Bayesian Causal Inference

Maximilian Kurthen

Master's Thesis
Max Planck Institute for Astrophysics
Within the Elite Master Program
Theoretical and Mathematical Physics
Ludwig Maximilian University of Munich
Technical University of Munich

Supervisor: PD Dr. Torsten Enßlin

Munich, September 12, 2018

Abstract

In this thesis we address the problem of two-variable causal inference. This task refers to inferring an existing causal relation between two random variables (i.e. $X \to Y$ or $Y \to X$) from purely observational data. We begin by outlining a few basic definitions in the context of causal discovery, following the widely used *do-Calculus* [Pea00]. We continue by briefly reviewing a number of state-of-the-art methods, including very recent ones such as *CGNN* [Gou+17] and *KCDC* [MST18].

The main contribution is the introduction of a novel inference model where we assume a Bayesian hierarchical model, pursuing the strategy of Bayesian model selection. In our model the distribution of the cause variable is given by a Poisson lognormal distribution, which allows to explicitly regard discretization effects. We assume Fourier diagonal covariance operators, where the values on the diagonal are given by power spectra. In the most shallow model these power spectra and the noise variance are fixed hyperparameters. In a deeper inference model we replace the noise variance as a given prior by expanding the inference over the noise variance itself, assuming only a smooth spatial structure of the noise variance. Finally, we make a similar expansion for the power spectra, replacing fixed power spectra as hyperparameters by an inference over those, where again smoothness enforcing priors are assumed.

Based on our assumptions we describe an algorithmic forward model in order to produce synthetic causal data. These synthetic datasets are being used as benchmarks in order to compare our model to existing State-of-the-art models, namely *LiNGAM* [Hoy+09], *ANM-HSIC* [Moo+16], *ANM-MML* [Ste+10], *IGCI* [Dan+10] and *CGNN* [Gou+17]. We explore how well the above methods perform in case of high noise settings, strongly discretized data and very sparse data. Our model (BCM) shows to perform generally reliable in case of the synthetic data sets. While it is able to provide an accuracy close to the ANM methods in case of high noise and strongly discretized data, which deliver the best performance here, it is able to outperform other methods in case of very sparse (10 samples) synthetic data. We further test our model on the *TCEP* benchmark set, which is a widely used benchmark with real world data. Here our model can provide an accuracy comparable to state-of-the-art algorithms and is able to outperform other methods in a setting where only a small number of samples (20 samples) are available.

Contents

List of Symbols **VI**

1. Introduction **1**

1.1. Motivation and Significance of the Topic 1

1.2. Structure of the Work 2

1.3. Related Work 2

2. Problem Setting and Related Work **3**

2.1. Problem Setting 3

2.2. Related Work 6

2.2.1. Additive Noise Models - Principles 7

2.2.2. LiNGAM - an ICA-based Approach 7

2.2.3. Additive Noise Models with Regression and Residuum Independence Scoring 7

2.2.4. Empirical Bayes - Additive Noise with MML Scoring 9

2.2.5. Kernel Deviance Measures 10

2.2.6. Information Geometric Causal Inference 10

2.2.7. Learning Methods - CGNN 11

3. Bayesian Inference Model **13**

3.1. Our Contribution and the Formalism of IFT 13

3.2. A Shallow Inference Model 13

3.2.1. Distribution of the Cause Variable 14

3.2.2. Functional Relation of Cause and Effect 18

3.2.3. Computing the Bayes factor 19

3.3. Inference of the Noise Variance 20

3.3.1. Unknown Variance of the Noise 20

3.3.2. Marginalization of the Noise Variance 21

3.4. Inference of Power Spectra 23

3.4.1. Unknown Power Spectra 23

3.4.2. Marginalization of the Cause Distribution Power Spectrum 24

3.4.3. Marginalization of the Power Spectrum for the Causal Mechanism . . 24

3.4.4. Evidence for the Deep Hierarchical Model 25

3.4.5. Simultaneous Laplace Approximation 27

3.4.6. Bayes Factor for the Deep Model Selection 30

4. Implementation and Benchmarks **31**

4.1. Sampling Causal Data via a Forward Model 31

4.2. Implementation of the Bayesian Causal Inference 32
4.2.1. Implementation of the Shallow Model 32
4.2.2. Issues with the Implementation of the Deeper Models 35
4.3. Methods for Comparison 37
4.4. Results 37
4.4.1. Default Synthetic Data 37
4.4.2. High Noise Data 38
4.4.3. Strongly Discretized Data 38
4.4.4. Very Sparse Data 39
4.4.5. Real World Data 39

5. Discussion and Conclusion **42**

A. Explicit Calculations **46**
A.1. Derivatives of γ_ξ 46
A.1.1. The Term $\log\left|2\pi\widehat{e^{\tau_\beta}}\right|$ 46
A.1.2. The Term $\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta$ 46
A.2. Derivatives of γ_ζ 47
A.2.1. The Term $\boldsymbol{y}^\dagger(\tilde{F}[\tau_f]+\widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y}$ 47
A.2.2. The Term $\log\left|\tilde{F}[\tau_f]+\widehat{e^{\eta(\boldsymbol{x})}}\right|$ 47
A.3. The Determinant $|\Gamma_\xi|$ 48
A.4. The Determinant $|\Gamma_\zeta|$ 48

B. Details on the Computational Implementation **49**

C. Scatter Plots for the Benchmark Data **50**

D. Description of the TCEP Benchmark Set **56**

List of Figures

2.1. Example for a DAG representing a causal structure 3
2.2. Models for causal relations in the 2 variable case 5

3.1. Overview over the Bayesian hierarchical model (shallow) 14
3.2. Overview over the Bayesian hierarchical model with inference of the noise variance 21
3.3. Overview over the deep Bayesian hierarchical model 26

4.1. Different samples of fields obtained via algorithm 1 33
4.3. Synthetic Datasets sampled via alg. 1 34
4.4. Demonstration of the computational minimization 36
4.5. Real World Data TCEP benchmark set with 20 randomly chosen samples per dataset 41

C.1. Synthetic Datasets sampled via alg. 1 with $\varsigma^2 = 0.2$ 50
C.2. Synthetic Datasets sampled via alg. 1 with $\varsigma^2 = 1.0$ 51
C.3. Synthetic Datasets sampled via alg. 1 with $n_{\mathrm{bins}} = 16$ 52
C.4. Synthetic Datasets sampled via alg. 1 with $n_{\mathrm{bins}} = 8$ 53
C.5. Synthetic Datasets sampled via alg. 1 with $\tilde{N} = 30$ 54
C.6. Synthetic Datasets sampled via alg. 1 with $\tilde{N} = 10$ 55

D.1. TCEP benchmark dataset 61

List of Symbols

β	a function, $\beta \in \mathbb{R}^{[0,1]}$
δ_z	The Dirac delta distribution centered at $z \in \mathbb{R}$, i.e. $\delta_z = \delta(\cdot - z)$
η	logarithmic noise variance, $\varsigma(x)^2 = e^{\eta(x)}$
$\boldsymbol{\epsilon}$	N-vector of noise samples $(\epsilon_1, ..., \epsilon_N)^T$
$\boldsymbol{\lambda}$	n_{bins}-vector of Poisson means, $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_{n_{\text{bins}}})^T$
$\Lambda(z)$	$N \times N$ matrix, with entries $\Lambda(z)_{ij} = (\mathcal{F}^\dagger \widehat{e^{\tau_f}\delta_z} \mathcal{F})_{x_i x_j}$
ς^2	variance of the noise in the i.i.d. noise case
$\sigma_{\beta/\eta/f}$	hyper parameters controlling the strength of the smoothness or slope priors
$\tau_{\beta/f}$	logarithmic power spectra, $P_{\beta/f}(q) = e^{\tau_{\beta/f}(q)}$
B	covariance matrix of the distribution for β
$\boldsymbol{d}$	tuple of the measurements $\boldsymbol{x}, \boldsymbol{y}$
$\mathcal{E}$	covariance matrix of the distribution for $\boldsymbol{\epsilon}$
f	a function representing the causal mechanism, $f \in \mathbb{R}^{[0,1]}$
F	covariance matrix of the distribution for f
$\tilde{F}$	$N \times N$ matrix, with entries $\tilde{F}_{i,j} = F(x_i, x_j)$
G	$N \times N$ matrix: $G = (\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}$
$\mathcal{F}$	Fourier transformation operator
$\mathcal{H}(\cdot)$	Information Hamiltonian, $\mathcal{H}(\cdot) = -\log \mathcal{P}(\cdot)$
$\boldsymbol{k}$	n_{bins}-vector of counts $(k_1, \cdots, k_{n_{\text{bins}}})^T$
$\mathcal{N}(u \vert m, U)$	Normal distribution in u, centered at m, with covariance U
$\mathcal{P}(\cdot)$	A probability distribution, sometimes also being used directly as the corresponding probability density function
$\mathcal{P}(A \vert B)$	the conditional probability of A given B
$\boldsymbol{x}$	N-vector of measurements $(x_1, \cdots, x_N)^T$
$\boldsymbol{y}$	N-vector of measurements $(y_1, \cdots, y_N)^T$
$\int \mathcal{D}[u]$	Notation for a path-integral over a function or field u
$\sim$	Notation for a random variable being distributed according to some probability distribution
$\hookleftarrow$	Notation for a specific value of a random variable being drawn as a sample from some probability distribution
$\vert \cdot \vert$	The determinant when applied to operators or matrices
$\hat{\cdot}$	Notation for a vector or field raised to a diagonal matrix or operator

1. Introduction

1.1. Motivation and Significance of the Topic

Causal Inference regards the problem of drawing conclusions about how some entity we can observe does - or does not - influence or is being influenced by another entity. Having knowledge about such law-like causal relations enables us to predict what will happen ($\hat{=}$ the effect) if we know how the circumstances ($\hat{=}$ the cause) do change. For example, one can draw the conclusion that a street will be wet (the effect) whenever it rains (the cause). Knowing that it will rain, or indeed observing the rainfall itself, enables one to predict that the street will be wet. Less trivial examples can be found in the fields of epidemiology (identifying some bacteria as the cause of a sickness) or economics (knowing how taxes will influence the GDP of a country).

As [PJS17] remark, the mathematical formulation of these topics has only recently been approached. Especially within the fields of data science and machine learning specific tasks from causal inference have been attracting much interest recently. [HHH18] propose that causal inference stands as a third main task of data science besides description and prediction. Judea Pearl, best known for his Standard Reference *Causality: Models, Reasoning and Inference*, recently claimed that the task of causal inference will be the next "big problem" for Machine Learning [Pea18]. Such a specific problem is the two variable causal inference, also addressed as the *cause-effect problem* by [PJS17]. Given purely observational data from two random variables, X and Y, which are directly causally related, the challenge is to infer the correct causal direction. In the example of rain and wetness of a street, this would mean, we are given two-dimensional observation samples corresponding to (*Rainfall*, *Wetness of street*). The samples itself could be ("**it rains**", "**the street is wet**"), ("**it doesn't rain**", "**the street is wet**"), ("**it doesn't rain**", "**the street is dry**"). We now have to conclude the true causal direction which is obviously *Rainfall* → *Wetness of street.*

Having only observational data means we can not intervene into the data, e.g. use a garden hose to see if the street gets wet when we simulate rainfall. In such a setting, inferring the true direction might seem to be a futile task. Indeed, inferring the true direction in the above example would be impossible if we only had observed the samples ("**it rains**", "**the street is wet**"), ("**it doesn't rain**", "**the street is dry**"). It was just the sample ("**it doesn't rain**", "**the street is wet**") that allowed us to discard the hypothetical direction *Wetness of street* → *Rainfall.* If we know that either one or the other direction have to be true, we can therefore conclude the true causal direction. Interestingly, this is an incorporation of a fundamental asymmetry between cause and effect which does always hold and can be exploited to tackle such an inference problem. Given two random variables, X and Y which are related causally, $X \rightarrow Y$ ("X causes Y"), there exists a fundamental independence between the distribution of the cause $\mathcal{P}(X)$ and the mechanism which relates the cause X to the effect Y. This independence however does not hold in the reverse direction. Most of the proposed methods for the inference of such a causal direction make use of this asymmetry in some way, either by considering the independence directly [Dan+10], [Moo+16], or by taking into account the algorithmic complexity for the description of the factorization $\mathcal{P}(X)\mathcal{P}(Y|X)$ and comparing it to the complexity of the reverse factorization $\mathcal{P}(Y)\mathcal{P}(X|Y)$.

1.2. Structure of the Work

The rest of the thesis will be structured as following. In Chapter 2 we will first outline and specify our problem setting. We will attempt to define a self-contained framework of definitions for the necessary models of causality, following the *do-Calculus* introduced by [Pea00]. We also will review existing methods here, namely *Additive Noise Models*, *Information Geometric Causal Inference* and *Learning Methods.*

Chapter 3 will describe our inference model which is based on a hierarchical Bayesian model. In Section 3.2 we will introduce a first, shallow inference model. Here we assume covariance operators, determining distributions for the causal mechanism and the cause variable distribution itself, as well as the noise variance, to be given.

In Section 3.3 we will relax the fixed assumptions by allowing the noise variance to be determined by a field itself which only is governed by a prior distribution controlling the slope of the field. This will be expanded in 3.4 by allowing for arbitrary power spectra which are assumed to be random variables distributed by smoothness enforcing priors.

In Chapter 4 we will accompany the theoretical framework with experimental results. As the computational implementation of the deeper models (3.3, 3.4) showed to be problematic, we will limit the considerations here to an implementation of the shallow inference model proposed in 3.2. To that end we begin by outlining a "forward model" which allows to sample causally related data in 4.1. We describe a specific algorithm for the inference model in 4.2, which is then tested on various benchmark data (4.4). To provide a reference, the performance is compared to state-of-the-art methods introduced in 2.2.

We conclude by assessing that our model generally can show competitive classification accuracy and propose possibilities to further advance the model.

1.3. Related Work

2. Problem Setting and Related Work

2.1. Problem Setting

We begin by briefly defining key concepts in causal inference using the *do-Calculus*, introduced by [Pea00]. This outline attempts to be as short as possible as necessary for the present thesis, however still self contained.

Definition 1. *A* ***causal structure*** *of some random variables* $\mathcal{V} = \{X_1, ..., X_n\}$ *is a directed acyclic graph (*DAG*)* $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ *with the elements of* $\mathcal{V}$ *as nodes and the edges* $\mathcal{E}$ *representing functional relations between the variables.*

Returning to the example of rain causing a street to be wet from the beginning, one could describe this situation by a causal structure with two vertices (R for *rainfall* and W for *wetness of the street*) and one directed edge (R, E). [1] To make this more illustrating, we expand the model by considering a sprinkler next to the street, with a state S either being on or off, as another possible cause for the street to be wet, and we consider the cloudiness of the sky (C) as a "reason" for the rainfall. The causal structure is therefore given by the DAG $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{C, R, S, W\}$ and $\mathcal{E} = \{(C, R), (R, W), (S, W)\}$. The full graph structure is depicted in Fig. 2.1.

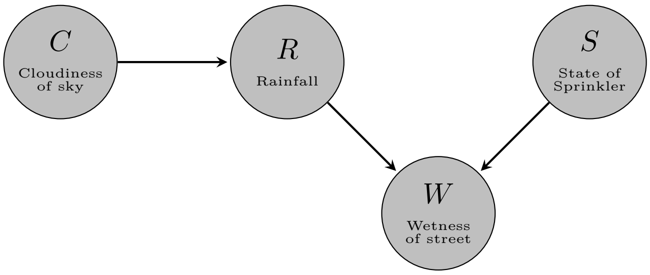


Figure 2.1.: An example for a DAG representing a causal structure, the nodes are random variables and the arrows give the directed edges, representing functional relations.

Definition 2. *A pair* $\mathcal{M} = (\mathcal{D}, \Theta_{\mathcal{D}})$ *of a graph* $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ *and its* ***parameters*** $\Theta_{\mathcal{D}}$ *defines a* ***causal model****. The parameters* $\Theta_{\mathcal{D}}$ *assign to each* $X_i \in \mathcal{V}$ *an equation of the form*

$$x_i = f(pa_i, u_i) \tag{2.1}$$

[1] Using the standard notation for directed graphs here, where an edge (a, b) indicates the direction "a to b".

and a probability distribution $\mathcal{P}(U_i)$ where Pa_i denote the parents of X_i w.r.t. to the graph $\mathcal{D}$ and U_i are unobserved disturbances having influences on X_i. x_i, pa_i, u_i denote realizations of X_i, Pa_i and U_i respectively.

In the above example (Fig. 2.1), the vertices C, S do not have any (observed) parental nodes, so their values are only determined by unobserved noise ($c = f_c(u_c)$ and $s = f_s(u_s)$). We think of the rainfall as an effect of the sky's cloudiness, so the corresponding equation would be $r = f_r(c, u_r)$, where we allow for some unobserved nuisance. The wetness of the street finally does have two separate observed causes - the rainfall and the sprinkler. This corresponds to an equation of the form $w = f_w(r, s, u_w)$, again allowing for some unobserved "noise" influencing the measured state of the street wetness. Now we can define an intervention in the model which represents an external manipulation of some variable:

Definition 3. *Given some causal model $\mathcal{M}$, with a random variable $X_i \in \mathcal{V}$ the atomic intervention* ***do***$(X_i = x_i)$ *is defined by setting the value of X_i to x_i, removing the equation regarding $x_i = f(pa_i, u_i)$ from the parameters of the model and substituting $X = x_i$ in all other equations.*

Assuming for a moment here we had the power to make it rain, we could set the variable $R =$**rainfall** in our example. This would remove the equation $r = f_r(c, u_r)$ from the parameters and set $w = (R = \textbf{rainfall}, s, u_w)$

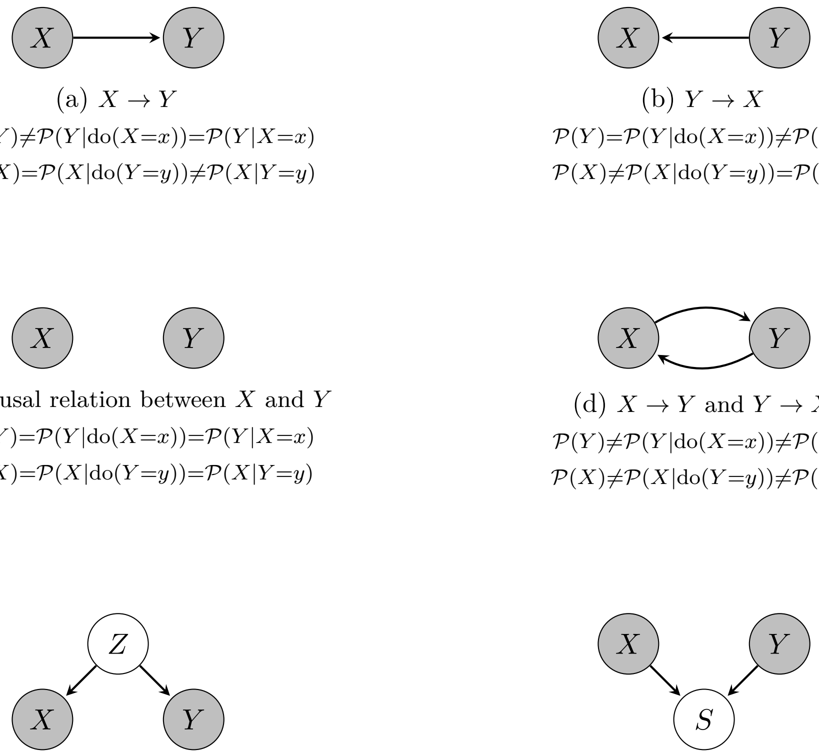
Using this *do*-formalism we can provide a definition for one variable being the cause of another one:

Definition 4. *Given two random variables X, Y with a joint probability distribution $\mathcal{P}(X, Y)$ and a corresponding conditional distribution $\mathcal{P}(Y|X)$ we say X* ***causes*** *Y (denoted by $X \rightarrow Y$) iff $\mathcal{P}(y|\mathrm{do}(x)) \neq \mathcal{P}(y|\mathrm{do}(x'))$ for some x, x' being realizations of X and y being a realization of Y*

For a last time returning to the above example, the probability for the street being wet, when we force rainfall, $\mathcal{P}(w|\mathrm{do}(R = \textbf{it rains})$ will probably be much higher than the corresponding one for manipulating the system such that it doesn't rain, $\mathcal{P}(w|\mathrm{do}(R = \textbf{it doesn't rain})$. In contrast to this, watering the street with a garden hose and thus making it wet will most likely not influence the probability of rain, $\mathcal{P}(r|\mathrm{do}(w = \textbf{the street is wet}) = \mathcal{P}(r|\mathrm{do}(w = \textbf{the street is dry})$

Fig. 2.2 shows some possibilities in which way two variables can be causally related. In case (c) there is no causal relation at all. Note that case (d) is not consistent with our definition of causality as it cannot be modeled with a DAG. Case (e) corresponds to a *confounding* variable, i.e. a variable that influences both of the other ones. A standard example in the statistical literature would be the human birth rate and the stork population, where one can find a significant correlation [Mat00]. As it is however unlikely that one of these variables directly causes the other one, there might be some confounding variable as industrialization or environmental health (and therefore standards of living) which is causing the other ones [ST05].

Figure 2.2.: Models for causal relations in the 2 variable case, reproduced from [Moo+16]



(a) $X \to Y$

$\mathcal{P}(Y) \neq \mathcal{P}(Y|\text{do}(X=x)) = \mathcal{P}(Y|X=x)$

$\mathcal{P}(X) = \mathcal{P}(X|\text{do}(Y=y)) \neq \mathcal{P}(X|Y=y)$

(b) $Y \to X$

$\mathcal{P}(Y) = \mathcal{P}(Y|\text{do}(X=x)) \neq \mathcal{P}(Y|X=x)$

$\mathcal{P}(X) \neq \mathcal{P}(X|\text{do}(Y=y)) = \mathcal{P}(X|Y=y)$

(c) no causal relation between X and Y

$\mathcal{P}(Y) = \mathcal{P}(Y|\text{do}(X=x)) = \mathcal{P}(Y|X=x)$

$\mathcal{P}(X) = \mathcal{P}(X|\text{do}(Y=y)) = \mathcal{P}(X|Y=y)$

(d) $X \to Y$ and $Y \to X$

$\mathcal{P}(Y) \neq \mathcal{P}(Y|\text{do}(X=x)) \neq \mathcal{P}(Y|X=x)$

$\mathcal{P}(X) \neq \mathcal{P}(X|\text{do}(Y=y)) \neq \mathcal{P}(X|Y=y)$

(e) (Hidden) confounder , $Z \to X$ and $Z \to Y$

$\mathcal{P}(Y) = \mathcal{P}(Y|\text{do}(X=x)) \neq \mathcal{P}(Y|X=x)$

$\mathcal{P}(X) = \mathcal{P}(X|\text{do}(Y=y)) \neq \mathcal{P}(X|Y=y)$

(f) V-collider , $X \to S$ and $Y \to S$

$\mathcal{P}(Y|S=s) \neq \mathcal{P}(Y|\text{do}(X=x), S=s) = \mathcal{P}(Y|X=x, S=s)$

$\mathcal{P}(X|S=s) \neq \mathcal{P}(X|\text{do}(Y=y), S=s) = \mathcal{P}(X|Y=y, S=s)$

Case (f) is often referred to as a *V-Collider* [Spi16] and illustrates the problem of *selection bias.* A car mechanic, only working with cars which do not start ($S = 0$), observes whether the start engine is broken ($X = 0$) and if the battery is empty ($Y = 0$). As in most cases one of these might be the case, but not the other one as well, he might draw the conclusion that X and Y are causally related (if $X = 1$, then usually $Y = 0$ and vice versa if $Y = 1$ then $X = 0$), if he does not consider his conditioning on the selection bias $S = 0$ [Moo+16]. I.e. the mechanic considers only $\mathcal{P}(X|Y), \mathcal{P}(Y|X)$ instead of $\mathcal{P}(X|Y, S), \mathcal{P}(Y|X, S)$ Not taking such a selection bias into account can thus lead to conclude a non-existing causal relation.

