Causal Discovery of Photonic Bell Experiments

by

Patrick Daley

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Physics - (Quantum Information)

Waterloo, Ontario, Canada, 2019

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

A causal understanding of a physical theory is vital. They provide profound insights into the implications of the theory and contain the information required to manipulate, not only predict, our surroundings. Unfortunately, one of the most broadly used and successful theories, quantum theory, continues to evade a satisfactory causal description. The progress is hindered by the difficulty of faithfully testing causal explanations in an experimental setting. This thesis presents a novel causal discovery algorithm which allows a direct comparison of a wide variety of causal explanations for experimental data. They include causal influences both classical and quantum mechanical in nature. First we provide relevant background information, predominately on quantum mechanics, quantum optics and statistical inference. Next, we review the framework of classical causality and the connection between a causal assumption and statistical model. We then present a novel causal discovery algorithm for noisy experimental data. Finally, we perform two Bell experiments and apply the newly developed algorithm on the resulting data.

The causal discovery algorithm operates on observational data without any interventions required. It utilizes the concept of predictive accuracy to assign a score to each causal explanation. This allows the simultaneous consideration of classical and quantum causal theories. In addition, this approach allows the identification of overly complex explanations as these perform poorly with respect to this criterion.

Both experiments are implemented using quantum optics. The first Bell experiment has a near maximally entangled shared resource state while the second has a separable resource state. The results indicate that a quantum local causal explanation bests describes the first experiment, whereas a classical local causal explanation is preferred for the second. A super-luminal or super-deterministic theory are sub-optimal for both.

Acknowledgements

My time at Waterloo has been challenging intellectually and personally. There are many people who supported me in one or both regards and made my time at Waterloo so great. I hope to begin to thank some of them here.

I'll begin by thanking my supervisor, Kevin Resch. Without your support and guidance this thesis would not exist. The example you set on balancing your passion for science and numerous other pursuits in life is reflected in the wonderful lab group you have created. Thank you to Rob Spekkens for all the fascinating meetings and conversations that would often run late and spill over to lunch. You have helped shape my views on science, reality and life. Thank you to Norbert Lütkenhaus for joining my committee and asking thought provoking questions. Thank you Rajibul Islam for agreeing to join my defence committee.

I would like thank Andrew Cameron, Sandra Cheng, Michael Grabowecky and Sacha Schwarz for providing feedback on this thesis. An extra few thanks must be given to Sacha for the sheer volume of work he edited and the whimsical and enthusiastic manner in which he did it. Thank you to other current and past members of QOQI: Matt Brown, Jean-Phillippe Maclean, Mike Mazurek, Morgan Mastovich and Ruoxuan Xu.

Many thanks to Angus Kan and Hannah Lobbezoo for the occasional chats about science and the frequent ones about climbing.

Thank you to Turner Silverthorne, Jon Tessier and Justin Mawle for the late night toasterside conversations. Thank you Adam Moniz for driving to Waterloo to visit me so many times and the many adventures. Thank you Natalie Villeneuve for everything you do and everything you are.

Finally, I would like to thank my family: Debbie, Richard, Simon, Katherine and Eric for their perpetual love and support.

Dedication

To my parents, Debbie and Richard.

Table of Contents

List of Tables viii

List of Figures ix

1 Applied Quantum Information Theory 1

1.1 Quantum Mechanics . . . 1

1.1.1 States . . . 1

1.1.2 Measurements . . . 3

1.1.3 Transformations . . . 4

1.2 Distance Measures . . . 4

1.2.1 Classical Distance Measures . . . 5

1.2.2 Quantum Distance Measures . . . 6

1.3 Bipartite Entanglement . . . 8

1.4 Parametric Inference . . . 9

1.4.1 Maximum Likelihood Estimation . . . 9

1.4.2 Bootstrap Estimation of Errors . . . 11

1.4.3 Quantum State Tomography . . . 12

1.5 Bell Inequalities . . . 15

1.5.1 Loopholes . . . 17

2 Experimental Quantum Optics **18**
2.1 Photonic Qubits . . . 19
2.2 Wave plates . . . 19
2.3 Single Qubit Operations and Measurements . . . 21
2.4 Entangled Photon Source . . . 23

3 Observational Causal Discovery **26**
3.1 Introduction . . . 26
3.2 Causal Models . . . 27
3.3 Noiseless Causal Discovery . . . 28
3.4 Noisy Causal Discovery . . . 29
3.4.1 Outline of Problem . . . 30
3.4.2 Statistical Model Selection . . . 32
3.4.3 Latent Variables . . . 34
3.4.4 Quantum Causal Structures . . . 35
3.5 Summary . . . 36

4 Bell Experiment **38**
4.1 Introduction . . . 38
4.2 Causal Discovery . . . 40
4.3 Results . . . 45
4.4 Discussion . . . 49
4.5 Supplementary Material . . . 50
4.5.1 Causal Structure Cardinality . . . 50
4.5.2 Alternative Loss Functions . . . 54
4.5.3 Likelihood Function . . . 54

5 Conclusion **60**

References **62**

List of Tables

4.1 The test error $\overline{\text{Err}}$ for each of the considered causal structure. The first column is for the experimental data when the entangled state is maximally depolarized. The second column is for the experimental data when the photons are left entangled. 48

4.2 The training error, err, for each of the considered causal structure. The first column is for the experimental data when the entangled state is maximally depolarized. The second column is for the experimental data when the photons are left entangled. 49

4.3 The test and training errors for SY causal structures with a latent variable of cardinality one through four. 52

4.4 The test errors $\overline{\text{Err}}$ for a selection of causal structures of the entangled experiment. Each column corresponds to a different loss function being used for the training and test error. 55

4.5 The test errors $\overline{\text{Err}}$ for a selection of causal structures of the dephased experiment. Each column corresponds to a different loss function being used for the training and test error. 56

List of Figures

1.1 A table defining three probability distributions: p, q, r over five events a, b, c, d, e. The distributions q, r have the same trace distance from p despite intuitively q being closer. The Euclidean distance and fidelity both capture that q is closer then r 6

2.1 Diagram of a tilted wave plate. The glass on the right is rotated to an angle which is non-orthogonal to the beam. This increases the distance the beam must travel through it. 20

2.2 Plot of the relative phase delay as a function of the AC voltage applied for a typical LCR. 21

2.3 Optical diagram of our polarization analyser. We can project onto an arbitrary pure qubit state by adjusting the angles of the half and quarter waveplates. 22

2.4 Two type II SPDC sources being coherantly pumped with their output modes overlaid. The two events: source A creating a pair and source B creating a pair are in a superposition if we post select only on situation where only one pair measured in the output modes. If origin source remains unknown, the resulting state over the two output modes labelled a, b is $|VH\rangle + |HV\rangle$. 24

2.5 A sagnac interferometer entangled photon source. 25

3.1 A single DAG, which as they are illustrated here, given different cardinalities of *Lambda* is compatible with different sets of distributions. 34

3.2 A graphic summarizing a hypothetical application of the causal discovery algorithm. Three possible causal explanations G_1, G_2, G_3 are considered. An estimator is found for each using the training data set $\mathcal{D}_1$. The testing loss is calculated using the second data set $\mathcal{D}_2$. A red box highlights the hypothetical lowest test loss, Err_{G_3}. In this example, the selected causal structure is G_3. 37

4.1 Examples of directed acyclic graphs for a Bell experiment. (a) A common causal structure. (b) A causal structure with an additional channel between between a setting and outcome. (c) A causal structure with the latent variable influencing the setting choice. 42

4.2 Experimental diagram. Maximally polarization entangled photons pairs are created through parametric down conversion in both paths of a Sagnac interferometer. One photon is sent to a polarization measurement, and the other photon first passes through a depolarizing channel comprised of two LCRs before also having its polarization measured. Coincidences between a photon being measured on both sides of the experiment are recorded. PPKTP, periodically-poled potassium titanyl phosphate; PBS, polarizing beamsplitter; LCR, liquid crystal retarder; HWP, half-wave plate; QWP, quarter-wave plate 46

4.3 The density matrix of the maximum likelihood estimate of the state of the source for (a) the entangled experiment without a dephasing channel and (b) the experiment with the dephasing channel. Blue represents a positive number while red represents a negative number. 47

4.4 The test and training error for various cardinalities K of the latent variable in a classical common cause structure. The data is from the experiments with (a) an entangled shared resource (b) a dephased shared resource. . . . 52

4.5 The test and training error for various cardinalities k of the latent variable in a ΛS causal structure. The data is from the experiments with (a) an entangled shared resource (b) a dephased shared resource. The dashed lines on the insets indicate the training and test error of the quantum common causal structure. The higher line is the test error and the lower is the test error. 53

Chapter 1

Applied Quantum Information Theory

A good introduction to quantum mechanics can be found in [34]. This chapter loosely follows this reference, summarizing the necessary elements of the formalism and highlighting some important properties.

In Section 1.1 we introduce the Hilbert space formalism including states, measurements and transformations. In Section 1.2 we discuss different metrics used to measure the distance between states. Section 1.3 discusses the important phenomena of entanglement. Section 1.4 outlines statistical techniques which are used to estimate the state of a system. Finally, in Section 1.5, we briefly discuss Bell inequalities, their violation by quantum mechanics, and how experiments tests of them are conducted.

1.1 Quantum Mechanics

1.1.1 States

The state of a system is commonly thought to be its position, momentum or temperature. However, in reality, it contains all the information describing how the system will interact with other systems, measurements and its environment. In quantum mechanics, a pure state is encoded as a vector in a projective complex Hilbert space. A complete Hilbert space is an inner product space which is complete under its induced metric

$$\mu(a, b) = \sqrt{\langle a - b, a - b \rangle}, \tag{1.1}$$

where $\langle a, a\rangle$ is the Hilbert space's inner product. Requiring it to be a projective Hilbert space is equivalent to saying that global phases on vectors are not observable and their length doesn't matter. Equivalently, we can write

$$|\psi\rangle \equiv ae^{i\varphi} |\psi\rangle \qquad \forall \quad \varphi, a \in \mathbb{R}. \tag{1.2}$$

Eq. (1.2) defines sets of equivalent vectors. We typically choose a unit length or *normalized* vector as the representative, however, we often leave the normalization to be implicit. A qubit is a system with a two-dimensional Hilbert space. This could be a subspace of a larger system. In general, Hilbert spaces can be of any dimension, even infinite. However, this thesis will only consider ones that are finite dimensional.

Until now we have considered *pure states*. These are states which cannot be thought of as a probabilistic mixture[1] of other states. Physically this means you cannot predict the measurement outcomes of the state for every measurement simply by knowing those of another set of states. The converse of a pure state is a *mixed state*. In contrast to classical physics, quantum physics allows states to exist in a *superposition* of different states. This is not the same as a probabilistic mixture of two states. For our encoding to distinguish between a statistical and a probabilistic mixture we must introduce the density matrix formalism. In this encoding an equal mixture of the states $|0\rangle, |1\rangle$ is represented by

$$|0\rangle\langle 0| + |1\rangle\langle 1| \tag{1.3}$$

while an equal superposition is

$$\big(\, |0\rangle + |1\rangle \,\big)\big(\, \langle 0| + \langle 1| \,\big) \tag{1.4}$$

In general, superpositions are denoted by the addition of vectors and probabilistic mixtures are the additions of the outer-products of the vectors. Notice that a superposition of pure states is again a pure state. In general, the density matrix of a probabilistic mixture of pure states is

$$\rho = \sum_i p_i \, |\psi_i\rangle\langle\psi_i| \tag{1.5}$$

where the system is in the pure state $|\psi_i\rangle$ with probability p_i. Since probabilities are always positive and sum to one, any density matrix is positive semi-definite with a trace equal to one. A trace-one density matrix is said to be normalized. Similarly to vectors, we sometimes leave the normalization implicit. We switch freely between the vector and density matrix description of pure states as they are in unique correspondence.

[1] This is also called a convex mixture.

In classical physics, if we know the state[2] of a joint system (A,B), to find the state of the subsystem A we marginalize over B, i.e.,

$$p(a) = \sum_b p(a, b). \tag{1.6}$$

In quantum mechanics, if ρ_{AB} is the state of the joint system (A,B) then the state of the subsystem A is given by

$$\rho_A = \mathrm{Tr}_B[\rho_{AB}], \tag{1.7}$$

where Tr_B is the partial trace over the Hilbert space of B. This can be viewed as being the quantum version of marginalizing a probability distribution.

1.1.2 Measurements

A general measurement in quantum mechanics is represented by a positive-operator valued measure (POVM). This is a set of positive operators, $\{E_i\}$, on the Hilbert space being measured, such that

$$\sum_i E_i = \mathbb{I}, \tag{1.8}$$

where $\mathbb{I}$ is the identity operator. Each element is associated with an outcome, for example a particular detector clicking. The probability of an outcome occurring is given by Born's rule

$$p(k|\rho) = \mathrm{Tr}[\rho E_k], \tag{1.9}$$

where $p(k|\rho)$ is the probability of the k^{th} outcome given that the system is in the state ρ.

Projective measurements form a special subclass of general measurements. We call a POVM projective when each operator is idempotent, i.e., $E_i^2 = E_i$. The positivity condition of POVMs implies that they are orthogonal projectors and thus can be written as

$$E_i = |\psi_i\rangle\langle\psi_i| \tag{1.10}$$

for some set of vectors $\{|\psi\rangle\}$. In the case of projective measurements, Born's rule, Eq, (1.9), simplifies to

$$p(k|\rho) = \langle\psi_k|\rho|\psi_k\rangle\,. \tag{1.11}$$

Furthermore, if the system is in a pure state, Eq. (1.11) reduces to the familiar form

$$p(k|\,|\phi\rangle\langle\phi|) = |\,\langle\phi|\psi_k\rangle\,|^2. \tag{1.12}$$

Projective measurements can be viewed as the pure state equivalent for measurements.

[2] A classical state is a probability distribution over the state space.

1.1.3 Transformations

In the Schrödinger picture, the state of a quantum system changes over time as it interacts with its surroundings and other systems. A transformation[3] should be represented by an operator[4] on the set of density operators. This operator should be completely positive and trace preserving (CPTP) in order to ensure that if acting on a density matrix, or the subspace of one, the resulting state is also a valid density matrix.

Evolution of a system can also be thought of as being a transformation. The transformation connects the system before and after the interaction. This idea can also be further generalized to a situation where the dimension of the density matrix entering and exiting may not be the same.

An important subclass of transformations are those which maintain pure states. These are called unitary transformation since they are described by a unitary operator

$$\begin{aligned} \Phi[\rho] &= U\rho U^\dagger, \\ U^\dagger U &= UU^\dagger = \mathbb{I}. \end{aligned} \tag{1.13}$$

The pure state after the transformation will be $U|\psi\rangle$ if $|\psi\rangle$ was the input. In general, we refer to the mathematical description of a transformation as a quantum channel.

1.2 Distance Measures

In experimental physics the desired state or measurement can never be implemented exactly. However, we still wish to know how close it is to the objective[5]. This is one of multiple uses for a distance measure. There is no universally preferred distance metric. A distance measure ideally returns a value with some physical meaning. It also should obey a set of useful and intuitive properties which extend our understanding of distance in Euclidean space. The appropriate choice may vary based on the application. Examples of desired properties are symmetry, triangle inequality and positivity:

$$\mu(A,B) = \mu(B,A) \qquad \text{symmetry} \tag{1.14}$$

$$\mu(A,C) \leq \mu(A,B) + \mu(B,C) \qquad \text{triangle inequality} \tag{1.15}$$

$$\mu(A,B) \geq 0 \qquad \text{positivity} \tag{1.16}$$

where μ is the distance measure, also called a *metric*.

[3] The terms channel and operation are sometimes used instead of transformation.

[4] I will use the term operation for a linear map with the same domain and codomain.

[5] After we have determined the state using tomography.

1.2.1 Classical Distance Measures

For classical probability distributions common metrics include the trace metric (L1 metric), log-likelihood, total variation distance and the Bhattacharyya distance. Here are the mathematical definitions of these metrics:

$$D(p,q) = \frac{1}{2}\sum_{x\in\Omega} |p_x - q_x| \qquad \text{trace distance,} \tag{1.17}$$

$$\text{loglik}(p,q) = -\sum_{x\in\Omega} p_x \log q_x \qquad \text{log-likelihood,} \tag{1.18}$$

$$TV(p,q) = -\max_{S\subset\Omega} |p(S) - q(S)| \qquad \text{total variation distance,} \tag{1.19}$$

$$BD(p,q) = -\log\left[\sum_{x\in\Omega} \sqrt{p_x}\sqrt{q_x}\right] \qquad \text{Bhattacharyya distance,} \tag{1.20}$$

where Ω is the sample space.

A similar notion of distance in Euclidean space is taking the dot product of two vectors. The classical fidelity[6] is calculated by taking the dot product between the square root of each vector. Mathematically, it is given by

$$F(p,q) = \sum_{x\in\Omega} \sqrt{p_x}\sqrt{q_x}. \tag{1.21}$$

The square root is taken to ensure the overlap between a probability distribution with itself is unity. Notice that the fidelity is the argument of the logarithm in the definition of the Battacharyya distance. The Bhattacharyya distance is one of the ways to take a fidelity and turn it into a true distance metric, in this case by taking the negative logarithm of the fidelity.

An alternative way to develop a distance measure, as opposed to extending concepts from Euclidean space, is using an operationally motivated quantity. For example, the degree to which one can discriminate, using any measurement, between two states. The total variation distance (TVD) has exactly this motivation. The event S in Eq. (1.19), which maximizes the difference in probabilities, is the best measurement an experimenter could perform in order to distinguish between the two distributions. The trace distance happens to be equal to the TVD. Therefore, it carries the same operational interpretation while being easier to compute.

