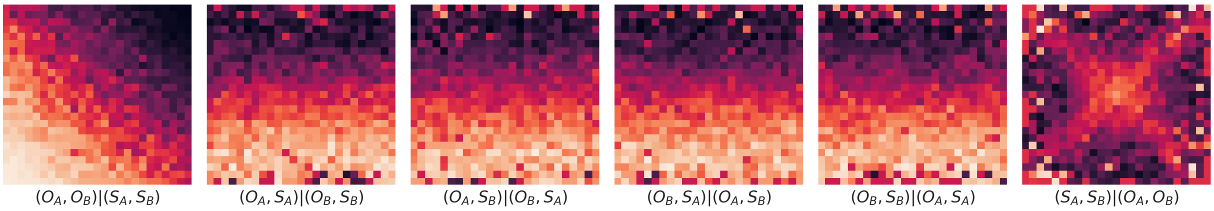


Figure 4.4: Heatmaps associated to the simulated data.

Correlation Matrices and Heatmaps

Figure 4.4 provides an alternative view on relations of type $(2, 2)$ in the simulated data. The six plots in this figure are heatmaps picturing the entries of a matrix that I refer to as a "correlation matrix". To understand these plots, one must first understand the notion of a correlation matrix.

Let $\mathbf{X} = \{X_1, X_2\}$ and $\mathbf{Y} = \{Y_1, Y_2\}$ be two sets of continuous variables. A "correlation matrix" $M_{\mathbf{X}|\mathbf{Y}}$ is a matrix whose elements represent the correlation coefficients between X_1 and X_2 in different partitions of the joint distribution of Y_1 and Y_2. To compute the matrix, each of the conditioning variables (i.e., Y_1 and Y_2) is binned into m bins. Then, the correlation coefficients between the conditioned variables (i.e., X_1 and X_2) are computed for each bin. Algorithm 1 summarizes the said procedure.

Algorithm 1 Construction of a Correlation Matrix

```
Input: $(X_1, X_2, Y_1, Y_2)$ (a sample from joint distribution, where $X$s are conditioned variables and $Y$s are conditioning variables)
Bin the $(Y_1, Y_2)$ into $m^2$ bins $\{\mathcal{B}_{ij}(Y_1, Y_2)\}$
for each bin do
    Find $X_1^{(ij)}$ and $X_2^{(ij)}$ via conditioning $X_1$ and $X_2$ on the bin
    Compute the correlation coefficient $c_{ij}$ between $X_1^{(ij)}$ and $X_2^{(ij)}$
    Set $\mathrm{M}_{\mathbf{X}|\mathbf{Y}}[i, j] = c_{ij}$
end for
Output: $\mathrm{M}_{\mathbf{X}|\mathbf{Y}}$ (correlation matrix)
```

Hence, the dimension of the correlation matrix is $m \times m$, and its elements are in the interval $[-1, +1]$. The matrix not only reveals whether two variables are conditionally dependent, but it also describes how intensely the two variables are correlated in different ranges of the conditioning variables.

Figure 4.4 visualizes the six correlation matrices computed for the simulated data. In this figure, the positivity and negativity of the matrices' elements are displayed by the gradient of colors in the heatmap: darker colors correspond to values closer to minus one, while lighter colors correspond to values closer to plus one. I focus on the first plot and explain what it represents; the other plots can be interpreted similarly.

The first plot shows the conditional dependence of the two outcomes given the two settings. The x-axis and y-axis of this plot correspond to the values of S_A and S_B; hence, each axis is in the interval $[0, 1]$. The plot indicates that the outcomes are conditionally dependent given the settings: if both settings are close to zero (i.e., if position measurements are performed at both wings), the outcomes will be perfectly correlated. In contrast, if both settings are close to one (i.e., if momentum measurements are performed at both wings), the outcomes will be anti-correlated.

Discovery Algorithms

Before discussing the CGNN algorithm, it is instructive to check the predictions of some traditional discovery algorithms when applied to our simulated data. Figure 4.5 depicts the causal graphs predicted by six popular discovery algorithms, namely the PC (constraint-based) Spirtes et al. (2000), GES (score-based) Chickering (2002), MMPC (hybrid) Tsamardinos, Aliferis & Statnikov (2003), IAMB (score-based) Tsamardinos, Aliferis, Statnikov & Statnikov (2003), GS (score-based) Margaritis (2003), and LiNGAM (FCM-based) Shimizu et al. (2006). Except for the LiNGAM algorithm, all the algorithms predict that there is no causal connection between the four variables.

4.3 The CGNN Algorithm

The CGNN is a powerful discovery algorithm that combines lots of advanced techniques from other discovery algorithms. This section introduces the internal structure of the algorithm and explains the steps it follows to evaluate a candidate model. It is worth

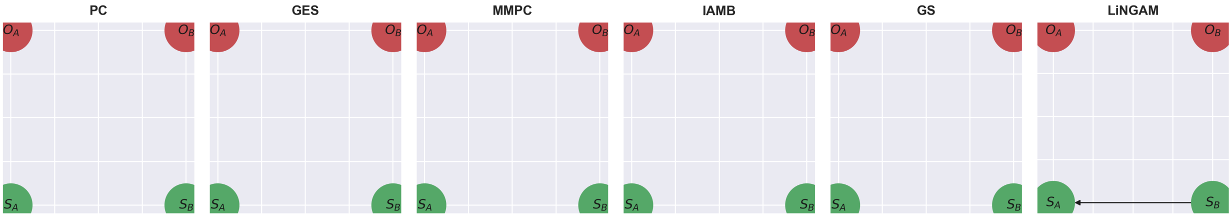


Figure 4.5: The graphs predicted by six discovery algorithms. Almost all the algorithms predict that the variables are causally independent.

clarifying how the CGNN is linked to causal modeling and Machine Learning frameworks.

- The CGNN is "FCM-based": it posits a set of deterministic functions to represent the causal mechanisms of a scenario. Nonetheless, the algorithm handles non-deterministic scenarios because it postulates non-deterministic noise terms along with deterministic functions.

- The CGNN is "pairwise-based": it is obtained from a graphical generalization of a pairwise algorithm with a similar name. The pairwise variant of the CGNN is tailored to distinguish cause from effect, i.e., to distinguish between $X \rightarrow Y$ and $X \leftarrow Y$.

- The CGNN is "generative": it generates synthetic data for a given candidate. The data represents the joint distribution induced by the candidate.

- The CGNN is "score-based": it assigns a score to a given candidate where the score indicates the quality of the candidate's synthetic data.

- The CGNN is "ML-based": it uses artificial neural networks to estimate the causal mechanisms of a scenario.

4.3.1 A Typical Scenario

Let $\mathbf{X} = \{X_1, \ldots, X_p\}$ be a set of one-dimensional continuous variables which are causally related by an underlying causal graph $\mathcal{G}_{true}$. Given an "observational" sample data $\mathscr{D}$ from the joint distribution $\mathbb{P}_{\mathbf{X}}$ and a set of candidate graphs $\mathbf{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_s\}$, the goal is to

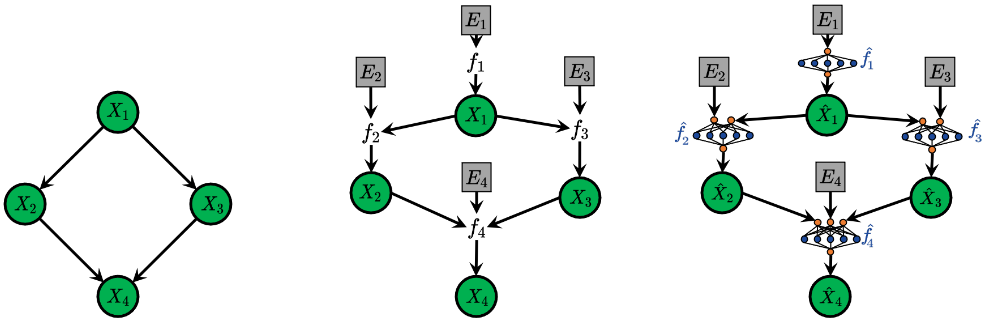


Figure 4.6: The CGNN uses generative neural networks to estimate causal mechanisms. Left: candidate graph, middle: candidate FCM, and right: CGNN's estimation.

discover the graph which is more probable to be $\mathcal{G}_{true}$. The CGNN is designed to deal with such a scenario and solves the problem in several steps.

For now, let us assume the scenario under consideration involves no latent variable, i.e., the "causal sufficiency" assumption holds. Therefore, the scenario involves p variables interacting causally. An FCM for such a scenario postulates p noise terms $\{E_i\}$ and p functions $\{f_i\}$ to represent the causal mechanisms of the variables. In other words, f_i shows how the variable X_i is generated from its causal parents:

$$X_i := f_i(X_{\text{pa}(X_i;\mathcal{G})}, E_i). \tag{4.11}$$

Figure 4.6 depicts an example of such a scenario. The first figure on the left shows a candidate causal graph over the four variables, and the second figure shows the corresponding FCM. The structural equations postulated by this FCM are as follows:

$$\begin{aligned} X_1 &= f_1(E_1) \\ X_2 &= f_2(E_2, X_1) \\ X_3 &= f_3(E_3, X_1) \\ X_4 &= f_4(E_4, X_2, X_3) \end{aligned} \tag{4.12}$$

The CGNN models the said scenario by estimating the functional forms of the causal mechanisms and modeling the noise terms. To model the noise terms, the algorithm draws samples from a continuous probability distribution such as Gaussian. To estimate the mechanisms, it exploits neural networks, i.e., the functions f_i are estimated by neural networks. Once the estimation is completed, the algorithm generates data in accordance with the topological order of the candidate causal graph:

$$\hat{X}_i = \hat{f}_i(\hat{X}_{\mathrm{pa}(X_i;\mathcal{G})}, E_i). \tag{4.13}$$

Before explaining how this can be done systematically, I wish to make a few clarifying remarks.

- The CGNN evaluates the candidates in $\mathbf{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_s\}$ one by one by assigning a score to each of them. The score assigned to a candidate represents the predictive power of that candidate, i.e., the candidate's capability to reproduce the ground truth data.

- The set of candidates represents the "search space" of the algorithm. In principle, the search space must contain all possible directed acyclic graphs (henceforth, DAGs) definable over the variables in the scenario. However, evaluating all definable DAGs is impossible, given the possibility of infinite latent variables. Even without latent variables, estimating a causal graph from observational data is computationally expensive because the number of possible DAGs super-exponentially grows with the number of nodes (Robinson 1977). For instance, the number of possible DAGs for a scenario with 2, 3, 4, and 5 variables is 3, 25, 543, and 29281, respectively.

- The CGNN does not have an inherent method for the search space restriction, and the user is supposed to provide a set of probable candidates to be evaluated by the algorithm. Nonetheless, there exist some general strategies for restricting the search space. For instance, before applying the CGNN, one can apply a constraint-based algorithm and restrict the CGNN's search space to edges about which the latter algorithm is uncertain.

- Regardless of how the search space is defined, the CGNN scoring workflow for a candidate remains the same. Therefore, I focus on only one candidate and explain how the CGNN assigns a score to it.

4.3.2 Algorithm Steps

Algorithm 2 summarizes the steps carried out during the CGNN scoring workflow. The algorithm takes two ingredients as input: (1) a sample dataset from the joint distribution of a scenario and (2) a candidate DAG to be evaluated by the algorithm. The output is the score of the given candidate. The supplementary comments of each step are indicated by the letter c and are explained in the following paragraphs.

Algorithm 2 The CGNN Scoring Workflow

```
Input: $\mathscr{D}$ (ground truth data) and $\mathcal{G}$ (candidate DAG)
Sort out $\mathscr{D}$ with respect to the topological order of $\mathcal{G}$          ▷ c1
Split $\mathscr{D}$ into training, validation, and test subsets                        ▷ c2
Initialize an FCM $\mathscr{F}_\Theta$ with model parameters $\Theta$                  ▷ c3
while in training phase do                                                             ▷ c4
    Generate synthetic data $\hat{\mathscr{D}}_\Theta$ by the model $\mathscr{F}_\Theta$   ▷ c5
    Quantify the training loss $\mathscr{L}_{tr}(\Theta) = \boldsymbol{\mathcal{M}}_{tr}(\mathscr{D}_{tr}, \hat{\mathscr{D}}_\Theta)$   ▷ c6
    Quantify the validation loss $\mathscr{L}_{va}(\Theta) = \boldsymbol{\mathcal{M}}_{va}(\mathscr{D}_{va}, \hat{\mathscr{D}}_\Theta)$   ▷ c6
    Update $\Theta$ by minimizing the training loss                                    ▷ c7
    if early stopping is satisfied then                                                ▷ c8
        Terminate the training phase
    else
        Continue
    end if
end while
while in test phase do                                                                 ▷ c9
    Generate synthetic data $\hat{\mathscr{D}}_{\Theta^\star}$                         ▷ c5
    Quantify the test loss $\mathscr{L}_{te} = \boldsymbol{\mathcal{M}}_{te}(\mathscr{D}_{te}, \hat{\mathscr{D}}_{\Theta^\star})$   ▷ c6
end while
Compute the final score $S = -\bar{\mathscr{L}}_{te}$                                  ▷ c10
Output: $S$ (candidate score)
```

c1. Topological Ordering

The very first step is to sort the data $\mathscr{D}$ by the topological ordering of the graph $\mathcal{G}$. The topological ordering is found iteratively by first finding the source nodes (i.e., nodes without parents), followed by the children of the source nodes, followed by the children of the children of the source nodes until the algorithm reaches the sink nodes (i.e., nodes without children). For instance, the topological ordering of the graph in Figure 4.6 is $[X_1, X_2, X_3, X_4]$ or $[X_1, X_3, X_2, X_4]$.

c2. Splitting Data

Data splitting is a common technique for training and evaluating Machine Learning models. The technique involves dividing a dataset into separate subsets, each serving a different purpose. Typically, the data is split into three disjoint subsets: training, validation, and test sets. Specifically, the data can be represented as follows:

$$\mathscr{D} = (\mathscr{D}_{tr}, \mathscr{D}_{va}, \mathscr{D}_{te}). \tag{4.14}$$

c3. Initializing Model

As mentioned earlier, the CGNN postulates an FCM over the variables of a scenario and uses generative neural networks to estimate the functional form of the causal mechanisms. Thus, for a candidate graph $\mathcal{G}$ with p variables, the CGNN initializes an FCM $\mathscr{F}$ composed of p neural networks. The initialized model has a large number of free parameters, the values of which are set randomly at this step. These parameters, denoted by Θ, are the "weights" and the "biases" of the neural networks and are to be optimized during the training phase.

c4. Training Phase

To train an ML model is to optimize its parameters in light of the training data. The optimization is an iterative process, i.e., it is done in several repetitive steps known as "epochs". In each epoch, $\mathscr{D}_{tr}$ is broken down into small data "batches." This is done to overcome the issues that could arise due to storage limitations of a computer system, especially when $\mathscr{D}_{tr}$ is large. The first training epoch ends when all data points in $\mathscr{D}_{tr}$ are used exactly once. At this stage, the validation phase starts in which the model performance is evaluated in light of $\mathscr{D}_{va}$. Once done, the subsequent training epoch starts.

The training phase can be repeated as many times as the user specifies, i.e., the number of training epochs is a hyperparameter that can be selected by the user. It is important to set a suitable value for these parameters because a too-large and a too-small number of training epochs can respectively lead to overfitting and underfitting, which are both problematic and unwanted. To avoid such problems, one can either (1) find a suitable value for the number of training epochs before starting the algorithm or (2) monitor the model performance on both $\mathscr{D}_{tr}$ and $\mathscr{D}_{va}$ and terminate the training phase if there is a risk of overfitting. We used the second strategy in the current project by implementing an early-stopping algorithm.

c5. Synthesizing Data

In each training epoch, the model $\mathscr{F}_{\Theta}$ is used to generate synthetic data. Assuming the neural estimator of the i^{th} mechanism has one hidden layer with n_h hidden units, the estimation of X_i takes the following form:

$$\hat{X}_i = \hat{f}_i\left(\hat{X}_{\text{pa}(X_i;\mathcal{G})}, E_i\right) = \sum_{k=1}^{n_h} \bar{w}_k^i \boldsymbol{\sigma}\left(\sum_{j \in \text{pa}(X_i;\mathcal{G})} w_{jk}^i \hat{X}_j + w_k^i E_i + b_k^i\right) + \bar{b}^i \tag{4.15}$$

where $\hat{f}_y$ and $\hat{Y}$ are respectively the neural estimations of the true mechanism f_y and the true variable Y. Moreover, $\boldsymbol{\sigma}$ is the "activation function", $\bar{w}_k, \hat{w}_k, w_k \in \mathbb{R}$ are the "weights," and $b_k, \hat{b} \in \mathbb{R}$ are the "biases" of the neural network.

Remark 4.2. *Equation 4.15 formulates a standard regression problem wherein the goal is to estimate a target (i.e., X_i) in terms of a set of features (i.e., $\hat{X}_{\text{pa}(X_i;\mathcal{G})}$ and E_i). Hence, in principle, the CGNN could use other regression models instead of neural networks.*

Remark 4.3. *The nonlinearity of the activation function allows the CGNN to estimate its mechanisms with the nonlinear equation $Y = f(X, E)$. From the perspective of regression analysis, the CGNN is more general than algorithms such as LiNGAM and ANM, which postulate more restrictive forms for causal mechanisms. The said models postulate $Y = \alpha X + \beta + E$ and $Y = f(X) + E$, respectively. Such a generality comes at the price of non-identifiability, especially if one does not limit the number of hidden layers and hidden units. I will return to this point later.*

Synthesizing data is carried out not only in the training phase but also in the test phase. The rationale of the test phase remains unchanged, except that the weights and biases are already optimized in the test phase and are no longer updated.

c6. Quantifying Loss

Loss quantification is a crucial step carried out in all three phases, i.e., the training, validation, and test. A loss function, or a cost function, is a function that maps an event onto a real number representing a "cost" associated with the event. The CGNN seeks a candidate that generates the most "close" synthetic data $\hat{\mathscr{D}}$ to the original data $\mathscr{D}$. Thus, the *cost* is defined as the distance between the joint distributions of $\hat{\mathscr{D}}$ and $\mathscr{D}$. To measure the distance, one needs a "distance measure" $\boldsymbol{\mathcal{M}}$, a function that estimates the distance from the two datasets.

The loss quantification differs depending on the learning phase (i.e., training, validation, or test). In the training phase, the synthetic data $\hat{\mathscr{D}}$ is compared with the training data $\mathscr{D}_{tr}$ and the training loss is computed by the distance measure $\boldsymbol{\mathcal{M}}_{tr}$, i.e., $\mathscr{L}_{tr}(\Theta) = \boldsymbol{\mathcal{M}}_{tr}(\mathscr{D}_{tr}, \hat{\mathscr{D}}(\Theta))$. Similarly, the validation and test losses are quantified using their corresponding datasets and distance measures.

Remark 4.4. *In the Machine Learning literature, it is customary to take a single distance measure for all three phases, i.e., $\boldsymbol{\mathcal{M}}_{tr} = \boldsymbol{\mathcal{M}}_{va} = \boldsymbol{\mathcal{M}}_{te}$. That is, the model is tested by the same criterion with which it has been trained and validated. The original CGNN follows this custom and uses the "Maximum Mean Discrepancy" (Gretton et al. 2006) for all the* *three phases. We abandon this convention and use different measures for reasons that will be clarified later.*

c7. Updating Parameters

The training phase of an ML model is nothing but solving an optimization problem wherein the goal is to minimize the loss function. The theoretical ground of such an optimization is based on the principle of "empirical risk minimization." In artificial neural networks, the optimization is achieved by the "stochastic gradient descent" and "backpropagation" algorithms. I do not intend to get involved with the technical details of Machine Learning, so I merely sketch the overall procedure.

In each training epoch, the gradient of the training loss $\mathscr{L}_{tr}(\Theta)$ is calculated with respect to the model parameters Θ. The said gradient provides the direction in which the training loss can be reduced maximally in the space of the model parameters. The model parameters are updated precisely in the said direction, i.e., $\Theta_{new} = \Theta_{old} - \eta \nabla_{\Theta} \mathscr{L}_{tr}$. Here, η is the "learning rate" of the algorithm, an adjustable hyperparameter that controls the step size at each epoch while moving toward a minimum of a loss function.