In this work we however only consider case (a) and (b), i.e. X being a cause of Y and Y being a cause of X and deciding which is the true one. Having access to the distributions $\mathcal{P}((X)|\text{do}(Y))$ and $\mathcal{P}(Y|\text{do}(X))$ would render this an easy task, as one could simply check via the definition 4 which direction holds. However, such distributions are usually not available, as manipulation of the systems of interest is often not possible. Instead we assume to have continuous observations in form of samples (x_i, y_i) being drawn from $\mathcal{P}(X, Y)$ and want to know which of the cases, (a) or (b) in Fig. 2.2 holds for the underlying process. We state the problem as following:

Problem 1. ***Prediction of causal direction for two variables***
Input*: A finite number of sample data* $\mathbf{d} \equiv (\boldsymbol{x}, \boldsymbol{y})$*, where* $\boldsymbol{x} = (x_1, ..., x_N), \boldsymbol{y} = (y_1, ..., y_N)$
Output*: A predicted causal direction* $\mathcal{D}_{X \to Y} \in \{-1, 1\}$ *where* -1 *represents the prediction "*$Y \to X$*" and* 1 *represents "*$X \to Y$*"*

2.2. Related Work

Approaches to tasks in causal inference from purely observational data are often divided into three groups ([SZ16; MST18]), namely constraint-based, score-based and asymmetry-based methods. Sometimes this categorization is extended by considering learning methods as a fourth, separate group. Two of these categories, constraint-based methods and score-based methods are basically searching for the true DAG representing some structure and rely on tests of conditional independence using conditioning on external variables. As such are not available in the two-variable case, those models are of little interest for the present task.

A third category exploits an inherent asymmetry between cause and effect. This asymmetry can be framed in different terms. One way is to use the concept of algorithmic complexity - given a true direction $X \to Y$, the factorization of the joint probability into $\mathcal{P}(X, Y) = \mathcal{P}(X)\mathcal{P}(Y|X)$ will be less complex than the reverse factorization $\mathcal{P}(Y)\mathcal{P}(X|Y)$ This approach is often used by *Additive Noise Models* (ANMs). Another way is to state that the mechanism relating cause and effect should be independent of the cause [Dan+10]. This formulation is employed by the concept of *Information Geometric Causal Inference.*

We will also consider *Causal Generative Neural Networks* (CGNN) as an example of learning methods. For the following, let $\mathcal{X}, \mathcal{Y}$ be some measurable spaces. We consider some random variables X and Y on $\mathcal{X}$ and $\mathcal{Y}$.

2.2.1. Additive Noise Models - Principles

A large family of inference models assume *additive noise*, i.e. in the case $X \to Y$, Y is determined by some function f, mapping X to Y, and some collective noise variable E_Y, i.e. $Y = f(X) + E_Y$, where X is independent of E_Y. Hypothetically, the same can be done in the backwards direction: $X = g(Y) + E_X$. [Moo+16] show that the resulting joint distribution $\mathcal{P}(X, Y)$ of such an ANM is either induced by the forward or the backward model, but generally not by both. If this is the case the model is said to be *identifiable.*

2.2.2. LiNGAM - an ICA-based Approach

An early adoption of this principle, *LiNGAM* (Linear Non-Gaussian Additive Noise Model, [Shi+06]) models the structure with linear functions, i.e. writing $x = c_x y + e_x$ or $y = c_y x + e_y$. This can be modeled using a vector calculus

$$\begin{pmatrix} x \\ y \end{pmatrix} = C \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_x \\ e_y \end{pmatrix} \tag{2.2}$$

As the restriction in the given problem setting allows only one direction of $X \to Y, Y \to X$ to be true, one of c_x, c_y can be set to zero, meaning C is a triangular matrix. Introducing the mixing matrix $A \equiv (\mathbb{1} - C)^{-1}$, the above relation can be re-written as

$$\begin{pmatrix} x \\ y \end{pmatrix} = A \begin{pmatrix} e_x \\ e_y \end{pmatrix} \tag{2.3}$$

The authors assume non-Gaussian error terms e_x, e_y which makes it possible to employ the technique of *independent component analysis* (ICA) to estimate the mixing matrix A which yields an estimation of the component matrix $C = \mathbb{1} - A^{-1}$. Finally a strictly lower triangular permutation is sought, the permutation matrix gives the causal ordering of the variables.

The model is proposed for a multi-variable case, the two-variable case is actually just a special case here. Even though we do not make the two main assumptions (namely linear relations and non-Gaussianity of the errors) in the course of specifying our model, it is mentioned at this place as LiNGAM has become a standard reference for causal inference benchmarks [Ste+10; MST18].

2.2.3. Additive Noise Models with Regression and Residuum Independence Scoring

More recent approaches usually deal with the possibility of non-linear causal mechanisms. As mentioned above, there is a fundamental asymmetry between cause and effect which incorporates in the fact that the distribution $\mathcal{P}(X)$ of the cause is independent from the true causal mechanism f itself. This functional mechanism is represented by the conditional distribution $\mathcal{P}(Y|X)$. A way to exploit this is to make a regression $\hat{f}$ for f and calculate the residual $\hat{f}(X) - Y$. If $\hat{f}$ would be an exact regression of f and $X \rightarrow Y$, this residual would now be fully independent of X.

However the regression is only somewhat precise (Gaussian Process Regression is usually employed here [Hoy+09; Moo+16]) and not the whole distribution $\mathcal{P}(X)$ but only a finite number of samples $(x_1, ..., x_N)$ is given. Therefore one cannot expect to measure complete independence between $(x_1, ..., x_N)$ and $(\hat{f}(x_1) - y_1, ..., \hat{f}(x_N) - y_N)$ but only a higher independence score in the true direction. A number of proposed methods use the (empirical) *Hilbert Schmidt Independence Criterion* (HSIC) to estimate the the independence. This measure is defined using the formulation of *Reproducing Kernel Hilbert Spaces* (RKHS), in which probability distributions can be bijectively embedded as elements.

An explicit introduction of this framework would not be in scope within this thesis, however we want to give a consistent, self-contained overview of the key concept used by recent publications in causal inference. The following outline is based on [Gre+05] and [Gre+07].

For a measurable space $\mathcal{X}$, the RKHS $\mathcal{H}$ is a Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, in which the evaluation functionals $\delta_x : f \mapsto f(x)$ are continuous for all $x \in \mathcal{X}$. As guaranteed by the *Riesz representation theorem* theorem, one can always represent such an evaluation at some $x \in \mathcal{X}$ by taking the inner product with a unique element of the Hilbert space:

$$f(x) = \delta_x(f) = \langle k_x, f \rangle_{\mathcal{H}} \tag{2.4}$$

The function k_x therefore defines a kernel via $k_x(y) = \langle k_y, k_x \rangle_{\mathcal{H}} \equiv k(x, y)$ and thus fully specifies the RKHS, allowing us to write $\mathcal{H}_k$ A probability distributions $\mathcal{P}$ over $\mathcal{X}$ can be embedded into such a RKHS $\mathcal{H}_{\mathcal{X}}$ by the *mean embedding*:

Definition 5. *For a kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, a random variable $X \sim \mathcal{P}$ in $\mathcal{X}$, the mean embedding $\mu_k(\mathcal{P})$ is defined by:*

$$\mu_k(\mathcal{P}) \equiv \mathbb{E}_{X \sim \mathcal{P}}[k(\cdot, X)] \tag{2.5}$$

given samples $(x_1, ..., x_N)$ from X an estimation of the mean embedding is given by

$$\hat{\mu}_k(\mathcal{P}) \equiv \frac{1}{N} \sum_{i=1}^{N} k(\cdot, x_i) \tag{2.6}$$

An inner product with the mean embedding is therefore given by Eq. 2.4 via

$$\langle f, \mu_k(\mathcal{P}) \rangle_{\mathcal{H}_k} \equiv \mathbb{E}_{X \sim \mathcal{P}}[f(X)] \tag{2.7}$$

Now one can define the HSIC [Gre+05]

Definition 6. *Given kernels* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *and* $y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$*, the* **HSIC** *is defined as*

$$\begin{aligned} \mathrm{HSIC}(X,Y)_{k,l} \equiv & ||\mu_{k\otimes l}(\mathcal{P}(X,Y)) - \mu_{k\otimes l}(\mathcal{P}(X)\mathcal{P}(Y)||_{\mathcal{H}_{k\otimes l}} = \\ = & \mathbb{E}_{\substack{X,X'\sim\mathcal{P}(X)\\ Y,Y'\sim\mathcal{P}(Y)}}[k(X,X')l(Y,Y') \\ & + \mathbb{E}_{X,X'\sim\mathcal{P}(X)}[k(X,X')]\mathbb{E}_{Y,Y'\sim\mathcal{P}(Y)}[l(Y,Y')] \\ & - 2\mathbb{E}_{\substack{X\sim\mathcal{P}(X)\\ Y\sim\mathcal{P}(Y)}}[\mathbb{E}_{X'\sim\mathcal{P}(X)}[k(X,X')]\mathbb{E}_{Y'\sim\mathcal{P}(Y)}[l(Y,Y')]] \end{aligned} \tag{2.8}$$

An estimation based on finite sample data $\boldsymbol{x} = (x_1, ..., x_N), \boldsymbol{y} = (y_1, ..., y_N)$ *, is given by* $\widehat{\mathrm{HSIC}}(\boldsymbol{x}, \boldsymbol{y})_{k,l} = (N-1)^{-2}\mathrm{tr}(KHLH)$*, with the Gram matrices* $K_{ij} = k(x_i, x_j), L_{ij} = l(x_i, x_j)$ *and the centering matrix* $H_{ij} = \delta_{ij} - N^{-1}$ *.*

This method, in the following called *ANM-HSIC*, performed strongly in recent benchmarks ([Moo+16; Gou+17; MST18]). The authors [Moo+16] use squared exponential (also called Gaussian) kernels here, i.e. $k(x,y) = e^{-\gamma(x-y)^2}$, where they allow the bandwidth γ to be estimated from the data itself.

2.2.4. Empirical Bayes - Additive Noise with MML Scoring

Other models, also based on the additive noise model, i.e. the assumption $Y = f(X) + E$, use *Bayesian model selection.* Here one compares the probability $\mathcal{P}(X \to Y|\boldsymbol{d})$ to the probability of the competing direction $\mathcal{P}(Y \to X|\boldsymbol{d})$, where again $\boldsymbol{d}$ denotes the observed samples, $\boldsymbol{d} = (\boldsymbol{x}, \boldsymbol{y})$. The ratio of these model probabilities

$$\mathcal{O}_{X\to Y} = \frac{\mathcal{P}(X \to Y|\boldsymbol{d})}{\mathcal{P}(Y \to X|\boldsymbol{d})} \tag{2.9}$$

is often referred to as the odds ratio or *Bayes factor* (see e.g. [BS09]).

Using Bayes Theorem

$$\mathcal{P}(X \to Y|\boldsymbol{d}) = \frac{\mathcal{P}(X \to Y)\mathcal{P}(\boldsymbol{d}|X \to Y)}{\mathcal{P}(\boldsymbol{d})} \tag{2.10}$$

and the fact that the prior probabilities of the competing models should be equal ($\mathcal{P}(X \to Y) = \mathcal{P}(Y \to X)$) we can express 2.9 in terms of the marginal likelihoods:

$$\mathcal{O}_{X\to Y} = \frac{\mathcal{P}(\boldsymbol{d}|X \to Y)}{\mathcal{P}(\boldsymbol{d}|Y \to X)} \tag{2.11}$$

Such a model is employed by [Ste+10] which assume a Gaussian mixture model for the distribution of the cause $\mathcal{P}(X)$ and a Gaussian Process with a squared exponential kernel for the causal mechanism. Here, the authors include a regression step for the function f representing the causal mechanism. Numerical quantities for the resulting terms are given by an expansion based on the concept of *Minimum Message Length* (MML). We will therefore refer to this approach as *ANM-MML*.

2.2.5. Kernel Deviance Measures

In a very recent (April 2018) publication, [MST18] introduced the method of *Kernel Conditional Deviance*. Here again the asymmetry in the algorithmic complexity between the factorizations of the joint probability $\mathcal{P}(X, Y)$, $\mathcal{P}(X)\mathcal{P}(Y|X)$ and $\mathcal{P}(Y)\mathcal{P}(X|Y)$ is considered. The authors reason that in case the true causal direction is $X \to Y$, it holds that

$$K(\mathcal{P}(Y|x_i)) = K(\mathcal{P}(Y|x_j)) \quad \forall i, j, \tag{2.12}$$

which however is not true in the other direction:

$$K(\mathcal{P}(X|y_i)) \neq K(\mathcal{P}(X|y_j)) \quad \forall i, j \tag{2.13}$$

Above, $K(\mathcal{P})$ denotes the *Kolmogorov complexity* of $\mathcal{P}$ which is, loosely speaking, the length of a program that encodes the distribution $\mathcal{P}$[GV08]. As this is not further specified, the Kolmogorov complexity itself is uncomputable and to be understood in a conceptional way. Based on this thought and using the variance in the of the conditional mean embedding of distributions in the RKHS, one derives at the estimator

$$S^{\text{KCDC}}_{X \to Y} = \frac{1}{N} \sum_{i=1}^{N} \left(||\mu_{Y|X=x_i}||_{\mathcal{H}_\mathcal{Y}} - \frac{1}{N} \sum_{j=1}^{N} ||\mu_{Y|X=x_j}||_{\mathcal{H}_\mathcal{Y}} \right) \tag{2.14}$$

and equivalently in the other direction $S^{\text{KCDC}}_{Y \to X}$, with roles of X and Y, resp. x and y switched. The direction predicted is the one with the lower deviance, i.e. $X \to Y$ if $S^{\text{KCDC}}_{X \to Y} < S^{\text{KCDC}}_{Y \to X}$, $Y \to X$ otherwise. The authors measure the performance of their algorithm in a experimental setup, describing mostly perfect predictions in case of synthetic data and very good ($\approx 74\%$) accuracy for the real world TCEP-benchmark.

2.2.6. Information Geometric Causal Inference

The concept that the distribution of the cause variable $\mathcal{P}(X)$ and the causal mechanism relating cause and effect, represented by the conditional distribution $\mathcal{P}(Y|X)$, represent independent mechanisms of nature is also the foundation of the approach of *Information Geometric Causal Inference* (IGCI). This approach has been introduced by [Dan+10], where, instead of approximating algorithmic complexity, the orthogonality of independent distributions in information space is exploited. The authors show that their approach even works in a deterministic, noise-free scenario. In such a scenario, one has $X = f(Y)$ and because of the bijectivity one also can state $Y = f^{-1}(X)$. The concept of IGCI is the thought that given independence between $\mathcal{P}(X)$ and f, the covariance of $\mathcal{P}(X)$ and $|\log(f')|$ being considered as random variables, should vanish. Here the covariance w.r.t. some reference distribution is considered, in the simplest case, the uniform distribution on $[0, 1]$. The authors arrive at the score

$$C^{\text{IGCI}}_{X \to Y} \equiv \int \mathrm{d}x \log(|f'(x)|)\mathcal{P}(x) - \int \mathrm{d}y \log(|g'(y)|)\mathcal{P}(y) = -C^{\text{IGCI}}_{Y \to X} \tag{2.15}$$

and infer $X \to Y$ whenever $C^{\text{IGCI}}_{X\to Y} > 0$ and $Y \to X$ otherwise. The authors show that Eq. 2.15 is is equivalent to the difference of Shannon entropies, i.e.

$$C^{\text{IGCI}}_{X\to Y} = S(\mathcal{P}(X)) - S(\mathcal{P}(Y)) \tag{2.16}$$

with the Shannon entropy

$$S(\mathcal{U}) \equiv -\int \mathrm{d}x\,\mathcal{U}(x)\log\mathcal{U}(x) \tag{2.17}$$

which they estimate on finite sample data $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^N$ via:

$$\hat{S}(\boldsymbol{x}) \equiv \psi(N) - \psi(1) + \frac{1}{N-1}\sum_{i=1}^{N-1} \log|x_{i+1} - x_i| \tag{2.18}$$

where $\psi(x) = \frac{d}{dx}\log\Gamma(x)$, the *Digamma* function. In 2.18, the samples are assumed to be in non-decreasing order, i.e. $x_i \leq x_{i+1}$. Further, the convention $\log(0) = 0$ is assumed, so that repeated samples with $x_{i+1} - x_i = 0$ do not contribute to $\hat{S}(\boldsymbol{x})$.

As a main drawback we consider the restriction to noiseless (or almost noiseless, the authors show that their work can be extended to models with small noise) case. Also the independence with respect to certain reference distributions is essential, as pointed out by the authors themselves. Furthermore, as [SZ16] remarks, the introduced method also relies on $\mathcal{P}(X)$ and $|\log(f')|$ being complex enough that they can be assessed for empirical results.

2.2.7. Learning Methods - CGNN

A recent publication [Gou+17] introduced a neural network-based approach for causal inference. The authors use neural networks with one hidden layer and a *ReLU* activation function (see e.g. [Bis06] for a detailed introduction on neural networks), taking samples from one variable, e.g. X, as the input and fitting the output to the other variable Y. The loss function is the (empirical) *Maximum Mean Discrepancy*, introduced by [Gre+07]:

Definition 7. *Given a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and random variables X, Y on $\mathcal{X}$, the* **MMD** *can be defined as*

$$\text{MMD}_k(X,Y) = ||\mathbb{E}_X k(\cdot, X) - \mathbb{E}_Y k(\cdot, Y)||_{\mathcal{H}_k} \tag{2.19}$$

and can be estimated from samples $\boldsymbol{x} = (x_1, ..., x_N) \hookleftarrow X$, $\boldsymbol{y} = (y_1, ..., y_{N'}) \hookleftarrow Y$ as the **empirical MMD***, given by:*

$$\widehat{\text{MMD}}_k(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{N^2}\sum_{ij}^{N} k(x_i,x_j) + \frac{1}{N^2}\sum_{ij}^{N'} k(y_i,y_j) - \frac{2}{NN'}\sum_{i}^{N}\sum_{j}^{N'} k(x_i,y_j) \tag{2.20}$$

The authors use a Gaussian kernel where the bandwidth γ is a hyperparameter to be set. After a training phase, in which the neural network is being tuned to predict the distribution for Y given the samples from X, the empirical MMD between the given samples $\boldsymbol{y}$ and the samples $\hat{\boldsymbol{y}}(\boldsymbol{x})$ predicted by neural network is being measured. The same is done for the reverse direction, $Y \to X$. The direction with the smaller MMD is then the preferred one. A known advantage of this method is that neural networks are *universal approximators*, meaning that basically every function can be approximated. The authors find that their method performs quite well, outperforming all other methods in the real-world data *TCEP*-benchmark.

3. Bayesian Inference Model

3.1. Our Contribution and the Formalism of IFT

Our contribution incorporates the concept of Bayesian model selection. As already briefly outlined in 2.2.4, this concept compares two competing models, in our case $X \rightarrow Y$ and $Y \rightarrow X$, and asks for the ratio of the marginalized likelihoods,

$$\mathcal{O}_{X\rightarrow Y} = \frac{\mathcal{P}(\boldsymbol{d}|X \rightarrow Y, M)}{\mathcal{P}(\boldsymbol{d}|Y \rightarrow X, M)}$$

Where M denotes the hyperparameters which are assumed to be the same for both models.

In the setting of the present causal inference problem, such an approach has already been used by [Ste+10]. In contrast to the above publication we will use the formalism of *information field theory* (IFT), introduced by [EFK09].

IFT considers *signal fields* s which reflect a physical state ψ, $s = s[\psi]$ and follows some probability $s \hookleftarrow \mathcal{P}(s)$. Such signal fields usually have infinite degrees of freedom, this makes them an adequate choice to model our distribution of the cause variable and the function relating cause and effect.

Throughout the following we will consider $X \rightarrow Y$ as the true underlying direction which we derive our formalism on. The derivation for $Y \rightarrow X$ will follow analogously by switching the variables.

3.2. A Shallow Inference Model

We will begin with deriving in 3.2.1 the distribution of the cause variable, $\mathcal{P}(X|X \rightarrow Y, M)$ where M defines a set of assumptions and hyperparameters we impose on the model and are yet to be specified. In 3.2.2 we continue by considering the conditional distribution $\mathcal{P}(Y|X, X \rightarrow Y, M)$. Combining those results, we compute then the full Bayes factor in 3.2.3.

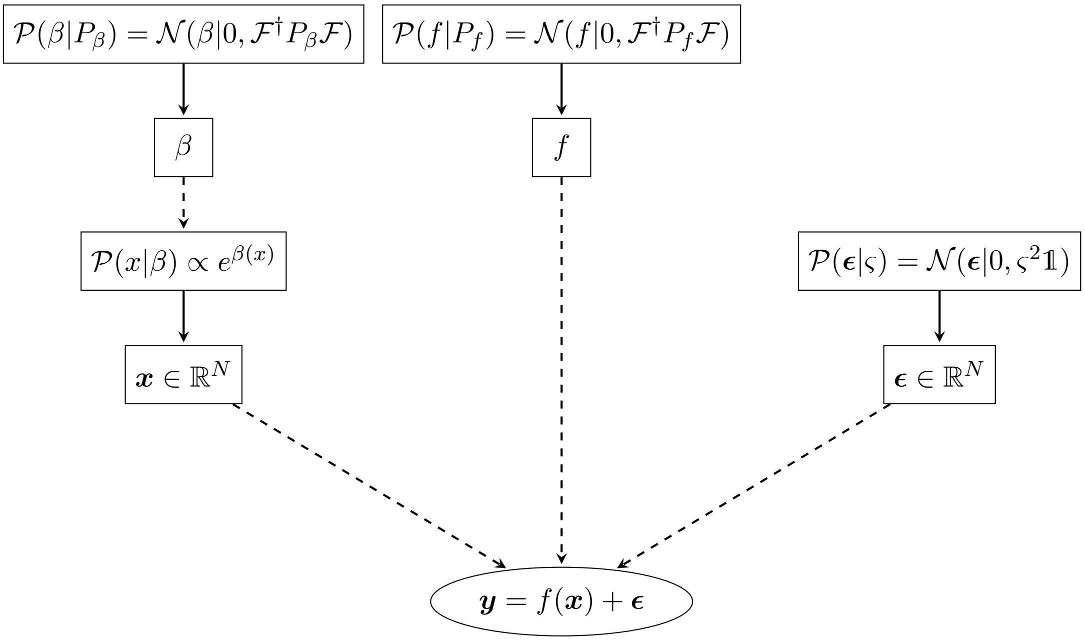

$\mathcal{P}(\beta|P_\beta) = \mathcal{N}(\beta|0, \mathcal{F}^\dagger P_\beta \mathcal{F})$
$\mathcal{P}(f|P_f) = \mathcal{N}(f|0, \mathcal{F}^\dagger P_f \mathcal{F})$
β
f
$\mathcal{P}(x|\beta) \propto e^{\beta(x)}$
$\mathcal{P}(\boldsymbol{\epsilon}|\varsigma) = \mathcal{N}(\boldsymbol{\epsilon}|0, \varsigma^2 \mathbb{1})$
$\boldsymbol{x} \in \mathbb{R}^N$
$\boldsymbol{\epsilon} \in \mathbb{R}^N$
$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$

Figure 3.1.: Overview over the most shallow Bayesian hierarchical model considered, for the case $X \to Y$

3.2.1. Distribution of the Cause Variable

Basic Considerations

Without imposing any constraints, we reduce our problem to the interval $[0, 1]$ by assuming that $\mathcal{X} = \mathcal{Y} = [0, 1]$. This can always be ensured by rescaling the data. Now we make the assumption that in principle, the cause variable X follows a lognormal distribution.

$$\mathcal{P}(x|\beta) \propto e^{\beta(x)} \tag{3.1}$$

with $\beta \in \mathbb{R}^{[0,1]}$, being some signal field which follows a zero-centered normal distribution, $\beta \sim \mathcal{N}(\beta|0, B)$.
Here we write B for the covariance operator $\mathbb{E}_{\beta \sim \mathcal{P}(\beta)}[\beta(x_0)\beta(x_1)] = B(x_0, x_1)$. We note that this is equivalent to the definition of β as a Gaussian Process which would be $\beta(x) = \mathcal{GP}(0, B(x, x'))$ using the notation of [RW06].

