Assessing causality in financial time series

Anna Zaremba

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London.

Department of Computer Science
University College London

January 31, 2022

To my Dad, Piotr Zaremba.

Dla mojego Taty, Piotra Zaremby.

Statement of Originality

I, Anna Zaremba, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

We develop new classes of semiparametric multivariate time series models based on Multi-Output Gaussian Processes and warped Multi-Output Gaussian Processes. These describe relationships between a current vector of observations and the lagged history of each marginal time series. We encode a serial dependence structure through mean and covariance functions and introduce a more complex dependence structure using copulas to couple each warped marginal Gaussian process. Within this class of models our primary goal is to detect causality and to study the interplay between the causal structure and the dependence structure. We do not, however, require true representation of the data generating process, but we model structural hypotheses regarding how causality may have manifested in the observed vector valued processes. With our framework we test the dependence with regards to the structures that are specified, and can use testing for causality under different model assumptions as a way to explore the data and the potentially complex dependence relationships. To perform the testing we consider several families of causality testing and develop compound tests which first require estimation/calibration of the mean and covariance functions parametrising the nonparametric vector valued time series. Our approach allows very general nonlinear dependence and causal relationships which are not often considered in classical parametric time series models, including causality in higher order information and joint extreme dependence features. We provide a generic framework which can be applied to a variety of different problem classes and discuss a number of examples to illustrate the ideas developed.

Throughout, we will consider, without loss of generality, two multivariate time series denoted by $\mathbf{X}_t \in \mathbb{R}^d, \mathbf{Y}_t \in \mathbb{R}^{d'}$ where one may assume, for instance, that these have been generated by observing partial realisations of a generalised diffusion processes:

$$d\mathbf{X}_t = \mu_X(t, \mathbf{X}_t^{-k}, \mathbf{Y}_t^{-l}, \mathbf{Z}_t^{-m})dt + \mathbf{\Sigma}_X(t, \mathbf{X}_t^{-k}, \mathbf{Y}_t^{-l}, \mathbf{Z}_t^{-m})dW_t \tag{1}$$

$$d\mathbf{Y}_t = \mu_Y(t, \mathbf{X}_t^{-k}, \mathbf{Y}_t^{-l}, Z_t^{-m})dt + \mathbf{\Sigma}_Y(t, \mathbf{X}_t^{-k}, \mathbf{Y}_t^{-l}, \mathbf{Z}_t^{-m})dW_t', \tag{2}$$

where $\{\mathbf{Z}_t\}$, which may or may not be included, is some real process that we will call side information,

dW_t, dW'_t are two different Brownian motions, possibly with marginal serial correlation and/or instantaneous cross-correlation. All of those processes are only partially observed, and may be sampled at irregular intervals. The form of drift and volatility in Equations (1 - 2) means that the processes $\{\mathbf{X}_t\}$ and $\{\mathbf{Y}_t\}$ can be conditionally dependent on each other, and this dependence can be introduced through both the drift and the volatility. Such generalised diffusion models can induce in the marginal process between $\{\mathbf{X}_t\}$ and $\{\mathbf{Y}_t\}$ different types of extremal dependence, depending on the forms of the drift and volatility functions.

We propose a smooth stochastic process statistical model to capture the smooth variation of the partially observed time series represented by data $\{\mathbf{X}_t\}_{t>0}$, $\{\mathbf{Y}_t\}_{t>0}$, $\{\mathbf{Z}_t\}_{t>0}$ using multiple output Warped Gaussian Process models. In this work we are interested in partial observations of these processes, for which the partially observed time series of $\{\mathbf{X}_t\}$ and $\{\mathbf{Y}_t\}$ will have different types of extremal dependence characteristics. We wish to detect the presence or absence of statistical causality where such extremal dependence features may or may not obfuscate the ability to detect causality in nonlinear partially observed time series models.

The rationale for developing a semiparametric solution for modelling the partially observed time series is that we may accommodate, through the use of Gaussian Process models, a wide variety of features for the hypotheses about the trends and volatility and importantly their possible causal structures, which can be formally tested in our framework. Furthermore the use of Warped Gaussian Process models allows to incorporate higher order dependence such as extremal tail dependence features.

Statistical Causality.

The notion of causality that lies at the centre of our research is the concept of statistical causality, based on comparing two predictive models. Quoting Wiener [1956]: *For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.* The null hypothesis of no causal relationship from time series $\{\mathbf{X}_t\}$ to $\{\mathbf{Y}_t\}$ means that including the past of $\{\mathbf{X}_t\}$ does not improve the prediction of future of $\{\mathbf{Y}_t\}$. In a most general form this can be written as equality of conditional distribution of Y, conditioning on either set of explanatory variables $(\mathbf{X}_t^{-k}, \mathbf{Y}_t^{-l}, \mathbf{Z}_t^{-m})$ denote past of the $\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Z}_t$ time series up to lags k, l, m respectively):

$$
\begin{aligned}
H_0 : \quad & p(\mathbf{Y}_t \mid \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) = p(\mathbf{Y}_t \mid \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}), \quad \forall_{t\in\mathbb{Z}} \\
H_1 : \quad & p(\mathbf{Y}_t \mid \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) \neq p(\mathbf{Y}_t \mid \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}), \quad \forall_{t\in\mathbb{Z}}.
\end{aligned} \tag{3}
$$

The type of casual dependence that is described by statistical causality is a mechanism that occurs at multiple lags over time - which could have been triggered by a sequence of processes, not an individual one. It can help to gain an insight into both cross-sectional and temporal dynamics of the data analysed.

Warped Multi-Output Gaussian Processes.

A Gaussian process is a Markov process, such that all finite dimensional distributions are Gaussian. While Gaussian processes models can accommodate wide range of properties and are very attractive for their easy implementation and optimisation, but they do not allow higher order dependence such as extremal tail dependence features. One way to generalise Gaussian process models so that higher order dependence can be handled, is to apply a transformation to the joint collection of Gaussian processes for each marginal time series model. We apply mean-variance transformation that results in the transformed variables having multivariate skew-t distributions and being finite dimensional realisations of a general multivariate skew-t process.

Motivation for the Model Choice.

There are numerous advantages of using Gaussian Processes, beginning with: ease of optimisation and interpretability of hyperparameters, flexibility, richness of covariance functions, allowing for various model structures. Using a likelihood ratio type test with a GP is a very natural choice, as estimating GP model parameters is often done on the basis of maximising likelihood, and therefore this estimation can be incorporated into the compound version of the likelihood ratio test (Generalised Likelihood Ratio Test, GLRT). From Gaussian variables, GPs inherited the property of being fully specified by the mean and the covariance, and so testing for model equivalence inherently means testing for equivalence of the mean and covariance functions. But many popular kernels do not have the ARD property, and using them for a likelihood ratio test settings gives no easy way to account for causal structures in covariance. Consequently, it is using GLRT with an ARD-GP that gives a uniformly most powerful test with an unparalleled flexibility: known asymptotic distribution under the null, explicit evaluation and in a closed form, and usefulness also for misspecified models. The proposed use of copula warping allows introduction of additional dependence, in particular tail dependence, while keeping the likelihood in closed form.

Application.

We provide a generic framework which can be applied to a wide range of problems, and which can be readily tailored or further extended. The illustrative examples included demonstrate how a range of

data properties can be encoded in the model, and how they might affect the detection of causality.

We present two real data application: to commodity futures data and inflation and interest rates. We show how the framework can be used in practice, and how it can be combined with, or enhance, more common approaches to analysing financial time series. Our observations are in line with financial interpretations, but they also offer additional insight and pose thought-provoking questions.

Structure of the thesis.

This thesis presents the research as it evolved: starting from an overview of a range of the causality methods already known, and demonstrating out why they are unsatisfactory. Subsequently, a new approach is presented – a method based on Gaussian processes, that was developed to solve the drawbacks of the methods presented in the first part. Afterwards, an extension is proposed to widen the range of dependence structures, as well as marginal properties of the data that can be incorporated.

Chapter 1 introduces the topic of the thesis, and reviews relevant literature. Chapter 2 discusses philosophical roots of the concept of statistical causality, as well as alternative notions of causality. After illustrating some of the varied ways of conceptual representation of causality, we present four distinct ways of modelling statistical causality. Chapter 3 contains background on the models considered: Gaussian processes, copulas and selected distributions. Chapter 4 describes inference procedures used: assessing hypothesis tests, generalised likelihood ratio test, permutation tests, and likelihood ratio test.

The second part, *New Perspectives on Causality Representation and Inference*, presents the main contribution of our work. It starts with Chapter 5 containing the theoretical background for describing and testing causality with GP models. Chapter 6 extends the model from the previous chapter by introducing mean-variance transformation that results in a warped GP model, which can describe causality in the presence of skewness and tail dependence. Chapter 7 describes how synthetic data has been simulated, details the algorithm for approximating likelihood in the warped GP, and provides information on other relevant algorithms and the software used to implement our method. Chapter 8 presents an extensive experiment section, which aim to show, firstly, the good behaviour of the proposed procedures (model sensitivity and misspecification analysis), secondly, good power of the test for a range of structures, and, thirdly, the interaction of causality and tail dependence. Applications to real-world data are described in Chapter 9, where time series for commodities and currency markets are analysed.

Finally, Chapter (10 presents the conclusions and directions for further development, and Chapter (A) provides supplementary material.

Impact Statement

Causal analysis is widely used in all areas of science, social science, engineering. Understanding of causal relationships is a crucial step in the analysis of data, regardless of its type. The novel methodology of modelling and testing statistical causality based on warped multiple output Gaussian Processes is applicable to a very wide range of multivariate nonlinear time series. The work in this thesis will open discussion about how different types of dependence, causality included, can interact with each other, and how statistical properties of the data alter those interactions and the recognition of causality.

This thesis also develops new classes of nonparametric multivariate time series models based on warped multiple output Gaussian Processes. These classes are extensions of widely popular Gaussian Process methods, and as such can be of interest to researchers in the field. Their properties, and the way they are constructed, have been inspired by copula methods, and skew-t copula in particular – important tools in financial risk management.

Outside of academia, this research could be of interest especially to the innovative branches of financial and technological sector, open to inventions and new ways of building investment strategies and managing risk. Insurers could utilise this work for risk pooling, asset managers could apply it for dynamic asset allocation, economist – for forecast and stress testing, banks and hedge funds – for interest rates trading. The fact that causality can be detected even for misspecified models, together with flexibility inherited from Gaussian Process component, allows the user to tailor the framework to their needs.

Our work includes an exploration of how statistical causality can be understood and how it compares to other notions of causality. Our aim is therefore to provide a reference point for understanding statistical causality that readers from a wide range of backgrounds could consult.

List of Publications

Measures of Causality in Complex Datasets with Application to Financial Data, Anna Zaremba, Tomaso Aste, *Entropy*, 16(4), 2309–2349, 2014;

Statistical Causality for Multivariate Nonlinear Time Series via Gaussian Process Models, Anna Barbara Zaremba, Gareth William Peters, *Methodology and Computing in Applied Probability*, *accepted*, 2022;

On-chain analytics for sentiment-driven statistical causality in cryptocurrencies, Ioannis Chalkiadakis, Anna Zaremba, Gareth W. Peters, and Michael J. Chantler, *Blockchain: Research and Applications*, *accepted*, 2022;

Non-Linear Multiple Output Warped Gaussian Process Models and Testing For Causality, Anna Zaremba, Gareth W Peters, *submitted*;

Gaussian Process statistical causality analysis between public news sentiment and crypto prices, Anna Zaremba, Gareth W Peters, Ioannis Chalkiadakis, *working paper*;

Research Presentations

2016 International Workshop on Spatial and Temporal Modeling from Statistical, Machine Learning and Engineering perspectives, July 2016, Tachikawa, Japan

12th International Confrence on Monte Carlo and Quasi-Monte Carlo Methos in Scientific Computing, August 2016, Stanford, California

Salzburg Workshop on Dependence Models & Copulas, September 2016, Salzburg, Austria

10th International Conference of the ERCIM WG on Computational and Methodological Statistics, *11th International Conference on Computational and Financial Econometrics*, December 2017, London, UK

15th International Conference on Computational and Financial Econometrics, December 2021, London, UK

International Visits

Institute of Statistical Mathematics, Tokyo, Japan, July 2016 - August 2016

Acknowledgements

> *I was taught that the human brain was the crowning glory of evolution so far, but I think it's a very poor scheme for survival.*
>
> Kurt Vonnegut,

Thank you to those who helped me with research and writing the thesis, and thank you to those who helped me survive the process!

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr Gareth W. Peters for the continuous support of my PhD study and related research, for his patience, motivation, support, and immense knowledge. The beginnings of our work together were overwhelming, as Gareth was providing interesting ideas and insightful comments faster that I was able to think. But his guidance helped me in all the time of research and writing of this thesis, and I could not have imagined having a better advisor and mentor for my PhD study. We are currently working on developing some of the ideas further, and I am very excited about all our future research collaboration!

I would like to thank Prof. Dr Tomaso Aste for his guidance, patience and supervision during earlier years of my PhD. He has helped me to find a topic I was truly passionate about, and to build a vision of what I wanted to achieve in my research. In my later years of PhD, I was especially grateful for the words "I'm happy to support you in any way you need."

My sincere thanks to the UK PhD Centre in Financial Computing for funding provided for my doctoral research. It would have been difficult for me to start my study without help and support from Prof. Philip Treleaven, the director of the UK PhD Centre in Financial Computing.

Very special thanks to Yonita Carter and Sarah Turnbull for their support, encouragement and advice, invaluable when I was facing additional challenges in working on my research.

My Dad has been the biggest inspiration in all of the years I have spent in education. When I was small he has been inventing stories about imaginary worlds as well as mathematical games. He would

answer all of my questions when I was at school, offer assistance with my Bachelor's thesis, and even with my Masters thesis. During my PhD he was the only person who would happily listen to me talking about my research for as long as I would wish to talk!

My older sister has also been a great source of support and inspiration. She has completed her PhD a few years ago – and I just cannot be the least educated child in my family! My mum has always been offering both support, as well as words of wisdom and reality. There is no - she would say - duty to complete a PhD, and one can be a good statistician even without it.

I am forever grateful to Maciek Makowski for being the biggest rock during such a big part of my PhD journey, for all his support, patience, encouragement, help and advice. I will never be able to repay for all he has done for me.

Very special thanks and appreciation to Claudia Cruz, whose appearance in my life made me a happier, stronger, and altogether a better human. She gave me the strength that I needed to face all the challenges that life threw at me at the last stages, and it is because of her support I was able to complete the journey from a mere Researcher to a fully fledged Doctor.

Last but not least, very special thanks to Kacper Chwiałkowski for inspiration and discussions at the earlier stages of my PhD, and for friendship throughout. and to Dorota Toczydłowska for discussions, support, motivation and friendship.

My Grandpa Jerzy did not see me starting my PhD, but I am sure he would have been proud. After all he has already found me a PhD supervisor (in marine palaeontology) before I finished high school...

Contents

List of Figures

List of Tables

Part I

Background and literature review

Chapter 1

Introduction

> *In all disciplines in which there is systematic knowledge of things with principles, causes, or elements, it arises from a grasp of those: we think we have knowledge of a thing when we have found its primary causes and principles, and followed it back to its elements.*
>
> Aristotle, *Physics Book 1*

This chapter presents an overview of the thesis. Starting with the motivation for pursuing the subject of causality, we then show the origins and the evolution of the concept, and how our treatment of it relates to other works. Subsequently, in the Contribution section we describe the novelty in our research, and we conclude with description of the structure of the thesis.