During the training phase, the training loss gets smaller and smaller implying that the synthetic distribution gets closer and closer to the original distribution. This usually continues until the algorithm reaches a point where updating the parameters no longer leads to a loss reduction. The ultimate hope is to capture the global optima $\Theta^{\star}$, i.e., the optimal set of parameters that minimizes the training loss function: $\Theta^{\star} = \text{argmin}_{\Theta} \mathscr{L}_{tr}(\Theta)$.

c8. Early Stopping

"Early stopping" refers to regularization techniques used in many ML algorithms to avoid overfitting. In a nutshell, an early-stopping algorithm continuously monitors the quality of the learning procedure and terminates the training phase if the model stops being optimized. How and when to terminate the training phase depends on how an early-stopping algorithm is defined. It is common to check the model performance on the validation dataset and adapt the algorithm through adjustable hyperparameters such as “patience” (i.e., the number of epochs the algorithm waits before terminating the training phase).

Remark 4.5. *The original CGNN neither has a validation phase nor exploits an early stopping technique. We added these details to reach our candidates to their maximum predictive power without worrying about the risk of overfitting.*

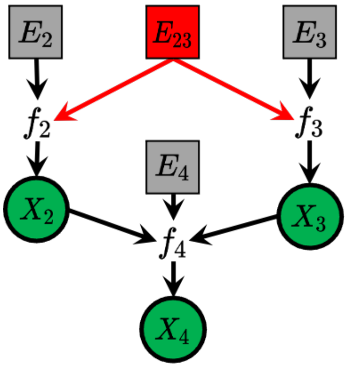
c9. Test Phase

At the end of the training phase, the model parameters are optimized and will no longer be updated. At this stage, the test phase begins. The test phase is carried out in several epochs to obtain stable numerical results. In a test epoch, the CGNN generates synthetic data and compares it with the test data. The final $\mathscr{L}_{te}$ is computed by averaging over all the test epochs.

c10. Final Score

The last step is assigning a final score to the candidate $\mathcal{G}$. If the algorithm is executed only once, the final score is minus the mean of the test loss over the test epochs. However, it is recommended to execute the whole algorithm multiple times and select the best result as an indicator of the candidate’s performance. This strategy is used to lower the risk of getting stuck at local optima. In the current project, we executed the algorithm four times per candidate.

Remark 4.6. *The original CGNN in Goudet et al. (2018) contains more details than what was presented here. In particular, it (1) computes the importance of edges in a DAG and prunes the unimportant ones, (2) optimizes a DAG structure by a hill-climbing algorithm, and (3) penalizes a DAG if it attains too many edges. We do not need these additional tools because our goal differs from the original CGNN.*



$$\begin{cases} X_2 = f_2(E_2, E_{23}) \\ X_3 = f_3(E_3, E_{23}) \\ X_4 = f_4(E_4, X_2, X_3) \end{cases}$$

Figure 4.7: Latent confounder: the CGNN postulates a shared noise term between two observed nodes if a latent confounder influences the nodes.

4.3.3 Latent Variables

The algorithm presented in the previous section is based on the causal sufficiency assumption, i.e., the scenario under consideration was assumed to involve no latent variables. This part briefly explains how the CGNN handles latent variables. Let Λ be a latent common cause for the observed nodes X_i and X_j. The CGNN models Λ through a noise term E_{ij} shared between the two nodes, where the values of E_{ij} are drawn from a continuous probability distribution such as Gaussian. Figure 4.7 illustrates the FCM corresponding to such a scenario. The rationale behind this strategy is that a noise term E_i represents the causal effect of all unknown variables affecting a single observed node X_i. Yet, if the latent Λ affects both X_i and X_j, its causal effect must be shared between the two observed nodes via a shared noise term.

The only mode the original CGNN considers for a latent variable is to be an "exogenous common cause." That is, Λ is always assumed to be a common cause for some observed nodes, but it is assumed not to be affected by any other observed nodes. Since some of the proposed causal graphs for the EPR-Bell scenario include endogenous latent variables, we generalize the above strategy as follows. Let Λ be a latent common cause for the observed nodes X_i and X_j, which itself is affected by an observed node X_k. The latent Λ is still modeled by a noise term E_{ij} shared between X_i and X_j. However, the values of E_{ij} depend on X_k, i.e., $E_{ij} = f_\Lambda(E_\Lambda, X_k)$, where f_Λ and E_Λ respectively represent the causal

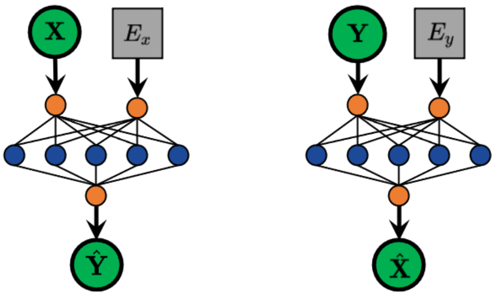


Figure 4.8: The pairwise variant of the CGNN trains two neural networks. The left network generates $\hat{\mathscr{D}} = (X, \hat{Y})$, while the right network generates $\hat{\mathscr{D}} = (\hat{X}, Y)$.

mechanism and the noise term corresponding to Λ. This leads to a new FCM that treats a latent variable like an observed variable. We use a neural network to estimate f_Λ and draw samples from a Gaussian distribution to model E_Λ.

4.3.4 Discussion

Before finalizing this section, I should address a few critical remarks about the methodology of the CGNN algorithm and pairwise methods in general. To make a start, let us focus on a bivariate scenario composed of two variables X and Y in which it is always the case that either X is a cause for Y ($X \to Y$), or Y is a cause for X ($Y \to X$). Given a sample from the joint distribution of X and Y, the goal is to discover the true causal arrow.

The pairwise variant of the CGNN is designed to deal with such a question. It starts by considering two neural networks corresponding to the two possible causal arrows (i.e., $X \to Y$ and $Y \to X$). The first network presumes $Y = f_Y(E_y, X)$ represents the true causal mechanism of the pair, while the second network presumes $X = f_X(E_x, Y)$ represents the true mechanism, as depicted in Figure 4.8. After training both neural networks, the algorithm generates two synthetic data and compares them with the true data $\mathscr{D} = (X, Y)$. Finally, the network that leads to a lower loss is chosen as the indicator of the true causal arrow.

The idea behind the sketched strategy is that the direction of causation manifests itself in the quality of the synthetic data: the network with the correct causal arrow leads to synthetic data that is closer to the true data, i.e., the true causal direction helps a network having a higher predictive power. Goudet et al. (2018) have shown that this idea successfully distinguishes cause from effect in many real-world examples of cause-effect pairs. Despite such a success, it is well-known that standard causal modeling techniques take the two causal models $X \rightarrow Y$ and $Y \rightarrow X$ indistinguishable, for the two models are Markov-equivalent, i.e., they imply the same statistical pattern of (in)dependencies. If so, the predictive power of $Y = f_Y(E_y, X)$ and $X = f_X(E_x, Y)$ must be the same. Yet, the critical question is how pairwise methods can distinguish between indistinguishable cases.

The answer lies in the fact that pairwise methods are intentionally designed to be sensitive to different footprints that the direction of causation leaves on data distributions. While the most reliable footprint is the set of conditional (in)dependencies, there are other footprints for identifying the causal direction. The difference between the complexities required to estimate the mechanisms is an instance of such a causal footprint: the estimation of the cause (from the effect) usually requires more complexity than the estimation of the effect (from the cause).[2]

Therefore, by keeping the complexity of the two mechanisms the same, the predictive power of the model that estimates the effect would be higher than the model that estimates the cause. In the context of the CGNN, the complexity of mechanisms is reflected in the number of hidden layers and units used by the neural networks. Therefore, the predictive power of two neural networks can be compared only after fixing the number of hidden units and hidden layers.

Another related point is that increasing the complexity usually leads to increased predictive power. For example, a neural network with several hidden layers is more powerful than a network with only one hidden layer. In standard ML problems, where the aim is to maximize the predictive power of a specific model, there is nothing wrong with increasing the complexity as long as one ensures that issues such as overfitting or lack of memory

[2] While the validity of this assumption is confirmed in many empirical studies, there is no rigorous theoretical proof for it. Nonetheless, check Janzing (2007, 2019) for novel justifications of why this assumption holds in practice.

do not occur. In contrast, pairwise models keep the complexity of both models as low as possible because too-high complexity makes them equally powerful, making it impossible to distinguish between cause and effect. In other words, high complexity makes the causal footprint dim and even disappear. Therefore, a compromising procedure is needed to optimize the complexity of the models: not so low that none of the two models can have a good estimation and not so much that both become equally powerful. Section 4.5 explains the compromising procedure we used for the current project.

4.4 Distance Quantification

The original CGNN utilizes the "Maximum Mean Discrepancy" (MMD) as its loss function, i.e., it uses the MMD to quantify the distance between the original data and the synthetic data generated by a candidate graph. At the beginning of the current project, we also used the MMD as the loss function. However, after applying it to quantum data, we realized that the MMD is not general enough to effectively distinguish a dataset containing quantum correlations from a completely uncorrelated one. To put it differently, the MMD cannot effectively distinguish correlations that satisfy Bell inequalities from correlations that violate Bell inequalities.[3]

This observation led us to a long journey to find alternative distance measures and heuristically define our own criteria for distance quantification. In this section, I introduce the MMD and our alternative measures. In addition, I explain how we systematically built a custom loss function for the EPR-Bell scenario.

4.4.1 MMD: definition and problem

Suppose $\mathbb{P}$ and $\hat{\mathbb{P}}$ are two probability distributions, defined over $\mathbb{R}^p$, from each of which we have two data samples $\mathscr{D}$ and $\hat{\mathscr{D}}$ of size n. A distance measure $\boldsymbol{\mathcal{M}}$ is a function that

[3] It is not the case that the MMD cannot detect any difference between these datasets. Instead, its sensitivity to quantum correlations is so low that the MMD cannot be used alone for training a model on quantum data.

receives $\mathscr{D}$ and $\hat{\mathscr{D}}$ as input and returns a non-negative real number that estimates the distance between $\mathbb{P}$ and $\hat{\mathbb{P}}$ as output. To this end, a distance measure compares the two data in light of a set of statistical features.

The MMD is a distance measure defined based on the framework of the kernel mean embedding introduced in Section 2.1.2. To quantify the distance, the MMD maps the two distributions to an intermediate space $\mathcal{H}_k$ and builds the so-called "kernel mean embedding" $\mu_k(.)$ of each distribution within the said space. It then compares the two embeddings to compute the distance between the distributions:

$$\mathrm{MMD}_k(\mathbb{P},\hat{\mathbb{P}}) = \left\|\mu_k(\mathbb{P}) - \mu_k(\hat{\mathbb{P}})\right\|_{\mathscr{H}_k}, \tag{4.16}$$

where k is a "kernel function", a real-valued symmetric function whose role is to measure the similarities between two vectors by computing their inner products. The above equation is suitable for distance quantification directly from the distributions. To estimate the distance from the data samples, one has to use the empirical variant of the former equation:[4]

$$\widehat{\mathrm{MMD}}_k(\mathscr{D},\hat{\mathscr{D}}) = \frac{1}{n^2}\sum_{i,j}^{n} k\left(x_i, x_j\right) + \frac{1}{n^2}\sum_{i,j}^{n} k\left(\hat{x}_i, \hat{x}_j\right) - \frac{2}{n^2}\sum_{i,j}^{n} k\left(x_i, \hat{x}_j\right). \tag{4.17}$$

There exist theoretical guarantees about the ability of the MMD to distinguish between *any* two distributions if the kernel function has a set of desirable properties. In particular, for a characteristic kernel function such as Gaussian $k\left(x,\hat{x}\right) = \exp\left(-\gamma \left|x - \hat{x}\right|_2^2\right)$, it holds that:

$$\lim_{n\to\infty} \widehat{\mathrm{MMD}}_k(\mathscr{D},\hat{\mathscr{D}}) = 0 \Longleftrightarrow \mathbb{P} = \hat{\mathbb{P}}. \tag{4.18}$$

The above result says that in the presence of an infinite number of data observations, the MMD vanishes if and only if the two samples come from an identical distribution. To put it differently, the MMD is guaranteed to distinguish non-identical distributions if it uses a characteristic kernel and accesses infinite-sized samples from the two distributions.

[4] There are many interesting details about the underlying ideas of kernel methods. I skip these details because they are beyond the scope of the present discussion. However, see Gretton et al. (2005) for a systematic approach toward these notions.

Despite this exciting guarantee, when we applied the MMD to quantum data, we noticed that the MMD could not effectively distinguish between conditionally correlated and uncorrelated data. To understand the reason for this contradictory behavior, note that the above guarantee holds if the sample sizes are arbitrarily large. In practice, it is impossible to infinitely increase the data size because the computational complexity increases so much that the computations become infeasible.

The said limitation dramatically lowers the sensitivity of the MMD to statistical features representing the very nature of quantum entanglement. Consequently, there is a risk that the MMD underestimates the distance between two datasets that are marginally close but conditionally very distant. Such a deficiency poses a severe problem to the applicability of the CGNN in adjudicating between different candidates in the EPR-Bell scenario. Because, as will be illustrated in Section 4.6, most candidates can reproduce marginal relations, while only a few of them can reproduce conditional relations. This situation led us to build alternative measures for distance quantification between two distributions.

4.4.2 Alternative Measures

Table 4.2 represents the eight criteria we exploited for distance quantification in the current project. Some of these criteria are entirely innovative, but others are inspired by the statistics literature. The “backpropagation” column in this table indicates whether a measure supports the backpropagation algorithm, i.e., whether the output of the measure is differentiable with respect to its inputs. A measure that supports backpropagation can be used in the training phase of the ML algorithm because its differentiability allows the algorithm to update the model parameters by minimizing the loss values. A measure that does not support backpropagation cannot be used in the training phase, but it can be used in the test phase.

We used the first three measures for the training phase and all eight for the validation and test phases. Hence, the first three measures are “seen” by the models (i.e., the models have seen them during the training phase and have used them to update their parameters),

Name	Backpropagation	Training	Validation & Test
MMD_cdt	Yes	Yes	Yes
CorrD	Yes	Yes	Yes
CndD	Yes	Yes	Yes
MMD_Fr	Yes	No	Yes
CorrD_N	Yes	No	Yes
NpMom	Yes	No	Yes
BinnedD	No	No	Yes
EDF_Marg	No	No	Yes

Table 4.2: The measures used for distance quantification between two distributions.

while the other five are "unseen." In the following, I give a brief explanation of each distance measure:

- **MMD_CDT**: the standard MMD measure defined in Equation 4.17. We took the programming codes of this measure from the Causal Discovery Toolbox (CDT) (Kalainathan & Goudet 2019).

- **CorrD**: a measure that considers all possible permutations of relations of types $(2, 2)$ and $(1, 3)$. To this end, it compares the correlation matrices of the two distributions by calculating the (1) joint distributions of two variables conditioned on the remaining two and (2) distribution of one variable conditioned on the remaining three.

- **CndD**: a measure that considers all possible permutations of relations of types $(1, 3)$, $(2, 2)$, $(3, 1)$, and $(4, 0)$. For each relation, it computes the corresponding conditional distributions for the two datasets. Then, it compares the two distributions by comparing their first n moments.

- **MMD_Fr**: a measure that approximates the standard MMD using the idea of Fourier coefficients introduced by Lopez-Paz et al. (2015). It is faster than the standard MMD but, similar to the standard MMD, cannot effectively detect conditional dependencies in quantum data.

- **CorrD_N**: a measure that considers all possible permutations of relations of types $(2,0)$, $(3,0)$, and $(4,0)$. It compares the two distributions by calculating the correlations between their marginals.

- **NpMom**: a measure that considers the relations of type $(4,0)$. To compare the two distributions, it compares their first n moments.

- **BinnedD**: a measure that bins the distributions into multidimensional histograms. To compare the two distributions, it compares the patterns of the histograms.

- **EDF_Marg**: a measure that is based on the distance between the empirical distribution function (EDF) of marginals in the two distributions. For each variable (such as O_A), it calculates the distance between the EDFs in the two datasets and then takes the average over all the variables.

4.4.3 Loss Function

Each of the eight measures listed in Table 4.4 has a set of hyperparameters controlling its computational speed and sensitivity to different conditional relations. To build loss functions that overcome the problem that the MMD faces, one should first tune the hyperparameters of the distance measures and then combine the distance measures in an appropriate way. This part explains the systematic steps we followed to build our loss functions.

Simulate Calibration Data

Besides the original quantum data, which is conditionally correlated, we simulate two calibration datasets: (1) a correlated dataset whose distribution is precisely the same as the original data and (2) an uncorrelated dataset whose marginal distribution is the same as the original data but is conditionally uncorrelated. Let us denote the original and the first calibration data by $\mathscr{D}_{C_1}$ and $\mathscr{D}_{C_2}$. Also, let $\mathscr{D}_U$ denote the second calibration data.

These three datasets are used to check the sensitivity of each distance measure to different types of conditional relations.

Calibrate Distance Measures

A distance measure takes two data samples and returns a non-negative output. The raw output does not straightforwardly indicate whether the two data are close or far; rather, it should be calibrated to be understandable. To this end, we calculate two reference quantities for each distance measure $\boldsymbol{\mathcal{M}}_j$ to represent lower and upper bounds on its outputs. I refer to these quantities as "effective zero" and "benchmark" and denote them by z_j and b_j, respectively:

- $z_j = \boldsymbol{\mathcal{M}}_j(\mathscr{D}_{C_1}, \mathscr{D}_{C_2})$,
- $b_j = \boldsymbol{\mathcal{M}}_j(\mathscr{D}_{C_1}, \mathscr{D}_U)$

The signal-to-noise ratio (SNR) is computed from these quantities by $s_j = (b_j - z_j)/\Delta b_j$, where the standard deviation Δ is taken over several data batches. Note that the SNR represents how reliably a measure can distinguish between correlated and uncorrelated datasets so that it remains stable in the presence of statistical fluctuations. Hence, the higher the SNR, the more reliable the measure outputs.

Tune Distance Measures

Each distance measure has some hyperparameters to be tuned. For instance, e.g., the dimension of the correlation matrices created by the CorrD is one of its hyperparameters. Besides measure-specific hyperparameters, the batch size of each measure is another hyperparameter to be tuned. The higher the batch size, the higher the SNR, but the lower the computational speed. Nonetheless, depending on the mathematical operations a distance measure performs internally, it may not be able to handle a high batch size.

To tune the hyperparameters of our measures (including the measure-specific ones and the batch sizes), we exploited a trade-off strategy between the computational speed and the SNR of each measure. Regarding the batch sizes, we found that all measures could handle a large batch size (i.e., 8000) except for the MMD, which required a low batch size (i.e., 500).

Define Loss Functions

Given synthetic data $\hat{\mathscr{D}}$ generated by a candidate, the original quantum data $\mathscr{D}_{C_1}$, and a set of distance measures $\{\boldsymbol{\mathcal{M}}_j\}$, we define our loss functions as follows:

$$\mathscr{L} = \sum_j w_j \mathscr{L}_j \text{ where } \mathscr{L}_j = \frac{1}{m_j} \sum_{k=1}^{m_j} \frac{\boldsymbol{\mathcal{M}}_j(\hat{\mathscr{D}}^{(k)}, \mathscr{D}_{C_1}^{(k)})}{z_j} \tag{4.19}$$

with w_j and m_j representing the weight and the number of batches corresponding to measure $\boldsymbol{\mathcal{M}}_j$. Furthermore, $\hat{\mathscr{D}}^{(k)}$ and $\mathscr{D}_{C_1}^{(k)}$ denote the k^{th} batch of data $\hat{\mathscr{D}}$ and $\mathscr{D}_{C_1}$, respectively.

There are two remarks regarding the above strategy. First, assigning non-equal weights to measures enforces the model to put more emphasis on the measures with higher weights. Especially in the training phase, this forces the model parameters to be updated in a desirable direction. In the training phase, we assigned the weights based on the SNRs, i.e., $w_j = \frac{s_j}{\sum_j s_j}$. In the test phase, we assigned equal weighting. Second, one of our measures (i.e., the MMD) requires a different batch size. The above strategy enables one to combine distance measures with different batch sizes and minimize them simultaneously.

4.5 Candidates

The CGNN does not have an inherent method for restricting its search space. Instead, the algorithm expects the user to specify a set of appropriate candidates for the given data so that the algorithm can adjudicate between the specified candidates. Therefore, to apply the CGNN to the EPR-Bell scenario, we must determine a set of candidates in advance.

Fortunately, the quantum foundations literature contains many proposed models for the underlying causal structure in the EPR-Bell scenario. Chapter 3 investigated some of these models in the context of the causal problem of entanglement. This includes interpretations of Quantum Mechanics that violate the CFC assumption (such as superluminal, superdeterministic, and retrocausal), models that violate the CMC assumption (such as separate common causes and interactive common causes), and models that undermine the legitimacy of the classical causal modeling framework for studying quantum correlations (such as quantum causal models and violations of the independence of mechanisms).