We postulate statistical homogeneity [1]). for the covariance, that is

$$\mathbb{E}_{\beta \sim \mathcal{P}(\beta)}[\beta(x)] = \mathbb{E}[\beta(x + t)] \tag{3.2}$$

$$\mathbb{E}_{\beta \sim \mathcal{P}(\beta)}[\beta(x)\beta(y)] = \mathbb{E}[\beta(x + t)\beta(y + t)] \tag{3.3}$$

i.e. first and second moments should be independent on the absolute location. The *Wiener-Khintchine Theorem* now states that the covariance has a spectral decomposition, i.e. it

[1] This property is called *stationarity* in the context of stochastic processes (see e.g. [Cha16]

is diagonal in Fourier space, under this condition (see e.g. [Cha16]). Denoting the Fourier transform by $\mathcal{F}$, i.e. in the one dimensional case, $\mathcal{F}[f](q) = (2\pi)^{-\frac{1}{2}} \int \mathrm{d}x \, e^{-iqx} f(x)$. Therefore, the covariance can be completely specified by a one dimensional function:

$$(\mathcal{F}B\mathcal{F}^{-1})(k,q) = \delta(k-q)P_\beta(k) \tag{3.4}$$

Here, $P_\beta(k)$ is called the *power spectrum* in the formalism of IFT [EFK09].

A Poisson Lognormal Approach for Handling Discretization

Building on these considerations we now regard the problem of discretization. Measurement data itself is usually not purely continuous but can only be given in a somewhat discretized way (e.g. by the measurement device itself or by precision restrictions imposed from storing the data). Another problem is that many numerical approaches to inference tasks, such as Gaussian Process regression, use finite bases as approximations in order to efficiently obtain results [Moo+16; Ste+10]. Here, we aim to directly confront these problems by imposing a formalism where the discretization is inherent.

So instead of taking a direct approach with the above formulation, we use a Poissonian approach and consider an equidistant grid $\{z_1, ..., z_{n_\text{bins}}\}$ in the $[0,1]$ interval. This is equivalent to defining bins, where the z_j are the midpoints of the bins. We now take the measurement counts, k_i which gives the number of x-measurements within the i-th bin. For these measurement counts we now take a Poisson lognormal distribution as an Ansatz, that is, we assume that the measurement counts for the bins are Poisson distributed, where the means follow a lognormal distribution. We can model this discretization by applying a response operator $R : \mathbb{R}^{[0,1]} \to \mathbb{R}^{n_\text{bins}}$ to the lognormal field. This is done in the most direct way via employing a delta distribution

$$R_{jx} \equiv \delta(x - z_j) \tag{3.5}$$

$$\tag{3.6}$$

In order to allow for a more compact notation we will use an index notation from now on, e.g. $f_x = f(x)$ for some function f or $O_{xy} = O(x,y)$ for some operator O. Whenever the indices are suppressed, an integration (in the continuous case) or dot product (in the discrete case) is understood, e.g. $(Of)_x \equiv O_{xy}f_y = \int \mathrm{d}y O_{xy} f_y = \int \mathrm{d}y O(x,y) f(y)$ In the following we will use bold characters for finite dimensional vectors, e.g. $\boldsymbol{\lambda} \equiv (\lambda_1, ..., \lambda_{n_\text{bins}})^T$. By inserting such a finite dimensional vector in the argument of a function, e.g. $\beta(\boldsymbol{x})$ we refer to a vector consisting of the function evaluated at each entry of $\boldsymbol{x}$, that is $(\beta(\boldsymbol{z}) \equiv (\beta(z_1), ..., \beta(z_{n_\text{bins}}))$. Later on we will use the notation $\hat{\cdot}$ which raises some vector to a diagonal matrix $(\hat{\boldsymbol{x}}_{ij} \equiv \delta_{ij} x_i)$. We will use this notation analogously for fields, e.g. $(\hat{\beta}_{uv} \equiv \delta(u-v)\beta(u))$. Now we can state the probability distribution for the measurement counts k_j:

$$\mathcal{P}(k_j|\lambda_j) = \frac{\lambda_j^{k_j} e^{-\lambda_j}}{k_j!} \tag{3.7}$$

$$\lambda_j = \mathbb{E}_{(k|\beta)}[k_j] = \rho e^{\beta_{z_j}} = \int dx R_{jx} e^{\beta_x} = \rho(Re^{\beta})_j \tag{3.8}$$

$$\boldsymbol{\lambda} = \rho \boldsymbol{R} e^{\beta} = \rho e^{\beta(\boldsymbol{z})} \tag{3.9}$$

$$\mathcal{P}(\boldsymbol{k}|\boldsymbol{\lambda}) = \prod_j \frac{\lambda_j^{k_j} e^{-\lambda_j}}{k_j!} = \prod_j \frac{(R_j e^{\beta})^{k_j} e^{-R_j e^{\beta}}}{k_j!} = \frac{(\prod_j (R_j e^{\beta})^{k_j}) e^{-\mathbf{1}^\dagger \boldsymbol{R} e^{\beta}}}{\prod_j k_j!} \tag{3.10}$$

$$\mathcal{P}(\boldsymbol{x}|\boldsymbol{k}) = \frac{1}{N!} \tag{3.11}$$

Eq. 3.11 follows from the consideration that given the counts $(k_1, ..., k_{n_{\text{bins}}})$ for the bins, only the positions of the observations $(x_1, ..., x_N)$ is fixed, but the ordering is not. The N observations can be ordered in $N!$ ways.

Now considering the whole r-vector of bin counts $\boldsymbol{k}$ at once, we get

$$\mathcal{P}(\boldsymbol{k}|\beta) = \frac{e^{\sum_j k_j \beta(z_j)} e^{-\rho^\dagger e^{\beta(\boldsymbol{z})}}}{\prod_j k_j!} = \frac{e^{\boldsymbol{k}^\dagger \beta(\boldsymbol{z}) - \boldsymbol{\rho}^\dagger e^{\beta(\boldsymbol{z})}}}{\prod_j k_j!} \tag{3.12}$$

$$\tag{3.13}$$

Marginalizing β we get

$$\begin{aligned}
\mathcal{P}(\boldsymbol{x}|P_\beta, X \to Y) &= \frac{1}{N!} \int \mathcal{D}[\beta] \mathcal{P}(x|\beta, X \to Y) \mathcal{P}(\beta|P_\beta) \\
&= \frac{1}{N!} |2\pi B|^{-\frac{1}{2}} \int \mathcal{D}[\beta] \frac{e^{\boldsymbol{k}^\dagger \beta(\boldsymbol{z}) - \boldsymbol{\rho}^\dagger e^{\beta(\boldsymbol{z})}}}{\prod_j k_j!} e^{-\frac{1}{2}\beta^\dagger B^{-1} \beta} = \\
&= \frac{|2\pi B|^{-\frac{1}{2}}}{N! \prod_j k_j!} \int \mathcal{D}[\beta] e^{-\gamma[\beta]}
\end{aligned} \tag{3.14}$$

where

$$\gamma[\beta] \equiv -\boldsymbol{k}^\dagger \beta(\boldsymbol{z}) + \boldsymbol{\rho}^\dagger e^{\beta(\boldsymbol{z})} + \frac{1}{2} \beta^\dagger B^{-1} \beta. \tag{3.15}$$

We approach this integration by a saddle point approximation. In the following we will denote the functional derivative by ∂, i.e. $\partial_{f_z} \equiv \frac{\delta}{\delta f(z)}$.

Taking the first and second order functional derivative of γ w.r.t. β we get (

$$\partial_\beta \gamma[\beta] = -\boldsymbol{k}^\dagger + \rho(e^{\beta(\boldsymbol{z})})^\dagger + \beta^\dagger B^{-1} \tag{3.16}$$

$$\partial_\beta \partial_\beta \gamma[\beta] = \widehat{\rho e^{\beta(\boldsymbol{z})}} + B^{-1} \tag{3.17}$$

The above derivatives are still defined in the space of functions $\mathbb{R}^{[0,1]}$, that is

$$\begin{aligned} \boldsymbol{k}_u^\dagger &\equiv \sum_{j=1}^{n_{\text{bins}}} k_j (\tilde{R}_j)_u \\ (\widehat{\rho e^{\beta(\boldsymbol{z})}})_{uv} &= \rho \sum_{j=1}^{n_{\text{bins}}} (\tilde{R}_j)_u (\tilde{R}_j)_v e^{\beta(u)} \end{aligned}$$

i.e. a diagonal operator with $e^{\beta(\boldsymbol{z})}$ as diagonal entries.

Let β_0 denote the function that minimizes the functional γ, i.e.

$$\left.\frac{\delta\gamma[\beta]}{\delta\beta}\right|_{\beta=\beta_0} = 0 \tag{3.18}$$

We expand the functional γ up to second order around β_0:

$$\begin{aligned} \int \mathcal{D}[\beta] e^{-\gamma[\beta]} &= \int \mathcal{D}[\beta] e^{-\gamma[\beta_0] - (\frac{\delta\gamma[\beta]}{\delta\beta}|_{\beta=\beta_0})^\dagger \beta - \frac{1}{2}\beta^\dagger (\frac{\delta^2\gamma[\beta]}{\delta\beta^\dagger\beta}|_{\beta=\beta_0})\beta + \mathcal{O}(\beta^3)} \\ &\approx e^{-\gamma[\beta_0]} \left| 2\pi \left(\frac{\delta^2\gamma[\beta]}{\delta\beta^2}|_{\beta=\beta_0} \right)^{-1} \right|^{\frac{1}{2}} \\ &= e^{+\boldsymbol{k}^\dagger \boldsymbol{\beta_0} - \boldsymbol{\rho}^\dagger e^{\boldsymbol{\beta_0}} - \frac{1}{2}\beta_0^\dagger B^{-1}\beta_0} \left| \frac{1}{2\pi} (\widehat{\rho e^{\boldsymbol{\beta_0}}} + B^{-1}) \right|^{-\frac{1}{2}} \end{aligned} \tag{3.19}$$

where we dropped higher order terms of β, used that the gradient at $\beta = \beta_0$ vanishes and evaluated the remaining Gaussian integral.

Plugging the result (3.19) into (3.14) and using

$$|2\pi B|^{-\frac{1}{2}} \left| \frac{1}{2\pi} (\widehat{\rho e^{\boldsymbol{\beta_0}}} + B^{-1}) \right|^{-\frac{1}{2}} = \left| B(\widehat{\rho e^{\boldsymbol{\beta_0}}} + B^{-1}) \right|^{-\frac{1}{2}} = \left| \rho B \widehat{e^{\boldsymbol{\beta_0}}} + \mathbb{1} \right|^{-\frac{1}{2}} \tag{3.20}$$

we get:

$$\mathcal{P}(\boldsymbol{x}|P_\beta, X \to Y) \approx \frac{1}{N!} \frac{e^{+\boldsymbol{k}^\dagger \boldsymbol{\beta_0} - \boldsymbol{\rho}^\dagger e^{\boldsymbol{\beta_0}} - \frac{1}{2}\beta_0^\dagger B^{-1}\beta_0}}{\left| \rho B \widehat{e^{\boldsymbol{\beta_0}}} + \mathbb{1} \right|^{\frac{1}{2}} \prod_j k_j!} \tag{3.21}$$

$$\mathcal{H}(\boldsymbol{x}|P_\beta, X \to Y) \approx \mathcal{H}_0 + \frac{1}{2}\log|\rho B \widehat{e^{\boldsymbol{\beta_0}}} + \hat{\mathbb{1}}| + \log(\prod_j k_j!) - \boldsymbol{k}^\dagger \boldsymbol{\beta_0} + \boldsymbol{\rho}^\dagger e^{\boldsymbol{\beta_0}} + \frac{1}{2}\beta_0^\dagger B^{-1}\beta_0 \tag{3.22}$$

where $\mathcal{H}(\cdot) \equiv -\log(\mathcal{P}(\cdot))$ is called the *information Hamiltonian* and $\mathcal{H}_0$ collects all terms which do not depend on the data $\boldsymbol{d}$.

3.2.2. Functional Relation of Cause and Effect

Similarly to β, we suppose a Gaussian distribution for the function f, relating Y to X:

$$\mathbb{R}^{[0,1]} \ni f \sim \mathcal{N}(0|f, F) \tag{3.23}$$

Proposing a Fourier diagonal covariance F once more, determined by a power spectrum P_f:

$$(\mathcal{F}F\mathcal{F}^{-1})(k,q) = \delta(k-q)P_f(k) \tag{3.24}$$

we assume additive Gaussian noise, using the notation $f(\boldsymbol{x}) \equiv (f(x_1), ..., f(x_N))^T$ and $\boldsymbol{\epsilon} \equiv (\epsilon_1, ..., \epsilon_N)^T$, we have

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon} \tag{3.25}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\epsilon|0, \mathcal{E}) \tag{3.26}$$

$$\mathcal{E} \equiv \operatorname{diag}(\varsigma^2, \varsigma^2, ...) = \varsigma^2 \mathbb{1} \in \mathbb{R}^{N\times N} \tag{3.27}$$

that is each independent noise sample is drawn from a zero-mean Gaussian distribution with given variance ς^2.

Knowing the noise $\boldsymbol{e}$, the cause $\boldsymbol{x}$ and the causal mechanism f completely determines $\boldsymbol{y}$ via 3.25. Therefore, $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, f, \boldsymbol{\epsilon}, X \to Y) = \delta(\boldsymbol{y} - f(\boldsymbol{x}) - \boldsymbol{\epsilon})$, where $\delta(\cdot)$ denotes the *Dirac delta distribution.* We can now state the conditional distribution for the effect variable measurements $\boldsymbol{y}$, given the cause variable measurements $\boldsymbol{x}$. Marginalizing out the dependence on the relating function f and the noise $\boldsymbol{\epsilon}$ we get:

$$\begin{aligned}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, P_f, \varsigma, X \to Y) &= \int \mathcal{D}[f] \, \mathrm{d}^N \boldsymbol{\epsilon} \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, f, \boldsymbol{\epsilon}, X \to Y)\mathcal{P}(\boldsymbol{\epsilon}|\varsigma)\mathcal{P}(f|P_f) \\ &= \int \mathcal{D}[f] \, \mathrm{d}^N \boldsymbol{\epsilon} \delta(\boldsymbol{y} - f(\boldsymbol{x}) - \boldsymbol{\epsilon})\mathcal{N}(\boldsymbol{\epsilon}|0, \mathcal{E})\mathcal{N}(f|0, F)\end{aligned} \tag{3.28}$$

We will now use the Fourier representation of the delta distribution, specifically $\delta(x) = \int \frac{\mathrm{d}q}{2\pi} e^{iqx}$

$$\delta(\boldsymbol{y} - f(\boldsymbol{x}) - \boldsymbol{\epsilon}) = \int \frac{\mathrm{d}^N \boldsymbol{q}}{(2\pi)^N} e^{i\boldsymbol{q}^\dagger(\boldsymbol{y} - \boldsymbol{\epsilon} - f(\boldsymbol{x}))} = \int \frac{\mathrm{d}^N \boldsymbol{q}}{(2\pi)^N} e^{i\boldsymbol{q}^\dagger(\boldsymbol{y} - \boldsymbol{\epsilon} - f(\boldsymbol{x}))} \tag{3.29}$$

Once more we employ a vector of response operators, mapping $\mathbb{R}^{\mathbb{R}}$ to $\mathbb{R}^N$,

$$\boldsymbol{R}_x \equiv (R_{1x}, ..., R_{Nx})^T = (\delta(x - x_1), ..., \delta(x - x_N))^T \tag{3.30}$$

This allows to represent the evaluation $f(\boldsymbol{x}) = \boldsymbol{R} \dagger f$, i.e. as a linear dot-product. Using the well known result for Gaussian integrals with linear terms (see e.g. [GBR13]),

$$\int \mathcal{D}[u] e^{-\frac{1}{2}u^\dagger A u + b^\dagger u} = \left|\frac{A}{2\pi}\right|^{-\frac{1}{2}} e^{\frac{1}{2}b^\dagger A b} \tag{3.31}$$

we are able to analytically do the path integral over f,

$$\begin{aligned}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, P_f, \varsigma, X \to Y) &= |2\pi F|^{-\frac{1}{2}} \int \mathcal{D}[f]\, \mathrm{d}^N\boldsymbol{\epsilon} \frac{\mathrm{d}^N\boldsymbol{q}}{(2\pi)^N} e^{i\boldsymbol{q}^\dagger(\boldsymbol{y}-\boldsymbol{\epsilon}-\boldsymbol{R}^\dagger f)-\frac{1}{2}f^\dagger F^{-1} f} \mathcal{N}(\boldsymbol{\epsilon}|0,\mathcal{E}) \\ &= \int \mathrm{d}^N\boldsymbol{\epsilon} \frac{\mathrm{d}^N\boldsymbol{q}}{(2\pi)^N} e^{i\boldsymbol{q}^\dagger(\boldsymbol{y}-\boldsymbol{\epsilon})+(-i)^2\frac{1}{2}\boldsymbol{q}^\dagger\boldsymbol{R}^\dagger F\boldsymbol{R}\boldsymbol{q}} \mathcal{N}(\boldsymbol{\epsilon}|0,\mathcal{E}) \qquad (3.32)\end{aligned}$$

Now we do the integration over the noise variable, $\boldsymbol{\epsilon}$, by using the equivalent of Eq. 3.31 for the vector-valued case:

$$\begin{aligned}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, P_f, \varsigma, X \to Y) &= |2\pi\mathcal{E}|^{-\frac{1}{2}} \int \mathrm{d}^N\boldsymbol{\epsilon} \frac{\mathrm{d}^N\boldsymbol{q}}{(2\pi)^N} e^{i\boldsymbol{q}^\dagger(\boldsymbol{y}-\boldsymbol{\epsilon})-\frac{1}{2}\boldsymbol{q}^\dagger\boldsymbol{R}^\dagger F\boldsymbol{R}\boldsymbol{q}-\frac{1}{2}\boldsymbol{\epsilon}^\dagger\mathcal{E}^{-1}\boldsymbol{\epsilon}} \\ &= \int \frac{\mathrm{d}^N\boldsymbol{q}}{(2\pi)^N} e^{i\boldsymbol{q}^\dagger\boldsymbol{y}-\frac{1}{2}\boldsymbol{q}(\boldsymbol{R}^\dagger F\boldsymbol{R}+\mathcal{E})\boldsymbol{q}} \qquad (3.33)\end{aligned}$$

In the following we will write $\mathbb{R}^{N\times N} \ni \tilde{F} = \boldsymbol{R}^\dagger F\boldsymbol{R}$, with entries $\tilde{F}_{ij} = F(x_i, x_j)$.[2] The integration over the Fourier modes $\boldsymbol{q}$, again via the multivariate equivalent of 3.31, will give the preliminary result:

$$\begin{aligned}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, P_f, \varsigma, X \to Y) &= \int \frac{\mathrm{d}^N\boldsymbol{q}}{(2\pi)^N} e^{i\boldsymbol{q}^\dagger\boldsymbol{y}-\frac{1}{2}\boldsymbol{q}(\tilde{F}+\mathcal{E})\boldsymbol{q}} \\ &= (2\pi)^{-\frac{N}{2}} \left|\tilde{F}+\mathcal{E}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}+\mathcal{E})^{-1}\boldsymbol{y}} \qquad (3.34)\end{aligned}$$

3.2.3. Computing the Bayes factor

Now we are able to calculate the full likelihood of the data $\boldsymbol{d} = (\boldsymbol{x}, \boldsymbol{y})$ given our assumptions P_β, P_f, ς for the direction $X \to Y$ and vice versa $Y \to X$. As we are only interested in the ratio of the probabilities and not in the absolute probabilities itself, it suffices to calculate the Bayes factor:

$$\begin{aligned}O_{X\to Y} &= \frac{\mathcal{P}(\boldsymbol{d}|P_\beta, P_f, \varsigma, X \to Y)}{\mathcal{P}(\boldsymbol{d}|P_\beta, P_f, \varsigma, Y \to X)} \\ &= \exp[\mathcal{H}(\boldsymbol{d}|P_\beta, P_f, \varsigma, Y \to X) - \mathcal{H}(\boldsymbol{d}|P_\beta, P_f, \varsigma, X \to Y)] \qquad (3.35)\end{aligned}$$

Above we used again the information Hamiltonian $\mathcal{H}(\cdot) \equiv -\log \mathcal{P}(\cdot)$

Making use of (3.21) and (3.34) we get, using the calculus for conditional distributions on the Hamiltonians, $\mathcal{H}(A, B) = \mathcal{H}(A|B) + \mathcal{H}(B)$

$$\begin{aligned}\mathcal{H}(\boldsymbol{d}|P_\beta, P_f, \varsigma, X \to Y) &= \mathcal{H}(\boldsymbol{x}|P_\beta, X \to Y) + \mathcal{H}(\boldsymbol{y}|\boldsymbol{x}, P_f, \varsigma, X \to Y) \\ &= \mathcal{H}_0 + \log(\prod_j k_j!) + \frac{1}{2}\log|\rho B\widehat{e^{\boldsymbol{\beta_0}}} + \mathbb{1}| - \boldsymbol{k}^\dagger\boldsymbol{\beta_0} + \\ &+ \boldsymbol{\rho}^\dagger e^{\boldsymbol{\beta_0}} + \frac{1}{2}\beta_0^\dagger B^{-1}\beta_0 + \frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}+\mathcal{E})^{-1}\boldsymbol{y} + \frac{1}{2}\left|\tilde{F}+\mathcal{E}\right| \qquad (3.36)\end{aligned}$$

[2] This type of matrix, i.e. the evaluation of covariance or kernel at certain positions, is sometimes called a Gram matrix.

Where we suppressed the dependence of $\tilde{F}, \beta_0$ on $\boldsymbol{x}$ (for the latter, the dependence is not explicit, but rather implicit as β_0 is determined by the minimum of the $\boldsymbol{x}$-dependent functional γ).

We omit stating $\mathcal{H}(\boldsymbol{d}|P_\beta, P_f, \varsigma, Y \rightarrow X)$ explicitly as the expression is just given by taking (3.36) and switching $\boldsymbol{x}$ and $\boldsymbol{y}$ or X and Y, respectively.

3.3. Inference of the Noise Variance

3.3.1. Unknown Variance of the Noise

So far we assumed the variance of the noise to be known and identical, $\epsilon_i \sim \mathcal{N}(\epsilon_i|0, \varsigma^2)$. We want to relax this condition, allowing the noise variance to be a priori unknown and position dependent by substituting $\varsigma^2 = \varsigma(x)^2 \rightarrow e^{\eta(x)}$, where $\eta \in \mathbb{R}^{[0,1]}$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, \mathcal{E}(\boldsymbol{x})) \tag{3.37}$$

$$\mathcal{E}(\boldsymbol{x}) = \mathrm{diag}(e^{\eta(x_1)}, e^{\eta(x_2)}, ..., e^{\eta(x_N)}) = \widehat{e^{\eta(\boldsymbol{x})}} \tag{3.38}$$

$$\Rightarrow |\mathcal{E}| = \prod_i^N e^{\eta(x_i)} = e^{\mathbf{1}^\dagger \eta(\boldsymbol{x})} \tag{3.39}$$

We further argue that there exists a spatial correlation in the noise level, meaning that points x, x' which are close are affected by a similar noise variance $e^{\eta(x)}, e^{\eta(x')}$. We incorporate this consideration by constraining the derivative $\nabla e^{\eta(x)}$ with a prior:

$$\mathcal{P}(\eta|\sigma_\eta) = Z[\sigma_\eta]^{-1} e^{-\frac{1}{2\sigma_\eta}\int \mathrm{d}q |\nabla \eta(q)|^2} \propto e^{-\frac{1}{2\sigma_\eta}\eta^\dagger \nabla^\dagger \nabla \eta} \tag{3.40}$$

where we dropped the normalization factor $Z[\sigma_\eta] = \int \mathcal{D}[\eta] e^{-\frac{1}{2\sigma_\eta}\eta^\dagger \nabla^\dagger \nabla \eta} = \left|\frac{\nabla^\dagger \nabla}{2\pi\sigma_\eta}\right|^{-\frac{1}{2}}$ as it depends only on the hyper parameter σ_η which we can set to a canonical value.