[6] This is also called the Bhattacharyya constant or affinity.

	a	b	c	d	e
p	1/10	2/10	1/10	4/10	2/10
q	1/10	1/10	2/10	4/10	2/10
r	0/10	1/10	2/10	3/10	3/10

$$D(p,q) = 1/10 \qquad \|p-q\|_2 = \sqrt{1/5} \qquad F(q,r) = 98.3$$
$$D(p,r) = 1/10 \qquad \|p-r\|_2 = \sqrt{1/2} \qquad F(q,r) = 87.4$$

Figure 1.1: A table defining three probability distributions: p, q, r over five events a, b, c, d, e. The distributions q, r have the same trace distance from p despite intuitively q being closer. The Euclidean distance and fidelity both capture that q is closer then r

The trace distance does have its disadvantages. It only considers the largest discrepancy in the probabilities for an event. This results in the rest of the distribution being ignored. A distance measure like this is useful in bounding errors, however, it does not capture what many scenarios would consider to be a distance. The fidelity in comparison does not have this issue. Figure 1.1 illustrates the advantage of fidelity in this regard.

The fidelity in contrast, lacks a clear operational motivation. It does however have a geometric motivation, as noted earlier, by virtue of the dot product.

1.2.2 Quantum Distance Measures

The two most commonly used distances in quantum information are the quantum trace distance and quantum fidelity:

$$D(\rho,\sigma) = \frac{1}{2}\operatorname{Tr}\{\rho-\sigma\} \qquad \text{quantum trace distance,} \tag{1.22}$$
$$F(\rho,\sigma) = \operatorname{Tr}\left\{\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right\} \qquad \text{quantum fidelity.} \tag{1.23}$$

As their names suggest, they are equivalent to their classical counterparts when applied to mutually diagonalizable[7] states. The trace distance retains its interpretation as the greatest amount a measurement could distinguish between the two states since

$$D(\rho, \sigma) = \max_{\{E_m\}} D(p, q), \tag{1.24}$$

where $p_m = \mathrm{Tr}\{\rho E_m\}$, $q_m = \mathrm{Tr}\{\sigma E_m\}$. This maximization is taken over all possible POVM's. The quantum trace distance value is the maximal classical trace distance between the two distributions created by doing a measurement. The maximum can actually always be obtained by a two-outcome measurement. This means it is exactly how likely the two states can be distinguished by the best possible POVM however this still has the same problem as a classical fidelity. In order to illustrate this, let us consider the following quantum states:

$$\rho = 1/10\Big[1\,|1\rangle\langle 1| + 2\,|2\rangle\langle 2| + 1\,|3\rangle\langle 3| + 4\,|4\rangle\langle 4| + 2\,|5\rangle\langle 5|\Big],$$
$$\sigma = 1/10\Big[1\,|1\rangle\langle 1| + 1\,|2\rangle\langle 2| + 2\,|3\rangle\langle 3| + 4\,|4\rangle\langle 4| + 2\,|5\rangle\langle 5|\Big],$$
$$\psi = 1/10\Big[1\,|2\rangle\langle 2| + 2\,|3\rangle\langle 3| + 3\,|4\rangle\langle 4| + 2\,|5\rangle\langle 5|\Big].$$

The distributions in the classical example in Figure 1.1 are on the main diagonal so the quantum/classical trace and fidelity agree. This means despite intuitively σ being closer then ψ to ρ they have the same trace distance. Agreeing with the classical analogy they do have different fidelities with ρ, 98% and 87% respectively.

Quantum fidelity still lacks the information theoretic motivation, however, it can easily be related to a metric through the Bhattacharyya distance, Bures metric or others. It retains a geometric interpretation with the inner product through Uhlmann's theorem [34] which states that the fidelity is equal to the maximal magnitude of the inner product between any two purification[8] of the two density matrices.

In the experimental work contained in this thesis we do not consider bounding inequalities or error correction and wish to use a more intuitive notion of distance and thereby use fidelity as a measure of our success at preparing states in the laboratory.

[7] These can be loosely interpreted as being classical probability distributions.

[8] A purification of a density matrix is a pure state with a partial trace equal to the density matrix.

1.3 Bipartite Entanglement

The previous sections introduced quantum mechanics and some general machinery to work with the theory. Now it is time to discuss one of its hallmark features: entanglement. Entanglement is at the heart of most distinctly quantum phenomena including quantum teleportation, remote state preparation and Bell inequality violations.

A quantum state ρ_{AB} is said to be *separable* if it can be decomposed as

$$\rho_{AB} = \sum_k p_k \rho_A \otimes \rho_B, \tag{1.25}$$

where $p_k \geq 0$ and $\sum_k p_k = 1$. A state which fails this criterion is called *entangled*.

For pure states a more physically motivated way to view entanglement is that the joint system exists in a pure state but its subsystems do not. This motivates a measure of entanglement for pure states called the **entropy of entanglement**, defined as

$$E[|\psi_{AB}\rangle] = -\operatorname{Tr}\{\rho_B \log \rho_B\}, \tag{1.26}$$

where $\rho_B = \operatorname{Tr}_B |\psi_{AB}\rangle\langle\psi_{AB}|$. This is simply the von Neumann entropy of the subsystem. This measure turns out to be the unique measure of entanglement for pure states [36]. A brief argument explaining this claim is as follows. Any ensemble of pure entangled states are reversibly connected by local operations to an ensemble of singlets [6]. Since local operations cannot increase entanglement, the two ensembles (of size m, n respectively) must have the same entanglement. Finally, the amount of entanglement is an intrinsic property and scales linearly with the number of systems in the ensemble. This together gives the constraint

$$\mu[|\psi\rangle] = \frac{n}{m}\mu[|S\rangle] \tag{1.27}$$

for any valid entanglement measure, μ, on pure states. If we set the entanglement of a singlet to 1, then we get the entanglement of another state to be the entropy of entanglement. The entanglement of a singlet could be set to another number giving another measure of entanglement, however, it would just be a constant times the entropy of entanglement for every state.

The situation for mixed states is more complicated. The entanglement entropy is no longer a valid measure since some separable mixed states, a maximally mixed state for example, would have non-zero entanglement entropy. There no longer exists a single unique measure. A widely used measure is the **entanglement of formation**, which is given by

$$E_f[\rho] = \inf \sum_i p_i E[|\psi_i\rangle], \tag{1.28}$$

where the infimum is taken over all pure state decompositions, $\{p_i, |\psi_i\rangle\}$, of ρ and $E[|\psi_i\rangle]$ is the entanglement entropy of the pure state. The ensemble $\{p_i, |\psi_i\rangle\}$ is a decomposition of the density matrix if $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$. It can readily be seen that this agrees with the entanglement entropy for pure states.

The entanglement of formation in the case of entanglement between a pair of qubits can be explicitly calculated (as opposed to numerically minimizing) using concurrence [52]. In this case, concurrence and the entanglement of formation are related by a monotonically increasing function. Therefore, concurrence can be used as a metric itself or simply converted to entanglement of formation. The explicit formula for concurrence is giving by

$$C[\rho] = \max\{0, \lambda_1 - \lambda_2 - \lambda_3 - \lambda_4\}, \tag{1.29}$$

where the λ's are the positive square roots of the eigenvalues of the operator $\rho(\sigma_y \otimes \sigma_y)\rho^*(\sigma_y \otimes \sigma_y)$ with λ_1 being the largest. In Eq. (1.29) ρ^* is the complex conjugate of ρ. See [26] for more information on concurrence and entanglement beyond the bipartite case.

1.4 Parametric Inference

An experiment, in the statistical sense, is sampling a set of random variables. The set of samples is referred to as *data*. The experimenter attempts to infer the distribution they were sampled from, the *state*, or some property of it. In most scenarios, a model of the statistical noise is assumed. This is normally an educated assumption based on knowledge gathered from previous experiments. After assuming a statistical noise model, the family the distribution belongs to, for example normal distributions, is known. The experimenter only needs to determine its parameters. These parameters allow us to predict the results of future observations of the random variables. The techniques needed for this analysis are contained within the field of parametric inference. These methods are used daily, often unwarily, by experimental physicists. In this section, we explicitly outline those most commonly used for the experimental work in this thesis. An interested reader is recommended [10, 47] for further reading on this topic.

1.4.1 Maximum Likelihood Estimation

A parametric model, $\mathcal{M}$, for a random vector is a set

$$\mathcal{M} = \big\{ f(x;\theta) : \theta \in \Theta \big\}, \tag{1.30}$$

where $\Theta \subset \mathbb{R}^n$ is the parameter space and the function f is over the range of the random vector. In the case of a normal distribution of unknown mean and variance, the parametric model is

$$\mathcal{M} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+ \right\}, \tag{1.31}$$

where $\mathbb{R}^+$ is the set of positive real numbers.

In order to determine the parameters, the distribution is sampled independently many times. Mathematically, this is a set of independent identically distributed random variables

$$X_1, X_2, ..., X_n \sim f(x; \theta). \tag{1.32}$$

The results of these measurements is a real vector $\vec{x} = (x_1, x_2, ..., x_n)$[9], for example the results of flipping a coin multiple times, colloquially known as "data". A way of calculating a single "best guess" for a parameter, for example the bias of a coin, from the data is called a point estimator. Maximum likelihood estimation (MLE) offers a well motivated way to find an estimate. This decision rule selects the value of the parameter which maximizes the likelihood of the observed outcomes being measured. The likelihood of observing the data $\vec{x}$ for a given value of the parameters θ is given by

$$\mathcal{L}[x_1, x_2, ..., x_n|\theta] = \prod_i f(X_i = x_i|\theta), \tag{1.33}$$

where $f \in \mathcal{M}$. Maximizing this function is also equivalent to maximizing the log-likelihood which often is an easier problem to solve.

A nice property of this estimate is that, in the limit of infinitely many samples, the probability of it giving the correct answer converges to one. This property is called consistency. Another important property is its asymptotic normality. Furthermore, the Cramer-Rao bound ensures the asymptotic efficiency of the estimate among unbiased estimators. Both these topics will be further discussed in the next section.

There exist many methods for generating point estimates of parameters. The framework of decision theory [46] is used to compare them. A brief introduction of this theory is provided as it further motivates the use of a MLE.

The risk associated with an estimate is a measure of its quality. For example by taking the square error

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta\big[(\theta - \hat{\theta})^2\big], \tag{1.34}$$

[9]Upper case letter will be used to denote random variables and lower case letters are used to denote the outcomes of a measurement of the random variable.

where θ is the true value, $\hat{\theta}$ is the estimate and the expectation value is taken with respect to the true value of θ. This gives the risk in using the particular estimate, also referred to as a rule, if the real value of the parameter is θ. The risk function would take a different value for each possible state of reality θ. An important class of rules are those which are **minimax**. A rule is minimax if out of all the possible rules the max value of its risk function is smaller then all others. This highlights the goal to minimize the worst case scenario. An important result for MLE is that this decision rule is asymptotically minimax [47]. It should be noted that the MLE when applied to finite data is not minimax. It sometimes is even dominated by another estimator for classical and quantum point estimation [17]. We choose to continue using MLEs in this thesis due to their physical interpretation, relative simplicity and prevalence in the literature.

1.4.2 Bootstrap Estimation of Errors

A point estimator, MLE or otherwise, is always incorrect in the sense that it has a zero probability of returning the true value. A point estimator is a single "best guess" to a question with infinitely many potential answers[10]. To alleviate this issue, a region of estimates must be given. This practice also allows an investigation into the precision of an estimate by measuring the size of the region.

Prior to finding the region a confidence level must be decided. An $\alpha\%$ confidence interval is a region estimator which returns a subset of parameters values which contains the true value for $\alpha\%$ percent of the occasions it is calculated[11]. The following section will outline one method of calculating confidence regions.

The estimators used in this thesis are all MLEs and these are always asymptotically normal and unbiased. Instead of reporting confidence regions, the mean (μ) and variance (σ^2) of the normal distribution can be given. For example, $\mu \pm \sigma$. This allows the calculation of any desired confidence region at a later time. In the asymptotic case, the mean is the MLE point estimator so we need only calculate the variance. The well developed method of **parametric bootstrap estimation** [15] can be used to calculate this in an extremely wide variety of scenarios. Other techniques exist to estimate a MLEs variance, for example the delta method which relies on calculating the Fisher Information. However, the bootstrap method is far easier when even minimal computational power is available. This technique is used throughout the thesis alongside any point estimate.

[10] I'm implicitly assuming the parameter space is an interval of real numbers as this is essentially always the case.

[11] I'm implicitly using a frequentist interpretation of probability.

The scenario concerns again the case where there are independent identically distributed random variables, $\vec{x} = x_1, x_2, ..., x_n$, sampled according to some model with an unknown parameter value, θ. We wish to find the variance of some statistic (S), for example the fidelity or CHSH parameter[12]

$$\mathbb{V}[S(x_1, x_2, ..., x_n)] = \mathbb{E}[(S - \mathbb{E}[S])^2]. \tag{1.35}$$

The variance and expectation value are taken with respect to the true distribution. An easy way to calculate the variance would be to simply sample the n random variables for B more data sets $\{\vec{x}^{(k)}\}_{k=1}^{B}$ then calculate the statistic on each set and look at the variance over the B examples; this is referred to as *simulation*. Unfortunately, sampling the distribution again is often not an option. In the parametric inference case, we can instead do a ML estimation ($\hat{\theta}$) of the model's parameters (θ) and computationally sample from the distribution $f(x; \hat{\theta})$.

It should be noted again that these results are only asymptotically true. The MLE statistic generally has an unknown distribution and is biased for finite sample sizes which invalidates the assumptions we used to find confidence regions [7]. In the cases encountered in this thesis we have large sample sizes, in the thousands, so we are comfortable using asymptotic results as a good approximation.

1.4.3 Quantum State Tomography

Quantum state tomography remains a research area with constantly evolving and refining methods. The problem concerns estimating the density matrix given the outcomes of a set of measurements. Some popular techniques for point estimation include linear inversion, MLE and Bayesian mean estimation. In this thesis we use a MLE method.

The set of all d-dimensional quantum states is parameterized by d^2 real parameters, $\theta = (\theta_1, \theta_2, \cdots, \theta_{d^2})$, such that

$$\rho = \sum_{k=1}^{d^2} \theta_k B_k, \tag{1.36}$$

where B_k is a set of operators that span, over $\mathbb{R}$, all positive trace one operators. Finding an estimate for the density matrix is a problem of parametric point estimation. I will use θ and ρ to label interchangeably without having a particular basis in mind, however in any application a basis would obviously have to be specified.

[12] The CHSH parameter is further discussed in section 1.5.

Consider a set of POVMs $\{M_i\}_{i=1}^n$. Each individual POVM represents a measurement with m possible outcomes. For each outcome $j \in m$ of a POVM, a quantum state will give a prediction for the expected frequency Eq. (1.9). The process of measuring each of these POVMs consecutively can be rewritten as performing one larger POVM with $n \times m$ outcomes. Consequently we will perform the analysis only for the case of tomographic reconstruction from the measurement outcomes of one POVM. To avoid losing any generality, consider the case of a POVM with arbitrarily many outcomes, i.e.,

$$f(k;\theta) = \mathrm{Tr}[\rho E_k], \tag{1.37}$$

where E_k is the positive operator associated with the k^{th} outcome. The measurements are assumed to have been implemented perfectly and the dimension of the state is known[13].

It is now time to find the likelihood function. In most experiments, we do not simply measure the state one time, we instead measure an ensemble. If the number of measured systems is known to be N then the likelihood of the state ρ is

$$\mathcal{L}[\rho] = \prod_k \mathrm{Tr}[\rho E_k]^{N_k}, \tag{1.38}$$

where N_k is the number of times the outcome was k. In the scenarios considered in this thesis, the number of states is unknown. The systems are produced at a uniform rate thus the probability of observing a particular number of counts follows a Poissonian distribution. The likelihood of a certain measurement result $N_1, N_2, \cdots, N_{d^2}$ is thus

$$\mathcal{L}[\rho] = \prod_k \Pr[\mathrm{Poi}[\mathrm{Tr}[R\rho E_k]] = N_k] \tag{1.39}$$

$$= \prod_k \Pr[\mathrm{Poi}[Rf(k;\theta)] = N_k], \tag{1.40}$$

where R denotes the mean of the Poisson distribution describing the number of states created. We could stop here and numerically maximize the likelihood in Eq. (1.40) however it is common in the literature to maximize an equation that is approximately the same. A Poissonian distribution with a large mean is approximately a normal distribution with the same mean and a variance equal to the mean. This allows us to rewrite the likelihood in Eq. (1.40) as

$$\mathcal{L}[\rho] = \prod_k C_k \exp\left\{\frac{(N_k - Rf_k)^2}{2Rf_k}\right\}, \tag{1.41}$$

[13]These assumptions are obviously not true in general. In our situation they are however an acceptable approximation.

where the C_k's are normalization coefficients. It should be noted that the normalization coefficients depend on the parameters being estimated. Experiments often (for example see [27]) estimate these, using the same data, prior to calculating the MLE. They are estimated by a sample mean of one data point. This is substituted into Eq. (1.41) so that the normalization coefficients become constants with respect to the maximization. This can be viewed as making an additional approximation to the true likelihood function. In this estimated error likelihood scenario discussed above, the log-likelihood is then

$$\ell[\rho] = \frac{(N_k - Rf_k)^2}{2Rf_k} + C \tag{1.42}$$

In this thesis I directly numerically maximize the likelihood in Eq. (1.40) although the estimates do not vary substantially if the approximated likelihood in Eq. (1.42) is maximized instead.

Confidence regions can be constructed for the density matrix directly using the parametric bootstrap, however, we often are more interested in the fidelity of the state with a target state and the uncertainty of that statistic.