The CGNN provides a unified arena to judge between several of the above proposals. The former statement should not be construed as the claim that the CGNN discovers *the* underlying causal structure of the EPR-Bell scenario because many proposed models even cannot be sketched within the current framework. Nevertheless, the framework provides a good insight into the overall performance, strengths, and weaknesses of a given candidate in reproducing quantum statistics. Table 4.3 contains the names and the graph of the candidates examined in the current project. [5]

The first six candidates can be straightforwardly implemented within the CGNN framework by specifying their corresponding causal graphs. There are two particular candidates marked in red, which require non-standard implementations, i.e., specific changes must be inserted into the CGNN so that it can evaluate these candidates. In the following, I briefly introduce each candidate.

Remark 4.7. *There is a conceptual difference between the standard and non-standard candidates: the essence of the standard models lies in their "causal graph", while the essence of the non-standard models lies in the specific "causal mechanisms" they prescribe.*

[5] Besides these, we studied several other candidates. However, most of these candidates are not conceptually independent from the candidates presented in this chapter. Thus, I do not focus on them and put a summary of their performance in the GitHub repository.

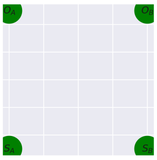
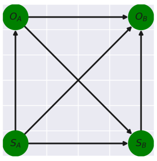
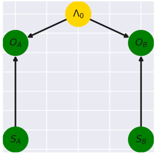
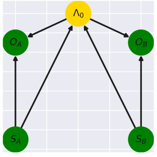
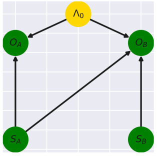
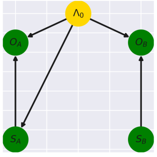
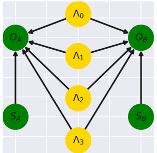
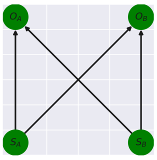
Abb.	Name	Graph
DIS	disconnected DAG	
SAT	saturated DAG	
CCC	classical common cause	
NSS	non-separability of states	
SLM	superluminal model	
SDM	superdeterministic model	
SCC	separate common causes	
VIM	violation of independence of mechanisms	

Table 4.3: The candidates studied in the current project.

4.5.1 Standard Implementations

DIS

A model in which no causal influence is exchanged between the nodes because they are all disconnected. Therefore, the values of each node are generated from the noise terms alone. The quality of the synthetic data generated by the DIS should be very low and can be considered as a lower bound for other candidates. That is, all candidates are expected to generate synthetic data with a quality higher than the DIS.

SAT

A model that is saturated with the maximum possible number of edges between the nodes so that the graph remains a DAG. Since most nodes are causally informed of the other nodes, the quality of data generated by the SAT should be very high and can be considered as an upper bound for other candidates.

CCC

A popular model used to derive the CHSH inequality (Clauser et al. 1969). In this model, each setting directly affects its outcome, and a common cause influences the two outcomes. The common cause is "classical" in the sense that a classical random variable represents its values. The CGNN models the said variable with a one-dimensional Gaussian distribution. The expectation would be that the quality of data generated by the CCC is better than the DIS, but it should not be able to reproduce quantum statistics completely. For more explanation of such a view, please refer to Section 2.3.3.

NSS

A model that postulates that the underlying physical states of two entangled particles are non-separable even when the particles are spacelike separated. According to this view, when two objects become entangled, a non-localized state is created that spreads throughout the spacetime regions between the two wings of the experiment. Such a non-localized state is denoted by Λ in Table 4.3. Here, both setting variables can influence Λ, and Λ can influence the two outcomes. For more explanation of such a view, please refer to Section 3.3.

Remark 4.8. *The NSS can be interpreted alternatively as a retrocausal model. Consider a scenario wherein the choices of the experiments (i.e., the setting variables) can change the underlying state of the entangled pair. Since the entangled state is prepared before starting the experiment, the causal influence of the settings must transfer to* Λ *retrocausally. For more explanation of such a model, see, for example, Wood & Spekkens (2015, p. 22).*

SLM

A model in which one of the settings causally influences the outcome of the other wing. Since the two wings are spacelike separated, the causal influence is transferred superluminally.

Remark 4.9. *The graph presented here is not the only possibility for a superluminal model; there are several other possibilities belonging to the family of superluminal models. I examine only one member of the family.*

SDM

A model in which the choices that one or both experimenters make for their measurements are not entirely free but depend on the underlying state of the entangled pair. The same underlying state is responsible for correlating the outcomes. The SDM violates the intervention assumption because it does not consider the "controllable" settings as exogenous variables.

Remark 4.10. *Figure 3.4 is another example of a superdeterministic model. Since the two models belong to the same family, I do not discuss the results of the second model here. However, it is worth mentioning that I applied the CGNN to both models, and the results were similar as they should have been.*

4.5.2 Non-standard Implementations

SCC

The SCC is a model that aims to show that quantum correlations do not conflict with Reichenbach's principle of common cause. The original variant of this model is formulated for a scenario wherein both settings and outcomes are binary-valued, Redei et al. (see, e.g. 2013, p. 162). I have explained this model in Chapter 3 and do not intend to repeat the details. Instead, I explain how to generalize the discrete variant of the model and implement it within the CGNN framework.

The discrete SCC postulates four latent common causes Λ_{ij} for the two outcomes, corresponding to the four possible setting choices. In each experiment run, when the experimenters choose $(S_A = i, S_B = j)$, Λ_{ij} is activated, and the other three common causes are deactivated. To extend the idea into a continuous scenario, one can partition the joint distribution of the settings into four bins $\mathcal{B}_{ij}$ and define four common causes corresponding to the four bins. Similar to the discrete variant, in each experiment run, Λ_{ij} is activated whenever the experimenters' choices locate in $\mathcal{B}_{ij}$, and the other common causes are deactivated.

In the context of the CGNN, the four variables Λ_{ij} are modeled by four independent Gaussian distributions and sampled in every experiment run. To generate synthetic data, depending on which bin the settings locate in, only one of the Λ_{ij} is allowed to influence the outcomes, and the values of the remaining common causes are discarded.

Remark 4.11. *Choosing four common causes is not a must. In fact, the number of common causes is a hyperparameter for the SCC. We chose four because we observed that the SCC could generate high-quality data even with such a small number of bins.*

VIM

The VIM refers to a model that violates one of the underlying assumptions of the causal modeling framework, namely the "autonomy of causal mechanisms" or "independence of causal mechanisms." Chapter 3 has already discussed this principle and has explained the motivations for questioning its validity in the quantum realm. Here, I propose a model that violates the principle and explain how to implement it within the CGNN framework.

Recall that, for a scenario with n causal variables, the CGNN posits n neural networks with the reasoning that each network estimates the causal mechanism of exactly one variable. Yet, consider a situation in which a single causal mechanism is responsible for simultaneously generating the two outcomes of the EPR-Bell scenario. In such a situation, the outcomes are not two separate events but a single event generated by a single causal mechanism. The functional form of such a mechanism is $O_A, O_B = f(S_A, S_B)$. In other words, not only do we not have two independent causal mechanisms, but also we have one causal mechanism that generates two variables.

To implement such a model within the CGNN, three neural networks (instead of four) are needed. The first two networks are standard and are used to estimate S_A and S_B. The third network is non-standard and is used to estimate O_A and O_B simultaneously. The latter has three inputs, i.e., (E, S_A, S_B), and two outputs, i.e., (O_A, O_B).

$$\begin{aligned} S_A &= f_{S_A}(E_{S_A}) \\ S_B &= f_{S_B}(E_{S_B}) \\ O_A, O_B &= f_O(E_O, S_A, S_B) \end{aligned} \tag{4.20}$$

Remark 4.12. *The model proposed here is not the only possible model for violating the independence of causal mechanisms. As before, I examine only one member of a family.*

4.5.3 Hyperparameter Tuning

Hyperparameters are parameters of an ML model that cannot be inferred during the training phase but can affect the learning quality. For instance, the numbers of hidden layers and hidden units are hyperparameters of a neural network. Depending on the internal structure of an ML model and the purpose for which the model is designed, the values of some hyperparameters play a critical role. In contrast, some hyperparameters are less

Abbreviation	Name	Selected Value
n_{layers}	number of hidden layers	1
n_{units}	number of hidden units	40
bs_{mmd}	batch size of MMD measure	500
bs_{other}	batch size of other measures	8000
n^{tr}_{epochs}	number of training epochs	10000
n^{te}_{epochs}	number of test epochs	20
$n_{patience}$	patience of early stopping	100
n_{runs}	number of algorithm's executions	4
η	learning rate	0.01

Table 4.4: Hyperparameters of the CGNN

important since they do not impact the learning quality of the models. One must therefore have strategies for tuning the hyperparameters depending on the purpose of the model under consideration. Table 4.4 sketches CGNN's hyperparameters and the values we selected for each.

Remark 4.13. *Besides the hyperparameters listed in Table 4.4, we tuned the hyperparameters of the eight distance measures in Section 4.4.*

In the case of the CGNN, the first four parameters in Table 4.4 are more critical than the rest. The batch sizes are important because they control the behavior of the loss functions by controlling the behavior of the distance measures, as was explained in Section 4.4. In what follows, I describe how we tuned the first two hyperparameters.

Hidden Layers and Hidden Units

n_{layers} denotes the number of hidden layers that the CGNN uses in its neural networks to estimate the causal mechanisms of a scenario. Similarly, n_{units} denotes the total number of hidden units the CGNN uses in its hidden layers. Increasing n_{layers} and n_{units} increases the model complexity and improves the learning quality of the models. However, high complexity can dim the causal footprints and make distinguishing cause from effect impossible. In contrast, low complexity can lead to underfitting, an instance where a model is so weak that it neither fits the training data nor generalizes to new data.

To combat the risk of dimming the causal footprints and underfitting the models at the same time, we took the SAT as a reference model and examined its performance for different n_{layers} and n_{units}.[6] We found that only one hidden layer is sufficient for the SAT to reproduce quantum statistics; thus, we set $n_{layers} = 1$. To optimize n_{units}, we implemented separate SAT models with one hidden layer but with different hidden units in the interval $[5, 80]$. We observed that the SAT performance improves with increasing the number of hidden units up to $n_{units} = 40$. Beyond this value, adding more hidden units did not contribute to improving the learning quality. So, we selected 40 as the optimal number.

4.6 Results

Previous sections created the theoretical and computational grounds for building a judgment framework between well-known causal models for the EPR-Bell scenario. The present section discusses the results of this judgment. That is, under the assumption that the CGNN is a valid framework for judging between different candidate models for the EPR-Bell scenario, I address which model(s) are preferred. Aside from this central question, I open the black box of the CGNN and present deeper insights about the performance of each candidate.

4.6.1 Performance by Seen Measures

Figure 4.9 represents the candidates' loss values at the end of the training and test phases. The blue and the red bars indicate the training and the test loss values, respectively, i.e., they represent the distance between the synthetic data of each candidate $\hat{\mathscr{D}}$ with the training data $\mathscr{D}_{tr}$ and test data $\mathscr{D}_{te}$. The error bars show the fluctuations of the losses in different epochs; the shorter an error bar, the more stable the corresponding candidate. The losses depicted in this figure are calculated by the seen measures $\boldsymbol{\mathcal{M}}_{tr}$ (i.e., MMD_cdt, CorrD, and CndD). Hence, following the standard convention in ML, the same criteria

[6] Recall that the SAT should generate one of the best synthetic data.

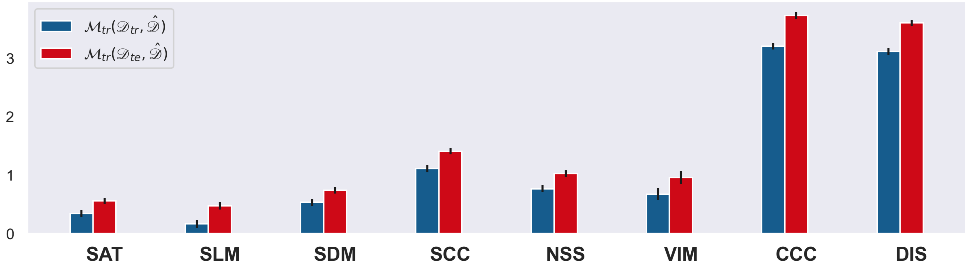


Figure 4.9: Performance on the seen measures: Loss values on the training and test datasets, where both losses are calculated using the seen measures.

are used for model training and model testing. Nonetheless, I will relax the convention in other figures. The figure suggests several points, listed below.

Firstly, the figure confirms a well-known point in ML: the training loss is larger than the test loss. The reason is that an ML model never sees the test data during the training phase, and hence its performance on the test data is slightly worse than its performance on the training data. The figure also demonstrates that the differences between the training and the test losses are not too high, suggesting that none of the models are overfitted.

Secondly, as expected, the CCC and the DIS reveal poor performance in reproducing quantum statistics in both the training and the test phases. Given that the candidates could fit to the training data unrestrictedly (i.e., their number of training epochs was unrestricted unless a candidate stopped learning), the poor performance in the training phase signals an essential feature of our framework. Namely, a candidate must have particular causal edges to learn all conditional patterns in quantum data, and this is true regardless of how many iterations the candidate fits to the training data. Because of the lack of these edges, the CCC and the DIS cannot fully learn the patterns in the quantum data.

Thirdly, apart from the CCC and the DIS, other candidates are more or less capable of learning and reproducing the quantum data. The SLM performs the best among the candidates, even better than the SAT. After the SLM and the SAT, the performance order is the SDM, the VIM, the NSS, and the SCC.

Fourthly, if we are to select only one candidate, we must declare that the CGNN prefers the SLM. However, we need further analysis since our goal is more than a simple model

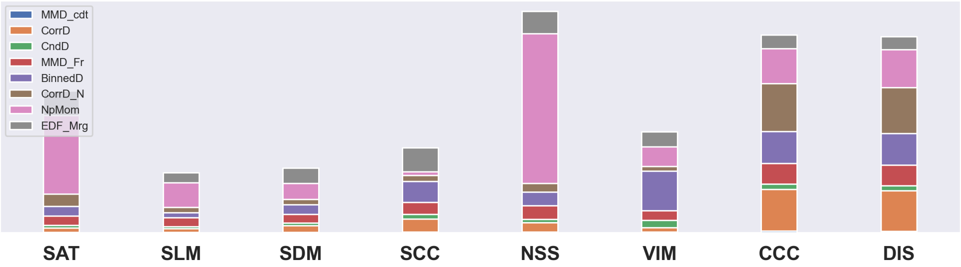


Figure 4.10: Performance on all measures

selection.

4.6.2 Performance by all Measures

Figure 4.10 depicts another way of adjudicating between the candidates. In this figure, the synthetic data of each candidate $\hat{\mathscr{D}}$ is compared with the test data $\mathscr{D}_{te}$, but in light of all the eight measures in Table 4.2. Recall that the first three measures are "seen" by the candidates during the training phase, while the remaining five are "unseen" measures.

There are two motivations for monitoring the candidates' performance by the seen and unseen measures. On the one hand, in the absence of a universal criterion for distance quantification between two distributions, it is safer to exploit a wide range of measures to ensure that our distance quantification is sensitive to different types of conditional and marginal relations. On the other hand, I aim to find out which candidate(s) are such well-suited to quantum data that they can learn its patterns even if they have not explicitly seen some criteria. A successful candidate in this context not only performs well on unseen data but also performs well on unseen measures. The figure suggests several points, listed as follows.

Firstly, the SLM and the SDM reveal the best performance and perform similarly in this context. In contrast, the DIS and the CCC reveal the worst performance and perform similarly.

Secondly, the CCC and the DIS cannot minimize the CorrD, while other candidates are more or less successful in this task. Note that the CorrD is the important criterion that

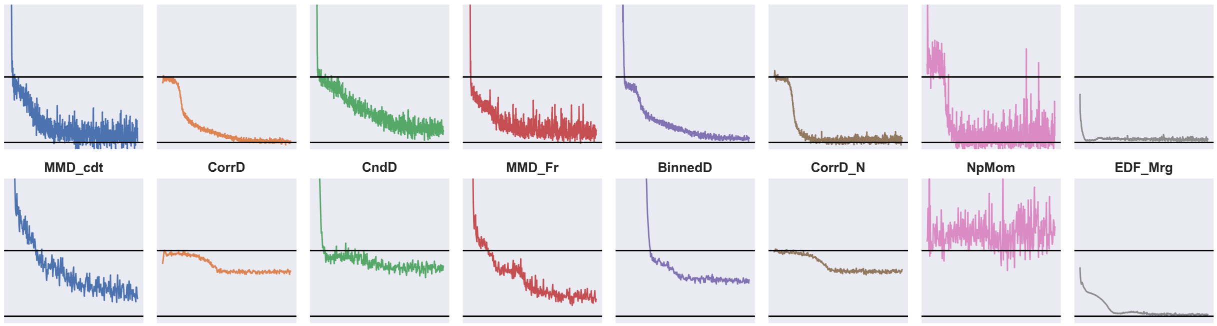


Figure 4.11: Learning progress of the SLM vs. CCC. The first and the second rows correspond to the SLM and the CCC.

evaluates the candidates in reproducing relations of type $(2, 2)$ and $(1, 3)$.

Thirdly, in contrast to the previous figure, the NSS and the SAT reveal poor performance in the current figure, primarily due to the inability of these two models to minimize the NpMom criterion. Given that the NpMom is sensitive to relations of type $(4, 0)$, the figure indicates that the NSS and SAT models are mainly incapable of learning these relations.

Fourthly, in a more general view, it appears that minimizing some of the measures is "easy" for all or most candidates, while there are some measures that a few candidates can minimize. The subsequent figure is dedicated to this concept.

4.6.3 Loss Progress

Figure 4.11 compares the learning progress of a successful candidate (SLM) versus an unsuccessful (CCC) one. The figure depicts how the two candidates evolve in reducing the eight distance measures during the training phase. For each candidate, there are eight sub-figures corresponding to eight distance measures. Each sub-figure has two black lines indicating the calibration quantities of the corresponding measure: the lower and upper lines represent the "effective zero" and the "benchmark," respectively. Hence, even the worst candidate (i.e., the DIS) should be able to reduce all the measures to the upper lines, while few candidates are expected to be so successful that they can reduce the distances to the lower lines. The figure suggests several points, listed as follows.

Firstly, in the limit of many training epochs, the SLM can reduce all the eight distance

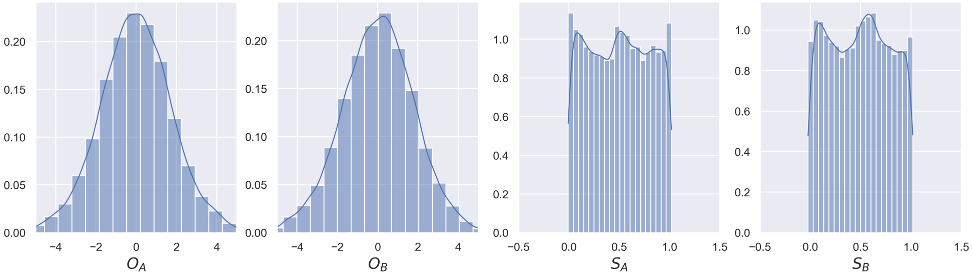


Figure 4.12: The marginals generated by the DIS.

measures to their lower bounds, while the CCC can reduce only some of them.

Secondly, the figure shows the problem of the MMD loss function described in Section 4.4. To see the point, note that while the quality of the CCC synthetic data is very low and most distance measures certify such low quality, the two measures MMD_cdt and MMD_Fr are progressed towards the lower lines. Consequently, if our distance quantification was merely based on the MMD, the CGNN could hardly distinguish between low-quality (e.g., CCC) and high-quality (e.g., SLM) datasets.

Thirdly, some distance measures shown in this figure are more sensitive to marginal relations, while others are more sensitive to conditional relations. A measure such as the EDF_Marg belongs to the first category, while measures such as CorrD and BinnedD belong to the second category. To see the reason, note that the unsuccessful CCC succeeds in minimizing the first category while it cannot minimize the second category.

4.6.4 Marginal Distributions

Figure 4.12 examines the synthetic data of the DIS from the perspective of marginal distributions. Comparing this figure with Figure 4.1 reveals that the DIS has successfully learned the relations of type $(1, 0)$. Such success should not surprise us because no causal information must be exchanged between the nodes of a graph to generate these marginal relations. Note that the same result holds for other candidates, i.e., they learn the marginal distributions equally well.