We remark that it would be possible to introduce a similar inference structure for the function η as for the functions β and f. This would mean a functional normal distribution for η with a Fourier diagonal covariance matrix given by a power spectrum P_η which itself can be assumed to be set as a fixed hyperparameter (or to be just constrained by some prior distribution and made part of the inference). However at this point we choose to refrain from this step in order to avoid the resulting model from being unnecessary complex.

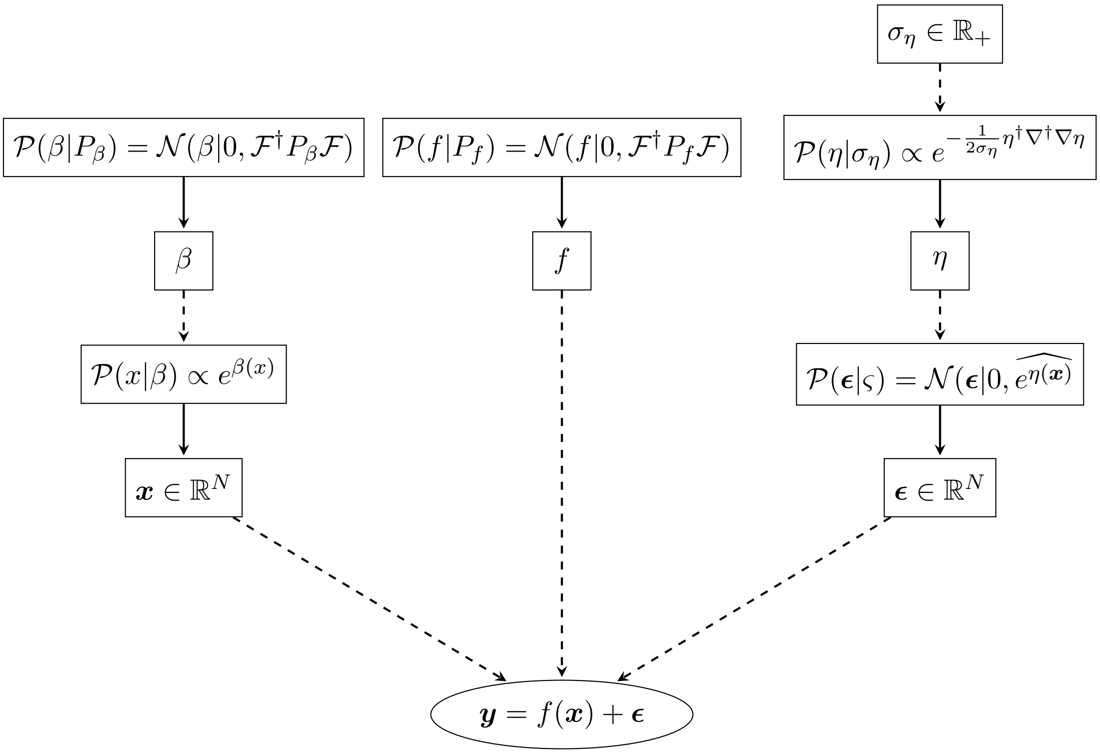

$\sigma_\eta \in \mathbb{R}_+$
$\mathcal{P}(\beta|P_\beta) = \mathcal{N}(\beta|0, \mathcal{F}^\dagger P_\beta \mathcal{F})$
$\mathcal{P}(f|P_f) = \mathcal{N}(f|0, \mathcal{F}^\dagger P_f \mathcal{F})$
$\mathcal{P}(\eta|\sigma_\eta) \propto e^{-\frac{1}{2\sigma_\eta}\eta^\dagger \nabla^\dagger \nabla \eta}$
β
f
η
$\mathcal{P}(x|\beta) \propto e^{\beta(x)}$
$\mathcal{P}(\boldsymbol{\epsilon}|\varsigma) = \mathcal{N}(\boldsymbol{\epsilon}|0, \widehat{e^{\eta(\boldsymbol{x})}}$
$\boldsymbol{x} \in \mathbb{R}^N$
$\boldsymbol{\epsilon} \in \mathbb{R}^N$
$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$

Figure 3.2.: Overview over the Bayesian hierarchical model, for the case $X \rightarrow Y$. The power spectra P_β, P_f are given as hyperparameters, but the noise variance is part of the inference.

3.3.2. Marginalization of the Noise Variance

Marginalizing out the noise variance in (3.34) gives

$$\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, P_f, \sigma_\eta, X \rightarrow Y) = \int \mathcal{D}[f,\eta]\, \mathrm{d}^N \boldsymbol{\epsilon} \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, \tau_f, \eta, X \rightarrow Y)\mathcal{P}(\boldsymbol{\epsilon}|\eta)\mathcal{P}(f|P_f)\mathcal{P}(\eta|\sigma_\eta) \propto$$
$$\propto \int \mathcal{D}[\eta] \left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} - \frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta\right) \tag{3.41}$$

again omitting the normalization factors depending only on σ_η and the $\frac{1}{\sqrt{2\pi}^N}$ factor.

In order to tackle the integration we again employ the Laplace approximation, introducing the energy functional $\gamma_\eta : \mathbb{R}^{[0,1]} \rightarrow \mathbb{R}$:

$$\gamma_\eta[\eta] \equiv \frac{1}{2}\log\left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right| + \frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} + \frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta \tag{3.42}$$

We now need to compute the first and second order derivatives of $\gamma_\eta(\eta)$ w.r.t. η. Especially the derivatives of the logarithmic determinant $\log\left(\left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right|\right)$ are non-obvious, so we will explicitly derive these. We begin by remarking that for some non-singular matrix or operator A, we have the identity [SN10]

$$\log|A| = \operatorname{tr}\log(A) \tag{3.43}$$

Applying the chain rule here gives the formula:

$$\frac{\partial|A[t]|}{\partial t} = |A[t]|\operatorname{tr}\left[A^{-1}\left(\frac{\partial A[t]}{\partial t}\right)\right] \tag{3.44}$$

We further make the abbreviation:

$$G[\eta] \equiv (\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}})^{-1} \tag{3.45}$$

Finally we will use the relation:

$$\begin{aligned} 0 = \frac{\partial \mathbb{1}}{\partial t} = \frac{\partial A[t]A[t]^{-1}}{\partial t} &= \frac{\partial A[t]}{\partial t}A[t]^{-1} + A[t]\frac{\partial A[t]^{-1}}{\partial t} \\ \Rightarrow \frac{\partial A[t]^{-1}}{\partial t} &= -A[t]^{-1}\frac{\partial A[t]}{\partial t}A[t]^{-1} \end{aligned} \tag{3.46}$$

We therefore get:

$$\partial_{\eta_u}\log\left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right| = \operatorname{tr}\left(G[\eta]\,\partial_{\eta_u}(\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})\right) = \operatorname{tr}\left(G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\right) \tag{3.47}$$

$$\begin{aligned} \partial_{\eta_v}\partial_{\eta_u}\log\left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right| &= \operatorname{tr}\left((\partial_{\eta_v}G)\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}} + G[\eta](\partial_{\eta_v}\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}})\right) \\ &= \operatorname{tr}\left(-G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}} + \delta_{uv}G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\right) \end{aligned} \tag{3.48}$$

The derivatives of the other terms are rather straight-forward:

$$\partial_{\eta_u}\boldsymbol{y}^\dagger(\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} = -\boldsymbol{y}^\dagger G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G[\eta]\boldsymbol{y} \tag{3.49}$$

$$\partial_{\eta_v}\partial_{\eta_u}\boldsymbol{y}^\dagger(\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} = \boldsymbol{y}^\dagger\left(2G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G[\eta] - G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\delta_{uv}G[\eta]\right)\boldsymbol{y} \tag{3.50}$$

$$\begin{aligned} \partial_{\eta_u}\frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta &= \frac{1}{\sigma_\eta}(\eta\nabla^\dagger\nabla)_u \\ \partial_{\eta_v}\partial_{\eta_u}\frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta &= \frac{1}{\sigma_\eta}(\nabla^\dagger\nabla)_{uv} \end{aligned} \tag{3.51}$$

We thus have the gradient:

$$\partial_{\eta_u}\gamma_\eta[\eta] = \frac{1}{2}\operatorname{tr}\left(G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\right) - \frac{1}{2}\boldsymbol{y}^\dagger G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G[\eta]\boldsymbol{y} + \frac{1}{\sigma_\eta}(\eta\nabla^\dagger\nabla)_u \tag{3.52}$$

and the curvature

$$\begin{aligned}\partial_{\eta_v}\partial_{\eta_u}\gamma_\eta[\eta] =&\mathrm{tr}\left(-G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}+\delta_{uv}G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\right)\\ &+\frac{1}{2}\boldsymbol{y}^\dagger\left(2G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G[\eta]-G[\eta]\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\delta_{uv}G[\eta]\right)\boldsymbol{y}\\ &+\frac{1}{\sigma_\eta}(\nabla^\dagger\nabla)_{uv}\equiv\\ \equiv&\Gamma_\eta[\eta]\end{aligned}\tag{3.53}$$

We write the the density in 3.41 in an exponential form, using the energy defined in 3.42:

$$\mathcal{P}(\boldsymbol{y}|\boldsymbol{x},P_f,\sigma_\eta,X\to Y)\propto\int\mathcal{D}[\eta]e^{-\gamma_\eta[\eta]}\tag{3.54}$$

And expand up to second order in η around $\eta_0=\underset{\eta\in\mathbb{R}^{[0,1]}}{\mathrm{argmin}}\gamma_\eta[\eta]$:

$$\begin{aligned}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x},P_f,\sigma_\eta,X\to Y)\propto\int\mathcal{D}[\eta]&e^{-\gamma_\eta[\eta_0]-(\frac{\delta\gamma_\eta[\eta]}{\delta\eta}|_{\eta=\eta_0})^\dagger\eta-\frac{1}{2}\eta^\dagger(\frac{\delta^2\gamma_\eta[\eta]}{\delta\eta^\dagger\eta}|_{\eta=\eta_0})\eta+\mathcal{O}(\eta^3)}\\ \approx&e^{-\gamma_\eta[\eta_0]}\left|2\pi(\frac{\delta^2\gamma_\eta[\eta]}{\delta\eta^2}|_{\eta=\eta_0})^{-1}\right|^{\frac{1}{2}}\\ =&e^{-\frac{1}{2}\log\left|\tilde{F}+\widehat{e^{\eta_0(\boldsymbol{x})}}\right|-\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}+\widehat{e^{\eta_0(\boldsymbol{x})}})^{-1}\boldsymbol{y}-\frac{1}{2\sigma_\eta}\eta_0\nabla^\dagger\nabla\eta_0}\left|\frac{1}{2\pi}\Gamma_\eta[\eta_0]\right|^{-\frac{1}{2}}\end{aligned}\tag{3.55}$$

Using the result above, we can state the information Hamiltonian for the causal direction $X\to Y$:

$$\begin{aligned}\mathcal{H}(\boldsymbol{d}|P_\beta,P_f,\sigma_\eta,X\to Y)&=\mathcal{H}(\boldsymbol{x}|P_\beta,X\to Y)+\mathcal{H}(\boldsymbol{y}|\boldsymbol{x},P_f,\sigma_\eta,X\to Y)\\ &=\mathcal{H}_0+\log(\prod_j k_j!)+\frac{1}{2}\log|\rho B\widehat{e^{\boldsymbol{\beta_0}}}+\mathbb{1}|-\boldsymbol{k}^\dagger\boldsymbol{\beta_0}+\boldsymbol{\rho}^\dagger e^{\boldsymbol{\beta_0}}\\ &+\frac{1}{2}\log\left|\frac{1}{2\pi}\Gamma_\eta[\eta_0]\right|+\frac{1}{2}\beta_0^\dagger B^{-1}\beta_0+\frac{1}{2}\log\left|\tilde{F}+\widehat{e^{\eta_0(\boldsymbol{x})}}\right|\\ &+\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}+\widehat{e^{\eta_0(\boldsymbol{x})}})^{-1}\boldsymbol{y}+\frac{1}{2\sigma_\eta}\eta_0\nabla^\dagger\nabla\eta_0\end{aligned}\tag{3.56}$$

3.4. Inference of Power Spectra

3.4.1. Unknown Power Spectra

So far we assumed the power spectra P_β, P_f to be given. We can approach a deeper model selection by expanding our inference on the power spectra itself and just assuming a somewhat smooth structure for the power spectra as a prior distribution. Assuming smoothness on a logarithmic scale gives:

$$
\begin{aligned}
&P_{\beta/f}(q) = e^{\tau_{\beta/f}(q)} && (3.57)\\
&\mathcal{H}(\tau_{\beta/f}|\sigma_{\beta/f}) = -\log(Z[\sigma_{\beta/f}])\frac{1}{2}\int \mathrm{d}\log(q)\sigma_{\beta/f}^{-2}\left(\frac{\partial^2\tau_{\beta/f}(q)}{\partial(\log q)^2}\right)^2 && (3.58)\\
&\mathcal{P}(\tau_{\beta/f}|\sigma_{\beta/f}) = Z[\sigma_{\beta/f}]^{-1}e^{-\frac{1}{2\sigma_{\beta/f}}\tau_{\beta/f}^{\dagger}\Delta^{\dagger}\Delta\tau_{\beta/f}} && (3.59)
\end{aligned}
$$

here,

$$
Z[\sigma_{\beta/f}] \equiv \int \mathcal{D}[\tau]e^{-\frac{1}{2\sigma_{\beta/f}}\tau^{\dagger}\Delta^{\dagger}\Delta\tau} = \left|\frac{\Delta^{\dagger}\Delta}{2\pi\sigma_{\beta/f}}\right|^{-\frac{1}{2}} \tag{3.60}
$$

serves as a normalization factor.

3.4.2. Marginalization of the Cause Distribution Power Spectrum

We can now perform a marginalization of the power spectrum $P_\beta = e^{\tau_\beta}$ in (3.14). We can use that $B = (\mathcal{F}^{\dagger}\widehat{e^{\tau_\beta}}\mathcal{F})$ and thus $|B| = |\mathcal{F}^{\dagger}||\widehat{e^{\tau_\beta}}||\mathcal{F}| = |\widehat{e^{\tau_\beta}}|$

$$
\begin{aligned}
\mathcal{P}(\boldsymbol{x}|\sigma_\beta, X \to Y) = &\int \mathcal{D}[\tau_\beta]\mathcal{P}(\boldsymbol{x}|\tau_\beta, \sigma_\beta, X \to Y)\mathcal{P}(\tau_\beta|\sigma_\beta) \propto\\
&\propto \int \mathcal{D}[\beta, \tau_\beta]\frac{|2\pi\widehat{e^{\tau_\beta}}|^{-\frac{1}{2}}}{\prod_j k_j!}e^{\boldsymbol{k}^{\dagger}\beta(\boldsymbol{z})-\boldsymbol{\rho}^{\dagger}e^{\beta(\boldsymbol{z})}-\frac{1}{2}\beta^{\dagger}\mathcal{F}^{\dagger}\widehat{e^{-\tau_\beta}}\mathcal{F}\beta-\frac{1}{2\sigma_\beta}\tau_\beta^{\dagger}\Delta^{\dagger}\Delta\tau_\beta}
\end{aligned} \tag{3.61}
$$

where we used $B^{-1} = (\mathcal{F}^{\dagger}\widehat{e^{\tau_\beta}}\mathcal{F})^{-1} = \mathcal{F}^{\dagger}\widehat{e^{-\tau_\beta}}\mathcal{F}$

3.4.3. Marginalization of the Power Spectrum for the Causal Mechanism

We perform a similar marginalization over τ_f for the likelihood of $\boldsymbol{y}$, given $\boldsymbol{x}$. We begin by noting:

$$
\begin{aligned}
\tilde{F}_{ij} &= F(x_i, x_j) = (\mathcal{F}^{\dagger}\widehat{e^{\tau_f}}\mathcal{F})(x_i, x_j)\\
&= \int \frac{1}{2\pi}\,\mathrm{d}q\,\mathrm{d}q' e^{ix_iq}e^{-ix_jq'}e^{\tau_f(q)}\delta(q-q')\\
&= \int \frac{1}{2\pi}\,\mathrm{d}q e^{iq(x_i-x_j)}e^{\tau_f(q)}\\
&= \mathcal{F}[e^{\tau_f}](x_i - x_j)
\end{aligned} \tag{3.62}
$$

From (3.41) we now get

$$\begin{aligned}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x},\sigma_f,\eta)_{X\to Y} = &\int \mathcal{D}[\tau_f]\mathcal{P}(\boldsymbol{y}|\boldsymbol{x},\tau_f,\eta)_{X\to Y}\mathcal{P}(\tau_f|\sigma_f) \propto \\ &\propto \int \mathcal{D}[\tau_f]|\widehat{e^{\eta(\boldsymbol{x})}}|^{\frac{1}{2}} \left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}+\widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y}-\frac{1}{2\sigma_f}\tau_f^\dagger\Delta^\dagger\Delta\tau_f}\end{aligned} \tag{3.63}$$

$$\begin{aligned}\mathcal{H}(\boldsymbol{y},\tau_f|\boldsymbol{x},\sigma_f,\eta)_{X\to Y} + \mathcal{H}_{\prime} - \frac{1}{2}\log|\widehat{e^{\eta(\boldsymbol{x})}}| + \frac{1}{2}\log\left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right| \\ + \frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} - \frac{1}{2\sigma_f}\tau_f^\dagger\Delta^\dagger\Delta\tau_f\end{aligned} \tag{3.64}$$

where we again collected data-independent factors as $\frac{1}{\sqrt{2\pi}^N}$ and $Z[\sigma_f]^{-1}$ in $\mathcal{H}_{\prime}$.

This leads to the marginalized likelihood for $\boldsymbol{y}$ given $\boldsymbol{x}$ for the deep model,

$$\begin{aligned}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x},P_f,\sigma_\eta,X\to Y) = &\int \mathcal{D}[f,\eta]\,\mathrm{d}^N\boldsymbol{\epsilon}\mathcal{P}(\boldsymbol{y}|\boldsymbol{x},\tau_f,\eta,X\to Y)\mathcal{P}(\boldsymbol{\epsilon}|\eta)\mathcal{P}(f|\sigma_f)\mathcal{P}(\eta|\sigma_\eta) \propto \\ &\propto \int \mathcal{D}[\eta]\left|\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} + \frac{1}{2}\mathbf{1}^\dagger\eta(\boldsymbol{x})\right. \\ &\left.-\frac{1}{2\sigma_f}\tau_f^\dagger\Delta^\dagger\Delta\tau_f - \frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta\right)\end{aligned} \tag{3.65}$$

3.4.4. Evidence for the Deep Hierarchical Model

We summarize the background assumptions (which includes the model, the hyper parameters and the causal direction) into a hypothesis H_1 and H_{-1} respectively:

$$\begin{aligned}H_1 &\equiv (\sigma_\beta,\sigma_f,\sigma_\eta,X\to Y) \\ H_{-1} &\equiv (\sigma_\beta,\sigma_f,\sigma_\eta,Y\to X)\end{aligned} \tag{3.66}$$

For the full evidence, given the Hypothesis H_1:

$$\begin{aligned}\mathcal{P}(\boldsymbol{d}|H_1) =&\mathcal{P}(\boldsymbol{x}|\sigma_\beta,X\to Y)\mathcal{P}(\boldsymbol{y}|\boldsymbol{x},\sigma_f,\sigma\eta,X\to Y) \propto \\ &\propto \int \mathcal{D}[\beta,\tau_\beta,\tau_f,\eta]\left|\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}}\right|^{-\frac{1}{2}} \frac{|2\pi\widehat{e^{\tau_\beta}}|^{-\frac{1}{2}}}{\prod_j k_j!}\times \\ &\times \exp\left(\boldsymbol{k}^\dagger\beta(\boldsymbol{z}) - \boldsymbol{\rho}^\dagger e^{\beta(\boldsymbol{z})} - \frac{1}{2}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta\right. \\ &-\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} + \frac{1}{2}\mathbf{1}^\dagger\eta(\boldsymbol{x}) - \\ &\left.-\frac{1}{2\sigma_\beta}\tau_\beta^\dagger\Delta^\dagger\Delta\tau_\beta - \frac{1}{2\sigma_f}\tau_f^\dagger\Delta^\dagger\Delta\tau_f - \frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta\right)\end{aligned} \tag{3.67}$$

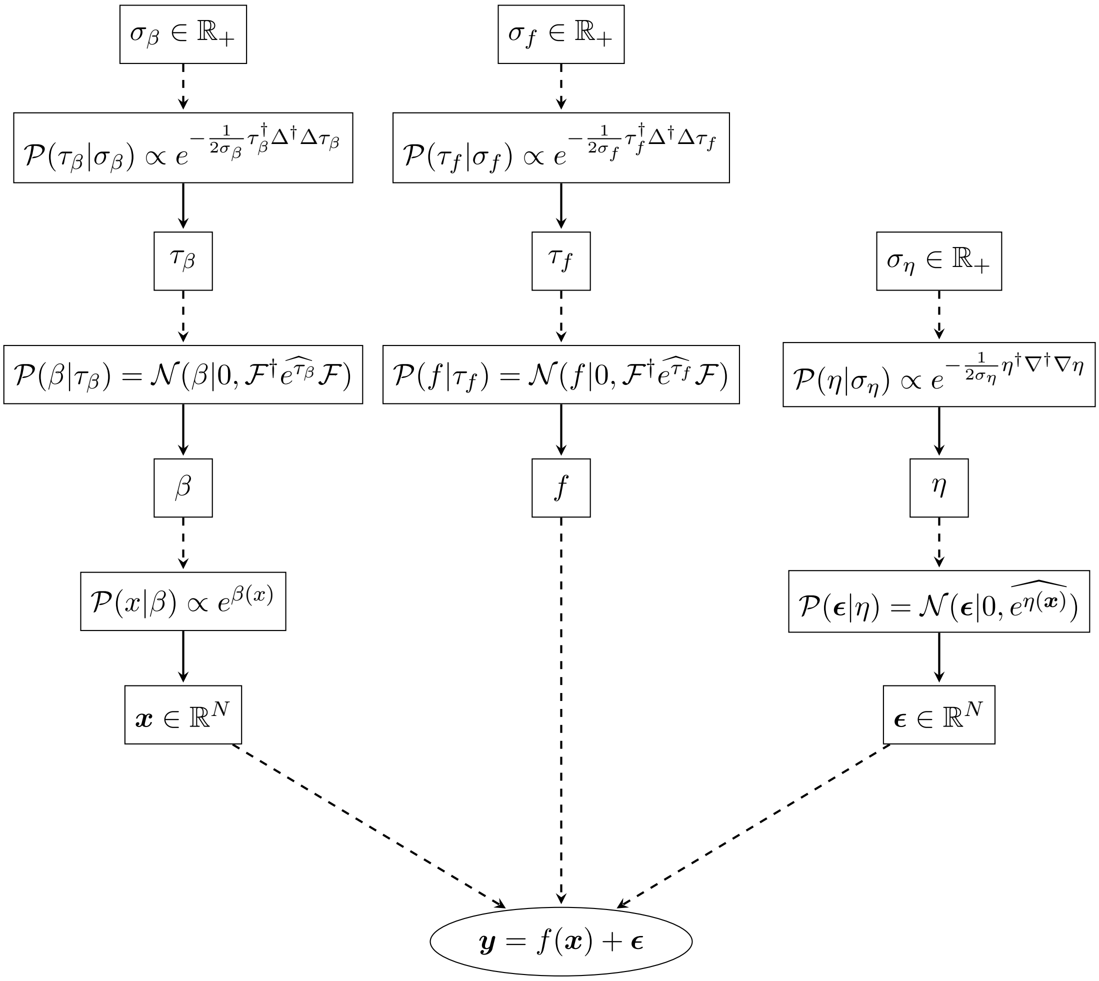

$\sigma_\beta \in \mathbb{R}_+$
$\sigma_f \in \mathbb{R}_+$
$\mathcal{P}(\tau_\beta|\sigma_\beta) \propto e^{-\frac{1}{2\sigma_\beta}\tau_\beta^\dagger \Delta^\dagger \Delta \tau_\beta}$
$\mathcal{P}(\tau_f|\sigma_f) \propto e^{-\frac{1}{2\sigma_f}\tau_f^\dagger \Delta^\dagger \Delta \tau_f}$
τ_β
τ_f
$\sigma_\eta \in \mathbb{R}_+$
$\mathcal{P}(\beta|\tau_\beta) = \mathcal{N}(\beta|0, \mathcal{F}^\dagger \widehat{e^{\tau_\beta}} \mathcal{F})$
$\mathcal{P}(f|\tau_f) = \mathcal{N}(f|0, \mathcal{F}^\dagger \widehat{e^{\tau_f}} \mathcal{F})$
$\mathcal{P}(\eta|\sigma_\eta) \propto e^{-\frac{1}{2\sigma_\eta}\eta^\dagger \nabla^\dagger \nabla \eta}$
β
f
η
$\mathcal{P}(x|\beta) \propto e^{\beta(x)}$
$\mathcal{P}(\boldsymbol{\epsilon}|\eta) = \mathcal{N}(\boldsymbol{\epsilon}|0, \widehat{e^{\eta(\boldsymbol{x})}})$
$\boldsymbol{x} \in \mathbb{R}^N$
$\boldsymbol{\epsilon} \in \mathbb{R}^N$
$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$

Figure 3.3.: Overview over the Bayesian hierarchical model for the case $X \to Y$. Here, besides the model itself, only strength parameters for the smoothness of the fields are fixed as Hyperparameters (H_1).