The fidelity of the state ρ with a constant target state σ is some function of the parameters θ that describe the state, i.e.,

$$F(\rho, \sigma) = g(\theta). \tag{1.43}$$

The form of the function g depends on the target state. The equivariance property of an MLE states that

$$\widehat{g(\theta)} = g(\hat{\theta}). \tag{1.44}$$

where $\hat{\theta}$ is the MLE of the parameter and $\widehat{g(\theta)}$ is the MLE of the function. This means the ML estimate of the fidelity can be calculated by finding the fidelity of the ML estimate of the state.

The confidence intervals for the fidelity can be calculated by re-sampling the counts. The distribution from which the counts are sampled has the MLE as its parameter, i.e.,

$$\tilde{N}_k \sim \text{Poi}[R \,\text{Tr}[\hat{\rho} E_k]], \tag{1.45}$$

where $\hat{\rho}$ is the MLE for the density matrix and $\tilde{N}_k$ is the random variable to be sampled B times to get the simulated counts for the k^{th} outcome. Each measurement is sampled, then the set of counts, $\{\tilde{n}_k^{(b)}\}_k$, are used to find a MLE for the state which can then be used to calculate the fidelity. This is repeated B times according to the parameteric bootstrap procedure to find the variance of the fidelity statistic.

Finally, a mention of tomographic completeness should be made. Imagine you roll a dice and only record if its even or odd. If you knew the probability of the result exactly you still would not know the state and there would be a large set of equally likely results in the case of finite statistics so your MLE estimate would identify a subspace. A tomographically complete measurement is one where if we knew the exact probabilities of each outcome we could uniquely identify the state. In this case of finite run statistics this means our MLE estimate will be more likely to be unique and will converge as sample size increase to a unique solution. Mathematically the POVMs span[14] the space of all valid density matrices. There are no additional complications using the MLE approach while doing 'over-complete' tomography. This is the case when there exists a non-trivial subset of POVMs which are themselves tomographically complete.

1.5 Bell Inequalities

One of the most surprising and unsettling properties of quantum theory is its ability to violate a Bell inequality. As the name suggests, these inequalities were originally derived by John Stewart Bell in 1964 [5] to highlight the tension between local realism and quantum theory. They were modified soon after to be robust to noise and imperfections of experiments by Clauser-Horne-Shimony-Holt (CHSH) [12]. The CHSH inequality will be the subject of this discussion, however, we will adopt a more modern approach as found in the paper by Wood and Spekkens in 2015 [51].

Consider two systems, which may have previously interacted, that are now sent to separated labs, often referred to as Alice and Bob. The labs decide their measurement settings independently. In addition, this must be done in a manner such that the choice of measurement setting in one lab and the measurement outcome in the other are space-like separated. The settings of Alice and Bob's measurement device are denoted s, t and the outcome of their measurement x, y respectively. We consider the scenario where Alice and Bob implement only binary outcome measurements and each have only two measurement options. Two assumptions that are central to classical physics are:

1. Correlations have a causal explanation (Reichenbach's principle);
2. Space-like separated events cannot causally influence each other (relativistic causality).

[14] The span is taken over the real numbers.

These place restrictions on the possible correlations an experiment can observe. These assumptions will be referred to as R-RC (Reichenbach-Relativistic Causality) as a shorthand. In the experiment described above, these assumptions imply the restriction

$$p(xy|st) = \sum_{\lambda} p(\lambda)p(x|s,\lambda)p(y|t,\lambda), \tag{1.46}$$

where λ is the state of the two particles. Bell experiments ask if the underlying distribution, also called the 'true distribution', can be factored in such a way by sampling it numerous times. The distribution of these samples is unimaginatively called the *sample distribution*.

A distribution that factors according to Eq. (1.46) also obeys the following inequality

$$\mathbb{E}[x_0y_0] + \mathbb{E}[x_1y_0] + \mathbb{E}[x_0y_1] - \mathbb{E}[x_1y_1] \leq 2, \tag{1.47}$$

where $\mathbb{E}[x_sy_t] = \sum_{s,t} xy\ p(xy|st)$ is the expectation value of the random variable XY, when the Alice and Bob's settings are s, t respectively. The left side of this inequality is called the **CHSH statistic**.

A naive way of conducting a Bell experiment would be to check if the sample distribution obeys Eq. (1.47). However, this assumes that the sample distribution is exactly the underlying true distribution. This assumption is far from guaranteed if finitely many measurements are made. This is akin to assuming that a coin is unbiased after flipping it twice and observing heads and tails each once.[15]

Despite early Bell experiments [3, 48] relying on this assumption in their analysis, this troubling assumption is not required. A Bell experiment is exactly a well studied scenario called hypothesis testing, where R-RC is the null hypothesis. Modern Bell experiments are analyzed this way [29]. In a hypothesis test, a set of potential sample distribution is defined called the rejection region. After performing the experiment, if the sample distribution is found to be inside the rejection region, we reject the null hypothesis. A hypothesis test often utilizes a statistic whose value is bounded on the set of distributions which obey the null hypothesis to define the rejection region. The CHSH statistic in Eq. (1.47) is precisely this. For example, define the rejection region as all sample distributions with a CHSH statistic greater then 2. A test statistic, like the CHSH parameter, can define a continuum of rejection regions this way.

An important measure of the quality of a hypothesis is its *size* [47]. The *size* is an upper bound on the probability that the null hypothesis is rejected despite it being true.[16]

[15] An even more alarming example is the case where the coin is heads both times. The conclusion would be that the coin would always return heads for the next infinitely many coin tosses.

[16] This outcome is called a type 1 error.

An interesting quantity for an experiment is the maximum size a hypothesis test could have while still rejecting the null hypothesis. This is called the *p-value*. This is the value Bell experiments report.

So far we haven't mentioned quantum mechanics in this section, however the reason this particular scenario is of interest is because quantum mechanics allows distributions that violate Eq. (1.47). These violating distributions are only predicted when the joint state is entangled. A violation would force us to reject the null hypothesis of R-RC. Note that this doesn't mean we accept quantum mechanics. The entire discussion of Bell experiment could be done without reference to quantum physics as it is a question about the validity of Reichenbach's principle and relativistic causality. Quantum mechanics, due to its impeccable agreement with reality, guides us to this particular scenario to test our assumptions.

1.5.1 Loopholes

Experimental attempts at violating a Bell inequality are often complicated by *loopholes* which prevent the conclusion to reject R-RC. An ideal Bell experiment would assume only these two principles (or just local realism) and then place a constraint on an experiment's results. This constrained region is then used as the null hypothesis. A loophole is an additional assumption that must be made due to the particular experimental implementation [29]. The null hypothesis in an experiment with loopholes is local causality and these additional assumptions. Someone could rightfully argue that local causality is still valid, if they instead accept one of the other assumptions to be false. Common examples of loopholes include memory, failure of fair sampling and failure of locality. A loophole-free violation was only first achieved as recently as 2015 [40, 24, 20]. The experiments within this thesis do not attempt to be 'loophole free' as the presence of them does not interfere with the conclusions we draw.

Chapter 2

Experimental Quantum Optics

This thesis concerns itself with the analysis of Bell-type experiments. We generated this data by performing quantum optical experiments. This chapters presents an overview of the required concepts from this field with a focus on the polarization degree of freedom. The general information contained in this chapter is loosely based on a variety of introductory textbooks and review papers [8, 23, 28, 30]. An interested reader is encouraged to read these references for a more broad and technical understanding.

There exists many possible physical systems upon which to encode a qubit or a higher dimensional quantum state. Common examples include trapped ions, superconducting systems, photons and spin states of nuclei. The ideal implementation depends on the desired application. In a Bell experiment, the required elements are: a source of entangled qubits, single qubit operations and projective measurements. Photons, in particular the polarization degree of freedom, are our system of choice. They interact comparatively little with each other and the environment which is advantageous for maintaining coherence of the encoded state. This is a also drawback when attempting two-qubit gates, however, we do not require this in our experiments. In our case, this implementation offers relatively simple and stable systems at room temperature. In addition, polarization entangled qubits can be readily produced using spontaneous parametric down-conversion (SPDC). Single qubit unitaries and projective measurements can be accomplished using only linear optical components: polarizing beamsplitters and linear retarders.

2.1 Photonic Qubits

A photon is an excitation of the electromagnetic field. The state space of a photon is infinite-dimensional. Its polarization degree of freedom however is two-dimensional. We only deal with photons effectively identical in all ways except polarization[1], allowing us to fully describe a photon's state by only its polarization state. The computational basis is taken to be the orthogonal horizontal and vertical polarizations. An arbitrary pure state can be written as

$$|\psi\rangle = \cos(\theta)\,|H\rangle + e^{i\phi}\sin(\theta)\,|V\rangle\,, \tag{2.1}$$

where θ, ϕ are real numbers. Other common states include: diagonal, anti-diagonal, right circular and left circular, which in that order are defined as:

$$|D\rangle = \frac{1}{\sqrt{2}}\big(\,|H\rangle + |V\rangle\,\big), \tag{2.2}$$

$$|A\rangle = \frac{1}{\sqrt{2}}\big(\,|H\rangle - |V\rangle\,\big), \tag{2.3}$$

$$|R\rangle = \frac{1}{\sqrt{2}}\big(\,|H\rangle + i\,|V\rangle\,\big), \tag{2.4}$$

$$|L\rangle = \frac{1}{\sqrt{2}}\big(\,|H\rangle - i\,|V\rangle\,\big). \tag{2.5}$$

Polarization states with real coefficients in the H/V basis are called linearly polarized, for example $|H\rangle\,, |V\rangle\,, |D\rangle\,, |A\rangle$.

2.2 Wave plates

An arbitrary single qubit unitary[2] on the polarization state of a photon can be performed with three pieces of uniaxial birefrigent glass by adjusting the angle of their optical axes[41]. A birefrigent material has a 'fast' axis and a perpendicular 'slow' axis. Light travelling along the slow axis will get a relative delay and thereby picks up a relative phase. Mathematically, this is a map which implements the transformation

$$\begin{aligned} |f\rangle &\to |f\rangle \\ |s\rangle &\to e^{i\phi}\,|s\rangle\,, \end{aligned} \tag{2.6}$$

[1] We actually only need to assume that our transformations and measurements are blind to any other properties of the photon.

[2] We actually only need to implement all operators in SU(2) since global phase doesn't matter.

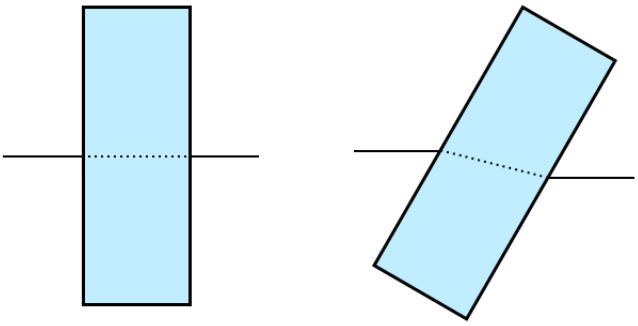

Figure 2.1: Diagram of a tilted wave plate. The glass on the right is rotated to an angle which is non-orthogonal to the beam. This increases the distance the beam must travel through it.

where $|f\rangle, |s\rangle$ are polarization states aligned along the fast and slow axes respectively.

A half waveplate (HWP) and quarter waveplate (QWP) cause a relative phase delay of $e^{i\pi}$ and $e^{i\pi/2}$ respectively. These are the most commonly used elements to manipulate polarization. The action of the waveplate on a polarization state can be changed by rotating the fast axis relative to the horizontal axis. The unitaries in the H, V basis, up to an irrelevant global phase factor, for a HWP and QWP are [13]:

$$HWP[\theta] = \begin{pmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{pmatrix} \tag{2.7}$$

$$QWP[\theta] = \begin{pmatrix} \cos^2(\theta) - i\sin^2(\theta) & (1+i)\cos(\theta)\sin(\theta) \\ (1+i)\cos(\theta)\sin(\theta) & -i\cos^2(\theta) + \sin^2(\theta) \end{pmatrix}, \tag{2.8}$$

where θ denotes the angle between the fast axis and the horizontal axis.

The final polarization manipulation device we will talk about is a variable retarder. The half and quarter waveplate assign a fixed relative delay. In some situations however the ability to adjust the delay in Eq. (2.1) from 0 to 2π is needed. A simple way to achieve this is to take a piece of birefrigent glass, a quarter waveplate for example, and rotate about an axis perpendicular to the beam, as shown in Figure 2.1, to change the amount of glass the beam passes and thereby the relative delay.

If the amount of retardation must be changed quickly, then an electro-optic device is more appropriate. We use a nematic liquid crystal retarder (LCR) in this situation [9]. These devices are comprised of crystals where the orientation affects the retardation of the

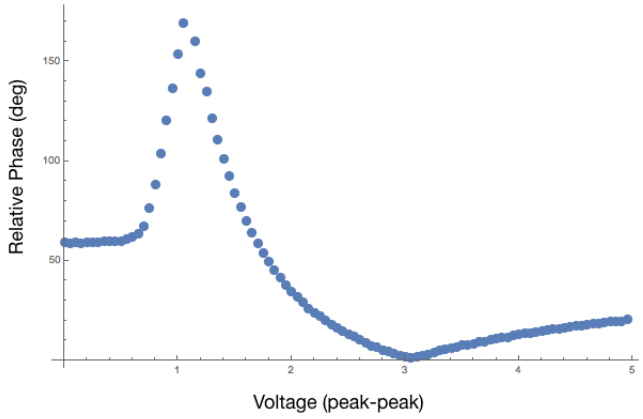

Figure 2.2: Plot of the relative phase delay as a function of the AC voltage applied for a typical LCR.

device. The crystals can be rotated by applying a voltage across the device allowing rapid control of the retardation.

2.3 Single Qubit Operations and Measurements

The detectors in the experiment are not polarization sensitive. Instead we use a polarizing beamsplitter (PBS) to couple the polarization degree of freedom to different spatial paths. A PBS reflects vertically polarized light and transmits horizontal. A detector can then be placed along the two paths. This allows us to perform projective measurements onto the $|H\rangle, |V\rangle$ basis.

An arbitrary unitary transformation on a polarization qubit can be done using a Q-H-Q[3] device [41]. In our experiments we are more interested in transforming an arbitrary polarization to $|H\rangle$ or $|V\rangle$ so that we can perform a general projective measurement when combined with a PBS. This operation can be performed using two waveplates: either H-Q or Q-H [13].

We use an H-Q device followed by a PBS as seen in Figure 2.3. The probability the polarization analyser would measure the state $|\psi\rangle$ at the transmitted port, using Born's rule and the unitary representations of the HWP and QWP as given in Eq. (2.7) and

[3] A Q-H-Q device is a QWP, followed by a HWP, then a second QWP.

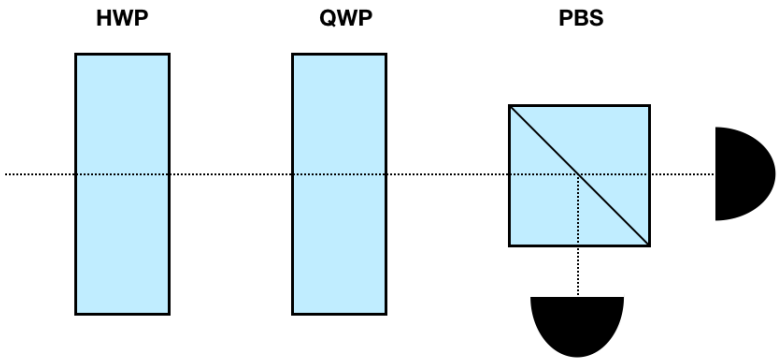

Figure 2.3: Optical diagram of our polarization analyser. We can project onto an arbitrary pure qubit state by adjusting the angles of the half and quarter waveplates.

Eq. (2.8), is

$$|\langle H| QWP[\phi] HWP[\theta] |\psi\rangle|^2, \tag{2.9}$$

where θ, ϕ are the angles that the QWP and HWP are positioned. To find the angles necessary to project onto a given polarization state $|\psi\rangle$ we numerically maximize Eq. (2.9) over θ, ϕ.

Optical fibres in general do not preserve the polarization of the light that enters and will perform a unitary transformation. To solve this, polarization maintaining fibers can be used, or an arbitrary unitary operation needs to be implemented to reverse any polarization change. We use the latter approach. As opposed to using a traditional Q-H-Q device for our arbitrary unitary we use a H-Q device followed by a piece of glass on a mount that can be rotated about an axis perpendicular to the laser beam. A H-Q device brings an arbitrary polarization to a linear polarization. So we can set it to ensure that whatever polarization exits the fibre when horizontal light enters is then returned to horizontal. This means the combination of a fibre and H-Q device preserves $|H\rangle$ and $|V\rangle$, however the relative phase between them is in general non-zero. The final tilted birefrigent glass, with its fast axis aligned horizontally, allows this phase to be compensated.

We also wish to implement non-unitary channels, in particular a dephasing channel

$$\rho \to p\rho + (1-p)\sigma_x \rho \sigma_x \tag{2.10}$$

where $0 \leq p \leq 1$ is the amount of dephasing and σ_x is a Pauli matrix. This can be accomplished using a LCR. The LCR in Figure 2.2 would act as an identity gate when

driven with an approximately 3V square wave and as a HWP when driven with a 1V square wave. As seen from Eq. (2.7) with $\theta = 45°$, a HWP with its fast axis 45 degrees from horizontal implements a Pauli σ_x gate. By fixing the fast axis of the LCR 45 degrees from horizontal, we can switch between an identity and σ_x gate quickly by adjusting the voltage. This voltage can be switched many times within the time one measurement is taken allowing us to simulate a dephasing channel. The relative amount of time spent at each voltage will dictate the amount of dephasing with an even mixture resulting in complete dephasing.

