Bayesian Causal Inference and Quasi Periodic Signal Analysis

Master's Thesis at the Faculty of Physics
of the
Ludwig Maximilians University, Munich

submitted by
Andrija Kostić
born in Leskovac, Serbia, on 11.03.1995

Munich, Germany, 15.09.2020

Supervisor:
P.D. Dr. Torsten Enßlin

Bayesianische Inferenz von Kausalität und Analyse quasiperiodischer Signale

Master Arbeit an der Fakultät für Physik
der
Ludwig-Maximilians-Universität München

vorgelegt von
Andrija Kostić
geboren in Leskovac am 11. März 1995

München, den 15. September, 2020

Betreuer:
P.D. Dr. Torsten Enßlin

Abstract

The work done in this thesis addresses three topics in Bayesian inference. First, the metric Gaussian variational inference (MGVI) method to approximately solve large inference problems is investigated. Correction terms are calculated with aim to improve the approximation, however for a high computational cost. Second, causal inference based on Bayesian causal models is further developed, exploiting MGVI for obtaining evidence lower bound estimates for different causal models. The methods developed perform equally reliably as existing state of the art methods. Third, photon count data from a recent cosmic X-ray burst is examined for the presence of quasi periodic signals that could reveal crust properties of the emitting magnetars. No significant quasi periodic signals are detected in this Bayesian analysis, once again involving MGVI.

Contents

1 Introduction

Inferring parameters of the posterior distribution in Bayesian inference can be rather challenging. Especially if the posterior has a complicated deep hierarchy of priors. Therefore, it is often only possible to just approximate the posterior using several available techniques. There are many different approaches to solving this problem with varying degrees of precision. One of the biggest challenges is to have a good balance between computational efficiency and precision of the approximation. The recently developed metric Gaussian variational inference algorithm [36] offers a great balance of the two. That is why in the first part of this work an attempt is made to understand this algorithm better and improve on it. It is shown that in order to be fully consistent within the framework algorithm sits in, additional terms need to be added to the inference machinery, although the computational cost for those terms seems to be high. These terms are calculated and their impact in few available cases is analysed.

The second and third part of this thesis try to use the metric Gaussian variational inference algorithm in solving two problems.

The first problem considered is a very common in the field of causal inference. Here a relation between "cause" and "effect" is sought for. For example if somebody is given certain number of joint measurements of the height of mercury in a barometer and pressure of the surrounding air, the goal would be to determine whether increased pressure would cause the change in the height of mercury or is the reverse true. In this case, it is clear that the "cause" would be pressure and the "effect" would be change in height of mercury, since upon changing the height of mercury would not change the surrounding pressure of the air. There could as well be a common cause which would influence both the air pressure and the height of mercury. For instance raising the temperature could cause the fluid of mercury to extend a bit increasing its height and changing the air pressure as well. The algorithm we develop tries to distinguish between different causal scenarios for different benchmark datasets. Furthermore, a comparison of its performance with state-of-the-art algorithms is given as well. The main contribution of this part is that a fully Bayesian approach within the formalism of Information Field Theory [35] is developed with a comparable performance to state-of-the-art methods. These algorithms, in many cases, are not completely Bayesian, but rely on certain heuristic evaluation in order to make a decision.

For the final, third part, a problem of extracting quasi periodic signals from Poisson data is considered. The motivation for pursuing this came from receiving a completely fresh dataset which was obtained from a detection of a new soft gamma repeater event occurred during the course of this thesis. Soft gamma repeaters are defined as compact objects with very strong magnetic fields of the order of $10^{13} - 10^{15} G$, usually referred to as magnetars, because of their strong magnetic fields. The interest in these objects was raised in particular, after a discovery of certain frequencies present in their lightcurves, which are believed to be a consequence of crust-core interaction within these compact objects coupled through the strong magnetic fields. Therefore, detecting and studying quasi periodic oscillations in their light curves, could offer a new observational windows for probing the internal structure of these exotic objects, as indeed simulations suggest. Hence, a potential of possibly discovering new quasi periodic oscillation frequencies seemed to be a great opportunity to pass by.

In conclusion, this thesis has a rather diverse content, but all of it with a common thread that starts from the pursuit for understanding the metric Gaussian variational inference algorithm and then trying to drive it towards its limitations.

In order to better understand the content in the rest of the text, next section serves as a quick introduction into the formalism of the Information Field Theory. Hopefully, it provides the reader with enough information in order to understand the content of the chapters that follow.

1.1 A short introduction to Information Field Theory

In order to introduce concepts of information field theory it is instructive to consider first plain information theory.

Within information theory, information I about certain quantity of interest, s, can be represented through the conditional probability $p(s|I)$. This definition is useful because the language of conditional probabilities allows for a consistent and optimal way of updating our state of knowledge of quantity s through the use of Bayes theorem [3] upon receiving some new piece of information about s. For example, we can call that new information d and the theorem yields:

$$p(s|d, I) = \frac{p(d|s, I)p(s|I)}{p(d|I)} \tag{1.1}$$

One can think about this new information as measurements of the quantity s, i.e. the data we work with in order to infer quantity s. For example, the information I can suggest that data d comes from a linear measurement of s with an additive noise n, independent of s, whose distribution is a zero centered Gaussian. In other words

$$d = R(s) + n, \tag{1.2}$$

$$p(n|I) = \mathcal{G}(n, N) \tag{1.3}$$

$$= |2\pi N|^{-1/2} \exp\left(-\frac{1}{2} n^T N^{-1} n\right), \tag{1.4}$$

where N represents the noise covariance, R the measurement response, and $|\,.\,|$ an operation of taking a determinant. Now, the information I can also tell us that the distribution of s, before we receive any data from the measurement, is a multivariate Gaussian

$$p(s|I) = \mathcal{G}(s, S) \tag{1.5}$$

where by S we denote the positive-definite symmetric covariance of the Gaussian distribution. Using the fact that noise is independent and additive the likelihood term from Bayes theorem (equation (1.1)), $p(d|s, I)$, can be obtained by marginalizing the noise. The likelihood term is

$$p(d|s, I) = \mathcal{G}(d - R(s), N), \tag{1.6}$$

where now $R(s)$ becomes the mean of this Gaussian. In order to find the posterior distribution of s, $p(s|d, I)$, it is useful to introduce information Hamiltonians. An information Hamiltonian is defined as the following quantity:

$$\mathcal{H}(x|y) \equiv -\ln(p(x|y)) \tag{1.7}$$

Using this definition and assuming the response R can be represented as a matrix, one obtains

$$\begin{aligned}
\mathcal{H}(d,s) &= \mathcal{H}(d|s) + \mathcal{H}(s) = -\ln p(d|s) - \ln p(s) && (1.8)\\
&= \frac{1}{2}\left(d - R(s)\right)^T N^{-1}\left(d - R(s)\right) + \frac{1}{2}\ln|2\pi N| + \frac{1}{2}s^T S^{-1} s + \frac{1}{2}\ln|2\pi S| && (1.9)\\
&= \frac{1}{2}\left(d - R(s)\right)^T N^{-1}\left(d - R(s)\right) + \frac{1}{2}s^T S^{-1} s + \frac{1}{2}\ln|2\pi N| + \frac{1}{2}\ln|2\pi S| && (1.10)\\
&= \frac{1}{2}\left(s^T(R^T N^{-1} R + S^{-1})s + d^T N^{-1} R s + s R^T N^{-1} d\right) && (1.11)\\
&+ \frac{1}{2}\left(d^T N^{-1} d + \mathrm{Tr}\ln N + \mathrm{Tr}\ln S + N_d \ln 2\pi + N_s \ln 2\pi\right), && (1.12)
\end{aligned}$$

where N_d and N_s represent the dimensions of the spaces on which d and s are defined, respectively. Now, redefining $j = R^T N^{-1} d$ and $D^{-1} = (R^T N^{-1} R + S^{-1})$ we obtain

$$\mathcal{H}(d,s) = \frac{1}{2}\left(s - Dj\right)^T D^{-1}\left(s - Dj\right) + \mathcal{H}_0, \tag{1.13}$$

where we have accumulated all the s independent terms inside $\mathcal{H}_0$, i.e.

$$\mathcal{H}_0 = \frac{1}{2}\left(d^T N^{-1} d + \mathrm{Tr}\ln N + \mathrm{Tr}\ln S + N_d \ln 2\pi + N_s \ln 2\pi\right). \tag{1.14}$$

These terms will become especially important later on in section 4.

From above it can be seen that the posterior $p(s|d,I)$ is also a Gaussian:

$$p(s|d,I) = \mathcal{G}(s - m, D),$$

with mean m and covariance D given as

$$\begin{aligned}
m &= Dj\\
D &= (R^T N^{-1} R + S^{-1})^{-1}.
\end{aligned} \tag{1.15}$$

Now, in order to make a transition towards information field theory one has to go from the discrete to a continuum limit. More precisely:

$$s_i \to s^x,$$

where $x \in \mathbb{R}^n$. The distinction from the previous case is that now the field is a continuous quantity which we evaluate on certain grid space $\mathbb{R}^n$ and not a discretized quantity which is only defined on the chosen grid. Furthermore, this indexation allows us to distinguish from the vectors of the dual space and this will become important when considering the limit $n \to \infty$. To make this more precise, a field s can be represented as

$$s = s^x e_x, \tag{1.16}$$

where e_x can be chosen to be the functional basis $e_x(y) = \delta(x - y)$. A scalar product can be chosen as

$$r^\dagger s = \int dx\ \overline{r(x)}\ s(x) = \overline{r_x} s^x, \tag{1.17}$$

therefore, field component would be simply

$$s(x) = s^y e_y(x) = \int dx s^y \delta(x-y) = s^x, \tag{1.18}$$

from which it is clear that s^x indeed represents a value of the field at a grid position x.

In this notation the matrices S_{ij} and R_{ij} become operators and are understood through their action on the corresponding fields:

$$\begin{gathered} S_{ij} \to S^{xy} \\ R_{ij} \to R^i_y \\ S^{-1}(s) \equiv (S^{-1})_{xy} s^y = g_{xz} \int dy (S^{-1})^z_y s^y \\ R(s) \equiv R^i_x s^x = \int dy R^i_y s^y, \end{gathered}$$

where the metric g_{xy} is used as a connection to the dual space. Furthermore, indexation of the response $R \equiv R^i_y$ reflects the fact that the data space is still kept discrete and finite, while the space of the field s is kept continuous.

It is worthwhile mentioning that the choice of the metric g_{xy}, and hence a particular coordinate system, doesn't influence the quantities we're interested in. The information Hamiltonian which we use for the variational inference approach is a scalar quantity invariant under coordinate transformation, while the mean (look back for example at equation (1.15))

$$m^x = D^{xy} R^i_y N^{-1}_{ij} d^j \tag{1.19}$$

has all indices occurring in contracting pairs. Therefore the result obtained from the inference machinery is completely coordinate independent. This would allow for convenient reparameterizations as shown in [28, 36] which help to increase the numerical efficiency of the metric Gaussian variational inference approach.

Of course, at the end of the day we're doing our computations on a computer machine, but it is still important to make connections between the continuum and discretized limit in order to make them consistent with each other. A good example of this is for example when simulating particle interactions which source continuous fields (for example the gravitational field) or when solving partial differential equations. The latter has already been considered in the IFT context, and the reader is directed towards the work described in [30] for further details.

Now, another thing that has to be addressed before continuing to section 4.2 is the choice of priors. A reasonable prior should constrain the quantity of interest we want to infer from its measurement data d to sensible range, think of s from above. This choice of constraints is dictated by knowledge I. Furthermore, the prior should be as ignorant as possible towards a priori indistinguishable outcomes. For example, we may not wish to single out a particular point in space on which field s is defined. This would then translate into its prior covariance being translational invariant

$$S^{xy} = C(x-y), \tag{1.20}$$

with some correlation kernel C. Then, the Wiener-Khinchin theorem [1, 2] suggests this

covariance will be diagonal in Fourier space:

$$S^{xy} = (\mathbb{F}^{-1})^x_m S^{ml} (\mathbb{F}^{-1})^y_l$$
$$\text{with } S^{ml} = \begin{cases} 0 & l \neq m \\ V P_s(k) & l = m, \end{cases}$$

where $(\mathbb{F}^{-1})^x_m$ and $(\mathbb{F}^{-1})^y_l$ denote the inverse and adjoint inverse Fourier transform. V denotes the total volume of the domain of the field s and $P_s(k)$ its power spectrum. In the following chapters we would utilize the notation and concepts introduced briefly here. Much of this text closely followed descriptions given in [34, 38] to which the reader is referred for more details.

2 Model selection with Bayesian reasoning

It is not for pure convenience that the IFT formalism is developed within the framework of Bayesian reasoning. As it was shown by Cox back in 1946 [3], this framework arrises as a natural formulation for problems where one needs to draw conclusions from incomplete information at hand. Namely, there is a unique correspondence between reasoning in logic and Bayesian inference if one constrains himself to representing degrees of plausibility by real numbers and accepts few reasonable desiderata. In this natural setting it is possible to formulate any problem and incorporate into this formulation all the available knowledge about the given problem. In fact, following Jaynes' [8] advice, this is exactly what one should do. Excellent sources for learning the matter, besides the Jaynes' book (*ibid.*) are the lecture notes by Ariel Caticha [22] and Torsten Enßlin's notes on Information Theory and Information Field Theory [34]. The structure of this chapter would be closely following these three references.

Just for the sake of building the narrative for the rest of the text that follows here I will briefly describe a very simple problem and demonstrate the power of Bayesian formalism in testing different hypotheses. Later on, I will follow the example of David MacKay [14] and further motivate the use of Bayesian approach as a method which has an implicitly built in Occam's razor. This will be important later on for understanding the content of section 4, especially section 4.4. Along the way, motivation for the metric Gaussian variational inference approach (MGVI) will be offered from a different perspective and hopefully serve as a good introduction to section 3.

2.1 A simple problem

Coin tossing

In this chapter, I will consider a problem of inferring whether the given coin being tossed is fair or not. Consider the following problem statement:

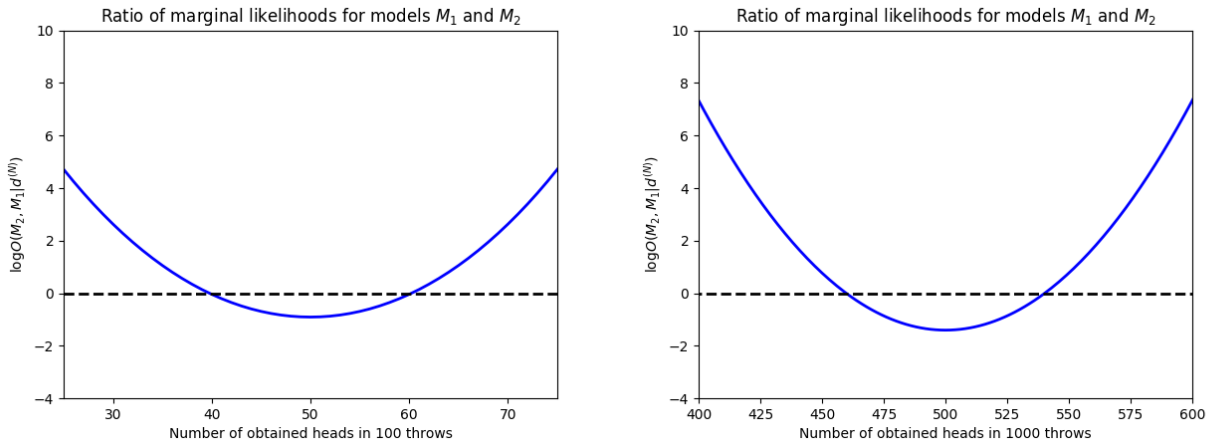
We are given a sequence of 0's and 1's, which represent tails and heads respectively. This we will store in a variable $d^{(N)} = \{d_i\}_{i=1}^N$*, with* $d_i \in \{0, 1\}$*. We assume that outcome of* d_i *and* d_j *for* $i \neq j$ *is independent. From this data, the problem is to calculate the evidence for the following two models:*

- *Model* M_1 = *"Given the coin is fair, i.e. the probability of obtaining 1 is* $\theta = 0.5$*"*
- *Model* M_2 = *"Given the coin is not fair, to an unknown degree. In this case we denote the probability of obtaining heads with* $\theta \in \Omega_\theta^{(M_2)} = [0, 1] \setminus \{2^{-1}\}$

Now, it is clear how the Bayes theorem will be applied:

$$\begin{aligned} p(M_1|d^{(N)}) &= p(d^{(N)}|M_1)p(M_1)/p(d^{(N)}) \\ p(M_2|d^{(N)}) &= p(d^{(N)}|M_2)p(M_2)/p(d^{(N)}). \end{aligned} \tag{2.1}$$

Hence it is necessary to determine the likelihoods and choose suitable priors for the two cases. It is reasonable to assume equal priors for both models $p(M_1) = p(M_2)$, since before data enters the picture we don't have any preference towards any of the models.



(a) $N = 100$ (b) $N = 1000$

Figure 2.1: Ratio $\log_{10}(O(M_1, M_2|d^{(N)}))$ shown for two different cases, with $N = 100$ (figure 2.1a) and $N = 1000$ (figure 2.1b).

Regarding the likelihood of obtaining the particular sequence $d^{(N)}$ we have

$$p(d^{(N)}|M_i) = p(d_1|M_i)p(d_2|d_1, M_i)\dots p(d_N|d_1, d_2, \dots, d_{N-1}, M_i), \tag{2.2}$$

and remembering that tosses are mutually independent,

$$p(d^{(N)}|M_i) = p(d_1|M_i)p(d_2|M_i)\dots p(d_N|M_i), \tag{2.3}$$

with $M_i \in \{M_1, M_2\}$. What can also be seen from the above equation is that the particular ordering in which the heads or tails appear is not important, only the number of their occurrences. The likelihood becomes

$$p(d^{(N)}|\theta, M_i) = \theta^{N_\mathrm{h}}(1-\theta)^{N_\mathrm{t}}, \tag{2.4}$$

where N_h and N_t represent the number of heads and tails in the observed sequence. Of course $N_\mathrm{t} = N - N_\mathrm{h}$. Furthermore, note that in the above equation 2.4, the information about θ is very easily incorporated in the conditional expression. That is another nice property of the Bayesian approach, since by using the language of conditional probabilities it is very clear what assumptions are made and what dependencies are assumed.

Continuing further, it is possible to immediately write down what is the expression for the likelihood for the model M_1:

$$p(d^{(N)}|M_1) = 2^{-N}. \tag{2.5}$$

For obtaining the likelihood for the data given the model M_2 one has:

$$\begin{aligned} p(d^{(N)}|M_2) &= \sum_{\theta\in\Omega_\theta^{(M_2)}} p(d^{(N)}|\theta, M_2)p(\theta|M_2) \\ &= \int_{\theta\in\Omega_\theta^{(M_2)}} p(d^{(N)}|\theta, M_2)p(\theta|M_2) \\ &= \int_0^1 \mathrm{d}\theta\, \theta^{N_\mathrm{h}}(1-\theta)^{N-N_\mathrm{h}} \\ &= \beta(1+N_\mathrm{h}, 1+N_\mathrm{t}) \stackrel{\mathsf{def}}{=} \frac{\Gamma(1+N_\mathrm{h})\Gamma(1+N_\mathrm{t})}{\Gamma(N_\mathrm{h}+N_\mathrm{t}+2)} \\ &= \frac{N_\mathrm{h}!N_\mathrm{t}!}{(N+1)!}. \end{aligned} \tag{2.6}$$

From first to second line in the above equation, the fact that θ is a continuous variable was used, as well as the fact that the set $\{2^{-1}\}$ is of measure zero. Therefore, comparing these two model likelihoods would give us an estimate of how well the data is favoring model M_1 or model M_2:

$$O(M_1, M_2|d^{(N)}) = p(d^{(N)}|M_2)/p(d^{(N)}|M_1) = 2^N N_\mathrm{h}!N_\mathrm{t}!/(N+1)!\ . \tag{2.7}$$

The quantity given in eq. (2.7) is also known as the Bayesian odds factor and the method is called *Marginal Likelihood method* since we marginalized over the parameter θ. In the figure 2.1, one can see how the $\log(O(M_1, M_2|d^{(N)})$ (logarithm of the equation 2.7) behaves in the case of $N = 100$ and $N = 1000$. What can be noticed by comparing the two figures is that in the case of $N = 1000$ the range where the fair coin model (M_1) is more probable given the data is narrower than in the case for $N = 100$. For example if one would see around 460 heads in 1000 throws then it would be difficult to distinguish between the two models, but seeing 425 heads in 1000 throws would make model $M_2 \approx 10^4$ times more probable than model M_1.

Now, it would not always be possible to perform the marginalization over our parameter space. Reasons could be that the space is simply infinite dimensional or that the function we have to integrate simply doesn't have an analytic solution. In these cases there are multiple different ways of obtaining a preference towards one of the hypotheses. For example, *Minimal Message Length* (MML) score [16] and *Empirical Bayes* methods to name a few. All of them have one thing in common and that is that all of them have a built in notion of 'Occam's razor'. This principle of 'Occam's razor' will be described briefly in next chapter and serve to motivate the metric Gaussian variational inference (MGVI) approach [36] from a different perspective.

2.2 Bayesian Occam's factor

Bayesian inference problems in essence can be split into two stages. First one being the inference of the unknown parameters of the model at hand, and the second one consisting of estimating the uncertainties of the inferred parameters. In cases where there are multiple competing models, the second stage is what separates the Bayesian approach from other inference schemes.

"The best" model M, can be defined as the one for which $p(M|d, I)$ is maximal, given some background knowledge I and additional information d (think of a particular data set). But,

usually we don't have access to this quantity, since in order to do so, we have to marginalize over all possible parameterizations for the given model M. This task is hardly possible, with one of the special cases described in previous section. Therefore, what is often done is the following.

Let the competing models be denoted by M_i, with $i = 1, 2, \ldots$ used to label them. Each of them has some parameterization $\boldsymbol{\theta}$, which represents a vector, a prior probability distribution it assigns these parameters to, $p(\boldsymbol{\theta}|M_i, I)$, and a likelihood $p(d|\boldsymbol{\theta}, M_i, I)$. Here, within I certain background knowledge about the problem is collected. Then, in the context of the Bayes theorem, the posterior probability for parameters $\boldsymbol{\theta}$ for some model M_i after data d is received will be:

$$p(\boldsymbol{\theta}|d, M_i, I) = \frac{p(d|\boldsymbol{\theta}, M_i, I)p(\boldsymbol{\theta}|M_i, I)}{p(d|M_i, I)} \tag{2.8}$$

In case this can't be evaluated analytically, as it was in section 2.1, often some gradient method is used for finding the maximum of the posterior distribution $\boldsymbol{\theta}_{\mathrm{MP}}$. Error estimates on this can be obtained by looking at the value of the Hessian matrix for example, which is defined as

$$H_{ij}(\boldsymbol{\theta}) = - \left. \frac{\partial}{\partial\theta^i\partial\theta^j} \ln p(\boldsymbol{\theta}|d, M_i, I) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}} \tag{2.9}$$

where the $\partial_i \equiv \frac{\partial}{\partial\theta^i}$ denotes a directional derivative along the given component of $\boldsymbol{\theta}$. Note the ln and the minus in front in the definition. When restated in the language of information Hamiltonians from the section 1.1, the Hessian will be exactly a matrix filled with second derivatives of the corresponding information Hamiltonian giving an estimate to the measure of local curvature at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{MP}}$.

Now, using this definition, the ln-posterior can then be expanded around the $\boldsymbol{\theta}_{\mathrm{MP}}$ giving

$$\begin{aligned} \ln p(\boldsymbol{\theta}|d, M_i, I) = \ln p(\boldsymbol{\theta}_{\mathrm{MP}}|d, M_i, I) + \frac{1}{p(\boldsymbol{\theta}_{\mathrm{MP}}|d, M_i, I)} \, \partial_i p(\boldsymbol{\theta}|d, M_i, I)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}} \, \delta\theta^i \\ + \frac{1}{2} \, \partial_i\partial_j \ln p(\boldsymbol{\theta}|d, M_i, I)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}} \, \delta\theta^i\delta\theta^j + o((\delta\theta)^3), \end{aligned} \tag{2.10}$$

and since $\boldsymbol{\theta}_{\mathrm{MP}}$ is an extremum the first derivative is vanishing giving in total

$$\ln p(\boldsymbol{\theta}|d, M_i, I) = \ln p(\boldsymbol{\theta}_{MP}|d, M_i, I) + \frac{1}{2} \, \partial_i\partial_j \ln p(\boldsymbol{\theta}|d, M_i, I)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{MP}}} \, \delta\theta^i\delta\theta^j + o((\delta\theta)^3). \tag{2.11}$$

This can be reshuffled to give

$$p(\boldsymbol{\theta}|d, M_i, I) \approx p(\boldsymbol{\theta}_{\mathrm{MP}}|d, M_i, I) \exp\left(-\frac{1}{2}\delta\theta^i H_{ij} \delta\theta^j\right). \tag{2.12}$$

Equation above only shows that locally, up to second order, the posterior is well approximated by a Gaussian distribution, a property which will be as well utilized by the MGVI algorithm. Note that this can often be done at a point which lies at an extremum of the posterior distribution. But, it should be understood that this extremum point is not uniquely defined, since one might as well make a non-linear reparameterization under which this extremum point will no longer be an extremum at all. For most of the cases however the likelihood is chosen such to assure the posterior is sufficiently concentrated around some given point, which could then be reached by the usual gradient based approaches. Otherwise, if this is not

possible, more sophisticated approaches are needed which can explore the complete posterior distribution given enough time (like the Hamiltonian Monte Carlo approach for example [5]).

With the uncertainty of the chosen $\boldsymbol{\theta}_{\mathrm{MP}}$ quantified through the Hessian estimate, now one can turn towards model comparison.

At this stage the optimal Bayesian way of finding the evidence for the given model M_i would be to simply marginalize over all possible parameterizations:

$$p(d|M_i, I) \int_{\Omega_{\boldsymbol{\theta}}} p(\boldsymbol{\theta}, d|M_i, I)\mathrm{d}\boldsymbol{\theta} = \int_{\Omega_{\boldsymbol{\theta}}} p(d|\boldsymbol{\theta}, M_i, I)p(\boldsymbol{\theta}|M_i, I)\mathrm{d}\boldsymbol{\theta} \tag{2.13}$$

But, since this is not always possible to do, the expression above can only be approximated. If the model is chosen such that posterior is sufficiently well concentrated around its maximum, $\boldsymbol{\theta}_{\mathrm{MP}}$, the integral above will have the biggest contribution from the surrounding of this point. Therefore:

$$p(d|M_i, I) \approx \underbrace{p(d|\boldsymbol{\theta}_{\mathrm{MP}}, M_i, I)}_{\text{goodness of fit}} \underbrace{p(\boldsymbol{\theta}_{\mathrm{MP}}|M_i, I)\sigma_{\theta|d,I}}_{\text{Occam's factor}}, \tag{2.14}$$

where the goodness of fit factor comes in through $p(d|\boldsymbol{\theta}_{\mathrm{MP}}, M_i, I)$ but is complemented with the product $p(\boldsymbol{\theta}_{\mathrm{MP}}|M_i, I)\sigma_{\theta|d,I}$, with $\sigma_{\theta|d,I}$ the total posterior uncertainty of $\boldsymbol{\theta}_{\mathrm{MP}}$. For example this can be estimated by calculating the determinant of the Hessian matrix at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{MP}}$ as defined above, giving back the volume of this posterior which lies under the Gaussian approximation.

In case the background knowledge I suggests to take a uniform prior $p(\boldsymbol{\theta}|M_i, I) = \sigma_{\theta|I}^{-1}$ the Occam's factor from the equation above will be

$$p(\boldsymbol{\theta}_{\mathrm{MP}}|M_i, I)\sigma_{\theta|d,I} = \frac{\sigma_{\theta|d,I}}{\sigma_{\theta|I}}, \tag{2.15}$$

which represents the ratio of the posterior volume factor to the prior volume factor. For example if a model at hand is a rather complex one, i.e. the space of parameters $\boldsymbol{\theta}$ it can choose from is rather large, then it will be more penalized than a simpler model with a smaller parameter space which can still explain the data reasonably well. Reason being that the prior volume for the complex model is larger than the one of the simpler one ($\sigma_{\theta_c|I} > \sigma_{\theta_s|I}$).