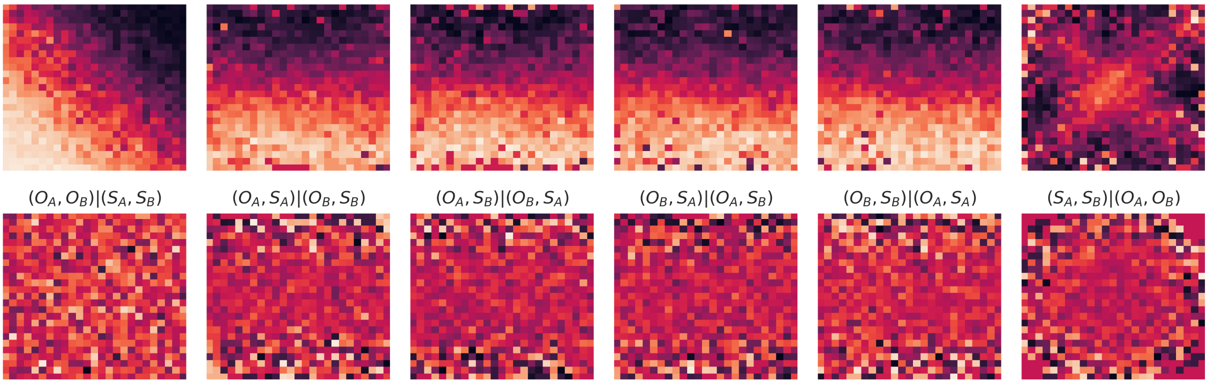


Figure 4.13: Heatmaps of the SLM vs. CCC. The first and the second rows correspond to the SLM and the CCC.

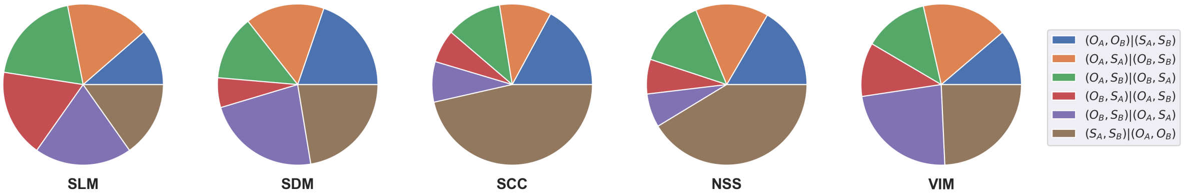


Figure 4.14: Portions of the six conditional relations of type $(2, 2)$.

4.6.5 Heatmaps

Figure 4.13 compares the SLM and the CCC from the perspective of the six relations of type $(2, 2)$. Comparing this figure with Figure 4.4 leads to a qualitative conclusion that the SLM accurately learns relations of type $(2, 2)$, while the CCC cannot learn any of these relations. For instance, the first plot in the second row reveals that the conditional $(O_A, O_B|S_A, S_B)$ in the original data is by no means similar to the same conditional in the CCC data.

4.6.6 Conditional Portions

Figure 4.14 compares the candidates in learning the six relations of type $(2, 2)$. In the previous figures, we saw that all the candidates, except the DIS and CCC, can learn relations of type $(2, 2)$. However, the present figure studies the six relations separately in the sense that it visualizes the contribution of these six relations in the average distance of type $(2, 2)$. The motivation for making such a fine-grained comparison originates from the fact that proposals for the EPR-Bell scenario mostly care about the relation $(O_A, O_B|S_A, S_B)$ and ignore other relations of type $(2, 2)$ such as $(S_A, S_B|O_A, O_B)$. Through this analysis, I aim to examine how symmetrically each candidate deals with the six relations.

The SLM (and the SDM to some extent) exhibits the best performance in this context because it reproduces the six relations symmetrically regardless of the permutation order of the variables. In contrast, reproducing $(S_A, S_B|O_A, O_B)$ for the SCC, NSS, and VIM is more challenging than reproducing $(O_A, O_B|S_A, S_B)$. This observation supports what was stated in the prior paragraph: since the primary goal of these models is to reproduce $(O_A, O_B|S_A, S_B)$, they pay less attention to the quality of $(S_A, S_B|O_A, O_B)$.

4.7 Conclusion

The cornerstone of this chapter is the CGNN algorithm, a classical causal discovery algorithm that combines techniques from Machine Learning and causal modeling. While the original CGNN has nothing to do with the issues in the quantum foundations, by inserting a set of changes to the algorithm, I adapted it to a new framework for studying the causal problem of entanglement. The new framework overcomes the shortcomings of traditional discovery algorithms, particularly the problem of constraint-based algorithms.

As was mentioned in Chapter 3, constraint-based algorithms cannot do justice to the causal problem because they rely on conditional independence relations to estimate the causal structure of a scenario. While conditional independence relations usually encapsulate reliable causal information, they are not the only resource for causal information. In the EPR-Bell scenario, it is essential to look at the full (i.e., joint) distribution rather than just conditional independencies because the latter does not provide enough information. That is why Wood & Spekkens (2015) proposed the need for discovery algorithms that are sensitive to the strength of correlations.

The framework discussed in this chapter follows the proposal of Wood & Spekkens (2015) and provides an inherently empirical tool to judge between a wide range of models for the causal problem. It is, however, important to emphasize that there are many other models for the EPR-Bell scenario that cannot be sketched with the current framework. Thus no claim can be made about such models within the current framework.

Among the candidates examined in this project, the SLM and (to some extent) the SDM delivered the best performance. By the best performance, I do not just mean getting the lowest value for the loss function. I compared the learning quality of the candidates from different perspectives using various distance measures, each being sensitive to some statistical features in the datasets. For instance, while candidates such as the NSS, VIM, and SCC can learn conditional relations, their learning quality is not very high. The said candidates are not successful, especially in minimizing the "unseen measures" (see Figure 4.10) and symmetrical dealing with conditional relations (see Figure 4.14).

4.7.1 Objections

Two objections might be raised to the current framework and the way I interpreted the results. I address these objections in what follows.

Empirical Approaches

The first objection is whether the current framework is inherently empirical, as I claim. Can it be argued that the results obtained in this project are assumption-free and that mere data leads to the results? The most crucial part of the framework about which this concern can be raised is the distance measures used to adjudicate among the candidates. The question is whether the data alone dictates the distance measures. Mainly because each measure has several adjustable hyperparameters that cannot be inferred from the data alone, i.e., additional assumptions are needed to tune these hyperparameters.

To some extent, the above objection is fair. Choosing appropriate distance measures and tuning each measure requires many assumptions that are not necessarily data-driven. Nonetheless, I argue that the framework is inherently empirical because all the complications regarding the selection and tuning of the distance measures originate from the fact that there is no universal measure for distance quantification between two probability distributions. If such a universal measure existed, we could take two datasets and estimate the absolute distance between them. In that case, we would no longer need additional assumptions for distance quantification: the candidate with the most compliance with the actual causal graph would reproduce the closest data to the quantum data. In the absence of such a universal criterion, we must define our judgment criteria so widely that the quality of the synthetic data of each candidate can be tested from different statistical perspectives. Therefore, it can be said that this objection is mainly related to the lack of a universal distance measure rather than the methodological foundations of the framework.

Interpretations of Quantum Mechanics

The second objection concerns an apparent conflict between interpretations of Quantum Mechanics and the results presented in this chapter. It is generally believed that there is an empirical equivalence between many interpretations of Quantum Mechanics and the standard (i.e., operational) quantum theory. A famous example is the variant of Bohmian mechanics defended by Dürr et al. (2013) that, despite all its ontological differences with the standard quantum theory, it makes similar predictions at the operational level. The question is whether the results discussed in this chapter contradict these empirical equivalence results.

I argue that there is no contradiction. To explain the reason, I must first return to the foundations of pairwise discovery algorithms. As detailed in Section 4.3, pairwise algorithms are designed to distinguish between Markov-equivalent models. For example, the pairwise variant of the CGNN distinguishes between two candidates $X \rightarrow Y$ and $Y \rightarrow X$ by fitting two functions $Y = f(X)$ and $X = f(Y)$ and comparing the quality of the data generated by both candidates. Given that the two candidates are Markov-equivalent, the two functions must yield the same quality if they acquire as much complexity as they need. However, a pairwise algorithm does not maximize the complexity of the two functions because its ultimate goal is not model-fitting. Instead, the goal is to find a candidate that generates the highest quality data with the lowest complexity with the justification that the closer a candidate is to the actual causal graph, the less complexity it requires to reproduce the data.

With such a description, I do not claim that the results presented in this chapter indicate that quantum interpretations are empirically non-equivalent. Nor do I claim that some interpretations cannot reproduce some statistical aspects of the original quantum data. However, I do claim that some candidates require less complexity to learn the patterns in quantum data, and these models are preferred from the CGNN perspective. Note that it is enough to increase the complexity of each candidate (i.e., the number of hidden layers and units) to observe the empirical equivalence. However, I do not increase the complexity of the candidates arbitrarily because the goal is to find the candidate(s) which can reproduce the quantum data with the lowest complexity.

Another related observation to the present discussion is that Daley et al. (2022) recently proposed a method to empirically adjudicate between four candidate models for a Bell-type scenario, including a quantum causal model (QCM), the SLM, the SDM, and the CCC. Similar to the ideas presented in this chapter, the authors compared the candidates based on their training and test errors. They demonstrated that the candidates achieved different scores and concluded that the QCM outperformed the others[7]. To justify why empirically-equivalent candidates exhibit different performances in their framework, the authors referred to the phenomenon of overfitting. They argued that the SLM and SDM are more prone to overfitting (compared to the QCM) because these candidates violate the CFC assumption, and hence the conditional independencies they generate are achieved by fine-tuning the parameters. In contrast, the causal graph of the QCM implies conditional independence, and therefore it is less prone to overfitting. According to this argument, the two candidates reveal poor performance because they mistake statistical fluctuations for

[7] It is worth noting that the method in Daley et al. (2022) differs from the one presented in this chapter in several aspects, such as the discrete nature of their scenario, the use of optimization algorithms instead of Machine Learning, and the judging criteria considering only $(O_A, O_B|S_A, S_B)$ rather than the joint distribution.

real features during training.

Unfortunately, I did not find a way to directly implement a QCM within the CGNN framework. However, I can state that the performance discrepancy among our candidates is not due to overfitting. This is so because our framework combats the risk of overfitting by early-stopping, and as shown in Figure 4.9, none of the candidates overfitted. Therefore, I contend that the performance discrepancy in our framework is due to the complexity discrepancy between the candidates. As such, the fact that the SLM outperforms the other candidates can be taken as evidence that the SLM requires less complexity to learn the quantum data.

4.7.2 Lessons

This chapter has two important lessons for future research. The first lesson concerns the need for modifying the CFC in light of quantum correlations, while the second concerns the advantages of Machine Learning in studying the causal problem of entanglement.

CFC Modification

The CFC asserts that two causally dependent variables are statistically dependent, or conversely, two statistically independent variables are causally independent. The CFC is usually justified by the argument that there is no reason to believe that causal connections conspiratorially hide themselves from observers, so causally connected variables exhibit statistical dependencies.

When applied to our EPR-Bell scenario, the CFC implies that all four variables are causally disconnected because each pair is marginally independent. There are, of course, some physical mechanisms connecting the variables, the footprints of which are revealed only after examining the conditional relations between the variables. However, the question is why the CFC does not consider such an important possibility. Why does the CFC not check the conditional dependencies of two marginally independent variables before declaring them causally disconnected?

The phenomenon I am describing above is the existence of random variables that are marginally independent but conditionally dependent. Such variables are not specific to Quantum Mechanics. For instance, in the collider structure $X \rightarrow Z \leftarrow Y$, it holds that $X \perp\!\!\!\perp_{\mathbf{s}} Y$, while $X \perp\!\!\!\perp_{\mathbf{s}} Y|Z$.I believe that there are three possible answers to the question posed above: (1) the no-conspiracy argument (which suggests that it is reasonable to assume that causally connected variables do not hide their causal connections); (2) practical considerations (checking conditional dependencies between two marginally independent variables can be impractical because there may be infinitely many conditioning variables to consider); and (3) the CFC is not problematic in typical examples such as the collider structure (in such cases, the CFC concludes causal independence from marginal independence, and this is a correct conclusion).

Nonetheless, the critical point about quantum correlations is that the pattern of marginal independencies and conditional dependencies does not occur in a collider. In the EPR-Bell scenario, the two outcomes are marginally independent while conditionally dependent, given the two settings. Notably, the settings are not child nodes of the outcomes as in a collider. The CFC becomes problematic in facing such a scenario. It appears that the CFC needs some modifications in light of quantum correlations. To address this, we need a modified version of the CFC that considers both marginal *and* conditional relations between variables to make a decision about the causal connection between the variable.

Advantages of Machine Learning

The CGNN algorithm generalized in this chapter is only one of several approaches that exploit Machine Learning techniques. There are many other ideas that the quantum foundations community can utilize to address challenges related to quantum entanglement phenomena. One promising idea is the development of custom loss functions whose minimization leads to solving a desired complex problem. This approach may be particularly useful in scenarios where multiple quantum systems interact simultaneously or when both observational and interventional data are available. In the next chapter, I will introduce another algorithm for quantum systems, which is again based on Machine Learning techniques and is more closely aligned with the concepts of quantum causal models. This algorithm can estimate causal relationships in quantum networks with multiple interacting quantum systems.

Chapter 5

Quantum Causal Discovery with Machine Learning[1]

In the previous chapter, I presented a framework for studying the causal problem of quantum entanglement using Machine Learning (ML). While this framework provides a useful tool for evaluating candidate models of the EPR-Bell scenario, the lack of an independent ground truth for verifying the adjudication results makes independent verification impossible. This is due to the inherent nature of the causal problem of entanglement, where apart from quantum data, there is no other information about the causal structure of the scenario.

In contrast, this chapter focuses on supervised Machine Learning tasks and examines quantum scenarios whose causal structures are already known. This enables the evaluation of Machine Learning algorithms using independent ground truth. The cornerstone of the chapter is the Randomized Causation Coefficient (RCC), a pairwise discovery algorithm originally designed for distinguishing cause from effect in classical scenarios. I extend the RCC in several ways to create a quantum causal discovery algorithm that handles bivariate and multivariate quantum scenarios. The resulting algorithm performs well in

[1] The present project is the result of my collaboration with **Jan Dziewior** from the Max-Planck-Institute for Quantum Optics, 85748 Garching, Germany; Faculty of Physics, LMU Munich, 80797 Munich, Germany. Jan Dziewior has simulated the quantum experiments.
The source code of this project can be found here: `https://github.com/OmidCharrakh/Quantum_RCC/`

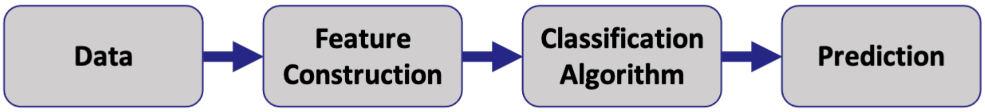


Figure 5.1: The idea of discriminative algorithms.

various simulated scenarios.

5.1 Introduction

Consider a bivariate scenario wherein there are only two variables, and it is always the case that one variable is the cause and the other is the effect. A pairwise causal discovery algorithm aims to distinguish cause from effect by examining a data sample from the joint distribution of the scenario. To achieve this goal, a pairwise algorithm looks at the hidden asymmetries (in the given data sample) that should arise because of the directionality of causal relationships. There are two types of pairwise algorithms: generative and discriminative.

A "generative" algorithm estimates the functional form of mechanisms for both causal possibilities and generates synthetic data. Such an algorithm estimates the direction of causation by comparing the original data sample with the synthetic samples. Examples of generative algorithms include CGNN (Goudet et al. 2018), LiNGAM (Shimizu et al. 2006), and IGCI (Janzing et al. 2012).

A "discriminative" algorithm converts the cause-effect problem into a classification problem to be tackled by Machine Learning algorithms (Goudet et al. 2019). Such an algorithm builds a classifier, the output of which estimates the direction of causation. The RCC algorithm (Lopez-Paz et al. 2015) discussed in the present chapter is an example of discriminative pairwise models. Other examples include ProtoML (Almeida 2019) and Jarfo (Minnaert 2019).

Figure 5.1 depicts how a discriminative algorithm works. The "feature construction" is the most critical step in a discriminative algorithm. In this step, a set of features are extracted from the given data sample so the classifier can recognize the causal direction from these extracted features. Following Goudet et al. (2019), there are three approaches for feature construction: (1) manually constructing causally relevant features, (2) automatically identifying causally relevant features from the training set, and (3) embedding the sample into a fixed-size feature vector.

The RCC is a strong algorithm that follows the third approach for feature construction. To this end, the RCC extracts a finite number of features by computing the empirical kernel embedding of distributions in the reproducing kernel Hilbert space. In short, the main idea is to systematically extract a set of causally relevant features out of the samples from the joint distribution of cause and effect variables and let a classifier learn the relationship between the extracted features and the direction of causation.

The original RCC is limited to one-dimensional random variables, i.e., it presumes that both the cause and effect variables are one-dimensional. Once trained, the original RCC can answer queries such as "Does altitude cause a change in temperature or vice versa?" given a sample over these variables. The authors showed that the RCC reveals one of the best performances in recognizing the causal direction in classical scenarios.

This chapter seeks to extend the methodology of the RCC to quantum scenarios. The quantum scenarios discussed in this chapter consider the transmission of a quantum state between two qubits through potentially noisy channels. The aim is to learn the direction of the channel, and thus the causal direction, from the qubits' states without explicitly performing interventions. Since the original RCC is limited to probability distributions over classical variables, it is nontrivial whether it applies to probabilistic descriptions of quantum states (i.e., density matrices). To investigate this question, I extend the algorithm work with three-dimensional random variables. I show that the algorithm can learn the direction of causation in most cases.

In addition to bivariate quantum scenarios, I extend the RCC to multivariate quantum scenarios. In the latter scenarios, multiple local labs exist where each lab transmits its qubit state to other labs through potentially noisy channels. For such scenarios, I combine the discriminative RCC with the generative CGNN and build an algorithm that takes the states of all labs as inputs and returns the underlying causal graph as output. The resulting

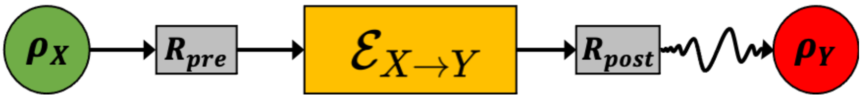


Figure 5.2: A quantum channel transmitting the quantum state of a qubit from Alice's lab to Bob's lab.

algorithm does not require tomographically complete data to recover the underlying causal graph.

The results obtained in this chapter suggest that the relevance of causal modeling and Machine Learning techniques in the quantum realm is not restricted to fundamental discussions. These techniques have practical advantages in quantum engineering and process tomography.

The remainder of the chapter is organized as follows. Section 5.2 discusses the physical scenarios simulated in this chapter. Section 5.3 describes how the original RCC works. Section 5.4 explains how I generalize the methodology of the RCC and the CGNN algorithms and build a discovery algorithm applicable to quantum scenarios. Section 5.5 contains the results of applying the discovery algorithm to the simulated quantum scenarios. Section 5.6 explains how the algorithms presented in this chapter relate to previous works in quantum causal discovery and suggests ideas for future works by drawing the implications of the current model.

5.2 Physical Scenario

Consider two experimenters, Alice and Bob, in two local labs, each with a qubit. The experiment comprises several runs, in which one of the experimenters must send the state of her/his qubit to another lab. Figure 5.2 depicts one of these runs wherein Alice sends her state to Bob through a quantum channel. The state that Bob receives is not the same as the initial state that Alice sends, for the state undergoes changes along the way.

Here, we are dealing with a cause-effect scenario among quantum states where Alice's state ρ_X is a cause for Bob's state ρ_Y. The reason why ρ_X can be construed as a cause for

ρ_Y is that any change in the former brings out changes in the latter, while the converse is not true. Therefore, the direction of the quantum channel indicates the direction of causation.

The primary goal of the current project is to find out whether it is possible to determine the direction of causation from the states. That is, whether it is possible to treat ρ_X and ρ_Y like classical random variables and infer the causal direction from the asymmetries hidden in their statistics. Meanwhile, the goal is to identify the causal direction from "observational" rather than "interventional" data. That is, we are not allowed to intervene explicitly on one of the states and analyze the effect of the intervention on the other variable. Rather, we are only allowed to get a sample $\mathcal{S} = (\rho_X, \rho_Y)$ from the joint distribution over the states and determine whether the sample is compatible with $\rho_X \to \rho_Y$ or $\rho_X \leftarrow \rho_Y$.

5.2.1 States and Channels

In Quantum Mechanics, the state of a d−dimensional quantum system is represented by a $d \times d$ complex matrix, which is known as a density matrix. A density matrix ρ is a positive semi-definite and Hermitian operator of trace one that is defined on the Hilbert space corresponding to the system. The density matrix can be viewed as a probability distribution over pure states, with $\rho = \sum_i p_i |\psi_i\rangle \langle\psi_i|$ where p_i is the probability of a given pure state $|\psi_i\rangle$ occurring.