3.4.5. Simultaneous Laplace Approximation

The energy

Once more we want to tackle the integration of Eq. 3.67 via a Laplace approximation. In this case however, we cannot divide the integration over β, τ_β into two separate ones, same as for the integration over η, τ_f. We will therefore simultaneously find the arguments which maximize the expression under the integral. To do so we rewrite the above expression into an exponential form under the integrals, separating the integrations over β, τ_β and η, τ_f:

$$\begin{aligned}\mathcal{P}(\boldsymbol{d}|H_1) =& \frac{1}{\prod_j k_j!}\left[\int \mathcal{D}[\beta, \tau_\beta] \exp\left(-\frac{1}{2}\log\left|2\pi\widehat{e^{\tau_\beta}}\right| + \boldsymbol{k}^\dagger\beta(\boldsymbol{z}) - \boldsymbol{\rho}^\dagger e^{\beta(\boldsymbol{z})}\right.\right.\\ &\left.\left.-\frac{1}{2}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta - \frac{1}{2\sigma_\beta}\tau_\beta^\dagger\Delta^\dagger\Delta\tau_\beta\right)\right]\left[\int \mathcal{D}[\eta, \tau_f]\exp\left(-\frac{1}{2}\log\left|\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}}\right|\right.\right.\\ &\left.\left.-\frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} + \frac{1}{2}\boldsymbol{1}^\dagger\eta(\boldsymbol{x}) - \frac{1}{2\sigma_f}\tau_f^\dagger\Delta^\dagger\Delta\tau_f - \frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta\right)\right] \equiv\\ &\equiv \frac{1}{\prod_j k_j!}\left[\int \mathcal{D}[\beta, \tau_\beta]e^{-\gamma_\xi[\beta,\tau_\beta]}\right]\left[\int \mathcal{D}[\eta, \tau_f]e^{-\gamma_\zeta[\eta,\tau_f]}\right] \end{aligned} \tag{3.68}$$

Above we exploited that the probability density under the integrals factorizes into terms depending on either β, τ_β or η, τ_f, i.e. either on fields that determine the cause distribution or the causal mechanism, but not on both. We write $\mathbb{R}^{[0,1]} \oplus \mathbb{R}^{\mathbb{R}_+} \ni \xi \equiv (\beta, \tau_\beta)$ and $\mathbb{R}^{[0,1]} \oplus \mathbb{R}^{\mathbb{R}_+} \ni \zeta \equiv (\eta, \tau_f)$.

We introduced the energy functionals:

$$\gamma_\xi[\beta, \tau_\beta] \equiv \frac{1}{2}\log\left|2\pi\widehat{e^{\tau_\beta}}\right| - \boldsymbol{k}^\dagger\beta(\boldsymbol{z}) + \boldsymbol{\rho}^\dagger e^{\beta(\boldsymbol{z})} + \frac{1}{2}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta + \frac{1}{2\sigma_\beta}\tau_\beta^\dagger\Delta^\dagger\Delta\tau_\beta \tag{3.69}$$

$$\begin{aligned}\gamma_\zeta[\eta, \tau_f] \equiv& \frac{1}{2}\log\left|\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}}\right| + \frac{1}{2}\boldsymbol{y}^\dagger(\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y} - \frac{1}{2}\boldsymbol{1}^\dagger\eta(\boldsymbol{x})\\ &+ \frac{1}{2\sigma_f}\tau_f^\dagger\Delta^\dagger\Delta\tau_f + \frac{1}{2\sigma_\eta}\eta\nabla^\dagger\nabla\eta\end{aligned} \tag{3.70}$$

We consider the fields that minimize the energies:

$$\xi_0 \equiv \underset{\xi\in\mathbb{R}^{[0,1]}\oplus\mathbb{R}^{\mathbb{R}_+}}{\operatorname{argmin}} \gamma_\xi[\xi] \tag{3.71}$$

$$\zeta_0 \equiv \underset{\zeta\in\mathbb{R}^{[0,1]}\oplus\mathbb{R}^{\mathbb{R}_+}}{\operatorname{argmin}} \gamma_\zeta[\zeta] \tag{3.72}$$

Again we perform a second-order expansion of γ_ξ, γ_ζ around ξ_0, ζ_0 w.r.t. ξ and ζ respectively. In the following we will denote the curvatures by $\Gamma_\xi[\xi] \equiv \frac{\delta^2}{\delta\xi^\dagger\delta\xi}\gamma_\xi[\xi]$ and $\Gamma_\zeta[\zeta] \equiv \frac{\delta^2}{\delta\zeta^\dagger\delta\zeta}\gamma_\zeta[\zeta]$.

Omitting terms of higher order and exploiting the vanishing gradient (and thus first order) we end up with:

$$\begin{aligned}\mathcal{P}(\boldsymbol{d}|H_1) =& \frac{1}{\prod_j k_j!}\left[\int \mathcal{D}[\xi] e^{-\gamma_\xi[\xi_0]-(\frac{\delta}{\delta_\xi}\gamma_\xi[\xi]|_{\xi=\xi_0})\xi-\frac{1}{2}\xi^\dagger\Gamma_\xi[\xi]|_{\xi=\xi_0}\xi+\mathcal{O}(\xi^3)}\right]\times \\ &\times\left[\int \mathcal{D}[\zeta] e^{-\gamma_\zeta[\zeta_0]-(\frac{\delta}{\delta_\zeta}\gamma_\zeta[\zeta]|_{\zeta=\zeta_0})\zeta-\frac{1}{2}\zeta^\dagger\Gamma_\zeta[\zeta]|_{\zeta=\zeta_0}\zeta+\mathcal{O}(\zeta^3)}\right]\approx \\ \approx& \frac{1}{\prod_j k_j!}\left|\frac{1}{2\pi}\Gamma_\xi[\xi]|_{\xi_0}\right|^{-\frac{1}{2}}\left|\frac{1}{2\pi}\Gamma_\zeta[\zeta]|_{\zeta_0}\right|^{-\frac{1}{2}} e^{-\gamma_\xi[\xi_0]-\gamma_\zeta[\zeta_0]}\end{aligned} \tag{3.73}$$

To specify the approximation we need to compute the derivatives of first and second order

Derivatives of γ_ξ

We give the explicit computations in A.1, stating only the results here. We have the gradient:

$$\partial_{\beta_u}\gamma_\xi[\beta,\tau_\beta] = -\boldsymbol{k}_u + \rho(e^{\beta(\boldsymbol{z})})^\dagger + \left(\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\right)_u \tag{3.74}$$

$$\partial_{\tau_{\beta_u}}\gamma_\xi[\beta,\tau_\beta] = \frac{1}{2} - \frac{1}{2}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}\delta_u}\mathcal{F}\beta + \left(\frac{1}{\sigma_\beta}\tau_\beta^\dagger\Delta^\dagger\Delta\right)_u \tag{3.75}$$

The curvature, $\Gamma_\xi[\beta,\tau_\beta]$ has non-vanishing mixed derivatives,

$$\Gamma_\xi[\beta,\tau_\beta] = \begin{pmatrix} \frac{\delta^2\gamma_\xi}{\delta\beta^\dagger\delta\beta} & \frac{\delta^2\gamma_\xi}{\delta\beta^\dagger\delta\tau_\beta} \\ \frac{\delta^2\gamma_\xi}{\delta\tau_\beta^\dagger\delta\beta} & \frac{\delta^2\gamma_\xi}{\delta\tau_\beta^\dagger\delta\tau_\beta} \end{pmatrix} \tag{3.76}$$

We have the terms

$$\partial_{\beta_u}\partial_{\beta_v}\gamma_\xi[\beta,\tau_\beta] = \left(\widehat{\rho e^{\beta(\boldsymbol{z})}}\right)_{uv} + \left(\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\right)_{uv} \tag{3.77}$$

$$\partial_{\tau_{\beta_u}}\partial_{\tau_{\beta_v}}\gamma_\xi[\beta,\tau_\beta] = \frac{1}{2}\delta_{uv}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}\delta_z}\mathcal{F}\beta + \left(\frac{1}{\sigma_\beta}\Delta^\dagger\Delta\right)_{uv} \tag{3.78}$$

$$\partial_{\tau_{\beta_u}}\partial_{\beta_v}\gamma_\xi[\beta,\tau_\beta] = -\left(\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}\delta_v}\mathcal{F}\right)_u = \partial_{\beta_v}\partial_{\tau_{\beta_u}}\gamma_\xi[\beta,\tau_\beta] \tag{3.79}$$

Exploiting that for some block matrix we have ([Sil00]):

$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D-CA^{-1}B) = \det(D)\det(A-BD^{-1}C) \tag{3.80}$$

under the condition that both, A and D, are invertible. we get:

$$\det\Gamma_\xi = \left(\left|\frac{\delta^2\gamma}{\delta\beta^\dagger\delta\beta}\right|\left|\frac{\delta^2\gamma}{\delta\tau_\beta^\dagger\delta\tau_\beta} - \frac{\delta^2\gamma}{\delta\tau_\beta^\dagger\delta\beta}\left(\frac{\delta^2\gamma}{\delta\beta^\dagger\delta\beta}\right)^{-1}\frac{\delta^2\gamma}{\delta\beta^\dagger\delta\tau_\beta}\right|\right) \tag{3.81}$$

Derivatives of γ_ζ

Making the definitions

$$\begin{aligned}&\Lambda : \mathbb{R} \to \mathbb{C}^{N\times N}\\ &\Lambda(u)_{ij} = (\mathcal{F}^\dagger \widehat{e^{\tau_f}\delta_u}\mathcal{F})_{x_i x_j}\end{aligned} \tag{3.82}$$

Again we refer to the explicit computation of the derivatives in A.2, only giving the results here For the gradient we get:

$$\partial_{\eta_u}\gamma_\zeta[\eta,\tau_f] = \frac{1}{2}\mathrm{tr}\left(G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\right) - \frac{1}{2}\boldsymbol{y}^\dagger G\widehat{e^{\eta(\boldsymbol{x})}G}\delta_{\boldsymbol{x}u}\boldsymbol{y} - \frac{1}{2}(\boldsymbol{x}) + \left(\frac{1}{\sigma_\eta}\eta^\dagger\nabla^\dagger\nabla\right)_u \tag{3.83}$$

$$\partial_{\tau_{f_u}}\gamma_\zeta[\eta,\tau_f] = \frac{1}{2}\mathrm{tr}\left(G\Lambda_u\right) - \frac{1}{2}\boldsymbol{y}^\dagger G\Lambda_u G\boldsymbol{y} + \left(\frac{1}{\sigma_f}\tau_f^\dagger\Delta^\dagger\Delta\right)_u \tag{3.84}$$

$$\tag{3.85}$$

And for the curvature,

$$\Gamma_\zeta[\eta,\tau_f] = \begin{pmatrix}\frac{\delta^2\gamma_\zeta}{\delta\eta^\dagger\delta\eta} & \frac{\delta^2\gamma_\zeta}{\delta\eta^\dagger\delta\tau_f}\\ \frac{\delta^2\gamma_\zeta}{\delta\tau_f^\dagger\delta\eta} & \frac{\delta^2\gamma_\zeta}{\delta\tau_f^\dagger\delta\tau_f}\end{pmatrix}$$

We have the terms

$$\begin{aligned}\partial_{\eta_u}\partial_{\eta_v}\gamma_\zeta[\eta,\tau_f] &= \frac{1}{2}\mathrm{tr}\left(-G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}} + \delta_{uv}G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}\right)\\ &+ \frac{1}{2}\boldsymbol{y}^\dagger\left(2G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G - G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\delta_{uv}G\right)\boldsymbol{y} + \left(\frac{1}{\sigma_\eta}\nabla^\dagger\nabla\right)_{uv} \end{aligned}\tag{3.86}$$

$$\begin{aligned}\partial_{\tau_{f_u}}\partial_{\tau_{f_v}}\gamma_\zeta[\eta,\tau_f] &= \frac{1}{2}\mathrm{tr}\left(-G\Lambda_u G\Lambda_v + G\Lambda_u\delta_{uv}\right)\\ &+ \frac{1}{2}\boldsymbol{y}^\dagger\left(2G\Lambda_u G\Lambda_v G - G\Lambda_u\delta_{uv}G\right)\boldsymbol{y} + \left(\frac{1}{\sigma_f}\Delta^\dagger\Delta\right)_{uv}\end{aligned}\tag{3.87}$$

$$\partial_{\eta_u}\partial_{\tau_{f_v}}\gamma_\zeta[\eta,\tau_f] = -\frac{1}{2}\mathrm{tr}\left(G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\Lambda_v\right) + \boldsymbol{y}^\dagger G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\Lambda_v G\boldsymbol{y} = \partial_{\tau_{f_v}}\partial_{\eta_u}\gamma_\zeta[\eta,\tau_f] \tag{3.88}$$

Using again 3.80, we get:

$$\det\Gamma_\zeta = \left(\left|\frac{\delta^2\gamma}{\delta\tau_f^\dagger\delta\tau_f}\right|\left|\frac{\delta^2\gamma}{\delta\eta^\dagger\delta\eta} - \frac{\delta^2\gamma}{\delta\eta^\dagger\delta\tau_f}\left(\frac{\delta^2\gamma}{\delta\tau_f^\dagger\delta\tau_f}\right)^{-1}\frac{\delta^2\gamma}{\delta\tau_f^\dagger\delta\eta}\right|\right) \tag{3.89}$$

We do not refrain the explicit terms at this place and refer to A.2 where these are explicitly given.

3.4.6. Bayes Factor for the Deep Model Selection

Plugging in the approximations above we can state the information Hamiltonian for the evidence in a deep model selection, i.e. the only fixed hyperparameters are strength parameters for the smoothness enforcing priors $\sigma_\beta, \sigma_f, \sigma_\eta$:

$$\begin{aligned}\mathcal{H}(\boldsymbol{d}|H_1) \approx& \mathcal{H}_0 + \log \prod_j k_j! + \frac{1}{2} \log \left| \frac{1}{2\pi} \Gamma_\xi[\xi]|_{\xi_0} \right| + \\ &+ \frac{1}{2} \log \left| \frac{1}{2\pi} \Gamma_\zeta[\zeta]|_{\zeta_0} \right| + \gamma_\xi[\xi_0] + \gamma_\zeta[\zeta_0]\end{aligned} \tag{3.90}$$

The model evidence for the reverse direction is again obtained by switching $\boldsymbol{x}$ and $\boldsymbol{y}$ in all terms. The Bayes factor is then given by:

$$\mathcal{O}_{X\to Y} = \exp[-\mathcal{H}(\boldsymbol{d}|H_1) + \mathcal{H}(\boldsymbol{d}|H_{-1})] \tag{3.91}$$

If $\mathcal{O}_{X\to Y} > 1$, the Bayesian Model Selection suggests the causal direction $X \to Y$, if $\mathcal{O}_{X\to Y} < 1$, the other direction, $Y \to X$ is suggested instead. In case of equality, $\mathcal{O}_{X\to Y} = 1$, no direction is preferred. If one admits independence of the variables, i.e. neither $X \to Y$ or $Y \to X$, the Bayes factor being equal to 1 or very close would an indicator for this latter case.

4. Implementation and Benchmarks

4.1. Sampling Causal Data via a Forward Model

To estimate the performance of our algorithm and compare it with other existing approaches, a benchmark dataset is of interest to us. Such benchmark data is usually either real world data, or synthetically produced one. While we will use the *TCEP* benchmark set [Moo+16] in 4.4.5, we also want to use our outlined formalism to generate artificial data representing causal structures. Based on our derivation for cause and effect we implement a forward model to generate data $\boldsymbol{d}$ as following:

Algorithm 1. *Sampling of causal data via forward model*
Input: Power spectra P_β, P_f,
noise variance ς^2, number of bins n_{bins},
approximate * desired number of samples $\tilde{N}$
Output: N samples $(d_i) = (x_i, y_i)$ generated from a causal relation of either $X \to Y$ or $Y \to X$

1. Draw a sample field $\beta \in \mathbb{R}^{[0,1]}$ from the distribution $\mathcal{N}(\beta|0, B)$

2. Set an equally spaced grid with n_{bins} points in the interval $[0,1]$: $\boldsymbol{z} = (z_1, ..., z_{n_{\text{bins}}}), z_i = \frac{i-0.5}{n_{\text{bins}}}$

3. Calculate the vector of Poisson means $\boldsymbol{\lambda} = (\lambda_1, ...\lambda_{n_{\text{bins}}})$ with $\lambda_i \propto e^{\beta(z_i)}$

4. At each grid point $i \in \{1, ..., n_{\text{bins}}\}$, draw a sample k_i from a Poisson distribution with mean λ_i: $k_i \sim \mathcal{P}_{\lambda_i}(k_i)$

5. Set $N = \sum_{i=1}^{n_{\text{bins}}} k_i$,

6. For each $i \in \{1, ..., n_{\text{bins}}\}$ add k_i times the element z_i to the set of measured x_j. Construct the vector $\boldsymbol{x} = (..., \underbrace{z_i, z_i, z_i}_{k_i \text{ times}}, ...)$

7. Draw a sample field $f \in \mathbb{R}^{[0,1]}$ from the distribution $\mathcal{N}(f|0, F)$. Rescale f s.th. $f \in [0,1]^{[0,1]}$.

8. Draw a multivariate noise sample $\boldsymbol{\epsilon} \in \mathbb{R}^N$ from a normal distribution with zero mean

and variance ς^2, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, \varsigma^2)$

9. Generate the effect data $\boldsymbol{y}$ by applying f to $\boldsymbol{x}$ and adding $\boldsymbol{\epsilon}$: $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$)

10. With probability $\frac{1}{2}$ return $\boldsymbol{d} = (\boldsymbol{x}^T, \boldsymbol{y}^T)$, otherwise return $\boldsymbol{d} = (\boldsymbol{y}^T, \boldsymbol{x}^T)$,

[*] As we draw the number of samples from Poisson distribution in each bin, we cannot deterministically control the total number of samples

Comparing the samples for different power spectra (see Fig. 4.1), we decide to sample data with power spectra $P(q) = \frac{1000}{q^4+1}$ and $P(q) = \frac{1000}{q^6+1}$, as these seem to resemble ”natural” mechanisms.

4.2. Implementation of the Bayesian Causal Inference

4.2.1. Implementation of the Shallow Model

Based on our derivation in 3.2 we propose a specific algorithm to decide the causal direction of a given dataset and therefore give detailed answer for problem 1. Basically the task comes down to find the minimum β_0 for the saddle point approximation and calculate the terms given in Eq. 3.36:

Algorithm 2. *2-variable causal inference*
Input: Finite sample data $\boldsymbol{d} \equiv (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{N\times 2}$, Hyperparameters $P_\beta, P_f, \varsigma^2, r$
Output: Predicted causal direction $\mathcal{D}_{X\to Y} \in \{-1, 1\}$ where -1 represents the prediction ”$Y \to X$” and 1 represents $X \to Y$

1. Rescale the data to the $[0, 1]$ interval. I.e. $\min\{x_1, ..., x_N\} = \min\{y_1, ..., y_N\} = 0$ and $\max\{x_1, ..., x_N\} = \max\{y_1, ..., y_N\} = 1$

2. Define an equally spaced grid of $(z_1, ..., z_{n_{\text{bins}}})$ in the interval $[0, 1]$

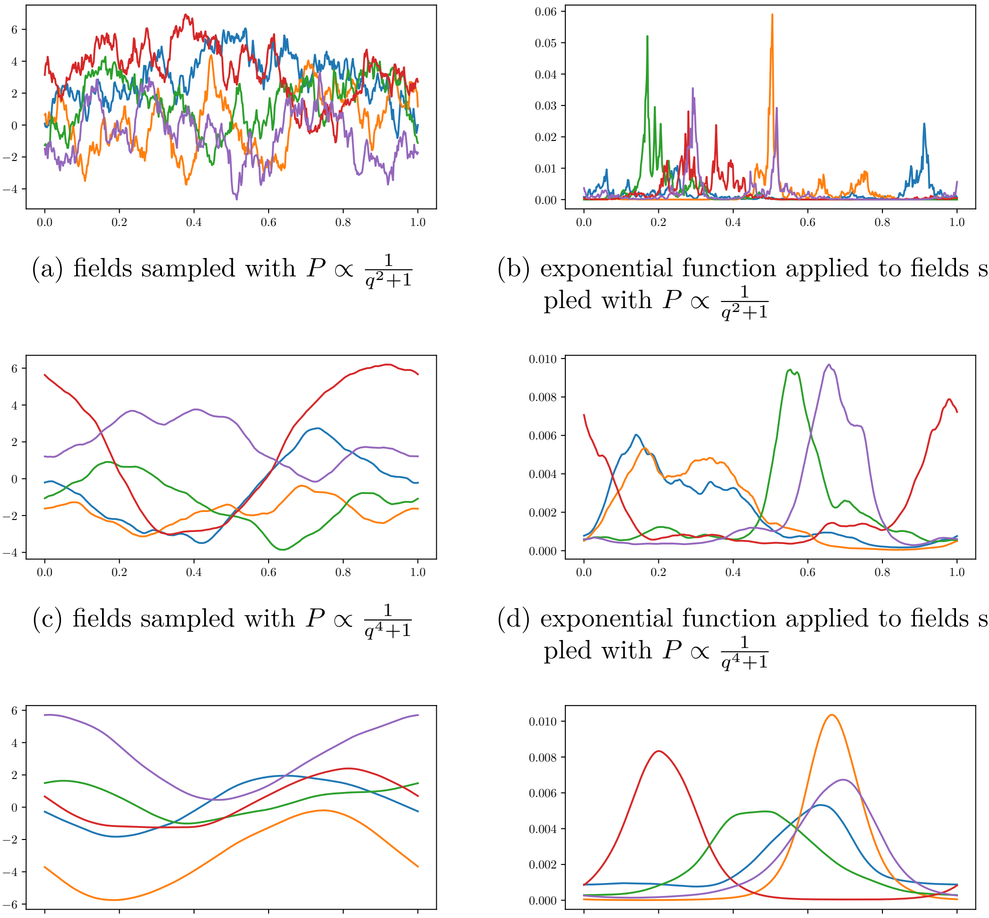
3. Calculate matrices $\boldsymbol{B}, \boldsymbol{F}$ representing the covariance operators B and F evaluated at the positions of the grid, i.e. $\boldsymbol{B}_{ij} = B(z_i, z_j)$