Another appealing interpretation of eq. (2.15) is given in [14], where it is claimed that the logarithm of the Occam's factor is nothing else than the amount of information that is obtained about parameters $\boldsymbol{\theta}$ upon receiving the data. This interpretation is interesting because it fits well within the picture of information Hamiltonians which are central pieces in the IFT formalism. It will be seen later on in section 4.4 that the method developed within this work for model comparison indeed has this factor built in with a similar interpretation as given in [14].

Hopefully up to this point the reader is convinced that estimating the uncertainties of the inferred quantities of interest is of extreme importance in Bayesian reasoning. Therefore, having Bayesian inference algorithms which offer not only the posterior mean estimates but as well the appropriate uncertainty estimates is crucial in order to properly evaluate performance of different models. This is precisely what the MGVI algorithm offers, utilizing not the Hessian as it was done here, but a different metric which seems to capture the available information inside the data better. Hence, in next section this algorithm is described in detail with an attempt to understand it properly and improve on it.

3 Metric Gaussian Variational Inference

The MGVI algorithm [36] is an algorithm for performing variational Bayesian inference in standardized coordinates. It does so by approximating the true posterior with a series of Gaussian approximations. Correlations between the inferred parameters are taken into account during the optimization procedure itself, instead of just at the end of the inference scheme. This allows self-consistent optimization for the corresponding mean and uncertainties. It is possible to achieve this due to the implicit parameterization of the covariance for the Gaussian approximation, which exploits local properties of the true posterior through the use of the inverse Fisher information metric.

Given this implicit covariance, the mean of approximating Gaussian distribution is shifted to better capture the true posterior. This is done by minimizing Kullback-Leibler (KL) divergence between the posterior and the approximation w.r.t. the mean. In turn, this allows for enhanced scaling properties to very high dimensional problems, since for sufficiently Gaussian posteriors, the scaling will be linear with the size of the dataset, due to the fact that only mean is optimized for. After the KL minimization step is done, procedure is reiterated until sufficient level of convergence is achieved.

It should be as well noted that MGVI algorithm can't capture well posteriors with a multi-modal structure, as any other variational inference approach. Also, if the model considered is too non-linear, meaning that the likelihood is chosen such that too many non-linear operations have to be performed in order to evaluate it, the higher order terms will become important and must be taken into account. These higher order terms are explicitly calculated in this chapter and their impact is briefly discussed for the few available cases.

In the following sections we motivate the approximating distribution chosen for the MGVI algorithm and describe how it is used for minimizing the KL. Afterwards, the higher order terms are computed and their impact is discussed.

3.1 Approximating distribution

Starting from the Bayes theorem one has

$$\begin{aligned} p(\theta|d) &= \frac{p(d|\theta)p(\theta)}{p(d)} \\ \mathcal{H}(\theta|d) &\mathrel{\widehat{=}} \mathcal{H}(d|\theta) + \mathcal{H}(\theta), \end{aligned} \tag{3.1}$$

where θ denotes the set of degrees of freedom of our model that need to be inferred, d is the data we're given and the sign $\mathrel{\widehat{=}}$ denotes that the term $\mathcal{H}(d)$ is dropped from the above equation since it will not be important for the variational inference, being constant w.r.t. θ.

In order to perform the variational inference, the complex posterior distribution $p(\theta|d)$ is approximated by a simpler one, in this case a Gaussian. The degree of discrepancy between the approximating distribution and the posterior is measured through the KL divergence:

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}(\mathcal{Q}(\theta|d)||p(\theta|d)) &= \int \mathcal{Q}(\theta|d) \ln \frac{\mathcal{Q}(\theta|d)}{p(\theta|d)} \mathrm{d}\theta \\ &= \langle \mathcal{H}(\theta|d) + \mathcal{H}(d) \rangle_{\mathcal{Q}(\theta|d)} - \langle \mathcal{Q}(\theta|d) \rangle_{\mathcal{Q}(\theta|d)} - \mathcal{H}(d) \\ &\mathrel{\widehat{=}} \langle \mathcal{H}(\theta, d) \rangle_{\mathcal{Q}(\theta|d)} - \langle \mathcal{Q}(\theta|d) \rangle_{\mathcal{Q}(\theta|d)} \end{aligned} \tag{3.2}$$

where the $\mathcal{H}(d)$ term is dropped from the second to the third row, denoted by $\widehat{=}$, in order to emphasize it is not important for the inference. These kind of terms will however become important later on in section 4. The last term in eq. (3.2) is the self entropy of the Gaussian approximation. Note that the conditional dependence of the approximating distribution on d is kept in the expressions in order to emphasize the mean of this approximation has to be inferred from the data.

Now, expressing the exact form of the approximation with

$$\mathcal{Q}(\theta|d) = \mathcal{G}(\theta - \bar{\theta}, \Theta), \tag{3.3}$$

the $\mathcal{D}_{\mathrm{KL}}$ becomes

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(\mathcal{Q}(\theta|d)||p(\theta|d)) &= \langle \mathcal{H}(\theta,d)\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} + \langle \ln \mathcal{G}(\theta-\bar{\theta},\Theta)\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} \\
&= \langle \mathcal{H}(\theta,d)\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} - \frac{1}{2}\langle (\theta-\bar{\theta})^\dagger \Theta^{-1}(\theta-\bar{\theta}) + \ln|2\pi\Theta|\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} \\
&= \langle \mathcal{H}(\theta,d)\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} - \frac{1}{2}\left(\langle (\theta-\bar{\theta})^\dagger \Theta^{-1}(\theta-\bar{\theta})\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} + \langle \ln|2\pi\Theta|\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)}\right) \\
&= \langle \mathcal{H}(\theta,d)\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} - \frac{1}{2}\left(\mathrm{Tr}\,\mathbb{1} + \langle \ln|2\pi\Theta|\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)}\right) \\
&= \langle \mathcal{H}(\theta,d)\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} - \frac{1}{2}\langle \ln|2\pi e\Theta|\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)}
\end{aligned} \tag{3.4}$$

with e being the basis of the ln and hence allowing for absorption of $\mathrm{Tr}\,\mathbb{1}$ term into the determinant term. From this one can then continue and perform the variation of $\mathcal{D}_{\mathrm{KL}}$ w.r.t. the parameters of the approximating distribution, $(\bar{\theta}, \Theta)$. In the following an assumption that $\Theta \neq \Theta|_{\theta=\bar{\theta}}$ is made. Reason behind this is to motivate certain properties of the covariance which we wish to keep when the transition to a mean dependent covariance is taken.

Now, variation w.r.t. $\bar{\theta}$ is performed as follows

$$\begin{aligned}
\frac{\partial}{\partial\bar{\theta}}\mathcal{D}_{\mathrm{KL}} &= \int \frac{\partial}{\partial\bar{\theta}}\mathcal{G}(\theta-\bar{\theta},\Theta)\mathcal{H}(\theta,d)\mathrm{d}\theta \\
&= \int \frac{\partial}{\partial\bar{\theta}}\mathcal{G}(\theta',\Theta)\mathcal{H}(\theta'+\bar{\theta},d)\mathrm{d}\theta' \\
&\text{with} \quad \theta' = \theta - \bar{\theta},
\end{aligned} \tag{3.5}$$

and observing

$$\frac{\partial}{\partial\bar{\theta}}\mathcal{G}(\theta',\Theta) \equiv \partial_{\bar{\theta}}\mathcal{G}(\theta',\Theta) = -\partial_{\theta'}\mathcal{G}(\theta',\Theta),$$

the eq. (3.5) becomes

$$\begin{aligned}
\partial_{\bar{\theta}}\mathcal{D}_{\mathrm{KL}} &= -\int \partial_{\theta'}\mathcal{G}(\theta',\Theta)\mathcal{H}(\theta'+\bar{\theta},d)\mathrm{d}\theta' \\
&= \underbrace{\left[(-\mathcal{H}(\theta'+\bar{\theta},d))\mathcal{G}(\theta',\Theta)\right]\big|_{-\infty}^{\infty}}_{=0} + \int \mathcal{G}(\theta',\Theta)\partial_{\theta'}\mathcal{H}(\theta'+\bar{\theta},d)\mathrm{d}\theta' \\
&= \int \mathcal{G}(\theta',\Theta)\partial_{\theta'}\mathcal{H}(\theta'+\bar{\theta},d)\mathrm{d}\theta',
\end{aligned} \tag{3.6}$$

with the first term on the second line, obtained after partial integration, being zero, since $(-\mathcal{H}(\theta'+\bar{\theta},d))$ is bounded from above w.r.t. θ' and $\mathcal{G}(\theta',\Theta) \to 0$ as $\theta' \to \pm\infty$. Therefore,

tracing back the change from θ to θ' and imposing the extremum condition on eq. (3.6), we have the following constraint equation for the mean

$$0 \stackrel{!}{=} \partial_{\bar{\theta}} \mathcal{D}_{\mathrm{KL}} = \langle \partial_\theta \mathcal{H}(\theta, d) \rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)}. \tag{3.7}$$

Regarding the variation of the KL w.r.t. the covariance Θ one has

$$\begin{aligned} \partial_\Theta \mathcal{D}_{\mathrm{KL}} &= \partial_\Theta \left(\langle \mathcal{H}(\theta, d) \rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)} - \frac{1}{2} \langle \ln|2\pi e \Theta| \rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)} \right) \\ &= \frac{1}{2} \left(\frac{\partial^2}{\partial \bar{\theta} \partial \bar{\theta}^\dagger} \langle \mathcal{H}(\theta, d) \rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)} - \partial_\Theta \langle \ln|2\pi e \Theta| \rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)} \right). \end{aligned}$$

Performing the same reasoning as in the case of variation w.r.t. the mean, the derivatives in front of $\langle \mathcal{H}(\theta, d) \rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)}$ can enter the expectation value. Hence we have

$$\partial_\Theta \mathcal{D}_{\mathrm{KL}} = \frac{1}{2} \left(\left\langle \frac{\partial^2}{\partial \theta \partial \theta^\dagger} \mathcal{H}(\theta, d) \right\rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)} - \partial_\Theta \left(\ln|2\pi e \Theta| \right) \right). \tag{3.8}$$

Note that the second term in the above equation is without the expectation value, since the integration was performed over θ and due to the Gaussian being normalized, integration results in multiplication with identity. Here, it was used that $\Theta \neq \Theta(\theta)$. Now, the second term in eq. (3.8) can be evaluated with the use of Jacobi's formula yielding simply

$$\partial_\Theta \left(\ln|2\pi e \Theta| \right) = \Theta^{-1}. \tag{3.9}$$

Inserting this result back in eq. (3.8) we have

$$\partial_\Theta \mathcal{D}_{\mathrm{KL}} = \frac{1}{2} \left(\left\langle \frac{\partial^2}{\partial \theta \partial \theta^\dagger} \mathcal{H}(\theta, d) \right\rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)} - \Theta^{-1} \right), \tag{3.10}$$

and imposing the extremum condition the final constraint equation for the covariance is given through its inverse as

$$\Theta^{-1} = \left\langle \frac{\partial^2}{\partial \theta \partial \theta^\dagger} \mathcal{H}(\theta, d) \right\rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)}. \tag{3.11}$$

This relation can then be expanded and gives the following equation:

$$\Theta^{-1} = \left\langle \frac{\partial}{\partial \theta} \mathcal{H}(\theta, d) \frac{\partial}{\partial \theta^\dagger} \mathcal{H}(\theta, d) \right\rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)} - \left\langle \frac{1}{p(\theta, d)} \frac{\partial^2}{\partial \theta \partial \theta^\dagger} p(\theta, d) \right\rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)}. \tag{3.12}$$

The solution to eq. (3.12) produces a plausible covariance only if the second term is smaller than the first term. Reason being that the covariance needs to be positive-definite, which first term provides and second term prevents if it becomes too large. Therefore, what is proposed instead is to drop the second term, and approximate the inverse covariance by only the first term

$$\Theta^{-1} \approx \left\langle \frac{\partial}{\partial \theta} \mathcal{H}(\theta, d) \frac{\partial}{\partial \theta^\dagger} \mathcal{H}(\theta, d) \right\rangle_{\mathcal{G}(\theta - \bar{\theta}, \Theta)}. \tag{3.13}$$

The neglected second term will be small anyhow for posteriors which are sufficiently Gaussian.

This then serves as an inspiration to make the parameterization of the covariance for the MGVI algorithm. Namely, what is proposed is to approximate the inverse covariance through the use of metric $M^{-1}_{(\theta,d)}$ given as

$$\begin{aligned} M_{(\theta,d)} &\equiv M_{d|\theta} + M_\theta, \\ M_{d|\theta} &= \left\langle \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta} \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta^\dagger} \right\rangle_{p(d|\theta)}, \\ M_\theta &= \left\langle \frac{\partial \mathcal{H}(\theta)}{\partial \theta} \frac{\partial \mathcal{H}(\theta)}{\partial \theta^\dagger} \right\rangle_{p(\theta)}. \end{aligned} \tag{3.14}$$

The expression on the second line represents the Fisher information metric of the likelihood, and its inverse gives a lower bound to the uncertainty of our mean estimate, $\bar{\theta}$. More precisely, the statement is known as the Cramer-Rao bound [6, 10] and it is given as

$$M^{-1}_{d|\theta} \leq \langle (\theta - \bar{\theta})(\theta - \bar{\theta})^\dagger \rangle_{p(d|\theta)}. \tag{3.15}$$

The expression on the last line in eq. (3.14) represents the Fisher information metric of the prior. This term gives the lower bound to the prior variance of the mean [4]:

$$M^{-1}_{\theta} \leq \langle (\theta - \bar{\theta})(\theta - \bar{\theta})^\dagger \rangle_{p(\theta)}. \tag{3.16}$$

Using the two inequalities as shown in eq. (3.16) and eq. (3.15), the authors of [36] claim that the metric

$$M_{(\bar{\theta},d)} \equiv (M_{d|\theta})|_{\theta=\bar{\theta}} + (M_\theta)|_{\theta=\bar{\theta}} \tag{3.17}$$

should fulfill the following inequality for sufficiently Gaussian posteriors:

$$M^{-1}_{(\bar{\theta},d)} \lesssim \langle (\theta - \bar{\theta})(\theta - \bar{\theta})^\dagger \rangle_{p(\theta|d)}. \tag{3.18}$$

This then suggests that the approximating Gaussian should be taken as

$$\mathcal{G}(\theta - \bar{\theta}, \Theta) \equiv \mathcal{G}(\theta - \bar{\theta}, M^{-1}_{(\bar{\theta},d)}), \tag{3.19}$$

thus fully specifying the approximating Gaussian. Note that by $(.)|_{\theta=\bar{\theta}}$ we emphasize that this metric is evaluated at the position of the current mean estimate. This then allows for consistently updating the metric as the posterior mean is inferred during the variational inference. Since the form of the covariance is exactly known, it can be implicitly stored and calculated when needed for uncertainty estimation.

Besides the above, few more reasonable justifications are given in the original paper [36] for having this parameterization.

Firstly, this choice of the metric allows for positive-definiteness to be fulfilled since both $(M_{d|\theta})|_{\theta=\bar{\theta}} \equiv M_{d|\bar{\theta}}$ and $(M_\theta)|_{\bar{\theta}} \equiv M_{\bar{\theta}}$ are Hermitian operators. This property will not be affected when the standardization of the prior is performed, since it corresponds to a unitary transformation of coordinates.

Secondly, it captures the important features of the posterior in two extremes. In the limit of scarce data, likelihood will become completely uninformative, for example in very large inference problems, degrees of freedom the model has can severely surpass the number of constraints data gives. A similar problem will be encountered in section section 4.2.2. In this

case it is of crucial importance that the method is able to be close to the true uncertainty estimates for all the inferred parameters. Indeed in the case of no likelihood contribution at all, only the prior metric $M_{\bar{\theta}}$ will remain which indeed will give us the correct uncertainty estimate for this regime. The other limit is the limit of highly informed setup, or the one of almost infinite data. Here, the Fisher metric of the likelihood will be contributing a lot more than the prior metric to the covariance of the approximating Gaussian. This demonstrates that the choice of parameterization made for the MGVI algorithm allows the approximating Gaussian distribution to fulfill the Bernstein-von Mises theorem [13]. Namely, the theorem states that it must be true that in highly informed settings the prior information is irrelevant. Even though it is not completely clear how the approximation will cope with a regime in between these two, confirming that it has reasonable behaviour in the limits considered here is re-assuring and hence motivates to proceed further.

Before continuing, a brief restatement of the complete parameterization for the MGVI algorithm is given below with a change in notation which would prove useful later when the higher order terms are calculated. We have:

$$\begin{aligned}
\mathcal{Q}(\theta|d) &= \mathcal{G}(\theta - \bar{\theta}, \Theta|_{\theta=\bar{\theta}}), \\
\Theta^{-1}|_{\theta=\bar{\theta}} &= M_{d|\bar{\theta}} + M_{\bar{\theta}} \\
M_{d|\bar{\theta}} &= \left\langle \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta} \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta^\dagger} \right\rangle_{p(d|\theta)} \Bigg|_{\theta=\bar{\theta}} \\
M_{\bar{\theta}} &= \left\langle \frac{\partial \mathcal{H}(\theta)}{\partial \theta} \frac{\partial \mathcal{H}(\theta)}{\partial \theta^\dagger} \right\rangle_{p(\theta)} \Bigg|_{\theta=\bar{\theta}},
\end{aligned} \tag{3.20}$$

where the $\Theta|_{\theta=\bar{\theta}}$, represents the covariance of the approximating Gaussian, and $\Theta^{-1}|_{\theta=\bar{\theta}}$ the corresponding inverse. There is an emphasis in both these expressions on the fact that the metric is evaluated at the point of the current mean $\bar{\theta}$, in order to remind the reader the metric is treated implicitly. The terms on the second and third line of eq. (3.20) are the same as before, as given in eq. (3.14). This can then be translated to the standardized coordinates by performing a transformation:

$$\begin{aligned}
\theta &= \mathrm{CDF}^{-1}_{p(\theta)} \left(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi) \right) \\
&\equiv f(\xi)
\end{aligned} \tag{3.21}$$

where:

$$\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi) = \int_{\xi_0}^{\xi} \mathcal{G}(\xi', \mathbb{1}) \mathrm{d}\xi' \tag{3.22}$$

with ξ_0 denoting the lower bound of the domain of ξ. The $\mathrm{CDF}^{-1}_{p(\theta)}$ represents the inverse cdf transform for the distribution $p(\theta)$ which is defined through the choice of prior structure in the model. Finally the complete operation was denoted by f to emphasize that this is a mapping from the ξ space to θ space. The coordinates ξ would from here on be referred to as the standardized coordinates.

In the standardized coordinates now the approximating Gaussian and the metric will be:

$$\begin{aligned}
\mathcal{Q}(\xi|d) &= \mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}}) \\
\Theta^{-1}|_{\xi=\bar{\xi}} &= \underbrace{J^\dagger_{\bar{\xi}} M_{d|\bar{\xi}} J_{\bar{\xi}}}_{=M_{d|\bar{\theta}}} + \mathbb{1}
\end{aligned} \tag{3.23}$$

where by $\mathcal{Q}(\xi|d)$ the approximating Gaussian has been denoted, and the Jacobian $J_{\bar{\xi}} = |\frac{\partial f}{\partial \xi}|$, reducing the covariance to a rather convenient form from the numerical perspective, since the prior metric reduces to $\mathbb{1}$. With this in mind, the stage is set for diving into the details of how the KL is actually minimized.

3.2 Minimizing the KL

Going now back to the eq. (3.2), it can be seen that the coordinate transformation as performed in eq. (3.21) will not affect the result, since KL is a scalar quantity. Therefore, we can rewrite eq. (3.2) as

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}})||p(\theta(\xi)|d)) \mathrel{\widehat{=}} \langle \mathcal{H}(\xi, d)\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta|_{\xi=\bar{\xi}})}, \tag{3.24}$$

where the self-entropy term was dropped since it is not important if the KL is minimized only w.r.t. the mean $\bar{\xi}$. Then, the gradient of the KL w.r.t. the mean is given by

$$\frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \bar{\xi}} = \langle \mathcal{H}(\xi, d)\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta|_{\xi=\bar{\xi}})}. \tag{3.25}$$

The gradient is calculated using the stochastic estimate of the expectation value, with samples drawn from the approximating distribution $\mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}})$. In other words, with the use of the identity from eq. (3.6), we have

$$\begin{aligned}
\frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \bar{\xi}} &= \left\langle \frac{\partial}{\partial \xi} \mathcal{H}(\xi, d) \right\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta|_{\xi=\bar{\xi}})} \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{H}}{\partial \xi}(\xi, d)\bigg|_{\xi=\xi_i} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{H}}{\partial \xi}(\xi, d)\bigg|_{\xi=\bar{\xi}+\Delta\xi_i}, \\
\text{with} \quad \xi_i &\hookleftarrow \mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}}) \quad \text{and} \quad \Delta\xi_i \hookleftarrow \mathcal{G}(\xi, \Theta|_{\xi=\bar{\xi}}).
\end{aligned} \tag{3.26}$$

Sampling can then be performed by utilizing the following generative process

$$\begin{aligned}
\xi' &\hookleftarrow \mathcal{G}(\xi', \mathbb{1}), \\
n' &\hookleftarrow \mathcal{G}(n', M_{d|\bar{\xi}}), \\
\Delta\xi' &= J_{\bar{\xi}}\xi' + n',
\end{aligned} \tag{3.27}$$

with $\Delta\xi' \hookleftarrow \mathcal{G}(\Delta\xi', (\Theta|_{\xi=\bar{\xi}})^{-1})$ and

$$\Theta|_{\xi=\bar{\xi}} = \left(J_{\bar{\xi}}^{\dagger} M_{d|\bar{\xi}} J_{\bar{\xi}} + \mathbb{1}\right)^{-1}.$$

Now, in order to have the samples distributed with correct covariance given by $\Theta|_{\bar{\xi}}$ one needs to solve the following equation

$$\Delta\xi = \Theta|_{\xi=\bar{\xi}} \Delta\xi'.$$

Now, since the metric $\Theta|_{\xi=\bar{\xi}}$ is constructed to be positive-definite, the Conjugate Gradient algorithm can be used to solve the above equation [7]. This then provides us with $\Delta\xi \hookleftarrow \mathcal{G}(\Delta\xi, \Theta|_{\xi=\bar{\xi}})$ which can be translated to follow the correct mean $\bar{\xi}$

$$\xi \equiv \bar{\xi} + \Delta\xi \hookleftarrow \mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}}).$$

Afterwards, using these samples the gradient of the KL is estimated and a minimization step can be performed. For performing the minimization step we used Newton Conjugate Gradient method. This procedure is then repeated until sufficient level of convergence is achieved.

3.3 Higher order terms

One of the important steps in the MGVI algorithm is to use the stochastic estimate of the gradient of $\mathcal{D}_{\mathrm{KL}}$ w.r.t. the mean (see eq. (3.26)). While doing so, only the first order terms are taken into account, i.e. only the gradient of the $\mathcal{H}(\xi, d)$ is estimated. In order to take into account the higher order terms, the variation of the metric $\Theta|_{\xi=\bar{\xi}}$ w.r.t. the current mean estimate, $\bar{\xi}$, has to evaluated as well. That is what will be done in the following.

First, we rewrite once more the KL-divergence between posterior and approximating Gaussian $\mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}})$

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}}), p(\xi|d)) = \langle \mathcal{H}(\xi|d) + \ln(\mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}}))\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta|_{\xi=\bar{\xi}})}. \tag{3.28}$$

In the following the $(.)|_{\xi=\bar{\xi}}$ will be omitted from $\Theta|_{\xi=\bar{\xi}}$ for notational convenience, but the dependence of course should be kept mentally. Now, in order to optimize for the mean, one looks at the extremum of the KL-divergence given above w.r.t. the mean $\bar{\xi}$:

$$\begin{aligned}\frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \bar{\xi}} &= \int_{\Omega_\xi} \frac{d\mathcal{G}(\xi - \bar{\xi}, \Theta)}{d\bar{\xi}} \left[\underbrace{\mathcal{H}(\xi|d) + \ln(\mathcal{G}(\xi - \bar{\xi}, \Theta))}_{=F(\xi,\bar{\xi},d,\theta)}\right] + \underbrace{\int_{\Omega_\xi} \mathcal{G}(\xi - \bar{\xi}, \Theta) \frac{d \ln \mathcal{G}(\xi - \bar{\xi}, \Theta)}{d\bar{\xi}}}_{=0, \text{ since Gaussian is normalized}} \\ &= \left\langle \frac{\partial}{\partial \xi} F(\xi, \bar{\xi}, d, \Theta) \right\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta)} + \int_{\Omega_\xi} F(\xi, \bar{\xi}, d, \Theta) \frac{\partial \mathcal{G}(\xi - \bar{\xi}, \Theta)}{\partial \Theta} \frac{\partial \Theta}{\partial \bar{\xi}}. \end{aligned} \tag{3.29}$$

Note that from the first to the second line a partial integration was done. Now, the second term in the above equation appears since the covariance is parameterized through the mean in the MGVI algorithm. This term will be treated properly after evaluating the first term of the above equation:

$$\begin{aligned}\left\langle \frac{\partial}{\partial \xi} F(\xi, \bar{\xi}, d, \Theta) \right\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta)} &= \left\langle \frac{\partial}{\partial \xi} \mathcal{H}(\xi|d) \right\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta)} + \underbrace{\langle \Theta^{-1}(\xi - \bar{\xi})\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta)}}_{=0, \text{ since } \xi - \bar{\xi} \text{ has zero mean}} \\ &\quad + \left\langle \underbrace{\frac{\partial}{\partial \xi} \left[\frac{1}{2} \ln |2\pi\Theta|\right]}_{=0, \text{ since } |\Theta| \neq |\Theta(\xi)|} \right\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta)} \\ &= \left\langle \frac{\partial}{\partial \xi} \mathcal{H}(\xi|d) \right\rangle_{\mathcal{G}(\xi-\bar{\xi},\Theta)}. \end{aligned} \tag{3.30}$$

In the first line of eq. (3.30), the term $\ln(\mathcal{G}(\xi - \bar{\xi}, \Theta))$ was immediately expanded and the derivative was calculated explicitly. Hence, one can see that the first term would contribute only through the gradient field of the Hamiltonian. But, when it comes to the second term, additional terms appear. Let us first focus on evaluating the Gaussian variation w.r.t. the covariance. Using coordinate notation and substitution

$$q(\xi, \bar{\xi}, \Theta) = -(1/2)(\xi - \bar{\xi})^\dagger \Theta^{-1} (\xi - \bar{\xi}),$$

one has

$$\begin{aligned}\frac{\partial \mathcal{G}(\xi-\bar{\xi},\Theta)}{\partial\Theta_{ij}} &= \left[\frac{\partial}{\partial\Theta_{ij}}(|2\pi\Theta|)^{-1/2}\right]\exp(q(\xi,\bar{\xi},\Theta))\\ &+ |2\pi\Theta|^{-1/2}\frac{\partial}{\partial\Theta_{ij}}\left[-\frac{1}{2}(\xi-\bar{\xi})^l(\Theta^{-1})_{lm}(\xi-\bar{\xi})^m\right]\exp(q(\xi,\bar{\xi},\Theta))\\ &= \left(-\frac{1}{2}\right)\left(|2\pi\Theta|^{-1/2}(\Theta^{-1})_{ij}\exp(q(\xi,\bar{\xi},\Theta)\right)\\ &+ \left(-\frac{1}{2}\right)|2\pi\Theta|^{-1/2}\left((\xi-\bar{\xi})^l\frac{\partial(\Theta^{-1})_{lm}}{\partial\Theta_{ij}}(\xi-\bar{\xi})^m\right)\exp(q(\xi,\bar{\xi},\Theta))\\ &= \frac{1}{2}\mathcal{G}(\xi-\bar{\xi},\Theta)\left(-(\Theta^{-1})_{ij}+(\Theta^{-1})_{li}(\xi-\bar{\xi})^l(\Theta^{-1})_{jm}(\xi-\bar{\xi})^m\right). \end{aligned} \tag{3.31}$$

In the above equation, passing from the second equality sign to the last, the following identity was used:

$$\begin{aligned}\frac{\partial}{\partial\Theta_{ij}}((\Theta^{-1})^{ln}\Theta_{nm}) &= 0\\ &= \frac{\partial}{\partial\Theta_{ij}}((\Theta^{-1})^{ln})\Theta_{nm}+(\Theta^{-1})^{ln}\frac{\partial}{\partial\Theta_{ij}}(\Theta_{nm})\\ &= \frac{\partial(\Theta^{-1})^{ln}}{\partial\Theta_{ij}}\Theta_{nm}+(\Theta^{-1})^{ln}\delta_i^n\delta_j^m,\end{aligned}$$

from the above one has

$$\frac{\partial(\Theta^{-1})^{ln}}{\partial\Theta_{ij}} = -(\Theta^{-1})^{li}(\Theta^{-1})^{jn}.$$

Regarding the variation of covariance w.r.t. the mean in the MGVI parameterization one has (following the definitions in eq. (3.20)), again in coordinate form

$$\begin{aligned}\frac{\partial\Theta^{-1}_{ij}}{\partial\bar{\xi}^l} &= \frac{\partial}{\partial\bar{\xi}^l}((M_{d|\bar{\xi}})_{ij})+\frac{\partial}{\partial\bar{\xi}^l}((M_{\bar{\xi}})_{ij})\\ &\widehat{=} \frac{\partial}{\partial\bar{\xi}^l}\left\langle\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^i}\left(\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^j}\right)^\dagger\right\rangle_{p(d|\bar{\xi})}\\ &= \int_{\Omega_d}\frac{\partial^2\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^i\partial\bar{\xi}^l}\left(\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^j}\right)^\dagger p(d|\bar{\xi})+\text{adjoint}\\ &+ \int_{\Omega_d}\frac{\partial p(d|\bar{\xi})}{\partial\bar{\xi}^l}\frac{\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^i}\left(\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^j}\right)^\dagger\\ &= \int_{\Omega_d}p(d|\bar{\xi})\left(\frac{\partial^2\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^i\partial\bar{\xi}^l}\left(\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^j}\right)^\dagger+\text{adjoint}+\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^l}\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^i}\left(\frac{\partial\mathcal{H}(d|\bar{\xi})}{\partial\bar{\xi}^j}\right)^\dagger\right)\\ &= 2\Gamma_{ij,l}+T_{ijl},\end{aligned}$$

with

$$\Gamma_{ij,l} = \left\langle \frac{\partial^2 \mathcal{H}(d|\bar{\xi})}{\partial \bar{\xi}^i \partial \bar{\xi}^l} \left(\frac{\partial \mathcal{H}(d|\bar{\xi})}{\partial \bar{\xi}^j} \right)^\dagger \right\rangle_{p(d|\bar{\xi})}$$
$$T_{ijl} = \left\langle \frac{\partial \mathcal{H}(d|\bar{\xi})}{\partial \bar{\xi}^l} \frac{\partial \mathcal{H}(d|\bar{\xi})}{\partial \bar{\xi}^i} \left(\frac{\partial \mathcal{H}(d|\bar{\xi})}{\partial \bar{\xi}^j} \right)^\dagger \right\rangle_{p(d|\bar{\xi})}. \tag{3.32}$$

In the above eq. (3.32) the variation of the prior metric was omitted from passing to the second line, since once the form of the prior is given it is possible to explicitly calculate that variation[1]. Also, in the last line, it was recognized that the Amari's 0-connection and three index tensor appear which make a connection towards the information geometric formulation [15]. This could be interesting, since the formalism Amari develops in his information geometric approach to statistical inference borrows many concepts from differential geometry which allows for inducing some global properties of the manifolds by only working on local scales. This connection between local and global properties could prove useful for the variational inference as performed by MGVI, since it could offer more insight into how to use the local information provided by the Gaussian approximation. Therefore, it would be interesting to make the connection deeper and see whether through this the current MGVI parameterization can be better understood and expanded upon. This however will not be discussed in this thesis and is left for future work.