2.4 Entangled Photon Source

The final resource we require is a source of polarization entangled qubits. We used the same source as previous experiments in the lab [45, 21] which in turn were based on the sources in [16, 50]. See [45] for an in-depth description of the device and its alignment procedures. This subsection will provide a working overview of its function.

We use a periodically poled potassium-titanyl-phosphate (PPKTP) crystal to do degenerate type II down-conversion. If it is pumped with a horizontally polarized blue laser (404nm), a small fraction of the photons will down-convert into pairs of red photons (808nm). When considering only the polarization degree of freedom, the mathematical description of this process is

$$|H\rangle \to |HV\rangle \,. \tag{2.11}$$

Once these pairs of qubits are available, we are able to do single qubit unitaries to create any two-qubit product state. However single qubit gates on the state in Eq. (2.11) are not sufficient to produce entanglement. To create entanglement, we use two sources and create a state which is in an equal superposition of each source's individual output. To accomplish this, we coherently pump both sources with the same blue laser and map their outputs into two modes as shown in Figure 2.4. Due to the low efficiency of down-conversion, the probability of only one of the two sources down-converting in a short time interval[4] dominates multi-pair events. We ignore multi-pair events for the remainder of this analysis.

If the left source down-converts the resulting state over the modes a, b is $|V_1\rangle |H_1\rangle$. If the right source down-converts the resulting state is $|H_2\rangle |V_2\rangle$. They are coherently pumped so these two options are in a superposition giving the state

$$|V_1\rangle |H_1\rangle + |H_2\rangle |V_2\rangle \,. \tag{2.12}$$

[4]The time interval is set to 3ns as each detector has approximately a 1ns uncertainty.

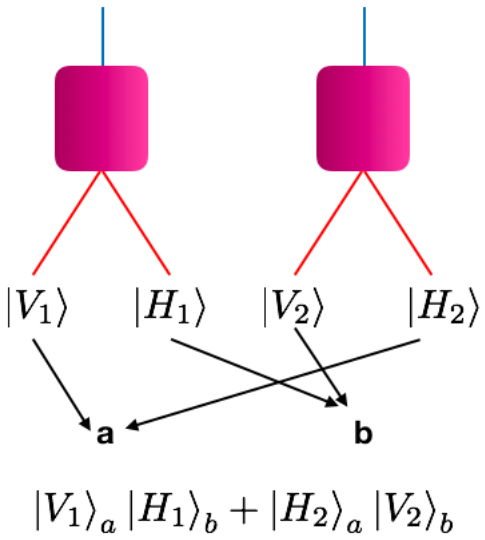

Figure 2.4: Two type II SPDC sources being coherantly pumped with their output modes overlaid. The two events: source A creating a pair and source B creating a pair are in a superposition if we post select only on situation where only one pair measured in the output modes. If origin source remains unknown, the resulting state over the two output modes labelled a, b is $|VH\rangle + |HV\rangle$.

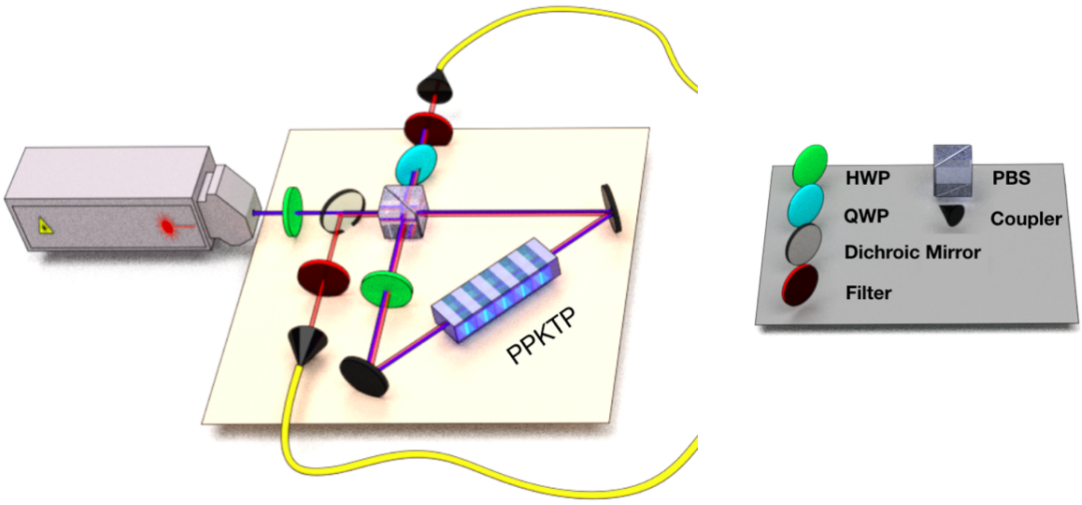

Figure 2.5: A sagnac interferometer entangled photon source.

If the two sources are identical, we cannot tell which source the photons came from so the output state becomes

$$|V\rangle |H\rangle + |H\rangle |V\rangle , \tag{2.13}$$

which is a maximally entangled state.

The source we use is based upon the same idea. As opposed to using two different sources, we place the PPKTP inside a sagnac interferometer. The two sources discussed earlier are formed from light propagating clockwise and counter clockwise around the interferometer. A diagonally polarized blue laser (404nm) is sent into the interferometer. The horizontal component travels counterclockwise while the vertical polarization goes clockwise. Independent of which direction had down-conversion, one photon will go to each coupler on the output of the source. The output of the source is the state in Eq. (2.13). This can be transformed using single qubit unitaries into an arbitrary maximally entangled state. We can also reduce the amount of entanglement by applying a depolarizing channel to one of the qubits.

Chapter 3

Observational Causal Discovery

This section outlines a novel method of observational causal discovery developed for the experiment presented in Chapter 4. The algorithm begins by converting the causal explanations to statistical models. It then utilizes concepts in the field of statistical learning to select a model. In contrast to existing methods [11, 43], it can also be applied to quantum causal models without any modifications. This allows the direct comparison of quantum and classical causal explanations. The core principle upon which the algorithm is based is that the causal explanation which minimizes the predictive error should be favoured. This limits philosophical arguments and embraces a pragmatic operational approach.

3.1 Introduction

Correlation does not imply causation. This common saying is correct in spirit but not in practice[1]. Particularly in recent years, using correlations to infer causal information is precisely the goal of many studies. The objective of these studies is referred to as causal discovery. In the field of medicine for example, a drug's causal influence on the patient's recovery may be of interest. Ideally, a randomized trial is conducted. This is an example of interventional causal discovery. In such a scenario the experimenter forces a variable, in this case drug use, to take on a certain value independent of any potential causal influences. This removes the possibility of confounding causes. Unfortunately,

[1]It would however be problematic to assume that two variables could correlate without any sort of causal mechanism. Reichenbach's principle states that this is prohibited and it is an important underlying assumption.

experiments of this nature are not always possible. For example, consider a study looking at the effect of diabetes on hair growth. It would be impractical and unethical to randomly assign participants with diabetes. The causal influences must be inferred from purely observational data. We focus on this paradigm of causal discovery.

In quantum foundations we may wish to avoid making interventions for a few different reasons. We may want to avoid assuming an intervention was made successfully due to the assumptions on the device's mechanics this would require. In particular it would be difficult to compare classical and quantum causal models as the mechanics of the intervention would depend on the theory. Minimizing the number of physical assumptions is crucial. Another reason an observational analysis may be preferred is that when testing theories like superdeterminisnm and local causation it can be extremely difficult to ensure we truly did manage to randomize the variable.

Methods of doing causal discovery with interventions have previously been investigated in quantum foundations [19, 37, 38], however, to the author's knowledge this work is the first example of observational causal inference. In addition, it is also the first method able to compare classical and quantum causal explanations of an experiment.

3.2 Causal Models

This section will review the essentials of classical causal models and is based on the treatment by Pearl [35].

We consider experiments where each run is independent. An experiment consists of a set of variables which are sampled multiple times. A *causal model* for such an experiment specifies the dependency[2] of each variable on the values of the others. Often, the exact nature of the dependencies aren't of interest. Instead, what of interest is determining the set of variables upon which the variable has a non-trivial dependence. This specification is called a *causal structure*. A succinct way of representing a causal structure is a directed acyclic graph (DAG). Each variable in the experiment is a vertex. If there exists a direct causal influence between two variables, there is an edge between their respective vertices. A graph representing a causal structure is directed since causal influence is not symmetric. Actually, if A causally influences B, then B never can influence A because causation can only propagate forward in time. By similar logic, the graphs must be acyclic. Causal discovery attempts to determine which causal *structure* best describes an experiment.

[2] A dependency is not the same as a correlation. It is a conditional probability.

A variable A in a DAG is called a *parent* of another variable B if there is an edge going from A to B. Given a causal structure, a causal model can be specified. A causal model specifies the probability distribution for each variable given the value of the variable's parents. A related concept, which connects the languages of probability and causality, is the causal Markov condition

$$p(x_1, x_2, \cdots, x_n) = \prod_{k=1}^{n} p(x_k|\mathrm{Pa}(X_k)), \tag{3.1}$$

where $\mathrm{Pa}(x_k)$ is the set of parents of the variable x_k. The set of parameters in the causal model is the set of probability distributions $p(x_k|\mathrm{Pa}(X_k))$.

A causal model specifies a joint distribution over all the variables in an experiment. A probability distribution is *compatible* with a causal structure if there exists a causal model for the structure which results in the same joint distribution. There are clearly multiple non-equivalent causal models for the same causal structure. A causal structure still places constraints on the probability distributions compatible with it. The most common constraints are called conditional independence constraints. The set of all conditional independence constraints for a DAG can be found using a method called d-separation. If two DAGs give rise to the same set of conditional independence constraints, they are called *Markov equivalent*. Two Markov equivalent DAGs may however have different sets of compatible probability distributions. Two DAGs with the same compatible distributions are called *observationally equivalent*. A causal discovery algorithm can never do better then selecting a class of observationally equivalent DAGs. Many common algorithms only attempt to learn the Markov equivalency class.

3.3 Noiseless Causal Discovery

An interesting paradigm of causal discovery considers when the exact underlying joint probability distribution is known. This would rarely be the case since the experiment would have to be without any noise, which requires infinitely many samples. This is also referred to as infinite run data since in the case where the experiment is conducted infinitely many times, the sample mean is equal to the true underlying distribution without any noise. Even in this scenario, determining the observational equivalence class (OEC) is non-trivial. For a particular DAG, infinitely many causal models would have to be searched in order to determine if they result in the observed distribution. In addition, the probability distribution will be in general compatible with multiple OECs. Finding the

Markov equivalency class (MEC) is a much simpler problem. This can be accomplished by testing if the distribution obeys the finite set of independence relations which define the MEC. Generally, a probability distribution will also be compatible with multiple MECs. A standard principle used to resolve this problem is *faithfulness*. This principle states that the compatible MEC with the maximal number of independence conditions should be chosen. This implies that the probability distribution obeys every independence condition required by the selected MEC but no additional ones. A more complicated and connected DAG has fewer independence conditions. For this reason the faithfulness can be viewed as an "Occam's razor" argument giving preference to the simplest compatible explanation.

The situation described above changes noticeably when the probability distribution is not observed directly, but instead sampled finitely many times. This is the paradigm relevant to most situations and the subject of the remaining sections.

3.4 Noisy Causal Discovery

This section outlines how observational causal model discovery on finite, also called *noisy*, data can be viewed as a problem of statistical model selection.

The approach differs highly from the one used in the exact data case. A causal structure is treated similarly to a noise model. Including a noise model in an analysis will increase its predictive accuracy, if the data is in fact distributed as specified by that noise model. This observation is the motivation behind the causal discovery algorithm. The analysis is performed multiple times, assuming a different causal structure each time. The structure whose corresponding estimate results in the highest predictive accuracy is chosen. This can be viewed as assigning a score to each causal structure similar to the approach in the purely classical greedy equivalence search algorithm [11].

The most pertinent estimation problem for the experiments we consider is predicting the outcomes of one set of variables based on the values of another set. This is a regression problem. We could blindly fit, i.e. without any assumptions, a regression function on the data. However, by making causal assumptions we limit the possible forms of the regression function. Our causal discovery method relies on this fact. If restricting the regression function to be compatible with a particular causal structure increases the predictive accuracy, then it is likely that the distribution being sampled is compatible with that causal structure. Notice that a condition similar to faithfulness is also achieved without having to add it as an additional assumption. The compatible models which are more complicated, and thereby are compatible with more distributions, would be prone to over-fitting. This

is analogous to the following scenario. If the data is sampled from a linear function with noise, the best fit out of all polynomial functions, which include linear function, essentially connects the data points. This results in over-fitting and harms the functions predictive accuracy on future data sets.

A final important observation is that any reference to a true distribution can be removed from the interpretation of the causal discovery algorithm. This is done by arguing that if restricting to only distributions compatible with a certain causal structure helps, relative to the other causal structures, then this same restriction should be made for similar experiments in the future. This doesn't require the experimenter to claim the true distribution is compatible with that causal structure. The experimenter could only report that performing the analysis according to this causal structure is superior than according to the other proposed experiments. A future researcher, performing a similar experiment, after reading the result would then be motivated to in their analysis restrict only to distributions compatible with a similar causal structure.

The following sections will explicitly outline the regression problem, the constraints resulting from a causal structure assumption and the ensuing model selection.

3.4.1 Outline of Problem

Consider the following general type of experiment. There is a set of variables and we wish to predict the values of one subset $\vec{X}$ given an observation of the value of a second subset $\vec{S}$. These sets are called *outcomes* and *settings* respectively. They are sometimes referred to as *responses* and *predictors*. The predictions are calculated using data from previous runs of the experiment. The variables in the experiment obey a joint probability distribution $p(\vec{x}, \vec{s})$, referred to as the underlying or "true" distribution. A data point is a measurement of a random variable, distributed according to $p(\vec{x}, \vec{s})$.

The experimenter may wish to provide a single outcome as the prediction for a given setting. Alternatively, they could attempt to predict the probabilities of each outcome. The second case is what is considered here. To pursue this second case it's helpful to "dummy encode" our outcome variables in order to infer the probabilities. To perform this encoding, if $\vec{X}$ has m possible outcomes labelled $x_1, x_2, \cdots, x_m$, let $\vec{F} \in \mathbb{R}^m$. Define the random variable $\vec{F}$ such that if $\vec{X}$ equals its i^{th} outcome, the i^{th} component of $\vec{F}$ is 1 and the rest are zero. For example, if $\vec{X} = x_3$, then $\vec{F} = (0, 0, 1, 0 \cdots, 0)$. The joint probability distribution $P(\vec{f}, \vec{s})$ of the new random variable $\vec{F}$ and the settings $\vec{s}$ can be determined from $p(\vec{x}, \vec{s})$. An illustrative relationship between these two random variables is that

$$\mathbb{E}[\vec{F}] = (p(\vec{X} = x_1), p(\vec{X} = x_2), \cdots, p(\vec{X} = x_m)). \tag{3.2}$$

By estimating $\vec{F}$, we effectively estimate the probability of each outcome of $\vec{X}$ as desired.

Mathematically, the problem is to find a function $f(\vec{f}|\vec{s})$, which estimates $\mathbb{E}\big[P(\vec{f}|\vec{s})\big]$ given

$$(\vec{f}^{(1)}, \vec{s}^{(1)}), (\vec{f}^{(2)}, \vec{s}^{(2)}), \cdots, (\vec{f}^{(m)}, \vec{s}^{(m)}) \sim P(\vec{f}, \vec{s}). \tag{3.3}$$

The observations of the independent identically distributed random variables in Eq. (3.3) are the data collected from the experiment and is denoted $\mathcal{D}$. The random variables $\vec{F}^{(k)}$ are called frequencies since physically they are the observed frequencies of each outcome during a particular run of the experiment.

The form of the probability density $P(\vec{f}, \vec{s})$ depends on the manner by which the data is collected. As an example, in the hypothetical ideal case where the data is collected without noise

$$P(\vec{F}, \vec{S}) = \begin{cases} p(\vec{S} = s_1) & \vec{F} = \big(p(\vec{X} = x_1|\vec{s} = s_1), \cdots, p(\vec{X} = x_m|\vec{s} = s_1)\big), \vec{S} = s_1 \\ p(\vec{S} = s_2) & \vec{F} = \big(p(\vec{X} = x_1|\vec{s} = s_2), \cdots, p(\vec{X} = x_m|\vec{s} = s_2)\big), \vec{S} = s_2 \\ & \vdots \\ p(\vec{S} = s_n) & \vec{F} = \big(p(\vec{X} = x_1|\vec{s} = s_n), \cdots, p(\vec{X} = x_m|\vec{s} = s_n)\big), \vec{S} = s_n \\ 0 & \text{else} \end{cases}, \tag{3.4}$$

where $p(\vec{x}, \vec{s})$ is the exact underlying infinite run distribution. Collecting data of this form is akin to finding a way to measure, in one observation without uncertainty, the bias of a coin toss as opposed to flipping it multiple times. Other common situations include the data being sampled from a Gaussian or Poissonian distribution with the appropriate mean. Independent of the data collection method, we expect

$$\mathbb{E}_P(\vec{F}|\vec{s}) = \big(p(\vec{X} = x_m|\vec{s}), p(\vec{X} = x_{m-1}|\vec{s}), \cdots, p(\vec{X} = x_1|\vec{s})\big). \tag{3.5}$$

For example, consider the case where $\vec{S}$ is trivial and $\vec{X}$ is the result of a coin toss. If the sampling is done by flipping the coin, let $\vec{F} = (0, 1)$ for heads and $\vec{F} = (1, 0)$ for tails. The outcome of $\vec{F}$ for each toss is distributed according to $P[\vec{F} = (0, 1)] = p(\text{heads}), P[\vec{F} = (1, 0)] = p(\text{tails})$ and zero otherwise. The expectation value of F with respect to this distribution p is

$$p[\vec{F} = (0, 1)] * (0, 1) + p[\vec{F} = (1, 0)] * (1, 0) = \big(p[\vec{F} = (1, 0)], p[\vec{F} = (0, 1)]\big). \tag{3.6}$$

This agrees with Eq. (3.5).