A qubit is a 2−dimensional quantum system, the state of which is represented by a 2×2 complex matrix. Since the space of matrices is a vector space, there are bases of matrices that can be used to decompose any matrix. For qubits, such a basis consists of three Pauli matrices, allowing a density matrix to be expressed as a 3−dimensional "Bloch vector" (Bertlmann & Krammer 2008):

$$\rho = \frac{1}{2}\left(I + r_x\sigma_x + r_y\sigma_y + r_z\sigma_z\right) \tag{5.1}$$

Remark 5.1. *From a statistical perspective, a classical bit can be represented by a 1-dimensional binary-valued random variable, while a qubit can be represented by a 3-*

dimensional continuous random variable.

To describe the dynamics of quantum systems in the most general way, the "quantum operations formalism" is usually used (see, e.g., Nielsen & Chuang 2010). The formalism provides a powerful tool to calculate how state transformation occurs when a quantum system is subjected to changes. Let a quantum system with the initial state ρ_i undergo a transformation represented by $\mathcal{E}$. According to the quantum operations formalism, the relationship between the initial and final states is expressed by the following equation:

$$\rho_f = \mathcal{E}(\rho_i) = \sum_k E_k \rho_i E_k^\dagger. \tag{5.2}$$

Equation 5.2 is known as the "operator-sum representation" of the quantum operation $\mathcal{E}$, where operators $\{E_k\}$ are called the "operation elements" for the quantum operation $\mathcal{E}$. In order to be a transformation legitimated by the axioms of Quantum Mechanics, $\mathcal{E}$ should be a completely positive (CP) and non-trace-increasing map on the space of density matrices. This implies that the operational elements must satisfy the condition $\sum_k E_k^\dagger E_k \leq I$.

A "quantum channel" is a trace-preserving quantum operation, i.e., the input and outputs of a quantum channel have the same trace. Therefore, for a quantum channel, it holds that $\sum_k E_k^\dagger E_k = I$. The trace-preserving property ensures that if a density matrix enters a quantum channel, the output is again a density matrix.

In Figure 5.2, the state sent by Alice undergoes the following transformations on the way to reaching Bob:

- First, an initial rotation is applied to the state: $\rho_2 = R_{pre} \rho_1 R_{pre}^\dagger$.
- Then, the state enters a quantum channel: $\rho_3 = \mathcal{E}_{A \to B}(\rho_2)$.
- Next, a secondary rotation is applied: $\rho_4 = R_{post} \rho_3 R_{post}^\dagger$.
- Finally, a Gaussian noise is applied: $\rho_5 = \rho_4 + \vec{N}$.

The channel $\mathcal{E}_{A \to B}$ is a probabilistic mixture of six different channels: (1) identity, (2) dephasing on the σ_z basis, (3) dephasing on a random basis, (4) replacement by the pure

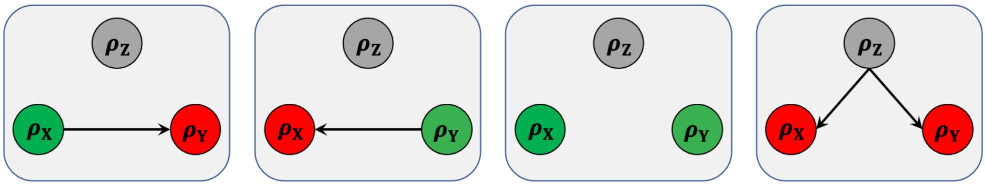


Figure 5.3: The four causal labels of the second scenario.

state σ_z, (5) replacement by a random pure state, and (6) replacement by white noise. Therefore, the relationship between the channel input and output is formulated as,

$$\rho_3 = \sum_{k=1}^{6} w_k E_k \rho_2 E_k^\dagger \quad \text{where} \quad \sum_{k=1}^{6} w_k = 1. \tag{5.3}$$

Remark 5.2. *The above stages involve several adjustable parameters, such as the weights assigned to the six channels, the mean and variance of rotation strength in R_{pre} and R_{post}, and the level of noise introduced. Adjusting these parameters allows for the simulation of different data-generating processes, resulting in a range of physically distinct scenarios.*

5.2.2 Scenarios

Throughout the chapter, I consider three scenarios: two bivariate and one multivariate. In what follows, I briefly explain each.

Scenario 1

The first scenario is the cause-effect scenario depicted in Figure 5.2. Given a sample $\mathcal{S} = (\rho_X, \rho_Y)$, the goal is to predict the causal label of the sample. Two causal labels are possible in this scenario: "forward" ($\rho_X \rightarrow \rho_Y$) and "backward" ($\rho_X \leftarrow \rho_Y$).

Scenario 2

The second scenario is also bivariate. Given a sample $\mathcal{S} = (\rho_X, \rho_Y)$, the aim is to predict the causal label. The causal labels are of four types (see Figure 5.3):

1. Forward: ρ_X causes ρ_Y

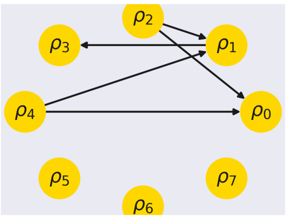


Figure 5.4: An example of a multivariate scenario with eight variables.

2. Backward: ρ_Y cases ρ_X

3. Disconnected: ρ_X and ρ_Y are disconnected causally.

4. Latent Common Cause: there is no causal edge between ρ_X and ρ_Y, but a latent common cause ρ_Z influences both.

Scenario 3

The third scenario involves multiple local labs, where each lab sends its qubit state to others through various channels. This is a multivariate scenario, denoted as $p-$variate, where p is the number of local labs. The aim is to recover the causal graph of the scenario using a $p-$variate sample $\mathcal{S} = (\rho_1, \ldots, \rho_p)$ from the joint distribution of the variables. To simulate this scenario, an $8-$variate example is considered, which is shown in Figure 5.4.

Remark 5.3. *In addition to quantum operations used in the bivariate scenarios (i.e., rotations, channels, and noise), the multivariate scenario comprises an additional quantum operation, namely quantum gates. The quantum gates combine the causal effects of incoming edges to a node. I use a probabilistic mixture of three quantum gates: an averaging gate and two controlled NOT (cNOT) gates. Similar to the quantum channels, the weights of quantum gates are adjustable parameters.*

5.3 The RCC Algorithm

In the introduction, I mentioned that discriminative algorithms have a "feature construction" step in which appropriate statistical features are extracted or constructed from a given probability distribution. The RCC is the cornerstone of the present chapter because its feature construction is carried out systematically with minimal assumptions on the underlying data-generating process. In this section, I explain how the RCC constructs its features and trains a classifier on top of the constructed features.

5.3.1 Featurization

The feature construction of the RCC is based on the kernel mean embedding framework. In this framework, probability distributions are mapped to a reproducing kernel Hilbert space so that calculations on the distributions can be performed on the image of the distributions in the said space (see Section 2.1.2 for the details). The authors prove that for shift-invariant characteristic kernels (such as the Gaussian kernel), it is possible to obtain a low-dimensional representation of distributions using the normalized Fourier transform of the kernel. In particular, given the sample $\mathcal{S}$ from distribution $\mathbb{P}$, the empirical low-dimensional representation of the distribution is computed as:

$$\hat{\mu}_k(\mathcal{S}) = \sqrt{\frac{2}{m}} \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} [\cos(w_1^\intercal x + b_1), \ldots, \cos(w_m^\intercal x + b_m)] \in \mathbb{R}^m \tag{5.4}$$

where w_i is drawn from the normalized Fourier transform of k and $b \sim \mathcal{U}[0, 2\pi]$. Moreover, $|\mathcal{S}|$ denotes the sample size, and m is the desired number of features to be constructed from the given sample. Hence, Equation 5.4 "featurizes" $\mathcal{S}$ by constructing m features from it.

In a bivariate scenario, the RCC exploits the following equation to featurize a joint distribution $\mathbb{P}_{XY}$ using sample $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$:

$$\nu(\mathcal{S}) = (\hat{\mu}_k(\mathcal{S}_x), \hat{\mu}_k(\mathcal{S}_y), \hat{\mu}_k(\mathcal{S}_{xy})) \in \mathbb{R}^{3m} \tag{5.5}$$

where $\mathcal{S}_x = \{x_i\}_{i=1}^n$, $\mathcal{S}_y = \{y_i\}_{i=1}^n$, and $\mathcal{S}_{xy} = \{(x_i, y_i)\}_{i=1}^n$. That is, $3m$ features are constructed, corresponding to the marginal distribution $\mathbb{P}_X$, marginal distribution $\mathbb{P}_Y$, and joint distribution $\mathbb{P}_{XY}$. Additionally, in a multivariate scenario, the RCC exploits the following equation to featurize the conditional joint distribution $\mathbb{P}_{XY|Z}$ using the sample $\mathcal{S} = \{(x_i, y_i, z_i)\}_{i=1}^n$, where $\mathcal{S}_x = \{x_i\}_{i=1}^n$, $\mathcal{S}_y = \{y_i\}_{i=1}^n$, and $\mathcal{S}_{xyz} = \{(x_i, y_i, z_i)\}_{i=1}^n$:

$$\nu(\mathcal{S}) = (\hat{\mu}_k(\mathcal{S}_x), \hat{\mu}_k(\mathcal{S}_y), \hat{\mu}_k(\mathcal{S}_{xyz})) \in \mathbb{R}^{3m} \tag{5.6}$$

5.3.2 Classification

Bivariate

Suppose a set of bivariate labeled data $\mathscr{D} = \{(\mathcal{S}_i, l_i)\}$ in which $\mathcal{S}_i = (X_i, Y_i) = (x_{ij}, y_{ij})_{j=1}^{n_i}$ is a sample drawn from the joint distribution $\mathbb{P}_{X_iY_i}$ and l_i characterizes the causal relationship between X_i and Y_i.

The goal is to build a classification model $\mathbf{\Pi}$ that estimates the causal label l_i from the sample $\mathcal{S}_i$. The model does not estimate the label directly from the sample but from the vector of features constructed from the sample:

$$\mathbf{\Pi} : \nu(\mathcal{S}_i) \longrightarrow \hat{l}_i \tag{5.7}$$

The label estimation formulated in this way is a supervised learning problem. To obtain such a classification model, large amounts of labeled data must be given to $\mathbf{\Pi}$ so that the model learns the relationship between samples and labels. This is done in the training phase of the model, wherein each sample is featurized and fed to the model twice, one with $\mathcal{S}_i = (X_i, Y_i)$ and the other with $\mathcal{S}_i' = (Y_i, X_i)$. Once the model is trained, the test phase begins. In the test phase, each sample is first featurized and then given to $\mathbf{\Pi}$ so that the model predicts the causal label.

$$\left((X_i, Y_i), l_{xy}^{(i)}\right), \left((Y_i, X_i), l_{yx}^{(i)}\right) \tag{5.8}$$

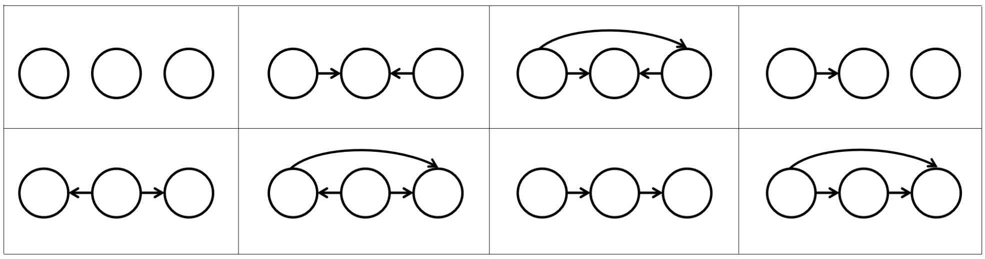


Figure 5.5: The eight possible DAGs on three variables under the assumption that node pairs have maximally one common cause.

Remark 5.4. *To build an arbitrarily large training dataset, the original RCC paper proposes a heuristic generative process to generate synthetic data samples. The said process is based on a Gaussian mixture model whose hyperparameters are tuned in accordance with the scenario under consideration. I skip discussing this process because it does not apply to quantum scenarios. Instead, I propose an alternative generative model in the next section.*

Multivariate

The method sketched above is applicable to bivariate scenarios because the number of causal labels definable over two variables is very limited. For more than two variables, the number of possible DAGs super-exponentially grows with the number of nodes; hence, it is almost impossible to associate each causal DAG with one classification label. Consequently, the sketched strategy must be modified to recover the causal graph in a multivariate scenario.

First, notice that a causal edge $X_i - X_j$ between nodes X_i and X_j in a DAG $\mathcal{G}$ can be characterized by three labels: "forward" ($X_i \rightarrow X_j$), "backward" ($X_i \leftarrow X_j$), and "disconnected" ($X_i \perp\!\!\!\perp X_j$). Denote these labels by +1, −1, and 0. Recovering the causal graph in a multivariate scenario can be converted into a three-class classification problem in which the classifier $\mathbf{\Pi}$ predicts the causal label of all pairs X_iX_j in a graph.

Nonetheless, the feature construction in the multivariate scenario for pair X_iX_j must be computed conditioned on other variables. That is, instead of featurizing the marginal joint distribution of the pair, conditional joint distributions must be featurized using Equation 5.6.

Such a featurization scheme would be more feasible if one imposes the simplifying constraint that each pair of variables has at most one common cause. Therefore, to featurize pair X_iX_j, it is enough to featurize conditional relations type $(2,1)$, i.e., $X_iX_j|X_k$. In such a situation, the causal DAG among every node triple is one of the eight graphs shown in Figure 5.5.

To synthesize a training dataset for the classifier $\mathbf{\Pi}$, the authors of the RCC suggest to randomly draw N graphs from the eight possible DAGs in Figure 5.5 and generate samples $S_i = (X_i, Y_i, Z_i)$ using the drawn graphs $\mathcal{G}_i$:

$$\begin{aligned} &\left((X_i, Y_i, Z_i), +l_{xy}\right), \left((Y_i, Z_i, X_i), +l_{yz}\right), \left((X_i, Z_i, Y_i), +l_{xz}\right), \\ &\left((Y_i, X_i, Z_i), -l_{xy}\right), \left((Z_i, Y_i, X_i), -l_{yz}\right), \left((Z_i, X_i, Y_i), -l_{xz}\right), \end{aligned} \tag{5.9}$$

where l_{xy}, l_{yz}, and l_{xz} characterize the causal relationships between (X_i, Y_i), (Y_i, Z_i), and (X_i, Z_i) in graph $\mathcal{G}_i$.

Once the classifier $\mathbf{\Pi}$ is trained, it can take a sample $\mathcal{S} = (X_i, X_j, X_k)$ as input and estimate the causal label between X_i and X_j given X_k. More precisely, the classifier returns the probabilities of three labels (i.e., forward, backward, and disconnected), and we take the label with the highest probability as the most probable label.

To recover the full causal DAG from the sample $\mathcal{S} = (X_1, \ldots, X_p)$, it is enough to predict the label of each pair X_iX_j separately. To this end, $p-2$ conditional relations like $X_iX_j|X_k$ must be considered to predict the label of the pair X_iX_j. To estimate the label X_iX_j, one must predict the probabilities of all three labels and average over for all $p-2$ conditional relations. Finally, the label with the highest probability can be selected as the label of X_iX_j.

5.4 Quantum Causal Discovery

The current project generalizes the RCC in two ways. On the one hand, I use three-dimensional variables instead of one-dimensional variables. Thus, I use the RCC to learn the causal direction among quantum states rather than classical variables. Therefore, in Equations 5.5 and 5.6, I replace X, Y, Z respectively by ρ_X, ρ_Y, and ρ_Z. Note that no part of the kernel mean embedding framework prevents us from such a generalization.

On the other hand, as mentioned in Remark 5.4, the generative process used by the original RCC does not apply to the multivariate quantum scenario. Thus, my second generalization introduces an alternative procedure for synthesizing training data for the RCC classifier. The said procedure is based on the CGNN algorithm. However, as we shall see, its implications go beyond the CGNN and RCC algorithms and are tied to discussions in quantum tomography with incomplete data.

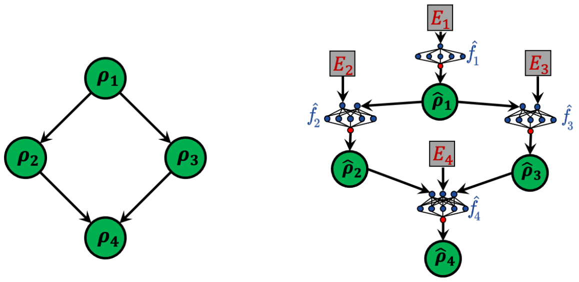
5.4.1 Generative Learning

Let $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_p\}$ represent the states of qubits in a $p-$variate quantum scenario. The aim is to build a generative model $\mathscr{F}$ that takes an arbitrary causal graph $\mathcal{G}^{(l)}$ as input and generates a synthetic sample $\hat{\mathcal{S}}^{(l)} = (\hat{\rho}_1^{(l)}, \ldots, \hat{\rho}_p^{(l)})$:

$$\mathscr{F} : \mathcal{G}^{(l)} \longrightarrow \hat{\mathcal{S}}^{(l)}. \tag{5.10}$$

To achieve such a generative model, I take the CGNN algorithm as a basis and insert two modifications into it. To start, imagine that graph nodes in the CGNN algorithm represent states of qubits rather than values of classical random variables. Figure 5.6 depicts the corresponding FCM initialized by the CGNN. In this model, neural networks map the Bloch vectors of parent nodes into the Bloch vectors of child nodes. That is, each neural network estimates the functional form of a quantum channel from parents to children, i.e.,

$$\hat{\rho}_i = \hat{\mathcal{E}}_i(\rho_{\mathrm{pa}(\rho_i;\mathcal{G})}, E_i). \tag{5.11}$$



$$\begin{cases} \hat{\rho}_1 = \hat{f}_1(E_1) \\ \hat{\rho}_2 = \hat{f}_2(E_2, \hat{\rho}_1) \\ \hat{\rho}_3 = \hat{f}_3(E_3, \hat{\rho}_1) \\ \hat{\rho}_4 = \hat{f}_4(E_4, \hat{\rho}_2, \hat{\rho}_3) \end{cases}$$

Figure 5.6: The generalized CGNN: the nodes represent quantum states, while the edges represent quantum channnels.

Therefore, the CGNN plays the role of a generative model that receives a causal graph $\mathcal{G}^{(l)}$ as input and generates a synthetic sample $\hat{\mathcal{S}}^{(l)}$. However, the problem with this strategy is that such a model can synthesize samples merely for graph $\mathcal{G}^{(l)}$. To synthesize samples for another given graph, one has to initialize another FCM whose causal structure corresponds to the new graph. Therefore, to synthesize samples for m causal graphs, m FCMs must be initialized. What is worse is that each initialized model requires a separate training dataset to be trained on. Therefore, m real datasets must be available to train the m initialized FCMs.

Such a training strategy is practically ineffective because the number of possible DAGs in a multivariate scenario can be extremely large, and it is almost impossible to collect true data for all these possibilities. For instance, for an 8−variate scenario, $783,702,329,343$ causal graphs are definable!

To overcome this problem, I insert a second modification to the standard CGNN algorithm, according to which the algorithm initializes only one FCM $\mathscr{F}_\Theta$ for all possible DAGs definable over the p variables. In such a model, each node is connected to all other nodes except itself. In the 8−variate scenario, this means the following functional relationships:

$$\begin{aligned} &\hat{\rho}_1 = \hat{f}_1(E_1, \hat{\rho}_2, \dots \hat{\rho}_8) \\ &\vdots \\ &\hat{\rho}_8 = \hat{f}_8(E_8, \hat{\rho}_1, \dots, \hat{\rho}_7) \end{aligned} \tag{5.12}$$

The model $\mathscr{F}_\Theta$ is not inherently causal since it posits that variables depend on each other

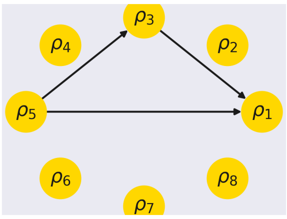


Figure 5.7: A graph in which only three nodes are causally related while the other nodes are causally disconnected.

regardless of their causal relationships. However, by selecting a fragment of the model that is compatible with a particular causal DAG, it is possible to convert it into a causal model while leaving the rest of the model acausal.

Assuming the model is already trained and its parameters Θ are optimized, one can activate a fragment of the model that is consistent with the DAG $\mathcal{G}$ and freeze the remaining mechanisms. The activated fragment resembles the functional causal model initialized by the standard CGNN and can generate synthetic data for the given DAG. However, the frozen fragment is acausal and should not be modified when synthesizing data for the given DAG.