Now we have all the terms necessary to calculate the full form of the second term of eq. (3.29). Plugging the results of eq. (3.32) and eq. (3.31) we have (in coordinate notation)

$$\int_{\Omega_\xi} F(\xi, \bar{\xi}, d, \Theta) \frac{\partial \mathcal{G}(\xi - \bar{\xi}, \Theta)}{\partial \Theta_{ij}} \frac{\partial \Theta_{ij}}{\partial \bar{\xi}^m} = \int_{\Omega_\xi} F(\xi, \bar{\xi}, d, \Theta) \frac{\partial \mathcal{G}(\xi - \bar{\xi}, \Theta)}{\partial \Theta_{ij}} \frac{\partial \Theta_{ij}}{\partial (\Theta^{-1})_{kl}} \frac{\partial (\Theta^{-1})_{kl}}{\partial \bar{\xi}^m}, \tag{3.33}$$

where

$$\frac{\partial \mathcal{G}(\xi - \bar{\xi}, \Theta)}{\partial \Theta_{kl}} \frac{\partial \Theta_{kl}}{\partial (\Theta^{-1})_{ij}} \frac{\partial (\Theta^{-1})_{ij}}{\partial \bar{\xi}^m} = \underbrace{\frac{\partial \mathcal{G}(\xi - \bar{\xi}, \Theta)}{\partial \Theta_{kl}} \left[-(\Theta^{ki})(\Theta^{jl}) \right]}_{=\#} \left[2\Gamma_{ij,m} + T_{ijm} \right],$$

with

$$\begin{aligned} \# &= \frac{1}{2} \mathcal{G}(\xi - \bar{\xi}, \Theta) \left[(-(\Theta^{-1})_{kl})(-\Theta^{ki})(\xi^{jl}) + (\xi - \bar{\xi})^i \delta_i^j \delta_i^j (\xi - \bar{\xi})^j \right] \\ &= \frac{1}{2} \mathcal{G}(\xi - \bar{\xi}, \Theta) \left[\Theta^{ij} + |(\xi - \bar{\xi})|^2 \delta_j^i \right], \end{aligned}$$

hence, the eq. (3.33) becomes

$$\int_{\Omega_\xi} \left(\frac{1}{2} \right) \mathcal{G}(\xi - \bar{\xi}, \Theta) F(\xi, \bar{\xi}, d, \Theta) (\Theta^{ij} (2\Gamma_{ij,m} + T_{ijm}) + |\xi - \bar{\xi}|^2 (2\Gamma_{ii,m} + T_{iim})). \tag{3.34}$$

Now, if one would take a look at a fully Gaussian case, i.e. for a model $d = Rs + n$, where the prior and noise distribution are Gaussian with zero mean, $\mathcal{G}(\xi, S)$, $\mathcal{G}(n, N)$ respectively, one

[1]Typically one chooses a simple form for the prior, therefore the form of this variation is usually also simple

has for the likelihood $p(d|\xi) = \mathcal{G}(d - R\xi, N)$ and hence the posterior metric would be

$$\begin{aligned} M = M_{d|\xi} + M_{\xi} &= \left\langle \frac{\partial \mathcal{H}(d|\xi)}{\partial \xi} \frac{\partial \mathcal{H}(d|\xi)}{\partial \xi^\dagger} \right\rangle_{p(d|\xi)} + \left\langle \frac{\partial \mathcal{H}(\xi)}{\partial \xi} \frac{\partial \mathcal{H}(\xi)}{\partial \xi^\dagger} \right\rangle_{p(\xi)} \\ &= R^\dagger N^{-1} R + S^{-1}, \end{aligned} \tag{3.35}$$

which is expected. Hence, in the fully Gaussian case the contribution of the eq. (3.34) would be zero, since none of the terms in eq. (3.35) depend on the mean. But, if one takes a look at the Poissonian case, there the Fisher metric explicitly depends on the mean:

$$\begin{aligned} p(d|\lambda) &= \frac{\lambda^d \exp(-\lambda)}{d!} \\ \left\langle \frac{\partial \mathcal{H}(d|\lambda)}{\partial \lambda} \frac{\partial \mathcal{H}(d|\lambda)}{\partial \lambda^\dagger} \right\rangle_{p(d|\lambda)} &= \left\langle - \frac{\partial^2}{\partial \lambda^2} \mathcal{H}(d|\lambda) \right\rangle_{p(d|\lambda)} \\ &= \frac{1}{\lambda^2} \langle d \rangle_{p(d|\lambda)} \\ &= \frac{1}{\lambda}. \end{aligned} \tag{3.36}$$

Here a 1D Poisson distribution was considered. In a multidimensional case the Fisher metric would be a corresponding diagonal matrix. It is then expected that in the case of eq. (3.36), one would have a non-vanishing contribution from the terms in eq. (3.34). It would be interesting to see how exactly the correction terms look like in this case, but due to the time limitation of the thesis the focus was turned to the topics explained in the next sections, hence this was left for future work. What will most likely be worthwhile to do in order to gain insight is to implement the higher order terms as depicted in eq. (3.34) and try to calculate them numerically for different small scale problems. It can be immediately seen that the computational cost for going to larger problems is high, since one needs to evaluate not only the expectation value of the complicated objects but the trace terms as well.

With this, the section on the MGVI algorithm is concluded. Now, the focus turns to applying this method to different settings and testing its capabilities.

4 Bayesian Causal Inference

In this chapter an attempt is made to tackle the problem of causal inference within the *Information Field Theory* (IFT) formalism. Specifically, the focus is on the problem of bivariate causal inference ($X \to Y$, $Y \to X$) as well as inferring the existence of a confounder ($X \leftarrow Z \to Y$), given an observed dataset (X, Y). Bivariate causal discovery is especially interesting since the usual methods of statistical independence testing are not useful in this regime. Even more so, the problem of inferring the existence of a confounder Z requires both inferring this latent variable while probing for the correct causal structure. Here the solution of the problem is pursued through the use of Bayesian hierarchical modelling and Additive Noise Models. For different generative models ($X \to Y$, $Y \to X$, $X \leftarrow Z \to Y$) a Bayesian inference algorithm is developed, which tries to learn all the quantities of interest non-parameterically within the IFT context. A method for estimating evidences for each of the considered models is developed and it is used as a model score for deciding which causal direction has most support by the data. In the end, a comparison is made with previously published methods and it is shown that comparable accuracy can be achieved on the benchmark tests considered therein.

4.1 Introduction

The pioneering works of Pearl [11, 31] offered various methods for causal discovery. The *do*-calculus approach which was developed as a result, provided a mathematical foundation within which the problem of causality inference can be clearly formulated. This formalism relies on the possibility to do interventions, which are a mathematical description for how the original system which generated the data would behave under perturbations that affect some of the variables while keeping the others constant. All this is used in order to obtain additional information on the causal relationship between the random variables, allowing us to make further statements about causal (in)dependence. Still, on many real world systems interventions are not possible. For example most astronomical systems or the Earth climate are not accessible to interventions for the purpose of studying causalities. Therefore, developing methods which would be able to tackle the problem of causal discovery from just observational data on such systems is still an interesting open problem. One of the common approaches to causality inference is to perform conditional independence tests of the observed variables [11, 12, 18]. This approach however is not very useful in the bivariate scenario, where the observed variables ($\mathbf{d} = (X, Y)$) are expected to be dependent in most of the cases.

In this paper we try to instead use the Bayesian formalism, restricting our decision process between the $X \to Y$, $Y \to X$ and $X \leftarrow Z \to Y$ causal models which we believe are the most important causal structures for the case with data $\mathbf{d}$. In order to infer all the necessary quantities needed to make a decision we use Bayesian hierarchical modelling coupled with a variational inference approach. The Bayesian hierarchical models we consider here are generative models that mirror the nature of *Structural Equation Models* (SEMs). This setting is especially useful since SEMs contain more information than the family of all intervention distributions together with the corresponding observational distributions (see proposition 9 in [24] and discussion around it). Therefore, we believe that our Bayesian hierarchical models give us enough predictive power to discover the corresponding SEM and hence the appropriate causal structure.

Besides this, we restrict our models with the requirement for noise to be additive. The rea-

son behind using this assumption is that in this setting the identifiability between $X \to Y$ and $Y \to X$ causal structures can be guaranteed in the limit of an infinite amount of data, especially for nonlinear mappings between the random variables [19, 27]. This can also be extended to the case of $X \leftarrow Z \to Y$ as pursued in [23], where it is shown that partial identifiability can be guaranteed for the case when the mappings are taken to be invertible. Nonetheless, we proceed with applying our confounder model and partially address the problem of identifiability.

The structure of the following chapters is the following. In section 4.2 a description of all considered models is given. Afterwards, in section 4.3 a short overview of the variational inference scheme is given and in section 4.4 an outline of the calculation for the model evidences is described. It is then discussed how this can be used to decide between generative models. We give a concise description of the datasets we used to test our methods in section 4.5 and discuss the results of those tests in section 4.6.

4.2 Inference models

The building blocks of the inference machinery developed here are in essence Gaussian processes through which we encode our prior knowledge. This is possible to achieve, since the knowledge present inside the specific choice of a generative model can be built into the likelihood. Hence, leaving us with a very simple standardized Gaussian random field prior for the inference [28]. Besides some of the numerical advantages this approach brings [36], it also allows us to interpret the implemented model directly as an SEM. Major consequence of this is that we can claim identifiability of the models considered here by restricting us to certain classes of SEMs. More precisely, under the assumption of additive noise and allowing for nonlinear mappings to be possible between the random variables the identifiability conditions can be met [19, 23, 26]. Therefore, it is expected that in a generic case one of the three causal scenarios $X \to Y$, $Y \to X$ and $X \leftarrow Z \to Y$ can be discovered by the models developed here. With this in mind we continue with elaborating in more detail the models compared here.

4.2.1 Models: $X \to Y$ and $Y \to X$

During the course of this thesis, several bivariate models were implemented and tried on the benchmark tests described in section 4.5. Below, a description is given of two different models which proved to be the most promising in terms of their overall performance on the benchmark tests.

4.2.1.1 Version 1

In this section we will be concerned with inferring the underlying causal direction given the dataset $\mathbf{d} = (X, Y)$. If indeed there is a causal relationship between the observed variables X and Y then we would expect

$$Y := f_X(X) + \epsilon_X \tag{4.1}$$

if X causes Y ($X \to Y$), with f being a certain deterministic mapping between the spaces of X and Y, and ϵ_X denoting some noise realisation. The symbol ":=" indicates value assignment and is borrowed from the computer programming notation. This emphasizes the fact that in order to obtain the Y data, first the X and ϵ_X have to be generated, which aligns with the

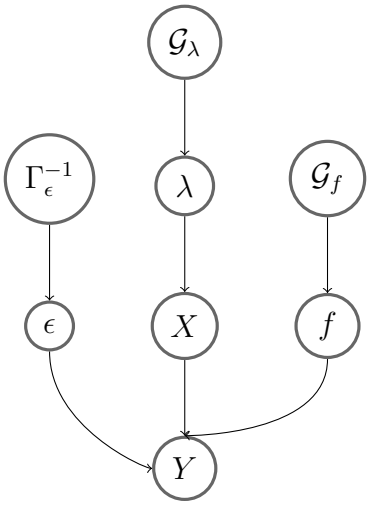


Figure 4.1: Graph structures for the bivariate model $X \to Y$. The graph for $Y \to X$ is obtained by exchanging X and Y.

assumed causal direction in this case. Or, the second case if the opposite causal direction ($Y \to X$) is true

$$X := f_Y(Y) + \epsilon_Y$$

with once again a deterministic mapping f_Y and a noise realisation ϵ_Y. Therefore, the problem boils down to inferring the probability distribution of the cause variable, the mapping and the noise realization. Depending on whether the model we are looking at is $X \to Y$ or $Y \to X$ we would be inferring the probability distribution of X, mapping f_X and ϵ_X or the probability distribution of Y, mapping f_Y and ϵ_Y respectively. In the following, since these two cases are completely symmetric, we would focus on the $X \to Y$ direction and it should be understood that the opposite $Y \to X$ direction would follow completely analogously.

Since the data is expected to be discretized to a certain degree the probability density of the cause variable X is taken as in [29]. Namely, we assume here that the probability distribution of the cause variable is described by a Poisson process. This process can be characterized with a rate field, which we will denote with λ. Since no further information on the generation process for X is given, this is a minimalistic assumption. Now, the λ field is expected to be positive, therefore a log-normal prior is taken and the generative model for λ can be expressed as

$$\begin{aligned} \lambda &= \exp(s_\lambda), \\ \text{with} \quad p(s_\lambda | M_b^{(1)}) &= \mathcal{G}(s_\lambda, S_\lambda). \end{aligned} \tag{4.2}$$

In the eq. (4.2) a new field s_λ has been introduced and exponentiated in order to enforce positiveness, and it is Gaussian distributed with a Gaussian process kernel S_λ. In other words, S_λ represents simply the prior covariance. Finally, within the $M_b^{(1)}$ all the information about the assumptions made is kept. The superscript $^{(1)}$ is there to distinguish assumptions made in this section from the ones made in section 4.2.1.2 and the subscript 'b' is to stand short for 'bivariate', since in this section we are dealing with the bivariate model only. The assumptions will be clearly stated as the model pieces are discussed, and once the whole model is presented all the assumptions will be restated together once more.

The prior covariance is assumed to be unknown and therefore has to be inferred from the data. In order to achieve this, few assumptions are made. The first one is reflecting the fact that we don't want to single out any particular value for the cause variable a priori and therefore statistical homogeneity is assumed for the prior structure of the λ field. According to the Wiener-Khinchin theorem [1, 2], the covariance S_λ is then going to be diagonal in Fourier space

$$(S_\lambda)_{kk'} \propto \delta(k - k')p_{S_\lambda}(|k|), \tag{4.3}$$

with the power spectrum $p_{S_\lambda}(|k|)$. Therefore, instead of inferring the full covariance S_λ the problem boils down to inferring the power spectrum $p_{S_\lambda}(|k|)$. Now, since the MGVI algorithm assumes standardized coordinates (refer to section 3 to see why this is beneficial) it is best to rewrite the problem in terms of the amplitude spectrum

$$(A_\lambda)_{kk'} \propto \delta(k - k')\sqrt{p_{S_\lambda}(|k|)} \equiv \delta(k - k')p_\lambda(|k|). \tag{4.4}$$

To see this is an equivalent formulation, remember that S_λ is real positive-definite and symmetric. Therefore, it can be decomposed as $S_\lambda = A_\lambda A_\lambda^T$ always. Inserting this back in eq. (4.3), eq. (4.4) follows immediately. Now, one can rewrite the generative model for the λ field as a Gaussian process

$$\begin{aligned} \lambda(\xi_\lambda) &= \exp(\mathbb{F}^{-1}A_\lambda\xi_\lambda), \\ \text{with} \quad s_\lambda(\xi_\lambda) &= \mathbb{F}^{-1}A_\lambda\xi_\lambda, \end{aligned}$$

where $\mathbb{F}^{-1}$ is the inverse Fourier transformation, and $\xi_\lambda \hookleftarrow \mathcal{G}(\xi_\lambda, \mathbb{1})$ represents the standardized prior. Once more the problem reduces further, from inferring the covariance S_λ to inferring the amplitude spectrum A_λ, or in other words the field $p_\lambda(|k|)$. Hence, the next task is to obtain a model for $p_\lambda(|k|)$ which is flexible enough to be compatible with different scenarios met in practice and is as well able to properly capture the uncertainty of the λ field.

First of all, the amplitude spectrum is a positive quantity. Therefore, the generative model for $p_\lambda(|k|)$ has to be written as

$$p_\lambda(|k|) = \exp(\gamma_\lambda(|k|)) \tag{4.5}$$

with $\gamma_\lambda(|k|)$ modelled non-parametrically through an integrated Wiener process in order to restrain from choosing a particular functional basis. Since the amplitude spectrum $p_\lambda(|k|)$ has to be positive, its logarithm is modelled instead, i.e. the $\gamma_\lambda(|k|)$ field. Furthermore, inspired by many different physical processes, the dependency of this logarithmic amplitude on the k-modes is modelled in a logarithmic coordinate system. Therefore, the whole amplitude forward model is done in double-logarithmic scale. Note that this implies the zero mode has to be constrained through other means, since it cannot be represented on the logarithmic scale. This is done through identifying the zero mode with an overall scaling factor, with a log-normal prior imposed to ensure its positiveness

$$(A_\lambda)_{00} \equiv \alpha_0 = \exp(\mu_{\alpha_0} + \sigma_{\alpha_0}\xi_{\alpha_0}), \tag{4.6}$$

with α_0 denoting the zero mode value, and $(\mu_{\alpha_0}, \sigma_{\alpha_0})$ being the mean and standard deviation of this Gaussian process with an excitation ξ_{α_0}.

Coming back now to the $\gamma_\lambda(|k|)$ term. Its precise form is given by

$$\gamma_\lambda(|k|) = c_0 + m|k| + \eta \int_{|k_0|}^{|k|} \int_{|k_0|}^{|k'|} \xi_W(|k''|) d|k'| d|k''|,$$
$$\tilde{U} = \int_{k\neq 0} e^{2\gamma_\lambda(|k|)} d|k| \ , \tag{4.7}$$

with c_0 representing an overall integration constant which has to be fixed and will be dealt with shortly. The parameter m is describing the expected slope with an imposed Gaussian prior, and the η parameter represents the strength of the smooth deviations from the linear part of the equation. For this parameter a log-normal prior is chosen. Finally, the precise shape of these deviations from the slope m is captured by the Gaussian excitation field $\xi_W \hookleftarrow \mathcal{G}(\xi_W, \mathbb{1})$. All of these parameters are inferred from the data using the MGVI approximation. The last term in eq. (4.7), $\tilde{U}$, is nothing else than the total power contained in all modes except the $k = 0$ mode.

In order to fix the integration constant c_0 in eq. (4.7), we set the expectation value of the total real space fluctuations of the λ field, which are given through

$$p_\lambda(|k|) = a \ \frac{\exp(\gamma_\lambda(|k|))}{\sqrt{\tilde{U}}}, \ \forall k \neq 0, \tag{4.8}$$

where a represents the strength of the expected real space fluctuations of λ. The value of a is inferred from the data as well and has an imposed log-normal prior. Summarizing the prior structures for all of these parameters together we have

$$\begin{aligned} \alpha_0 &= \exp(\mu_{\alpha_0} + \sigma_{\alpha_0}\xi_{\alpha_0}) \\ m &= \mu_m + \sigma_m \xi_m \\ \eta &= \exp(\mu_\eta + \sigma_\eta \xi_\eta) \\ a &= \mu_a + \sigma_a \xi_a \\ \text{with} \ \ \xi_j &\hookleftarrow \mathcal{G}(\xi_j, \mathbb{1}) \text{ for } i \in \{\alpha_0, m, \eta, a\}. \end{aligned} \tag{4.9}$$

Therefore, this can also be summarized by setting $\xi_{A_\lambda} = (\xi^{(\lambda)}_{\alpha_0}, \xi^{(\lambda)}_m, \xi^{(\lambda)}_\eta, \xi^{(\lambda)}_a, \xi_W)$. The $(.)^{(\lambda)}$ superscript is there to emphasize that these Gaussian excitations are assigned to the forward model of the λ field. In total, the full forward model of the λ field is

$$\begin{aligned} \lambda(\boldsymbol{\xi}_\lambda) &= \exp(\mathbb{F}^{-1}(A_\lambda(\xi_{A_\lambda})(\xi_\lambda))), \\ \text{with} \ \ \boldsymbol{\xi}_\lambda &= (\xi_{A_\lambda}, \xi_\lambda). \end{aligned} \tag{4.10}$$

For further details on modelling the correlation structure reader is referred to the following publications [33, 37].

Since with eq. (4.10) all the necessary quantities for the generative model of the $\lambda(\boldsymbol{\xi}_\lambda)$ field have been introduced, we now continue towards defining the probability distribution for the cause variable. This can be written as

$$p(X|\lambda, M_b^{(1)}) = \prod_{i=1}^{N_X} \frac{(\lambda_{(i)}(\boldsymbol{\xi}_\lambda))^{x_i} \exp(-\lambda_{(i)}(\boldsymbol{\xi}_\lambda))}{x_i!}, \tag{4.11}$$

where by N_X the total number of bins is denoted at which counts of X were collected, by x_i the number of counts within a given bin i and the Poisson statistics was assumed for the bin counts. The quantity $\lambda_{(i)}(\boldsymbol{\xi}_\lambda)$ represents the λ field evaluated at the position of bin i.

Rephrasing the likelihood in terms of an information Hamiltonian, eq. (4.11) becomes

$$\begin{aligned}\mathcal{H}(X|\lambda, M_b^{(1)}) &\equiv -\ln p(X|\lambda, M_b^{(1)}) \\ &= \sum_{i=1}^{N_X} \left(\lambda_{(i)}(\boldsymbol{\xi}_\lambda) - x_i \ln \lambda_{(i)}(\boldsymbol{\xi}_\lambda) + \ln(x_i!)\right)\end{aligned} \tag{4.12}$$

with $\mathcal{H}(X|\lambda, M_b^{(1)})$ denoting the information Hamiltonian.

For the second part of the model we look now at eq. (4.1) and drop the subscript X from f_X and ϵ_X, since the case for the $Y \to X$ causal direction is completely analogous. We assume Gaussian statistics for the field f, representing the mapping between the cause and the effect variable, and noise ϵ. This then allows to write down the likelihood for Y as

$$\begin{aligned}p(Y|X, f, M_b^{(1)}) &= \int p(Y|X, f, \epsilon, M_b^{(1)}) p(\epsilon|M_b^{(1)}) d\epsilon = \int \delta(Y - f(X)) \mathcal{G}(\epsilon, \Sigma) \\ &= \mathcal{G}(Y - f(X), \Sigma),\end{aligned}$$

where by Σ the unknown noise covariance is denoted. This covariance is assumed to be diagonal and with unknown noise variance on its diagonal denoted by σ_ϵ^2. In other words

$$\Sigma = \sigma_\epsilon^2 \, \mathbb{1}_\Sigma, \tag{4.13}$$

with $\mathbb{1}_\Sigma$ denoting a unit operator on the space on which Σ is defined. The values of the unknown variance σ_ϵ^2 are constrained with an inverse gamma prior through the following generative process

$$\sigma_\epsilon^2(\xi_\epsilon) = \mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_\epsilon, \beta_\epsilon)}\left(\mathrm{CDF}_{\mathcal{G}(\xi, \mathbb{1})}(\xi_\epsilon)\right) \tag{4.14}$$

with

$$\mathrm{CDF}_{\mathcal{G}(\xi, \mathbb{1})}(\xi_\epsilon) = \int_{-\infty}^{\xi_\epsilon} \mathcal{G}(\xi, \mathbb{1}) d\xi,$$

and the $\mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_\epsilon, \beta_\epsilon)}$ represents the inverse transform of the CDF of the inverse gamma distribution with parameters $(\alpha_\epsilon, \beta_\epsilon)$. Namely, it is the inverse of

$$\mathrm{CDF}_{\Gamma^{-1}(\alpha_\epsilon, \beta_\epsilon)}(\xi_\Gamma) = \int_{-\infty}^{\xi_\Gamma} \Gamma^{-1}(\xi; \alpha_\epsilon, \beta_\epsilon) d\xi. \tag{4.15}$$

Choosing the inverse gamma distribution as a prior for $(\sigma_\epsilon)^2$ enforces positivity and allows for higher values to be sampled because of its heavy tails. What is expected from the generative model is to start initially with higher values for the variance, slowly reducing them as the model converges towards a plausible posterior mean.