One may ask if all of this complicating dummy variable encoding is necessary. The reason that dummy coding is necessary is that the causal models we are considering do not

need to be deterministic. If A causally influences B, that does not mean knowing the value A is sufficient to predict the value of B with certainty. We instead consider a more general scenario where different values of A result in different probability distributions for B. This means a causal structure places restrictions on the expectation values of the distribution we sample. The next section outlines methods to measure how "good" a particular restriction is.

3.4.2 Statistical Model Selection

Consider a data set $\mathcal{D} = \{(s^{(k)}, f^{(k)})\}_{k=1}^{m}$ as described in Eq. (3.3). We wish to find an estimator, also called a *regression function* $\hat{r}$, which when given a value of $\vec{s}$ returns a prediction for $\vec{f}$. This function is called an estimator since it can be viewed as an estimate of $\mathbb{E}[P(\vec{f}|\vec{s})]$. There are multiple methods for calculating regression functions. Here we focus on calculating regression functions by minimizing a loss function. A loss function specifies how close an estimate is to the true value, similar to choosing a metric. A common loss function is squared error, the regression function that results from minimizing this loss is given by

$$\hat{r} = \underset{r}{\operatorname{argmin}}\ L(r, \mathcal{D}) \tag{3.7}$$

$$= \underset{r}{\operatorname{argmin}} \sum_{k=1}^{m} \left(r(s^{(k)}) - f^{(k)}\right)^2, \tag{3.8}$$

where $\mathcal{D} = \{(s^{(k)}, f^{(k)})\}_{k=1}^{m}$ is the data that is being fitted. This data set will be referred to as the *training data* since we can view the data as educating or training our choice of estimator.

Associated to each causal structure is the set of its compatible probability distributions. This set is given by

$$M_{_G} = \{P(x_1, \ldots, x_n) \mid \exists\ P_1, \ldots, P_n\ s.t.\ P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_i(x_i|\mathrm{Pa}(X_i))\}, \tag{3.9}$$

where G is the DAG over a set of variables $X_1, X_2, ..., X_n$, $\mathrm{Pa}(X_i)$ are the parents of the vertex X_1 in the graph G, and $P_1, ..., P_n$ are all valid conditional probability distributions. Alternatively, we can consider the set of *conditional* probability distributions, of the outcome variables conditioned on the settings, which are compatible with the causal structure. This set is denoted $J_{_G}$. In this problem, the set of compatible conditional probability distributions of the form $P(\vec{f}|\vec{s})$, which we denote $M_{_G}$, is of interest.

The estimate of $\mathbb{E}[P(\vec{f}|\vec{s})]$ for a causal structure assumption G is the function within M_G which minimizes the loss on the data $\mathcal{D}$. The minimal value of the loss is called the *training error*

$$\text{err} = L(r, \mathcal{D}). \tag{3.10}$$

Explicitly, the estimate is given by

$$\hat{r}_G = \min_{r \in M_G} L(r, \mathcal{D}). \tag{3.11}$$

The question is now framed as deciding which estimate $\hat{r}_G$ is best. The field of statistical learning provides techniques to achieve this goal. Chapter 7 of [22] is highly recommended to the reader for additional information on this topic.

Firstly, we need to define a quantity to measure the success of an estimator. A clear choice is the *generalization error*. Upon receiving an independent new data pair $(\vec{f}, \vec{s})$, the generalization error is the expected discrepancy between the prediction $\hat{r}(\vec{s})$ and the observed value $\vec{F}$. There is no reason to use a loss function that differs from the one chosen in Eq. (3.11) to find the regression function. Mathematically, the generalization error is given by

$$\text{Err} = \mathbb{E}_{\vec{f},\vec{s}}[L(\hat{r}(\vec{s}), \vec{f})]. \tag{3.12}$$

We clearly cannot directly calculate this expectation value since it would require knowledge of the underlying true distribution which is what we are trying to determine in this problem.

There are many methods of estimating the generalization error including Akaike Information Criterion (AIC) [1], Bayesian Information Criterion (BIC) [39], cross validation [44] and bootstrap covariance estimation [14]. These methods are effective when taking an independent second data set is not possible or when there are too few data points to fit and test the model. This is often the case when the data points are human subjects. The experiment we consider in this thesis, and indeed most in this field, easily have the ability to generate more data so we need not concern ourselves with these methods. We simply take a second data set, $\mathcal{D}_2 = \{(s_2^{(k)}, y_2^{(k)})\}_{k=1}^m$, and calculate the average loss on this set. We call this quantity the *test error*

$$\overline{\text{Err}} = L(\hat{r}, \mathcal{D}_2) \tag{3.13}$$

$$= \sum_{k=1}^{m} (\hat{r}(s_2^{(k)}) - f_2^{(k)})^2. \tag{3.14}$$

The only assumption made here is that an average is a good estimate of an expectation value. This assumption is well motivated when there are multiple data points. In fact, this assumption is also necessary for any other method described above.

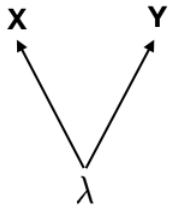

Figure 3.1: A single DAG, which as they are illustrated here, given different cardinalities of *Lambda* is compatible with different sets of distributions.

	Symbol	Definition
Training Error	err	Eq. 3.10
Generalization Error	Err	Eq. 3.12
Test Error	$\overline{\text{Err}}$	Eq. 3.14

Table 3.1: Summary of the error quantities.

3.4.3 Latent Variables

Latent variables are a common inclusion in a causal description of an experiment. These are variables that are not measured during the experiment but have causal influences on variables that are.

We need only consider latent variables that have a causal influence on two or more observed variables. A causal structure with a latent variable which only affects one event is observably equivalent to the same DAG with the latent variable and its influence removed. In our framework, any potential latent variables are considered part of a causal structure assumption. A causal structure for an experiment with observed variable $\vec{X}, \vec{S}$ is a set of latent variables $\vec{\Lambda}$ and a DAG on $(\vec{X}, \vec{S}, \vec{\Lambda})$.

A notable complication arises from the fact that the cardinality of the latent variable is unknown. The exact values do not matter. The number of options, i.e. the latent variable's cardinality, however is important. The set of compatible distributions can depend on this cardinality. To illustrate this, consider the following example. There are two binary variables X, Y and one latent variable Λ. Consider two causal structures $G1, G2$ with the same DAG (Figure 3.1). The first structure's latent variable can only take on one value while the second structure's latent variable can take on two distinct values λ_1, λ_2. No

probability distribution compatible with $G1$ allows correlations between X, Y since

$$\begin{aligned} J_{G1} &= \left\{ p(x,y) \mid p(x,y) = \sum_{\lambda=1}^{1} p(x|\lambda)p(y|\lambda) \right\}, \\ &= \left\{ p(x,y) \mid p(x,y) = p(x|\lambda=1)p(y|\lambda=1) \right\}, \\ &= \left\{ p(x,y) \mid p(x,y) = p(x)p(y) \right\}. \end{aligned} \tag{3.15}$$

However, distributions where both binary variables are perfectly correlated are compatible with $G2$. For example $p(x=0,y=0) = 1/2, p(x=1,y=1) = 1/2$ is compatible since if we let $p(\lambda=1) = 1/2, p(\lambda=2) = 1/2$ and $p(x=0|\lambda=1) = 1, p(y=0|\lambda=1) = 1, p(x=1|\lambda=2) = 1, p(y=1|\lambda=2) = 1$ then

$$\sum_{\lambda=1}^{2} p(x|\lambda)p(y|\lambda) \tag{3.16}$$

is exactly this distribution.

We have established that different specifications of a latent variable's cardinality can correspond to distinct causal assumptions.[3] Often the experimenter would not like to specify a cardinality and instead test a *cardinality-agnostic* causal assumption. One potential way of accomplishing this is to assume the cardinality is large enough that any increase would not change the set of compatible distributions. In practice this sufficient cardinality can be quite large which makes calculations difficult. Additionally, if for example the "true" distribution has a binary latent variable, then forcing the causal assumption to have a large cardinality would cause over-fitting. When we wish to include a latent variable, we consider every dimension up to this sufficient cardinality separately. If the causal assumption for every cardinality performs worse then some other causal assumption, we rule out the cardinality-agnostic causal assumption entirely.

3.4.4 Quantum Causal Structures

The framework for quantum causal structures is less developed than its classical counterpart and remains an active area of research [2, 4]. Fortunately, independent of framework,

[3] No other information about the latent variable ever needs to be specified.

a quantum causal structure still places a restriction on the compatible probability distributions for an experiment. This can be incorporated seamlessly into the causal discovery algorithm. The algorithm allows one to compare multiple classical, quantum and hybrid causal explanations directly. A hybrid quantum-classical causal structure is considered in the experiment in Chapter 4. The causal structure has a latent quantum common cause between the two response variables. This "quantum common cause" is propagated by a two qubit quantum state. One qubit effects each response. The constraints that this model applies to the set of possible regression functions M_{QLC} is found in Eq. (4.11). Besides this example, no other quantum causal structures are directly considered in this thesis. The reader should note that other causal structure frameworks being developed, such as generalized probabilistic theory causal structures [25, 18, 49], could be incorporated into the causal discovery algorithm.

3.5 Summary

This section and Figure 3.2 summarize the observational causal discovery algorithm. In this chapter, the situation where there are multiple independent runs of an experiment was considered. An experiment consisted of different events which were split into two categories: predictors and responses. All the events had finitely many possible outcomes. The goal is to determine the causal structure which best describes the data. In order to quantify each causal structure's quality, their ability to solve a prediction problem is considered. The objective of this problem was to find a regression function which predicts the probabilities of each possible response given an observation of the predictors. By assuming a causal structure during the analysis the accuracy of the regression function can be increased. A causal structure assumption results in a constrained set of regression functions that can be fit to training data. A regression function was found for each causal structure by minimizing a loss function on the training data. The experimenter collects a second independent set of data called the test data. The different regression functions are compared by measuring their success on this second data set. This quantity is called the test error Eq. (3.14). The test error acts as the score for each regression function and thereby each causal structure. The causal discovery algorithm selects the causal structure with the lowest test error.

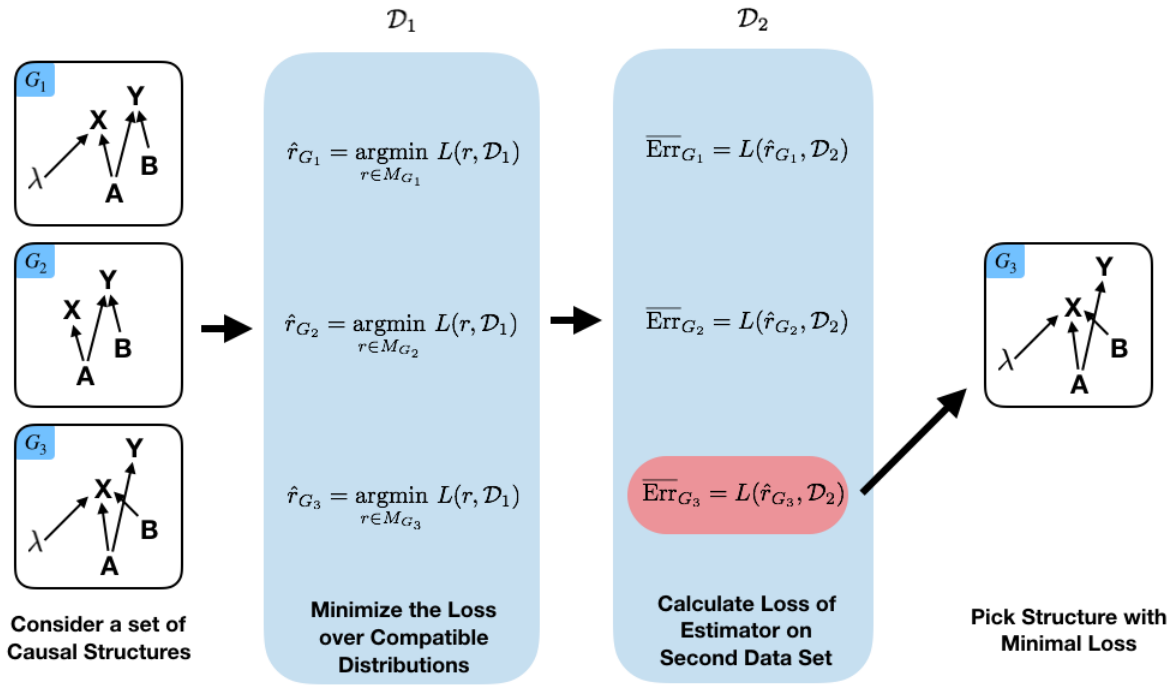

Figure 3.2: A graphic summarizing a hypothetical application of the causal discovery algorithm. Three possible causal explanations G_1, G_2, G_3 are considered. An estimator is found for each using the training data set $\mathcal{D}_1$. The testing loss is calculated using the second data set $\mathcal{D}_2$. A red box highlights the hypothetical lowest test loss, Err_{G_3}. In this example, the selected causal structure is G_3.

Chapter 4

Bell Experiment

This chapter applies the causal discovery algorithm developed in Chapter 3 to a Bell scenario. Two photonic Bell experiments are conducted and the causal discovery algorithm is performed on the resulting data. A paper is in preparation based on the content contained in this chapter. The contributions to this work are as follows:

- **Robert Spekkens** and **Kevin Resch** proposed studying Bell experiments from a causal perspective.
- **Patrick Daley** and **Kevin Resch** designed the experiment.
- **Patrick Daley** and **Robert Spekkens** developed the causal discovery algorithm.
- **Patrick Daley** conducted the experiment and performed the data analysis.
- **Patrick Daley** is the sole author of the content contained in this chapter.

4.1 Introduction

Correlation and causation are distinct yet intertwined concepts. As humans, we interpret the world from a causal perspective. If we notice that the lights turn off after a switch is flipped, we would conclude that one event *caused* the other[1]. We would not think the two events are *correlated* for no reason. Furthermore, causal explanations are of practical

[1] Another explanation is that a third event caused them both.

importance. They tell us which variables to manipulate in order to affect the target quantity. Essentially, a causal understanding allows us to control our surroundings rather than only predict them.

Traditionally, experiments and physical theories only consider correlations between measurable properties. In recent years, effort has been placed on bridging the gap between these two notions. By observing correlations, passively or by implementing interventions, one can determine the compatible causal explanations through the use of a causal discovery algorithm.

Physicists have also been studying, somewhat indirectly, the connection between correlations and causal influences in the realm of quantum mechanical systems for 50 years through Bell experiments. In the past few years this connection has been studied in a more conscientious manner. A standard Bell experiment is a hypothesis test with local realism as the null hypothesis. However, the same statistical results hold if the null hypothesis is instead local causality [51]. A more recent interpretation of a Bell experiment is that the causal structure we expect to be true, based on physical principles like relativistic causality and free will, is not compatible with the correlations we observe. This conclusion indicates that classical causality requires a reformulation in order to explain experiments on quantum mechanical objects, in particular, those that involve entangled systems. This was an important first step, however, a more formal investigation of causality in the quantum world is needed. Wood and Spekkens [51] applied the classical principles of noiseless causal discovery to the correlations predicted by quantum theory for a Bell experiment. Their analysis shows that not only does the intuitive explanation, local causality, fail to explain the correlations but so does every other classical causal explanation. Unfortunately, these results are only applicable to exact distributions whereas in practice we can only sample a distribution finitely many times and thereby will always have some noise. The method must be further developed in order to be applicable to experiments.

Recently, experiments have been designed and conducted which extend the principles of a Bell experiment by using different causal structures as the null hypothesis [38, 53]. This work, by design, does not allow us to ever accept a certain causal explanation. It also does not consider quantum causal explanations. In Chapter 3 we proposed a method of comparing different causal explanations, classical or quantum, on purely statistical grounds. This allows us to perform statistical model selection in order to determine the most appropriate causal explanation for a set of options, in contrast to conducting a series of hypothesis tests with each as the null hypothesis. We apply this causal discovery algorithm on data from a photonic Bell scenario experiment. We do not close to locality loophole so we call the standard causal structure classical common cause, where the common cause is a latent variable. The two alternative classical causal structures we consider: an additional channel between

Alice's setting and Bob's outcome meant to reminiscent of allowing super-luminal causal influences, and the second is a super-deterministic structure where the latent variable can also influences Alice's setting choice (see Figure 4.1). We also consider the alternative where as oppose to allowing more channels we modify the theory of causality to a quantum version. For practical purposes we only consider these structures. Any other causal structure, however, could be included without requiring any modifications to the theory.

We begin by applying the causal discovery algorithm to a Bell scenario in section 4.2. Section 4.3 describes the experimental setup used to generate the data sets and the results of the causal discovery algorithm. Section 4.5 includes technical details on the algorithm and additional data analysis.

4.2 Causal Discovery

A causal explanation, a term which we use synonymously with causal structure, can be succinctly represented by a directed acyclic graph (DAG). The framework of causal models [35] allows us to associate each causal explanation with a set of compatible probability distributions. The data analysis of an experiment often involves estimating the distribution the data was sampled from, or a related quantity. Different causal structure assumptions result in different sets of permissible estimates of this quantity of interest. If multiple causal structures are considered, each results in its own estimator. Each estimator, and thereby each causal explanation, can then be compared by measuring its predictive accuracy. This approach to causal analysis is accomplished with minimal physical assumptions. In particular, no assumptions regarding the space-time location of the measurements, free will of the experimenters or relativistic causality are required. These principles may motivate the choice of a particular set of causal explanations to investigate but they are not assumptions. In this analysis, however, we do assume that individual runs of the experiment are independent. In future work, this assumption could be revisited and potentially relaxed. The causal discovery algorithm is presented in the context of Bell experiments but it could be applied more broadly.