To illustrate the idea, consider Figure 5.7 in which only three variables are causally related. The structural equations for such a DAG are

$$\begin{aligned} \hat{\rho}_1 &= \hat{f}_1(E_1, \vec{0}, \hat{\rho}_3, \vec{0}, \hat{\rho}_5, , \vec{0}, \vec{0}, \vec{0}) \\ \hat{\rho}_3 &= \hat{f}_3(E_3, \vec{0}, \vec{0}, \hat{\rho}_5, \vec{0}, \vec{0}, \vec{0}, \vec{0}) \\ \hat{\rho}_k &= \hat{f}_k(E_k, \vec{0}, \vec{0}, \vec{0}, \vec{0}, \vec{0}, \vec{0}, \vec{0}) \quad \forall k \in \{2, 4, 5, 6, 7, 8\} \end{aligned} \tag{5.13}$$

where vectors $\vec{0} = [0, 0, 0]^\intercal$ are used whenever there is no causal connection between the nodes. These vectors are used to temporarily destroy the edges that are absent in the given DAG.

Figure 5.8 depicts how the model $\mathscr{F}_\Theta$ synthesizes data for these structural equations. In this figure, only a fragment of $\mathscr{F}_\Theta$ that is compatible with the given graph is activated while the other edges are temporarily destructed.

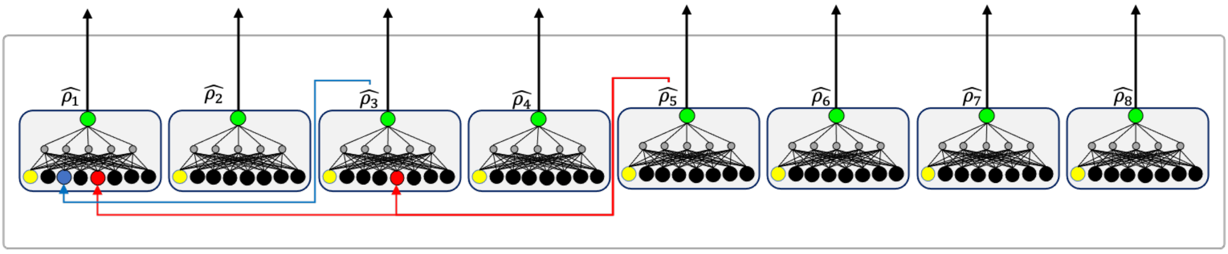


Figure 5.8: An example of activating and freezing a model.

Model Training

Now I describe how to train the model $\mathscr{F}_\Theta$. To this end, a training dataset $\mathscr{D}_{tr} = \{(\mathcal{S}^{(l)}, \mathcal{G}^{(l)})\}_{l=1}^{m}$ consisting of m pairs are needed, where $\mathcal{S}^{(l)}$ and $\mathcal{G}^{(l)}$ are true data samples and causal DAGs. There are two interesting points about the required samples and graphs, both essential from the perspective of quantum tomography.

First, there is no need to use an excessive number of $783,702,329,343$ true data samples to train $\mathscr{F}_\Theta$. It suffices to provide the model with a "sufficient number" of training pairs, which means the number that enables the model to reconstruct the rest of the data. The user can determine the appropriate number based on the needs she expects from the model.

Second, it is unnecessary to use complete data samples to train $\mathscr{F}_\Theta$. Instead, one can train the model with samples from the joint distribution over a subset of variables. For example, in the 8−variate scenario, 3−variate data samples like $\mathcal{S}^{(l)} = \left(\rho_i^{(l)}, \rho_j^{(l)}, \rho_k^{(l)}\right)$ can be used for training. To accomplish this, one should convert the model in a manner depicted in Figure 5.8 and minimize the following training loss function:

$$\mathscr{L}_{tr}(\Theta) = \sum_{l=1}^{m} \left\| \mathcal{S}^{(l)} - \hat{\mathcal{S}}^{(l)}(\Theta) \right\| \Longrightarrow \nabla_\Theta \mathscr{L}_{tr} \overset{!}{=} 0 \Longrightarrow \Theta^\star = \mathrm{argmin}_\Theta \, \mathscr{L}_{tr}. \tag{5.14}$$

The loss minimization scheme described in Equation 5.14 optimizes the model for synthesizing 3−variate samples. Although the model generates 8−variate samples, the generated data is guaranteed to be accurate up to the level of 3−variate joint distributions. One can, of course, train a more powerful generative model that captures joint distributions over more than three variables by training the model on more complete data samples.

For the current project, I trained the model $\mathscr{F}_\Theta$ using merely 3−variate samples on the eight graphs shown in Figure 5.5. While I could expand the training data and exploit more sophisticated training samples, I decided to stick to 3−variate samples for two reasons.

On the one hand, from an experimental point of view, it is much easier to perform joint measurements on a small number of qubits. Collecting 3−variate data samples is much easier than, say 6−variate samples. Thus, as a quantum causal discovery task, the preference is to rely on samples that are easier to be collected in a practical scenario.

On the other hand, recall that the primary purpose of the present generative model is to synthesize samples for the RCC classifier. Given the simplifying constraint of the RCC that there is maximally one common cause for each pair of nodes, it is sufficient to synthesize samples that are accurate with respect to conditional relations of type $(2, 1)$ and hence 3−variate joint distributions. Hence, I did not expand the training data because the parsimonious 3−variate samples led to achieving the project goal, i.e., to recover the causal graph by the RCC.

5.4.2 Pipeline

This section summarizes the training and prediction pipelines used in the current project. I used 2688 3−variate samples for training $\mathscr{F}_\Theta$. The said samples are used twice: (1) for training the generative $\mathscr{F}$ and (2) for training the discriminative $\mathbf{\Pi}$. Once the generative $\mathscr{F}$ is trained, it can synthesize 8−variate samples for 8−node DAGs. For the current project, I synthesized 400 8-variate samples associated with 400 random 8−node DAGs.

The raw graph created based on the RCC predictions may not be very accurate at the global (i.e., graphical) level because it is based on the local information between the nodes. To improve the quality of the graph, I apply two post-processing steps that utilize the $\mathscr{F}$ model. These steps are inspired by Goudet et al. (2018) and are described below.

In the first step, less important edges are removed. To determine the importance of an edge, I synthesize data samples with and without that edge using $\mathscr{F}$ and compare the quality of the samples with the original data. If the quality difference between the two

Algorithm 3 The RCC Training Workflow

```
Input: $\mathscr{D}_{tr} = \{(\mathcal{S}_{tr}^{(l)}, \mathcal{G}_{tr}^{(l)})\}_{l=1}^{m}$ (training data) and $\{\mathcal{G}^{(l)}\}_{l=1}^{n}$ ($p$−node DAGs)
Initialize a discriminative model $\mathbf{\Pi}$ and a generative model $\mathscr{F}$.
Train $\mathscr{F}$ on $\mathscr{D}_{tr}$.
Synthesize $p$-variate samples for the $p$−node graphs using the trained $\mathscr{F}$:
    $$\hat{\mathscr{D}} = \{(\hat{\mathcal{S}}^{(l)}, \mathscr{F}(\mathcal{G}^{(l)}))\}$$
Initialize training data for the discriminative model $\mathscr{D}_{tr}^{\Pi} = \{\}$.
for each 3−variate sample $\mathcal{S}_{ijk} \in \mathscr{D}_{tr} \cup \hat{\mathscr{D}}$ do
    Compute the six permutations in Equation 5.9.
    Featurize the permutations: $f_{ijk} = \nu(\mathcal{S}_{ijk})$.
    Add pair $(f_{ijk}, l_{ij})$ to $\mathscr{D}_{tr}^{\Pi}$.
end for
Train $\mathbf{\Pi}$ on $\mathscr{D}_{tr}^{\Pi}$.
Output: $\mathbf{\Pi}$ (trained discriminator)
```

samples is not significant, the edge is removed. Otherwise, it is kept (the significance level is determined in the validation phase). In the second step, the remaining edges are randomly reversed, and the corresponding data quality is checked. If the quality improves significantly, the edge is reversed; otherwise, it is kept in its original direction.

5.5 Results

This section evaluates the performance of the RCC in simulated quantum scenarios. As mentioned earlier, the evaluation is carried out in three quantum scenarios, the first two being bivariate and the third being multivariate. Before describing the RCC performance, let us describe the simulated data for these scenarios.

For the bivariate scenarios, 5000 data samples are simulated, and each of them is physically different. In other words, the quantum operations connecting ρ_X and ρ_Y vary in each sample. Recall that each quantum operation contains adjustable physical parameters, such as the strength of rotations, noise, and the weights to combine different types of quantum channels. To simulate each sample, random values for these parameters are drawn, and the sample is generated using these values. After simulating the samples, they are split

Algorithm 4 The RCC Prediction Workflow

```
Input: $\mathcal{S} = (\rho_1, \dots, \rho_p)$ ($p-$variate sample), $\mathbf{\Pi}$ (trained discriminator)
Initialize an $p \times p$ adjacency matrix $\mathcal{A}_{raw}$
for each pair $(\rho_i, \rho_j)$ do
    Extract all 3−variate samples $\mathcal{S}_{ijk}$.
    Featurize the samples: $f_{ijk} = \nu(\mathcal{S}_{ijk})$ .
    Predict probabilities for causal labels: $\vec{P}_{ijk} = P^{ijk}_{\rightarrow}, P^{ijk}_{\leftarrow}, P^{ijk}_{\perp\!\!\!\perp} = \mathbf{\Pi}(f_{ijk})$
    Average the probabilities over conditioning variables: $P_{ij} = \sum_k \vec{P}_{ijk}$.
    Fill out $\mathcal{A}_{raw}[i, j]$ in accordance to $\max(P_{ij})$.
end for
Convert $\mathcal{A}_{raw}$ into $\hat{\mathcal{G}}_{raw}$
if post-processing then:
    Post-process $\hat{\mathcal{G}}_{raw}$ to obtain $\hat{\mathcal{G}}_{post}$.
    Return $\hat{\mathcal{G}}_{post}$.
else:
    Return $\hat{\mathcal{G}}_{raw}$.
end if
Output: $\hat{\mathcal{G}}_{post}$ (predicted graph)
```

into training and test sets with a $70 : 30$ ratio. The classifier is trained and tested on the training and test sets.

The multivariate scenario is physically different from the first two scenarios. In this scenario, each node has a unique set of physical parameters that remain constant across the samples. For example, in all samples, the quantum channel of the i^{th} node takes the constant form $\mathcal{E}^{(i)}(\rho) = \sum_k w_k^{(i)} E_k \rho E_k^{\dagger}$, and the weights $w_k^{(i)}$ do not vary. This assumption is well-motivated, especially if an underlying quantum process exists over all nodes. In such a situation, the physical parameters of each node remain constant because they encapsulate the functional form of the causal mechanism of the nodes.

Another point to note regarding the multivariate scenario is that the type of training samples is different from the validation and test samples. Specifically, 2688 3−variate samples are simulated for training, while 134 and 135 8−variate samples are simulated for validation and testing, respectively.

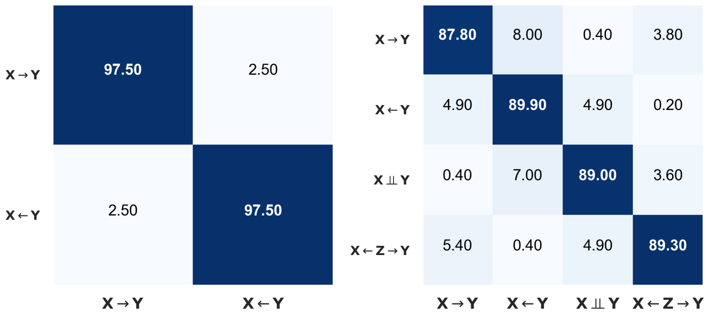


Figure 5.9: Confusion matrices for the bivariate scenarios. Left: performance in the first scenario; Right: performance in the second scenario.

5.5.1 Bivariate

Figure 5.9 displays the confusion matrices for the first and second bivariate scenarios. The rows and columns of the matrices correspond to the true and predicted causal labels, respectively. The element $m_{ij} = M[i,j]$ in each matrix represents the percentage of samples with the true label i that are predicted with label j. As shown in the figure, the RCC accurately learns the causal labels in both scenarios, with an accuracy of 97.50% and 89.0%, respectively.

The fact that the physical parameters of the test samples are different from the training samples shows that the RCC detects causal labels irrespective of the underlying parameters of a scenario. Said differently, the patterns the RCC learns during the training phase are independent of the underlying parameters of data-generating processes; these are the asymmetries in the probability distributions. Thus, as long as these asymmetries exist in a sample, the algorithm can detect the direction of causation, but it may fail if the asymmetries are weak or absent.

Figure 5.10 depicts a realization of the previous claim. In this figure, the accuracy of the RCC is shown for different quantum channels. According to the figure, the accuracy of the RCC for the identity channel is approximately 50%, i.e., the RCC is absolutely unable to distinguish the causal direction in highly symmetric schemes such as the identity channel.

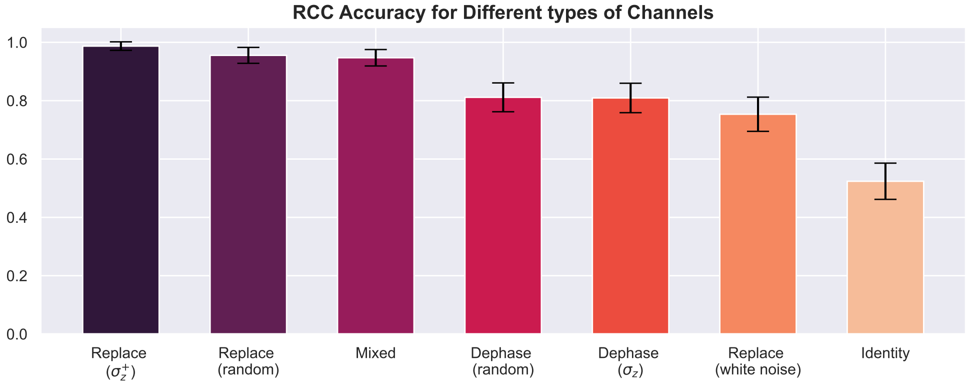


Figure 5.10: The RCC accuracy for different channels.

Remark 5.5. *From a causal modeling perspective, the direction of the identity channel is not identifiable because the channel leads to a linear model with Gaussian noise. Such models are identifiable if and only if the noise is non-Gaussian. For non-Gaussian noises, the functional relationship between ρ_X and ρ_Y is described by a LiNGAM, which has a theoretical guarantee for its identifiability (see Shimizu et al. 2006).*

5.5.2 Multivariate

Figure 5.11 shows the confusion matrices of the multivariate scenario before and after post-processing the prediction. Recall that three causal labels can be assigned to a pair of nodes: forward (→), backward (←), and disconnected ($\perp\!\!\!\perp$). The matrices indicate the performance in classifying these three labels in the 135 testing graphs. Thus, the performance is computed based on $\binom{8}{2} \times 3 \times 135$ predictions.

Before and after post-processing, 6.65 and 3.62 edges are misclassified per graph, i.e., 84.0% and 85.6% accuracy. Figure 5.12 shows an example of one of the predicted graphs.

5.6 Discussion

In the previous section, I showed the success of the RCC in different quantum scenarios. Specifically, combining the discriminative RCC with the generative CGNN delivered a

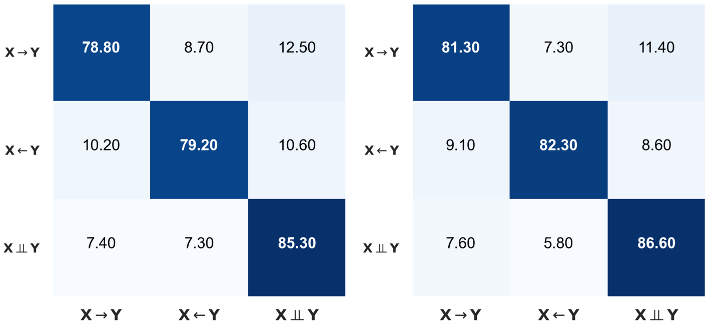


Figure 5.11: Confusion matrices for the multivariate scenario. Left: before post-processing; Right: after post-processing.

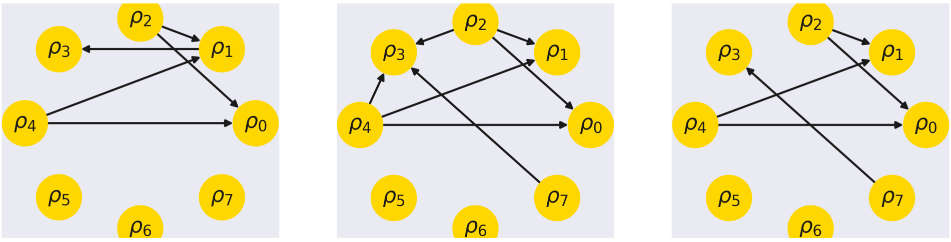


Figure 5.12: Example of a multivariate predication. Left: true graph, middle: raw prediction, and right: post-processed prediction.

powerful algorithm that can recover the causal graph of multivariate quantum scenarios with limited training data. In this section, I will further describe the method and its consequences. I begin with a brief overview of previous work in quantum causal discovery and then discuss the implications of the current method. Finally, I present ideas for future work.

5.6.1 Past

In recent years, several conceptual frameworks have been proposed to study the causal relationships among quantum systems. The central idea of the said frameworks is that causal relations among quantum systems do not follow the classical causal modeling framework and, therefore, generalized principles of causal modeling must be employed to describe causal relations in the quantum realm. Examples include Henson et al. (2014), Pienaar & Brukner (2015), Costa & Shrapnel (2016), Allen et al. (2017), Barrett et al. (2019, 2021).

While a solid ground has been formed at the theoretical level to study causality in the quantum realm, very few works have been done at the algorithmic level. The available algorithms mostly apply to simple scenarios such as when the number of variables is small (see, e.g., Ried et al. 2015, Fitzsimons et al. 2015, Chiribella & Ebler 2019). The following two examples are more interesting because they pose fewer restrictions on the scenario under consideration.

The first example is the algorithm introduced by Giarmatzi (2019). The algorithm is a general quantum causal discovery algorithm formulated based on the "process matrix formalism" introduced by Pienaar & Brukner (2015). For a multivariate quantum scenario, the algorithm receives the process matrix as input and reconstructs the causal graph through the sparsity patterns in the process matrix. The algorithm is general in that it handles several causal circumstances. For instance, it detects whether the given process matrix is "causally ordered" (i.e., whether the causal structure of the scenario admits a DAG) or whether the process matrix represents a "Markovian" system (i.e., whether there are latent common causes between the variables). Despite such generality, the fact that the input is a process matrix means that before causal discovery, a full quantum process tomography must be performed. Since process tomography for high-dimensional quantum systems requires a large number of measurements, the algorithm can hardly be exploited in a real scenario.

The second example is the algorithm introduced by Bai et al. (2022). The algorithm aims to solve the "causal unraveling" problem, formulated as follows. Given an unknown quantum process with inputs $X_1, \ldots, X_n$ and outputs $Y_1, \ldots, Y_m$, determine whether the process can be broken into a set of interactions from a partition of inputs to a partition of outputs and if so, determine which inputs and outputs are involved in each interaction. In other words, the aim is to explore the causal order between the inputs and outputs of an unknown quantum process through a series of measurements. The authors provide an efficient algorithm that does not require full quantum process tomography. The key idea is to perform a series of independence tests between the density matrices of the inputs and outputs and recover the causal orders step by step. Hence, the algorithm can be construed as a constraint-based method inherently similar to the PC algorithm (Spirtes et al. 2000).

The method presented in this chapter is most similar to the approach taken by Bai et al. On the one hand, both methods do not require a complete tomographic characterization of the process under consideration. Instead, they both infer the causal directions from pairs of variables. On the other hand, my multivariate scenario can be reformulated as a causal unraveling problem. For this, it is enough to divide the variables of the multivariate scenario into two subsets (input and output) and pose that input variables are exogenous (i.e., causal interactions always start from the input variables and end with output variables). By doing so, causal discovery in my method is equivalent to finding causal orders in Bai et al. method. Nonetheless, the use of Machine Learning in quantum causal discovery is unique to the method presented here. To the best of my knowledge, neither Bai et al. (2022) nor any quantum causal discovery algorithms use Machine Learning as an organic part of the algorithm.[2]

[2] Notice that Machine Learning approaches are used indirectly in tasks such as performing process tomography more efficiently (see, e.g., Torlai et al. 2020).

5.6.2 Present

The results obtained in this chapter have important implications in various research fields. I divide these fields into Quantum Foundations, Artificial Intelligence, and Quantum Engineering. In the following, I address each of the parts.

Quantum Foundations

The main idea of the current project is to apply classical causal modeling tools to the state of quantum systems instead of classical quantities such as measurement outcomes. Consequently, I implicitly posit a causal model in which the variables are quantum states and the mechanisms are quantum channels. The causal model constructed in this way is conceptually similar to quantum causal models (QCMs). To see the reason, let us focus on the framework proposed by Costa & Shrapnel (2016) as an example of a QCM.