1.1 Motivation

Causality has been an object of research since Ancient times, but while scientists were interested in discovering causes of particular phenomena, studying causality as an abstract concept of its own was for a long time a domain of philosophy. That changed in the mid-XX century, when Wiener asserted that causality can be estimated, followed by Granger who proposed a practical implementation of Wiener's concept. By the beginning of XXI century the sheer number of publications describing, measuring and testing causality became too vast for a single person to follow.

When I started my PhD, I felt that I had found an important topic: causality is a ubiquitous concept, relevant and applicable in many fields, with the possibility to be mathematically defined and tested in numerous ways. To make it more fascinating, the philosophical connection is important, since the methods of causality do not generally reflect the everyday understanding of the word "causality", and understanding of what it is that they actually model becomes essential. This dissertation concentrates on one particular conceptual representation of causality, called "statistical causality" (a short introduction to other methods is provided in the Section 2.1).

My research on causality started from re-discovering Granger causality, an elegant concept which I first encountered, and ignored, in the Econometrics lecture during my Masters studies in Financial Engineering, when it was introduced in the context of model selection in a linear regression problem. When I was trying to narrow down my research interests, I considered two approaches to causality, one proposed by Granger, and a very different one – by Judea Pearl (please see Section (2.1) for more details about different notions of causality). The decision was made on the basis of the area of application I was primarily interested in – I wanted to work with time series, and financial time series in particular. The fact that Granger causality was also a research area favoured by my supervisor, Prof. Tomaso Aste, also played a role. The MRes part of my PhD program has mostly been devoted to building my knowledge of Granger causality and its extensions. I have chosen four methods to study in depth, the choice being based on the popularity of the methods and on the utility to a range of different fields. Scientific findings from the MRes project are in the article "Measures of Causality in Complex Datasets with Application to Financial Data", [Zaremba and Aste, 2014]. They have helped to shape the primary research questions for this thesis.

Granger causality has been proposed for linear regression problems and is therefore limited in its scope. To allow working with nonlinear relationships I have been researching several generalisations of Granger causality as well as alternative formulations, and settled on methods from the fields of econometrics, information theory and machine learning. The three methods I chose to compare with the classical Granger causality were: transfer entropy (TE), kernel ridge regression (KRR) and a nonparametric conditional dependence measure based on the normalised conditional cross-covariance operator (which I will refer to as HSNCIC, for Hilbert Schmidt Normalised Conditional Independence Criterion). A short summary of basic properties of the four methods is presented in Table (1.1).

All four methods were analysed in terms of, among others, recognising linear and nonlinear causality in multivariate time series, estimation, parameter selection, and ability to estimate causality in the presence of nonstationarity. The investigation of theoretical properties has been supplemented by experiments on simulated data structures with known causal structure, and on real data. The publication "Measures of Causality in Complex Datasets with Application to Financial Data"([Zaremba and Aste, 2014]) includes an analysis of economic data and application to equity and currency data. The economic data consisted of United States Consumer Price Index and US Dollar 1 month BBA LIBOR, both from Thomson Reuters and with the sampling period from December 1995 to June 2013, end of month data. The currency application used daily data for the period from 19 July 2010 to 22 July 2013 for six curry trade currency pairs (AUDJPY, CADJPY, NZDJPY, AUDCHF, CADCHF, NZDCHF), Standard & Poor's 500 Index, and the Chicago Board Options Exchange Market Volatility Index – the last one used as side information.

Section (2.2) of Chapter (2), Overview and Comparison of Existing Causality Methods, provides full details of the analysis of the four causal methods under consideration. Out of all the four methods, kernel ridge regression performed the best in terms of not rejecting the true hypothesis of lack of causality (small type I error) and rejecting the false hypothesis of lack of causality, i.e. spurious causality (small type II error). This method achieved small type I error as well as small type II error regardless of whether the data generating mechanism was linear or not, and whether the causality was direct or indirect. What is more, kernel ridge regression based test statistic can be estimated in a way that is computationally stable and efficient for small samples and high dimensionality. The parameter optimisation, however, has to be achieved through methods such as cross-validation. Cross-validation is not only computationally expensive, it does not allow the model to adapt parameters to each set of data, and in consequence it enforces at least some level of local stationarity; together with the fact that the optimised parameters cannot easily be interpreted in terms of structural properties of the data, the model might not adapt well to data out of the learning sample. The other three methods suffer from a range of disadvantages. Granger causality in the original form is unable to recognise nonlinear causality for a range of data. Typically the estimators used for transfer entropy are based on binning procedures, or clustering procedures like nearest neighbours, and they are less numerically stable than estimators for kernel ridge regression: exhibiting high errors of both type I and type II for small samples, and suffering from the curse of dimensionality[1]. HSNCIC should, as a kernel method, be computationally efficient for high-dimensional data, but in our experiments it has failed to reject the hypothesis of no causality when used on four dimensional systems (high type II error). Neither transfer entropy nor HSNCIC parameters can be interpreted in terms of data properties. There is no literature on TE, HSNCIC, or on krr for that matter, that would provide estimators with asymptotic properties known, outside of special cases, and as such permutation tests have to be used for significance assessment. Finally, none of the four methods are able to deal with structural properties such as nonstationarity[2] and long memory without having to change the model.

The objective of my research became to improve on the kernel ridge regression, without sacrificing its main strengths, and to expand the methodology to allow better understanding of causality as a part of dependence structure, as well as support special structures in the **marginal model** (model describing distribution of each of the time series separately) and in the **joint model**. I wanted to research the following directions:

1. Parametrisation of different forms of statistical causality, in particular second order aspects of the

[1]The curse of dimensionality is a bigger problem for the estimators based on binning, than based on nearest neighbours. In my experience, naive binning (histogram based approaches) typically fail in four dimensions. Approaches that use vector quantisation are more suited for higher dimensional data systems, but they are more difficult to implement, and do not scale well with large data, [Faes et al., 2011, Dimitriadis et al., 2016]

[2]Transfer entropy is not intrinsically unable to deal with nonstationary data, as it is possible to use density estimators for nonstationary data. This approach does not seem to be used, but instead it has been proposed to use Partial Symbolic Transfer Entropy [Papana et al., 2016] or use computationally expensive estimators on ensemble of realisations [Gómez-Herrero et al., 2015, Wollstadt et al., 2014].

	GC	TE	krr	HSNCIC
multivariate time series	+	+	+	+
nonlinear causality	−	+	+	+
Markov structure	+	+	+	+
known conditional distribution	+	−	−	−
test statistics in closed form	+	+	+	+
known asymptotic distribution (under null hypothesis)	+	−	+	−
causality in mean and in other statistical features	−	±	+	±
parameters interpretable in terms of model features	+	−	−	−
marginal distribution: heteroscedasticity	−	−	−	−
marginal distribution: tail dependence	−	−	−	−
marginal distribution: long memory	−	−	−	−
dependence structure: asymmetry	−	−	−	−
dependence structure: leptokurtic tails	−	−	−	−
dependence structure: tail dependence	−	−	−	−
testing framework to assess power of the test	+	−	−	−

Table 1.1: Properties of the four methods for statistical causality, analysed in Zaremba and Aste [2014]

process such as linear and nonlinear causality in covariance. The challenge is to form models that are consistent with such parametrisations and allow to form tractable causal tests.

2. Methods for modelling and testing causality that could be applied in the multivariate time series context, with linear and nonlinear causality present. In this context, we would like to achieve this with models accounting for the following properties:

   (a) flexible class of models that can capture linear / nonlinear causality, while admitting Markov structure and knowledge of the conditional distribution of the model;

   (b) test statistics can be evaluated in closed form;

   (c) known asymptotic behaviour of the test statistic under the null;

   (d) statistically unbiased, efficient, consistent and computationally efficient parameter optimisation, for which parameters can be interpreted with respect to the structural properties of the model;

   (e) can detect causality in the mean, covariance function, or higher order moments.

3. Extending the marginal models to incorporate range of special structures, for example nonstationarity, heteroscedasticity, leptokurtic tails, long memory.

4. Broadening the framework, so that the joint model allows a wider range of dependence structures, for example: asymmetry, leptokurtic tails, tail dependence. What influence do such statistical features in underlying data generating mechanism have on the ability to perform inference, detect causality, and accuracy and power of the test?

5. How can one develop a testing framework to asses the power of the test for the models that meet the requirements from the points 1 through 4?

To address the proposed questions, and study each of these attributes in a common framework, without the need to change models for different types of data, a class of models is needed that is sufficiently rich, but also interpretable. Following the review of literature on the topic, we have concluded that models based on Gaussian processes (GP) allow us to address all of our research questions. GPs are nonlinear, semiparametric models that can be used for autoregressive multivariate time series and that have many desirable properties. Deriving from the properties of Gaussian distributions, they have known conditional distributions, which results in test statistics that can be evaluated in closed form, and known asymptotic behaviour of the test statistic under the null. Hyperparameter optimisation can be performed in an efficient and easily interpretable way. Since GP model hyperparameters can be optimised by maximising likelihood, then the likelihood ratio test – which is the uniformly most powerful test for comparing model fit – also provides optimum parameters and can be used as a compound test. What is worth emphasising, is that the Generalised Likelihood Ratio Test (GLRT) being a test for model selection, allows to test for a model that is most useful, rather than one that is well specified. Furthermore, the flexibility of a GP means that a wide range of data can be modelled through choice of functional form for the mean and covariance functions. The practical consequences is that the dependence is tested with regards to the structures that are specified, allowing testing for causality under many different model assumptions, which - crucially - means the framework can still capture causal relationships in misspecified context. The subsequent extension of the model is inspired by machine learning methods such as Gaussian mixture models, mean-variance transformations and GP warpings. This research proposes warping GP with a leptokurtic transform based on mean-variance mixing with an inverse gamma distribution.

The following acronyms will be used in this thesis:

GP	- time series
GPC	- random variable
wGP	- warped Gaussian Process
wGPC	- our framework using warped Gaussian Process for Causality
GLRT	- Generalised Likelihood Ratio Test

1.2 Notation

In this thesis we will use the following notation:

$\{Y_t\}$	time series
Y_t	random variable

$\mathbf{Y}_t$	column random vector
$y_t, \mathbf{y}_t$	realisation (univariate or multivariate)
$\mathbf{Y}_{t_1:t_2} = [Y_{t_1}, ..., Y_{t_2}]$	random vector
k, l, m	lag values, typically associated with, respectively, $\{X_t\}, \{Y_t\}, \{Z_t\}$
$\mathbf{Y}_t^{-l} = \left[\mathbf{Y}_{t-l+1}^{\mathrm{T}}, \mathbf{Y}_{t-l+2}^{\mathrm{T}}, ..., \mathbf{Y}_t^{\mathrm{T}}\right]$	random vector
$\mathbf{Y}^{-l} = \mathbf{Y}_{t_1:t_2}^{-l} = [\mathbf{Y}_{t_1-l+1:t_2-l+1}, ..., \mathbf{Y}_{t_1:t_2}]$	random vector
$\mathcal{F}_t$	filtration
$\mathcal{F}_t^Y$	natural filtration for process Y_t
$\mathbf{Q}_t$	random vector, used to denote joint distribution $[X_t, Y_t, Z_t]^T$, or its subset
$\mathbf{A}$	matrix
$A_{i,j}$	i, j-th element of a matrix $\mathbf{A}$
$a_{X,1}$	coefficient, or element of a matrix with alternative indexing
$\|\mathbf{Y}_{t_1:t_2}\|_2$	quadratic norm
$\|\mathbf{Y}_{t_1:t_2}\|_{\max} = \max_{t\in\{t_1,...,t_2\}} \{Y_{t_1}, ..., Y_{t_2}\}$	maximum norm
d	dimension
q	a difference in dimensionality between model A and B
p, p_x, p_y	number of lags
$X_t \perp\!\!\!\perp Y_t \mid Z_t$	conditional independence of random variables X_t, Y_t, conditioned on Z_t
$\pi(Y)$	density of random variable Y
$F(Y)$	distribution of random variable Y
π	constant 3.1415...
$\mathbb{P}(Y = y)$	probability of an event $Y = y$
$\zeta(t),\ t \in \{t_1, ..., t_2\}$	random permutation in the set $t \in \{t_1, ..., t_2\}$
$\mathbf{1}(M)$	characteristic function of a set M
$\mathcal{H}$	Hilbert space
$\langle\cdot,\cdot\rangle$	inner product
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$	inner product associated with Hilbert space $\mathcal{H}$
A	a linear operator
$\epsilon_t, \epsilon_t^X, \epsilon_t^Y$	(additive) noise; typically we will assume that its i.i.d. $\mathcal{N}(0, \sigma^2)$
$\mu(\cdot), \mu, \mu^X, \mu^Y$	mean function
$k(\cdot,\cdot), k, k^X k^Y$	covariance (kernel) function
$\boldsymbol{\mu}, \boldsymbol{\mu}^X, \boldsymbol{\mu}^Y$	mean vector

$\mathbf{K}, \mathbf{K}^X, \mathbf{K}^Y, \mathbf{K}^{XX}, \mathbf{K}^{YY}, \mathbf{K}^{XY}, \mathbf{K}^{YX}$	covariance (kernel, Gramm) matrix
$\oplus$	direct sum of matrices
$*$	convolution of two functions
$L^p, L^p(\mathcal{X})$	An L^p space of functions on $\mathcal{X}$, for which the p-th power of the absolute value is Lebesgue integrable
$C(u_1, ..., u_n)$	n-dimentional copula
$\mid x \mid$	absolute value of x
$\delta(\cdot, \cdot)$	Kronecker delta
$Vec(.)$	operator converting a matrix into a vector.

1.3 Related work

Over sixty years ago, Norbert Wiener formulated causality as a mathematical, rather than philosophical concept, one that was based on predictive models:

"*For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.*" Wiener [1956].

In 1963, Clive Granger, the Nobel laureate in economics from 2003, proposed what is often seen as the first testable definition of causality, consistent with Wiener's definition, Granger [1963]. Granger's concept was defined for stochastic processes and was consistent with time direction and consequently, with the philosophical and common sense understanding that cause precedes the effect. The context was that of linear, multivariate, stationary and nondeterministic time series with autoregressive representation.