Consider a Bell experiment where two parties, Alice and Bob, perform measurements in their respective laboratories. For each run, they record the measurement settings (s, t)[2] and outcomes (x, y). We consider the case where the outcomes are binary. The quantity we wish to estimate in our data analysis is the probability of each outcome if a particular setting is implemented. This estimate is called a regression function.

[2]Upper case letters denote random variables and the corresponding lower case letters denote their values.

Mathematically, X, Y, S and T are random variables distributed according to a joint probability distribution $p(x, y, s, t)$. We assume that there is no memory of previous outcomes and each run of the experiment is independent. The experimenter doesn't measure an outcomes x and y, but instead each outcome's relative frequency

$$\vec{f} = (f_{x=0,y=0}, f_{x=1,y=0}, f_{x=0,y=1}, f_{x=1,y=1}). \tag{4.1}$$

The relative frequencies can also be viewed as a random variable. The distribution for $\vec{F}, S, T$ can be calculated from $p(x, y, s, t)$ and we denote it $P(\vec{f}, s, t)$. The data consists of observations of independent identically distributed random variables

$$(\vec{F}^{(1)}, S^{(1)}, T^{(1)}), (\vec{F}^{(2)}, S^{(2)}, T^{(2)}), \cdots, (\vec{F}^{(m)}, S^{(m)}, T^{(m)}) \sim P(\vec{f}, s, t). \tag{4.2}$$

The experimenter, knowing only the data and unaware of the data collection process, tries to infer the causal influences between the variables X, Y, S and T. The situation has been considered when the distribution from which the data is being sampled $p(x, y, s, t)$ is known [51], however, this distribution cannot be calculated from the data. We will use the novel causal discovery algorithm outlined in Chapter 3.

This algorithm selects the causal structure which best predicts future outcomes (x, y) given a measurement setting (s, t). Mathematically, the quantity of interest is a function $\vec{r}(st)$ that returns a prediction for the relative frequencies $\vec{f}$ of the outcomes for each measurement setting. The component predicting the frequency of the outcome x, y is denoted $r_{xy}(st)$.

Two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ with m data points each, as specified in Eq. (4.2), are collected. We do not close to locality loophole so we call the standard causal structure classical common cause (CCC), where the common cause is a latent variable. The two alternative classical causal structures we consider: an additional channel between Alice's setting and Bob's outcome meant to reminiscent of allowing super-luminal causal influences (SY causal), and the second which we refer to as ΛS causal, is a super-deterministic structure where the latent variable can also influences Alice's setting choice (see Figure 4.1). We also consider the standard quantum causal explanation where the correlations are the result of performing measurements on a shared two-qubit system. This quantum causal structure is the quantum analogue of a common cause [2, 4]. Independent of the formalism of quantum causality chosen, our analysis only requires a specification of the set of regression functions compatible with the quantum causal structure. A set of compatible regression function can be considered a causal assumption instead of a DAG.

Every DAG for a set of events $X_1, X_2, \cdots, X_n$ has an associated set of probability

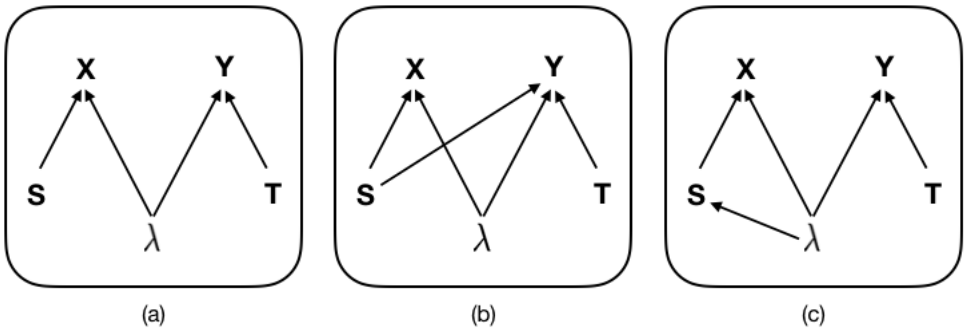

Figure 4.1: Examples of directed acyclic graphs for a Bell experiment. (a) A common causal structure. (b) A causal structure with an additional channel between between a setting and outcome. (c) A causal structure with the latent variable influencing the setting choice.

distributions that are compatible with it. This set is as follows:

$$J_{_G} = \{P(x_1, \ldots, x_n) \mid \exists\ P_1, \ldots, P_n\ s.t.\ P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_i(x_i|\mathrm{Pa}(X_i))\}, \tag{4.3}$$

where G is the DAG, $\mathrm{Pa}(X_i)$ are the parents of the vertex X_i in the graph G, and $P_1, ..., P_n$ are all valid conditional probability distributions. The set of allowed regression functions for the causal structure $M_{_G}$ is any conditional probability distribution $p(x, y|s, t)$ which results from a joint probability distribution in the set $J_{_G}$.

All of the structures considered in our analysis contain a latent variable. The cardinality of this variable must be defined in order to specify the set of compatible regression functions. We denote the DAG for the classical common causal structure with a latent variable of cardinality K as CCK. Similarly the DAGs for the ΛS causal and SY causal structures with cardinality K are denoted ΛSK and SYK respectively. We only consider the quantum causal model where the latent variable is a two-qubit state so the cardinality is always two and does not need to be specified explicitly. The quantum common cause DAG is denoted QCC.

Explicitly, the compatible joint distributions for the causal structures are given by

$$J_{_{CCK}} = \Big\{ P(x,y,s,t) = \sum_{\lambda=1}^{k} P_1(x|s,\lambda) P_2(y|t,\lambda) P_3(s) P_4(t) P_5(\lambda) \Big\}, \tag{4.4}$$

$$J_{_{SYK}} = \Big\{ P(x,y,s,t) = \sum_{\lambda=1}^{k} P_1(x|s,t,\lambda) P_2(y|t,\lambda) P_3(s) P_4(t) P_5(\lambda) \Big\}, \tag{4.5}$$

$$J_{_{\Lambda SK}} = \Big\{ P(x,y,s,t) = \sum_{\lambda=1}^{k} P_1(x|s,\lambda) P_2(y|t,\lambda) P_3(s|\lambda) P_4(t) P_5(\lambda) \Big\}, \tag{4.6}$$

$$J_{_{QCC}} = \Big\{ P(x,y,s,t) = \mathrm{Tr}\big[E_x^s E_y^t \ \rho\big] P_1(s) P_2(t) \Big\}, \tag{4.7}$$

where $P_1, ..., P_n$ are all valid probability distributions, ρ is a two-qubit quantum state and $\{E_x^s\}_{x=0,1}, \{E_y^t\}_{y=0,1}$ are qubit POVMs. The valid regression functions from which each estimator will be chosen are correspondingly

$$M_{_{CCK}} = \Big\{ \vec{r}(st) \ \Big| \ r_{xy}(s,t) = \sum_{\lambda=1}^{k} P_1(x|s,\lambda) P_2(y|t,\lambda) P_3(\lambda) \Big\}, \tag{4.8}$$

$$M_{_{SYK}} = \Big\{ \vec{r}(st) \ \Big| \ r_{xy}(s,t) = \sum_{\lambda=1}^{k} P_1(x|s,t,\lambda) P_2(y|t,\lambda) P_3(\lambda) \Big\}, \tag{4.9}$$

$$M_{_{\Lambda SK}} = \Big\{ \vec{r}(st) \ \Big| \ r_{xy}(s,t) = \sum_{\lambda=1}^{k} P_1(x|s,\lambda) P_2(y|t,\lambda) P_3(s|\lambda) P_4(\lambda) \Big(\sum_{\lambda'=1}^{k} P_3(s|\lambda') P_4(\lambda') \Big)^{-1} \Big\}, \tag{4.10}$$

$$M_{_{QCC}} = \Big\{ \vec{r}(st) \ \Big| \ r_{xy}(s,t) = \mathrm{Tr}\big[E_x^s E_y^t \ \rho\big] \Big\}. \tag{4.11}$$

For the moment, consider finding the "best" regression function within one model. Quality of fit is measured by a particular loss function. We use squared error[3](SE) as it is the most prevalent choice. The minimal loss is called the **training loss** and is given by

$$\mathrm{err}_{_G} = \min_{\vec{r} \in M_G} \sum_{i=1}^{m} \big| \vec{r}(s_1^{(i)}, t_1^{(i)}) - \vec{f}_1^{(i)} \big|^2, \tag{4.12}$$

[3]In the supplementary material we analyze the data using multiple different loss functions and the results of the causal discovery algorithm remain the same.

where the sum is taken over the first data set $\mathcal{D}_1$. The regression function which minimizes this loss is the chosen regression function for that DAG and is labelled $\hat{\vec{r}}_{G}$.

Returning to the question of comparing different models, naively one could always pick the model with the lowest training loss. However, this would result in many mistakes. For example, consider the case of nested models. We say a model A is nested in model B if every structure compatible with A is also compatible with B. The minimal training loss criterion would always select the most complicated model. This is because the regression function which minimizes a simpler causal structure will also be compatible with the more complicated version it is nested in. This conclusion is independent of the experimental data, which is clearly flawed. We will use the concept of expected testing loss to evaluate different models. This method finds the model which, if used, would best predict the future sampling from the experiment. The *generalization error* is the average quality of its future predictions. We should of course use the same loss function as we did before, the squared error loss. The generalization error for a model is given by

$$\mathbb{E}\left[\left|\hat{\vec{r}}(s,t) - \big(P(0,0|s,t), P(0,1|s,t), P(1,0|s,t), P(1,1|s,t)\big)\right|^2\right], \tag{4.13}$$

where $\hat{\vec{r}}(s,t)$ is the regression function that minimized the training loss and $p(x,y|s,t)$ is the true distribution. In Eq. (4.13) the expectation is taken with respect to the true underlying distribution, which if we already knew our work would be done. We estimate the generalization error by calculating the average error on the second data set $\mathcal{D}_2$. This quantity is called the test error

$$\overline{\mathrm{Err}}_{G} = \sum_{i=1}^{m} \left|\hat{\vec{r}}_{G}(s_2^{(i)}, t_2^{(i)}) - \vec{f}_2^{(i)}\right|^2, \tag{4.14}$$

where the sum is taken over the second data set.

The test error is calculated for each of the causal structures then the one with the lowest test error is chosen. In the analysis, we wish to compare causal explanations while remaining agnostic about any latent variable's cardinality. To accomplish this, test error or "score" assigned to the *cardinality-agnostic* causal structure is the minimal test error over all possible cardinalities. The following section outlines two experiments upon which this analysis was applied.

4.3 Results

This section describes a quantum optical Bell experiment and the results of the causal discovery algorithm applied to its data.

The state shared between Alice and Bob should be one which is known to imply non-simulability by classical causal theories. We choose the maximally entangled two-qubit state

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B). \tag{4.15}$$

To prepare this state, we encode the qubits in the polarization degree of freedom of photons. In this encoding, $|0\rangle$ and $|1\rangle$ are horizontally and vertically polarized photons respectively.

The polarization-entangled pairs of photons are produced using type-II spontaneous parametric down conversion. A 404 nm continuous wave laser pumps a periodically-poled potassium titanyl phosphate (PPKTP) crystal which produces degenerate 808 nm polarization entangled photons pairs. The PPKTP crystal is placed inside a Sagnac interferometer which results in the output state being entangled [45, 16, 50].

The polarization of each photon is manipulated using quarter- and half-wave plates. After the maximally entangled photon pairs have been produced, the polarization can be locally manipulated to create the desired Bell state.

Once the Bell state has been created each photon path is coupled into a single-mode fiber. The remaining operations in the experiment are agnostic to all degrees of freedoms except polarization, namely the spectral, temporal and spatial state. This allows these degrees of freedom to be ignored. A fibre directs the first photon to Alice and another directs the second photon to Bob. There, each photon encounters a measurement apparatus consisting of a quarter-wave plate, half-wave plate and a polarizing beamsplitter with each port coupled to a multi-mode fibre[4]. The four multimode fibres are connected to single-photon avalanche diodes. This allows each party to perform an arbitrary two-outcome projective measurement.

We label a measurement at the transmitted port of the PBS as a 0 outcome and the reflected port as a 1. For example, the number of counts registered at Alice's transmitted port is labelled N_0.

A second Bell experiment is conducted using a dephased version of the Bell state in Eq. (4.15). This state is not entangled and consequently we expect a classical common

[4]A multimode fibre is used to limit the dependency of the coupling efficiency on the angle of the wave plate resulting from deflections.

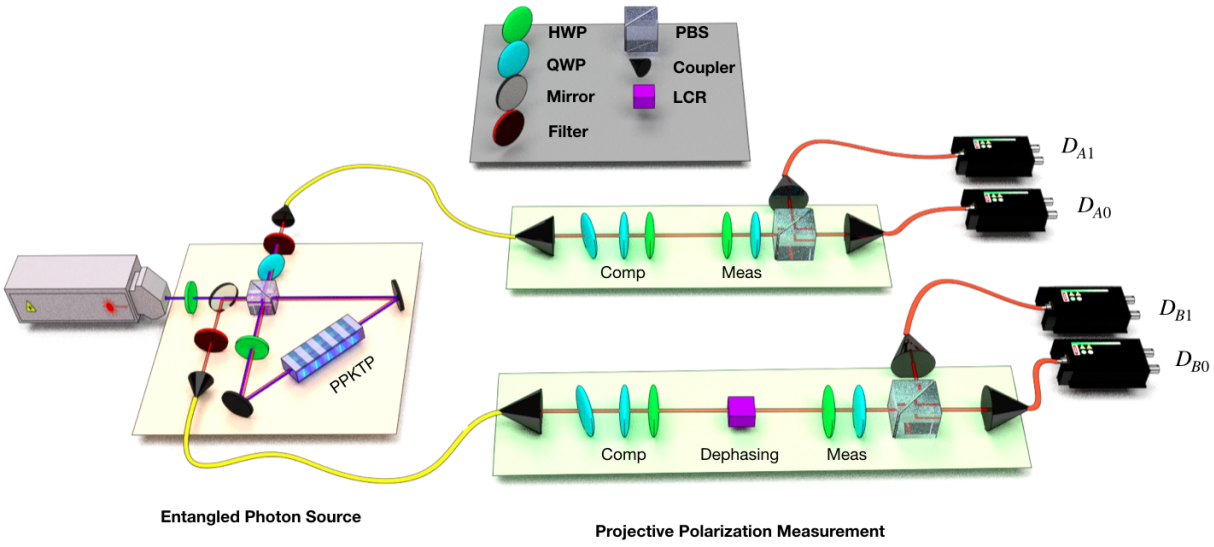

Figure 4.2: Experimental diagram. Maximally polarization entangled photons pairs are created through parametric down conversion in both paths of a Sagnac interferometer. One photon is sent to a polarization measurement, and the other photon first passes through a depolarizing channel comprised of two LCRs before also having its polarization measured. Coincidences between a photon being measured on both sides of the experiment are recorded. PPKTP, periodically-poled potassium titanyl phosphate; PBS, polarizing beamsplitter; LCR, liquid crystal retarder; HWP, half-wave plate; QWP, quarter-wave plate

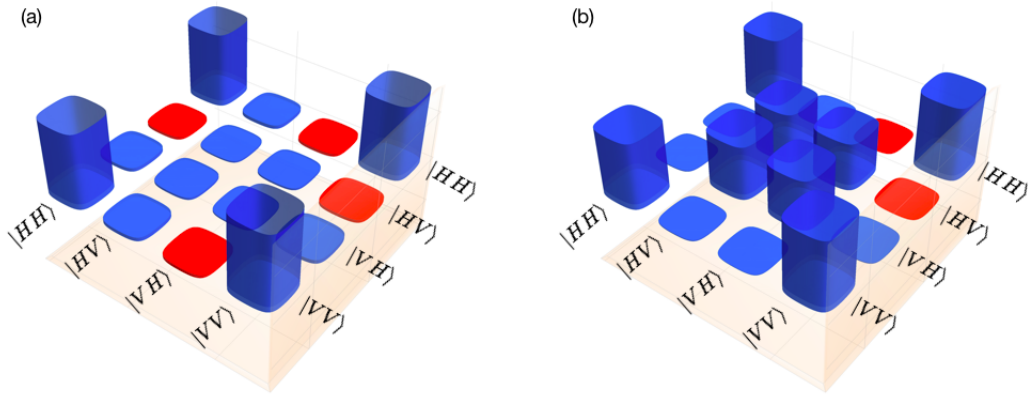

Figure 4.3: The density matrix of the maximum likelihood estimate of the state of the source for (a) the entangled experiment without a dephasing channel and (b) the experiment with the dephasing channel. Blue represents a positive number while red represents a negative number.

causal explanation to be preferred. We use liquid crystal retarders to implement a dephasing channel on one of the qubits to remove the entanglement in the state. The maximally dephasing channel consists of a σ_X gate being applied half the time, ideally creating the mixed state

$$\rho_{DP} = 0.5 \left|\Phi^+\right\rangle\!\left\langle\Phi^+\right| + 0.5 \left|\Psi^+\right\rangle\!\left\langle\Psi^+\right|, \tag{4.16}$$

where $|\Phi^+\rangle$ and $|\Psi^+\rangle$ are the standard Bell states. The maximally dephased state has no entanglement.