Regarding the field f, a generative process is as well taken for the prior. Similarly as in the case of the λ field, we do not want to choose any particular functional basis and we want the model to be sufficiently flexible in order to capture all the relevant features in the data. This is achieved through the following generative model

$$f(\xi_f, A_f) = \mathbb{F}^{-1}(A_f(\xi_{A_f})(\xi_f)) \tag{4.16}$$

where by A_f the amplitude spectrum for this filed is denoted with corresponding Gaussian excitation field ξ_{A_f} controlling its precise shape. Once again, this is done through utilizing five parameters in total, hence $\xi_{A_f} = (\xi^{(f)}_{\alpha_0}, \xi^{(f)}_m, \xi^{(f)}_\eta, \xi^{(f)}_a, \xi^{(f)}_W)$. Note that these parameters have completely analogous descriptions as mentioned earlier. For convenience, again we collect $\boldsymbol{\xi}_f = (\xi_f, \xi_{A_f})$. In total, the likelihood information Hamiltonian for the Y data, the effect variable, can be written down as

$$\mathcal{H}(Y|f, X, \sigma_\epsilon, M_b^{(1)}) = \frac{1}{2}\left(\frac{Y - f(\boldsymbol{\xi}_f)[X]}{\sigma_\epsilon(\xi_\epsilon)}\right)^T \left(\frac{Y - f(\boldsymbol{\xi}_f)[X]}{\sigma_\epsilon(\xi_\epsilon)}\right) + \frac{1}{2}\ln|2\pi\Sigma| \tag{4.17}$$

Collecting now the likelihoods from eq. (4.12) and eq. (4.17) together, the total likelihood information Hamiltonian for this model is going to be

$$\begin{aligned}\mathcal{H}((X,Y)|f(\boldsymbol{\xi}_f), \sigma_\epsilon(\xi_\epsilon), \lambda(\boldsymbol{\xi}_\lambda), M_b^{(1)}) =& \mathcal{H}(X|\lambda(\boldsymbol{\xi}_\lambda), M_b^{(1)}) + \mathcal{H}(Y|X, f(\boldsymbol{\xi}_f), \sigma_\epsilon(\xi_\epsilon), M_b^{(1)}) \\ =& \sum_i^n \left(\lambda_{(i)}(\boldsymbol{\xi}_\lambda) - x_i \ln \lambda_{(i)}(\boldsymbol{\xi}_\lambda)\right) \\ &+ \frac{1}{2}\left(\frac{Y - f(\boldsymbol{\xi}_f)[X]}{\sigma_\epsilon(\xi_\epsilon)}\right)^T \left(\frac{Y - f(\boldsymbol{\xi}_f)[X]}{\sigma_\epsilon(\xi_\epsilon)}\right) + \frac{1}{2}\ln|\Sigma| \\ &+ \mathcal{H}_0,\end{aligned} \tag{4.18}$$

where the terms irrelevant for the variational inference algorithm are collected in

$$\mathcal{H}_0 = \sum_{i=1}^N \ln(x_i!) + \frac{N_\Sigma}{2}\ln 2\pi. \tag{4.19}$$

Here, by N_Σ the dimensionality of noise covariance Σ is denoted. As promised, after the whole model is presented for this section, all the assumptions that were put in through $M_b^{(1)}$ are summarized as follows

- For all fields we assume statistical homogeneity. By the Wiener-Khinchin theorem [1, 2] then follows that the Fourier transform of the 2-point correlation operator of our fields is diagonal.

- For fields f and λ, a certain degree of smoothness is preferred in the model. This is enforced through their correlation structure terms A_f and A_λ respectively.

- The grid size N_X is assumed fixed and provided in advance before the inference of other unknown quantities is performed

- The noise variable ϵ is assumed to be Gaussian distributed, but with an unknown variance. The variance therefore becomes a model parameter which is inferred from the data. The variance is assumed to be a priori inverse gamma distributed. This guarantees the positiveness of the variance as well as allowing for larger variances by the prior having a heavy tail.

- The data pairs (X, Y) are assumed to be drawn iid from their joint distribution $P(X, Y)$.

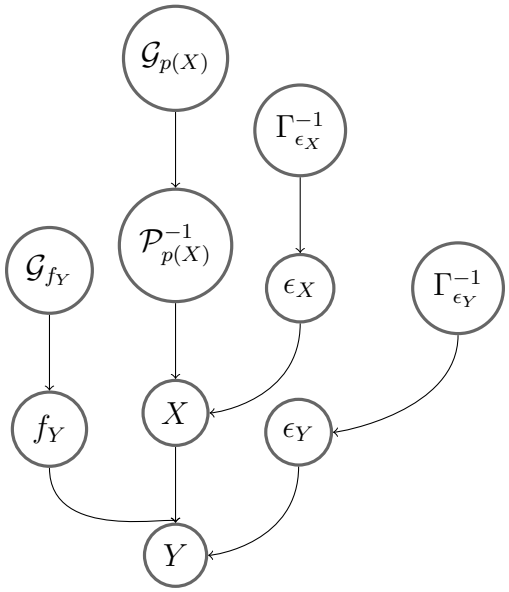


Figure 4.2: Graph structure for the version 2 bivariate model. Note that the model for the reverse direction $Y \to X$ would be obtained by just exchanging X and Y nodes.

Now, to make a connection with SEMs we rewrite the model as following

$$\begin{aligned} \ln \lambda &:= \mathrm{GP}(\boldsymbol{\xi}_\lambda), \\ \mathrm{Bin_N(X)} &:= \Pi_X(\lambda(\boldsymbol{\xi}_\lambda), N_X), \\ Y &:= \mathrm{f}(\boldsymbol{\xi}_\mathrm{f})[\mathrm{X}] + \epsilon(\sigma_\epsilon(\xi_\epsilon)), \end{aligned} \tag{4.20}$$

where GP stands for a Gaussian Process, and here represents the whole forward model of the $\lambda(\boldsymbol{\xi}_\lambda)$ field as described above. Π_X represents a Poisson process with rate $\lambda(\boldsymbol{\xi}_\lambda)$ as defined previously, and N_X is the number of bins taken for inferring the cause field (here denoted by X). This generative model is depicted as well in its equivalent graph form in fig. 4.1.

4.2.1.2 Version 2

Since the approach in the previous chapter assumes fixed grid size, N_X, for learning the probability density of the cause variable, for which a Poisson measurement process was assumed (refer to eq. (4.20)), it is destined to underperform for cases where the discretization of the measurement grid is different than the one assumed by the algorithm. For example, the observed data $(X^{(obs)}, Y^{(obs)})$ of a cause-effect could be

$$\begin{aligned} X^{(obs)} &\equiv \text{"Cement"[kg per m}^3\text{ of mixture]}, \\ Y^{(obs)} &\equiv \text{"Concrete compressive strength"[MPa]}. \end{aligned}$$

The $(X^{(obs)}, Y^{(obs)})$ pairs for this case are shown in fig. 4.3. The true causal direction in this case is $X \to Y$, since it is clear that the change of amount of cement in the mixture will affect the concrete compressive strength, while the reverse doesn't have to be true. As it can be seen the cause variable is discretized in this case, most likely due to the round-off during the measurements. Therefore fixing the number of bins $N_X = 512$ as it was done for the model described in section 4.2.1.1, will not be consistent with how data was measured.

In order to mend this problem, measurement noise is introduced here for the cause variable as well, opposite to what was done in the previous chapter where measurement noise was assumed only for the effect variable. This noise was taken to be Gaussian distributed. Furthermore, the probability distribution for the cause variable is inferred directly as a normalized log-normal field with unknown correlation structure.

In other words

$$\begin{aligned} p(X|M_b^{(2)}) &= \frac{\exp(s_X(\boldsymbol{\xi}_{p(X)}))}{\left\|\exp(s_X(\boldsymbol{\xi}_{p(X)})\right\|_1}, \\ \text{with} \quad s_X(\boldsymbol{\xi}_{p(X)}) &= \mathbb{F}^{-1}(A_X(\xi_{A_X})(\xi_X)). \end{aligned} \tag{4.21}$$

Here, as well as in the previous chapter we assume statistical homogeneity not to single out any particular value for the cause variable X a priori. The $\mathbb{F}^{-1}$ is the inverse Fourier transform, A_X represents the amplitude spectrum and the quantity ξ_{A_X} represents the excitation field controlling the exact realization of the amplitude spectrum. As before, it is parameterized through $\xi_{A_X} = (\xi_{\alpha_0}^{(X)}, \xi_m^{(X)}, \xi_\eta^{(X)}, \xi_W^{(X)}, \xi_a^{(X)})$. Once again, these excitations control the zero mode strength, the slope of the amplitude spectrum, the strength of smooth deviations from this slope, the exact shape of these deviations and expected strength of real space fluctuations respectively. The excitation field ξ_X controls the exact real space realization of the s_X field and we collect $\boldsymbol{\xi}_{p(X)} = (\xi_{A_X}, \xi_X)$. The l_1 norm of the field $\exp(s_X(\boldsymbol{\xi}_{p(X)}))$ is denoted with $\|.\|_1$. Finally, $M_b^{(2)}$ is there to keep track of the assumptions made. As in previous section, we restate them all together once the complete generative model is specified.

Now, in order to make the field $p(X|M_b^{(2)})$ a part of the forward model, we use the inverse cdf transform

$$\begin{aligned} \mathcal{P}_{p(X)}(U_X) &= \int^{U_X} p(x|M_b^{(2)})dx, \\ \text{with} \quad \mathcal{P}_{p(X)}^{-1} &\equiv (\mathcal{P}_{p(X)}(U_X))^{-1} \end{aligned} \tag{4.22}$$

representing the inverse cdf operator. The quantity U_X stands for the operator which provides the uniform sample used for generating corresponding sample of the cause variable X in the forward model. This operator is defined as

$$\begin{aligned} U_X(\xi_U) &= \int_{-\infty}^{\xi_U} \mathcal{G}(\xi, \mathbb{1})d\xi, \\ \text{with} \quad \xi_U &\hookleftarrow \mathcal{G}(\xi_U, \mathbb{1}), \end{aligned} \tag{4.23}$$

defining which value U_X will take, through the upper bound of the CDF integral defined on the first line in eq. (4.23).

With this, the generative model for the cause variable X has been defined. The complete set of the Gaussian excitation fields we collect into $\boldsymbol{\xi}_X = (\boldsymbol{\xi}_{p(X)}, \xi_U)$ which then defines fully the realization of $\mathcal{P}_{p(X)}^{-1}$.

Alongside the $\mathcal{P}_{p(X)}^{-1}$ operator, the measurement noise of the cause variable is also introduced and is assumed to be Gaussian distributed. More precisely

$$\begin{aligned} p(X^{(obs)}|\mathcal{P}_{p(X)}^{-1}, M_b^{(2)}) &= \mathcal{G}(X^{(obs)} - \mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X), \Sigma^{(X)}), \\ \text{with} \quad \Sigma^{(X)} &= (\sigma_\epsilon{}^{(X)})^2 \; \mathbb{1}_{\Sigma^{(X)}}. \end{aligned} \tag{4.24}$$

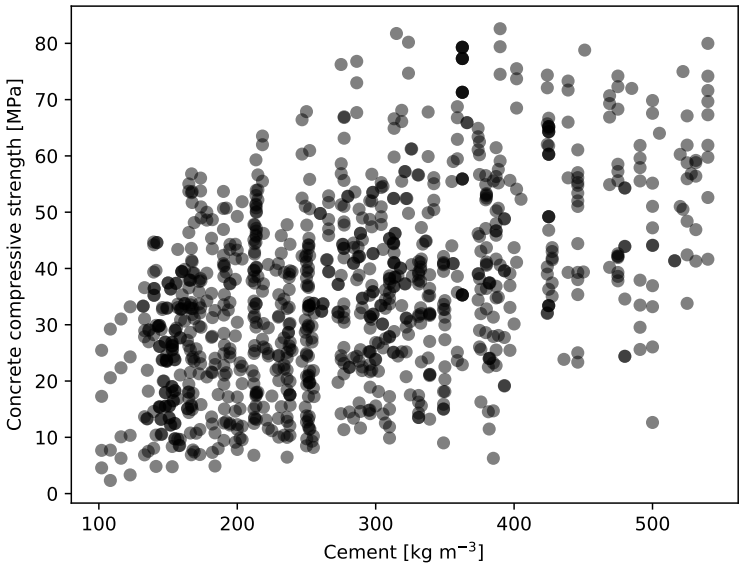


Figure 4.3: `pair0025` test case from the real world benchmark dataset `tcep` (see section 4.5)

By $X^{(obs)}$ we denote the observed value of the cause variable X in the given dataset. Similarly as in the last chapter, the noise covariance $\Sigma^{(X)}$ is assumed to be diagonal and with unknown noise variance on its diagonal denoted by $(\sigma_\epsilon{}^{(X)})^2$. This unknown variance is chosen to be inverse gamma distributed for the same reason as stated before (see discussion below eq. (4.15))

$$(\sigma_\epsilon{}^{(X)})^2(\xi_{\epsilon_X}) = \mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_{\epsilon_X},\beta_{\epsilon_X})}\left(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_{\epsilon_X})\right), \tag{4.25}$$

with

$$\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_{\epsilon_X}) = \int_{-\infty}^{\xi_{\epsilon_X}} \mathcal{G}(\xi,\mathbb{1})d\xi.$$

The parameters of the inverse gamma distribution are denoted by $(\alpha_{\epsilon_X}, \beta_{\epsilon_X})$ and the inverse cdf transform $\mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_{\epsilon_X},\beta_{\epsilon_X})}$ is defined in the same way as before, i.e. as the inverse of the operator shown in eq. (4.15).

With this, the full forward model for the causal variable is presented. In terms of a information Hamiltonian the likelihood for this model (eq. (4.24)) is written as

$$\begin{aligned}\mathcal{H}(X^{(obs)}|\mathcal{P}^{-1}_{p(X)}, \sigma_\epsilon{}^{(X)}, M_b^{(2)}) =& \frac{1}{2}\left(\frac{X^{(obs)} - \mathcal{P}^{-1}_{p(X)}(\boldsymbol{\xi}_X)}{\sigma_\epsilon{}^{(X)}(\xi_{\epsilon_X})}\right)^T\left(\frac{X^{(obs)} - \mathcal{P}^{-1}_{p(X)}(\boldsymbol{\xi}_X)}{\sigma_\epsilon{}^{(X)}(\xi_{\epsilon_X})}\right) \\ &+\frac{1}{2}\ln|\Sigma^{(X)}| \\ &+\mathcal{H}_0^{(X)},\end{aligned} \tag{4.26}$$

with

$$\mathcal{H}_0^{(X)} = \frac{N_{\sigma_\epsilon}^{(X)}}{2}\ln 2\pi. \tag{4.27}$$

Now, only the forward model for the observed effect variable $Y^{(obs)}$ remains to be specified. In fact, the proposed generative model here follows very closely the model proposed in the

previous section (see eq. (4.16) and discussion in the paragraph below it). The only difference is that now the cause variable, which lives in the domain of mapping f, is as well inferred. Summarizing quickly, the mapping is given by

$$f(\xi_f, A_f) = \mathbb{F}^{-1}(A_f(\xi_{A_f})(\xi_f)), \tag{4.28}$$

where A_f is the amplitude spectrum with the Gaussian excitation field ξ_{A_f} controlling its precise shape. Once again, $\xi_{A_f} = (\xi_{\alpha_0}^{(f)}, \xi_m^{(f)}, \xi_\eta^{(f)}, \xi_a^{(f)}, \xi_W^{(f)})$ and collecting $\boldsymbol{\xi}_f = (\xi_f, \xi_{A_f})$.

The measurement noise on the effect variable was taken to be Gaussian distributed as well and hence

$$\begin{aligned} p(Y^{(obs)}|f, \mathcal{P}_{p(X)}^{-1}, M_b^{(2)}) &= \mathcal{G}(Y^{(obs)} - f(\boldsymbol{\xi}_f)[\mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X)], \Sigma^{(Y)}), \\ \text{with } \Sigma^{(Y)} &= (\sigma_\epsilon{}^{(Y)})^2 \, \mathbb{1}_{\Sigma^{(Y)}}, \end{aligned} \tag{4.29}$$

and

$$(\sigma_\epsilon{}^{(Y)})^2(\xi_{\epsilon_Y}) = \mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_{\epsilon_Y}, \beta_{\epsilon_Y})}\left(\mathrm{CDF}_{\mathcal{G}(\xi, \mathbb{1})}(\xi_{\epsilon_Y})\right), \tag{4.30}$$

with the same definitions and notation for the corresponding operators and parameters as before in the case of $\sigma_\epsilon{}^{(X)}$ (look at eq. (4.25) and the discussion below). With this the likelihood information Hamiltonian for the effect variable is

$$\begin{aligned} \mathcal{H}(Y^{(obs)}|f, \mathcal{P}_{p(X)}^{-1}, \sigma_\epsilon{}^{(Y)}, M_b^{(2)}) &= \frac{1}{2}\left(\frac{Y^{(obs)} - f(\boldsymbol{\xi}_f)[\mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X)]}{\sigma_\epsilon{}^{(Y)}(\xi_{\epsilon_Y})}\right)^T \left(\frac{Y^{(obs)} - f(\boldsymbol{\xi}_f)[\mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X)]}{\sigma_\epsilon{}^{(Y)}(\xi_{\epsilon_Y})}\right) \\ &+ \frac{1}{2}\ln|\Sigma^{(Y)}| \\ &+ \mathcal{H}_0^{(Y)}, \end{aligned} \tag{4.31}$$

with

$$\mathcal{H}_0^{(Y)} = \frac{N_\Sigma^{(Y)}}{2}\ln 2\pi. \tag{4.32}$$

Collecting together the two information Hamiltonians from eq. (4.26) and eq. (4.31) the total likelihood information Hamiltonian for this model is

$$\begin{aligned} \mathcal{H}((X^{(obs)}, Y^{(obs)})|\mathcal{P}_{p(X)}^{-1}, f, \sigma_\epsilon{}^{(X)}, \sigma_\epsilon{}^{(Y)}, M_b^{(2)}) &= \mathcal{H}(X^{(obs)}|\mathcal{P}_{p(X)}^{-1}, \sigma_\epsilon{}^{(X)}, M_b^{(2)}) \\ &+ \mathcal{H}(Y^{(obs)}|f, \mathcal{P}_{p(X)}^{-1}, \sigma_\epsilon{}^{(Y)}, M_b^{(2)}) + \mathcal{H}_0, \end{aligned} \tag{4.33}$$

where

$$\begin{aligned} \mathcal{H}(X^{(obs)}|\mathcal{P}_{p(X)}^{-1}, \sigma_\epsilon{}^{(X)}, M_b^{(2)}) &= \frac{1}{2}\left(\frac{X^{(obs)} - \mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X)}{\sigma_\epsilon{}^{(X)}(\xi_{\epsilon_X})}\right)^T \left(\frac{X^{(obs)} - \mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X)}{\sigma_\epsilon{}^{(X)}(\xi_{\epsilon_X})}\right) \\ &+ \frac{1}{2}\ln|\Sigma^{(X)}|, \\ \mathcal{H}(Y^{(obs)}|f, \mathcal{P}_{p(X)}^{-1}, \sigma_\epsilon{}^{(Y)}, M_b^{(2)}) &= \frac{1}{2}\left(\frac{Y^{(obs)} - f(\boldsymbol{\xi}_f)[\mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X)]}{\sigma_\epsilon{}^{(Y)}(\xi_{\epsilon_Y})}\right)^T \left(\frac{Y^{(obs)} - f(\boldsymbol{\xi}_f)[\mathcal{P}_{p(X)}^{-1}(\boldsymbol{\xi}_X)]}{\sigma_\epsilon{}^{(Y)}(\xi_{\epsilon_Y})}\right) \\ &+ \frac{1}{2}\ln|\Sigma^{(Y)}| \end{aligned} \tag{4.34}$$

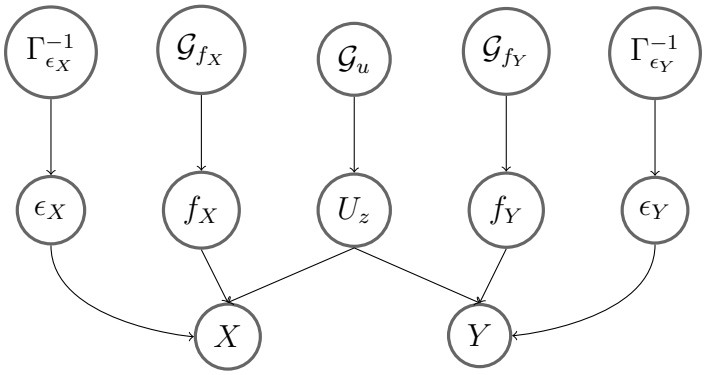


Figure 4.4: Graph structure for the confounder model $X \leftarrow Z \rightarrow Y$

with

$$\mathcal{H}_0 = \left(\frac{N_\Sigma^{(X)}}{2} + \frac{N_\Sigma^{(Y)}}{2} \right) \ln 2\pi. \tag{4.35}$$

The assumptions put into this model $M_b^{(2)}$ are completely the same as the ones put into $M_b^{(1)}$ from the previous section, with the exception of the fixed grid size for the measurement of the cause variable which is dropped in this case.

To make a connection with SEMs the model is rewritten as follows:

$$\begin{aligned}
\ln p(X) &:= \mathrm{GP}(\boldsymbol{\xi}_{\mathrm{p(X)}}), \\
U_X(\xi_U) &:= \mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_U), \\
\sigma_{\epsilon_X}(\xi_{\epsilon_X}) &:= \mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_{\epsilon_X},\beta_{\epsilon_X})}(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_{\epsilon_X})), \\
X^{(obs)} &:= \mathcal{P}^{-1}_{p(X)}(\boldsymbol{\xi}_X) + \epsilon_X(\sigma_{\epsilon_X}(\xi_{\epsilon_X})), \\
\\
\sigma_{\epsilon_Y}(\xi_{\epsilon_Y}) &:= \mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_{\epsilon_Y},\beta_{\epsilon_Y})}(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_{\epsilon_Y})), \\
Y^{(obs)} &:= f(\boldsymbol{\xi}_{f_Y})[\mathcal{P}^{-1}_{p(X)}(\boldsymbol{\xi}_X)] + \epsilon_Y(\sigma_{\epsilon_Y}(\xi_{\epsilon_Y})),
\end{aligned} \tag{4.36}$$

with GP denoting a Gaussian Process, which represents the full forward model for $p(X)$. The rest of the notation is consistent with what was discussed so far. The explicit dependencies on the excitation fields are there to emphasize that it is their values that control the exact realizations for this SEM. Corresponding graph is given in fig. 4.2.

4.2.2 Model: $X \leftarrow Z \rightarrow Y$

In this model a hidden (confounder) variable Z, that is not observed, is indirectly measured by causing two variables X and Y through an additive noise model. More precisely

$$\begin{aligned}
X &= f_X(Z) + \epsilon_X, \\
Y &= f_Y(Z) + \epsilon_Y,
\end{aligned}$$

where f_X and f_Y represent the mappings between the hidden variable Z and the observed data X and Y respectively, with ϵ_X and ϵ_Y being the corresponding noise realizations.

Now, under the assumption that the distribution of the hidden variable Z is continuous, one can always transform to a uniformly distributed variable U_z through the use of the probability integral transform

$$U_z = \int_{-\infty}^{z} p(z|M_c)dz, \tag{4.37}$$

with M_c reminding the reader of the assumption that $p(z)$ is taken to be continuous. As in previous sections, M_c will be kept where needed and collecting all the assumptions along the way. Once the complete confounder forward model is presented the assumptions will be restated once more.

The important point to be made here is that using the the probability integral transform to go from Z to U_z, there is no loss of information since this mapping is bijective, i.e. it has a well defined inverse. This is true as long as the cumulative distribution function of $p(z|M_c)$ is monotonically increasing. In case there are some regions where the cumulative distribution function is constant, this is no longer true. But, one can argue that since these regions correspond to $p(z|M_c) \equiv 0$, the actual probability to get a sample of the confounder variable in that range is 0, hence they will not play an important role in the forward model anyhow.

Therefore, transforming from $Z \to U_z$, all the subtleties the original distribution $p(z|M_c)$ had will be absorbed by the mappings f_X and f_Y. In other words the mappings can be redefined as

$$\begin{aligned} f_i^{(U_z)} &\overset{\text{def}}{=} f_i \circ \mathcal{F}_Z^{-1}, \\ \text{for} \quad \mathcal{F}_Z^{-1} &\overset{\text{def}}{=} \text{CDF}^{-1} : (0,1) \to \Omega_Z \\ \text{and} \quad i &\in \{X, Y\}. \end{aligned} \tag{4.38}$$

where CDF^{-1} stands for the inverse cdf transform, i.e. the inverse of operation done in eq. (4.37), while Ω_Z represents the domain of the confounder variable Z. Accepting this premise, we can then continue towards defining the rest of the confounder model.

The prior structure for the random variable U_z is taken to be a Gaussian process, such that

$$U_z(\xi_U) = \int^{\xi_U} \mathcal{G}(\xi, \mathbb{1})d\xi \tag{4.39}$$

with ξ_U specifying the U_z value sampled in the forward model.

The generative models for the mappings $f_X^{(U_z)}$ and $f_Y^{(U_z)}$ will be completely analogous to the previous sections. Namely, they will be modelled as fields having some unknown correlation structure. For convenience and increased notation clarity the $(.)^{(U_z)}$ superscript will be dropped in the following text, but it should be understood that the domain of the mappings is the domain of U_z. As in every instance so far, the generative model for the mappings can be expressed as

$$\begin{aligned} f_i(\xi_{A_{f_i}}, \xi_i) &= \mathbb{F}^{-1}(A_i(\xi_{A_i})(\xi_i)), \\ \text{for} \quad i &\in \{X, Y\}. \end{aligned} \tag{4.40}$$

The amplitude spectrum A_i is modelled completely analogously as it was done in eq. (4.16) and eq. (4.28), i.e. under the assumption of statistical homogeneity in order to be completely agnostic about values X and Y can take a priori. The excitation field of the amplitude spectrum is once more $\xi_{A_i} = (\xi_{\alpha_0}^{(f_i)}, \xi_m^{(f_i)}, \xi_\eta^{(f_i)}, \xi_a^{(f_i)}, \xi_W^{(f_i)})$, for $i \in \{X, Y\}$. The ξ_X and ξ_Y excitation

fields specify the exact real space realizations for the mappings f_X and f_Y respectively. Collecting these excitations together we write $\boldsymbol{\xi}_{f_X} = (\xi_{A_{f_X}}, \xi_X)$ and $\boldsymbol{\xi}_{f_Y} = (\xi_{A_{f_Y}}, \xi_Y)$. This then concludes the forward model for these mappings. Next, we turn to the noise variables.

Similarly to sections 4.2.1.1 and 4.2.1.2, the noise is assumed to be Gaussian distributed. This allows to rewrite the corresponding likelihoods as

$$\begin{aligned} p(X|f_X, U_z, M_c) &= \mathcal{G}(X - f_X(\boldsymbol{\xi}_{f_X})[U_z(\xi_U)], \Sigma^{(X)}), \\ \text{with} \quad \Sigma^{(X)} &= (\sigma_\epsilon{}^{(X)})^2 \, \mathbb{1}_{\Sigma^{(X)}} \\ p(Y|f_Y, U_z, M_c) &= \mathcal{G}(Y - f_Y(\boldsymbol{\xi}_{f_Y})[U_z(\xi_U)], \Sigma^{(Y)}), \\ \text{with} \quad \Sigma^{(Y)} &= (\sigma_\epsilon{}^{(Y)})^2 \, \mathbb{1}_{\Sigma^{(Y)}}, \end{aligned} \tag{4.41}$$

and

$$\begin{aligned} (\sigma_\epsilon{}^{(i)})^2(\xi_{\epsilon_i}) &= \mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_{\epsilon_i}, \beta_{\epsilon_i})} \left(\mathrm{CDF}_{\mathcal{G}(\xi, \mathbb{1})}(\xi_{\epsilon_i}) \right), \\ \text{for} \quad i &\in \{X, Y\}, \end{aligned} \tag{4.42}$$

with consistent notation and definitions as in sections 4.2.1.1 and 4.2.1.2 (see eq. (4.15) and discussion afterwards). Given this, the confounder forward model is completely specified.

Regarding the assumptions M_c, everything is completely similar as before (see the corresponding explanations for $M_b^{(1)}$ and $M_B^{(2)}$), with the addition of:

- The confounder variable Z is assumed to be continuously distributed with a non decreasing cumulative distribution function. This allows for performing a probability integral transform from Z to a uniformly distributed variable U_z, thus inferring only the mappings $f_X : (0,1) \to X$ and $f_Y : (0,1) \to Y$.