Let $\{X_t\}, \{Y_t\}$ be the two time series whose causal relationship will be analysed, and $\{Z_t\}$ will be a third time series called "side information". All three time series are stationary and non-deterministic, and can be represented with a basic form of autoregressive representation, for which we will temporarily introduce a shortened notation $\mathbf{Q}_t^T = [\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Z}_t]$, with $\mathbf{X}_t \in \mathbb{R}^{d_X}, \mathbf{Y}_t \in \mathbb{R}^{d_Y}, \mathbf{Z}_t \in \mathbb{R}^{d_Z}$, and so $\mathbf{Q}_t : \mathbb{R}^d$, $d = d_X + d_Y + d_Z$. Let $\mathcal{F}_t^X, \mathcal{F}_t^Y, \mathcal{F}_t^Z$ be natural filtrations for, respectively, $\mathbf{X}_t, \mathbf{Y}_t, \mathbf{Z}_t$.

$$\mathbf{A}_0 \mathbf{Q}_t = \sum_{i=1}^{p} \mathbf{A}_i \mathbf{Q}_{t-i} + \boldsymbol{\epsilon}_t, \qquad \mathbf{Q}_s \text{ is adapted to a filtration } \mathcal{F}_s^X \cup \mathcal{F}_s^Y \cup \mathcal{F}_s^Z \tag{1.1}$$

$$\mathbb{E}\left[\boldsymbol{\epsilon}_t^T \boldsymbol{\epsilon}_s\right] = \mathbf{1}_{t=s}, \tag{1.2}$$

where $\mathbf{A}_i$ are $d \times d$ matrices, and $\boldsymbol{\epsilon}_t^T = [\epsilon_{1,t} \cdots \epsilon_{dt}]$ is white noise. The assumption of noise being standardised is not critical, and in fact it is often relaxed. The maximum lag p, which can be infinite, has

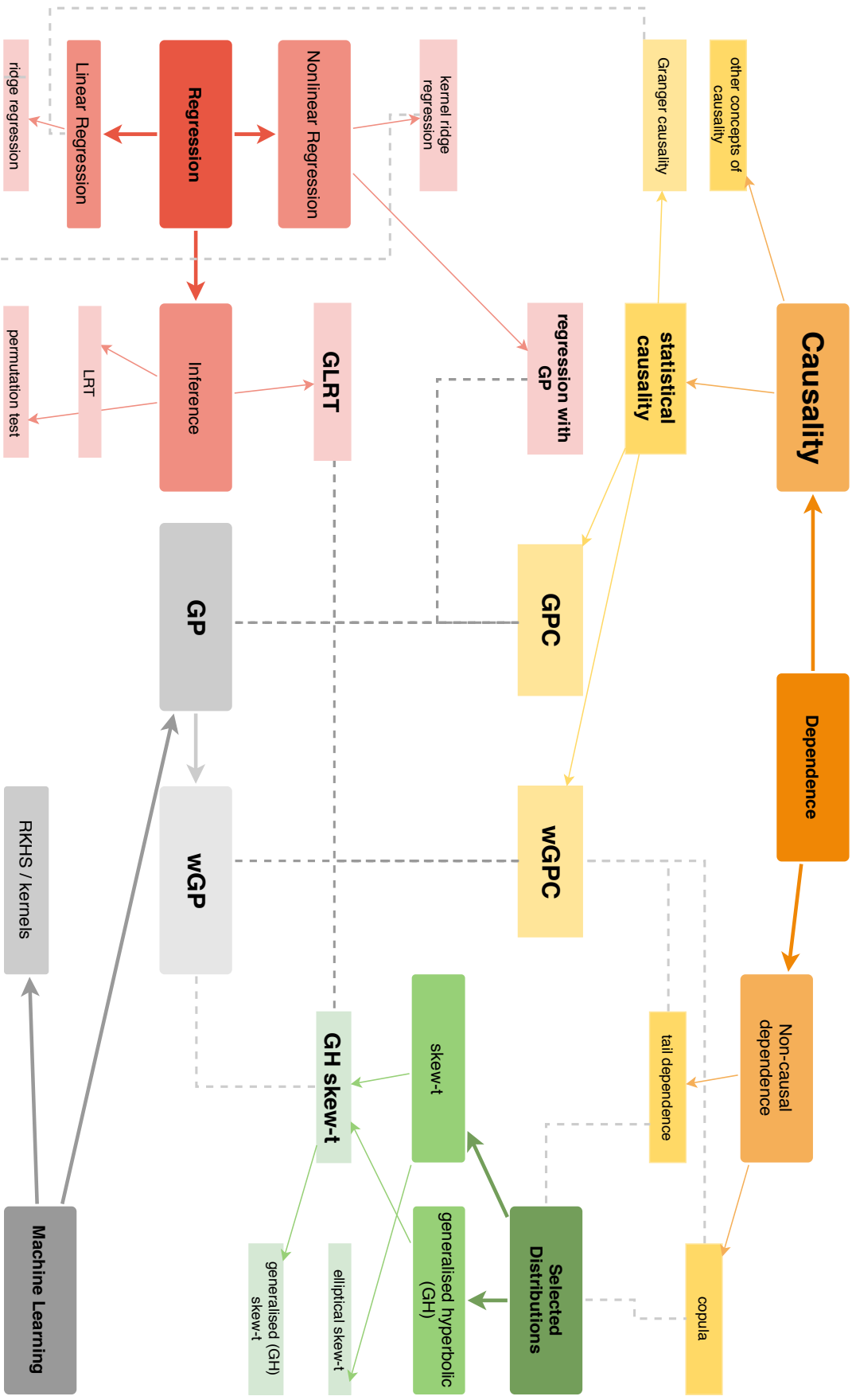


Figure 1.1: Concepts.

to be chosen as part of the model selection.

For the stochastic process $\mathbf{Y}_t$, we can use an optimal linear predictor $P_{Y,t}[X, Y, Z]$ based on the history of all three time series $\{\mathbf{X}_t\}, \{\mathbf{Y}_t\}, \{\mathbf{Z}_t\}$, and we compare it to $P_{Y,t}[Y, Z]$ – an optimal linear predictor based on the history of two time series $\{\mathbf{Y}_t\}, \{\mathbf{Z}_t\}$. The optimal linear predictors are chosen to minimise the prediction error variances:

$$V_Y[X, Y, Z] = \mathbb{E}\left[(\mathbf{Y}_t - P_{Y,t}[X, Y, Z])^2\right] \tag{1.3}$$

$$V_Y[Y, Z] = \mathbb{E}\left[(\mathbf{Y}_t - P_{Y,t}[Y, Z])^2\right] \tag{1.4}$$

Then Granger defines **causality of the process** $\{\mathbf{Y}_t\}$ **by the process** $\{\mathbf{X}_t\}$, within the set of the three time series $\{\mathbf{X}_t\}, \{\mathbf{Y}_t\}, \{Z_t\}$, denoted by $\{\mathbf{X}_t\} \Rightarrow \{\mathbf{Y}_t\}$, if $V_Y[Y, Z] - V_Y[X, Y, Z] > 0$. There is no causality in Granger sense, if $V_Y[Y, Z] - V_Y[X, Y, Z] = 0$, which is denoted by $\{\mathbf{X}_t\} \nRightarrow \{\mathbf{Y}_t\}$. Testing is performed for failing to reject or not the null hypothesis of lack of causality:

$$H_0: \quad V_Y[Y, Z] - V_Y[X, Y, Z] = 0, \quad \text{no causality of the process } \{\mathbf{Y}_t\} \text{ by the process } \{\mathbf{X}_t\} \tag{1.5}$$

$$H_1: \quad V_Y[Y, Z] - V_Y[X, Y, Z] > 0, \quad \text{causality of the process } \{\mathbf{Y}_t\} \text{ by the process } \{\mathbf{X}_t\} \tag{1.6}$$

The test statistic is based on **strength of causality** $\{\mathbf{X}_t\} \Rightarrow \{\mathbf{Y}_t\}$, denoted $C(X, Y)$:

$$C(X, Y) = 1 - \frac{V_Y[X, Y, Z]}{V_Y[Y, Z]}, \qquad 0 \leq C(X, Y) \leq 1. \tag{1.7}$$

In a vector autoregressive model, the value of the strength of causality is meaningful, because if the null hypothesis of no causality is true, then the strength of causality equals zero and high value of strength of causality will often be interpreted as meaning that the actual causal effect is "strong". While the latter is not a precise statement, high value of the estimate of $C(X, Y)$ is consistent with rejecting the null hypothesis with high confidence. However, a more popular test statistic is the following, called **measure of linear feedback** by Geweke [1982]:

$$L^{GC}_{X\to Y} = \log\left[\frac{V_Y[Y, Z]}{V_Y[X, Y, Z]}\right]. \tag{1.8}$$

If all of the processes are Gaussian, then Granger proposes using the test statistic $L^{GC}_{X\to Y}$, which is asymptotically chi-square distributed, [Granger, 1963, Whittle, 1953]. The estimator $\hat{L}^{GC}_{X\to Y}$ of the test statistic $L^{GC}_{X\to Y}$ is based on a approximations (finite dimensional and finite lag and finite sample) of prediction error variances ($\hat{V}(\cdot)$ is used for finite approximation), with N - sample size, p - maximum

number of lags, d - the dimensionality of the data:

$$\hat{L}^{GC}_{X\to Y} = (N - d - p) \log \left[\frac{\hat{V}_Y [Y, Z; p]}{\hat{V}_Y [X, Y, Z; p]} \right] \sim \chi^2_p \qquad \text{if } \{\mathbf{X}_t\} \nRightarrow \{\mathbf{Y}_t\}. \tag{1.9}$$

For completeness, before proceeding with, mostly, chronological description of the evolution of Granger's concept, we would like to point out two tests similar to the one from Equation (B.5). Those tests instead of using approximations of prediction error variances $\hat{V}(\cdot)$, use sums of squared residuals RSS. Let us denote the sum of squared residuals for the prediction $P_{Y,t}[X, Y, Z]$ as RSS_{XYZ}, and the sum of squared residuals for the prediction $P_{Y,t}[Y, Z]$ as RSS_{YZ}. The following two test statistics can be defined, [Hamilton, 1994, Hlaváčková-Schindler et al., 2007]:

$$L^{GC,2}_{X\to Y} = \frac{(RSS_{YZ} - RSS_{XYZ})/p}{RSS_{XYZ}/(N - 2p - 1)}, \qquad \hat{L}^{GC,2}_{X\to Y} \sim F_{p,T-2p-1} \text{ if } \{\mathbf{X}_t\} \nRightarrow \{\mathbf{Y}_t\} \tag{1.10}$$

$$L^{GC,3}_{X\to Y} = \frac{N(RSS_{YZ} - RSS_{XYZ})}{RSS_{XYZ}}, \qquad \hat{L}^{GC,3}_{X\to Y} \sim F_{p,T-2p-1} \text{ if } \{\mathbf{X}_t\} \nRightarrow \{\mathbf{Y}_t\}. \tag{1.11}$$

In later years, Granger's definition of causality became known as **Granger causality**, or less often Wiener-Granger causality.

Granger also discussed analysing causality in the frequency domain – this approach made Granger causality particularly useful in the field of neuroscience[3]. The frequency approach is applicable in stationary setting, and does not change the information content or explanatory power. We direct the reader to other works of Granger [1969, 1980], for spectral methods in causality, and also for discussion about strengths and shortcomings of Granger's method. The evolution of Granger's approach can be seen in the publication about causality between stock prices and currency exchange rates, Granger et al. [2000].

Of particular importance for the Granger causality in frequency domain are the works Geweke [1982, 1984b], who gave technical conditions for uniform boundedness of cross-power spectral density and guarantees for square summability of the regression coefficient. The focus of the thesis is the time domain, but the approach is applicable to the frequency domain, if one wanted to extend it.

In the following years, the concept of causality proposed by Granger became a very popular tool for economic analysis, see Sims [1972], Hamilton [1983], Thornton and Batten [1985], Joerding [1986], Lee [1992], Hiemstra and Jones [1994]. In 1972, Christopher Sims attempted to verify a hypothesis of key importance to the field of macroeconomics: Milton Friedman and Anna Schwartz's assertion that money (monetary disturbances) was a key factor affecting output, and that this relationship was causal. Sims has found that:

"(...) *the hypothesis that causality is unidirectional from money to income agrees with the postwar*

[3]Neuroscience has been one of the bigger fields of applications of methods of statistical causality. For a review of literature on application of Granger causality in neuroscience, please see: [Seth et al., 2015] and references therein, as well as [Bressler and Seth, 2011], Porta and Faes [2015], [Seth, 2010] and Barnett and Seth [2014]

U.S. data, whereas the hypothesis that causality is unidirectional from income to money is rejected. It follows that the practice of making causal interpretations of distributed lag regressions of income on money is not invalidated (on the basis of this evidence) by the existence of "feedback" from income to money." [Sims, 1972]

In his later publication, Sims analysed the relationship between money and income in a wider context, including interest rate data, [Sims, 1980a,b], and found out that the causal effect from income to money has disappeared. Typically, the interpretation is that side information (interest rate data) is a common factor, whose inclusion in the analysis implies that a variable (money) is an indirect rather than a direct cause of the second variable (income). In particular, Sims found that an innovation in the nominal rate of interest leads to a decline in output; King and Watson (1996) later referred to this as the "inverted leading indicator phenomenon". This work led Sims to retract his view on the Friedman-Schwartz hypothesis. As an "interesting working hypothesis", he adopted the idea that monetary policy actually has little to do with output fluctuations. Instead, he conjectured that the inverted leading indicator phenomenon reflects the operation of real shocks, with monetary disturbances playing only a minor role in fluctuations. The findings of a widely read publication (Litterman and Weiss, 1985) seem to provide support for Sims' hypothesis.

The aforementioned paper, Sims [1972], was the first substantial application to economic analysis and it has stimulated many interesting discussions in the field, both about the methodology and the applications. The statistical test that Sims has suggested a test that takes in consideration future values of the time series. The test has not been written with a precise mathematical notation:

We can always estimate a regression of Y on current and past X. But only in the special case where causality runs from X to Y can we expect that no future values of X would enter the regression if we allowed them. Hence, we have a practical statistical test for unidirectional causality: Regress Y on past and future values of X, taking account by generalised least squares or prefiltering of the serial correlation in w(t). Then if causality runs from X to Y only, future values of X in the regression should have coefficients insignificantly different from zero, as a group. [Sims, 1972]

Literature provided different formulation for testing Sim's causality, for example: Jacobs et al. [1979], Florens and Mouchart [1982], Eichler [2001], Chicharro [2014]. Sims has considered causality in the context of linear regression, and this is also the approach that Jacobs et al. [1979] has taken, while the other three sources consider more general nonparametric expression - in line with what we will later see as evolution of formulations of Granger causality concept. In bivariate case, Granger causality and Sims causality are equivalent, [Jacobs et al., 1979, Florens and Mouchart, 1982, Chicharro, 2014].

Jacobs et al. [1979] discuss the usefulness and difficulties in interpretation of Granger's and Sim's work. We will concentrate on the former, as Jacobs proves they are equivalent. Their critique addresses disparity between testable and intuitive definitions of causality. Furthermore, Jacobs et al. [1979] argue

that any specification error results in the causality test losing interpretability. Their arguments are illustrated on the following structural model:

$$X_t = a_{X,0}Y_t + a_{X,1}X_{t-1} + a_{X,2}Y_{t-1} + \epsilon_{Xt} \tag{1.12}$$

$$Y_t = a_{Y,0}X_t + a_{Y,1}X_{t-1} + a_{Y,2}Y_{t-1} + \epsilon_{Yt}. \tag{1.13}$$

We assume that $\epsilon_{Xt}, \epsilon_{Yt}$ are independent and serially uncorrelated random variables, with zero means, and variances $\sigma^2_{Xt}, \sigma^2_{Yt}$, respectively. The reduced form of the equations above is:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \mathbf{B} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \boldsymbol{\varepsilon}_t, \tag{1.14}$$

where, under assumption of $a_{X,0}a_{Y,0} \neq 1$, the matrix $\mathbf{B}$ can be expressed as:

$$\mathbf{B} = (1 - a_{X,0}a_{Y,0})^{-1} \begin{bmatrix} a_{X,1} + a_{X,0}a_{Y,1} & a_{X,2} + a_{X,0}a_{Y,2} \\ a_{Y,0}a_{X,1} + a_{Y,1} & a_{Y,0}a_{X,2} + a_{Y,2} \end{bmatrix} \tag{1.15}$$

and the new error term is:

$$\boldsymbol{\varepsilon}_t = (1 - a_{X,0}a_{Y,0})^{-1} \begin{bmatrix} 1 & a_{X,0} \\ a_{Y,0} & 1 \end{bmatrix} \begin{pmatrix} \epsilon_{X,t} \\ \epsilon_{Y,t} \end{pmatrix}. \tag{1.16}$$

The extent to which the time series $\{X_t\}$ affects $\{Y_t\}$ is assessed with one of the three hypotheses:

H_1:	$a_{Y,1} = a_{Y,0} = 0.$	Disturbance in the equation of X_t is not transferred into Y_t. [Jacobs et al., 1979] describe it as "X_t does not cause Y_t".
H_2:	$a_{Y,0} = 0.$	Current disturbance in the equation of X_t is not transferred into Y_t, also called lack of "instantaneous causality", or "Y_t is not contemporaneously exogenous".
H_3:	$a_{Y,1} + a_{Y,0}a_{X,1} = 0.$	Optimal linear prediction of Y_t does not depend on X_t, lack of Granger causality. Jacobs et al. [1979] describe it as "X_t is not informative about Y_t".