4. Find the $\beta_0 \in \mathbb{R}^{[0,1]}$ s.th. γ in Eq. 3.15 becomes minimal

5. Calculate the $\boldsymbol{d}$-dependent terms of the information Hamiltonian in Eq. 3.36 (i.e. all besides $\mathcal{H}_0$)

6. Repeat steps 4 and 5 with $\boldsymbol{y}$ and $\boldsymbol{x}$ switched

Figure 4.1.: Different sampled fields from $\mathcal{N}(\cdot|0, \mathcal{F}^\dagger P \mathcal{F})$ with the power spectrum $P(q) \propto \frac{1}{q^2+1}$ (top), $P(q) \propto \frac{1}{q^4+1}$ (middle), $P(q) \propto \frac{1}{q^6+1}$ (bottom). On the left, the field values themselves are plotted, on the right an exponential function is applied as in our formulation for $\lambda_j \propto e^{\beta(z_j)}$



(a) fields sampled with $P \propto \frac{1}{q^2+1}$

(b) exponential function applied to fields sampled with $P \propto \frac{1}{q^2+1}$

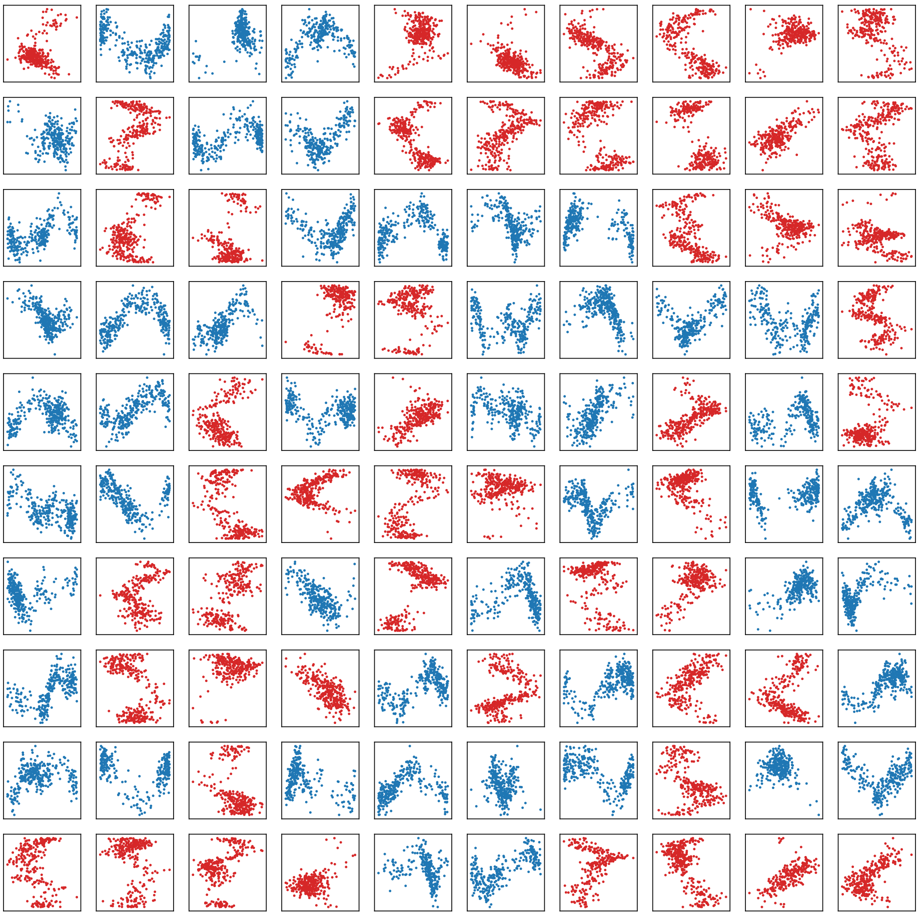
(c) fields sampled with $P \propto \frac{1}{q^4+1}$

(d) exponential function applied to fields sampled with $P \propto \frac{1}{q^4+1}$

(e) fields sampled with $P \propto \frac{1}{q^6+1}$

(f) exponential function applied to fields sampled with $P \propto \frac{1}{q^6+1}$

Figure 4.3.: Synthetic Datasets sampled via alg. 1. Blue scatter plots indicate a true causal direction of $X \to Y$, red scatter plots the direction $Y \to X$



```
7. Calculate the Bayes factor $\mathcal{O}_{X\to Y}$
8. If $\mathcal{O}_{X\to Y} > 1$, return 1, else return $-1$
```

We provide an implementation of alg. 2 in *Python* [2]. We approximate the operators B, F as matrices $\in \mathbb{R}^{n_{\text{bins}} \times n_{\text{bins}}}$, which allows us to explicitly numerically compute the determinants and the inverse. As the most critical part we consider the minimization of β, i.e. step 4 in 2. As we are however able to analytically give the curvature Γ_β and the gradient $\partial_\beta \gamma_\beta$, we can use a Newton-scheme here. We derive satisfying results (see Fig. 4.4) using the *Newton-CG* algorithm [NW06], provided by the *SciPy*-Library [J+01]. After testing our algorithm on different benchmark data, we choose the default hyperparameters as

$$P_\beta = P_f \propto \frac{1}{q^4 + 1}, \tag{4.1}$$

$$\varsigma^2 = 0.01, \tag{4.2}$$

$$r = 512, \tag{4.3}$$

$$\rho = 1. \tag{4.4}$$

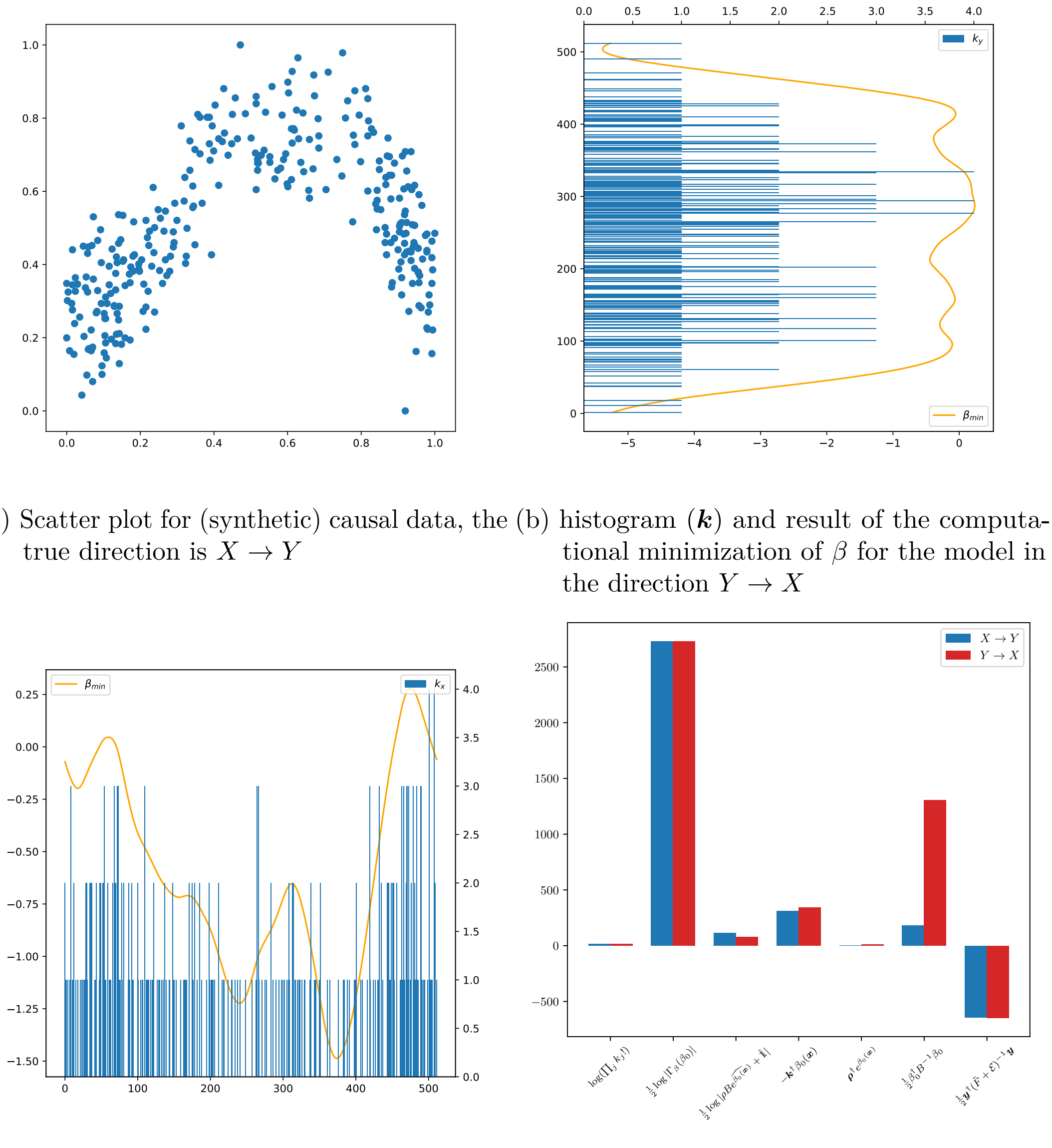
While fixing the power spectra might seem somewhat arbitrary, we remark that this corresponds to fixing a kernel e.g. as a squared exponential kernel, which is done in many publications (e.g. [MST18; Gou+17])

4.2.2. Issues with the Implementation of the Deeper Models

We tested the possibility to implement the deeper models. In the noise inference model, as described in 3.3, we need to go through two separate minimization phases. While the first minimization is the same one as 4 in 2, the second one, i.e. determining η_0, involves the explicit numerical inversion of matrices $\in \mathbb{R}^{n_{\text{bins}} \times n_{\text{bins}}}$ which depend on η when calculating the curvature $\Gamma_\eta[\eta]$. Using again a Newton-scheme which involves several computations of the curvature in each minimization step, this minimization becomes therefore vastly computationally expensive and such too slow to handle it efficiently.

Implementing the ”deep” model as outlined in 3.4 would now involve finding numerical representations for $\beta, \tau_\beta, \eta, \tau_f$. Besides the computational complexity which would behave even worse than mentioned above, one now has the difficulty that the minima β_0 and τ_β depend on each other, i.e. a change in τ_β will lead to a different β_0. The same holds for η and τ_f. We therefore leave these models at a theoretical state here and consider their implementation as an option for future work to be done.

[2] `https://github.com/MKurthen/BayesianCausalInference`



(a) Scatter plot for (synthetic) causal data, the true direction is $X \to Y$

(b) histogram ($\boldsymbol{k}$) and result of the computational minimization of β for the model in the direction $Y \to X$

(c) histogram ($\boldsymbol{k}$) and result of the computational minimization of β for the model in the direction $X \to Y$

(d) Numerical values for terms in $\mathcal{H}(\boldsymbol{d}|P_\beta, P_f, \varsigma, X \to Y)$ and $\mathcal{H}(\boldsymbol{d}|P_\beta, P_f, \varsigma, Y \to X)$

Figure 4.4.: Demonstration of the computational minimization

4.3. Methods for Comparison

We compare our outlined model, in the following called BCM, short for *Bayesian Causal Model*, to a number of state-of-the-art approaches. The selection of the considered methods is influenced by the ones in recent publications, e.g. [MST18; Gou+17].

As it is one of the oldest algorithms in the field and has been used as comparison in a variety of publications, we include the **LiNGAM** algorithm. We also use the ANM algorithm [Moo+16] with HSIC and Gaussian Process Regression (**ANM-HSIC**)

The **ANM-MML** approach [Ste+10] uses a Bayesian Model Selection, from the perspective of formulation it is the closest to the algorithm proposed within this thesis, at least to our best knowledge. This makes it an interesting reference and motivates the choice to include it here.

We further include the **IGCI** algorithm, as it differs fundamentally in its formulation from the ANM algorithms and has shown strong results in recent publications [Moo+16; Gou+17; MST18]. We employ the IGCI algorithm with entropy estimation for scoring and a Gaussian distribution as reference distribution.

Finally, **CGNN** [Gou+17] represents the rather novel influence of deep learning methods. As it proved to perform generally well in different scenarios [Gou+17; MST18], we include it in our comparison. We use the implementation provided by the authors, with itself uses Python with the *Tensorflow*[Aba+15] library. The most critical hyper-parameter here is, as the authors themselves mention, the number of hidden neurons. We set this number to a value of $n_h = 30$, as this is the default in the given implementation and delivers generally good results. We use 8 runs each, in our eyes this represents a adequate trade-off between unnecessary high computation time and bad performance for reasons of being too restrictive.

A comparison with the KCDC algorithm would be interesting, unfortunately the authors did not provide any computational implementation so far (September 2018).

4.4. Results

4.4.1. Default Synthetic Data

We arrive at the conclusion to choose spectra of the type $P(q) = \frac{1}{q^4+1}$, for both spectra. We further set n_{bins}=512, $\tilde{N}$=300 and ς^2=0.05 as default settings for producing synthetic causal data. The scatter plots for the resulting data are shown in 4.3.

The resulting accuracies are given below:

Table 4.1.: Accuracy for the default synthetic data benchmark.

Model	accuracy
BCM	0.98
LiNGAM	0.30
ANM-HSIC	1.00
ANM-MML	1.00
IGCI	0.65
CGNN	0.72

4.4.2. High Noise Data

As a first variation, we explore the influence of high and very high noise on the performance of the inference models. Therefore we set the parameter ς^2=0.2 for high noise and ς^2=1 for very high noise in 1, while keeping the other parameters set to the same values as in 4.4.1. The scatter plots for the resulting datasets are given in figures C.1 and fig C.2.

While our BCM algorithm is able to still perform somewhat reliable with a accuracy of $\geq 90\%$, especially the ANM algorithms do not remarkably suffer from the noise. This is likely due to the fact that the distribution of the true cause $\mathcal{P}(X)$ is not influenced by the high noise and this distribution is assessed in its own.

Table 4.2.: Accuracy for the high-noise synthetic data benchmark.

Model	ς^2=0.2	ς^2=1
BCM	0.94	0.90
LiNGAM	0.31	0.40
ANM-HSIC	0.98	0.94
ANM-MML	0.99	0.99
IGCI	0.60	0.58
CGNN	0.75	0.77

4.4.3. Strongly Discretized Data

As our model uses a Poissonian approach, which explicitly considers discretization effects of data measurement, it is of interest how the performance behaves when using a strongly discretization. We emulate such a situation by employing our forward model 1 with a very low number of bins. We keep all parameters as in 4.4.1 and set n_{bins}=16 and n_{bins}=8 for synthetic data with high and very high discretization. The visualization of the datasets is given in figures C.3 and C.4. Apparently, especially the ANM models do not suffer anyhow from the strong discretization. CGNN and IGCI perform significantly worse here. In the case of IGCI this can be explained by the entropy estimation, which simply removes non-unique samples. Our BCM algorithm is able to achieve over 90% accuracy here.

Table 4.3.: Accuracy for the strongly discretized synthetic benchmark data

Model	n_{bins}=16	n_{bins}=8
BCM	0.93	0.97
LiNGAM	0.23	0.21
ANM-HSIC	0.99	1.00
ANM-MML	1.00	1.00
IGCI	0.24	0.09
CGNN	0.57	0.22

4.4.4. Very Sparse Data

We explore another challenge for inference algorithms where we strongly reduce the number of samples. While we sampled about 300 observations with our other forward models so far, here we reduce the number of observed samples to 30 and 10 samples. Again we refer to the scatter plots in figures C.5 and C.6. In this case our model performs very well compared to the other models, in fact it is able to outperform them in the case of just 10 samples being given.

We note that of course our model does have the advantage that it ”knows” the hyperparameters of the underlying forward model. Yet we consider the results as encouraging.

Table 4.4.: Accuracy for very sparse data.

Model	30 samples	10 samples
BCM	0.92	0.75
LiNGAM	0.44	0.45
ANM-HSIC	0.91	0.71
ANM-MML	0.98	0.69
IGCI	0.48	0.40
CGNN	0.46	0.39

4.4.5. Real World Data

The most widely used benchmark set with real world data is the *Tübingen Cause Effect Pairs* dataset (TCEP) [Moo+16]. The collection currently (August 2018) consists of 108 datasets. However these include sets with multiple cause or effect variables. Excluding these reduced the collection to 102 datasets. As proposed by the maintainers, we use a weighted evaluation of accuracy here. As some of the datasets represent essentially the same mechanism and just have been collected with different means (e.g. ”latitude” and ”life expectancy at birth for different countries, female, 1995-2000” in `pair0057` vs. ”latitude” and life expectancy at birth for different countries, female, 1990-1995” in `pair0058`) they are assigned a reduced weight. A full description of the TCEP benchmark set is given in D, accompanied by scatter plots of the datasets in Fig. D.1.

As some of the contained datasets include a high number of samples (up to 11000) we randomly subsample large datasets to 500 samples each in order to keep computation time maintainable. We did not include the *LiNGAM* algorithm here, as we experienced computational problems with obtaining results here for certain datasets (namely `pair0098`). [Gou+17] report the accuracy of *LiNGAM* on the *TCEP* dataset at around 40%. Our model shows to perform generally comparable to established approaches as *ANM* and *IGCI*. *CGNN* performs best here, the accuracy which we evaluate at about 70% is however lower than the one reported by the authors [Gou+17] at around 80%. The reason for this is arguably to be found in the fact that we set all hyperparameters to fixed values, while [Gou+17] used a leave-one-out-approach to find the best setting for the hyperparameter n_h.

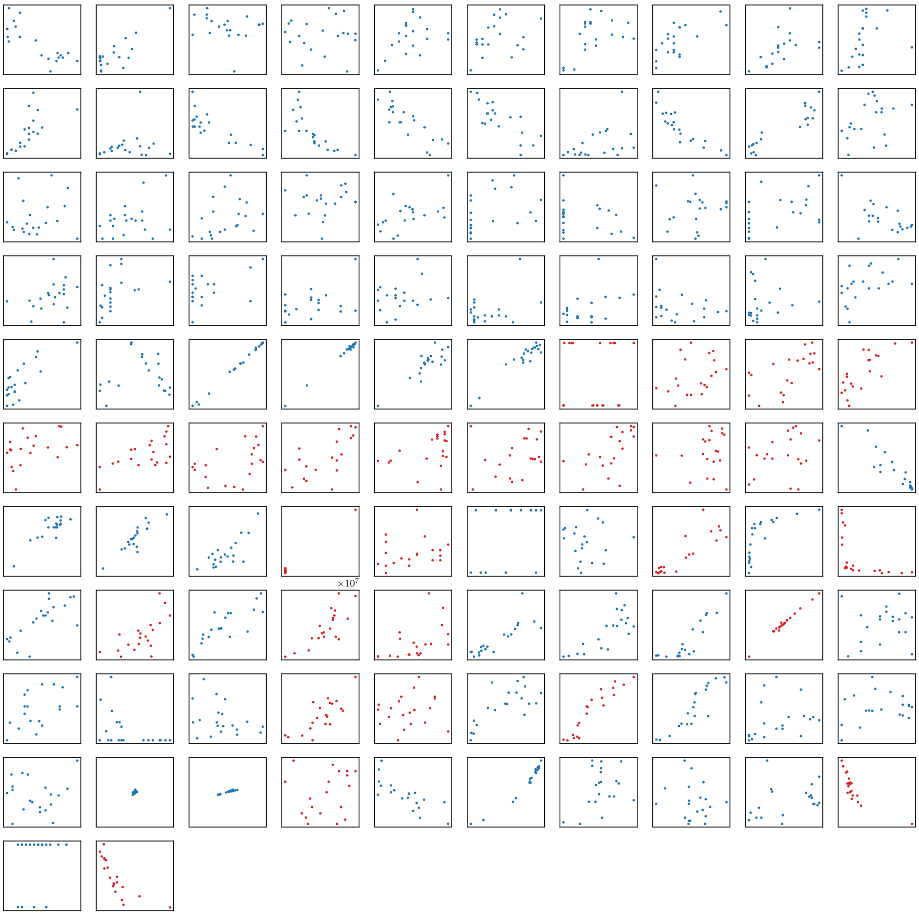
Motivated by the generally strong performance of our approach in the case of sparse data, we also explore a situation where real world data is only sparsely available. To that end, we subsample the TCEP datasets s.th. each 20 randomly chosen samples are kept. We give visual impression for this data in Fig. 4.5 The results are as well given in Table 4.5 The loss in accuracy of our model is remarkably small. In fact, *BCM* is able to outperform the other models considered here, even if not by a large margin.

Table 4.5.: Accuracy for TCEP Benchmark

Model	TCEP	TCEP with 20 samples
BCM	0.64	0.60
ANM-HSIC	0.63	0.41
ANM-MML	0.58	0.51
IGCI	0.66	0.59
CGNN	0.70	0.55

Figure 4.5.: Real World Data TCEP benchmark set with 20 randomly chosen samples per dataset. Blue scatter plots indicate a true causal direction of $X \to Y$, red scatter plots the direction $Y \to X$

.



5. Discussion and Conclusion

In this thesis we introduced a model for the 2-variable causal inference task. Our model builds on the formalism of information field theory which explicitly models the connection of finite dimensional measurement data to underlying infinite dimensional structures. In this regard, we employed the concept of Bayesian model selection and made the assumption of additive noise, i.e. $\boldsymbol{x} = f(\boldsymbol{y}) + \boldsymbol{\epsilon}$. In contrast to other methods which do so, such as ANM-MML, we do not model the cause distribution by a Gaussian mixture model but by a Poisson Lognormal statistic.

We could show that our model is able to provide reliable classification accuracy in the present causal inference task. Another difference from our model to existing ones is arguably to be found in the choice of the covariance operators. While most other publications use squared exponential kernels for Gaussian process regression, we choose a covariance which is governed by a $\frac{1}{q^4+1}$ power spectrum. This leads arguably to a different importance of structure at small scales which is emphasized more strongly by our covariance than in a squared exponential kernel.

As a certain weak point within our model we consider the approximation of the uncomputable path integrals via the Laplace approximation. A thorough investigation of error bounds, e.g. via [Maj15] is yet to be carried out. As an alternative, one can think about sampling-based approaches to approximate the integrals. A recent publication ([Cal+18]) introduced a harmonic-mean based sampling approach to approximate high dimensional integrals. Such a technique might be promising in the context of our formalism.

Another outlook is to be seen in the computational implementation of the deeper models discussed in 3.3 and 3.4. However the feasibility of this is certainly questionable. Especially with the outlined technique of the Laplace approximation the challenge persists in the computational minimization which is numerically highly complex in this case.

Yet, the implementation of our model with fixed noise variance and power spectra was able to deliver competitive results with regard to state-of-the-art methods in the benchmarks. In particular, our method seems to be slightly superior in the low sample regime, probably due to the more appropriate Poisson statistic used. We consider this as an encouraging result for a first work in the context of information field theory-based causal inference.

Bibliography

[Aba+15] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. 2015. URL: `https://www.tensorflow.org/`.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Berlin, Heidelberg: Springer-Verlag, 2006, pp. 225–231. ISBN: 0387310738.

[BS09] Jose M. Bernardo and A.F.M. Smith. *Bayesian Theory.* Wiley Series in Probability and Statistics. Wiley, 2009, pp. 389–401. ISBN: 9780470317716. URL: `https://books.google.de/books?id=11nSgIcd7xQC`.

[Cal+18] Allen Caldwell et al. "Integration with an Adaptive Harmonic Mean Algorithm". In: *ArXiv e-prints* (Aug. 2018). arXiv: `1808.08051`.

[Cha16] Christopher Chatfield. *The Analysis of Time Series: An Introduction, Sixth Edition.* Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2016, pp. 109–114. ISBN: 9780203491683. URL: `https://books.google.de/books?id=qKzyAbdaDFAC`.

[Dan+10] P. Daniusis et al. "Inferring deterministic causal relations". In: *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence.* Max-Planck-Gesellschaft. Corvallis, OR, USA: AUAI Press, July 2010, pp. 143–150.

[EFK09] Torsten A. Enßlin, Mona Frommert, and Francisco S. Kitaura. "Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis". In: *Phys. Rev. D* 80 (10 Nov. 2009), p. 105005. DOI: `10.1103/PhysRevD.80.105005`. URL: `https://link.aps.org/doi/10.1103/PhysRevD.80.105005`.