Below the total likelihood information Hamiltonian for this model is given

$$\begin{aligned} \mathcal{H}((X,Y)|U_z, \{\sigma_{\epsilon_i}, f_i\}_{i \in \{X,Y\}}, M_c) =& \mathcal{H}(X|U_z, \sigma_{\epsilon_X}, f_X, M_c) + \mathcal{H}(Y|U_z, \sigma_{\epsilon_Y}, f_Y, M_c) \\ =& \frac{1}{2} \left(\frac{X - f_X(\boldsymbol{\xi}_{f_X})[U_z(\xi_U)]}{\sigma_{\epsilon_X}(\xi_{\epsilon_X})} \right)^T \left(\frac{X - f_X(\boldsymbol{\xi}_{f_X})[U_z(\xi_U)]}{\sigma_{\epsilon_X}(\xi_{\epsilon_X})} \right) \\ &+ \frac{1}{2} \left(\frac{Y - f_Y(\boldsymbol{\xi}_{f_Y})[U_z(\xi_U)]}{\sigma_{\epsilon_Y}(\xi_{\epsilon_Y})} \right)^T \left(\frac{Y - f_Y(\boldsymbol{\xi}_{f_Y})[U_z(\xi_U)]}{\sigma_{\epsilon_Y}(\xi_{\epsilon_Y})} \right) \\ &+ \frac{1}{2} \left(\ln|\Sigma^{(X)}| + \ln|\Sigma^{(Y)}| \right) \\ &+ \mathcal{H}_0 \end{aligned} \tag{4.43}$$

with

$$\mathcal{H}_0 = \left(\frac{N_\Sigma^{(X)}}{2} + \frac{N_\Sigma^{(Y)}}{2} \right) \ln 2\pi \tag{4.44}$$

where $N_\Sigma^{(X)}$ and $N_\Sigma^{(Y)}$ denote the dimension of spaces on which $\Sigma^{(X)}$ and $\Sigma^{(Y)}$ operators are defined.

Finally everything can be restated in terms of a SEM as

$$\begin{aligned} U_z(\xi_U) &:= \int^{\xi_U} \mathcal{G}(\xi, \mathbb{1}) d\xi, \\ X &:= f_X(\boldsymbol{\xi}_{f_X})[U_z(\xi_U)] + \epsilon_X(\sigma_{\epsilon_X}(\xi_{\epsilon_X})), \\ Y &:= f_Y(\boldsymbol{\xi}_{f_Y})[U_z(\xi_U)] + \epsilon_Y(\sigma_{\epsilon_Y}(\xi_{\epsilon_Y})), \end{aligned} \tag{4.45}$$

Since up to this point, all the models considered in this thesis with their corresponding information Hamiltonians are specified, in the next section a brief summary of the variational inference algorithm is given which uses the information Hamiltonians defined in this section for inferring quantities of interest.

4.3 Variational inference

Now, that the models and their corresponding information Hamiltonians have been stated one can proceed with inferring all the fields of interest. This chapter serves as a short summary of the section 3, therefore here we outline the most important aspects of the MGVI algorithm.

4.3.1 Approximating distribution

In order to perform variational inference we use a Gaussian with a particular parameterization for its covariance

$$\begin{aligned} \mathcal{Q}(\theta|\mathbf{d}) &= \mathcal{G}(\theta - \bar{\theta}, \Theta|_{\theta=\bar{\theta}}) \\ \Theta^{-1}|_{\theta=\bar{\theta}} &= M_{\mathbf{d}|\bar{\theta}} + M_{\bar{\theta}} \\ M_{\mathbf{d}|\bar{\theta}} &= \left\langle \frac{\partial \mathcal{H}(\mathbf{d}|\theta)}{\partial \theta} \frac{\partial \mathcal{H}(\mathbf{d}|\theta)}{\partial \theta^\dagger} \right\rangle_{p(\mathbf{d}|\theta)} \Bigg|_{\theta=\bar{\theta}} \\ M_{\bar{\theta}} &= \left\langle \frac{\partial \mathcal{H}(\theta)}{\partial \theta} \frac{\partial \mathcal{H}(\theta)}{\partial \theta^\dagger} \right\rangle_{p(\theta)} \Bigg|_{\theta=\bar{\theta}}, \end{aligned} \tag{4.46}$$

where with $\mathcal{Q}(\theta|\mathbf{d})$ we denote the approximating distributions used for the parameters θ which are inferred and $\mathbf{d} \equiv (X, Y)$. We explicitly write that the covariance is taken at the position of the current mean $\bar{\theta}$ which emphasizes that there is a dependence of the covariance on the mean. This dependence is neglected during the inference, and the covariance is simply taken to be implicitly given through the expression in the second line of eq. (4.46). In case of a Gaussian posterior, this dependence is not present, but in general can not be neglected. The advantage of having an implicitly given covariance is that one only needs to infer the mean, which allows us to scale our algorithms to a very high dimensional setting. The metric $M_{\mathbf{d}|\bar{\theta}}$ is nothing else than the Fisher information metric of the likelihood. This motivates then a similar ansatz for the prior metric $M_{\bar{\theta}}$ as written on the last line of eq. (4.46).

Rewriting this in the same basis as the information Hamiltonians presented in section 4.2,

eq. (4.46) becomes

$$\begin{aligned}\mathcal{Q}(\xi|\mathbf{d}) &= \mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}}) \\ \Theta^{-1}|_{\xi=\bar{\xi}} &= \underbrace{J_{\bar{\xi}}^{\dagger} M_{\mathbf{d}|\bar{\xi}} J_{\bar{\xi}}}_{=M_{\mathbf{d}|\bar{\xi}}} + \mathbb{1},\end{aligned} \tag{4.47}$$

where $J_{\bar{\xi}} = |\frac{\partial f}{\partial \xi}|$, with $\theta = f(\xi)$. Now that all this has been specified we choose to measure the distance of our approximating distribution from the correct posterior with the use of the Kullback-Leibler divergence (KL). It can be shown that the KL divergence has a nice correspondence with the Gibbs free energy from statistical physics [20]. Therefore, minimizing the KL w.r.t. the fields of interest would give us approximate posterior mean fields. Furthermore, since the metric is available implicitly to us we can estimate the uncertainty of the inferred mean fields as well.

4.3.2 Minimizing the KL

In our notation the KL is given by

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{Q}(\xi|\mathbf{d})||p(\theta(\xi)|\mathbf{d})) = \langle \mathcal{H}(\mathbf{d}, \xi) \rangle_{\mathcal{Q}(\xi|\mathbf{d})} + \mathcal{H}_0. \tag{4.48}$$

The constant terms w.r.t. the mean are collected in $\mathcal{H}_0$ and are unimportant for the inference procedure. Then, the gradient of the KL w.r.t. the mean is given by

$$\frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \bar{\xi}} \mathrel{\widehat{=}} \left\langle \frac{\partial \mathcal{H}(\mathbf{d}, \xi)}{\partial \xi} \right\rangle_{\mathcal{Q}(\xi|\mathbf{d})}. \tag{4.49}$$

The gradient is calculated using the stochastic estimate of the expectation value, with samples drawn from the approximating distribution $\mathcal{Q}(\xi|\mathbf{d})$. In other words

$$\begin{aligned}\frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \bar{\xi}} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{H}}{\partial \xi}(\mathbf{d}, \xi)\Big|_{\xi=\xi_i} &= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{H}}{\partial \xi}(\mathbf{d}, \xi)\Big|_{\xi=\bar{\xi}+\Delta\xi_i}, \\ \text{with} \quad \xi_i \hookleftarrow \mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}}) \quad &\text{and} \quad \Delta\xi_i \hookleftarrow \mathcal{G}(\xi, \Theta|_{\xi=\bar{\xi}}).\end{aligned} \tag{4.50}$$

The sampling is done through utilizing the following generative process

$$\begin{aligned}\xi' &\hookleftarrow \mathcal{G}(\xi', \mathbb{1}), \\ n' &\hookleftarrow \mathcal{G}(n', M_{d|\bar{\xi}}), \\ \Theta|_{\xi=\bar{\xi}} &= \left(J_{\bar{\xi}}^{\dagger} M_{d|\bar{\xi}} J_{\bar{\xi}} + \mathbb{1}\right)^{-1} \\ \Delta\xi' &= J_{\bar{\xi}} \xi' + n'\end{aligned}$$

With $\Delta\xi' \hookleftarrow \mathcal{G}(\Delta\xi', \Theta|_{\bar{\xi}})$. In order to obtain samples distributed with covariance $\Theta|_{\bar{\xi}}$, the following equation has to be solved

$$\Delta\xi = \Theta|_{\xi=\bar{\xi}} \Delta\xi'.$$

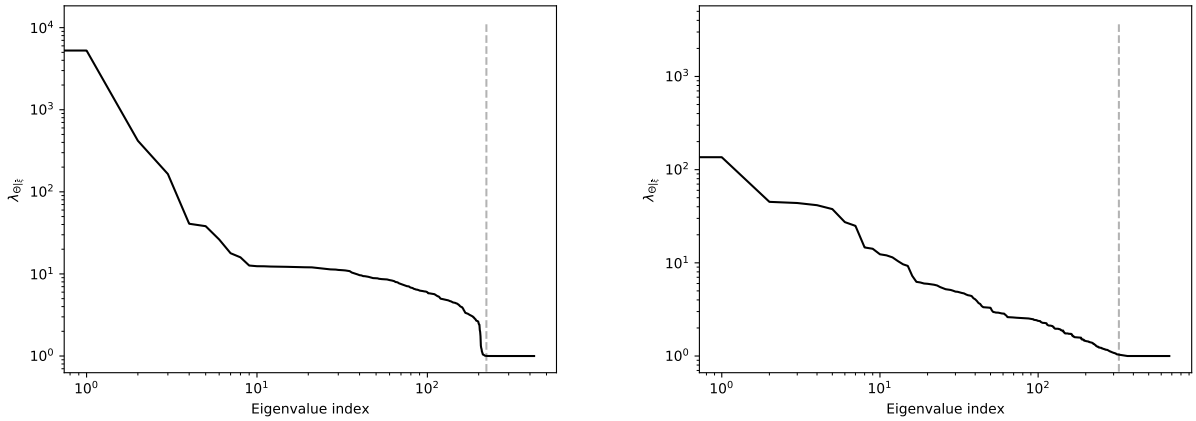


Figure 4.5: The characteristic eigenvalue spectrum of the metric $\Theta|_{\xi=\bar{\xi}}$. The dashed vertical line indicates a place where the eigenvalues $\lambda_{\Theta|_{\xi=\bar{\xi}}} \approx 1.0$ and hence all the ones to the right of this line don not contribute much to the evidence estimate. As it can be seen by comparing the two figures, the eigenvalue spectrum does not always strongly decrease. Thus, sometimes more eigenvalues need to be included in order to achieve sufficient precision in the calculation of Tr and Tr ln terms from eq. (4.58), which becomes computationally demanding.

This is done using conjugate gradient methods. With this we obtain $\Delta\xi \hookleftarrow \mathcal{G}(\Delta\xi, \Theta|_{\xi=\bar{\xi}})$ which can be shifted to follow the correct mean $\bar{\xi}$

$$\xi \equiv \bar{\xi} + \Delta\xi \hookleftarrow \mathcal{G}(\xi - \bar{\xi}, \Theta|_{\xi=\bar{\xi}})$$

With these samples the gradient of the KL can be estimated and a minimization step can be performed. For this step we used Newton conjugate gradient methods. Afterwards the procedure is repeated until sufficient degree of convergence is achieved.

4.4 Computing model evidences

Within the Bayesian framework it is possible to unambiguously define the degree of preference towards one model over the other given the data at hand. This factor is called the Bayesian odds factor. A very important feature of this factor is that it gives rise to Occam's razor naturally (see section 2.2). This is very convenient since it is expected that the models with more degrees of freedom (like for example the $X \leftarrow Z \rightarrow Y$ model) will in general explain the data better, than for example the bivariate models $X \rightarrow Y$, $Y \rightarrow X$. Inevitably in order to do so the $X \leftarrow Z \rightarrow Y$ model would have to use additional degrees of freedom which $X \rightarrow Y$ and $Y \rightarrow X$ don't have. We use the Bayesian odds factor as a decision mechanism for distinguishing which causal structure is the most probable given the dataset.

We extract a lower bound to the evidence factor directly from the KL divergence expression

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{Q}(\theta|d)||p(\theta|d)) = \int_{\Omega_\theta} \mathcal{Q}(\theta|d) \ln\left(\frac{\mathcal{Q}(\theta|d)}{p(\theta|d)}\frac{p(d)}{p(d)}\right) = \int_{\Omega_\theta} \mathcal{Q}(\theta|d) \ln\frac{\mathcal{Q}(\theta|d)}{p(\theta,d)} + \ln(p(d)) \tag{4.51}$$

where by $p(\theta|d)$ we denoted the true posterior approximated by $\mathcal{Q}(\theta|d)$ which is defined in section 4.3 (look at the discussion around eq. (4.46)). The factor which drops out the integration is the evidence factor we are looking for, $\ln p(d)$. From the above it follows

$$\begin{aligned}
\ln(p(d)) - \mathcal{D}_{\mathrm{KL}}(\mathcal{Q}(\theta|d)||p(\theta|d)) &= \langle \ln p(\theta, d)\rangle_{\mathcal{Q}(\theta|d)} - \langle \ln \mathcal{Q}(\theta|d)\rangle_{\mathcal{Q}(\theta|d)} \\
&= -\langle \mathcal{H}(d|\theta)\rangle_{\mathcal{Q}(\theta|d)} - \langle \mathcal{H}(\theta)\rangle_{\mathcal{Q}(\theta|d)} \\
&+ \left\langle \frac{1}{2}\left(\theta - \bar{\theta}\right)^{\dagger} \Theta^{-1}|_{\theta=\bar{\theta}} \left(\theta - \bar{\theta}\right) \right\rangle_{\mathcal{Q}(\theta|d)} + \left\langle \frac{1}{2} \ln|2\pi\Theta|_{\theta=\bar{\theta}}| \right\rangle_{\mathcal{Q}(\theta|d)} \\
&= -\langle \mathcal{H}(d|\theta)\rangle_{\mathcal{Q}(\theta|d)} - \langle \mathcal{H}(\theta)\rangle_{\mathcal{Q}(\theta|d)} \\
&+ \frac{N}{2}\ln(2\pi) + \frac{1}{2}\operatorname{Tr}\ln \Lambda_{\Theta|_{\theta=\bar{\theta}}} + \frac{1}{2}\operatorname{Tr}\mathbb{1}, \qquad (4.52)
\end{aligned}$$

where $\Lambda_{\Theta|_{\theta=\bar{\theta}}}$ is the diagionalized covariance $\Theta|_{\theta=\bar{\theta}}$ used in the MGVI approximation and N is the number of degrees of freedom present in the model. Note the constant terms in the equation above coming from the self-entropy term (third line of the eq. (4.52)). As it will be shown shortly, they are nicely absorbed by the prior term ($\langle \mathcal{H}(\theta)\rangle_{\mathcal{Q}(\theta|d)}$), therefore they do not pose a problem. The term $\operatorname{Tr}\ln \Lambda_{\Theta|_{\theta=\bar{\theta}}}$ is representing the determinant calculation of the posterior covariance (coming from the last term in the third line). Because of the way the splitting of the approximate posterior's metric is done (through the ansatz in the first of the equations listed in eq. (4.46)), the contribution to the eigenvalues of $\Lambda_{\Theta|_{\theta=\bar{\theta}}}$ comes through the eigenvalues of $M_{d|\bar{\theta}}$ and $M_{\bar{\theta}}$ from eq. (4.46). Of course, since $M_{d|\bar{\theta}}$ and $M_{\bar{\theta}}$ appear in the expression of the $\Theta^{-1}|_{\theta=\bar{\theta}}$ and we would like to find the eigenvalues in $\Lambda_{\Theta|_{\theta=\bar{\theta}}}$ which is the diagonalized $\Theta|_{\theta=\bar{\theta}}$, one needs to do one additional inversion after obtaining the eigenvalues.

Now, in order to actually calculate the eigenvalues it should be noted beforehand that it is possible to diagonalize $M_{d|\bar{\theta}}$ and $M_{\bar{\theta}}$ operators in the same basis. One of the ways to see this, is to remember that all the generative models implemented within `NIFTy` are assuming standardized coordinates for the prior as described in [28]. This allows us to write down the $\Theta^{-1}|_{\theta=\bar{\theta}}$ as

$$\Theta^{-1}|_{\bar{\xi}} = \underbrace{J_{\bar{\xi}}^{\dagger} M_{d|\bar{\xi}} J_{\bar{\xi}}}_{=M_{d|\bar{\theta}}} + \mathbb{1} \qquad (4.53)$$

as shown in the second line of eq. (4.47). From this one can directly get $\Lambda_{\Theta|_{\theta=\bar{\theta}}}$ and therefore evaluate the trace terms in eq. (4.52).

The ansatz in eq. (4.53) has one more nice property when it comes to calculating eigenvalues. Since the Fisher metric $M_{d|\bar{\theta}}$ of the problem at hand is a positive-definite operator, a lower bound of 1 can be imposed to all of the eigenvalues of $\Theta^{-1}|_{\xi=\bar{\xi}}$. This criterion is especially convenient when assessing whether all the relevant degrees of freedom were taken into account, when calculating the $\operatorname{Tr}\ln$ term from eq. (4.52).

From the numerical side, the actual calculations were done within `NIFTy` using the python wrap around the `ARPACK` package [9] available in `scipy.linalg` library. The characteristic spectrum of the eigenvalues obtained for $\Theta|_{\theta=\bar{\theta}}$ is given on fig. 4.5.

We show now how the marginalization of the prior yields exactly the same constant term which exactly cancels out the $(N/2)\ln 2\pi$ from (4.52). First, we note how the marginalized

Hamiltonian likelihood term looks like in standardized coordinates

$$\begin{aligned}\langle \mathcal{H}(d|\theta(\xi))\rangle_{\mathcal{Q}(\theta(\xi)|d)} &= \int D\theta(\xi)\mathcal{H}(d|\theta(\xi))\mathcal{Q}(\theta(\xi)|d) \\ &= \int \underbrace{D\xi\left|\frac{\partial\theta}{\partial\xi}\right|}_{\text{measure}}\mathcal{H}(d|\xi)\,\underbrace{\mathcal{Q}(\xi|d)\left|\frac{\partial\xi}{\partial\theta}\right|}_{\text{distribution}} \\ &= \int_{\Omega_\xi} D\xi\mathcal{H}(d|\xi)\mathcal{Q}(\xi|d) = \langle \mathcal{H}(d|\xi)\rangle_{\mathcal{Q}(\xi|d)},\end{aligned} \tag{4.54}$$

where for the $\mathcal{H}(d|\theta(\xi))$ it was used that it transforms as a function w.r.t. θ, and for the measure $D\theta$ and distribution $\mathcal{Q}(\theta(\xi)|d)$ in θ, the corresponding transformation laws were used. For passing from the second to the third line the Jacobians canceled and we were left with the expression on the last line. Now, using similar reasoning the marginalization of the prior Hamiltonian gives

$$\begin{aligned}\langle \mathcal{H}(\theta(\xi))\rangle_{\mathcal{Q}(\theta(\xi)|d)} = \langle \mathcal{H}(\xi)\rangle_{\mathcal{Q}(\xi|d)} &= \int_{\Omega_\xi}\left(\frac{1}{2}\xi^\dagger\xi\right)\mathcal{G}(\xi-\bar{\xi},\Theta|_{\xi=\bar{\xi}}) + \frac{N}{2}\ln(2\pi) \\ &= \frac{1}{2}\int_{\Omega_\xi}(\xi+\bar{\xi})^\dagger(\xi+\bar{\xi})\mathcal{G}(\xi,\Theta|_{\xi=\bar{\xi}}) + \frac{N}{2}\ln(2\pi) \\ &= \frac{1}{2}\left[\int_{\Omega_\xi}\xi^\dagger\xi\mathcal{G}(\xi,\Theta|_{\xi=\bar{\xi}}) + \bar{\xi}^\dagger\bar{\xi}\right] + \frac{N}{2}\ln(2\pi) \\ &= \frac{1}{2}\left[\mathrm{Tr}(\Theta|_{\xi=\bar{\xi}}) + \bar{\xi}^\dagger\bar{\xi}\right] + \frac{N}{2}\ln(2\pi) \\ &= \frac{1}{2}\left[\mathrm{Tr}(\Lambda_\Theta|_{\xi=\bar{\xi}}) + \bar{\xi}^\dagger\bar{\xi}\right] + \frac{N}{2}\ln(2\pi),\end{aligned} \tag{4.55}$$

where $\Lambda_{\Theta|_{\xi=\bar{\xi}}}$ is the diagonalized $\Theta|_{\xi=\bar{\xi}}$. Since Tr is invariant under unitary transformations, which diagonalization is, the Tr ln term can be as well transformed in this basis without loss of generality. Now it can be seen that the factor $(N/2)\ln(2\pi)$, from the above equation is exactly cancelling the one from before in eq. (4.52).

Combining everything up to now eq. (4.52) becomes

$$\begin{aligned}\ln(p(d)) - \mathcal{D}_{\mathrm{KL}}(\mathcal{Q}(\theta(\xi)|d)||p(\theta(\xi)|d)) = &-\langle \mathcal{H}(d|\xi)\rangle_{\mathcal{Q}(\xi|d)} \\ &- \frac{1}{2}\left[\mathrm{Tr}(\Lambda_\Theta|_{\xi=\bar{\xi}} - \mathbb{1}) + \bar{\xi}^\dagger\bar{\xi}\right] \\ &+ \frac{1}{2}\,\mathrm{Tr}\ln\Lambda_{\Theta|_{\xi=\bar{\xi}}}.\end{aligned} \tag{4.56}$$

It is worthwhile to comment on each of the terms appearing in the equation above. Starting from the lefthand side of the equation, one can already see that the bound for the evidence we will get depends on how close we come to the true posterior $p(\theta(\xi)|d)$ with our approximation $\mathcal{Q}(\theta(\xi)|d)$, which inherently makes sense since we're using a variational approach to inference. Therefore, it is of crucial importance to minimize the KL divergence as well as possible. In order to be consistent across different regression runs, the KL optimization was run for a sufficiently long time until a satisfying level of convergence was achieved.

The first term is nothing else than the marginalized likelihood in our posterior approximation, but with a minus sign in front. This term can be estimated through sampling

$$\langle \mathcal{H}(d|\xi)\rangle_{\mathcal{Q}(\xi|d)} \approx \frac{1}{n}\sum_{i=1}^{n} \mathcal{H}(d|\xi = \bar{\xi} + \xi_i), \tag{4.57}$$

where the samples ξ_i would be drawn according to $\mathcal{Q}(\xi|d)$ around the converged position $\bar{\xi}$. The number of samples used throughout this study is $n = 100$. From this, one can as well estimate the variance of the sample mean and give bounds to the evidence estimate. This plays an important role when comparing the models, because it could give us a measure of how strongly a given model is favoured w.r.t. the others considered.

The second term on the right-hand side of eq. (4.56) is completely negative since all the eigenvalues of $\Lambda_{\Theta|_{\xi=\bar{\xi}}} - \mathbb{1} \geq 0$ and $\bar{\xi}^\dagger\bar{\xi} \geq 0$. The minus sign in front of this term already hints towards its role. This term plays the role of the Occam's factor which is inherently present in all Bayesian forward models. For example, the $|\bar{\xi}|^2$ term penalizes each degree of freedom used in the model. The more complicated the model, the bigger the penalty it will get. The $\mathrm{Tr}(\Lambda_{\Theta|_{\xi=\bar{\xi}}} - \mathbb{1})$ plays a similar role. Namely, the terms which contribute to the trace are the ones coming from excited degrees of freedom, i.e. with eigenvalue $\lambda_{\Theta|_{\xi=\bar{\xi}}} > 1$. Since the metric is positive-definite, this term only reduces the evidence estimate. The last term, $\mathrm{Tr}\ln\Lambda_{\Theta|_{\xi=\bar{\xi}}}$, has the opposite sign to the previous one, hence working to increase the evidence bound. After a closer look, the balance between the Tr ln and Tr term will be held up until the second order expansion of the Tr ln term. This further adds up to the robustness of this approach and the ansatz assumed in the beginning for the MGVI algorithm.

Now, since one can not expect to minimize $\mathcal{D}_{\mathrm{KL}}$ exactly, we can only claim safely that this method would give us the lower bound to the $\ln(p(d))$, since $\mathcal{D}_{\mathrm{KL}} \geq 0$

$$\begin{aligned}\ln(p(d)) \geq &-\langle \mathcal{H}(d|\xi)\rangle_{\mathcal{Q}(\xi|d)} - \frac{1}{2}\left[\mathrm{Tr}(\Lambda_\Theta|_{\xi=\bar{\xi}} - \mathbb{1}) + \bar{\xi}^\dagger\bar{\xi}\right] \\ &+ \frac{1}{2}\,\mathrm{Tr}\ln\Lambda_{\Theta|_{\xi=\bar{\xi}}}\end{aligned} \tag{4.58}$$

The evidence estimate as depicted above and summarized in eq. (4.58) is used as a score for the models faced against data. The uncertainty bounds are obtained by evaluating the variance of the sample mean, with sample mean calculated in (4.57), corrected by an estimate for the upper bound to the error made for the Tr and Tr ln terms in eq. (4.58). The upper bound for the trace terms is calculated as follows

$$\begin{aligned}\Delta(\mathrm{Tr}\,\Lambda_{\Theta|_{\xi=\bar{\xi}}}) &\leq (N - \tilde{N})\lambda_{\Theta|_{\xi=\bar{\xi}}}(\tilde{N}), \\ \Delta(\mathrm{Tr}\ln\Lambda_{\Theta|_{\xi=\bar{\xi}}}) &\leq (N - \tilde{N})\ln\lambda_{\Theta|_{\xi=\bar{\xi}}}(\tilde{N}).\end{aligned} \tag{4.59}$$

Here $\tilde{N}$ denotes the index of the last calculated eigenvalue, which by construction of the `ARPACK` algorithm is the $\tilde{N}$th largest, guaranteeing $\lambda_{\Theta|_{\xi=\bar{\xi}}}(i) \leq \lambda_{\Theta|_{\xi=\bar{\xi}}}(\tilde{N})$ for $\tilde{N} < i < N$, with N being the total number of degrees of freedom present in the model. Therefore, in total, the uncertainty for the bound of the evidence is chosen to be

$$\Delta(\ln(p(d))) = \pm\sigma(\langle \mathcal{H}(d|\xi)\rangle_{\mathcal{Q}(\xi|d)}) - \Delta(\mathrm{Tr}\,\Lambda_{\Theta|_{\xi=\bar{\xi}}}) + \Delta(\mathrm{Tr}\ln\Lambda_{\Theta|_{\xi=\bar{\xi}}}). \tag{4.60}$$

This allows for some flexibility when comparing the model evidence estimates. Namely, one approach of would be to just look at the estimates of $\ln(p(d))$ as given in eq. (4.58) and claiming

preference for the model with a bigger estimate. But, we could as well look at the estimate intervals ($\ln(p(d))\pm\Delta(\ln(p(d)))$) and claim preference only if there is no overlap. This turns out to be especially important when evaluating the performance of all three models ($X \to Y$, $Y \to X$ and $X \leftarrow Z \to Y$) on the `ConSyn` dataset (see section 4.5). For that test case, a log-loss for each model was calculated as described in section 4.6.1 and this was used a quantity for model comparison.

4.5 Datasets

The datasets used in this paper are discussed below. They are referred to in the rest of the text through the names given herein. The results for corresponding datasets will be discussed later on in the results section (section 4.6).

`bcs_default`

This dataset was made by the forward model described in [29]. Here it served as a testing ground for the developed methods before applying them to the real world datasets. It should be noted that the generative models which generated this dataset and the ones presented here have a very similar structure, therefore it is expected that our methods perform well on this dataset.

`tcep`

The `tcep` is shorthand for *Tübingen Cause Effect Pairs* dataset. It was assembled in the paper [26] and consists of a variety of different Cause-Effect pairs, ranging from meteorological measurements to car industry problems (for example Weight - Consumption relation). It serves as a benchmark dataset for all new methods developed for causal discovery, hence we use the dataset here in order to evaluate how well our methods would cope with real-world problems.

Due to the high dimensionality of some of the test cases from this dataset they had to be neglected here. More precisely these are the test cases `pair0052`, `pair0053`, `pair0054`, `pair0055`, `pair0071` and `pair0105`. Further description of these datasets in particular is available in [26].