The hypothesis H_3 – Granger causality – becomes the same as H_1 if there is no instantaneous dependence.

Jacobs et al. [1979] point out, that due to identification issue, only the parameters of the reduced form – the four coefficients of the matrix $\mathbf{B}$ from Equation (1.14) – can be estimated, but not the six parameters of the structural model from Equations (1.12 - 1.13). In consequence, only the hypothesis H_3

can be estimated. We note that H_3 being false implies that H_1 is false, but H_3 being true does not imply that H_1 is true. This means that if the test statistic for H_3 is small, the data will support the hypothesis that X_t is not informative about the future of Y_t (X_t does not cause Y_t in the sense of Granger), but can have infinitesimal effect on Y_t. It is possible therefore to gather evidence against the hypothesis of lack of causality, but not in favour of it, and at the same time the evidence against this hypothesis can be attributed to a slight misspecification.

The argument of Jacobs et al. [1979] describes one of the biggest shortcomings of the Granger's method. While their argument has been demonstrated on a model with instantaneous effect, we would like to note that the same problem with identification will take place if any two, or more, lags were taken into consideration.

The relationship between Granger's and Sims' concepts of non-causality are also the subject of Florens and Mouchart [1982] and Florens and Mouchart [1985] who give conditions for their equivalence. What is most interesting from our point of view, is that they formulate non-causality in terms of conditional independence of σ-fields, or, heuristically, in terms of so-called "information sets". A continuation of their work has led to publication by Florens and Fougere [1996] who propose defining non-causality in continuous time, which we present below. Process Y_t is indexed by $t \in I \subset \mathbb{R}^+$. Y_t is a real valued measurable function defined on a probability space $(\Omega, \mathcal{A}, P)$. Three sets of filtrations are defined: minimal filtration $\mathcal{H}_t$ generated by the process Y_t, $\mathcal{G}_t$ to which Y_t is adapted but representing reduced information, and the full information $\mathcal{F}_t$ to which Y_t is also adapted, meaning:

$$\mathcal{H}_t \subset \mathcal{G}_t \subset \mathcal{F}_t, \ t \in I. \tag{1.17}$$

The notation $\mathcal{H}_t \subset \mathcal{G}_t \subset \mathcal{F}_t$ resembles the notation introduced for the Equation (1.1), although in the latter the filtrations $\mathcal{F}_t^X, \mathcal{F}_t^Y, \mathcal{F}_t^Z$ are explicitly introduced as generated by the processes X_t, Y_t, Z_t, while the filtrations $\mathcal{H}_t, \mathcal{G}_t, \mathcal{F}_t$ do not refer explicitly o stochastic processes, although they could be interpreted as associated with stochastic processes Y_t, (Y_t, Z_t) and (Y_t, X_t, Z_t).

Florens and Fougere [1996] defines **weak global non-causality** and **strong global non-causality** through equality of expectations and conditional independence:

$$H_1: \quad \mathbb{E}(Y_t \mid \mathcal{F}_s) = \mathbb{E}(Y_t \mid \mathcal{G}_s), \ \forall_{s,t} \in I, \qquad (\mathcal{F}_t) \text{ does not weakly globally cause } Y_t, \text{ given } (\mathcal{G}_t) \tag{1.18}$$

$$H_2: \quad Y_t \perp\!\!\!\perp \mathcal{F}_s \mid \mathcal{G}_s, \ \forall_{s,t} \in I, \qquad (\mathcal{F}_t) \text{ does not strongly globally cause } Y_t, \text{ given } (\mathcal{G}_t) \tag{1.19}$$

In our research we are interested in strong non-causality. Although we use discrete time in our definitions, continuous time can be straightforwardly introduced.

In 1994, a well-known tests for nonlinear causality has been proposed, based on nonparametric estimators of temporal relations within and across time series, Hiemstra and Jones [1994]. For the bivariate

model $\{X_t\}, \{Y_t\}$, the hypothesis of lack of causality was written as equality of conditional distributions:

$$H_0 : \qquad \pi\left(Y_t \mid \mathbf{X}_{t-p_x:t-1}, \mathbf{Y}_{t-p_y:t-1}\right) = \pi\left(Y_t \mid \mathbf{Y}_{t-p_y:t-1}\right), \tag{1.20}$$

where p_x denotes a lag for the time series $\{X_t\}$, p_y denotes a lag for the time series $\{Y_t\}$. The equality of the two conditional densities from the Equation (1.20) can be equivalently expressed, using Bayes law, as:

$$H_0^2 : \qquad \frac{\pi\left(Y_t, \mathbf{X}_{t-p_x:t-1}\mathbf{Y}_{t-p_y:t-1}\right)}{\pi\left(\mathbf{Y}_{t-p_y:t-1}\right)} = \frac{\pi\left(Y_t, \mathbf{Y}_{t-p_y:t-1}\right)}{\pi\left(\mathbf{Y}_{t-p_y:t-1}\right)} \frac{\pi\left(\mathbf{X}_{t-p_x:t-1}, \mathbf{Y}_{t-p_y:t-1}\right)}{\pi\left(\mathbf{Y}_{t-p_y:t-1}\right)}. \tag{1.21}$$

The joint distributions from Equation (1.21) can be expressed in terms of correlation integrals $C_W(\epsilon)$:

$$C_W(\epsilon) = P(\|W_1 - W_2\|_{\max} < \epsilon), \qquad W_1, W_2 \overset{i.i.d.}{\sim} W \tag{1.22}$$

Subsequently, Hiemstra and Jones [1994] argue that the null hypothesis from Equation (1.21) implies:

$$H_0^3 : \qquad \frac{C_{Y_t,\mathbf{X}_{t-p_x:t-1}\mathbf{Y}_{t-p_y:t-1}}(\epsilon)}{C_{\mathbf{Y}_{t-p_y:t-1}}(\epsilon)} = \frac{C_{Y_t,\mathbf{Y}_{t-p_y:t-1}}(\epsilon)}{C_{\mathbf{Y}_{t-p_y:t-1}}(\epsilon)} \frac{C_{\mathbf{X}_{t-p_x:t-1},\mathbf{Y}_{t-p_y:t-1}}(\epsilon)}{C_{\mathbf{Y}_{t-p_y:t-1}}(\epsilon)}. \tag{1.23}$$

The associated test statistic $F^{HJ}_{X\to Y}$ is defined as a difference between the left-hand and right-hand side ratios in the hypothesis H_0^3, Equation (1.23):

$$L^{HJ}_{X\to Y} = \left(\frac{C_{Y_t,\mathbf{X}_{t-p_x:t-1}\mathbf{Y}_{t-p_y:t-1}}(\epsilon)}{C_{\mathbf{Y}_{t-p_y:t-1}}(\epsilon)} - \frac{C_{Y_t,\mathbf{Y}_{t-p_y:t-1}}(\epsilon)}{C_{\mathbf{Y}_{t-p_y:t-1}}(\epsilon)} \frac{C_{\mathbf{X}_{t-p_x:t-1},\mathbf{Y}_{t-p_y:t-1}}(\epsilon)}{C_{\mathbf{Y}_{t-p_y:t-1}}(\epsilon)} \right). \tag{1.24}$$

Estimator for correlation integrals $C_W(\epsilon)$ are proposed

The test statistic based on the difference between the left and the right side of Equation (1.23) has normal distribution.

The Hiemstra-Jones test for nonlinear causality has gained popularity, however Diks and Panchenko [2005] give a counterexample for when the hypothesis H_0 from Equation (1.20) is true, but the hypothesis H_0^{HJ} from Equation (1.23) is not true.

In 2000, Schreiber [2000] published a paper where information-theoretical measure of **transfer entropy** was defined as a way to measure causality. Transfer entropy became one of the most popular nonlinear methods for statistical causality. It has been designed to measure departure from generalised Markov property, as in Equation (1.20): Transfer entropy can be defined in terms of Shannon entropy [Shannon, 1948], or related measures of mutual information, or conditional entropy – with the latter being the best conceptual representation of the departure from Markov property in Equation (1.20). Let

conditional (Shannon) entropy be defined as:

$$H(X \mid Y) = -\int\int \pi(x,y) \log \pi(x \mid y) dxdy, \tag{1.25}$$

then the transfer entropy is defined as:

$$L^{TE}_{X \to Y} = H(Y \mid Y_{t-k:t-1}) - H(Y \mid X_{t-k:t-1}, Y_{t-k:t-1}). \tag{1.26}$$

Transfer entropy is a generalisation of Granger causality that can be parametric and nonparametric. For variables that are distributed normally, transfer entropy is equivalent to Granger causality [Barnett et al., 2009] , and it can be seen as a likelihood ratio [Barnett and Bossomaier, 2012] , with known asymptotic properties. The popularity of transfer entropy is a result of the fact that it does not require any parametric model assumptions. However, if estimated as a nonparametric model, transfer entropy requires numerical estimation that is computationally expensive and potentially less practical than other methods. In low-dimensional case, a simple and popular solution is to use the histogram of the embedding vectors [Lungarella et al., 2007a]. Algorithms for calculating transfer entropy are typically based on algorithms for estimating entropy, and those include a data-efficient technique based on nearest neighbours estimators, [Lindner et al., 2011, Kaiser and Schreiber, 2002, Schreiber, 2000]. Please refer to [Hlaváčková-Schindler et al., 2007] for an in-depth reference for estimation of transfer entropy. Asymptotic behaviour of nonparametric transfer entropy cannot be easily modelled, which means that it has to be tested with, for example, a bootstrap test, (Chávez et al. [2003]).

Instead of modifying the null hypothesis or choosing a different test statistic for testing the hypothesis, a popular approach to generalising Granger causality is to start with a different predictive model. Ancona et al. [2004] took the linear autoregressive model from Equation (1.1), which we write below as two alternatives for explaining Y_t:

$$\text{Model A:} \qquad Y_t = \mathbf{a}^T_{A,Y} \mathbf{Y}_{t-p:t-1} + \epsilon^A_{Y,t} \tag{1.27}$$

$$\text{Model B:} \qquad Y_t = \mathbf{a}^T_{B,Y} \mathbf{Y}_{t-p:t-1} + \mathbf{b}^T_{B,Y} \mathbf{X}_{t-p:t-1} + \epsilon^B_{Y,t}. \tag{1.28}$$

It is proposed to alter the embedding vectors from Equations (1.27 - 1.28) using n-dimensional real vectors Φ, Ψ, such that $\Phi = (\phi_1, ..., \phi_n), \Psi = (\psi_1, ..., \psi_n)$ and $\phi_i(), \psi_i()$ are nonlinear radial basis functions (RBF). For n centres $\tilde{\mathbf{X}}^i, \tilde{\mathbf{Y}}^i$ chosen using a clustering procedure, the RBFs have the following form:

$$\psi_i(\mathbf{X}) = \exp\left(- \parallel \mathbf{X} - \tilde{\mathbf{X}}^i \parallel^2 / 2\sigma^2\right) \tag{1.29}$$

$$\phi_i(\mathbf{X}) = \exp\left(- \parallel \mathbf{Y} - \tilde{\mathbf{Y}}^i \parallel^2 / 2\sigma^2\right). \tag{1.30}$$

The nonlinear predictive models formulated with RBF are now:

$$\text{RBF Model A:} \qquad Y_t = \boldsymbol{\alpha}_{A,Y}^T \Phi\left(\mathbf{Y}_{t-p:t-1}\right) + \epsilon_{Y,t}^A \tag{1.31}$$

$$\text{RBF Model B:} \qquad Y_t = \boldsymbol{\alpha}_{B,Y}^T \Phi\left(\mathbf{Y}_{t-p:t-1}\right) + \boldsymbol{\beta}_{B,Y}^T \Psi\left(\mathbf{X}_{t-p:t-1}\right) + \epsilon_{Y,t}^B. \tag{1.32}$$

For the null hypothesis of lack of causality as before, the test statistic $L_{X\to Y}^{RBF}$ measures the difference of prediction errors $e_{Y,t}^A, e_{Y,t}^B$:

$$L_{X\to Y}^{RBF} = e_{Y,t}^A - e_{Y,t}^B. \tag{1.33}$$

If $L_{X\to Y}^{RBF} > 0$ then incorporating $\mathbf{X}_t$ improves the prediction of $\mathbf{Y}_t$ which can be interpreted as causal influence from X_t to Y_t. Analogously, one can test for causality in the opposite direction.

The growth of Machine Learning, and popularisation of kernels - which lead to many linear models being generalised to nonlinear by kernelisation, has brought important developments to the studies of statistical causality. One of the earliest methods of kernelisation was by some of the co-authors of the already mentioned Ancona et al. [2004], who extended the radial basis approach by changing the radial basis functions to radial kernels (Marinazzo et al. [2008b,a]). This meant a change from describing the feature maps and feature space explicitly to describing them implicitly using the reproducing kernels (please refer to Chapter 3 for an overview of kernels). To kernelise Granger causality Marinazzo et al. [2008b] used the geometric interpretation of linear least squares regression. The idea was to start from describing the residuals of the two compared models in terms of linear projections onto the spaces spanned by the regressors respective to the two models. Such a description would allow to formulate the metric used to quantify Granger causality in terms of inner product of regressors and hence enable the application of the kernel trick.