[GBR13] W. Greiner, D.A. Bromley, and J. Reinhardt. *Field Quantization.* Springer Berlin Heidelberg, 2013, p. 353. ISBN: 9783642614859. URL: `https://books.google.de/books?id=C-DVBAAAQBAJ`.

[Gou+17] Olivier Goudet et al. *Learning Functional Causal Models with Generative Neural Networks.* 2017. eprint: `arXiv:1709.05321`.

[Gre+05] Arthur Gretton et al. "Measuring Statistical Dependence with Hilbert-schmidt Norms". In: *Proceedings of the 16th International Conference on Algorithmic Learning Theory.* ALT'05. Singapore: Springer-Verlag, 2005, pp. 63–77.

[Gre+07] Arthur Gretton et al. "A Kernel Method for the Two-Sample-Problem". In: *Advances in Neural Information Processing Systems 19.* Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, 2007, pp. 513–520.

[GV08] P.D. Grünwald and P.M.B. Vitányi. "Algorithmic information theory". In: *ArXiv e-prints* (Sept. 2008). arXiv: `0809.2754 [cs.IT]`.

[Har42] R.V.L. Hartley. “A More Symmetrical Fourier Analysis Applied to Transmission Problems”. In: 30 (Apr. 1942), pp. 144–150.

[HHH18] Miguel A. Hernán, John Hsu, and Brian Healy. *Data science is science's second chance to get causal inference right: A classification of data science tasks.* 2018. eprint: `arXiv:1804.10846`.

[Hoy+09] Patrik O. Hoyer et al. “Nonlinear causal discovery with additive noise models”. In: *Advances in Neural Information Processing Systems 21.* Ed. by D. Koller et al. Curran Associates, Inc., 2009, pp. 689–696. URL: `http://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf`.

[J+01] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python.* 2001. URL: `http://www.scipy.org/`.

[Maj15] Piotr Majerski. “Simple error bounds for the multivariate Laplace approximation under weak local assumptions”. In: *arXiv preprint arXiv:1511.00302* (2015).

[Mat00] Robert Matthews. “Storks deliver babies (p= 0.008)”. In: *Teaching Statistics* 22.2 (2000), pp. 36–38.

[Moo+16] Joris M. Mooij et al. “Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks”. In: *Journal of Machine Learning Research* 17.32 (2016), pp. 1–102. URL: `http://jmlr.org/papers/v17/14-518.html`.

[MST18] Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. “Causal Inference via Kernel Deviance Measures”. In: *arXiv preprint arXiv:1804.04622* (2018).

[NW06] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization.* second. New York, NY, USA: Springer, 2006.

[Pea00] Judea Pearl. *Causality: Models, Reasoning, and Inference.* New York, NY, USA: Cambridge University Press, 2000, pp. 44, 70. ISBN: 0-521-77362-8.

[Pea18] Judea Pearl. “Theoretical impediments to machine learning with seven sparks from the causal revolution”. In: *arXiv preprint arXiv:1801.04016* (2018).

[PJS17] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms.* Cambridge, MA, USA: MIT Press, 2017.

[RW06] CE. Rasmussen and CKI. Williams. *Gaussian Processes for Machine Learning.* Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, Jan. 2006, pp. 13, 248.

[Shi+06] Shohei Shimizu et al. “A linear non-Gaussian acyclic model for causal discovery”. In: *Journal of Machine Learning Research* 7.Oct (2006), pp. 2003–2030.

[Sil00] John R Silvester. “Determinants of block matrices”. In: *The Mathematical Gazette* 84.501 (2000), pp. 460–467.

[SN10] Christopher S. Withers and Saralees Nadarajah. “log det A = tr log A”. In: 41 (Dec. 2010), pp. 1121–1124.

[Spi16] Rani Spiegler. “Can agents with causal misperceptions be systematically fooled?” In: (2016).

[ST05] TH Sparks and P Tryjanowski. “The detection of climate impacts: some methodological considerations”. In: *International Journal of Climatology* 25.2 (2005), pp. 271–277.

[Ste+10] Oliver Stegle et al. “Probabilistic latent variable models for distinguishing between cause and effect”. In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty et al. Curran Associates, Inc., 2010, pp. 1687–1695. URL: `http://papers.nips.cc/paper/4173-probabilistic-latent-variable-models-for-distinguishing-between-cause-and-effect.pdf`.

[SZ16] Peter Spirtes and Kun Zhang. “Causal discovery and inference: concepts and recent methodological advances”. In: *Applied Informatics* 3.1 (Feb. 2016), p. 3. ISSN: 2196-0089. DOI: `10.1186/s40535-016-0018-x`. URL: `https://doi.org/10.1186/s40535-016-0018-x`.

A. Explicit Calculations

A.1. Derivatives of γ_ξ

A.1.1. The Term $\log\left|2\pi\widehat{e^{\tau_\beta}}\right|$

$$\begin{aligned}
\partial_{\tau_{\beta_q}} \log\left|2\pi\widehat{e^{\tau_\beta}}\right| &= \left|2\pi\widehat{e^{\tau_\beta}}\right|^{-1} \partial_{\tau_{\beta_q}} \left|2\pi\widehat{e^{\tau_\beta}}\right| \\
&= \left|2\pi\widehat{e^{\tau_\beta}}\right|^{-1} \left|2\pi\widehat{e^{\tau_\beta}}\right| \operatorname{tr}\left(\widehat{e^{-\tau_\beta}} \partial_{\tau_{\beta_q}} \widehat{e^{\tau_\beta}}\right) \\
&= \operatorname{tr}\left(\widehat{e^{-\tau_\beta}} \widehat{e^{\tau_\beta}\delta_q}\right) \\
&= \operatorname{tr}(\widehat{\delta_q}) = 1
\end{aligned} \tag{A.1}$$

The second order derivative w.r.t. τ_β therefore vanishes,

$$\partial_{\tau_{\beta_r}} \partial_{\tau_{\beta_q}} \log\left|2\pi\widehat{e^{\tau_\beta}}\right| = 0 \tag{A.2}$$

As well as the second order mixed derivatives,

$$\partial_{\beta_u} \partial_{\tau_{\beta_q}} \log\left|2\pi\widehat{e^{\tau_\beta}}\right| = 0 = \partial_{\tau_{\beta_q}} \partial_{\beta_u} \log\left|2\pi\widehat{e^{\tau_\beta}}\right| \tag{A.3}$$

A.1.2. The Term $\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta$

As first order derivatives we have

$$\partial_{\beta_u}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta = (\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F})_u \tag{A.4}$$

$$\partial_{\tau_{\beta_q}}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta = \beta^\dagger\mathcal{F}^\dagger(\partial_{\tau_{\beta_q}}\widehat{e^{-\tau_\beta}})\mathcal{F}\beta = -\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}\delta_q}\mathcal{F}\beta \tag{A.5}$$

The above derivative w.r.t. to β is trivial, in the derivative w.r.t. τ_β we used that the Fourier transform is invariant under the derivative. The resulting expression $\widehat{e^{-\tau_\beta}\delta_q}$ is to be understood as the field analogue of a diagonal matrix where only a single entry (with the index q) is not zero. Therefore the second order derivatives are:

$$\partial_{\beta_v}\partial_{\beta_u}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta = (\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F})_{uv} \tag{A.6}$$

$$\partial_{\tau_{\beta_r}}\partial_{\tau_{\beta_q}}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta = \delta_{qr}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}\delta_q}\mathcal{F}\beta \tag{A.7}$$

$$\partial_{\tau_{\beta_q}}\partial_{\beta_u}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\beta = \partial_{\tau_{\beta_q}}(\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F})_u = -(\beta^\dagger\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}\delta_q}\mathcal{F})_u \tag{A.8}$$

A.2. Derivatives of γ_ζ

A.2.1. The Term $\boldsymbol{y}^\dagger(\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}\boldsymbol{y}$

The derivatives of the τ_f-dependent matrix elements $\tilde{F}_{ij}$ are

$$\partial_{\tau_{f_q}} \tilde{F}[\tau_f]_{ij} = \partial_{\tau_{f_q}} (\mathcal{F}^\dagger \widehat{e^{\tau_f}} \mathcal{F})_{x_i x_j} = (\mathcal{F}^\dagger \widehat{e^{\tau_f} \delta_q} \mathcal{F})_{x_i x_j} \equiv (\Lambda_q)_{ij} \tag{A.9}$$

$$\partial_{\tau_{f_r}} \partial_{\tau_{f_q}} \tilde{F}[\tau_f]_{ij} = \partial_{\tau_{f_r}} (\mathcal{F}^\dagger \widehat{e^{\tau_f} \delta_q} \mathcal{F})_{x_i x_j} = (\Lambda_q)_{ij} \delta_{qr} \tag{A.10}$$

For the sake of brevity we introduced $(\Lambda_q)_{ij} \equiv (\mathcal{F}^\dagger \widehat{e^{\tau_f} \delta_q} \mathcal{F})_{x_i x_j}$ above, which we will use from now on. We will further write $G \equiv (\tilde{F} + \widehat{e^{\eta(\boldsymbol{x})}})^{-1}$

$$\partial_{\tau_{f_q}} \boldsymbol{y}^\dagger G \boldsymbol{y} = -\boldsymbol{y}^\dagger G \Lambda_q G \boldsymbol{y} \tag{A.11}$$

$$\partial_{\tau_{f_r}} \partial_{\tau_{f_q}} \boldsymbol{y}^\dagger G \boldsymbol{y} = \boldsymbol{y}^\dagger \left(2G\Lambda_q G \Lambda_r G - G \Lambda_q \delta_{qr} G \right) \boldsymbol{y} \tag{A.12}$$

$$\partial_{\eta_u} \boldsymbol{y}^\dagger G \boldsymbol{y} = -\boldsymbol{y}^\dagger G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} G \boldsymbol{y} \tag{A.13}$$

$$\tag{A.14}$$

$$\partial_{\eta_v} \partial_{\eta_u} \boldsymbol{y}^\dagger G \boldsymbol{y} = \boldsymbol{y}^\dagger \left(2G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} G \widehat{e^{\eta(\boldsymbol{x})}} \delta_{\boldsymbol{x}r} G - G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} \delta_{uv} G \right) \boldsymbol{y} \tag{A.15}$$

$$\partial_{\eta_u} \partial_{\tau_{f_q}} \boldsymbol{y}^\dagger G \boldsymbol{y} = -2\boldsymbol{y}^\dagger (\partial_{\eta_u} G) \Lambda_q G \boldsymbol{y} = 2\boldsymbol{y}^\dagger G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} G \Lambda_q G \boldsymbol{y} \tag{A.16}$$

$$= \partial_{\tau_{f_q}} \partial_{\eta_u} \boldsymbol{y}^\dagger G \boldsymbol{y} \tag{A.17}$$

A.2.2. The Term $\log \left| \tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}} \right|$

$$\partial_{\tau_{f_q}} \log \left| \tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}} \right| = \operatorname{tr} \left(G \, \partial_{\tau_{f_q}} (\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}}) \right) = \operatorname{tr} \left(G \Lambda_z \right) \tag{A.18}$$

$$\partial_{\tau_{f_r}} \partial_{\tau_{f_q}} \log \left| \tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}} \right| = \operatorname{tr} \left((\partial_{\tau_{f_r}} G) \Lambda_{z'}) + G(\partial_{\tau_{f_r}} \Lambda_q) \right) = \operatorname{tr} \left(-G\Lambda_r G \Lambda_q + G \Lambda_r \delta_{qr} \right) \tag{A.19}$$

$$\partial_{\eta_u} \log \left| \tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}} \right| = \operatorname{tr} \left(G \, \partial_{\eta_u} (\tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}}) \right) = \operatorname{tr} \left(G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} \right) \tag{A.20}$$

$$\begin{aligned} \partial_{\eta_v} \partial_{\eta_u} \log \left| \tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}} \right| &= \operatorname{tr} \left((\partial_{\eta_v} G) \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} + G(\partial_{\eta_v} \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}}) \right) \\ &= \operatorname{tr} \left(-G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}v}} G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} + \delta_{vu} G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} \right) \end{aligned} \tag{A.21}$$

$$\partial_{\eta_u} \partial_{\tau_{f_q}} \log \left| \tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}} \right| = \operatorname{tr} \left((\partial_{\eta_u} G) \Lambda_q) \right) = \operatorname{tr} \left(-G \widehat{e^{\eta(\boldsymbol{x})} \delta_{\boldsymbol{x}u}} G \Lambda_q \right) \tag{A.22}$$

$$= \partial_{\tau_{f_q}} \partial_{\eta_u} \log \left| \tilde{F}[\tau_f] + \widehat{e^{\eta(\boldsymbol{x})}} \right| \tag{A.23}$$

A.3. The Determinant $|\Gamma_\xi|$

The determinant for the curvature of γ_ξ is given by

$$\begin{aligned}
|\Gamma_\xi(\beta,\tau_\beta)| =& \left|\frac{\delta^2\gamma}{\delta\beta^\dagger\delta\beta}\right| \left|\frac{\delta^2\gamma}{\delta\tau_\beta^\dagger\delta\tau_\beta} - \frac{\delta^2\gamma}{\delta\tau_\beta^\dagger\delta\beta}\left(\frac{\delta^2\gamma}{\delta\beta^\dagger\delta\beta}\right)^{-1}\frac{\delta^2\gamma}{\delta\beta^\dagger\delta\tau_\beta}\right| \\
=& \det_{uv}\left[\widehat{\rho e^{\beta(\boldsymbol{z})}}_{uv} + \left(\mathcal{F}^\dagger \widehat{e^{-\tau_\beta}}\mathcal{F}\right)_{uv}\right] \det_{uv}\left[\frac{1}{2}\delta_{uv}\beta^\dagger\mathcal{F}^\dagger\widehat{e^{\tau_\beta}\delta_x}\mathcal{F}\beta + \left(\frac{1}{\sigma_\beta}\Delta^\dagger\Delta\right)_{uv}\right. \\
& \left. - \left(\beta^\dagger\mathcal{F}^\dagger\widehat{e^{\tau_\beta}\delta_u}\mathcal{F}\right)_w \left(\widehat{\rho e^{\beta(\boldsymbol{z})}} + \left(\mathcal{F}^\dagger\widehat{e^{-\tau_\beta}}\mathcal{F}\right)\right)^{-1}_{wz} \left(\beta^\dagger\mathcal{F}^\dagger\widehat{e^{\tau_\beta}\delta_v}\mathcal{F}\right)_z\right] \qquad (A.24)
\end{aligned}$$

A.4. The Determinant $|\Gamma_\zeta|$

The full expression for the determinant of the curvature of γ_ζ is:

$$\begin{aligned}
|\Gamma_\zeta(\eta,\tau_f|) =& \left(\left|\frac{\delta^2\gamma}{\delta\tau_f^\dagger\delta\tau_f}\right| \left|\frac{\delta^2\gamma}{\delta\eta^\dagger\delta\eta} - \frac{\delta^2\gamma}{\delta\eta^\dagger\delta\tau_f}\left(\frac{\delta^2\gamma}{\delta\tau_f^\dagger\delta\tau_f}\right)^{-1}\frac{\delta^2\gamma}{\delta\tau_f^\dagger\delta\eta}\right|\right) \\
=& \det_{uv}\left[\frac{1}{2}\mathrm{tr}\left(-G\Lambda_x G\Lambda_y + G\Lambda_x\delta_{uv}\right) + \frac{1}{2}\boldsymbol{y}^\dagger\left(2G\Lambda_u G\Lambda_v G - G\Lambda_u\delta_{uv}G\right)\boldsymbol{y} + \left(\frac{1}{\sigma_f}\Delta^\dagger\Delta\right)_{uv}\right] \\
& \times \det_{uv}\left[\frac{1}{2}\mathrm{tr}\left(-G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}} + \delta_{uv}G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}\right)\right. \\
& \quad + \frac{1}{2}\boldsymbol{y}^\dagger\left(2G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G - G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}\delta_{uv}G\right)\boldsymbol{y} + \left(\frac{1}{\sigma_\eta}\nabla^\dagger\nabla\right)_{uv} \\
& \quad - \left(-\frac{1}{2}\mathrm{tr}\left(G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\Lambda_w\right) + \boldsymbol{y}^\dagger G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}u}}G\Lambda_w G\boldsymbol{y}\right) \times \\
& \times \left(\frac{1}{2}\mathrm{tr}\left(-G\Lambda_w G\Lambda_z + G\Lambda_w\delta_{wz}\right) + \frac{1}{2}\boldsymbol{y}^\dagger\left(2G\Lambda_w G\Lambda_z G - G\Lambda_w\delta_{wz}\right)\boldsymbol{y} + \left(\frac{1}{\sigma_f}\Delta^\dagger\Delta\right)_{wz}\right)^{-1} \\
& \left. \times \left(-\frac{1}{2}\mathrm{tr}\left(G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G\Lambda_z\right) + \boldsymbol{y}^\dagger G\widehat{e^{\eta(\boldsymbol{x})}\delta_{\boldsymbol{x}v}}G\Lambda_z G\boldsymbol{y}\right)\right] \qquad (A.25)
\end{aligned}$$

B. Details on the Computational Implementation

The implementation of algorithm 2 was done in *Python* [1] Using the notation of 3, we modelled $\boldsymbol{x}, \boldsymbol{y}$, as 1d-Numpy arrays with a length of n_{bins}.

The operators B, F are represented as 2d-Numpy arrays with a shape of $(n_{\text{bins}}, n_{\text{bins}})$, i.e. matrices in $\mathbb{R}^{n_{\text{bins}} \times n_{\text{bins}}}$. Using a Fourier transform (i.e. a Discrete Fourier transform) to represent the B and F would lead to the question how to handle the resulting imaginary parts. Instead we choose to use the *Hartley* transform ([Har42]) which is defined as

$$\mathcal{H}[f](q) = \frac{1}{2\pi} \int \mathrm{d}x (sin(xq) + cos(xq)) f(x) \tag{B.1}$$

Therefore, $\mathcal{H}[f] = \mathfrak{Re}(\mathcal{F}[f]) - \mathfrak{Im}(\mathcal{F}[f])$, and further $\mathcal{H}^{-1} = \mathcal{H}$. We thus implement the Hartley transform by calculating the discrete Hartley transform matrix, $\mathbb{H} = \mathfrak{Re}(\mathbb{F}) - \mathfrak{Im}(\mathbb{F})$. Here, $\mathbb{F}$ denotes the discrete Fourier transform (DFT) matrix, for which we use an implementation provided by the SciPy Module. We can now give numerical representations for B and F via $\mathbb{H}\widehat{P_\beta}\mathbb{H}$ and $\mathbb{H}\widehat{P_f}\mathbb{H}$.

[1] provided at `https://github.com/MKurthen/BayesianCausalInference`

C. Scatter Plots for the Benchmark Data

Figure C.1.: Synthetic Datasets sampled via alg. 1 with $\varsigma^2 = 0.2$. Blue scatter plots indicate a true causal direction of $X \to Y$, red scatter plots the direction $Y \to X$.

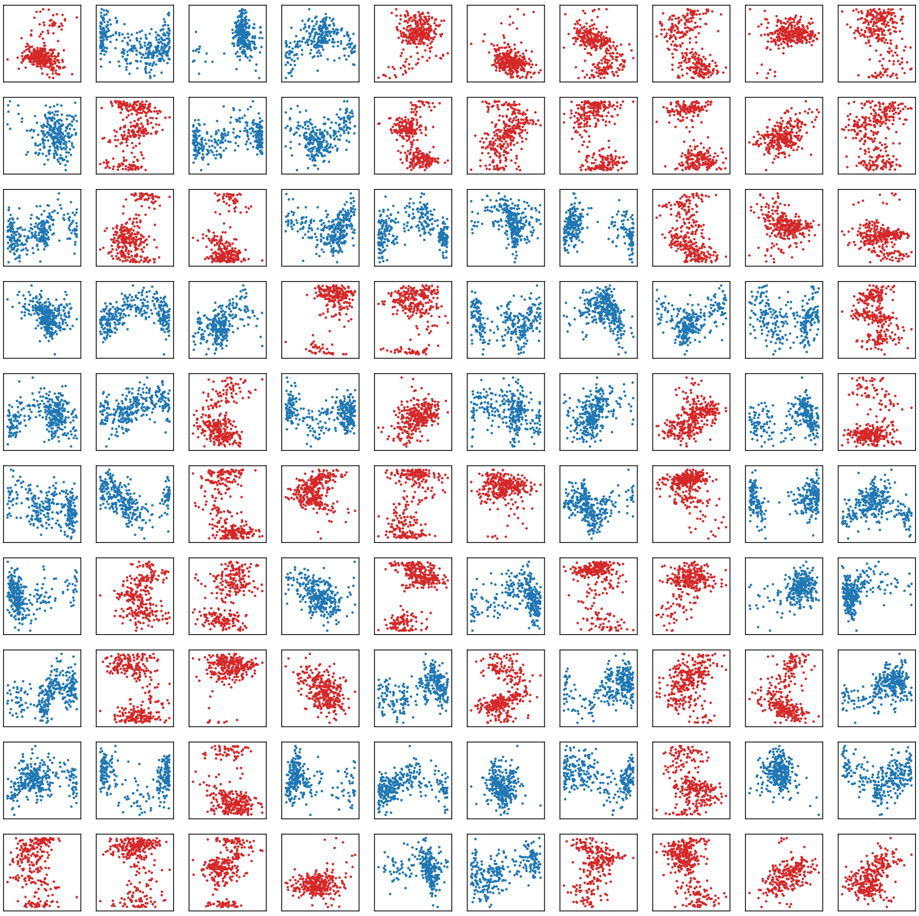


Figure C.2.: Synthetic Datasets sampled via alg. 1 with $\varsigma^2 = 1.0$. Blue scatter plots indicate a true causal direction of $X \rightarrow Y$, red scatter plots the direction $Y \rightarrow X$.

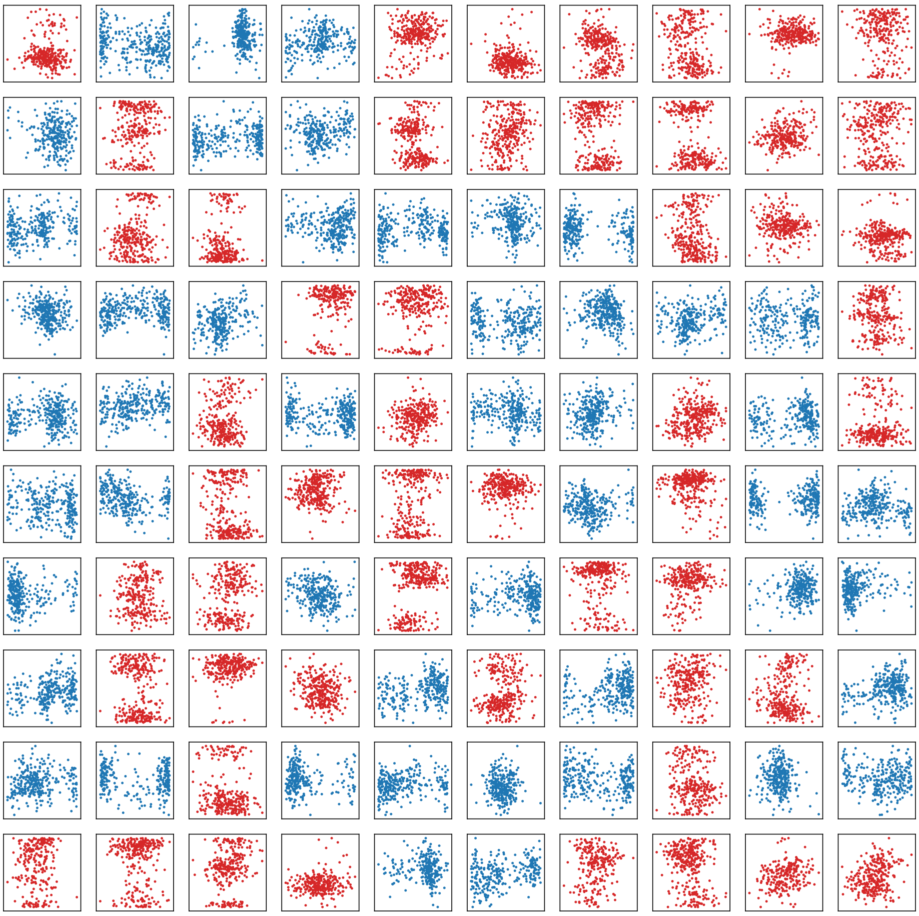


Figure C.3.: Synthetic Datasets sampled via alg. 1 with $n_{\text{bins}} = 16$. Blue scatter plots indicate a true causal direction of $X \to Y$, red scatter plots the direction $Y \to X$.