`SIM`, `SIM-G`, `SIM-ln`, `SIM-c`

The tests for these datasets were all taken from paper [26]. The especially interesting one for us here, out of the ones listed above, is the `SIM-c` dataset. It is a good testing ground for our confounder model. The graph which was used for generating the data is the same as the confounder generative model we consider $X \leftarrow Z \to Y$ with an additional edge $X \to Y$ or $Y \to X$.

ConSyn

This dataset serves as testing ground between the three implemented models $X \to Y$, $Y \to X$ and $X \leftarrow Z \to Y$. Two thirds of all test cases are evenly distributed amongst data generated by forward models for $X \to Y$ and $Y \to X$ as described in section 4.2.1.1. The last third of the dataset was generated with the use of a $X \leftarrow Z \to Y$ model, but a bit different than the one described in section 4.2.2. The difference is that now the probability density of the confounder variable Z is generated non-parametrically through the forward model described below

$$\begin{aligned} P_z(\boldsymbol{\xi}_z) &:= \exp\left(\mathbb{F}^{-1} A_z(\xi_{A_z})(\xi_{P_z})\right) \\ Z(\xi_z, P_z) &:= \mathrm{CDF}_{P_z}^{-1}(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_z)), \end{aligned} \tag{4.61}$$

with A_z representing the correlation structure of the field P_z with precise shape governed by ξ_{A_z} Gaussian excitation field. The ξ_z and ξ_{p_z} are Gaussian excitation fields controlling the exact real space realization for Z and $p(z)$ when acted on by the $\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}$ and $A_z(\xi_{A_z})$ operators respectively. Again, the abbreviation $\boldsymbol{\xi}_z = (\xi_{P_z}, \xi_{A_z})$ is made as in previous sections. Note that by taking the exponential in forward model for P_z we ensure its positiveness.

Next, the mappings from $Z \to X$ (f_X) and $Z \to Y$ (f_Y), are modelled as well non-parametrically with the following forward model:

$$\begin{aligned} f_i(\boldsymbol{\xi}_{f_i}) &:= \mathbb{F}^{-1} A_{f_i}(\xi_{A_{f_i}})(\xi_i), \\ &\text{for} \quad i \in \{X, Y\} \end{aligned} \tag{4.62}$$

with consistent notation as before. Finally the noise realizations are obtained through

$$\begin{aligned} (\sigma_{\epsilon_i}(\xi_{\sigma_i}))^2 &:= \mathrm{CDF}^{-1}_{\Gamma(\alpha_{\sigma_i},\beta_{\sigma_i})} \mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_{\sigma_i}) \\ \epsilon_i &:= \mathcal{G}(0, (\sigma_{\epsilon_i}(\xi_{\sigma_i}))^2 \cdot \mathbb{1}) \\ &\text{for} \quad i \in \{X, Y\} \end{aligned} \tag{4.63}$$

with ξ_{σ_i} for $i \in \{X, Y\}$ controlling the exact way in which σ_i is sampled from the inverse gamma distribution. All together, they give the following SEM

$$\begin{aligned} P_z(\boldsymbol{\xi}_z) &:= \exp\left(\mathbb{F}^{-1} A_z(\xi_{A_z})(\xi_{P_z})\right) \\ Z(\xi_z, P_z) &:= \mathrm{CDF}_{P_z}^{-1}(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_z)) \end{aligned}$$

$$\begin{aligned} \sigma_{\epsilon_i}(\xi_{\sigma_i}) &:= \mathrm{CDF}^{-1}_{\Gamma(\alpha_{\sigma_i},\beta_{\sigma_i})}(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_{\sigma_i})), \\ \epsilon_i(\sigma_i(\xi_{\sigma_i})) &:= \mathcal{G}(0, ((\sigma_i(\xi_{\sigma_i})^{-2}) \cdot \mathbb{1}), \\ f_i(\boldsymbol{\xi}_{f_i}) &:= \mathbb{F}^{-1} A_{f_i}(\xi_{A_{f_i}})(\xi_i), \\ &\text{for} \quad i \in \{X, Y\} \end{aligned}$$

$$\begin{aligned} X &:= \mathrm{f_X}(\boldsymbol{\xi}_{\mathrm{f_i}})[\mathrm{Z}(\xi_{\mathrm{z}}, \mathrm{P_z})] + \epsilon_{\mathrm{X}}(\sigma_{\mathrm{X}}(\xi_{\sigma_{\mathrm{X}}})), \\ Y &:= \mathrm{f_Y}(\boldsymbol{\xi}_{\mathrm{f_i}})[\mathrm{Z}(\xi_{\mathrm{z}}, \mathrm{P_z})] + \epsilon_{\mathrm{Y}}(\sigma_{\mathrm{Y}}(\xi_{\sigma_{\mathrm{Y}}})), \end{aligned} \tag{4.64}$$

where all the dependencies on the Gaussian excitation fields of this generative process were emphasized. On fig. 4.6, one of the generated datasets through this forward model is depicted.

On the top left figure the histogram of sampled values of the confounder variable Z can be seen, alongside the realization of its probability density field P_z. Figures on the top right and bottom left represent the mappings f_X and f_Y through which Z is mapped to the (X, Y) data space. Finally, on the bottom right the corresponding ground truth curve, which in its parametric form is simply $f \equiv (f_X(Z), f_Y(Z))$, is shown in red, while the observed data, obtained by adding Gaussian noise with variance $\sigma^2 = 10^{-4}$ to this curve, is shown with black circles.

This particular dataset is a good test case for several reasons. The first one being that this data case can be reasonably well explained through the $X \to Y$ or $Y \to X$ model. But, the $X \leftarrow Z \to Y$ model should have a better chance of explaining the higher density of observed values for (X, Y) caused by the loop on the bottom right of the lower right figure. The second reason is that the data above the looping point is rather sparse and the part of the ground truth curve (shown in red) is masked by the noise realization. Therefore, it is a good test for the $X \leftarrow Z \to Y$ model to try and reproduce this part of the red curve. Finally, the distribution of the confounder variable Z is clearly far from being close to uniform in this case which can be seen from the top left figure. Hence, the $X \leftarrow Z \to Y$ model has to adjust for this through mappings f_X and f_Y which are inferred. In section 4.6.1 a discussion of the performance of the algorithm described in section 4.2.2 is given for this particular test case.

The complete `ConSyn` dataset as well as the configuration file and the code for generating this data set are all available under the following link [42].

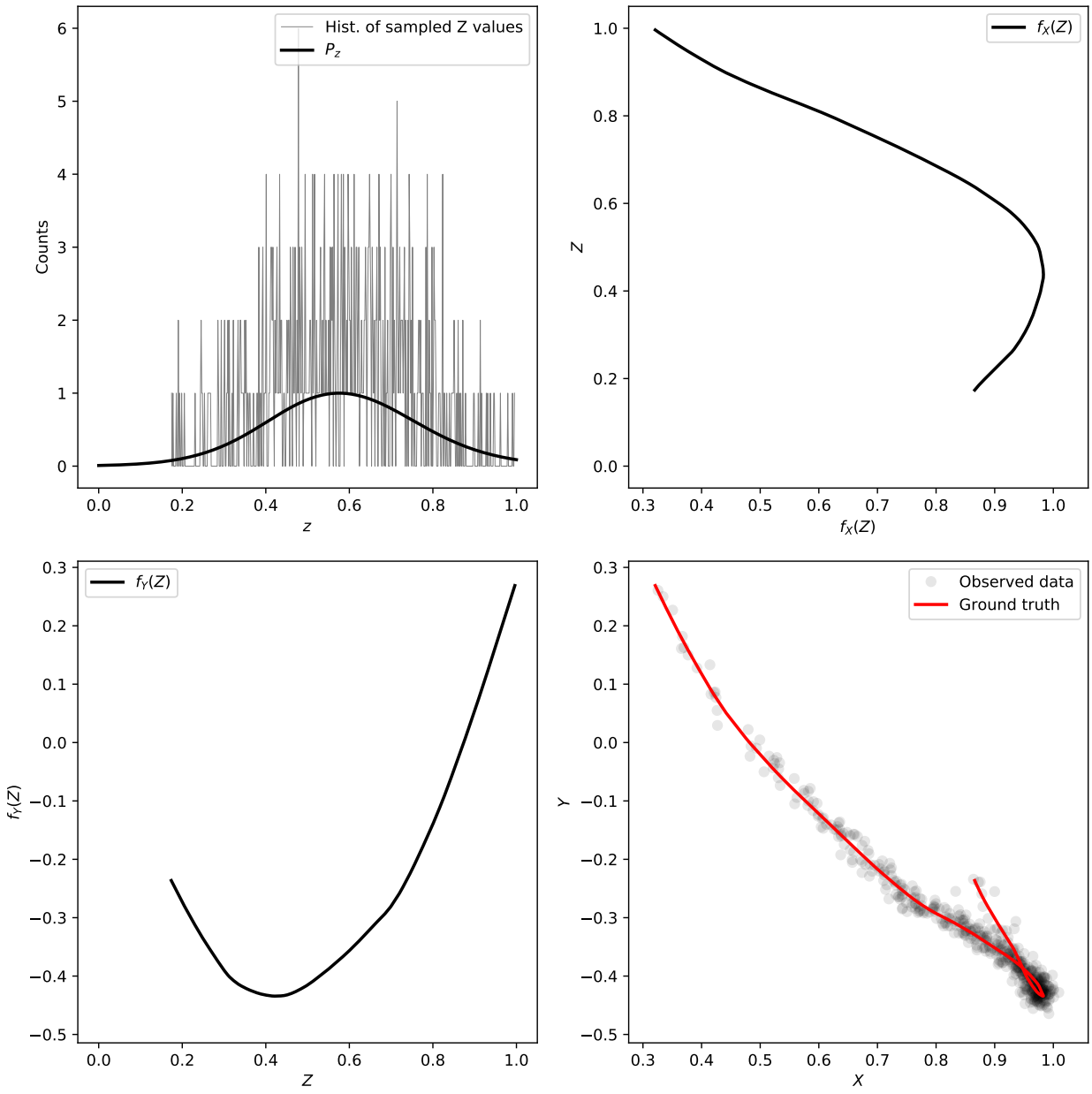


Figure 4.6: The generated data of the confounder model described in eq. (4.64), alongside with the ground truth field realizations for the `pair0087` test case.

Model	bcs_default	tcep	SIM-G	SIM-ln	SIM-c
Bivariate v1	99%	61%	57%	54%	42%
Bivariate v2	-	54%	54%	51%	-
BCI	98%	64%	-	-	-
ANM-HSIC	**100**%	63%	$\sim$ 75%	$\sim$ 80%	$\sim$ **81**%
ANM-MML	**100**%	58%	$\sim$ **82**%	$\sim$ **85**%	$\sim$ 62%
IGCI	65%	66%	$\sim$ 40%	$\sim$ 40%	50%
CGNN	72%	**70**%	-	-	-

Table 4.1: Results of testing the algorithms against the datasets described in section 4.5 upon enforcing to decide between $X \to Y$ and $Y \to X$ for all test cases. The '-' sign is denoting that no results are available at the moment and the sign $\sim$ is denoting approximate value since the exact was not reported in [26] and the values had to be read off the plots provided.

4.6 Results and Discussion

Results of all the performed tests are shown in table 4.1. Performance for all the considered models in this chapter is shown as well as the reported performance of algorithms developed in previous works [26, 29]. Note that the value of 50% corresponds to a chance level on all datasets but `ConSyn` where the chance level is 1/3.

From table 4.1 it can be seen that the best performing model on the `tcep` dataset, from the ones implemented in this work is the version 1 of the bivariate model denoted with 'Bivariate v1' in the table (see section 4.2.1.1). Even though the error for the observed values of the cause variable was neglected in this model, it still managed to outperform the 'Bivariate v2' model (see section 4.2.1.2) which did try to account for this as well.

One plausible reason which could cause this is the strong discretization of the cause variable present in 54% of the test cases in the `tcep` dataset. To obtain this percentage, the test cases were analyzed and labeled as 'discretized' if the number of unique values the true causal variable obtained in the dataset was smaller than half the size of the dataset. For example, imagine some dataset $(X_i, Y_i)_{i=1...N}$, where N is the total number of points. Suppose then the true causal direction is $X \to Y$. To check whether to label this dataset as 'discretized' by convention noted here, one looks at the number of unique values there are in the set $\{X_i\}_{i=1...N}$. Call this number N_u. Then if $N_u < (1/2)N$ the set will be labelled as discretized, otherwise not.

Note that if a strong discretization effect is present it would violate the assumption of smoothness for the λ field, in case of the Bivariate v1 model, and of the $\mathcal{P}_{p(X)}$, in case of the Bivariate v2 model, hence reducing the applicability of these two methods. After all the 'discretized' test cases were removed accuracy of both models increased, but not significantly. The Bivariate v1 model obtained an accuracy of 64% while Bivariate v2 increased by a similar amount, to 57%. On the other hand, in case of `bcs_default` dataset, where no discretization effects are present, both models perform very well.

One possible way of mending the problem of the discretized cause variable is to add an additional response operator for both models, which would inform the algorithm that the distribution of the cause variable is expected to be discrete. This is left to be done in the future.

Besides the discretization, another problem is the fact that we force our Bayesian algorithm to decide between these two models by comparing the evidence estimate from eq. (4.58). The

Model	Score
Bivariate v1 and Confounder model	38%
Bivariate v2 and Confounder model	**58**%

Table 4.2: Performance of different model combinations on the `ConSyn` data set. As before, 'Bivariate v1' stands for the model as described in section 4.2.1.1 and 'Bivariate v2' for the one described in section 4.2.1.2. The 'Confounder model' stands for the confounder model as described in section 4.2.2. Note that the level of accuracy for random choice is 30%.

completely correct thing to do here would be to give degrees of plausibility to each of the models using the information provided by the estimate from eq. (4.58). A similar approach is undertaken in section 4.6.1. This could be seen as an advantage and a weakness of the method developed here, since at the end of the day a decision between the causal directions has to be made.

4.6.1 Results of the `ConSyn` experiment

Before further discussing the evaluation of the performance for the $X \to Y$, $Y \to X$, and $X \leftarrow Z \to Y$ models on the `ConSyn` dataset, we give a short description of the reconstruction shown in fig. 4.7.

This figure shows the reconstruction of the confounder model as described in section 4.2.2 of the `pair0087` test case of the `ConSyn` dataset. This test case is also depicted independently on fig. 4.6. Note however that what is done before the inference is to shift the values of (X, Y) pairs to be in range $(0, 1)$ with the use of python's `sklearn.preprocessing.MinMaxScaler` library. Hence, the scales on the axes differ between fig. 4.6 and fig. 4.7. This rescaling makes the problem numerically more stable while preserving all the important features data has.

On the top left of fig. 4.7, the mean of the samples for the U_z variable (see eq. (4.37)) is shown alongside the ground truth of the probability distribution for the confounder variable Z which generated the data (see eq. (4.61)). Comparing the U_z samples with the actual samples drawn for the underlying confounder variable (top left in fig. 4.6) we see that they differ as expected. This discrepancy is to be absorbed inside the reconstructions of the mappings f_X and f_Y (for definitions see eq. (4.40)). These reconstructions are shown (dashed blue line) with the corresponding samples (light blue) on the top right and bottom left in fig. 4.7 alongside the ground truths of the actual realizations for these mappings (shown in black). As it can be seen the ground truth mappings are not fully recovered, but most of the features are captured with the reconstructions within the uncertainty of the posterior samples. Only at the turning point of the black curves does a bigger deviation occur. Finally, the bottom right figure shows the parametric $(f_X(Z), f_Y(Z))$ ground truth curve in red from which data samples were drawn with added noise on top. The posterior mean of the reconstruction is given in dashed blue, with samples in light blue coloring. Most of the data points are explained by the model, but the few which follow the small tail of the ground truth curve (just above the clump of data points at $X \approx 0.8$) are not captured within any of the posterior samples at all. This is just the reflection of the poor performance of the reconstruction in the region after the turning point of the ground truth mappings occurs as seen on the top right and bottom left figures. Investigating further how to improve in these scenarios is left for future work. Now we turn to explaining how the overall performance of the $X \to Y$, $Y \to X$ and $X \leftarrow Z \to Y$ models was calculated for the `ConSyn` dataset.

In the case of the `ConSyn` dataset we compare three different models against each other $X \to Y$, $Y \to X$ and $X \leftarrow Z \to Y$ model. Instead of just looking at the mean evidence estimate for these models (the right hand side of eq. (4.58)), denoted here by μ_{evid}, we also take into account the uncertainty of this mean, denoted here by σ_{evid}. This uncertainty is given by eq. (4.60).

This extra piece of information provided by σ_{evid} is taken into account through calculating the log-loss for each model. Calculation is done by assuming Gaussian statistics for the evidence estimate and taking $N = 10^5$ samples from a Gaussian distribution which is centered around μ_{evid} and has variance equal to σ_{evid} of the model at hand. More precisely

$$\begin{aligned} \omega_i &\hookleftarrow \mathcal{G}(\omega - \mu_{\text{evid}}, \sigma_{\text{evid}}), \\ \text{for} \quad i &= 1, \ldots, 10^5. \end{aligned} \tag{4.65}$$

The assumption of Gaussian statistics is suggested by the maximum entropy principle, since we only have information on the mean and the variance for the model evidence estimate.

Note however that if the errors for the calculated eigenvalues, shown in eq. (4.59), are too large (of the order $\approx 10\%$ of the mean value) then the bounds for the mean of evidence estimate will no longer be symmetrical enough, which will violate the Gaussian assumption. Therefore, it is of crucial importance for the evaluation of performance done on the `ConSyn` dataset to account for all relevant eigenvalues in the first place.

After drawing the N samples, $[\omega_i]_{i=1\ldots N}$, the frequency at which a given model has the biggest evidence value is recorded. In other words, the three lists of samples

$$\begin{aligned} &[\omega_i^{(1)}]_{i=1\ldots N} \\ \text{with} \quad &\omega_i \hookleftarrow \mathcal{G}(\omega - \mu_{\text{evid}}^{(1)}, \sigma_{\text{evid}}^{(1)}) \\ &\text{for the } X \to Y \text{ model,} \end{aligned}$$

$$\begin{aligned} &[\omega_i^{(2)}]_{i=1\ldots N} \\ \text{with} \quad &\omega_i \hookleftarrow \mathcal{G}(\omega - \mu_{\text{evid}}^{(2)}, \sigma_{\text{evid}}^{(2)}) \\ &\text{for the } Y \to X \text{ model,} \end{aligned}$$

$$\begin{aligned} &[\omega_i^{(3}]_{i=1\ldots N} \\ \text{with} \quad &\omega_i \hookleftarrow \mathcal{G}(\omega - \mu_{\text{evid}}^{(3)}, \sigma_{\text{evid}}^{(3)}) \\ &\text{for the } X \leftarrow Z \to Y \text{ model} \end{aligned} \tag{4.66}$$

are compared to each other element wise, without changing the order in which the samples were taken. After this comparison the number of times a given model had biggest sample value is stored. This value is then normalized by dividing it with N, the total number of samples taken, and a $-\log_2$ of this quantity is calculated. In this way, the log-loss is specified and can be used to compare the models to each other. The model with the smallest log-loss is deduced as the most plausible for the given test case.

Results of the evaluation for the Bivariate v1 and the Bivariate v2 models are shown in table 4.2. What can be seen from the table is that the Bivariate v1 model performs worse than Bivariate v2 on this test case, even though it had bigger score for the `tcep` dataset (refer to table 4.1).

One possible reason for this could be that simply some of the constant factors of the models (see eq. (4.19), eq. (4.35) and eq. (4.44)) are still missing in the implementation for the calculation of the evidence bounds for the considered models (see eq. (4.58)).

The other possibility could be that the KL was not properly minimized for the runs where the model failed to infer the correct direction. This can impair the evidence estimate as well, since the $\mathcal{D}_{\mathrm{KL}}$ term on the left hand side of eq. (4.56) will not be negligible and therefore reduce the evidence estimate further.

Both of these scenarios are left to be examined in the near future.

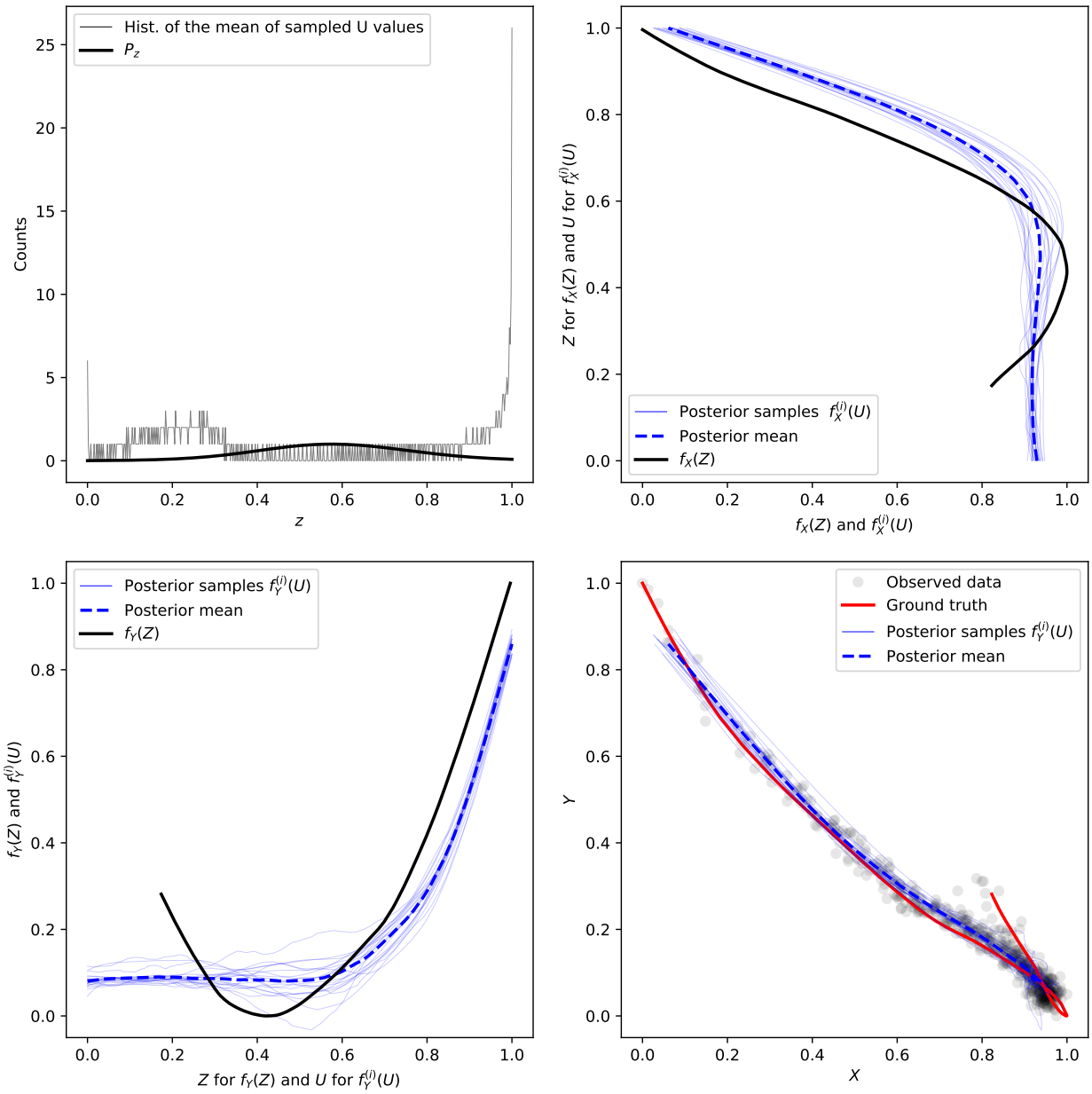


Figure 4.7: The inferred posterior mean fields of the mappings f_X and f_Y, as well as the histogram of the samples of the U_z hidden variable, for the test case `pair0087`.

5 Detecting Quasi Periodic Oscillations (QPOs)

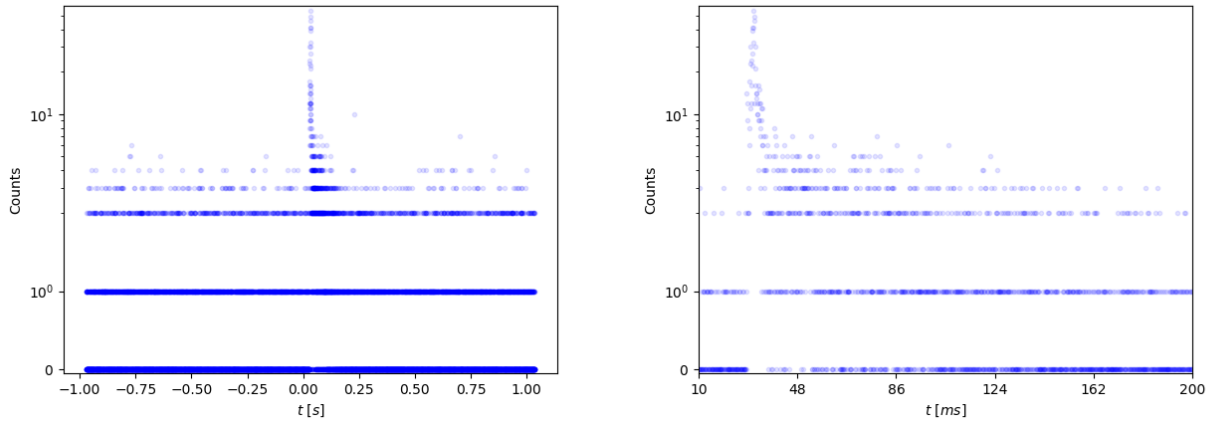
The work done here was initialized upon a private correspondence with a team from Institute of Astrophysics in Granada, Spain, which provided us with a dataset from the Atmosphere-Space Interactions Monitor boarded on the International Space Station [39] (refer to section 5.1 for more details). This detection showed potential of being high enough quality in order to be used to search for QPO signal, therefore a decision has been made to try the IFT methods on this problem and extract potential QPO signals.

The nature of the particular QPO signals we are looking for in this chapter traces back to the physics of soft gamma repeaters (SGRs). SGRs are defined as compact objects with very strong magnetic fields of the order of $10^{13} - 10^{15}$G, usually referred to as magnetars (look for example [17] for more details). The interest in these objects was raised in particular, after a discovery of certain QPO frequencies present in their lightcurves, which are believed to be a consequence of crust-core interaction within these compact objects coupled through the strong magnetic fields (*ibid.*). This in turn could offer a new observational windows for probing the internal structure of these exotic objects, as confirmed by simulations [21]. Therefore, every new characteristic frequency discovered would pave the way to new theoretical insights. Extending on the work of Pumpe et al. [32], where a search for quasi-periodic signals inside the light curves of the giant flares of SGR 1806-20 and SGR 1900+14 was done using a non-parametric Bayesian signal inference method called D^{3}PO, here I try to use the available methods of `NIFTy6` package [44] in order to discover new characteristic frequencies inside the dataset depicted on the fig. 5.1. In the next chapter (section 5.1) a description of the data is shown, followed by a section 5.2 where models are discussed and the corresponding results are given. At the end a brief discussion is given about the model performance and potential advancements are discussed.

5.1 Dataset

The dataset analyzed here was generated with the use of Atmosphere-Space Interactions Monitor, or ASIM for short. The instrument recorded data of total time of 2s, centered around the burst. To be more precise, the ASIM instrument consists of two independent instruments performing measurements. It has a low energy detector (LED), made of CdZnTe crystals, with a sensitivity range of 50 keV to 400 keV. The data which was used from the LED detector in this analysis is in range up to 350 keV. Alongside the low energy detector there is also a high energy one (HED), made of 12 HED scintillators, with sensitivity range from 300 keV to just over 30 MeV. Data from this detector was not included in the analysis. These LED and HED detectors have resolutions of 1μs and 28.7ns respectively allowing for a very high temporal resolution for the recorded event. The official name of the event is GRB 200415A, and it was simultaneosly detected by the Fermi satellite [41] and several others which allowed for precise localization of the event by the InterPlanetary Network (IPN) [40] to RA : 11.874°, DEC : $-25.194°$ (J2000). This places it just $\sim 1.5°$ above Earth's limb as seen by ASIM and only 1° from the edge of its field of view. A CCD image of the patch of the sky to which this event was localized is shown on fig. 5.2 .