Imagine a causal graph where each node represents a local lab with one input and one output. The nodes are connected by edges transmitting the outputs of parent labs to the inputs of child labs. Inside each lab, an experimenter exists who can perform any intervention on the input system legitimated by axioms of Quantum Mechanics. Following interventions, the experimenters emit the resulting systems as outputs. In such a model, nodes are represented by collections of CP maps characterizing the evolution of quantum systems before and after entering the labs. Moreover, causal mechanisms are represented by CPTP maps transferring the outputs of parent nodes to the inputs of child nodes.

To see the similarities between the above and current approaches, notice that the current project assumes that there is no intervention at the labs.[3] Therefore, a system entering and exiting a lab maintains its quantum state. Thus, the states themselves are sufficient for representing the nodes, and there is no need for CP maps. Regarding the causal mechanisms, both approaches utilize quantum channels to represent the transmissions of parent systems to child systems.

Given such a conceptual similarity, the success of the RCC and the CGNN in causal

[3] I make this simplifying assumption to investigate the learning problem within an observational scheme.

discovery among quantum systems should be construed as support for QCMs. In other words, the fact that standard causal modeling and Machine Learning tools can be used for quantum states suggests that the core ideas of causal modeling are valid in the quantum realm. Especially because other fundamental conceptions in causal modeling (such as the causal Markov condition and the faithfulness) are generalized by QCMs in a reasonably straightforward way, and quantum systems are shown to be compatible with these generalizations (see, e.g., Shrapnel 2016).

Therefore, it seems that many conflicts between Quantum Mechanics and causal modeling can be resolved if one applies the ideas of causal modeling at the level of quantum states rather than observed quantities. While this might be a fair conclusion, an unresolved challenge persists, especially if QCMs are supposed to be more than mathematical machinery for quantum computations. What would be the nature of a causal relationship prescribed between non-observable entities such as quantum states? Or, what is the ontological status of a causal mechanism acting among non-observables? In my opinion, this is the most challenging question for QCMs, and the answer directly relates to the discussions surrounding the reality of quantum states. To understand the status of causal mechanisms in a QCM, the status of quantum states must be determined first.

The status of the quantum state has been debated since the early days of Quantum Mechanics, and various answers have been given.[4]. I do not intend to enter into the debate about the reality of the quantum state because it is beyond the scope of the present discussion. However, I wish to emphasize the following two points.

First, understanding the nature of causal relationships in QCMs is not independent of understanding the nature of quantum states because these models study causal relationships not at the level of observables but at the level of quantum states. Of course, the

[4] According to the ontological models framework (Harrigan & Spekkens 2010), there are three possibilities for interpreting quantum states: (1) anti-realist ψ-epistemic view (quantum states are epistemic entities, but there is no deeper underlying reality), (2) realist ψ-epistemic (quantum states are epistemic, and there is an underlying ontic state), (3) realist ψ-ontic (there is an underlying reality, and quantum states correspond to states of physical reality). Note, however, that this characterization does not necessarily reflect all views. For example, Callender (2015) argues that there is a fourth view that takes quantum states as part of dynamical laws on top of ontic states.

fact that algorithms like the RCC can accurately learn the direction of causation between quantum states is good news for QCMs because it indicates the existence of such directionality at the level of states. However, to draw conclusions about the ontological status of the said directionality, it is necessary to clarify the ontological status of quantum states.

Second, as argued in Chapter 3, providing a "secure transition" from the quantum domain to the classical domain is a necessary condition to accept that a QCM tells a *causal* story about a scenario. The secure transition is meant to be an explanation of the classical limit and the origin of quantumness in an "unambiguous language" declared by Nils Bohr (Bohr 1949, p. 209). I believe the same doctrine applies to the results presented here. That is, to understand the ontological status of what the RCC has learned (i.e., causal directions among quantum states), one must provide a secure transition to translate the notion of causal direction from the level of states to the level of observables and interpret the results at the level of the latter. Such a secure transition might be accomplished by quantum decoherence (see, e.g., Zurek 2002). However, further studies should be conducted to confirm this hypothesis.

Artificial Intelligence

Apart from Quantum Mechanics, the implications of the current project can be studied in light of the interplay between Artificial Intelligence and causal modeling.

The first point is about discriminative causal learning algorithms such as the RCC. Unlike standard Machine Learning algorithms, these can be trained and tested on physically different datasets. For instance, in the Cause-Effect Pair Challenge (Guyon 2013) organized in 2013 on the Kaggle platform, the training and test datasets were selected from lots of different domains such as chemistry, climatology, ecology, economy, engineering, epidemiology, genomics, medicine, physics, and sociology. The high performance of an algorithm like the RCC on such data suggests that real-world data contains a series of universal causal signatures based on which the RCC could learn the direction of causation. If this conclusion is correct, Machine Learning helps one discover the causal footprints regardless of the underlying physical theories. Consequently, one can expect that future AI systems can learn causal facts about the physical world from the mere data (an example of such systems in causal imaging is studied by Lopez-Paz et al. 2017).

The next point is about generative causal learning algorithms such as the CGNN. Recall that in the variant of the CGNN utilized in this chapter, I use 3-variate samples for model training however synthesize 8-variate samples once the model is trained. Remarkably, the synthesized samples have high quality; they are distributionally very close to the corresponding true 8-variate samples. Therefore, even though the model never sees true 8-variate distributions during the training phase, it can estimate it from 3-variate distributions.[5]

This is an astonishing result; it is like giving the machine individual parts of a picture along with information about the relative position of each part and then asking the machine to put these parts together and build a new overall picture in accordance with a new configuration of the parts. The fact that the generative algorithm of this chapter was able to be trained well with minimal samples radiates the message that Machine Learning algorithms can work more efficiently with limited data if they have access to causal information.

Quantum Engineering

The project indicates that Machine Learning empowers one to causally characterize a quantum process even when limited information is available about the process. Therefore, unlike the discovery algorithm presented by Costa & Shrapnel (2016), which requires full quantum process tomography, the present algorithm performs causal discovery in the presence of limited training data. This feature is vital from a practical point of view, as the number of measurements required for a full process tomography grows exponentially with the number of systems involved.

[5] This strategy is justified only if one can postulate a “mother” generative process on top of all training and test data. That is, only if a single underlying process exists that generates all data samples, either 3-variate or 8-variate samples. The existence of such a process is guaranteed in the multivariate quantum scenario because all samples originate from a quantum process connecting the local labs.

Another interesting fact about the presented models is that they can be modified partially and used as an intermediate step for more expensive methods such as Costa & Shrapnel (2016). For example, penalization techniques can be employed to push the RCC classifier to spend less accuracy in "forward," and "backward" labels and instead increase the accuracy in "disconnected" label. The predictions of such a penalized model can then be used as the starting point for other algorithms. Considering the cost-efficiency of the present models, this strategy would be practically useful.

Apart from quantum process tomography, the current models can be trained to detect signaling relations for data transmission in quantum networks. Such an application is advantageous, especially in large quantum networks wherein full tomography is practically impossible.

Finally, a necessary condition for a causal discriminative algorithm to generalize well is seeing various types of asymmetries during the training phase. The more diverse the training data, the more generalizable the model in the test phase. Notably, the required asymmetries for training such algorithms can be partially obtained from other quantum experiments. Consequently, the methods described here would be even more affordable if Tübingen-like databases were created for real quantum experiments so that discriminative algorithms can employ them as part of their training data.

5.6.3 Future

The current project can be extended in several directions. I divide these directions into conceptual and technical levels and explain each separately.

Conceptual Level

1. As conjectured previously, there should be some universal causal signatures relying on which the RCC detects the causal direction. What is the interpretation of these signatures? What is the relationship between the features extracted in the feature construction stage and the underlying physical parameters of a distribution? One

strategy to answer the above questions is checking which features are most important to the RCC classifier. In the Machine Learning literature, many techniques have been invented for this task under the title of "input feature attribution." For example, in computer vision, there are methods to identify an image's pixels (i.e., features) based on which a classifier estimates a label. This technique is a subset of the "explainable AI" field, and it can be imagined to be extended to the field of causal modeling and quantum causal modeling (for a review of methods in the context of deep neural networks, see, e.g., Samek et al. 2021).

2. In classical causal modeling, "identifiability theorems" specify certain conditions under which the detection of causal direction from observational data is theoretically guaranteed. LiNGAM (Shimizu et al. 2006), IGCI (Janzing et al. 2012), ANM (Hoyer et al. 2008, Mooij et al. 2016), and PNL (Zhang & Hyvärinen 2009) are algorithms based on identifiability theorems. Extending these results to the quantum domain can lead to new discovery algorithms that guarantee the identifiability of the causal direction for specific quantum processes. I mentioned an exemplary model in Remark 5.5, but this can be followed more systematically.

Technical Level

1. The current method can be combined with constraint-based algorithms such as the algorithm presented in Bai et al. (2022). The RCC and the CGNN are based on pairwise methods, so their estimates are inherently local. On the contrary, the estimates of a constraint-based algorithm are inherently global. For this reason, combining such algorithms can increase their power.

2. The current project is based solely on observational schemes. To make it more practical, it should be combined with interventional schemes. That is, a principled way must be found to combine the information gathered through both schemes. Machine Learning ideas can be a breakthrough again. For example, Lee & Bareinboim (2018) has shown that reinforcement learning can be used to efficiently understand which

node and when to intervene. Another example is active learning, in which the machine asks the user to collect more information about the parts that the machine is more uncertain about (see, e.g., He & Geng 2008).

3. The current project focuses only on qubits. A more general algorithm would consider quantum systems with arbitrary input and output dimensions. Although this is a straightforward generalization, it guarantees the method's applicability for practical purposes.

Chapter 6

Conclusion

In this thesis, I have used Machine Learning to investigate the causal problem of entanglement. The first three chapters provided an overview of the research motivations, required tools, and models proposed in the literature to solve the causal problem. Chapters 4 and 5 introduced two Machine Learning algorithms for two quantum scenarios and discussed their results. Rather than repeating specific implications and suggested lines for future research in this chapter, I will provide more general explanations of the two main contributions of this thesis: (1) the implications of recent findings for the interventionist account of causation in the quantum domain, and (2) the relevance of Machine Learning for discussions related to the philosophy of physics. In the final section, I will briefly discuss the spatiotemporal problem of entanglement and its relationship with the causal problem.

6.1 Interventionism after Quantum Correlations

Let us take a step back and re-examine the causal problem of entanglement. The problem demonstrates that classical causal models cannot produce a graph that explains the statistics of scenarios such as EPR-Bell. Therefore, Causal Bayesian networks do not apply to the quantum domain due to conflicts in assumptions such as the causal Markov and causal faithfulness conditions. This raises the question of what consequences the causal problem has for the interventionist account of causation in the quantum domain.

One possible response is that the causal problem demonstrates the inadequacy of the interventionist view in the quantum domain. Therefore, alternative accounts of causation must be used to explain quantum correlations. For example, the causal process view or the counterfactual account of causation may be used to explain quantum correlations. I have no issue with this response and believe that a pluralistic view of causation is necessary due to the many peculiarities of Quantum Mechanics.

However, if we wish to retain the interventionist account in the quantum domain, perhaps because we believe it provides a precise way to determine effective strategies, this thesis becomes relevant. The results of Chapters 4 and 5 demonstrate that a version of interventionism can be preserved in the quantum domain if we are willing to accept one of the following two options: (1) causal relata in causal models for quantum systems are not classical variables, or (2) causal conditions such as the faithfulness condition for quantum systems require generalization. In the following, I will discuss these two conditions and their implications.

Extended Causal Relata

In standard models of causal Bayesian networks, the causal relata are classical random variables whose values indicate the properties of the variables under consideration. However, in the discovery algorithm presented in Chapter 5 and more generally in quantum causal models, the causal relata are respectively the states of quantum systems and collections of CP maps. If interventionism is to be applicable in the quantum domain, we must accept that the causal relata and causal mechanisms between them will be unconventional.

The extent to which such a model can be considered causal has already been discussed in Sections 3.6 and 5.6.2, where I put forward two criteria for a quantum causal model to be considered causal. The first criterion is whether the model provides a secure transition from the quantum domain to the classical domain. The second criterion is the ontological status that the desired model assigns to quantum states. I will not repeat the discussion of these criteria here.

However, recognizing such a model has at least two motivations. First, quantum systems lack measurement-independent properties, and thus their description using classical random variables is unlikely, if not impossible. Second, interventions on quantum systems can be described in the most general way through CPTP maps, and thus, causal relata must be defined in a way that allows such intervention. As Woodward argues, the problem of variable choice should be approached within a means/ends framework, where cognitive inquiries can have various goals or ends, and candidate criteria for variable choice are justified by showing that they are effective means to these ends (Woodward 2016*b*, p. 1051).

Extended Faithfulness

To maintain the interventionist account in the quantum domain, an alternative approach is to keep the causal relata in causal Bayesian networks classical and modify the classical conditions of causal modeling, specifically the faithfulness condition. In such a view, the variables and latent confounders in the EPR-Bell scenario should all be expressed with classical random variables, but the faithfulness should be modified, recognizing that at least the standard form of the faithfulness does not hold in this scenario. In such a view, the fact that certain interpretations of Quantum Mechanics violate the faithfulness is not only their weakness but shows their virtue in demonstrating the true source of conflict between causal models and quantum correlations, namely the faithfulness.

In Chapter 4, I followed this path and evaluated models regardless of the faithfulness. However, why is the faithfulness unreliable in the EPR-Bell scenario? I explained in detail in Section 4.7.2 why the faithfulness may need modifications when dealing with quantum correlations. Here, I will briefly review the argument.

One understanding of the faithfulness is to say that if two variables are causally dependent, they must be statistically dependent, or equivalently if two variables are statistically independent, they are causally independent. The first part of the statement ($\not\perp_{\mathbf{c}} \Rightarrow \not\perp_{\mathbf{s}}$) can be justified in the way that the trace of causal connections must be reflected in the statistics of variables; otherwise, it would seem that causal connections conspiratorially hide themselves from the observer. The second part of the statement ($\perp\!\!\!\perp_{\mathbf{s}} \Rightarrow \perp\!\!\!\perp_{\mathbf{c}}$) is logically equivalent to the first part and is more useful for the following argument.

Recall that in the scenario discussed in Chapter 4, all variables are statistically independent, two by two. According to the faithfulness, all these variables should be causally independent two by two, and no causal connection should exist between any of them. Does one commit to a conspiracy if she asserts that some variables (e.g., the two outcomes) are causally connected? I do not think so because the trace of entanglement (i.e., the correlation between outcomes) can only be detected when the conditional dependence of outcomes given the two settings is checked.

Quantum correlations reveal a generic circumstance in which two marginally independent variables may be causally connected because the two variables are conditionally dependent. The faithfulness ignores this generic case and, therefore, cannot hold in the EPR-Bell scenario. Thus, if causal Bayesian networks are to have classical causal relata in the EPR-Bell scenario, they must modify the faithfulness. However, modifying the faithfulness alone is not enough because this condition is usually linked with the Markov condition trough the d-separation criterion.

To modify the faithfulness and the d-separation criterion in the face of quantum correlations, we need to consider the connection point of the two options mentioned for interventionism. The faithfulness and d-separation should be modified to sketch the causal relationships between quantum systems at the level of states, not just observables. In other words, the revised criteria should be in a form that considers the existence of a non-classical common cause for the two outcomes of the EPR-Bell scenario. This means that the interventionist account in the quantum domain needs to consider the existence of variables and mechanisms that cannot be shown in a standard way (classical random variables). If we are willing to accept this conclusion, we can have a generalized version of interventionism and benefit from it for identifying effective strategies in the quantum domain.

6.2 The Relevance of Machine Learning in Quantum Mechanics

While Machine Learning is often used in empirical sciences, this thesis demonstrates its relevance for foundational discussions, particularly in the context of Quantum Mechanics. Chapters 4 and 5 provide insights that can be analyzed in three areas, which I will discuss below.

Machine Learning for Model Selection

Chapter 4 demonstrates how Machine Learning can serve as a framework for model selection, providing new perspectives on traditional debates in Quantum Mechanics. The findings reveal that different loss functions can be customized to target various conditional relations, enabling the evaluation of each model's ability to learn such relations. Additionally, the flexibility of artificial neural networks allowed for the implementation of models whose causal content goes beyond causal graphs (e.g., SCC and VIM in Table 4.3), allowing for comparisons within a single framework.

More broadly, the implications of the findings in Chapter 4 extend to discussions about interpretability in Machine Learning. While model precision is usually the primary focus of evaluation, it is sometimes necessary to delve deeper into the model's inner workings to answer more complex questions. Chapter 4 utilized various visualizations to accomplish this, but other questions can be tackled in a similar manner. For example, one might explore the relationship between a model's internal parameters and the physical parameters of the system being considered. This approach has been used by Weinstein (2017, 2018) to investigate highly symmetric patterns in neural network weights and their connection to the Almada et al. (2016) argument regarding the violation of the faithfulness due to physical symmetries. These methods could be extended to other quantum foundations discussions, such as contextuality scenarios and PR correlations.

Machine Learning for Causal Asymmetries

The discovery algorithm presented in Chapter 5 consists of a generalized version of the RCC (Lopez-Paz et al. 2015) and CGNN (Goudet et al. 2018) pairwise discovery algorithms. Pairwise algorithms are designed to distinguish Markov equivalent scenarios (e.g., $X \rightarrow Y$ versus $X \leftarrow Y$), revealing that conditional (in)dependencies are not the only resource for extracting causal information from observational data. Instead, different types of asymmetries can restore the direction of causation. The RCC algorithm can automatically extract such asymmetries from statistical data in a domain-free manner, which makes it possible to train on data from multiple domains and test on data from other domains.

The success of this algorithm suggests that causal relationships have universal signatures independent of different domains or underlying data-generating processes. Real-world data contains causal asymmetries that are universal, cannot be detected by traditional methods based on conditional dependencies, and do not require intervention to be detected. Machine Learning can draw out these asymmetries.

This raises questions about the physical origin of such asymmetries and the philosophical implications of such asymmetries for the epistemology of causation. Although a complete analysis has not been done, specific situations have been studied in the literature.

For example, Climenhaga et al. (2021) discussed causal asymmetries from a philosophical perspective, although the authors restricted asymmetries to noises. The authors noticed that causal asymmetries allow one to extract causal knowledge from purely observational data without intervention, thus taking the causal asymmetries as an alternative to randomized controlled trials. For these reasons, the authors argued that models based on causal asymmetries, "call for a revision of the epistemology of causation." Climenhaga et al. (2021, p.15)

Regarding the physical origin of causal asymmetries, Janzing et al. (2016) argued that such asymmetries can arise from a more general independence principle between causal mechanisms. According to the principle, the causal mechanism of the cause variable is informationally independent of the conditional mechanism of effect. The authors provided a thermodynamic interpretation of the independence principle, suggesting that causal asymmetries might be related to the thermodynamic arrow of time.

Machine Learning provides significant capabilities to detect causal asymmetries, which can bring about new foundational perspectives in the philosophy of physics and understanding of the notion of causation. Although more analysis is required to answer the questions raised, the success of Machine Learning algorithms in extracting causal asymmetries highlights the need to explore this area further.

Machine Learning for Causal Compositions

The generative algorithm introduced in Chapter 5 is an example of a Machine Learning model can use causal information to compose inputs. Although the model was trained on 3−variate samples, it was able to perform well on 8-variate samples in the test phase. How was the model able to generate high-quality 8-variate samples without ever having seen any?

The answer lies in the model's architecture. The architecture is designed to utilize causal information to piece together the training samples like a puzzle and form an overall picture of the data-generating process. This helps the model gain a better understanding of the training samples and use them more intelligently to estimate the underlying data-generating process.

The basic idea of the said method is based on the principle of compositionality and recent efforts to apply it in Machine Learning, particularly in natural language processing. These methods aim to test a trained model with data that is distributionally different from the training data but follows the same semantic and syntactic composition rules. Studies have shown that Machine Learning models that use causal reasoning principles deliver high performance in such scenarios (Lake et al. 2015).

Therefore, this thesis also highlights the relevance of Machine Learning in composition discussions through causal information. Although the example in Chapter 5 was a specific case, these methods can be extended to other fields in quantum foundations.

6.3 The Spatiotemporal Problem of Entanglement

While the focus of this dissertation was on the causal problem of entanglement, it is worth noting the spatiotemporal problem of entanglement and its connection to the causal problem. The spatiotemporal problem arises due to the conflict between quantum non-locality and the theory of Relativity. Specifically, it asks whether quantum non-locality violates Relativity and if not, how the apparent tension can be reconciled. The literature offers various answers to this question, and I do not to enter debates surrounding these answers. Instead, I want to clarify a natural link between these debates and and the causal problem of entanglement.