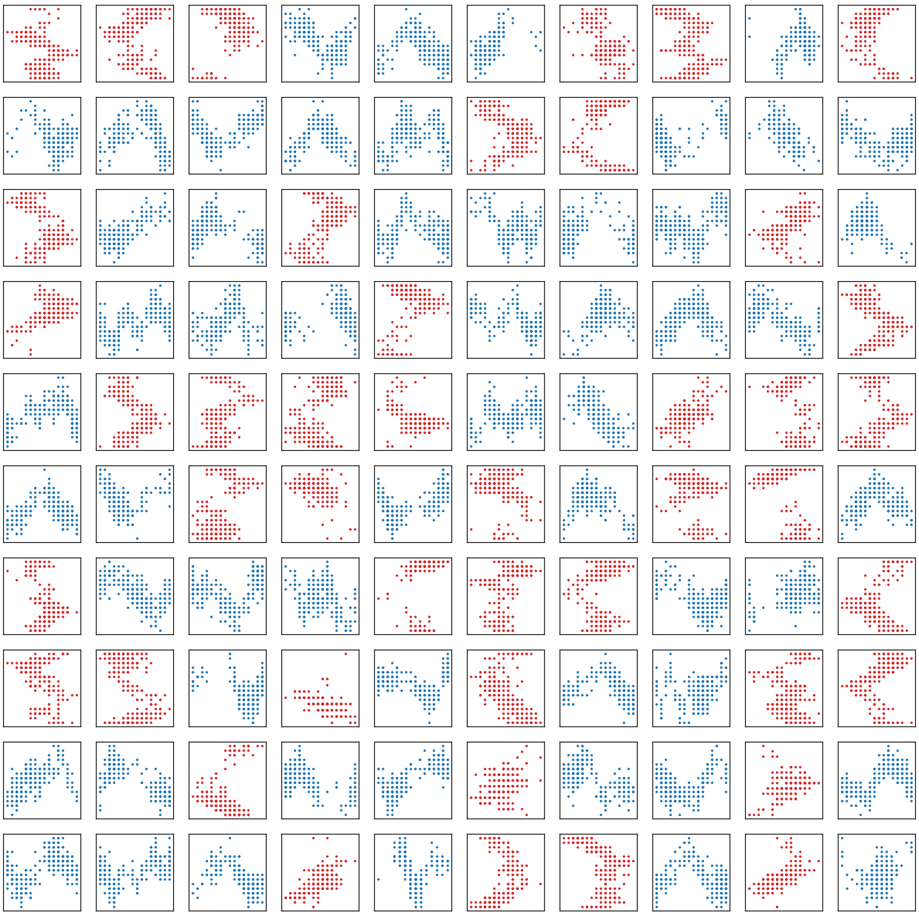


Figure C.4.: Synthetic Datasets sampled via alg. 1 with $n_{\text{bins}} = 8$. Blue scatter plots indicate a true causal direction of $X \rightarrow Y$, red scatter plots the direction $Y \rightarrow X$.

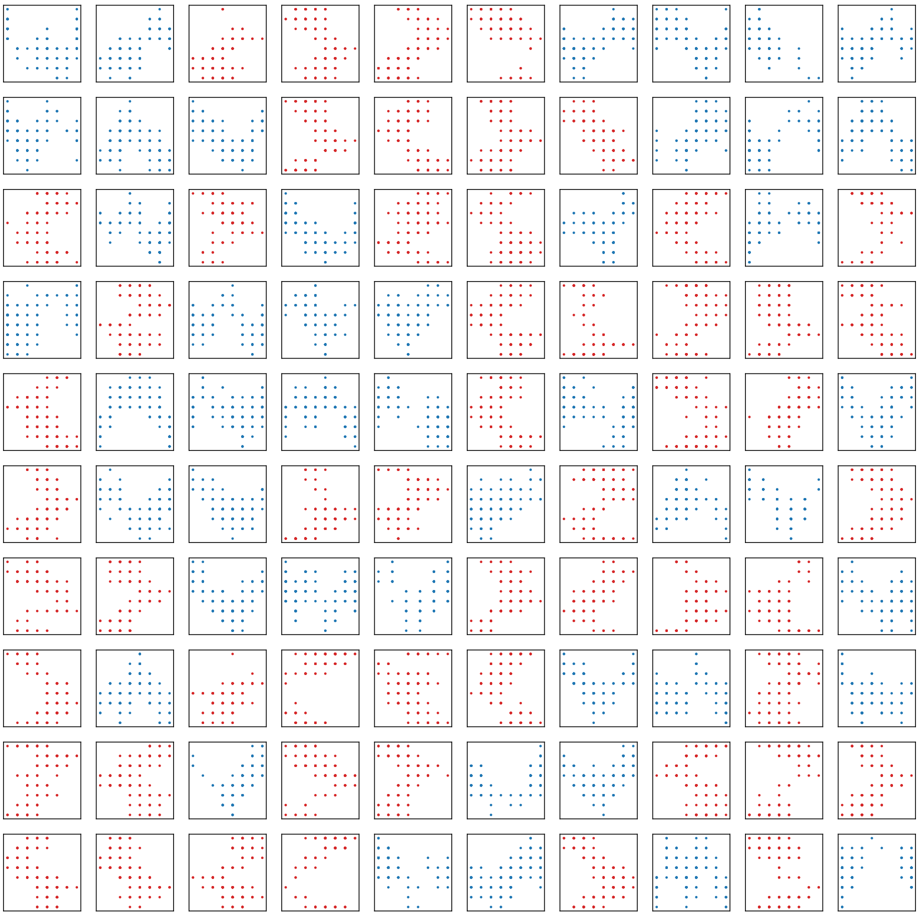


Figure C.5.: Synthetic Datasets sampled via alg. 1 with $\tilde{N} = 30$. Blue scatter plots indicate a true causal direction of $X \rightarrow Y$, red scatter plots the direction $Y \rightarrow X$.

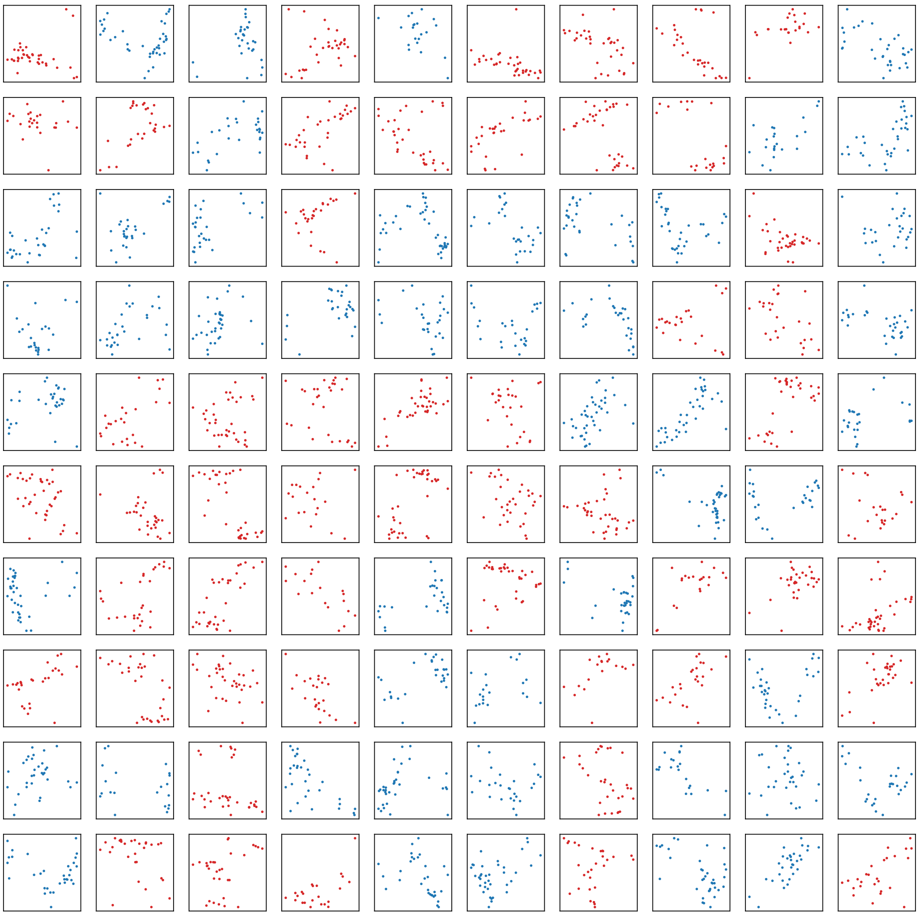
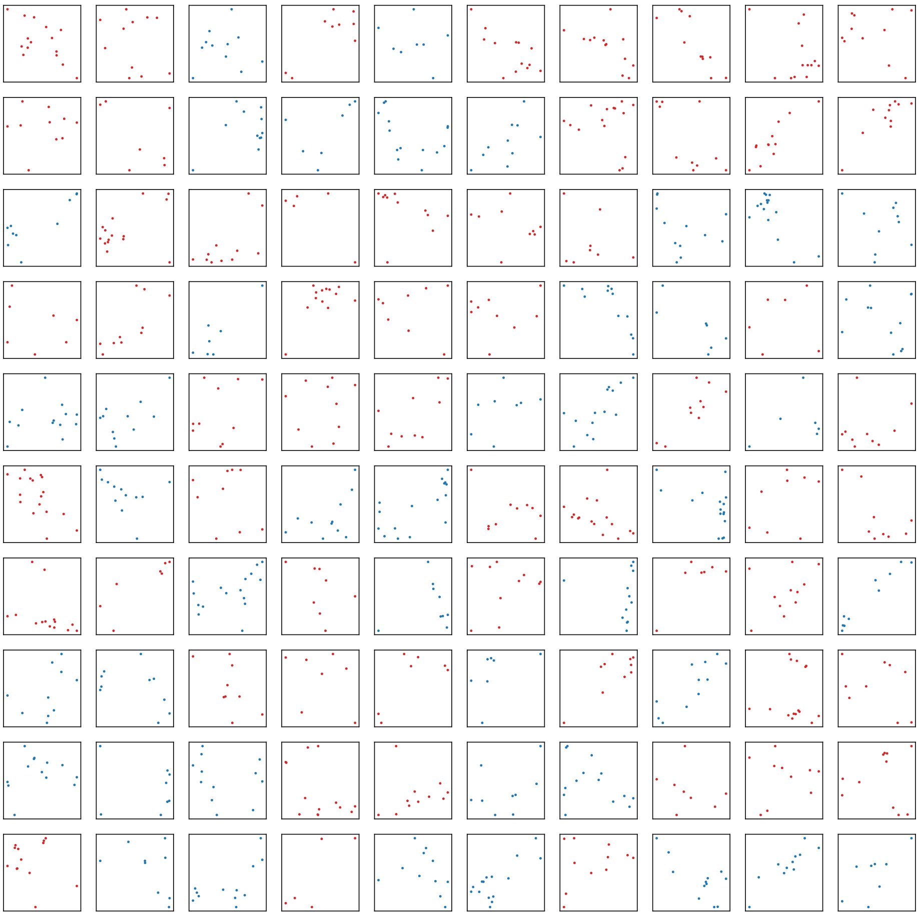


Figure C.6.: Synthetic Datasets sampled via alg. 1 with $\tilde{N} = 10$. Blue scatter plots indicate a true causal direction of $X \to Y$, red scatter plots the direction $Y \to X$.



D. Description of the TCEP Benchmark Set

The following is a description for the TCEP benchmark set. We omit the 6 datasets which contain more than 2 variables as they have not been used in the benchmark. Scatter plots are shown in fig. D.1.

dataset name	X	Y	source	causal direction
pair0001	Altitude	Temperature	DWD	$X \to Y$
pair0002	Altitude	Precipitation	DWD	$X \to Y$
pair0003	Longitude	Temperature	DWD	$X \to Y$
pair0004	Altitude	Sunshine hours	DWD	$X \to Y$
pair0005	Age	Length	Abalone	$X \to Y$
pair0006	Age	Shell weight	Abalone	$X \to Y$
pair0007	Age	Diameter	Abalone	$X \to Y$
pair0008	Age	Height	Abalone	$X \to Y$
pair0009	Age	Whole weight	Abalone	$X \to Y$
pair0010	Age	Shucked weight	Abalone	$X \to Y$
pair0011	Age	Viscera weight	Abalone	$X \to Y$
pair0012	Age	Wage per hour	census income	$X \to Y$
pair0013	Displacement	Fuel consumption	auto-mpg	$X \to Y$
pair0014	Horse power	Fuel consumption	auto-mpg	$X \to Y$
pair0015	Weight	Fuel consumption	auto-mpg	$X \to Y$
pair0016	Horsepower	Acceleration	auto-mpg	$X \to Y$
pair0017	Age	Dividends from stocks	census income	$X \to Y$
pair0018	Age	Concentration GAG	GAGurine (from R package MASS)	$X \to Y$
pair0019	Current duration	Next interval	geyser	$X \to Y$
pair0020	Latitude	Temperature	DWD	$X \to Y$
pair0021	Longitude	Precipitation	DWD	$X \to Y$
pair0022	Age	Height	arrhythmia	$X \to Y$
pair0023	Age	Weight	arrhythmia	$X \to Y$
pair0024	Age	Heart rate	arrhythmia	$X \to Y$
pair0025	Cement	Compressive strength	concrete_data	$X \to Y$

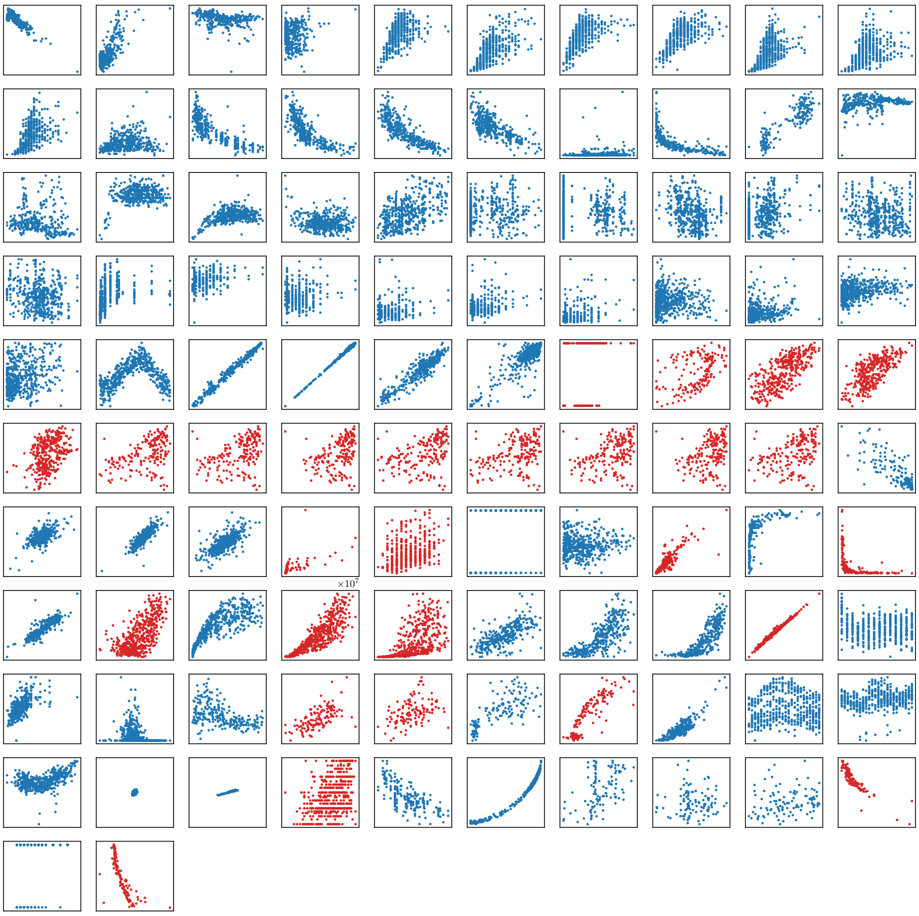
pair0026	Blast furnace slag	Compressive strength	concrete_data	$X \to Y$
pair0027	Fly ash	Compressive strength	concrete_data	$X \to Y$
pair0028	Water	Compressive strength	concrete_data	$X \to Y$
pair0029	Superplasticizer	Compressive strength	concrete_data	$X \to Y$
pair0030	Coarse aggregate	Compressive strength	concrete_data	$X \to Y$
pair0031	Fine aggregate	Compressive strength	concrete_data	$X \to Y$
pair0032	Age	Compressive strength	concrete_data	$X \to Y$
pair0033	Alcohol consumption	Mean corpuscular volume	liver disorders	$X \to Y$
pair0034	Alcohol consumption	Alkaline phosphotase	liver disorders	$X \to Y$
pair0035	Alcohol consumption	Alanine aminotransferase	liver disorders	$X \to Y$
pair0036	Alcohol consumption	Aspartate aminotransferase	liver disorders	$X \to Y$
pair0037	Alcohol consumption	Gamma-glutamyl transpeptdase	liver disorders	$X \to Y$
pair0038	Age	Body mass index	pima indian diabetes	$X \to Y$
pair0039	Age	Serum insulin	pima indian diabetes	$X \to Y$
pair0040	Age	Diastolic blood pressure	pima indian diabetes	$X \to Y$
pair0041	Age	Plasma glucose concentration	pima indian diabetes	$X \to Y$
pair0042	Day of the year	Temperature	B.Janzing	$X \to Y$
pair0043	Temperature at t	Temperature at t+1	ncep-ncar	$X \to Y$
pair0044	Pressure at t	Pressure at t+1	ncep-ncar	$X \to Y$
pair0045	Sea level pressure at t	Sea level pressure at t+1	ncep-ncar	$X \to Y$
pair0046	Relative humidity at t	Relative humidity at t+1	ncep-ncar	$X \to Y$
pair0047	Number of cars	Type of day	traffic	$Y \to X$
pair0048	Indoor temperature	Outdoor temperature	Hipel & Mcleod	$Y \to X$
pair0049	Ozone concentration	Temperature	Bafu	$Y \to X$

pair0050	Ozone concentration	Temperature	Bafu	$Y \to X$
pair0051	Ozone concentration	Temperature	Bafu	$Y \to X$
pair0056	Female life expectancy, 2000-2005	Latitude	UNdata	$Y \to X$
pair0057	Female life expectancy, 1995-2000	Latitude	UNdata	$Y \to X$
pair0058	Female life expectancy, 1990-1995	Latitude	UNdata	$Y \to X$
pair0059	Female life expectancy, 1985-1990	Latitude	UNdata	$Y \to X$
pair0060	Male life expectancy, 2000-2005	Latitude	UNdata	$Y \to X$
pair0061	Male life expectancy, 1995-2000	Latitude	UNdata	$Y \to X$
pair0062	Male life expectancy, 1990-1995	Latitude	UNdata	$Y \to X$
pair0063	Male life expectancy, 1985-1990	Latitude	UNdata	$Y \to X$
pair0064	Drinking water access	Infant mortality	UNdata	$X \to Y$
pair0065	Stock return of Hang Seng Bank	Stock return of HSBC Hldgs	Yahoo database	$X \to Y$
pair0066	Stock return of Hutchison	Stock return of Cheung kong	Yahoo database	$X \to Y$
pair0067	Stock return of Cheung kong	Stock return of Sun Hung Kai Prop.	Yahoo database	$X \to Y$
pair0068	Bytes sent	Open http connections	P. Stark & Janzing	$Y \to X$
pair0069	Inside temperature	Outside temperature	J.M. Mooij	$Y \to X$
pair0070	Parameter	Answer	Armann & Buelthoff	$X \to Y$
pair0072	Sunspots	Global mean temperature	sunspot data	$X \to Y$
pair0073	CO2 emissions	Energy use	UNdata	$Y \to X$
pair0074	GNI per capita	Life expectancy	UNdata	$X \to Y$

pair0075	Under-5 mortality rate	GNI per capita	UNdata	$Y \to X$
pair0076	Population growth	Food consumption growth	Food and Agriculture Organization of the Unite...	$X \to Y$
pair0077	Temperature	Solar radiation	B. Janzing	$Y \to X$
pair0078	PPFD	Net Ecosystem Productivity	Moffat A.M.	$X \to Y$
pair0079	Net Ecosystem Productivity	Diffuse PPFDdif	Moffat A.M.	$Y \to X$
pair0080	Net Ecosystem Productivity	Direct PPFDdir	Moffat A.M.	$Y \to X$
pair0081	Temperature	Local CO2 flux, BE-Bra	Mahecha, M.	$X \to Y$
pair0082	Temperature	Local CO2 flux, DE-Har	Mahecha, M.	$X \to Y$
pair0083	Temperature	Local CO2 flux, US-PFa	Mahecha, M.	$X \to Y$
pair0084	Employment	Population	spatial-econometrics.com	$Y \to X$
pair0085	Time of measurement	Protein content of milk	maths.lancs.ac.uk	$X \to Y$
pair0086	Size of apartment	Monthly rent	J.M. Mooij	$X \to Y$
pair0087	Temperature	Total snow	Snowfall in Whistler, from www.mldata.org	$X \to Y$
pair0088	Age	Relative spinal bone mineral density	bone dataset of R ElemStatLearn package	$X \to Y$
pair0089	root decomposition Oct (grassl)	root decomposition Oct (grassl)	Solly et al (2014). Plant and Soil, 382(1-2), ...	$Y \to X$
pair0090	root decomposition Oct (forest)	root decomposition Oct (forest)	Solly et al (2014). Plant and Soil, 382(1-2), ...	$Y \to X$
pair0091	clay cont. in soil (forest)	soil moisture	Solly et al (2014). Plant and Soil, 382(1-2), ...	$X \to Y$

pair0092	organic carbon in soil (forest)	clay cont. in soil (forest)	Solly et al (2014). Plant and Soil, 382(1-2), ...	$Y \rightarrow X$
pair0093	precipitation	runoff	MOPEX	$X \rightarrow Y$
pair0094	hour of day	temperature	S. Armagan Tarim	$X \rightarrow Y$
pair0095	hour of day	electricity load	S. Armagan Tarim	$X \rightarrow Y$
pair0096	temperature	electricity load	S. Armagan Tarim	$X \rightarrow Y$
pair0097	speed at the beginning	speed at the end	D. Janzing	$X \rightarrow Y$
pair0098	speed at the beginning	speed at the end	D. Janzing	$X \rightarrow Y$
pair0099	language test score	social-economic status family	nlschools dataset of R MASS package	$Y \rightarrow X$
pair0100	cycle time of CPU	performance	cpus dataset of R MASS package	$X \rightarrow Y$
pair0101	grey value of a pixel	brightness of the screen	D. Janzing	$X \rightarrow Y$
pair0102	position of a ball	time for passing a track segment	D. Janzing	$X \rightarrow Y$
pair0103	position of a ball	time for passing a track segment	D. Janzing	$X \rightarrow Y$
pair0104	time for passing 1. segment	time for passing 2. segment	D. Janzing	$X \rightarrow Y$
pair0106	time required for one round	voltage	D. Janzing	$Y \rightarrow X$
pair0107	strength of contrast	answer correct or not	Schuett, edited by D. Janzing	$X \rightarrow Y$
pair0108	time for 1/6 rotation	temperature	D. Janzing	$Y \rightarrow X$

Figure D.1.: TCEP benchmark dataset. Blue scatter plots indicate a true causal direction of $X \rightarrow Y$, red scatter plots the direction $Y \rightarrow X$.



Statement of authorship

I hereby declare that I completed this thesis on my own and that information which has been directly or indirectly taken from other sources has been noted as such.

Munich, September 12, 2018

(Signature)