The photon counts are shown on fig. 5.1. On the left figure (fig. 5.1a) each point represents the number of photons detected within a time interval of ~ 1.2ms, and on the right (fig. 5.1b) a zoom in plot is shown around the moment of the GRB 200415A event. It can be seen by



(a) $t \in [-1, 1]\,\mathrm{s}$ (b) $t \in [10, 200]\,\mathrm{ms}$

Figure 5.1: The dataset provided to us by the ASIM team, shown in different time ranges.

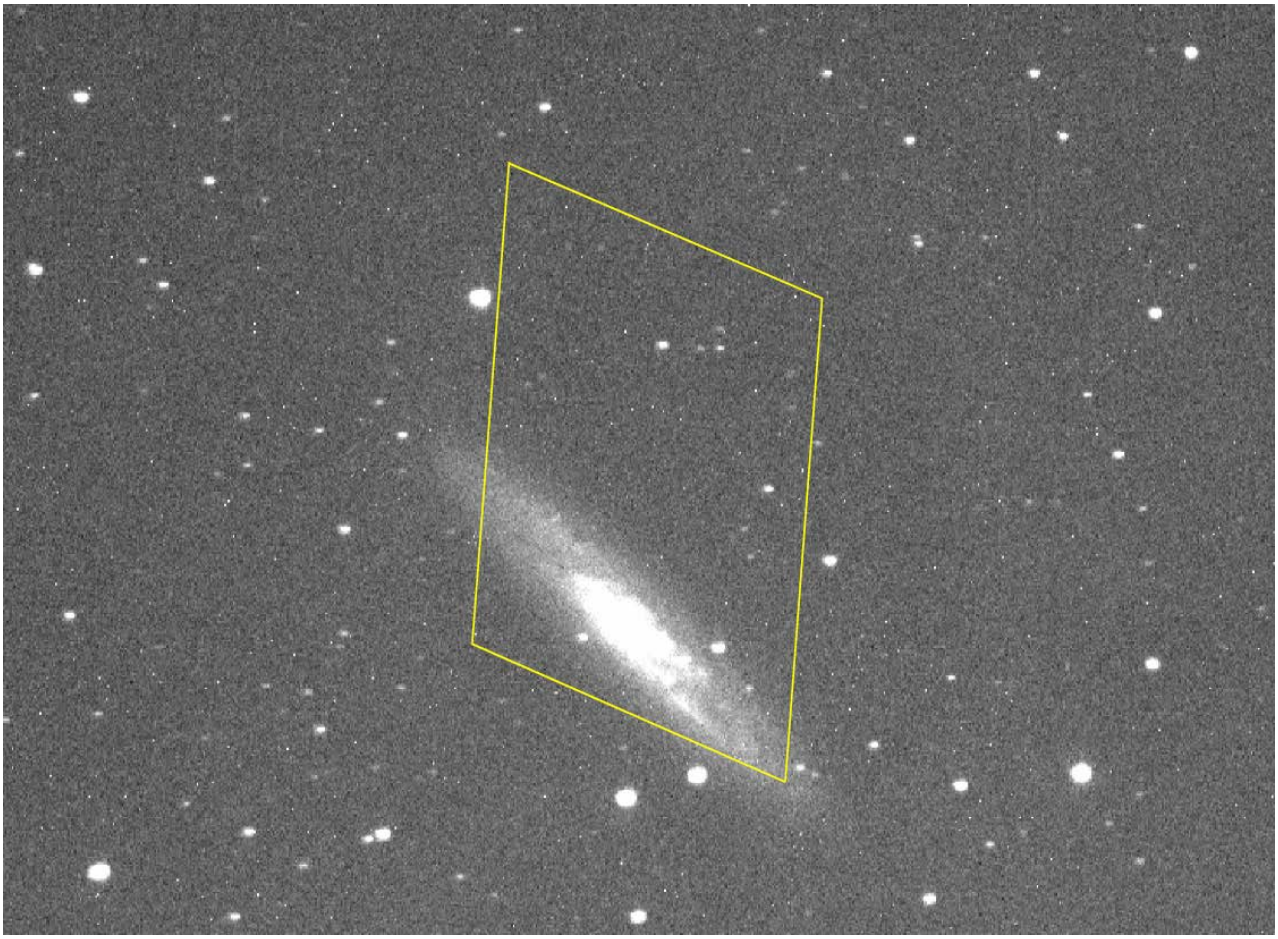


Figure 5.2: The IPN error box is plotted in yellow, localizing the GRB 200415A event to be within it. The galaxy in the center of the frame is the NGC253 galaxy. This picture was taken with BOOTES-1A wide field camera in Huelva (South Spain) on 23$^{\mathrm{rd}}$ July 2020 (04:00 UT).

eye that there are some recurrent peaks in photon counts at around 45, 55 , 65 ms and later on at around 75, 85, 95 ms, which would correspond to a frequency of order few hundred Hz. But, this could as well be a product of statistical Poisson noise hence one needs to be careful about making conclusions here.

5.2 Inferring the characteristic frequencies

In this chapter we elaborate in detail how the models for inferring QPO frequencies have been built. We distinguish between two different models. Since there was not much time during the course of the thesis to develop more sophisticated models, the ones presented here however offer few interesting insights into how the problem of QPO detection can be treated and extended. One of them even shows promising performance and insights into how one should proceed further.

5.2.1 Inverse Gamma model

The goal of the modelling essentially boils down to extracting characteristic frequencies of the sought for QPO signal from data with Poisson statistics. In order to achieve this the quasi periodic signal is modelled via

$$\lambda_{\mathrm{qpo}}(\boldsymbol{\xi}_{\lambda_{\mathrm{qpo}}}) = \lambda_0 \exp(\mathbb{F}^{-1}(A(\xi_{\mathrm{A}})\xi_{\lambda_{\mathrm{qpo}}})) + c_{\lambda_{\mathrm{qpo}}}(\xi_c), \tag{5.1}$$

where λ_{qpo} represents the signal field and it is nothing else than the rate of the Poisson process trying to explain the data, the λ_0 should serve to absorb the lattice size dependence of the signal and the $c_{\lambda_{\mathrm{qpo}}}$ serves to explain the photon counts coming from the background. Prior for this quantity has been chosen to follow the inverse gamma distribution, ensuring its positiveness and allowing for higher values if needed due to the heavy tails of this distribution

$$c_{\lambda_{\mathrm{qpo}}}(\xi_c) = \mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_c,\beta_c)}(\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_c)) \tag{5.2}$$

where

$$\mathrm{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_c) = \int_{-\infty}^{\xi_c} \mathcal{G}(\xi,\mathbb{1})\mathrm{d}\xi,$$

and the $\mathrm{CDF}^{-1}_{\Gamma^{-1}(\alpha_c,\beta_c)}$ represents the inverse cdf transform of the inverse gamma distribution with parameters (α_c, β_c). Values chosen for these parameters are depicted in table 5.1. The $\mathbb{F}^{-1}$ represents the inverse Fourier transform from the harmonic space to the data space. The symbol A encompasses the whole amplitude forward model of the signal field and the ξ_{A} and $\xi_{\lambda_{\mathrm{qpo}}}$ represent the Gaussian excitation fields, with $\boldsymbol{\xi}_{\lambda_{\mathrm{qpo}}} = (\xi_{\mathrm{A}}, \xi_{\lambda_{\mathrm{qpo}}})$. Since we expect excited frequencies, we can split the amplitude model into a smooth component, capturing large scale features in the data, and the component dedicated to modelling the peaks, which should be be able to explain the QPO signal if present inside the dataset. In other words

$$\begin{aligned} A(\xi_{\mathrm{A}}) &= A_{\mathrm{s}}(\xi_{\mathrm{s}})(1 + A_{\mathrm{p}}(\xi_{\mathrm{p}})) \\ \xi_{\mathrm{A}} &= (\xi_{\mathrm{s}}, \xi_{\mathrm{p}}) \end{aligned} \tag{5.3}$$

with A_{s} and A_{p} modelling the smooth component and peaks respectively with their corresponding Gaussian excitation fields ξ_{s} and ξ_{p}. The split, as done in eq. (5.3), allows for the

presence of peaks whenever the amplitude $A_{\mathrm{p}}(\xi_{\mathrm{p}}) > 1$, which gives much stronger preference to presence of lines than simply adding A_{s} and A_{p} in the forward model.

The smooth part is modelled under the assumption of statistical homogeneity, since by the smooth model the goal is to capture the large scale features in the data. Hence, there is no need to single out any particular time instance in the data at this point. Following the same reasoning as before (refer to eq. (4.3) and the discussion below it) we can construct a forward model for the smooth component A_{s}. Since here it is especially important to properly model the amplitude spectrum in order to capture the excited frequencies, the forward model for the amplitude spectrum will be once again explained but in less detail as this was already done in the previous section.

By applying the Wiener-Khinchin theorem [1, 2], the amplitude of the smooth part of the power spectrum can be written as

$$(A_{\mathrm{s}})_{\mathrm{kk}'} \propto (2\pi)\delta(k - k')p(|k|), \tag{5.4}$$

with enforced positivity of the amplitude spectrum

$$p(|k|) = \exp(\gamma(|k|)). \tag{5.5}$$

The field $\gamma(|k|)$ is modelled non-parametrically using an integrated Wiener process. This is preferred in order to restrain from choosing a particular functional basis for the realization of the $\gamma(|k|)$ field. Because the power spectrum $p(|k|)$ is positive, it's logarithm is modelled instead, i.e. the $\gamma(|k|)$ field. This is done in double-logarithmic scale. A consequence of choosing such coordinate system is that the zero mode has to be constrained through other means, because it is not possible to represent $k = 0$ mode on logarithmic scale, since it is located at $-\infty$. To achieve this, the zero mode is identified with an overall scaling factor, with a log-normal prior imposed to ensure its positiveness

$$(A_{\mathrm{s}})_{00} \equiv \alpha_0 = \exp(\mu_{\alpha_0} + \sigma_{\alpha_0}\xi_{\alpha_0}), \tag{5.6}$$

with α denoting the zero mode value, and $(\mu_{\alpha_0}, \sigma_{\alpha_0})$ being the mean and standard deviation of this Gaussian process with an excitation ξ_{α_0}.

Remembering now the precise form of the expression giving the realizations of $\gamma(|k|)$ field, we have

$$\begin{aligned}\gamma(|k|) &= c_0 + m|k| + \eta \int_{|k_0|}^{|k|} \int_{|k_0|}^{|k'|} \xi_W(|k''|)d|k'|d|k''|, \\ \tilde{U} &= \int_{k\neq 0} e^{2\gamma(|k|)}d|k| \ .\end{aligned} \tag{5.7}$$

In eq. (5.7) parameter m is describing the expected slope with an imposed Gaussian prior and η describes the strength of the smooth deviations from the linear part with a log-normal prior imposed. The precise shape of these deviations is governed by the Gaussian excitation field $\xi_W \hookleftarrow \mathcal{G}(\xi_W, \mathbb{1})$. All of these parameters are inferred from the data using the MGVI approximation. The term on the second line of eq. (5.7), $\tilde{U}$, is the total power contained in all modes except the $k = 0$ mode. This helps us then to fix the integration constant c_0 to the expected strength of real space fluctuations of the λ_{qpo} field

$$c_0 \equiv \frac{a}{\sqrt{\tilde{U}}} \tag{5.8}$$

with a defined through

$$p(|k|) = a\ \frac{\exp(\gamma(|k|))}{\sqrt{\tilde{U}}},\ \forall k \neq 0 \tag{5.9}$$

where with a the strength of the expected real space fluctuations of λ_{qpo} is denoted. The value of a is inferred from the data as well and it has an imposed log-normal prior. The complete prior structure for all of the parameters mentioned above is given by the following set of equations

$$\begin{aligned} \alpha_0 &= \exp(\mu_{\alpha_0} + \sigma_{\alpha_0}\xi_{\alpha_0}) \\ m &= \mu_m + \sigma_m \xi_m \\ \eta &= \exp(\mu_\eta + \sigma_\eta \xi_\eta) \\ a &= \mu_a + \sigma_a \xi_a \\ \text{with } \xi_j &\hookleftarrow \mathcal{G}(\xi_j, \mathbb{1}) \text{ for } j \in \{\alpha_0, m, \eta, a\} \end{aligned} \tag{5.10}$$

As earlier, this prior structure can be summarized through setting $\xi_{\text{s}} = (\xi_{\alpha_0}, \xi_m.\xi_\eta, \xi_a, \xi_W)$.

The values taken for the hyperparameters of this part of the model are shown in table 5.1. The chosen values reflect the fact that this part of the amplitude forward model should explain only the large scale features in the data, hence the low values for the (μ_η, σ_η) and (μ_a, σ_a). The priors for the slope (μ_m, σ_m) have been chosen such that certain degree of smoothness in enforced. The parameters controlling the zero mode strength $(\mu_{\alpha_0}, \sigma_{\alpha_0})$ are chosen to give broad enough range for the value of the zero mode, setting the mean to zero in order to allow the model for the background pick up the overall offset for the λ_{qpo} field. Note however that these are hyperparameters for the amplitude spectrum. Therefore, the chosen mean for the slope for example $\mu_m = -1.5$, has to be multiplied with a factor of 2 when talking about the slope of the powerspectrum. Finally, the hyperparameters for the background model $(\mu_{\alpha_c}, \sigma_{\alpha_c})$ were chosen such to have an expectation value of a few counts, since that is what can be seen from the data prior to the event, i.e. for times $t < 0$s.

For the second part of the model, the one that should be able to capture the power stored in excited frequencies, the A_{p} term from equation eq. (5.3), is modelled through

$$A_{\text{p}}(\xi_{\text{p}}) = \text{CDF}^{-1}_{\Gamma^{-1}(\alpha_p, \beta_p(\alpha_p))}(\text{CDF}_{\mathcal{G}(\xi, \mathbb{1})}(\xi_{\text{p}})), \tag{5.11}$$

where

$$\text{CDF}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_p) = \int_{-\infty}^{\xi_p} \mathcal{G}(\xi, \mathbb{1}) \text{d}\xi,$$

and the $\text{CDF}^{-1}_{\Gamma^{-1}(\alpha_p, \beta_p(\alpha_p))}$ represents the inverse cdf transform for the inverse gamma distribution, whose pdf is

$$p_{\Gamma^{-1}}(x; \alpha_p, \beta_p) = \frac{{\beta_p}^{\alpha_p}}{\Gamma(\alpha_p)} x^{-\alpha_p - 1} \exp(-\frac{\beta_p}{x}). \tag{5.12}$$

Note that in eq. (5.11) I have explicitly kept the α_p dependence of the β_p parameter inside the Γ^{-1} parameterization. Reason behind this is to compensate for the volume factors of the chosen time discretization. Namely, what we really want to do is to impose this prior not per k mode but per volume of the k space on which we're modelling our power spectrum amplitude. This can be achieved by utilizing the following property of the inverse gamma distribution

$$p_{\Gamma^{-1}}(x; \alpha_p, \beta_p) = {\beta_p}^{-1} p_{\Gamma^{-1}}(x/\beta_p; \alpha_p, 1).$$

Parameter	$(\mu_{\alpha_0}, \sigma_{\alpha_0})$	(μ_m, σ_m)	(μ_η, σ_η)	(μ_a, σ_a)	(α_c, β_c)	(α_p, β_p)
Value	$(0.0, 1.0)$	$(-1.5, 0.5)$	$(0.5, 0.05)$	$(0.1, 0.01)$	$(1.5, 0.75)$	$(1.0, \beta_p(\alpha_p))$

Table 5.1: The chosen hyperparameters for the IG model. Note that only the α_p value is given since β_p is determined from equation eq. (5.14) with proportionality factor of 1.0.

One can show that in order to have the probability of obtaining $x > a_p$, where a_p represents some high enough value for the power of the sought for characteristic frequencies (usually of the order $a_p \sim 10^1 - 10^2$ stronger than the smooth component), independent of the volume of the k space one looks at, the β_p needs to scale as

$$\beta_p(\alpha_p) \approx (\Delta V(k))^{-\frac{1}{\alpha_p}} \tag{5.13}$$

where $\Delta V(k) \equiv \Delta \log k$. This follows directly from looking at the cumulative distribution function of the inverse gamma distribution. In order to show this, consider the following.

Since we're interested in sampling high values for the power at the excited frequencies, we would be only worried about the tail of the inverse gamma distribution. There, the cumulative distribution function for the range $x > a_p$:

$$F_{\Gamma^{-1}(\alpha_p,\beta_p)}(a_p) = \int_{a_p}^{\infty} p_{\Gamma^{-1}}(x; \alpha_p, \beta_p)\mathrm{d}x \approx {\beta_p}^{\alpha_p} {a_p}^{-\alpha_p}, \text{ for } a_p \gg 1.$$

Now, if one looks at the volume $\Delta V(k)$, the probability of having peaks with power $x > a_p$ inside this volume would be given by:

$$\Delta V(k) F_{\Gamma^{-1}(\alpha_p,\beta_p)}(a_p),$$

since $F_{\Gamma^{-1}(\alpha_p,\beta_p)}$ is an additive quantity. This further implies that in order to have volume independent scaling, one needs to have:

$$\beta_p \approx \Delta V(k)^{-1/\alpha_p}, \tag{5.14}$$

as indicated in eq. (5.13).

This model here shall be referred to in the rest of the text as IG model (Inverse Gamma model) due to the fact that component modelling the power of excited frequencies has inverse gamma prior. Results of reconstructions of this model are shown on fig. 5.5 and fig. 5.3. Before continuing onto the next chapter, we quickly discuss the chosen hyperparameters of this part of the IG model, with values shown in table 5.1.

The chosen hyperparameters reflect the assumptions made throughout this model. The expected strength of the peaks is controlled by the inverse gamma prior parameters (α_p, β_p). We set $\alpha_p = 1$, to allow non-negligible probability mass to be present for the higher values of peak strengths. We do so, since the amplitude strength of the excited frequencies is expected to be at least $\approx 10^1 - 10^2$ times stronger than the amplitude spectrum of the smooth model A_s (refer back to eq. (5.3)).

5.2.2 High asperity model

Looking back at the assumptions made in the IG model it can be noticed that one of them could be too far from the nature of the signal we are after. Namely, the assumption that the

excited frequencies will happen at well defined frequencies with their power represented as δ-peaks superimposed on the smooth part of the spectrum could be far from the truth. In reality the excited frequencies will not be so well defined as to expect exactly δ-peaks in the power spectrum. Instead it is expected that the nearby k-modes will also be excited, although to a smaller degree than the central frequency. This results in having a finite spread of the peaks in the power spectrum instead of just a single well defined δ-peak.

In order to capture this feature as well, one needs to inform the algorithm that it is plausible to have a peak with power spread locally over several k-modes close to the excited frequency. In order to achieve this the amplitude spectrum model is reparametrized as described in the remainder of this section.

For consistency, consider again the model for λ_{qpo} as in eq. (5.1), with the smooth component of the amplitude spectrum as given in eq. (5.4)

$$A_{\text{kk}'} \propto (2\pi)\delta(k - k')p(|k|), \tag{5.15}$$

with

$$\begin{aligned} p(|k|) &= a\ \frac{\exp(\gamma(|k|))}{\sqrt{\tilde{U}}},\ \forall k \neq 0, \\ \gamma(|k|) &= m|k| + \eta \int_{|k_0|}^{|k|} \int_{|k_0|}^{|k'|} \xi_W(|k''|) d|k'| d|k''|, \\ \text{and}\ \ \tilde{U} &= \int_{k\neq 0} e^{2\gamma(|k|)} d|k|\ , \end{aligned} \tag{5.16}$$

and same description for the quantities that appear as before (please refer to the text below eq. (5.4)). As written down in eq. (5.16), it is not quite clear how one can make a generative model for $\gamma(|k|)$ if one needs to indeed perform a double integral of a Gaussian excitation ξ_W. Therefore, in the following a further elaboration on this is given and in the process a new parameter is introduced. This parameter will prove important for modelling the peaks occurring at the excited frequencies we're after.

Formally, the double integral in equation eq. (5.16) is a solution to the following stochastic differential equation

$$\begin{aligned} \partial_l^2 g(l) &= \xi_W, \\ \text{with}\ \ \xi_W &\hookleftarrow \mathcal{G}(\xi_W, \mathbb{1}), \end{aligned} \tag{5.17}$$

where an identification $l \equiv |k|$ has been made for notational convenience. In this form however the equation doesn't represent a Markov process and hence it is not clear how it can be used in the forward model. Therefore, what is proposed is, to make a system of two coupled stochastic differential equations, each being first order, and hence allowing for restoration of the Markov property. This can be done by modelling the derivative of $g(l)$, $h(l)$, as another process:

$$\begin{aligned} \partial_l g(l) - h(l) &= \sqrt{\mathcal{R}}\mathcal{W}\xi_g, \\ \partial_l h(l) &= \mathcal{W}\xi_h, \\ \text{with}\ \ \xi_i &\hookleftarrow \mathcal{G}(\xi_i, \mathbb{1})\ \ \text{for}\ i \in \{g, h\} \end{aligned} \tag{5.18}$$

Parameter	$(\mu_{\alpha_0}, \sigma_{\alpha_0})$	(μ_m, σ_m)	(μ_a, σ_a)	$(\mu_{\mathcal{W}}, \sigma_{\mathcal{W}})$	$(\mu_{\mathcal{R}}, \sigma_{\mathcal{R}})$	(α_c, β_c)
Value	$(0.7, 0.1)$	$(-2.5, 1.0)$	$(2.0, 0.5)$	$(1.0, 0.1)$	$(50.0, 5.0)$	$(1.5, 0.75)$

Table 5.2: The chosen hyperparameters for the High asperity model

where two new parameters $\mathcal{R}$ and $\mathcal{W}$ were introduced. The solution of this equation is

$$\begin{aligned} h(l) &= h(l_0) + \mathcal{W}\sqrt{(l - l_0)}\ \xi_h, \\ g(l) &= g(l_0) + \frac{(l - l_0)}{2}(h(l) + h(l_0)) + \mathcal{W}\sqrt{\frac{1}{12}(l - l_0)^3 + \mathcal{R}(l - l_0)}\ \xi_g, \end{aligned} \tag{5.19}$$

with l_0 representing a referent $|k_0|$ value which serves as a starting point for integration of the Wiener process. Since we're using the $\log - \log$ coordinates $k_0 \neq 0$. In this form, the eq. (5.19) corresponds to a generative model. This generative model is controlled by the values of $\mathcal{R}, \mathcal{W}$ and the precise realization depends on the Gaussian excitation fields ξ_g and ξ_h. As can be seen, the solution for the derivatives $h(l)$ is a Wiener process, and, in case $\mathcal{R} \to 0$, solution for the $g(l)$ will be an integrated Wiener process. This was precisely the regime in which the previous models were in. Otherwise, if $\mathcal{R} > 0$ it introduces an additional non-integrable Wiener process component into the forward model. In total, it can be concluded that $\mathcal{W}$ controls the total variance of the integrated Wiener process, while the parameter $\mathcal{R}$ determines the relative strength between the integrated and "non-integrated" part of the Wiener process. This then motivates the names *flexibility* and *asperity* for $\mathcal{W}$ and $\mathcal{R}$ parameters respectively, since the former controls the strength of deviations from linearity and latter measures the "roughness" of amplitude spectrum. Both of them are enforced to be positive with log-normal priors

$$\begin{aligned} \mathcal{W} &= \exp(\mu_{\mathcal{W}} + \sigma_{\mathcal{W}}\xi_{\mathcal{W}}), \\ \mathcal{R} &= \exp(\mu_{\mathcal{R}} + \sigma_{\mathcal{R}}\xi_{\mathcal{R}}). \end{aligned} \tag{5.20}$$

The hyperparameters μ_i and σ_i, for $i \in \{\mathcal{W}, \mathcal{R}\}$, denote the mean and standard deviation for the values of $\mathcal{W}$ and $\mathcal{R}$. In this way, both values for $\mathcal{W}$ and $\mathcal{R}$ can be informed by the data. With this, the η parameter from eq. (5.16) becomes obsolete.

Therefore, the amplitude spectrum forward model can now be restated as

$$\begin{aligned} p(|k|) &= a\ \frac{\exp(\gamma'(|k|))}{\sqrt{\tilde{U}}},\ \forall k \neq 0, \\ \gamma'(l) &= ml + g(l_0) + \frac{(l - l_0)}{2}(h(l) + h(l_0) + \mathcal{W}\sqrt{\frac{1}{12}(l - l_0)^3 + \mathcal{R}(l - l_0)}\ \xi_g, \\ h(l) &= h(l_0) + \mathcal{W}\sqrt{l - l_0}\ \xi_h, \end{aligned} \tag{5.21}$$

with $l \equiv |k|$. With this parameterization the variance for the integrated Wiener process component of the amplitude spectrum can be adjusted to the expected width of the peaks by choosing appropriate values for $(\mu_{\mathcal{W}}, \sigma_{\mathcal{W}})$. Furthermore, through adjusting the $(\mu_{\mathcal{R}}, \sigma_{\mathcal{R}})$, the amplitude spectrum model can be informed about the expected strength of the peaks. Note that since the ln-amplitude spectrum field is Gaussian distributed, the values of $(\mu_{\mathcal{W}}, \sigma_{\mathcal{W}})$ and $(\mu_{\mathcal{R}}, \sigma_{\mathcal{R}})$ have to be adjusted such that the $\exp(\gamma(|k|))$ field has the desired variance of the integrated Wiener process and the desired asperity. Therefore, log-normal moment matching

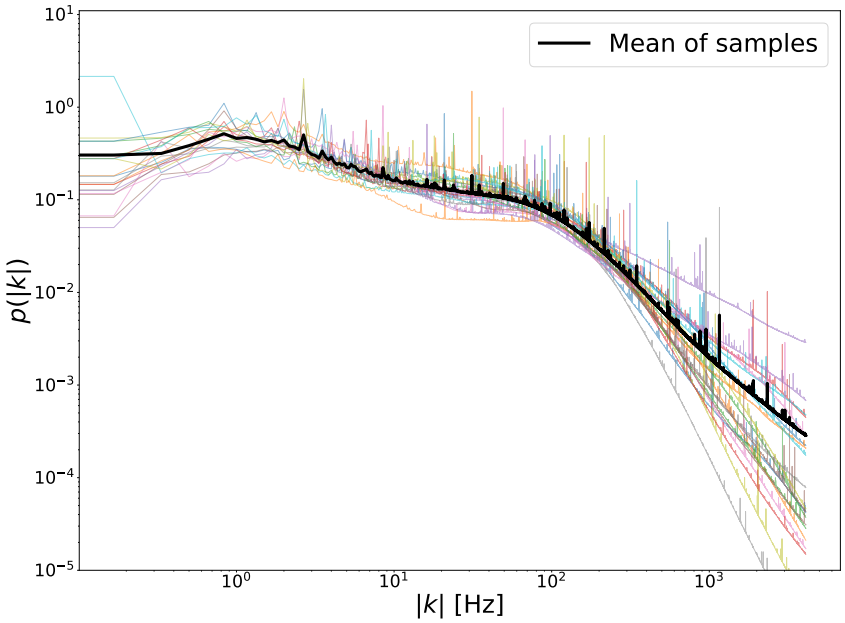


Figure 5.3: Posterior samples of the amplitude spectrum A_p of the IG model. Mean of the posterior samples is shown in black, and different colours emphasize different realizations which are sampled around the inferred posterior mean.

step has to be performed as well. For more details on the amplitude model please refer to [33, 37].

Due to the fact that the asperity parameter has to be adjusted to high values ($\approx 10^1 - 10^2$) the here presented model is termed 'High asperity model' (HA for short) and hence the name of the title of this section.

Before continuing onto the results section, a brief summary of the hyperparameters used for this model in the analysis of the ASIM data is given in table 5.2. The high value for the asperity mean $\mu_{\mathcal{R}}$ is followed by a strong steepness enforcement through μ_m in order to reduce the contribution of high k-modes, which are not expected to provide any statistically significant QPO signal.