In my opinion, the link between these two discussions is how causal models face the measurement problem. In particular, the way a causal model addresses the measurement problem has implications for the status of the wavefunction and how the model handles the transition from quantumness to classicality. If Relativity is viewed as a geometric constraint on the form of dynamical laws prescribed by a physical theory, then a model must have a Lorenz-invariant transition to be consistent with Relativity. Therefore, I propose that the measurement problem provides a natural connection between the two problems of entanglement, and this idea warrants further investigation in future research.

Bibliography

Abdi, H. (2007), 'The kendall rank correlation coefficient', *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA* pp. 508–510.

Adlam, E. (2022), 'Is there causation in fundamental physics? new insights from process matrices and quantum causal modelling', *arXiv preprint arXiv:2208.02721* .

Allen, J.-M. A., Barrett, J., Horsman, D. C., Lee, C. M. & Spekkens, R. W. (2017), 'Quantum common causes and quantum causal models', *Physical Review X* **7**(3), 031021.

Almada, D., Ch'ng, K., Kintner, S., Morrison, B. & Wharton, K. (2016), 'Are retrocausal accounts of entanglement unnaturally fine-tuned?', *International Journal of Quantum Foundations* **2**, 1–16.

Almeida, D. M. d. (2019), Pattern-based causal feature extraction, *in* 'Cause Effect Pairs in Machine Learning', Springer, pp. 321–329.

Araújo, M., Branciard, C., Costa, F., Feix, A., Giarmatzi, C. & Brukner, Č. (2015), 'Witnessing causal nonseparability', *New Journal of Physics* **17**(10), 102001.

Argaman, N. (2010), 'Bell's theorem and the causal arrow of time', *American Journal of Physics* **78**(10), 1007–1013.

Arntzenius, F. (2019), Reichenbach's common cause principle, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Spring 2019 edn, Metaphysics Research Lab, Stanford University.

Aronszajn, N. (1950), 'Theory of reproducing kernels', *Transactions of the American mathematical society* **68**(3), 337–404.

Aspect, A., Dalibard, J. & Roger, G. (1982), 'Experimental test of bell's inequalities using time-varying analyzers', *Physical review letters* **49**(25), 1804.

Bai, G., Wu, Y.-D., Zhu, Y., Hayashi, M. & Chiribella, G. (2022), 'Quantum causal unravelling', *npj Quantum Information* **8**(1), 1–9.

Barrett, J., Lorenz, R. & Oreshkov, O. (2019), 'Quantum causal models', *arXiv preprint arXiv:1906.10726* .

Barrett, J., Lorenz, R. & Oreshkov, O. (2021), 'Cyclic quantum causal models', *Nature communications* **12**(1), 1–15.

Bell, J. S. (1964), 'On the einstein podolsky rosen paradox', *Physics Physique Fizika* **1**(3), 195.

Bell, J. S. (1976), The theory of local beables, *in* 'Speakable and Unspeakable in Quantum Mechanics', Cambridge University Press, pp. 57–65.

Bertlmann, R. A. & Krammer, P. (2008), 'Bloch vectors for qudits', *Journal of Physics A: Mathematical and Theoretical* **41**(23), 235303.

Bohr, N. (1949), Discussion with einstein on epistemological problems in atomic physics, *in* 'Niels Bohr Collected Works', Vol. 7, Elsevier, pp. 339–381.

Branciard, C., Araújo, M., Feix, A., Costa, F. & Brukner, Č. (2015), 'The simplest causal inequalities and their violation', *New Journal of Physics* **18**(1), 013008.

Brask, J. B., Brunner, N., Cavalcanti, D. & Leverrier, A. (2012), 'Bell tests for continuous-variable systems using hybrid measurements and heralded amplifiers', *Phys. Rev. A* **85**.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.85.042116*

Butterfield, J. (1989), 'A space-time approach to the bell inequality', *Philosophical Consequences of Quantum Theory* pp. 114–144.

Callender, C. (2015), 'One world, one beable', *Synthese* **192**(10), 3153–3177.

Cartwright, N. (1979), 'Causal laws and effective strategies', *Noûs* pp. 419–437.

Cartwright, N. (1988), How to tell a common cause: Generalizations of the conjunctive fork criterion, *in* 'Probability and causality', Springer, pp. 181–188.

Cartwright, N. (2001), 'What is wrong with bayes nets?', *The monist* **84**(2), 242–264.

Cavalcanti, E. G. (2018), 'Classical causal models for bell and kochen-specker inequality violations require fine-tuning', *Physical Review X* **8**(2), 021018.

Chalupka, K., Perona, P. & Eberhardt, F. (2018), 'Fast conditional independence test for vector variables with large sample sizes', *arXiv preprint arXiv:1804.02747* .

Chickering, D. M. (2002), 'Optimal structure identification with greedy search', *Journal of machine learning research* **3**(Nov), 507–554.

Chiribella, G. & Ebler, D. (2019), 'Quantum speedup in the identification of cause–effect relations', *Nature communications* **10**(1), 1472.

Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. (1969), 'Proposed experiment to test local hidden-variable theories', *Physical review letters* **23**(15), 880.

Climenhaga, N., DesAutels, L. & Ramsey, G. (2021), 'Causal inference from noise', *Noûs* **55**(1), 152–170.

Costa, F. & Shrapnel, S. (2016), 'Quantum causal modelling', *New Journal of Physics* **18**(6), 063032.

Daley, P. J., Resch, K. J. & Spekkens, R. W. (2022), 'Experimentally adjudicating between different causal accounts of bell-inequality violations via statistical model selection', *Physical Review A* **105**(4), 042220.

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J. et al. (2013), Recent advances in deep learning for speech research at microsoft, *in* '2013 IEEE International Conference on Acoustics, Speech and Signal Processing', IEEE, pp. 8604–8608.

Dowe, P. (2000), *Physical Causation*, Cambridge Studies in Probability, Induction and Decision Theory, Cambridge University Press.

Dürr, D., Goldstein, S. & Zanghì, N. (2013), *Quantum Physics Without Quantum Philosophy*, Springer Science & Business Media.

Egg, M. & Esfeld, M. (2014), 'Non-local common cause explanations for EPR', *European Journal for Philosophy of Science* **4**(2), 181–196.

Einstein, A., Podolsky, B. & Rosen, N. (1935), 'Can quantum-mechanical description of physical reality be considered complete?', *Physical review* **47**(10), 777.

Evans, P. W. (2018), 'Quantum causal models, faithfulness, and retrocausality', *The British Journal for the Philosophy of Science* **69**(3), 745–774.

Fitzsimons, J. F., Jones, J. A. & Vedral, V. (2015), 'Quantum correlations which imply causation', *Scientific reports* **5**(1), 18281.

Friebe, C., Kuhlmann, M., Lyre, H., Näger, P. M., Passon, O. & Stöckler, M. (2018), *The Philosophy of Quantum Physics*, Springer.

Giarmatzi, C. (2019), A quantum causal discovery algorithm, *in* 'Rethinking Causality in Quantum Mechanics', Springer, pp. 125–150.

Glymour, C. (2006), Markov properties and quantum experiments, *in* 'Physical theory and its interpretation', Springer, pp. 117–126.

Glymour, C., Zhang, K. & Spirtes, P. (2019), 'Review of causal discovery methods based on graphical models', *Frontiers in genetics* **10**, 524.

Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. (2021), 'Revisiting deep learning models for tabular data', *Advances in Neural Information Processing Systems* **34**, 18932–18943.

Goswami, K., Giarmatzi, C., Kewming, M., Costa, F., Branciard, C., Romero, J. & White, A. G. (2018), 'Indefinite causal order in a quantum switch', *Physical review letters* **121**(9), 090503.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D. & Sebag, M. (2018), Learning functional causal models with generative neural networks, *in* 'Explainable and interpretable models in computer vision and machine learning', Springer, pp. 39–80.

Goudet, O., Kalainathan, D., Sebag, M. & Guyon, I. (2019), Learning bivariate functional causal models, *in* 'Cause Effect Pairs in Machine Learning', Springer, pp. 101–153.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B. & Smola, A. (2006), 'A kernel method for the two-sample-problem', *Advances in neural information processing systems* **19**, 513–520.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B. et al. (2005), 'Kernel methods for measuring independence', *Journal of Machine Learning Research* .

Guo, R., Cheng, L., Li, J., Hahn, P. R. & Liu, H. (2020), 'A survey of learning causality with data: Problems and methods', *ACM Computing Surveys (CSUR)* **53**(4), 1–37.

Guyon, I. (2013), 'Chalearn cause effect pairs challenge'.
URL: *http://www.causality.inf.ethz.ch/cause-effect.php*

Guyon, I. (2014), 'Chalearn fast causation coefficient challenge'.
URL: *https://www.codalab.org/competitions/1381*

Guyon, I., Statnikov, A. & Batu, B. B. (2019), *Cause effect Pairs in machine learning*, Springer.

Harrigan, N. & Spekkens, R. W. (2010), 'Einstein, incompleteness, and the epistemic view of quantum states', *Foundations of Physics* **40**(2), 125–157.

Hausman, D. M. (1999), 'Lessons from quantum mechanics', *Synthese* pp. 79–92.

Hausman, D. M. & Woodward, J. (1999), 'Independence, invariance and the causal markov condition', *The British journal for the philosophy of science* **50**(4), 521–583.

He, Y.-B. & Geng, Z. (2008), 'Active learning of causal networks with intervention experiments and optimal designs', *Journal of Machine Learning Research* **9**(Nov), 2523–2547.

Henson, J., Lal, R. & Pusey, M. F. (2014), 'Theory-independent limits on correlations from generalized bayesian networks', *New Journal of Physics* **16**(11), 113043.

Hofer-Szabó, G. & Vecsernyés, P. (2012), 'Noncommuting local common causes for correlations violating the clauser–horne inequality', *Journal of Mathematical Physics* **53**(12), 122301.

Hooft, G. (2009), 'Entangled quantum states in a local deterministic theory', *2nd Vienna Symposium on the Foundations of Modern Physics; arXiv preprint arXiv:0908.3408* .

Hossenfelder, S. & Palmer, T. (2020), 'Rethinking superdeterminism', *Frontiers in Physics* **8**, 139.

Howell, J. C., Bennink, R. S., Bentley, S. J. & Boyd, R. W. (2004), 'Realization of the einstein-podolsky-rosen paradox using momentum- and position-entangled photons from spontaneous parametric down conversion', *Phys. Rev. Lett.* **92**.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.92.210403*

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J. & Schölkopf, B. (2008), 'Nonlinear causal discovery with additive noise models', *Advances in neural information processing systems* **21**.

Janzing, D. (2007), 'On causally asymmetric versions of occam's razor and their relation to thermodynamics', *arXiv preprint arXiv:0708.3411* .

Janzing, D. (2019), 'The cause-effect problem: Motivation, ideas, and popular misconceptions', *Cause Effect Pairs in Machine Learning* pp. 3–26.

Janzing, D., Chaves, R. & Schölkopf, B. (2016), 'Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference', *New Journal of Physics* **18**(9), 093052.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniusis, P., Steudel, B. & Schölkopf, B. (2012), 'Information-geometric approach to inferring causal directions', *Artificial Intelligence* **182**, 1–31.

Janzing, D. & Schölkopf, B. (2010), 'Causal inference using the algorithmic markov condition', *IEEE Transactions on Information Theory* **56**(10), 5168–5194.

Kalainathan, D. & Goudet, O. (2019), 'Causal discovery toolbox: Uncover causal relationships in python', *arXiv preprint arXiv:1903.02278* .

Kochen, S. & Specker, E. P. (1975), The problem of hidden variables in quantum mechanics, *in* 'The logico-algebraic approach to quantum mechanics', Springer, pp. 293–328.

Koller, D. & Friedman, N. (2009), *Probabilistic graphical models: principles and techniques*, MIT press.

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S. & Barnes, L. E. (2017), Hdltex: Hierarchical deep learning for text classification, *in* '2017 16th IEEE international conference on machine learning and applications (ICMLA)', IEEE, pp. 364–371.

Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. (2015), 'Human-level concept learning through probabilistic program induction', *Science* **350**(6266), 1332–1338.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

Lee, S. & Bareinboim, E. (2018), 'Structural causal bandits: Where to intervene?', *Advances in neural information processing systems* **31**.

Leifer, M. S. (2006), 'Quantum dynamics as an analog of conditional probability', *Physical Review A* **74**(4), 042310.

Leifer, M. S. & Spekkens, R. W. (2013), 'Towards a formulation of quantum theory as a causally neutral theory of bayesian inference', *Physical Review A* **88**(5), 052130.

Lopez-Paz, D., Muandet, K., Schölkopf, B. & Tolstikhin, I. (2015), Towards a learning theory of cause-effect inference, *in* 'International Conference on Machine Learning', PMLR, pp. 1452–1461.

Lopez-Paz, D., Nishihara, R., Chintala, S., Scholkopf, B. & Bottou, L. (2017), Discovering causal signals in images, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 6979–6987.

Lévy–Leblond, J.-M. (1985), 'Discussion sections', *Symposium on the Foundations of Modern Physics: 50 Years of the Einstein-Podolsky-Rosen Gedankenexperiment* .

Margaritis, D. (2003), Learning bayesian network model structure from data, Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Maudlin, T. (2011), *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*, John Wiley & Sons.

Minnaert, B. (2019), Feature importance in causal inference for numerical and categorical variables, *in* 'Cause Effect Pairs in Machine Learning', Springer, pp. 349–358.

Mitchell, T. M. & Mitchell, T. M. (1997), *Machine learning*, McGraw-hill New York.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J. & Schölkopf, B. (2016), 'Distinguishing cause from effect using observational data: methods and benchmarks', *The Journal of Machine Learning Research* **17**(1), 1103–1204.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B. et al. (2017), 'Kernel mean embedding of distributions: A review and beyond', *Foundations and Trends® in Machine Learning* **10**(1-2), 1–141.

Näger, P. M. (2016), 'The causal problem of entanglement', *Synthese* **193**(4), 1127–1155.

Näger, P. M. (2020), 'A stronger bell argument for (some kind of) parameter dependence', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **72**, 1–28.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. & Ng, A. Y. (2011), Multimodal deep learning, *in* 'ICML'.

Nielsen, M. A. & Chuang, I. L. (2010), *Quantum computation and quantum information*, Cambridge university press.

Oreshkov, O., Costa, F. & Brukner, Č. (2012), 'Quantum correlations with no causal order', *Nature communications* **3**(1), 1–8.

Pearl, J. (1988), *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan kaufmann.

Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, 1 edn, Cambridge university press.

Pearl, J. (2009), *Causality*, Cambridge university press.

Pearl, J. & Verma, T. (1991), *A formal theory of inductive causation*, University of California (Los Angeles). Computer Science Department.

Peters, J., Janzing, D. & Schölkopf, B. (2017), *Elements of causal inference*, The MIT Press.

Pienaar, J. & Brukner, Č. (2015), 'A graph-separation theorem for quantum causal models', *New Journal of Physics* **17**(7), 073020.

Prechelt, L. (1998), Early stopping-but when?, *in* 'Neural Networks: Tricks of the trade', Springer, pp. 55–69.

Price, H. & Wharton, K. (2015), 'Disentangling the quantum world', *Entropy* **17**(11), 7752–7767.

Redei, M., Hofer-Szabo, G. & Szabo, L. (2013), *The Principle of the Common Cause*, Cambridge University Press.

Reichenbach, H. (1956), *The direction of time*, Vol. 65, University of California Press.

Ried, K., Agnew, M., Vermeyden, L., Janzing, D., Spekkens, R. W. & Resch, K. J. (2015), 'A quantum advantage for inferring causal structure', *Nature Physics* **11**(5), 414–420.

Robinson, R. W. (1977), Counting unlabeled acyclic digraphs, *in* 'Combinatorial mathematics V', Springer, pp. 28–43.

Rubino, G., Rozema, L. A., Feix, A., Araújo, M., Zeuner, J. M., Procopio, L. M., Brukner, Č. & Walther, P. (2017), 'Experimental verification of an indefinite causal order', *Science advances* **3**(3), e1602589.

Russell, B. (1912), On the notion of cause, *in* 'Proceedings of the Aristotelian society', Vol. 13, JSTOR, pp. 1–26.

Salmon, W. C. (1984), Scientific explanation: Three basic conceptions, *in* 'PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association', Vol. 1984, Philosophy of Science Association, pp. 293–305.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J. & Müller, K.-R. (2021), 'Explaining deep neural networks and beyond: A review of methods and applications', *Proceedings of the IEEE* **109**(3), 247–278.

Shen, D., Wu, G. & Suk, H.-I. (2017), 'Deep learning in medical image analysis', *Annual review of biomedical engineering* **19**, 221–248.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A. & Jordan, M. (2006), 'A linear non-gaussian acyclic model for causal discovery.', *Journal of Machine Learning Research* **7**(10).

Shrapnel, S. (2016), Using Interventions to Discover Quantum Causal Structure, PhD thesis, University of Queensland.

Shrapnel, S. (2019), 'Discovering quantum causal models', *The British Journal for the Philosophy of Science* .

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G. & Richardson, T. (2000), *Causation, Prediction, and Search*, MIT press.

Steel, D. (2020), 'Comment on hausman & woodward on the causal markov condition', *The British journal for the philosophy of science* .

Szabó, L. E. (2000), 'Attempt to resolve the epr-bell paradox via reichenbach's concept of common cause', *International journal of theoretical physics* **39**(3), 901–911.

Thearle, O., Janousek, J., Armstrong, S., Hosseini, S., Schünemann (Mraz), M., Assad, S., Symul, T., James, M. R., Huntington, E., Ralph, T. C. & Lam, P. K. (2018), 'Violation of bell's inequality using continuous variable measurements', *Phys. Rev. Lett.* **120**.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.120.040406*

Torlai, G., Wood, C. J., Acharya, A., Carleo, G., Carrasquilla, J. & Aolita, L. (2020), 'Quantum process tomography with unsupervised learning and tensor networks', *arXiv preprint arXiv:2006.02424* .

Tsamardinos, I., Aliferis, C. F. & Statnikov, A. (2003), Time and sample efficient discovery of markov blankets and direct causal relations, *in* 'Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 673–678.

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R. & Statnikov, E. (2003), Algorithms for large scale markov blanket discovery., *in* 'FLAIRS conference', Vol. 2, pp. 376–380.

Tsamardinos, I., Brown, L. E. & Aliferis, C. F. (2006), 'The max-min hill-climbing bayesian network structure learning algorithm', *Machine learning* **65**(1), 31–78.

Uhler, C., Raskutti, G., Bühlmann, P. & Yu, B. (2013), 'Geometry of the faithfulness assumption in causal inference', *The Annals of Statistics* pp. 436–463.

Verma, T. (1993), 'Graphical aspects of causal models', *Technical R eport R-191, UCLA* .

Vinh, N. X., Epps, J. & Bailey, J. (2010), 'Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance', *The Journal of Machine Learning Research* **11**, 2837–2854.

Weinberger, N. (2018), 'Faithfulness, coordination and causal coincidences', *Erkenntnis* **83**(2), 113–133.

Weinstein, S. (2017), 'Learning the einstein-podolsky-rosen correlations on a restricted boltzmann machine', *arXiv preprint arXiv:1707.03114* .

Weinstein, S. (2018), 'Neural networks as "hidden" variable models for quantum systems', *arXiv preprint arXiv:1807.03910* .

Wenger, J., Hafezi, M., Grosshans, F., Tualle-Brouri, R. & Grangier, P. (2003), 'Maximal violation of bell inequalities using continuous-variable measurements', *Phys. Rev. A* **67**.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.67.012105*

Williamson, J. (2009), 'Probabilistic theories of causality', *The Oxford Handbook of Causation* pp. 185–212.

Wood, C. J. & Spekkens, R. W. (2015), 'The lesson of causal discovery algorithms for quantum correlations: Causal explanations of bell-inequality violations require fine-tuning', *New Journal of Physics* **17**(3), 033002.

Woodward, J. (2007), 'Causation with a human face'.

Woodward, J. (2016*a*), Causation and Manipulability, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2016 edn, Metaphysics Research Lab, Stanford University.

Woodward, J. (2016*b*), 'The problem of variable choice', *Synthese* **193**, 1047–1072.

Zhang, J. & Spirtes, P. (2016), 'The three faces of faithfulness', *Synthese* **193**(4), 1011–1027.

Zhang, K. & Hyvärinen, A. (2009), On the identifiability of the post-nonlinear causal model, *in* 'Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence', AUAI Press, p. 647–655.

Zurek, W. H. (2002), 'Decoherence and the transition from quantum to classical-revisited', *Los Alamos Science* **27**, 86–109.