The first example of kernel trick in the context of statistical causality that we provide is from the publication by Marinazzo et al. [2008b]. In a bivariate model, we are interested in assessing causality between the time series $\{X_t\} \Rightarrow \{Y_t\}$, by studying the decrease in variance of prediction errors, see Equation (1.6). The test for improving prediction is based on the comparison of the two models from Equations (1.27 - 1.28). Fitting a linear regression means that a vector of responses is represented as a linear combination of the regressors, or as a projection onto the subspace spanned by the regressors. Let $\{Y_t\}, t \in [t_1, ..., t_2]$ be the $[t_2 - t_1] \times 1$ vector of responses we want to model, and matrices $\mathbf{K}^A$ and $\mathbf{K}^B$ be

defined as follows, based on the regressors of the two models, A and B:

$$\mathbf{K}^A = \left[\mathbf{Y}_{t_1-p:t_1-1} \cdots \mathbf{Y}_{t_2-p:t_2-1}\right]^T \left[\mathbf{Y}_{t_1-p:t_1-1} \cdots \mathbf{Y}_{t_2-p:t_2-1}\right] \tag{1.34}$$

$$\mathbf{K}^B = \left[\begin{bmatrix} \mathbf{Y}_{t_1-p:t_1-1} \\ \mathbf{X}_{t_1-p:t_1-1} \end{bmatrix} \cdots \begin{bmatrix} \mathbf{Y}_{t_2-p:t_2-1} \\ \mathbf{X}_{t_2-p:t_2-1} \end{bmatrix}\right]^T \left[\begin{bmatrix} \mathbf{Y}_{t_1-p:t_1-1} \\ \mathbf{X}_{t_1-p:t_1-1} \end{bmatrix} \cdots \begin{bmatrix} \mathbf{Y}_{t_2-p:t_2-1} \\ \mathbf{X}_{t_2-p:t_2-1} \end{bmatrix}\right]. \tag{1.35}$$

Let $H^A \subseteq \mathbb{R}^{t_2-t_1}$ be a p-dimensional space spanned by the matrix $\mathbf{K}^A$, and $H^B \subseteq \mathbb{R}^{t_2-t_1}$ be a $2p$-dimensional space spanned the matrix $\mathbf{K}^B$. Without loss of generality it can be assumed that the response vector $Y_{t_1:t_2}$ has mean zero and norm 1. Denote by P^A, P^B the projectors (projection matrices) that can be used to projecting the response vector onto the subspaces H^A, H^B, and denote by $\hat{\mathbf{Y}}^A_{t_1:t_2}, \hat{\mathbf{Y}}^B_{t_1:t_2}$ the fitted values, and by $u^A_{t_1:t_2}, u^B_{t_1:t_2}$ the prediction errors:

$$\hat{\mathbf{Y}}^A_{t_1:t_2} = P^A \mathbf{Y}_{t_1:t_2}; \qquad u^A_{t_1:t_2} = \|\mathbf{Y}_{t_1:t_2} - \hat{\mathbf{Y}}^A_{t_1:t_2}\|_2^2 = \mathbf{1} - \left(\hat{\mathbf{Y}}^A_{t_1:t_2}\right)^T \left(\hat{\mathbf{Y}}^A_{t_1:t_2}\right) \tag{1.36}$$

$$\hat{\mathbf{Y}}^B_{t_1:t_2} = P^B \mathbf{Y}_{t_1:t_2}; \qquad u^B_{t_1:t_2} = \|\mathbf{Y}_{t_1:t_2} - \hat{\mathbf{Y}}^B_{t_1:t_2}\|_2^2 = \mathbf{1} - \left(\hat{\mathbf{Y}}^B_{t_1:t_2}\right)^T \left(\hat{\mathbf{Y}}^B_{t_1:t_2}\right). \tag{1.37}$$

If $H^\perp = H^B \div H^A$, $P^\perp$ is a projection onto $H^\perp$, then $u^B_{t_1:t_2} = u^A_{t_1:t_2} - \|P^\perp \mathbf{Y}_{t_1:t_2}\|^2$. The strength of causality from Equation (1.7) will here be represented as:

$$C(X, Y) = \frac{u^A_{t_1:t_2} - u^B_{t_1:t_2}}{u^A_{t_1:t_2}} = \frac{\|P^\perp \mathbf{Y}_{t_1:t_2}\|_2^2}{\mathbf{1} - \left(\hat{\mathbf{Y}}^A_{t_1:t_2}\right)^T \left(\hat{\mathbf{Y}}^A_{t_1:t_2}\right)}. \tag{1.38}$$

Let us observe that the subspace $H^\perp$ is spanned by the matrix $\mathbf{K}^\perp = \mathbf{K}^B - P\mathbf{K}^B - \mathbf{K}^B P + P\mathbf{K}^B P$. Let $\mathbf{v}_1, ..., \mathbf{v}_p$ be eigenvectors of $\mathbf{K}^\perp$, and the $C(X, Y)$ can be represented as:

$$C(X, Y) = \sum_i^m \operatorname{corr}\left(\mathbf{v}_i, \mathbf{Y}_{t_1:t_2}\right). \tag{1.39}$$

Marinazzo et al. [2008b] suggest an additional step, where the eigenvalues $\mathbf{v}_i$ are filtered, so that only significantly big eigenvalues are used.

To allow modelling nonlinear causality, the **kernel trick** is used. The kernel trick is a simple and general principle based on the fact that kernels can be thought of as inner products. It can be stated as follows:

"*Any algorithm for vectorial data that can be expressed only in terms of dot products between vectors can be performed implicitly in the feature space associated with any kernel, by replacing each dot product by a kernel evaluation.*" [Schölkopf et al., 2004]

The matrix $\mathbf{K}^A$ now represents a Gramm matrix, such that $K^A_{ij} = k\left(\mathbf{Y}_{i-p:i-1}, \mathbf{Y}_{j-p:j-1}\right)$, for kernel function $k(\cdot, \cdot)$, and the matrix $\mathbf{K}^B$ now represents a Gramm matrix, such that $K^B_{ij} =$

$k\left(\left[\mathbf{Y}_{i-p:i-1}^T, \mathbf{X}_{i-p:i-1}^T\right]^T, \left[\mathbf{Y}_{j-p:j-1}^T, \mathbf{X}_{j-p:j-1}^T\right]^T\right)$. For the kernelised strength of causality, the condition of $H^A \subseteq H^B$ is not guaranteed, and so Marinazzo et al. [2008b] provide a further correction to account for that.

In 2012, Amblard et al. [2012b] proposed a generalisation of Granger causality using kernel ridge regression, a well-established methodology for generalising linear regression and introducing kernels, which has a clear interpretation and good computational properties. Ridge regression is a regularised least squares method, where the regularisation is performed by imposing penalty on the size of the regression coefficients [Friedman et al., 2001]. We use the notation from Equations (1.27 - 1.28), but add a time series $\{Z_t\}$ representing side information:

$$\text{Model A:} \qquad Y_t = \mathbf{a}_A^T \mathbf{Y}_{t-p:t-1} + \mathbf{c}_A^T \mathbf{Z}_{t-p:t-1} + \epsilon_{Y,t}^A \tag{1.40}$$

$$\text{Model B:} \qquad Y_t = \mathbf{a}_B^T \mathbf{Y}_{t-p:t-1} + \mathbf{b}_B^T \mathbf{X}_{t-p:t-1} + \mathbf{c}_B^T \mathbf{Z}_{t-p:t-1} + \epsilon_Y^B. \tag{1.41}$$

Equations (1.40 - 1.41) can be expressed in matrix notation, with the random vector $\mathbf{Y}_{t_1:t_2} = [Y_{t_1}, ..., Y_{t_2}]$, and matrix $\mathbf{Y}_{t_1,t_2;p} = \left[\mathbf{Y}_{t_1-p:t_1-1}^T, ..., \mathbf{Y}_{t_2-p:t_2-1}^T\right]$

$$\text{Model A:} \qquad \mathbf{Y}_{t_1:t_2} = \mathbf{a}_A^T \mathbf{Y}_{t_1,t_2;p} + \mathbf{c}_A^T \mathbb{Z}_{t_1,t_2;p} + \boldsymbol{\epsilon}_Y^A \tag{1.42}$$

$$\text{Model B:} \qquad \mathbf{Y}_{t_1:t_2} = \mathbf{a}_B^T \mathbf{Y}_{t_1,t_2;p} + \mathbf{b}_B^T \mathbb{X}_{t_1,t_2;p} + \mathbf{c}_B^T \mathbb{Z}_{t_1,t_2;p} + \boldsymbol{\epsilon}_Y^B. \tag{1.43}$$

The mechanism of ridge regression is the same for both models A and B, and regardless of weather the side information is present or not. Thus, we will refer to the optimisation problem using the notation $\mathbf{Q}_t$ for covariates (independent variables), where $\mathbf{Q}_t = [Y_t]$ (model A without side information), or $\mathbf{Q}_t = [X_t, Y_t]$ (model B without side information), or $\mathbf{Q}_t = [Y_t, Z_t]$ (model A with side information), or $\mathbf{Q}_t = [X_t, Y_t, Z_t]$ (model B with side information). Consequently, we will use different notation for the weights $\boldsymbol{\alpha}$.

Ridge regression attempts to find a solution, denoted $\boldsymbol{\alpha}^*$, that minimises the quadratic cost plus weighted sum of the squared coefficients:

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left\{ \sum_{t=t_1}^{t_2} \left(Y_t - \boldsymbol{\alpha}^T \mathbf{Q}_{t-p:t-1} \right)^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} \right\}. \tag{1.44}$$

with λ a chosen constant[4]. The solution – called **primal solution** – to the minimisation problem in Equation (1.44) is:

$$\text{primal solution:} \qquad \boldsymbol{\alpha}^* = \left(\mathbb{Q}_{(t_1,t_2;p)}^T \mathbb{Q}_{(t_1,t_2;p)} + \lambda \mathbf{1}_{t_2-t_1} \right)^{-1} \mathbb{Q}_{(t_1,t_2;p)}^T \mathbf{Y}_{t_1:t_2}. \tag{1.45}$$

[4]While for the purpose of this section we treat λ as a constant, the choice of the parameter λ itself can be interpreted as an actual parameter estimation or model selection, in Bayesian framework expressing prior belief about the behaviour of the function we seek, eg. smoothness. Furthermore, the regularisation can be applied separately to each of the parameters. [Friedman et al., 2001].

In the Appendix, Section (A.1) we show that the primal solution can be expressed in terms $\boldsymbol{\beta}^*$, which we will call dual weights, such that, $\boldsymbol{\alpha}^* = \mathbb{Q}^T_{(t_1,t_2;p)}\boldsymbol{\beta}^*$. The corresponding **dual solution** can be expressed as follows:

$$\text{dual solution:} \qquad \boldsymbol{\beta}^* = \left(\mathbb{Q}_{(t_1,t_2;p)}\mathbb{Q}^T_{(t_1,t_2;p)} + \lambda \mathbf{1}_{t_2-t_1}\right)^{-1} \mathbf{Y}_{t_1:t_2}. \tag{1.46}$$

The dual solution of ridge regression, from Equation (1.46), is in such a form that admits application of the kernel trick by "substituting" the matrix $\mathbb{Q}_{(t_1,t_2;p)}\mathbb{Q}^T_{(t_1,t_2;p)}$ with a Gramm matrix which we will denote $\mathbf{K}_\mathbb{Q}$ to obtain **kernel ridge regression**. $\mathbf{K}_\mathbb{Q}$ is such that $\{\mathbf{K}_\mathbb{Q}\}_{i,j} = k(\mathbf{Q}_{i-p:i-1}, \mathbf{Q}_{j-p:j-1})$, for a Mercer kernel function (semi-positive definite kernel function) $k(\cdot,\cdot)$. The optimal weights, fitted values and mean square of prediction error will for kernel ridge regression be as follows:

$$\text{krr optimal weights:} \qquad \boldsymbol{\beta}^{krr} = \left(\mathbf{K}_\mathbb{Q} + \lambda \mathbf{1}_{t_2-t_1}\right)^{-1} \mathbf{Y}_{t_1:t_2} \tag{1.47}$$

$$\text{krr fitted values:} \qquad \hat{\mathbf{Y}}_{t_1:t_2} = \mathbf{K}_\mathbb{Q}\boldsymbol{\beta}^{krr} \tag{1.48}$$

$$\text{krr MSE:} \qquad V\left(\hat{\mathbf{Y}}_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right) = \frac{1}{t_2 - t_1}\left(\mathbf{K}_\mathbb{Q}\boldsymbol{\beta}^{krr} - \mathbf{Y}_{t_1:t_2}\right)^T\left(\mathbf{K}_\mathbb{Q}\boldsymbol{\beta}^{krr} - \mathbf{Y}_{t_1:t_2}\right). \tag{1.49}$$

When kernel ridge regression is applied to model A, or model B, all of the steps above are applied, but with different definition of Q_t, and therefore different values of the covariance matrix $\mathbf{K}_\mathbb{Q}$. Denoting the fitted values as $\hat{\mathbf{Y}}^A_{t_1:t_2}$ and $\hat{\mathbf{Y}}^B_{t_1:t_2}$, we obtain the means square errors of kernel ridge regression prediction of the two models: $V\left(\hat{\mathbf{Y}}^A_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)$ and $V\left(\hat{\mathbf{Y}}^B_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)$, which are used in the test statistic in a similar manner to the strength of causality from Equation (1.7), and to the test statistic from Equation (1.8). Thus the test statistic based on the kernelised ridge regression, that [Amblard et al., 2012b] proposed is formulated as follows:

$$L^{krr}_{X\to Y} = \log \frac{V\left(\hat{\mathbf{Y}}^A_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)}{V\left(\hat{\mathbf{Y}}^B_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)}. \tag{1.50}$$

The hypotheses are:

$$H_0: \qquad L^{krr}_{X\to Y} = 0, \qquad \text{no causality from } \{X\} \text{ to } \{Y\} \tag{1.51}$$

$$H_1: \qquad L^{krr}_{X\to Y} > 0, \qquad \text{causality from } \{X\} \text{ to } \{Y\} \tag{1.52}$$

There is no explicitly known distribution for the test statistic, so such distribution has to be obtained numerically, using a permutation test. Let m be the number of permutations used and for $i \in \{1, ..., m\}$, let $\zeta_i(t)$, $t \in \{t_1, ..., t_2\}$ denote a random permutation of the time index, and $\zeta_i(X)$ denote a time series, where the original time ordered has been reorganised according to the permutation $\zeta_i(\cdot)$. Then the null hypothesis

is assessed by comparing the value of $L^{krr}_{X\to Y}$ to a histogram of values of $L^{krr}_{\pi_i(X)\to Y}$, and a p-value:

$$p(L^{krr}_{X\to Y} \mid H_0) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left(L^{krr}_{X\to Y} > L^{krr}_{X\to Y}\right) \tag{1.53}$$

where $\mathbf{1}(M)$ is a characteristic function for the set M. Amblard et al. [2012b] propose that the parameters of the kernel – the hyperparameters of the model – are chosen using cross-validation.

Another approach to kernelising Granger causality has been proposed by Sun [2008]. Taking Granger causality in the context of autoregressive time series as a starting point, Sun [2008] develops a nonlinear extension based on methods rooted in computer science and functional analysis. Sun proposed to replace the application of the kernel trick to the Granger causality metrics with a formulation of a kernel analogue of these metrics. As a starting point, we will again use the autoregressive model from Equations (1.27 - 1.28). Sun [2008], like Ancona et al. [2004], proposed reframing the model setting in terms of feature maps, but went further by moving from input spaces to feature spaces.