Typically in Bell experiments, there are two possible measurement settings for Alice and Bob. This is the minimal amount needed to violate a CHSH inequality and thereby reject local causality. We choose to instead implement six different measurement settings for Alice and Bob , to allow more refined inferences to be made. The variables s, t each range from 1 to 6 giving 36 unique combinations. The wave plate angles corresponding to each measurement setting are the same for Alice and Bob. The wave plate angles were chosen by ensuring that they implement projective operators which are uniformly spaced on the qubit Bloch sphere.

Two data sets, $\mathcal{D}_1, \mathcal{D}_2$, are collected. Each setting combination (s, t) for Alice and Bob is implemented once per data set. The coincidence counts $N_{00}, N_{10}, N_{01}, N_{11}$ are measured for 10 seconds on each setting with a 3 ns coincidence window. The relative frequencies of

Scores for the Causal Structures

	Dephased Experiment	Entangled Experiment
Classical Common Cause	$(2.0 \pm 0.2) \times 10^{-3}$	$(80 \pm 1) \times 10^{-3}$
SY Causal	$(3.5 \pm 0.3) \times 10^{-3}$	$(2.8 \pm 0.3) \times 10^{-3}$
ΛS Causal	$(2.3 \pm 0.2) \times 10^{-3}$	$(2.5 \pm 0.4) \times 10^{-3}$
Quantum Common Cause	$(2.1 \pm 0.2) \times 10^{-3}$	$(1.6 \pm 0.4) \times 10^{-3}$

Table 4.1: The test error $\overline{\text{Err}}$ for each of the considered causal structure. The first column is for the experimental data when the entangled state is maximally depolarized. The second column is for the experimental data when the photons are left entangled.

the form

$$f_{00} = \frac{N_{00}}{N_{00} + N_{10} + N_{01} + N_{11}}, \tag{4.17}$$

are recorded. Explicitly, the data sets used for the causal discovery algorithm are

$$\mathcal{D}_1 = \left\{(f^{st}_{00}, f^{st}_{10}, f^{st}_{01}, f^{st}_{11}, s, t) \mid s, t \in \{1, 2, 3, 4, 5, 6\}\right\}, \tag{4.18}$$

$$\mathcal{D}_2 = \left\{(g^{st}_{00}, g^{st}_{10}, g^{st}_{01}, g^{st}_{11}, s, t) \mid s, t \in \{1, 2, 3, 4, 5, 6\}\right\}, \tag{4.19}$$

where f_{xy}, g_{xy} denote the relative frequency of the outcome (x, y).

The same data used for the causal inference is used to calculate maximum likelihood estimates of the state. The entangled experiment produces a state with $97.9 \pm 0.07\%$ fidelity with the maximally entangled target state in Eq. (4.15). The state for the dephased experiment has $98.3 \pm 0.07\%$ fidelity with the target state in Eq. (4.16).

The regression function for each causal structure $\hat{\vec{r}}_G$ is calculated by minimizing the loss on the first data set. A Nelder-Mead optimization algorithm seeded 20 times is used for the minimization problem. The test error of the regression function on the second data set is calculated for each. The results are displayed in Table 4.1. Recall that the test error for the classical common cause, SY causal and ΛS causal structures is the minimum error over all possible cardinalities of the latent variable. For practical purposes search starting with the lowest cardinality and increase it until the test error begins to become worse. The data shows that the experiment with no entanglement in the distributed state prefers a classical

Training Error for the Causal Structures

	Dephased Experiment	Entangled Experiment
Classical Common Cause	$(1.5 \pm 0.2) \times 10^{-3}$	$(70 \pm 1) \times 10^{-3}$
SY Causal	$(0.2 \pm 0.1) \times 10^{-3}$	$(0.3 \pm 0.1) \times 10^{-3}$
ΛS Causal	$(1.1 \pm 0.2) \times 10^{-3}$	$(0.4 \pm 0.2) \times 10^{-3}$
Quantum Common Cause	$(1.5 \pm 0.2) \times 10^{-3}$	$(1.0 \pm 0.2) \times 10^{-3}$

Table 4.2: The training error, err, for each of the considered causal structure. The first column is for the experimental data when the entangled state is maximally depolarized. The second column is for the experimental data when the photons are left entangled.

common cause explanation while the entangled experiment prefers a quantum common cause explanation. The quantum common cause explanation performs better than the two classical alternatives to common cause.

Additional insight into why the SY causal and ΛS causal models are unfavourable can be gleamed by studying the training error (Table 4.2). For the Bell experiment with the entangled state, the training error of the SY causal and ΛS causal structures are both lower then the quantum causal structure's. This indicates that these explanations over-fit the data. Intuitively, this means that they not only are able to account for the correlations observed in a Bell experiment, but also more exotic correlations that are not present. This additional flexibility results in them fitting to statistical noise in the first data set, thereby harming their predictive accuracy on any future data set.

4.4 Discussion

Previous experiments have investigated potential classical causal descriptions of Bell scenarios from the perspective of hypothesis testing [38, 29]. This has resulted in the rejection of some classical causal explanations. However, SY causal and ΛS causal explanations could not be ruled out since they are able to explain the observed correlations. The concern with these theories is that they would allow additional correlations which the experiments do not seem to possess and hence seems to over-fit the data.

We perform a model selection analysis by extending the concepts used for causal discovery with idealized infinite run data [51] to realistic experiments where there is statistical noise. This new causal discovery algorithm was then applied on data produced by a photonic Bell experiment with a highly entangled shared resource. A quantum causal explanation was preferred over all considered classical causal structures including a SY causal and ΛS causal option. This was possible since the algorithm is able to identify theories that are overly complex for the data. This is a task that hypothesis tests fail to accomplish. In order to provide additional validation of the algorithm, a second Bell experiment was performed. This time, the shared resource was first dephased which removed all entanglement. The data from this second experiment preferred a classical common cause explanation and was slightly over-fitted by the quantum common cause structure. The result confirmed our postulate since a quantum mechanical explanation would allow for additional correlations which should not be observed in this experiment.

The experiment was conducted as a proof of principle for the causal discovery algorithm therefore many loopholes were left open. Future work should be done applying the algorithm to data generated by loophole free experiments. Further work is also required, theoretically and experimentally, to develop and test more classical and quantum causal structures.

4.5 Supplementary Material

4.5.1 Causal Structure Cardinality

The causal structures considered in our experiment all contain a latent variable. This does not pose a problem except requiring that the latent variable's cardinality must be specified. To circumvent this, the main results consider the test error *cardinality-agnostic* causal structure. This is defined as the minimum test error over all possible cardinalities of the latent variable. While these scores are sufficient for the main result, the scores for each individual cardinality are also interesting. This is in part due to the fact that two structures which differ only by the cardinality of their latent variable are nested. For example, $M_{CC3} \subset M_{CCK}$ for every $K > 3$. Any distribution, and thereby regression function, compatible with $CC3$ will also be compatible with the same causal structure with a latent variable which is allowed to take on more values. This allows us to further explore the concepts of relative over- and under-fitting.

We begin by exploring the classical common cause structures. Figure 4.4 plots the scores for the first eight cardinalities. Two interesting observations can be made. Firstly,

the inset on Figure 4.4(b) shows that for cardinalities greater than four, the training error continues to decrease but the test error begins to rise. This is an indication that the regression function for cardinalities greater then four are fitting to noise in the experiment. The training error for CC5 is smaller than for CC4 while the testing error is larger. In this situation we say the CC5 structure over-fits the data relative to the CC4. Secondly, for the entangled experiment (Figure 4.4(a)) the test error continues to decrease but remains significantly larger than that of the dephased experiment. This hints that the model is under-fitting however the test and training errors must be compared to another causal structure, say quantum common cause, to determine if this is due to over- or under-fitting. The training and test error for a quantum common cause structure, as seen in Tables 4.1 and 4.2, are both substantially lower indicating that the classical common cause description for all cardinalities relatively under-fits the data from the entangled experiment. We only explore up to CC15 due to long computational times, however, the training error stops decreasing after CC10 for the entangled experiment. The test error is already increasing after CC4 due to over-fitting in the dephased experiment so allowing a larger cardinality would not improve this situation.

The ΛS causal structures behave similarly to the classical common causal for the dephased experiment. Figure 4.5 plots the ΛS causal structures' scores. In contrast to the classical common causal case, the ΛS causal structures' scores behave similarly for both experiments. This indicates that the ΛS causal structures with $K \geq 7$ over-fit the entangled data relative to the ΛS6 structure. However, an additional statement can be made in the ΛS case. For $K < 7$ the ΛS structure under-fits relative to QCC since it trains and tests worse. For $K \geq 7$, the ΛS structure relatively over-fits the quantum causal structure. At no point it is the optimal causal structure.

Finally, the SY causal structures' are equivalent for cardinalities greater then $k = 2$. The training error for $SY2$, $SY3$, etc. are all the same since they are compatible with the same set of probability distributions and therefore will always have the same score, as seen in Table 4.3. For the quantum common cause structure we do not consider multiple "cardinalities". In future experiments more options could be considered for the quantum common cause structure. We showed every possible cardinality of the classical causal structures was outperformed by one choice of latent system dimension of the quantum common cause structure for the entangled experimental data. Larger cardinalities may further improve the score of QCC, but, this would not change the conclusions drawn in this thesis.

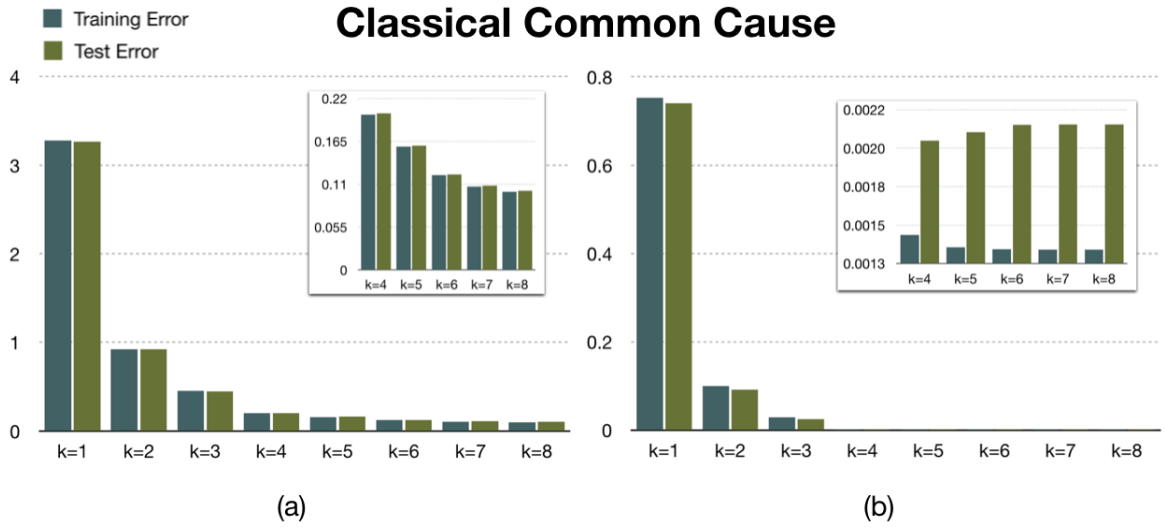

Figure 4.4: The test and training error for various cardinalities K of the latent variable in a classical common cause structure. The data is from the experiments with (a) an entangled shared resource (b) a dephased shared resource.

SY causal Causal Structure Scores

Causal Structures		Dephased Experiment	Entangled Experiment
SY1	err	0.751672	3.27512
	$\overline{\text{Err}}$	0.741152	3.27512
SY2	err	0.000195293	0.000302155
	$\overline{\text{Err}}$	0.00350377	0.00275417
SY3	err	0.000195293	0.000302155
	$\overline{\text{Err}}$	0.00350377	0.00275417
SY4	err	0.000195293	0.000302155
	$\overline{\text{Err}}$	0.00350377	0.00275417

Table 4.3: The test and training errors for SY causal structures with a latent variable of cardinality one through four.

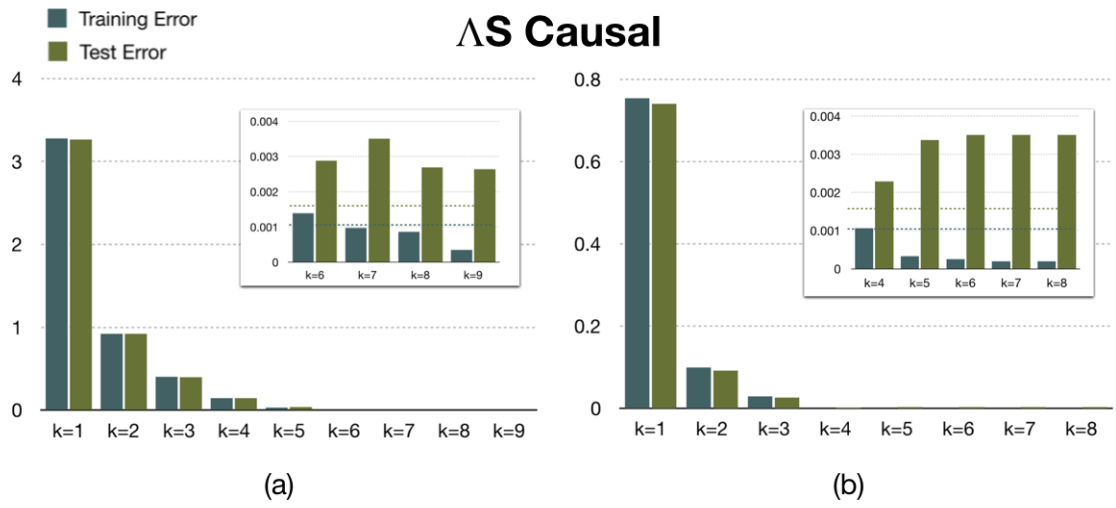

Figure 4.5: The test and training error for various cardinalities k of the latent variable in a ΛS causal structure. The data is from the experiments with (a) an entangled shared resource (b) a dephased shared resource. The dashed lines on the insets indicate the training and test error of the quantum common causal structure. The higher line is the test error and the lower is the test error.

4.5.2 Alternative Loss Functions

This section explores the results of the causal discovery algorithm with loss functions other than squared error. Squared error is the most common choice, but, it is still interesting to investigate the conclusion if other loss functions are chosen instead. We investigate three alternative loss functions: weighted squared error (WSE), approximate weighted squared error (aWSE) and the negative likelihood. Respectively these functions are

$$L(\vec{r},(\vec{f},s,t)) = \sum_{xy} \frac{(r_{xy}(st) - f_{xy}^{st})^2}{r_{xy}(st)(1-r_{xy}(st))}, \tag{4.20}$$

$$L(\vec{r},(\vec{f},s,t)) = \sum_{xy} (r_{xy}(st) - f_{xy}^{st})^2 \left(\frac{N_{xy}N_T}{(N_{xy}+N_T)^3}\right)^{-1}, \tag{4.21}$$

$$L(\vec{r},(\vec{f},s,t)) = \sum_{xy} -M \log\left[\frac{(r_{xy}(st) - f_{xy}^{st})^2}{r_{xy}(st)(1-r_{xy}(st))}\right] - \log\left[r_{xy}(st)(1-r_{xy}(st))\right], \tag{4.22}$$

where N_{xy} is the number of counts for the outcome xy, $N_T = \sum_{xy} N_{xy}$, $M = 144$ and f_{xy}^{st} is the frequency of the outcome xy. Note that we must assume the counts are Poissonian distributed in order in order to calculate these loss functions. This assumption was not needed for squared error loss.

The WSE and aWSE are simply the squared error divided by the variance of the frequency data f_{xy}^{st}. For the WSE, we assume the regression function accurately predicts the mean of the conditional probability distribution and use this along with the Poissonian assumption to calculate the variance. Alternatively, for aWSE the variance is calculated by propagating the sample mean estimated errors on the counts N_{xy}. The aWSE loss function was previously used in papers [32, 33] that dealt with similar data to our experiments. The negative log-likelihood function takes substantially more effort to calculate. The derivation of the log-likelihood function can be found in Section 4.5.3.

The test error when each of the loss functions are used, for the training error when finding the regression functions and test error, are presented in Tables 4.4 and 4.5. The same causal structures are preferred independent of the loss function used. This is a comforting result.

4.5.3 Likelihood Function

No assumptions were required regarding the statistical noise in the data in order to calculate the least squared loss function. However, one popular alternative loss function, the

Scores for the Entangled Experiment

Causal Structure	SE	aWSE	WSE	Log-Lik
CC1	3.26659	17.4234	$2.40667 * 10^6$	$3.55298 * 10^6$
CC2	0.924807	5.58679	419440	467941
CC3	0.444091	2.99268	232003	259018
CC4	0.200949	1.2423	86237.3	88032.1
CC5	0.163848	1.04827	69363.3	70969.2
SY1	3.26803	17.4276	$1.95627 * 10^6$	$3.55329 * 10^6$
SY2	0.00275417	0.0168795	887.004	882.854
ΛS1	3.26659	17.4234	$2.40667 * 10^6$	170.472
ΛS2	0.925138	5.58894	419598	-37.4337
ΛS3	0.397335	2.71821	198207	-120.814
ΛS4	0.147982	0.943622	63223.4	-221.593
ΛS5	0.0368944	0.278969	16660.5	-285.512
QCC	**0.00164609**	**0.0117338**	**586.114**	**-932.328**

Table 4.4: The test errors $\overline{\text{Err}}$ for a selection of causal structures of the entangled experiment. Each column corresponds to a different loss function being used for the training and test error.