5.3 Results and Discussion

Here a summary of the obtained results is given in the same order as the models themselves were introduced in section 5.2. The results were obtained by utilizing the MGVI algorithm as described in section 3 (for a brief summary of the most relevant pieces refer to section 4.3). For both models the number of samples used to perform the stochastic gradient descent was increased from 4 to 10 antithetic samples (for details about antithetic sampling please refer to [36]) as the conjugate gradient steps for KL minimization were performed, accumulating to 60 steps in total. Besides the results, the limitations for both models are discussed and suggestions for future work are made.

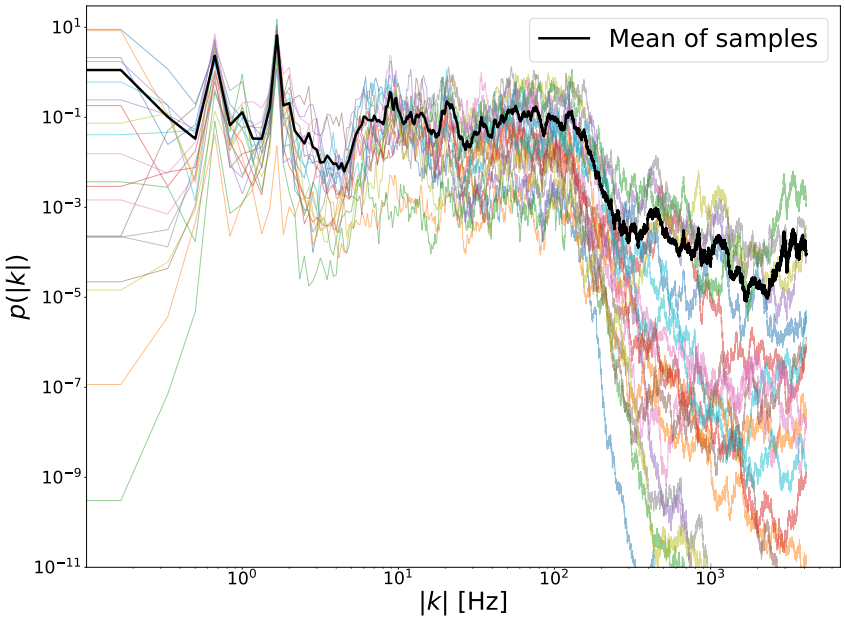


Figure 5.4: Posterior samples of the amplitude spectrum for the HA model. The black line shows the sample mean, and different colours emphasize different realizations which are sampled around the inferred posterior mean.

5.3.1 Results for the Inverse Gamma model

On fig. 5.3 and fig. 5.5 the results of applying the IG model to the dataset are shown. What can be seen from the reconstruction is that indeed the algorithm believes that in order to explain the counts seen in the data, a damped periodic (quasi-periodic) oscillation is plausible hence the power at certain frequencies is deemed to be higher than others as depicted in fig. 5.3. The corresponding QPO signal in the data space is plotted on fig. 5.5. As can be seen from the figure, the posterior samples, shown in red, indeed follow the regions where the higher number of counts is observed, hence demonstrating that the algorithm is able to capture these features of the data itself explaining them through the QPO signal.

Although the results here may seem appealing, what can also be noticed, especially on fig. 5.3, is that the posterior samples do not seem to have a clear preference towards any particular excited frequency. The posterior samples offer many different possibilities. This can be a consequence of the strong assumption that peaks are expected to be strongly localized, which, as discussed in section 5.2.2, can be far from reality.

Furthermore, the obtained reconstruction does not seem to be as robust as needed for clear astrophysical conclusions. The fig. 5.7c and fig. 5.7d illustrate this by showing two runs of this model, with different initial conditions denoted by the use of different seed. The converged reconstructions of the amplitude spectrum don't seem to have a clear preference towards any of the excited frequencies. It is left for future work to try and deduce what would be the cause of such behaviour.

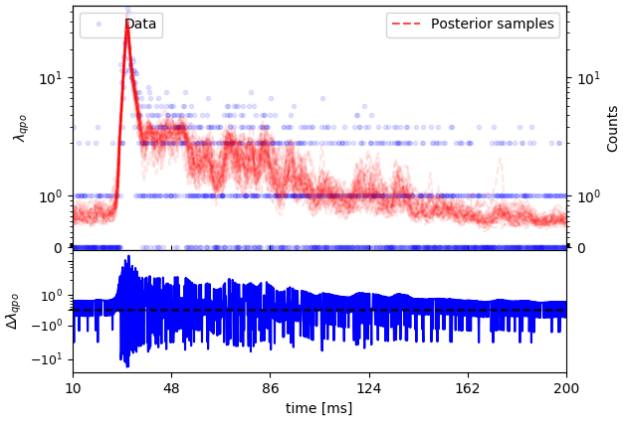
5.3.2 Results for the High Asperity model

The figure fig. 5.4 shows the reconstruction of the amplitude spectrum for the HA model. Comparing this reconstruction with the one done by the IG model (fig. 5.3) one can see that the samples show clearer preference towards peaks located at certain frequencies. Furthermore, the spread of peaks at the inferred excited frequencies is as well smaller, especially for the peaks located at $\approx 0.7, 1.0, 10$ and 20Hz. Note that since the time window of the observation was of the order of $\approx$ 2s. the first two detected peaks at ≈ 0.7 and ≈ 1.0Hz could be a consequence of the zero-padding which was necessary in order to decorrelate reconstruction of the QPO signal at the edges of the time domain. As far as the frequencies at 10 and 20Hz are concerned, the zero-padding should not have any effect. Therefore it could be plausible that these two frequencies are indeed present in the data itself. If one does a similar test as for the IG model, by changing initial conditions and reconstructing again the amplitude spectrum, these two frequencies seem to remain excited in the new reconstruction as well (see figures fig. 5.7a and fig. 5.7b). In order to make further conclusions a more robust analysis has to be performed for these results.

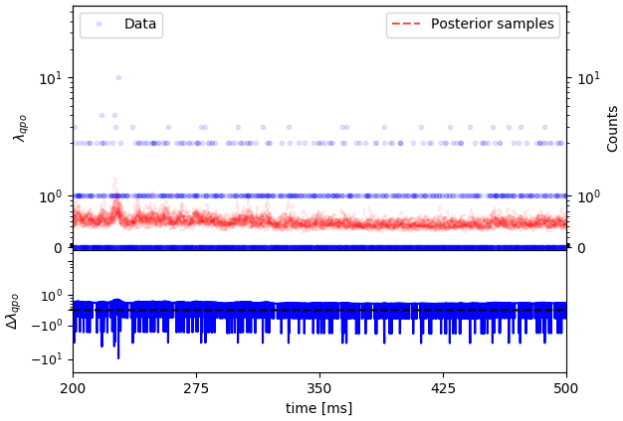
One possibility to check the significance of the detected lines is to shift the photon counts by random time shifts which are drawn from a Gaussian distribution of a width equivalent to few times $\approx 1/10$ Hz $= 0.1$s and $\approx 1/20$ Hz $= 0.05$s. By doing this, the original QPO signal which had excited frequencies at 10Hz and 20Hz will be preserved, while the data itself will be left with roughly the same statistics as it had before, i.e. shape of the light curve, total number of photons and Poisson statistics remain unchanged. Performing reconstruction now on this dataset, and repeating the procedure, the significance for the detected peaks at the inferred excited frequencies can be assessed. It is not clear however how many repetitions should be performed in order to claim certain level of significance and this is left for future work.

Besides the amplitude spectrum reconstruction, the data space reconstruction is given as well and it is available in fig. 5.6. A similar behaviour as for the IG model can be seen on these figures and not much information about the excited frequencies can be inferred except that the model seems to capture all the relevant features in the data by covering with the posterior samples all the counts above the inferred background level. What can also be noticed from fig. 5.6c and fig. 5.6b is that the model thinks the counts coming before $t < -50$ms and $t > 350$ms are completely due to the background which aligns well with what is expected from the nature of X-ray bursts. Namely, it is expected that it consists of three phases. The one immediately before the burst, the phase during the burst of flux and the phase after this which should quickly reduce down to the background due to the rapid weakening of the flux. This feature however is not noticeable on fig. 5.5c and fig. 5.5b, and it seems the model is capturing the background noise in these regions rather than the underlying quasi periodic signal.

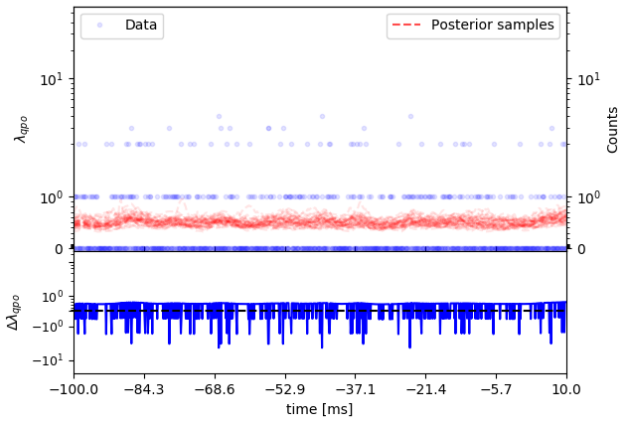
In conclusion, it seems that the approach used in the HA model should be the one that is to be utilized in future for extracting QPO signals from photon count data.



(a) $t \in [10, 200]$ ms

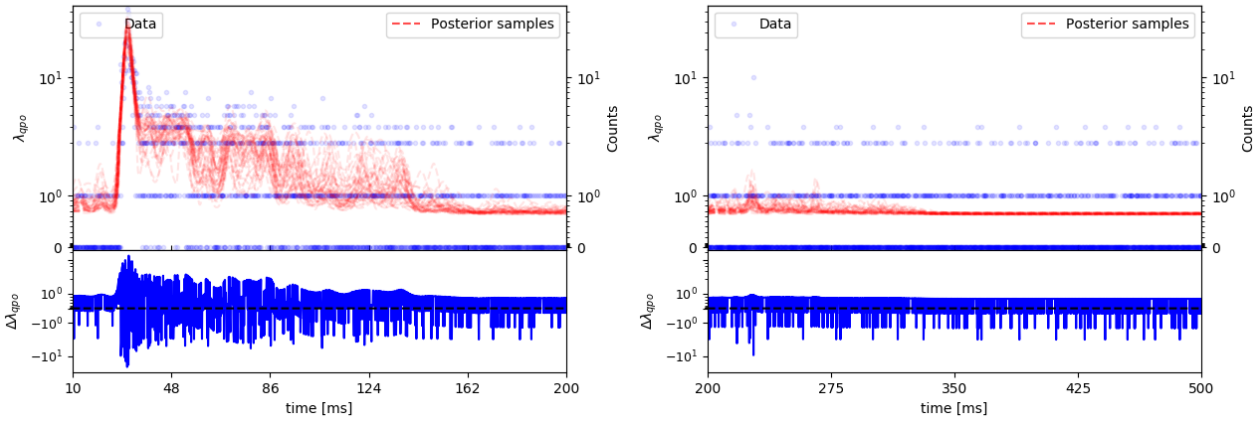


(b) $t \in [200, 500]$ ms



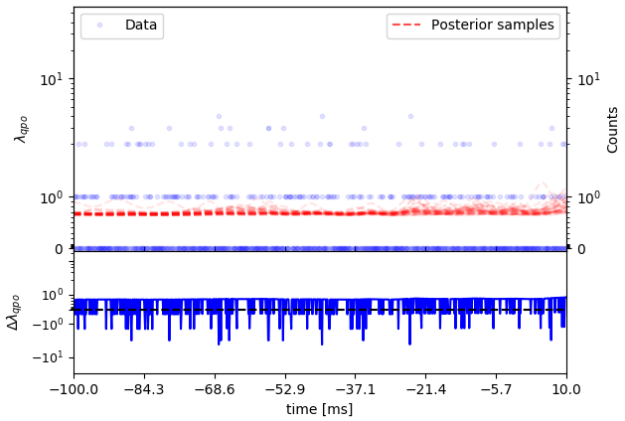
(c) $t \in [-100, 10]$ ms

Figure 5.5: Posterior samples of the reconstruction of the QPO signal using the IG model shown in red with data points (photon counts) shown in blue, within different time windows.



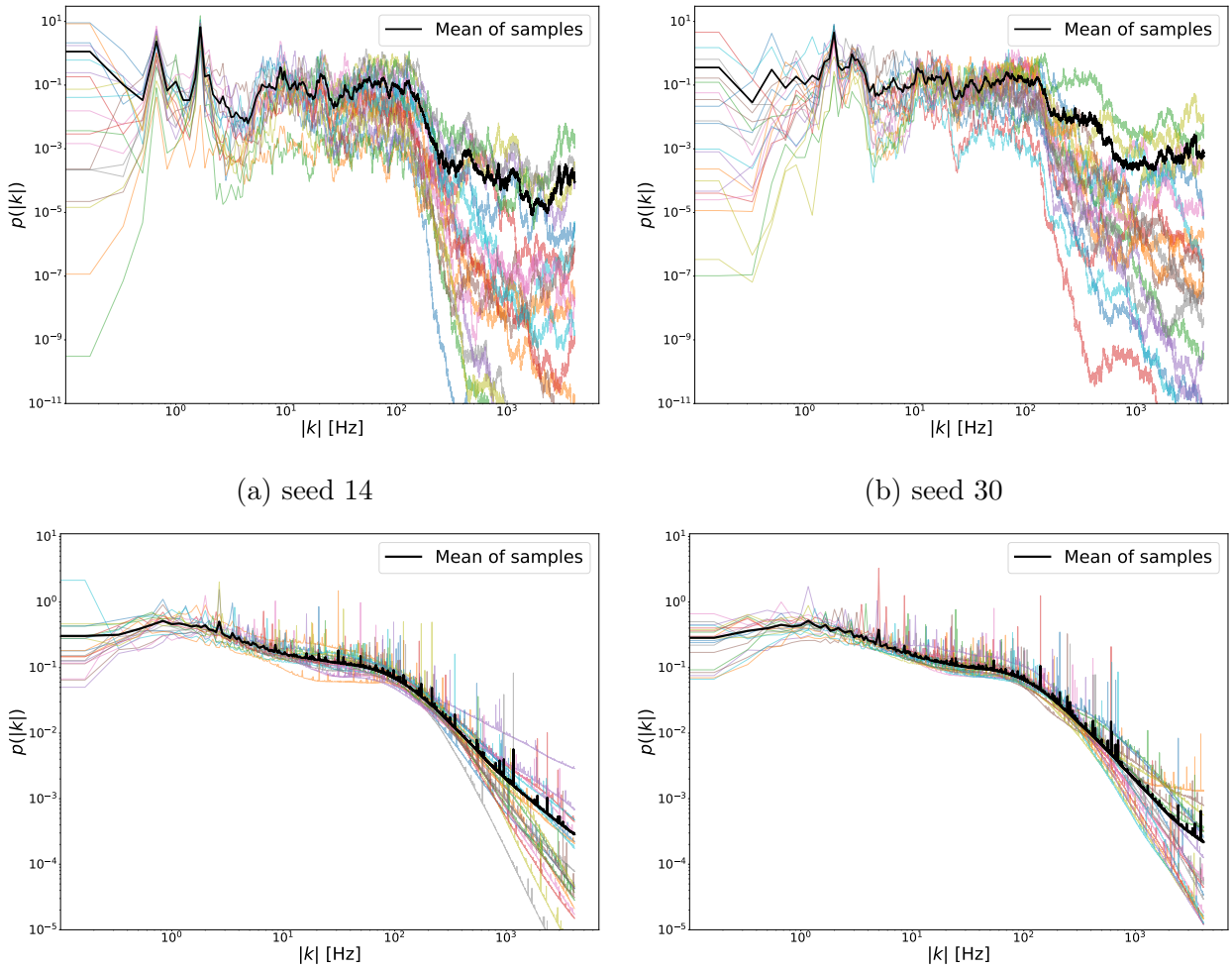
(a) $t \in [10, 200]$ ms

(b) $t \in [200, 500]$ ms



(c) $t \in [-100, 10]$ ms

Figure 5.6: Posterior samples of the reconstruction of the QPO signal using the HA model shown in red with data points (photon counts) shown in blue, within different time windows.



(a) seed 14

(b) seed 30

(c) seed 14

(d) seed 30

Figure 5.7: Comparison of different runs for both the IG and HA models, started with different initial conditions denoted by different seeds.

6 Conclusion

In this thesis three different topics within the field of Bayesian inference were addressed. In the first part of the thesis a novel metric Gaussian variational inference approach [36] was analysed in detail with aim to improve on it. The algorithm uses a particular type of a Gaussian approximation in order to perform variational inference of the quantities of interest. This variational inference is done through minimizing the KL divergence between the true posterior and the approximating Gaussian. The parameterization taken for this Gaussian was chosen such that its covariance has three important properties. First of all, it is positive-definite, since otherwise it will not represent a covariance. Second, it gives reasonable uncertainty estimates in limiting cases such as the case of infinite data and the opposite case of very scarce data. Third, that it allows for drawing samples from the approximate Gaussian without the need of explicitly constructing the covariance. The last one allowing for linear scaling with the increase of the inference problem. Therefore, the proposed covariance of the approximating Gaussian was chosen as a particular combination of Fisher information metric of the likelihood and the prior. MGVI algorithm then uses this Gaussian approximation to perform the minimization of KL through the use of stochastic gradient descent. In order to calculate the stochastic gradient an approximation is made by taking into account only the first order terms w.r.t. the derivatives of the joint information Hamiltonian of the problem. Therefore, what is done in this thesis is to calculate the higher order terms and asses their impact. These terms were calculated and it was noticed that they correspond to particular objects naturally appearing in information geometric approach to statistical inference [15, 25]. Although this interesting connection appears naturally it was immediately clear that the computational cost for using these terms for large scale inference problems will overshadow the increase in accuracy. The reason being the involvement of tracing operations between these information geometric objects and the Fisher metric. Nonetheless, it seems worthwhile to try and implement these correction terms for small scale problems in order to fully understand their effect on the variational inference. A project left to be done in the future. Last point that was made was that the correction terms vanish exactly in the case of fully Gaussian posterior, agreeing with claims made in the original paper.

In the second part of the thesis attention is turned to constructing a Bayesian inference algorithm for causal inference problems. Here the emphasis was on discovering the causal structure directly from observed data, without any interventions allowed. Here, a natural correspondence between the Bayesian hierarchical models and SEMs was exploited in order to motivate a particular type of generative models which are able to identify underlying causal structures. Namely, the models used were assuming additive noise and allowed for nonlinear mappings between the observed variables, hence constraining the problem and therefore allowing for distinguishing different causal structures from one another if present inside the observed dataset [19, 24]. In order to actually separate the causal models an estimate of the lower bound to the model evidence was calculated by making use of the particular parameterization the MGVI algorithm has. The novel thing that is introduced in this work is the use of IFT formalism alongside the MGVI method for performing causal inference nonparametrically. This allowed for developing models capable of discovering causal structures with a hidden common cause present between the observed variables as described in section 4.2.2. It was shown that the developed method performs comparably to the state-of-the-art approaches

on the standard benchmark datasets, such as the `tcep` dataset provided by [26]. Furthermore, as the need for new benchmark datasets was emphasized in [23, 26], here we use the developed Bayesian forward models to generate new test cases such as the `ConSyn` dataset described in section 4.5 and test our methods even further.

In the third and final part photon count data from a recent cosmic X-ray burst was analysed for the presence of quasi periodic signals. These quasi periodic signals offer a unique probe to the internal structure properties of the emitting magnetars [21]. The work that was done builds upon the work of [32] and extends it through the use of a more powerful inference algorithm made possible by the MGVI inference scheme. The novel method infers directly from the data the correlation structure of the quasi periodic signal as well, thus allowing for the discovery of excited frequencies which are expected to be present inside the data if the quasi periodic signal is indeed there. Two different models were developed and applied to the dataset. Neither of the models allowed for the discovery of any significant quasi periodic signals in this Bayesian analysis. Therefore, the end result of this part was to exploit different possibilities and offer an insight towards how the future analysis of these problems should be performed within the IFT context. Furthermore, a method of estimating uncertainties for the inferred quasi periodic signal is as well discussed and suggested for the future work.

All the implemented code is available under references [42] and [43].

References

[1] Norbert Wiener et al. "Generalized harmonic analysis". In: *Acta mathematica* 55 (1930), pp. 117–258.

[2] Alexander Khintchine. "Korrelationstheorie der stationären stochastischen Prozesse". In: *Mathematische Annalen* 109.1 (1934), pp. 604–615.

[3] Richard T Cox. "Probability, frequency and reasonable expectation". In: *American journal of physics* 14.1 (1946), pp. 1–13.

[4] MP Shutzenberger. "A generalization of the Fréchet-Cramér inequality to the case of Bayes estimation". In: *Bull. Amer. Math. Soc* 63.142 (1957).

[5] Simon Duane et al. "Hybrid monte carlo". In: *Physics letters B* 195.2 (1987), pp. 216–222.

[6] C Radhakrishna Rao. "Information and the accuracy attainable in the estimation of statistical parameters". In: *Breakthroughs in statistics*. Springer, 1992, pp. 235–247.

[7] Jonathan Richard Shewchuk et al. *An introduction to the conjugate gradient method without the agonizing pain*. 1994.

[8] Edwin T Jaynes. *Probability theory: the logic of science*. Washington University St. Louis, MO, 1996.

[9] Richard B Lehoucq, Danny C Sorensen, and Chao Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM, 1998.

[10] Harald Cramér. *Mathematical methods of statistics*. Vol. 43. Princeton university press, 1999.

[11] Judea Pearl et al. "Models, reasoning and inference". In: *Cambridge, UK: CambridgeUniversityPress* (2000).

[12] Peter Spirtes et al. *Causation, prediction, and search*. MIT press, 2000.

[13] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[14] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[15] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. Vol. 191. American Mathematical Soc., 2007.

[16] Peter D Grünwald and Abhijit Grunwald. *The minimum description length principle*. MIT press, 2007.

[17] Yuri Levin. "On the theory of magnetar QPOs". In: *Monthly Notices of the Royal Astronomical Society* 377.1 (2007), pp. 159–167.

[18] Jiji Zhang. "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias". In: *Artificial Intelligence* 172.16-17 (2008), pp. 1873–1896.

[19] Patrik O Hoyer et al. "Nonlinear causal discovery with additive noise models". In: *Advances in neural information processing systems*. 2009, pp. 689–696.

[20] Torsten A Enßlin and Cornelius Weig. "Inference with minimal Gibbs free energy in information field theory". In: *Physical Review E* 82.5 (2010), p. 051112.

[21] Michael Gabler et al. "Magneto-elastic oscillations of relativistic stars". In: *arXiv preprint arXiv:1007.0856* (2010).

[22] Ariel Caticha. "Entropic dynamics, time and quantum theory". In: *Journal of Physics A: Mathematical and Theoretical* 44.22 (2011), p. 225303.

[23] Dominik Janzing et al. "Identifying confounders using additive noise models". In: *arXiv preprint arXiv:1205.2640* (2012).

[24] Jonas Peters et al. "Causal discovery with continuous additive noise models". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2009–2053.

[25] Shun-ichi Amari. *Information geometry and its applications*. Vol. 194. Springer, 2016.

[26] Joris M Mooij et al. "Distinguishing cause from effect using observational data: methods and benchmarks". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1103–1204.

[27] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017.

[28] Jakob Knollmüller and Torsten A Enßlin. "Encoding prior knowledge in the structure of the likelihood". In: *arXiv preprint arXiv:1812.04403* (2018).

[29] Maximilian Kurthen and Torsten A Enßlin. "Bayesian Causal Inference". In: *arXiv preprint arXiv:1812.09895* (2018).

[30] Reimar H Leike and Torsten A Enßlin. "Towards information-optimal simulation of partial differential equations". In: *Physical Review E* 97.3 (2018), p. 033314.

[31] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[32] Daniel Pumpe et al. "Search for quasi-periodic signals in magnetar giant flares-Bayesian inspection of SGR 1806-20 and SGR 1900+ 14". In: *Astronomy & Astrophysics* 610 (2018), A61.

[33] Philipp Arras et al. "Unified radio interferometric calibration and imaging with joint uncertainty quantification". In: *arXiv preprint arXiv:1903.11169* (2019).

[34] Torsten Enßlin. *Lecture notes on Information Theory and Information Field Theory*. `https://wwwmpa.mpa-garching.mpg.de/~ensslin/lectures/lectures.html`. 2019.

[35] Torsten A Enßlin. "Information theory for fields". In: *Annalen der Physik* 531.3 (2019), p. 1800127.

[36] Jakob Knollmüller and Torsten A Enßlin. "Metric Gaussian Variational Inference". In: *arXiv preprint arXiv:1901.11033* (2019).

[37] Philipp Arras et al. "The variable shadow of M87". In: *arXiv preprint arXiv:2002.05218* (2020).

[38] Reimar Heinrich Leike. "Galactic dust and dynamics". In: *Ludwig Maximilians University, PhD thesis* (2020). URL: `http://nbn-resolving.de/urn:nbn:de:bvb:19-264215`.

[39] M Marisaldi et al. “GRB200521A: ASIM observation”. In: *GCN* 27815 (2020), p. 1.

[40] Pavel Minaev and Alexei Pozanenko. “GRB 200415A: magnetar giant flare or short gamma-ray burst?” In: *arXiv preprint arXiv:2008.12752* (2020).

[41] Fermi GBM Team et al. “GRB 200415A: Fermi GBM Final Real-time Localization”. In: *GCN* 27579 (2020), p. 1.

[42] Andrija Kostic. *Bayesian Causal Inference using NIFTy*. URL: `https://gitlab.mpcdf.mpg.de/ift/students/andrija-kostic`.

[43] Andrija Kostic. *Quasi Periodic Signal analysis using NIFTy*. URL: `https://gitlab.mpcdf.mpg.de/akostic/magnetar-qpos`.

[44] Martin Reinecke, et al. *NIFTy – Numerical Information Field TheorY*. Version nifty6. URL: `https://gitlab.mpcdf.mpg.de/ift/NIFTy`.

Acknowledgments

I would like to thank first of all to my supervisor Torsten Enßlin who allowed me to work on my thesis within the information field theory group of the Max Planck Institute for Astrophysics in Garching. I am immensely grateful especially because he was very encouraging and believed in my capabilities by taking me into the group even though I had not attended the Information Field Theory course at the university beforehand and had to learn the content from ground zero. He always had time and patience for my questions and offered invaluable guidelines whenever they were needed. He allowed me the freedom to undertake different research directions and therefore expand my views even further. He suggested valuable comments for this text which shaped it into its current form.

I am very thankful to Reimar Leike, a former PhD student in the IFT group, who offered me his time and knowledge selflessly. He was instrumental for a significant part of the methods I developed here. Always having time for a discussion and patience to let me figure out the problem on my own, sometimes spending few hours in front of the black board before clarifying all the ambiguities I had. Academically he is a great colleague full of ideas and insights, informally he is indeed a great friend to have. He as well provided comments and read pieces of this text, checking for computational errors wherever they arose.

Furthermore, my gratitude goes to Sebastian Hutschenreuther, also a former PhD student of the IFT group, who provided pieces of code needed for the full development for some of the models described in the text. He was always available and pointed out few very important typos in the text of this thesis hence making the content more comprehensible and transparent.

My thanks go as well to Philipp Arras and Philipp Frank, current PhD students of Torsten Enßlin, who provided crucial implementations inside of the `NIFTy` package used in this work. Philipp A. offered indispensable insight into the inner workings of the `NIFTy` code and helped me to understand how I should proceed with tailoring it to suit my needs.

I would also like to thank all the master students within the group, especially Gordian Edenhofer and Jakob Roth for helping me understand the contents of the Information Field Theory course by discussing, at times for hours, the solutions of the problem sets as well as the content of the lecture notes.

My gratitude goes to the whole IFT group for providing excellent environment for doing research and for being a source of inspiration and friendship.

Heartfelt thanks also to my fiance for being a dear, caring and encouraging companion. Thanks to my parents for all their support throughout my academic journey so far.

Finally I would also like to acknowledge all the support I received from the Deutscher Akademischer Austauschdienst (DAAD). DAAD organization financed my second year of master studies at the Ludwig Maximilians University and was extremely supportive and responsive during the uncertain times of the lockdown caused by the COVID-19 pandemic.