Let us assume here that $X_t \in \mathcal{X}, Y_t \in \mathcal{Y}$, that $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}$ are reproducing kernel Hilbert spaces [5] (RKHS) of functions on $\mathcal{X}$ and $\mathcal{Y}$ respectively, and that the nonlinear maps $\Phi_A, \Phi_B, \Psi_A : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$, $\Psi_B : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}\otimes\mathcal{X}}$. The two models A and B are described as follows:

$$\text{Model A:} \qquad \boldsymbol{\alpha}^T_{A,Y}\,\Phi_A\left(\mathbf{Y}_{t-p+1:t}\right) = \boldsymbol{\beta}^T_{A,Y}\,\Psi_A\left(\mathbf{Y}_{t-p:t-1}\right) + \epsilon^A_{Y,t} \tag{1.54}$$

$$\text{Model B:} \qquad \boldsymbol{\alpha}^T_{B,Y}\,\Phi_B\left(\mathbf{Y}_{t-p+1:t}\right) = \boldsymbol{\beta}^T_{B,Y}\,\Psi_B\left(\mathbf{Y}_{t-p:t-1}, \mathbf{X}_{t-p+1:t}\right) + \epsilon^B_{Y,t}. \tag{1.55}$$

Just as in the original Granger formulation, or in the case of kernel ridge regression, for the evaluation of the null hypothesis of lack of causality, the Models A and B are compared in terms of variance of prediction errors. Calculating those variances could involve explicit computation of the potentially infinitely-dimensional mappings Φ, Ψ. Instead, Sun's approach defines covariance operators which require only kernel evaluations. We refer the reader to Section (3.1.1) for more detailed definitions of covariance operators, and relevant background from functional analysis and machine learning. Cross-covariance operator $\Sigma_{XY} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ can be defined as a linear operator, which is analogous to covariance matrix, but is defined for feature maps:

$$\forall f \in \mathcal{H}_\mathcal{X}, gf \in \mathcal{H}_\mathcal{Y} \qquad\qquad \langle f, \Sigma_{XY}\, g\rangle_{\mathcal{H}_\mathcal{X}} = Cov(f(X), g(Y)) \tag{1.56}$$

If $X = Y$ then the cross-covariance operator can be called a covariance operator Σ_{XX}. Analogously to the conditional covariance matrix, and assuming that Σ^{-1}_{XX} exists , we can define conditional covariance

[5] Please refer to Chapter (3) for introduction to kernels, reproducing kernel Hilbert spaces (RKHS) and explanation how the feature maps are related to kernels.

operator $\Sigma_{YY|X}$:

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \tag{1.57}$$

The null hypothesis of no causality becomes a null hypothesis of equality of the trace norm of Σ_{YY} and trace norm of $\Sigma_{YY|X}$. This hypothesis is tested with a permutation test. Furthermore, Sun suggested recognition of causality using a permutation test and comparing the results for both directions. This is another example of using the kernel trick to transform a linear algorithm into a nonlinear one by using the nonlinear algorithm in a feature space.

Fukumizu et al. [2008] describes a closely related method (Sun was a co-author), originally based on the methods for testing independence [Gretton et al., 2008, Chwialkowski and Gretton, 2014]. In this method a normalised version of the cross-covariance operator from Equation (1.56) and conditional cross-covariance operator from Equation (1.57) are proposed for testing independence and conditional independence, and are said to be useful also for testing causality. The normalisation of the cross-covariance operator Σ_{YX} from Equation (1.56), which we will represent by V_{YX}, can be seen as analogous to the decomposition of cross-covariance to covariance of the marginals and correlation:

$$\Sigma_{YX} = \Sigma_{YY}^{\frac{1}{2}} V_{YX} \Sigma_{XX}^{\frac{1}{2}}. \tag{1.58}$$

Following the same pattern, we can also start with a conditional cross-covariance operator $\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$ and define a normalised cross-covariance operator, denoted as $V_{YX|Z}$:

$$\Sigma_{YX|Z} = \Sigma_{YY}^{\frac{1}{2}} V_{YX|Z} \Sigma_{XX}^{\frac{1}{2}}. \tag{1.59}$$

Fukumizu et al. [2008] then prove than under certain conditions on the properties of the kernels used (integrability, no degeneration), the normalised cross-covariance operator V_{YX} and normalised cross-covariance operators $V_{YX|Z}$ provide information about the independence of the variables:

$$V_{YX} = 0 \qquad \Longleftrightarrow \qquad Y \perp\!\!\!\perp X \tag{1.60}$$

$$V_{(Y,Z)(X,Z)|Z} = 0 \qquad \Longleftrightarrow \qquad Y \perp\!\!\!\perp X|Z. \tag{1.61}$$

The normalised version of the operator has the advantage that it is less influenced by the marginals than the non-normalised operator, while retaining all the information about dependence. Interestingly, Fukumizu et al. [2008] mention causality "both with and without time information" - which clearly indicates that their understanding of causality is different than the statistical causality, which requires the time ordering. However, the normalised conditional cross-covariance operator with time ordering can be used for testing

for statistical causality [Seth and Principe, 2011]:

$$L_{X\to Y}^{HSNCIC} = \|V_{(Y,Z)(X,Z)|Z}\|_{HS}^2 \tag{1.62}$$

The Hilbert-Schmidt norm (Section (3.1.1)) is used for the test statistic, and after [Seth and Principe, 2011], we will use the name Hilbert-Schmidt Normalised Conditional Independence Criterion (HSNCIC) for this test statistic. Just as in the case of the framework proposed by Sun [2008], testing is done with the permutation test.

In this context, it is Seth and Principe [2011] who focused on using methods of conditional independence, in particular HSNCIC in the context of causality. The relationship between conditional independence and causality has not been explicitly referred to very often, but this is one that lies at the centre of understanding the relationship between statistical causality and other philosophical concepts of causality. We refer to Florens and Mouchart [1982], Florens and Fougere [1996] for one of the first references of conditional independence methods for statistical causality. The second other big area where conditional independence is used for studying statistical causality is literature on copulas, that will be mentioned later.

The growth in prominence of machine learning, and one of its methods – graphical models – in particular, coincided with development of an alternative concept of causality. In 2000, Judea Pearl introduced to a wider audience the idea of using structural equations for describing causal mechanisms, which he called Structural Causal Model (SCM) Pearl [2000, 1993], Simon [1977], and later – a general theory of causation Pearl [2010]. Pearl's way of looking at causality is very different from statistical causality, and these differences are studied in Chapter (2), together with an overview of a wider range of causal concepts. At the time when Granger's method and its extensions were well established tools, Pearl presented new arguments in the discussion about identifiability of causal relationships. It wasn't only the case, according to Pearl, that the existing methods were unsatisfactory, but the field of statistics did not have language to deal with causal concepts. Pearl has proposed to distinguish between associations and causal relations, with observation and joint distribution being enough to establish the former, but not the latter.

Pearl [2000] defines causal model $M = \langle U, V, E\rangle$ in terms of endogenous variables U, exogenous variables V, and a set F of functions that describe mappings from the exogenous variables to endogenous variables. A probabilistic causal model is then described as a pair $\langle M, \mathbb{P}(u)\rangle$, of causal model M and the probability function defined over the set of exogenous variables U. The path diagram carries the information about potential causal relationships via edges and paths, and missing edges mean lack of correlation or direct effect. Graphical methods, in particular "d-connection" and "d-separation" can then be used to verify which conditional independence relationships will hold, based on the observed data.

The effect of intervention, which can be analysed with the use of "do operator" can also be studied using the path diagram. These tools allow to establish whether a causal relationship is identifiable or not.

For those relationship which are not identifiable on the basis of observational data cannot, the path diagram aids in designing which experiments are neccessary to achieve that identifiability.

Causality in the sense of Pearl does not require temporal ordering, but assumes ability to perform both observation and intervention (hence the name "intervention-based causality" is sometimes used). This is a very different setting than the one for which our statistical causality model is primarily designed – one set of historical data (time series), and no possibility to design the experiment. Pearl gives tools for analysing where a causal relationship is identifiable, and where not. If not, Pearl says additional experiments need to be designed and performed.

Statistical causality and causal inference do have common ground, with Granger causal graphs being the most popular link between the two Dahlhaus and Eichler [2003], Eichler [2012, 2007], Eichler and Didelez [2010], Fiedor [2014].

Eichler [2012] proposed building so called Granger causality graphs to represent Granger causal relationships. Granger causality graphs are mixed graphs, with time series represented as nodes, directed edges representing Granger causality, and undirected edges – contemporaneous dependence. A direct edge from X to Y will be drawn if an only if X Granger causes Y (the hypothesis of non-causality is rejected). The contemporaneous dependence is based on rejection of the hypothesis of conditional (contemporaneous) independence between X and Y. The Granger causal graphs introduced by Eichler share some ideas with Pearl's Structural Causal Models, in that it explores the use of tools from graphical model and structural equations for analysing causal relations. But while having some similarities, these two approaches are based on different causal concepts: Granger (statistical) causality versus "intervention based causality". Eichler [2012] builds graph with edges reflecting pairwise causal relations for one observation of multivariate time series, with explicit time ordering. Causality in the Structural Causal model is described in terms of a whole system We refer the reader to the Chapter (2) for more discussion of different concepts of causality.

Please note, that Eichler later shows that in the context of Granger causality and Granger causal graphs, one can still talk about interventions and identifiability. Using similar tools of graphical models as Pearl, Eichler show how can one start from statistical causality in the sense of Granger - that suggest a potential causal relationship, to proving whether a relationship can be identified as causal or not.

A very interesting link has been proposed by Billio et al. [2012], who build complex networks based on linear Granger causality and analyse evolution of these network using methods from complex network theory. Like Eichler [2012], they use nodes to represent time series, and directed edges to represent Granger causality. They build the network for data on monthly returns of hedge funds, banks, broker/dealers and insurance companies, and use measures of connectedness to describe changes in

systemic risk. They conclude that there is asymmetry in the degree of connectedness between the four sectors, with banks playing the biggest role in transmitting shocks.

One of the most important aspects of our work is the development of testing procedures with tests with known asymptotic properties and good power of the test (please refer to Chapter (4) for the definitions). The classical Granger causality allows the use of test statistics that are distributed according to F distribution or χ^2 distribution under the null hypothesis of no causality [Wald, 1943, Granger, 1963, 1969, Hlaváčková-Schindler et al., 2007]. The nonparametric Granger causality method by Hiemstra and Jones, based on correlation integrals, has been shown to be biased, but asymptotically normally distributed [Hiemstra and Jones, 1994]. Many of the methods for statistical causality do not allow formulation of test statistics with known asymptotic properties. Geweke proposed using bootstrap tests for his method of measuring causality in the frequency domain Geweke [1984a]. Bootstrap tests, or permutation tests [Pesarin, 2001] are typically used for testing hypotheses when distributions of the statistical tests are not known. For example, in the case of transfer entropy, whether the distribution of the test statistic is easily obtainable depends on additional assumptions. For Gaussian data, transfer entropy is equivalent to Granger causality [Barnett et al., 2009], and in broad parametric context the null hypothesis can be tested with log-likelihood ratio. However, in the nonparametric context, using transfer entropy requires permutation tests [Lindner et al., 2011, Gómez-Herrero et al., 2015, Zaremba and Aste, 2014]. As mentioned before, permutation tests have been used in the methods involving kernels Amblard et al. [2012b] Seth and Principe [2011] Sun [2008]. Amblard et al. [2012b], who propose using marginal likelihood of a GP model to test for causality, do not go far enough to formulate an actual likelihood ratio test Amblard et al. [2012a], but in their later publication employing kernel ridge regression, they propose permutation test Amblard et al. [2012b]. Our proposal relies on the generalised likelihood ratio test (GLRT) and necessitates nested model formulation MacKinnon [1983], Vuong [1989], Pesaran and Weeks [2001], which is not very restrictive, please refer to Chapter (4) for detailed definitions.

The GLRT is a composite hypothesis test that assesses whether a set of parameter θ belong to a particular set:

$$H_0 : \theta \in \omega \qquad \text{vs} \qquad H_1 : \theta \in \Omega - \omega \tag{1.63}$$

GLRT gives an asymptotic distribution of the test statistics, but it requires that the hypotheses are nested. For a random sample $X_1, X_2, ..., X_N$ from a distribution with pdf $\pi(x; \theta)$ and likelihood $L(\theta; x) = \pi(x; \theta)$ define:

$$\lambda = \left\{ \max_{\theta \in \omega} L(\theta; x) / \max_{\theta \in \Omega} L(\theta; x) \right\}. \tag{1.64}$$

For some constant A, we can use a test with critical region $\lambda \leq A$. If we define d as the difference in dimensionality of H_0 and $H_0 \cup H_1$, then we have that under the null the asymptotic distribution of the test statistic is distributed according to:

$$-2 \log \lambda \sim \chi^2_d. \tag{1.65}$$

Considering the general statistical test for non-causality seen as the lack of equality of conditional distributions, we observe that when assuming specific models, these tests can become equivalent to testing for particular properties or parameters. Employment of LRT and GLRT relies on the fact that a parametric model is used and that parameters are known (LRT) or estimated (GLRT) [Garthwaite et al., 2002]. When using GLRT, one has to ensure that the hypotheses are nested, which typically will be expressed in the form of restrictions on the parameter (or hyperparameter) space. In the case of our framework of Gaussian Processes for Causality (GPC) or warped Gaussian Processes for Causality (wGPC), the nesting of hypotheses is introduced through kernels with the property of automatic relevance determination (ARD). The ARD concept is typically used in the context of feature selection [MacKay, 1996, Neal, 2012] and sparse learning in Bayesian models Qi et al. [2004]. For resources about the alternatives to GLRT that allow for non-nested hypotheses, we suggest MacKinnon [1983], Vuong [1989], Pesaran [1990].

Having a test statistic with known distribution was essential when developing the generalised framework for statistical causality. As a consequence, there are several classes of distributions that are of highest importance for our method. Generalised hyperbolic distributions were introduced by Barndorff-Nielsen in 1977 [Barndorff-Nielsen, 1977]. We are mostly interested in a special case of generalised hyperbolic family: the skew-t distribution and the skew-t copula [Demarta and McNeil, 2005, Rachev, 2003, McNeil et al., 2015]. Cruz et al. [2015] describe the use of generalised skew-t distribution, where the generalisation pertains to allowing different degrees of freedom for each of the dimensions, see [Cruz et al., 2015, Luo and Shevchenko, 2010]. Fung and Seneta describe skew-t distribution based on skew elliptical distribution, which they argue have better tail properties Fung and Seneta [2010b], Azzalini and Capitanio [2003], Azzalini [2005]).

We are not aware of any models for statistical causality incorporating any type of the skew-t copula, although parametric and nonparametric copulas have been used to model statistical causality by Bouezmarni et al. [2012], Bahadori and Liu [2013], Hu and Liang [2014], Lee and Yang [2014], Hu et al. [2015, 2016]. Hu and Liang [2014] formulate statistical causality as a of log-likelihood ratio, which they express in terms of copula and copula density and test with bootstrap test. Later, Hu et al. [2016] propose a model where the Wald test can be used to test significance of causality. Lee and Yang [2014] use nonparametric and parametric copula tests for Granger causality, describe how causality in quantiles can be modelled and tested and how it can be useful for Value-at-Risk (VaR) analysis and risk analysis.

There is growing literature on use of GPs in the time series with financial applications, with the [Ghoshal and Roberts, 2016] being of particular interest, due to their use of ARD kernels to help with modelling multimodality. Other interesting sources are [Roberts et al., 2012],[Requeima et al., 2019] Girard et al. [2003], Brahim-Belhouari and Bermak [2004], Cunningham et al. [2012], Hernández-Lobato et al. [2013]. GP models are not common in the context of causality, and even less often in the context of statistical causality. Causality described in the publication by Cunningham et al. [2012] should not be understood as causality in the sense of any of the concepts mentioned before, but as a property of time ordering. A unique work is the one by Amblard et al. [2012a], where marginal log-likelihood of a GP model is suggested to test for causality. The statistic they proposed for testing causality by is similar to what we use in GPC; however Amblard et al. [2012a] stopped short of introducing the LRT or GLRT, and were not able to utilise most of the properties of GPs. In their following work, Amblard et al. [2012b] asserted that kernelised ridge regression is more practical as the GP framework does not allow permit defining instantaneous causality (coupling). We dispute this, since the use of multiple output GPs does allow for instantaneous causality; and also kernel ridge regression can be seen as a less flexible special case of GPs.