Scores for the Dephased Experiment

Causal Structure	SE	aWSE	WSE	Log-Lik
CC1	0.739839	3.9506	320741	-43.343
CC2	0.0918911	0.498272	23767	-349.571
CC3	0.0255474	0.138866	6497.2	-534.19
CC4	**0.00201981**	**0.0109851**	**514.139**	**-899.752**
CC5	0.00206802	0.0112432	527.72	-897.723
SY1	0.741152	3.95782	310194	-43.0681
SY2	0.00350377	0.019127	890.806	-43.0681
ΛS1	0.739839	3.9506	320741	-43.343
ΛS2	0.0918382	0.498002	23752.5	-349.651
ΛS3	0.0257356	0.139901	6541.83	-533.439
ΛS4	0.0022927	0.0125964	583.189	-857.935
ΛS5	0.00337661	0.0184894	857.56	-862.666
QCC	0.00206596	0.4491	527.352	-364.3

Table 4.5: The test errors $\overline{\text{Err}}$ for a selection of causal structures of the dephased experiment. Each column corresponds to a different loss function being used for the training and test error.

negative of the log=likelihood, does require such assumptions. This section outlines these assumptions and calculates the log-likelihood function.

We need to calculate the likelihood of the experiment returning a certain frequency f_{xy}^{st} for a given parameter choice. In our experiment the parameter choice is a conditional probability distribution $p(x, y|s, t)$.

We assume that rate of photon pair generation is constant in time and each so the errors on the number of counts during a given time window are Poissonian. This assumption was also verified by taking a series of measurements and comparing it visually to a Poissonian distribution. There are four detectors, two for Alice and Bob each. We call these detectors Alice transmitted, Alice reflected, Bob transmitted, Bob reflected and each will have a different efficiency[5] we label these $\eta_T^A, \eta_R^A, \eta_T^B, \eta_R^B$ respectively. Each measurement is performed by counting for 10 seconds. The mean number of photon pairs created by the source in 10 seconds is denoted R. For a given measurement setting (s, t), the likelihood of the coincidence counts being equal to N_{xy} is

$$\mathcal{L}[N_{xy}^{\alpha\beta}] = \Pr\Big[\text{Poi}[R\ \eta_\alpha^A \eta_\beta^B\ p(x, y|s, t)]\Big]\ =\ N_{xy}^{\alpha\beta}, \tag{4.23}$$

where $\text{PrPoi}[a] = b$ is the probability a random variable distributed according to $\text{Poi}[a]$ is measured to be b and $\alpha, \beta = T, R$ to denote which detectors were associated with the outcome (x, y).

In the limit where R is large then a Poissonian distribution is approximately normal.

$$\text{Poi}[\lambda] \approx \text{Normal}[\mu = \lambda, \sigma^2 = \lambda]. \tag{4.24}$$

This assumption is valid in our experiment since $R\eta_\alpha^A \eta_\beta^B \approx 1000$.

Our experiment implements a "rotation" of the detectors. Every measurement is performed four times. Each time the detector associated with each POVM element is changed. The counts from these four measurements are then summed and taken to be the counts for that measurement setting. Mathematically, the frequency of the outcome (x, y) of the averaged measurement is given by

$$F_{xy} = \frac{\sum_{\alpha,\beta} N_{xy}^{\alpha\beta}}{\sum_{\alpha,\beta,i,j} N_{ij}^{\alpha\beta}}. \tag{4.25}$$

[5]The efficiency includes the efficiency of the detector in addition to the coupling of the light into the fibre and any other sources loss that act linearly in photon number.

Let's now find the likelihood of the experimental data returning f_{xy}^{st}. Recall that each count $N_{xy}^{\alpha\beta}$ is a normal random variable so F_{xy} is a function of random variables. Using the Taylor approximation outlined in Section 7.3 of [31], this function of normally distributed random variables is also a normally distributed random variable with the following mean and variance

$$\mathbb{E}[F_{xy}] = p_{xy} \tag{4.26}$$

$$\mathbb{V}[F_{xy}] = \frac{p_{xy}(1-p_{xy})}{\tilde{\lambda}} \tag{4.27}$$

where $\tilde{R} = R\sum_{\alpha\beta}(\eta_\alpha^A \eta_\beta^B)$. Note that we could instead have assumed that these "frequency" random variables are distributed according to this normal distribution as oppose to assuming the distribution of the counts. In this case, the above discussion is viewed as supporting this assumption. The assumption of normality could be independently verified by taking measurements and looking at the shape of the distribution however we still would have to assume the dependence of our measurements and these probabilities. Either way the distribution for the frequency data points is

$$F_{xy} \sim \text{Normal}\left[p_{xy}, \frac{p_{xy}(1-p_{xy})}{\tilde{\lambda}}\right]. \tag{4.28}$$

The likelihood function of the parameters $\vec{p}, \tilde{R}$ given a data set $\mathcal{D} = \{\vec{f}^{(i)}, s^{(i)}, t^{(i)}\}_{i=1}^m$ is

$$L[\vec{p}, \tilde{R}|\vec{f}] = \prod_{i,x,y} \sqrt{\frac{\tilde{R}}{2\pi p_{xyst}^{(i)}(1-p_{xyst}^{(i)})}} \exp\left\{\frac{-\tilde{R}(p_{xyst}^{(i)} - f_{xy}^{(i)})^2}{2p_{xyst}^{(i)}(1-p_{xyst}^{(i)})}\right\}, \tag{4.29}$$

where $p_{xyst}^{(i)}$ is shorthand for $p(x,y|s^{(i)},t^{(i)})$ and $\vec{p}$ represents $p(x,y|s,t)$. Finally, the log-likelihood is given by

$$\ell[\vec{p}, \tilde{R}|\vec{f}] = -\sum_{s,t,x,y} \log\left[\frac{p_{xyst}^{(i)}(1-p_{xyst}^{(i)})}{\tilde{R}}\right] - \sum_{s,t,x,y} \frac{\tilde{R}(p_{xyst}^{(i)} - f_{xy}^{(i)})^2}{p_{xyst}^{(i)}(1-p_{xyst}^{(i)})}, \tag{4.30}$$

up to some constants which don't effect the minimization. To isolate $\vec{p}$ first find the MLE of $\tilde{R}$ and substitute it into Eq. (4.30). By taking the derivative with respect to $\tilde{R}$ and setting it to zero we can solve for $\tilde{R}$ in terms of $\vec{f}, \vec{p}$

$$\hat{\tilde{R}} = N\left(\sum_{s,t,x,y} \frac{(p_{xyst}^{(i)} - f_{xy}^{(i)})^2}{p_{xyst}^{(i)}(1-p_{xyst}^{(i)})}\right)^{-1}, \tag{4.31}$$

where N is the number of Alice's measurement settings, times the number of Bob's times the number of outcomes, which is 4. Substituting Eq. (4.31) into Eq. (4.30) we get

$$\ell[\vec{p}|\vec{f}] = -\sum_k \log\left[p_k(1-p_k)\sum_l \frac{(p_l - f_l)^2}{p_l(1-p_l)}\right] - N. \tag{4.32}$$

After some manipulation and dropping constants Eq. (4.32) becomes

$$\ell[\vec{p}|\vec{f}] = -\sum_k \log[p_k(1-p_k)] - N\log\left[\sum_k \frac{(p_k - f_k)^2}{p_k(1-p_k)}\right]. \tag{4.33}$$

This is the negative log-likelihood loss. Minimizing Eq. (4.33) with respect to $\vec{p}$ will gives the MLE which is also the regression function which minimizing the training error.

Chapter 5

Conclusion

Previous work investigating causality in Bell scenarios has focused on conducting hypothesis tests for various causal assumptions. The conclusion of these experiments is either a rejection of the causal hypothesis or no statement at all. In contrast, we simultaneously consider multiple causal structures and are able to provide a positive statement regarding the preferred causal explanation. Our results indicate that if the shared state has a sufficient amount of entanglement, a quantum common cause explanation is preferred. In the case where the state does not have entanglement, a classical common cause explanation is preferred. The use of an independent test data set to calculate the predictive accuracy of a causal structure allows the identification of unnecessarily complicated structures. The classical causal alternatives of adding an additional channel from Alice's setting to Bob's outcome and allowing the Alice's setting choice to be determined by the environment are examples of unnecessarily complicated structures for the data we observed.

In order to perform the desired analysis a new method of causal discovery was developed which could consider quantum and classical causal structures within the same framework. Existing methods of causal discovery for quantum mechanical experiments have relied on interventions. This potentially problematic reliance was removed.

This work represents still only an early step towards our understanding of causality's role in quantum mechanics and nature. Our causal discovery algorithm could be applied, without modifications, to much more complicated experimental scenarios. In addition, different quantum causal explanations could also be considered. An unintended feature of the algorithm is that it can also be applied to causal theories that have recently been developed in the framework of generalized probabilistic theories. A challenge in applying this algorithm broadly moving forward is that the computation time may become extraor-

dinarily long. Finding faster optimization algorithms to minimize the training error would be of great importance.

Another area of interest would be in investigating the relationship of our algorithm with the field of classical causal discovery. The causal discovery algorithm we developed when considering only classical causal structures is different from existing methods within this field. Ensuring the algorithm is compatible with quantum causal structures required us to use nontraditional methods which may offer unique advantages.

Finally, the concept of transferring causal assumptions to constraints on regression functions could be used for other open questions within quantum foundations. Any theory or assumption that can be tested must place some constraints on the observational statistics of an experiment. This is by no means a novel concept, but, using statistical learning to perform model selection rather then conducting hypothesis tests is seldom seen in quantum foundations and this analysis technique has substantial potential.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[2] John-Mark A. Allen, Jonathan Barrett, Dominic C. Horsman, Ciarán M. Lee, and Robert W. Spekkens. Quantum common causes and quantum causal models. *Phys. Rev. X*, 7:031021, Jul 2017.

[3] Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental realization of einstein-podolsky-rosen-bohm gedankenexperiment: A new violation of bell's inequalities. *Phys. Rev. Lett.*, 49:91–94, Jul 1982.

[4] Jonathan Barrett, Robin Lorenz, and Ognyan Oreshkov. Quantum Causal Models. *arXiv e-prints*, page arXiv:1906.10726, Jun 2019.

[5] J. S. Bell. On the einstein podolsky rosen paradox. *Physics Physique Fizika*, 1:195–200, Nov 1964.

[6] Charles H. Bennett, Herbert J. Bernstein, Sandu Popescu, and Benjamin Schumacher. Concentrating partial entanglement by local operations. *Phys. Rev. A*, 53:2046–2052, Apr 1996.

[7] Robin Blume-Kohout. Robust error bars for quantum tomography. *arXiv e-prints*, page arXiv:1202.5270, Feb 2012.

[8] Robert W. Boyd. *Nonlinear Optics, Third Edition*. Academic Press, Inc., Orlando, FL, USA, 3rd edition, 2008.

[9] Juan M Bueno. Polarimetry using liquid-crystal variable retarders: theory and calibration. *Journal of Optics A: Pure and Applied Optics*, 2(3):216–222, may 2000.

[10] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[11] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, March 2003.

[12] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed Experiment to Test Local Hidden-Variable Theories. , 23(15):880–884, Oct 1969.

[13] Jay Damask. Polarization optics in telecommunications. *Polarization Optics in Telecommunications: , Springer Series in Optical Sciences, Volume 101. ISBN 978-0-387-22493-0. Springer Science+Business Media, Inc., 2005*, 101, 01 2005.

[14] Bradley Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

[15] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[16] Alessandro Fedrizzi, Thomas Herbst, Andreas Poppe, Thomas Jennewein, and Anton Zeilinger. A wavelength-tunable fiber-coupled source of narrowband entangled photons. *Opt. Express*, 15(23):15377–15386, Nov 2007.

[17] Christopher Ferrie and Robin Blume-Kohout. Maximum likelihood quantum state tomography is inadmissible. *arXiv e-prints*, page arXiv:1808.01072, Aug 2018.

[18] Tobias Fritz. Beyond bells theorem: correlation scenarios. *New Journal of Physics*, 14(10):103001, oct 2012.

[19] Christina Giarmatzi and Fabio Costa. A quantum causal discovery algorithm. *npj Quantum Information*, 4(1):17, 2018.

[20] Marissa Giustina, Marijn A. M. Versteegh, Sören Wengerowsky, Johannes Handsteiner, Armin Hochrainer, Kevin Phelan, Fabian Steinlechner, Johannes Kofler, Jan-Åke Larsson, Carlos Abellán, Waldimar Amaya, Valerio Pruneri, Morgan W. Mitchell, Jörn Beyer, Thomas Gerrits, Adriana E. Lita, Lynden K. Shalm, Sae Woo Nam, Thomas Scheidl, Rupert Ursin, Bernhard Wittmann, and Anton Zeilinger. Significant-loophole-free test of bell's theorem with entangled photons. *Phys. Rev. Lett.*, 115:250401, Dec 2015.

[21] Deny R. Hamel. Realization of novel entangled photon sources using periodically poled materials, 2010.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2009.

[23] E. Hecht. *Optics*. Pearson education. Addison-Wesley, 2002.

[24] B. Hensen, H. Bernien, A. E. Dréau, A. Reiserer, N. Kalb, M. S. Blok, J. Ruitenberg, R. F. L. Vermeulen, R. N. Schouten, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, M. Markham, D. J. Twitchen, D. Elkouss, S. Wehner, T. H. Taminiau, and R. Hanson. Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526:682 EP –, 10 2015.

[25] Joe Henson, Raymond Lal, and Matthew F Pusey. Theory-independent limits on correlations from generalized bayesian networks. *New Journal of Physics*, 16(11):113043, nov 2014.

[26] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki. Quantum entanglement. *Rev. Mod. Phys.*, 81:865–942, Jun 2009.

[27] Daniel F. V. James, Paul G. Kwiat, William J. Munro, and Andrew G. White. Measurement of qubits. *Phys. Rev. A*, 64:052312, Oct 2001.

[28] Pieter Kok, W. J. Munro, Kae Nemoto, T. C. Ralph, Jonathan P. Dowling, and G. J. Milburn. Linear optical quantum computing with photonic qubits. *Reviews of Modern Physics*, 79(1):135174, Jan 2007.

[29] Jan-Åke Larsson. Loopholes in bell inequality tests of local realism. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424003, oct 2014.

[30] Ulf Leonhardt. *Essential Quantum Optics: From Quantum Measurements to Black Holes*. Cambridge University Press, 2010.

[31] Ryan Martin. *Data and Error Analysis*. PressBooks, Apr 2018.

[32] Michael D. Mazurek, Matthew F. Pusey, Kevin J. Resch, and Robert W. Spekkens. Experimentally bounding deviations from quantum theory in the landscape of generalized probabilistic theories. *arXiv e-prints*, page arXiv:1710.05948, Oct 2017.

[33] Mazurek, Michael. Testing classical and quantum theory with single photons, 2018.

[34] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, New York, NY, USA, 10th edition, 2011.

[35] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[36] Sandu Popescu and Daniel Rohrlich. Thermodynamics and the measure of entanglement. *Phys. Rev. A*, 56:R3319–R3321, Nov 1997.

[37] Katja Ried, Megan Agnew, Lydia Vermeyden, Dominik Janzing, Robert W. Spekkens, and Kevin J. Resch. A quantum advantage for inferring causal structure. *Nature Physics*, 11(5):414–420, May 2015.

[38] Martin Ringbauer, Christina Giarmatzi, Rafael Chaves, Fabio Costa, Andrew G. White, and Alessandro Fedrizzi. Experimental test of nonlocal causality. *Science Advances*, 2(8), 2016.

[39] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.

[40] Lynden K. Shalm, Evan Meyer-Scott, Bradley G. Christensen, Peter Bierhorst, Michael A. Wayne, Martin J. Stevens, Thomas Gerrits, Scott Glancy, Deny R. Hamel, Michael S. Allman, Kevin J. Coakley, Shellee D. Dyer, Carson Hodge, Adriana E. Lita, Varun B. Verma, Camilla Lambrocco, Edward Tortorici, Alan L. Migdall, Yanbao Zhang, Daniel R. Kumor, William H. Farr, Francesco Marsili, Matthew D. Shaw, Jeffrey A. Stern, Carlos Abellán, Waldimar Amaya, Valerio Pruneri, Thomas Jennewein, Morgan W. Mitchell, Paul G. Kwiat, Joshua C. Bienfang, Richard P. Mirin, Emanuel Knill, and Sae Woo Nam. Strong loophole-free test of local realism. *Phys. Rev. Lett.*, 115:250402, Dec 2015.

[41] R. Simon and N. Mukunda. Minimal three-component su(2) gadget for polarization optics. *Physics Letters A*, 143(4):165 – 169, 1990.

[42] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, 2016.

[43] Peter Spirts. Introduction to causal inference. *Journal of Machine Learning Research*, 11:1643–1662, 2010.

[44] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[45] Lydia Vermeyden. Fundamental tests of quantum mechanics using two-photon entanglement, 2014.

[46] Abraham Wald. Statistical decision functions. *Ann. Math. Statist.*, 20(2):165–205, 06 1949.

[47] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010.

[48] Gregor Weihs, Thomas Jennewein, Christoph Simon, Harald Weinfurter, and Anton Zeilinger. Violation of bell's inequality under strict einstein locality conditions. *Phys. Rev. Lett.*, 81:5039–5043, Dec 1998.

[49] Elie Wolfe, David Schmid, Ana Bel'en Sainz, Ravi Kunjwal, and Robert W. Spekkens. Quantifying bell: the resource theory of nonclassicality of common-cause boxes. 2019.

[50] F. N. C. Wong, J. H. Shapiro, and T. Kim. Efficient generation of polarization-entangled photons in a nonlinear crystal. *Laser Physics*, 16:1517–1524, November 2006.

[51] Christopher J Wood and Robert W Spekkens. 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, mar 2015.

[52] William K. Wootters. Entanglement of formation and concurrence. *Quantum Info. Comput.*, 1(1):27–44, January 2001.

[53] Magdalena Zych, Fabio Costa, Igor Pikovski, and Časlav Brukner. Bell's theorem for temporal order. *Nature Communications*, 10(1):3772, 2019.