Warped GPs have appeared in the machine learning literature in a number of different types of transformations [Snelson et al., 2004, Lázaro-Gredilla, 2012, Snoek et al., 2014, Adams and Stegle, 2008, Bornn et al., 2012], while relatable methods of spatial transformations have been widely studied in the spatial statistics literature [Anderes and Stein, 2008]. The most popular approach by Snelson et al. [2004], is to apply a nonlinear mapping, whose parameters can by learnt, such that transformed data is well modelled by a GP. In the statistics literature it is a standard practice to use a logarithm to transform the data at the stage of preprocessing, and Snelson et al. [2004] formalise this stage as part of a framework using GPs for modelling latent process. Let $\{X_t\}, \{Y_t\}$ denote observed time series, and $\{W_t\}$ denote latent process. With the notation from Section (3.1), with kernel k and assumption of the zero mean, we have a GP:

$$W_t \sim \mathcal{GP}(0, k; \Phi) \tag{1.66}$$

with the likelihood:

$$-\log \pi(W_t \mid X_t; \Phi) = \frac{1}{2}\log\det K_{1:T,1:T} + \frac{1}{2}\mathbf{X}_{1:T}^T K_{1:T,1:T}^{-1}\mathbf{X}_{1:T} + \frac{N}{2}\log 2\pi, \tag{1.67}$$

where $K_{1:T,1:T}$ is a covariance matrix, which can be defined elementwise as $\{K_{t_1,t_2}\}_{t_1,t_2=1}^T = k(X_{t_1}, X_{t_2})$. The

transformation, also called warping, is assumed to be a monotonic function f, such that:

$$W_t = f(Y_t; \Psi). \tag{1.68}$$

This results in the negative log-likelihood being described as follows:

$$-\log \pi\left(Y_t \mid X_t; \Phi, \Psi\right) = \frac{1}{2}\log\det K_{1:T,1:T} + \frac{1}{2} f(\mathbf{Y}_{1:T})^T K_{1:T,1:T}^{-1} f(\mathbf{Y}_{1:T}) - \sum_{t=1}^{T} \log \frac{\partial f(Y)}{\partial Y} \mid_{Y_t} + \frac{N}{2}\log 2\pi. \tag{1.69}$$

Subsequently, training and prediction is performed analogously to the standard GP model, but with the altered marginal likelihood.

Like Snelson et al., we also apply a transformation to the joint collection of GPs for each marginal time series model, although our approach differs in the class of transformations considered, and how we later use it. To the best of our knowledge, in literature such distortion mappings have only been applied marginally. In contrast, we develop a class of multivariate distortion or warping map. The mean-variance transformation applied in our method results in the transformed variables having multivariate skew-t distributions and being finite dimensional realisations of a general multivariate skew-t process whose likelihood can be obtained in a closed form.

Finally, across all of the approaches described earlier, only a handful of causal methods can allow application to data with more complicated structural forms. For example, [Papana et al., 2016] develops method of partial symbolic transfer entropy, to deal with nonstationarity that is a major issue for traditional transfer entropy. Chen claims that a spectrum estimator will allow recognition of Granger causality for a data with long memory, Chen [2006], while the spectral measures of causality by their nature should deal with periodicity. Candelon et al. [2013] describe method for testing causality in the tail. None of these models is flexible enough so that a common framework could be used, without the need to change models for different types of data.

1.4 Discussion on alternative modelling choices

There are multiple approaches to model selection, that can include inference procedures such as the likelihood ratio tests as well as the more popular ones based on the information criteria such as the Bayes factor [Kass and Raftery, 1995] or AIC [Akaike, 1974]. The approach favoured in this research is one of hypothesis testing rather than the alternative approach of ranking by information criterion. Furthermore, the models that are built are not interpreted as completely specified models of the processes being studied. Rather, the model specification is made in the context of the causal structures studied, even if this results in the model being misspecified for the distributional form, the trend or the covariance of the “true”

generating process of the data.

Adopting a hypothesis test, such as the Generalised Likelihood Ratio Test (GLRT) as an inference based procedure gives rise to three common challenges that must be considered when applying such tests in practice:

1. the need to evaluate the test statistic computationally in an efficient matter and in closed form,

2. being able to have an asymptotic distribution to perform the decision logic on (p-value), in order to avoid otherwise challenging time-series based bootstrap procedures that are computationally expensive and challenging to formulate in the general model settings proposed.

3. understanding and studying how the power and sensitivity of the model behaves.

The rationale for choosing the classical setting of GLRT, is that all three concerns can be addressed in the contexts of Gaussian Processes causal models, where causality was non-trivially entered into the model, such as in the trend and in the covariance, or both. Moreover, other warping factors were also possible to be included in these three considerations.

The Bayes factor approach to model selection provides a form of model ranking, whereas the objective of the framework adopted was to be able to make a binary decision on the basis of the model selection choice and to be able to understand and control Type I and Type II errors in this decision process.

Bayes factor would rank the models, but these rankings will not make distinct statements that model A is better than model B with respect to any control for errors of Type I or Type II. In contrast, making a decision on the probabilistic tail event with the hypothesis test will facilitate a more direct statement about plausibility of different model causal structures.

Under the GLRT framework we've been able to perform detailed studies of the Type I and Type II error, control that error through the decision of the hypothesis level of significance, and also study the power of those test in the specificity or sensitivity studies.

Another type of model one could consider are continuous time diffusive models (classical sources as Shreve [2004], Karatzas and Shreve [2012], Øksendal [2003]) as well as discrete time autoregressive models with continuous coefficients.

The reason why we did not focus on continuous time diffusion models is that when one works with non-standard and inhomogeneous drift and volatility functions, a significant amount of work is required to prove existence and uniqueness of the diffusion in the first place. That becomes even more complicated when trying to perform inference procedures. Doing stochastic calculus in these settings and discretising these processes is highly non-trivial when inhomogeneity in drift and volatility are present, but to capture causal structure in the manner I would like to introduce them requires such inhomogeneity in both drift and volatility, which would include state dependence, temporal dependence as well as dependence on

exogenous processes. Although one could specify such models, proving existence and uniqueness would be highly non-trivial, as the standard compact operator conditions would be very difficult if not impossible to verify except for very special cases of diffusion processes (please refer to Karatzas and Shreve [2012], Ikeda and Watanabe [2014], especially the theorem giving conditions under which strong solution with the pathwise uniqueness property exists in Karatzas and Shreve [2012]).

Another modelling choice is to use time series continuous coefficient models (most classical sources: Harvey and Stock [1985]). Difficulties with that framework pertain model identification and estimation, which are non trivial and may provide significant challenges when trying to build general causal structures in trend and covariance. Furthermore, it could also lead to lack of parsimony, which is another disadvantage compared to our chosen GP models, which even in ARD setting allows parsimonous representations.

1.5 Contributions

Expanding on recent results in the area of statistical causality, we have addressed many of the shortcomings of the methods that already exist. Our research advances how dependence structure and causality structure can be modelled, and will be of particular use in financial time series, or other types of data with a wide range of different structural properties.

Firstly, our framework allows to have aspects of causality, linear and nonlinear, in the mean and the covariance. We are not aware of research that studies these aspects in detail, and therefore we believe we add novelty to this literature, even if individual components of the framework are familiar to statisticians, the development of these in the context of mean and covariance causal structures is novel.

Secondly, analysing causal structure with Gaussian processes hasn't been done in the likelihood ratio framework, and in our research we propose a way to construct model nesting that allows for application of the likelihood ratio test. This model nesting is constructed to be applicable for assessing causality in the mean, or covariance, or both, and is achieved through ARD construction of the kernel (Automatic Relevance Determination). We explain that nested models are important, as the standard asymptotic distribution of χ^2 does not hold for non-nested hypotheses. Thus, we emphasise that the novelty does not lie in the asymptotic behaviour of the test, but in constructing a framework that allows to apply that test. Furthermore, with our GP model formulations the test statistic can be written in a closed form, can be computed point-wise, and is efficient to compute.

There are numerous advantages of using GPs, beginning with: ease of optimisation and interpretability of hyperparameters, flexibility, richness of covariance functions, allowing for various model structures. Using a likelihood ratio type test with a GP is a very natural choice, as estimating GP model parameters is often done on the basis of maximising likelihood, and therefore this estimation can be incorporated into the compound version of the likelihood ratio test (Generalised Likelihood Ratio Test, GLRT). From Gaussian variables, GPs inherited the property of being fully specified by the mean and the covariance,

and so testing for model equivalence inherently means testing for equivalence of the mean and covariance functions. But many popular kernels do not have the ARD property, and using them for a likelihood ratio test settings gives no easy way to account for causal structures in covariance. Consequently, it is using GLRT with an ARD-GP that gives a uniformly most powerful test with an unparalleled flexibility: known asymptotic distribution under the null, explicit evaluation and in a closed form, and usefulness also for misspecified models.

Thirdly, we demonstrate the ability to detect and identify causal structures in the mean and covariance, even in the presence of different types of model misspecifications. We undertake careful study of sensitivity and robustness of these testing frameworks to various features that one would encounter, like: sample size, parameter misspecification and structural misspecification. It is important as these studies demonstrate that one can reliably apply these tests in a general framework, even if the model is misspecified in those ways, and still have confidence that the inference procedure can detect these types of causality in mean and covariance incorporated in this framework reliably.

The fourth contribution of this dissertation is the proposal of a framework of Warped Gaussian Processes for Causality (wGPC) which allows for different types of causal and non-causal dependence to be described – important but rarely studied in the literature on causality or on classical parametric time series models. It can be customised in terms of mean functions, covariance functions, asymmetry, thicker tails; it allows for nonlinearity, autoregression, volatility clustering, nonstationarity, periodicity, tail dependence, etc. The wGPC framework can also be expended further, for example through introduction of different mean-variance transformations.

The extensive testing on synthetic and real-world data illustrates how causal analysis can form a part of a comprehensive time series analysis. Some of the experiments exemplify the interplay between different structures of time series (for instance, long memory and autoregression, causality and volatility clustering, causality and asymmetry or tail dependence), and how they all can influence understanding of dependence and causality.

1.6 Structure of the Dissertation

The first part of the thesis, *Background and literature review*, provides all of the material necessary to understand the context and the method proposed for modelling and testing statistical causality. The methods proposed in our work might be of interest to researchers from several fields, hence a detailed exposition of the concepts is helpful.

Chapter 1 introduces the topic of the thesis, related work and describes the structure.

Chapter 2 describes the original Granger's formulation of causality, explains how statistical causality differs from other notions of causality, and provides examples that demonstrate the strengths and weaknesses of existing methods for testing causality. Chapter 3 contains background on the models that are

used: GPs, copulas and selected distributions. Chapter 4 describes inference procedures used: assessing hypothesis tests, generalised likelihood ratio test, permutation tests, and likelihood ratio test.

The second part, *New Perspectives on Causality Representation and Inference*, presents the main contribution of our work. It starts with Chapter 5 containing the theoretical background for describing and testing causality with GP models. Chapter 6 extends the model from the previous chapter by introducing mean-variance transformation that results in a warped GP model, which can describe causality in the presence of skewness and tail dependence. Chapter 7, describes how synthetic data has been simulated, details the algorithm for approximating likelihood in the warped GP, and provides information on other relevant algorithms and the software used to implement our method. Chapter 8 presents a very extensive experiment section, which aim to show, firstly, the good behaviour of the proposed procedures (model sensitivity and misspecification analysis), secondly, good power of the test for a range of structures, and thirdly, the interaction of causality and tail dependence. Applications to real-world data are described in Chapter 9, where time series for commodities and currency markets are analysed.

The final section of the thesis: Chapter 10 *Conclusions*, summarise the research and offer directions for further development. The appendix in Chapter A contains theorem proofs, algorithms for likelihood estimation in the wGPC model, and other supplementary material.

Chapter 2

Overview and Comparison of Existing Causality Methods

> *Primary causes are unknown to us; but are subject to simple and constant laws, which may be discovered by observation, the study of them being the object of natural philosophy.*
>
> Jean-Baptiste-Joseph Fourier, *Théorie Analytique de la Chaleur*, 1822

> *It has been suggested that although such deeper relations need to be named, that name should not involve words like 'cause' or 'causality', as these words are too emotion-laden, involve too much preconception and have too long a history. (...) Provided I define what I personally mean by causation, I can use the term. I could, if I so wish, replace the word cause throughout my lecture by some other words, such as 'oshkosh' or 'snerd', but what would be gained?*
>
> Clive W.J. Granger, *Testing for causality : A personal viewpoint*, 1980

In modern statistics and data science there is a debate about where the different notions of causality arise and how they can be useful. An in-depth study of a causal method should start from putting that method into perspective: firstly, how it relates to other causal concepts, secondly, how it compares to different implementation of the same conceptual representations of causality. Chapter (1) introduced the concept of statistical causality and literature that is relevant to our research, in particular several methods that gained popularity or were important stepping stones. The current chapter takes a broader look at causality. We look at statistical causality as only one of concepts of causality that have been proposed, and we show different ways how the concept of causality has emerged and been studied over time. We refer to the history of philosophy and science to show how the concepts of causality were changing. We will pay special attention to the General Theory of Causation by Pearl, and will answer the following questions: How do Granger's statistical causality and Pearl's theory of causation cater to different needs?

What do they have in common or how they could be combined?

Finally we concentrate again on the representation that interests us the most – statistical causality – and take a closer look at four methods of measuring statistical causality that were introduced in Chapter (1). Those methods arise from different fields of statistics and data science, and exemplify what problems and potential issues lie in the process of testing causality. Their different properties, strengths and weaknesses are used to motivate why we turned to GPs to build a causal framework. This section is based, and contains excerpt from, the article "Measures of Causality in Complex Datasets with Application to Financial Data" [Zaremba and Aste, 2014].

2.1 Conceptual representations of causality

The ideas presented here appear in chronological order, and necessarily, we offer only a light sketch on the history of philosophical and scientific approaches to causality, as it too wide a topic for us to be able to cover in a more exhaustive way.

2.1.1 History of causal theory

Causality has been of interest to philosophers from the classical period, and from that time has been seen as a crucial tool for understanding the world. And just like humanity's ability to understand the world moved from the field of philosophy, theology to science, the same happened with what was seen as a domain of causal analysis. This is represented in Figure (2.1).

The first theory of causality, also known as the doctrine of four causes, is attributed to Aristotle – as presented in Physics and Methaphysics (see modern translations Aristotle et al. [2008] and Aristotle and Lawson-Tancred [1998]). For Aristotle, the theory of causality was a necessary tool to understand the world:

We suppose ourselves to possess unqualified scientific knowledge of a thing, as opposed to knowing it in the accidental way in which the sophist knows, when we think that we know the cause on which the fact depends as the cause of the fact and of no other, and further, that the fact could not be other than it is. (BWA, 111, Post. An. I.1, 71b 5–10; [Aristotle and McKeon, 1941])

The four causes were different types of answers to the "why?" question, and they relate to:

- the composition of the subject (**material cause**)
- the form of the subject (**formal cause**)
- the entity which is the source of the change that the subject undergoes (**efficient cause**)
- the final state - the result of the change/process or action undertaken (**final cause**).

Out of the four causes that Aristotle proposed, only the "efficient cause" aligns with the modern understanding of causality – whether it is the common usage, philosophical understanding, or mathematical

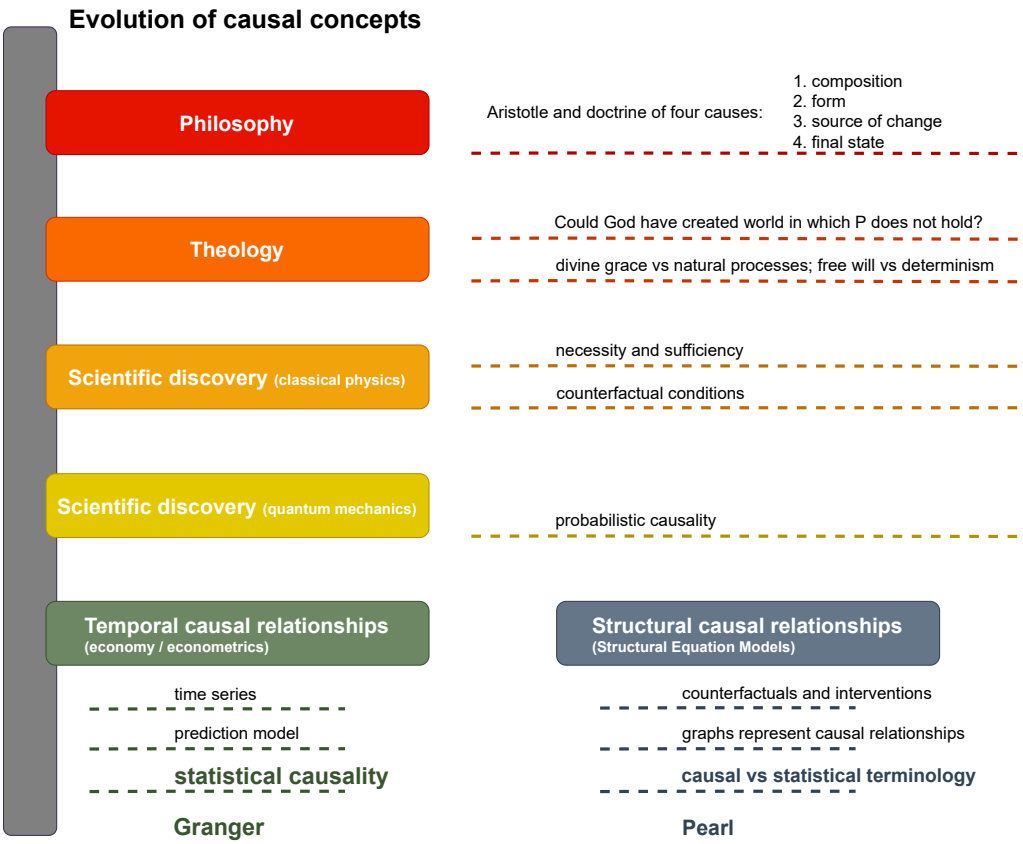


Figure 2.1: Causal concepts.

concept. For many centuries, however, the doctrine of four causes continued to be one of the most influential approaches to understanding causality.

In medieval times Aristotlean causality served as a point of reference, even though the model of the world changed, and the understanding of the world was described through the lens of theology. Aside from the natural processes, it was also the divine intervention or God's plan, that was crucial in the succession of the events:

Medieval thinkers believed that the world was created by God, and so a question like "Is proposition P contingent?" were seen as equivalent to the question "Could God have created a world in which P does not hold?". So our question can be reduced to one about divine power. [White, 2018]

The question "why?" was crucial for subsequent attempts to understand causality, but the focus moved to studying the processes and the relationship between the cause and the effect. The inspiration for studying causality was often coming from the sciences, especially physics and chemistry, for example: what causes fire, movement, gravitation, and in all these cases causality was described in deterministic terms. Indeed, from the XVIII century until XX century, the main approach to causality was deterministic, as seen in Hume [2008], Kant [2007], Mill [2015], Russell [1912]. In the discussion about the nature of causality, there were two aspects that were seen as most important [Granger, 1980]

- **Necessity** - if A occurs, then B must occur.

  Kant argued that there must be a necessary connection between the cause and the effect:

  [...] that something, A, should be of such a nature, that something else, B, should follow from it necessarily [Kant, 2007]

- **Sufficiency** - if B occurred, then A must have occurred.

  Hume defined causality based on the condition of sufficiency:

  We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or in other words, where, if the first object had not been, the second never had existed. [Hume, 2008]

Hume has also conjectured that the notion of time and space are necessary to the concept of causality. *Since therefore it is possible for all objects to become causes or effects to each other, it may be proper to fix some general rules, by which we may know when they really are so.*

1. *The cause and effect must be contiguous in space and time.*
2. *The cause must be prior to the effect.*
3. *There must be a constant union betwixt the cause and effect. It is chiefly this quality, that constitutes the relation. [Hume, 2010]*

The conditions of necessity and sufficiency described above do not capture all of the conditionals, or "if-clauses", relevant to deterministic causality. We can also look at **counterfactual conditionals** or "false if-clauses": if A had not happen, then B had not happen, or: if B had not happen, then A would not have happened. If we think about common use of the words "cause", "causal" and "causality", then all, or possibly a subset of the aforementioned conditionals will come to one's mind. We refer to [Frosch and Johnson-Laird, 2011] for a discussion about everyday "causation", but urge the reader to keep in mind Granger's words:

One interesting aspect of the philosophers' contribution is that they often try to discuss what the term causality means in "common usage", although they make no attempt to use common usage terms in their discussion. Rather than trying to decide what the public thinks they mean by such a difficult concept as causality, it may be preferable to try to influence common usage towards a sounder definition. [Granger, 1980].

Counterfactual analysis, which has long been present in the philosophical discourse about causality, received more attention after the publication by Lewis [1974]. The definition of causal dependence proposed by David Lewis is based on two conditionals:

An event E causally depends on C if, and only if, (i) if C had occurred, then E would have occurred, and (ii) if C had not occurred, then E would not have occurred. [Lewis, 1974]

In the 20th century development of science was again the reason for emergence of a new concept of causality. While Newton's laws of motion can be seen as describing physical phenomena in terms of deterministic causality, the deterministic approach was not suitable to model quantum physics. The development of modern probability theory, whose foundations were laid by Kolmogorov [1933], has provided tools for developing non-deterministic model of causality. Probabilistic approach to causality means that the effect is no longer seen as necessary, but more likely as a result of the cause. Reichenbach [1991] proposes the condition for A being the cause of B should be:

$$\mathbb{P}(B \mid A) > \mathbb{P}(B \mid \text{not } A). \tag{2.1}$$

The condition from Equation 2.1, does not explicitly include the notion of time, but the interpretation was one of a temporal asymmetry. The development of probability theory was followed by certain notions of probability entering into common usage, and consequently the common understanding of what causality means started to include probabilistic causality. A great example, from [Granger, 1980], is the assertion that smoking causes cancer: it is generally understood that while smoking will not always lead to cancer, it is one of the biggest contributory factors. This understanding of causality is what both the scientists as well as non-scientifically inclined people are able to accept.

In 1956, Norbert Wiener proposed a conceptual representation of causality that is central to our understanding of this concept – based on predictive models for stochastic processes:

For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one. [Wiener, 1956].

$$H_0^{Wiener}: \quad \text{Prediction } (Y_t \mid X_{t-k:t-1}, Y_{t-k:t-1}) \equiv \text{Prediction } (Y_t \mid Y_{t-k:t-1}). \tag{2.2}$$

2.1.2 Modern approaches to causality – two main concepts, and the link between them

Wiener's work has been one of the inspirations for Clive Granger and his groundbreaking work on causality, popularly called **Granger causality**. Granger has introduced a rigorous definition of causality, by clarifying the assumption about the data generating process, the model used for prediction, as well as the test statistic for assessing the null hypothesis. The data was assumed to be generated by a stationary, autoregressive process, the model used for prediction was linear regression, the null hypothesis was that of equivalence of variances of the two predictions, and the test statistic was based on the ratio of the two variances. For a formalisation, please refer to Chapter (1), Equations (1.1 - 1.6). In this section we will only repeat the definition that Granger [1980] introduced later, based on conditional distributions. For the

time series $\{X_t\}, \{Y_t\}$ and side information $\{Z_t\}$, the hypothesis of lack of causality $X \to Y \mid Z$ is written as equality of conditional distributions:

$$H_0 : \qquad P\left(Y_t \mid X_{t-k:t-1}, Y_{t-k:t-1}, Z_{t-k:t-1}\right) = P\left(Y_t \mid Y_{t-k:t-1}, Z_{t-k:t-1}\right). \tag{2.3}$$

The definition from Equation (2.3) lies at the centre of our understanding of causality, and therefore of this thesis. We will later see that the conditional independence is also important when building causal graphs.

Granger made certain assumptions, that he has called axioms, [Granger, 1980]:

Axiom A, Time ordering: The cause happens prior to the effect.

Axiom B, No redundant information: The cause contains unique information about the effect – it is not related via a deterministic function.

Granger has also pointed out that there is a third axiom which is "[...] *generally accepted, even though it is not necessarily true*", which he saw as central to the applicability of the concept of causality:

Axiom C, Consistency: The existence and direction of the causal relationship remain constant in time.

We strongly recommend reading Granger's opinion on other views on causality and discussion on the need for introducing a new definition. Statistical causality relies on the historical data that relates to observation. This, according to some, is not enough to be able to distinguish the true cause. Such sentiment was expressed by Hume:

If all our information derives from empirical observation, how can we be sure that any particular explanatory theory is the correct one?

The above quote can be seen as a starting point to Pearl's General Theory of Causation, [Pearl, 2000, 2010]. According to Pearl, it was not only the case that the existing methods were unsatisfactory, but the field of statistics did not have the language to deal with causal concepts. Pearl insisted on distinguishing between associations and causal relationships, with observation and joint distribution being enough to establish the former, but not the latter:

A useful demarcation line between associational and causal concepts crisp and easy to apply, can be formulated as follows. An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone. Examples of associational concepts are: correlation, regression, dependence, conditional independence, likelihood, collapsibility, propensity score, risk ratio, odds ratio, marginalization, conditionalization, "controlling for," and many more. Examples of causal concepts are: randomization, influence, effect, confounding, "holding constant," disturbance, error terms, structural coefficients, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, and attribution. The former can, while the latter cannot be defined in term of distribution functions. [Pearl, 2010]

Using the criterion from the above quote, causality in the sense of Granger is an associational concept. According to Pearl, given adequately large sample and precise measurements, one can in principle test associational assumptions, but not causal assumptions, which require experimental control. Pearl has revisited and reinvented the counterfactual theory, which before has been used mostly as a philosophical concept, and as such has not been formalised mathematically and lacked relevant language. Pearl postulated that causal analysis should allow to infer probabilities under static conditions, as well as how they change under dynamic conditions, by answering three types of questions:

1. Policy evaluation: what is the effect of potential intervention?
2. Probabilities of counterfactuals: can an event be identified as responsible for another event?
3. Mediation: can causal effect be assessed as direct or indirect?

The investigation starts from formulating a causal model and building a path diagram that represents causal assumptions. Inspection of this diagram, and the use of graphical model tools, allows one to decide next steps to obtain the target quantity – as shown in Figure (2.2). Pearl [2000] defines causal model $M = \langle U, V, E \rangle$ in terms of exogenous variables U, endogenous variables V, and a set F of functions that describe mappings from the exogenous variables to endogenous variables. A probabilistic causal model is then described as a pair $\langle M, \mathbb{P}(u) \rangle$, of causal model M and the probability function defined over the set of exogenous variables U. The path diagram carries the information about potential causal relationships via edges and paths, and missing edges mean lack of correlation or direct effect. Graphical methods, in particular "d-connection" and "d-separation" can then be used to verify which conditional independence relationships hold, based on the observed data. The effect of intervention, which can be analysed with the use of "do operator" can also be studied using the path diagram. These tools allow to establish whether a causal relationship is identifiable or not.

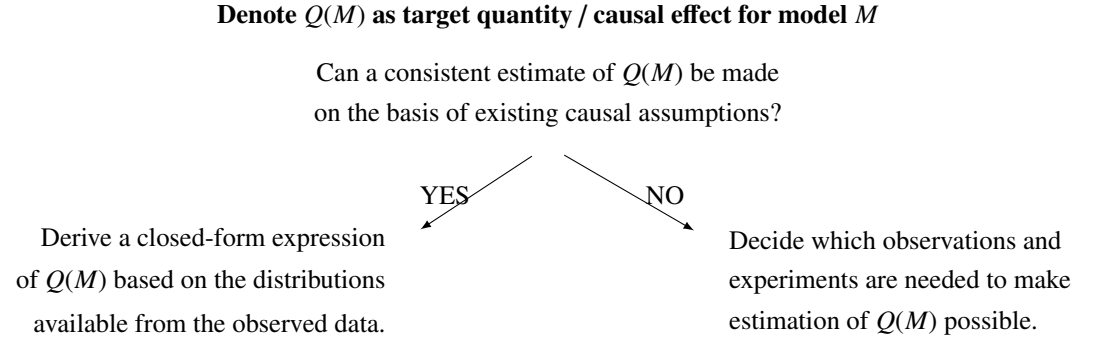


Figure 2.2: Obtaining the estimate of target quantity for model M (denoted $Q(M)$), based on the existing causal assumptions.

An example of analysis of intervention, and a use of a "do operator" can be observed on a simple case of a model described in Pearl et al. [2009]. Let X, Y, Z stand for observable random variables of interest,

and U_X, U_Y, U_Z represent variables or factors that might be observable or unobservable, but are not of interest for the model, they are called "exogenous variables", and might be seen as 'disturbances". The lower case x, y, z, u_X, u_Y, u_Z are realisations of the described variables. Let us assume that the observed variables can be described as:

$$\begin{aligned} \text{Model } M: \quad & z = f_Z(u_Z) \\ & x = f_X(z, u_X) \\ & y = f_X(x, u_Y) \end{aligned} \tag{2.4}$$

where U_X, U_Y, U_Z are assumed jointly independent, and the functions f_X, f_Y, f_Z are some functions whose specific form is not necessary to be known to analyse interactions between the variables of interest. It is only assumed that each of the functions is invariant to possible changes in other functions, i.e. is "autonomous", an property characterising a **structural** system of functions. The Equations (2.4) correspond to the path diagram from Figure (2.3 a).

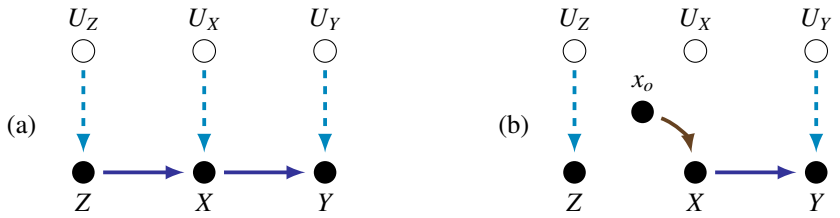


Figure 2.3: (a) Diagram representing the model from the Equations (2.4). (b) The same model but with an intervention $do(X = x_0)$, as per the Equations (2.5).

An important development in Pearl's approach is introduction of so called '*do*-calculus', which is a tool for modelling interventions and counterfactuals. The '*do*-operators', denoted as $do(x)$, are representing the operation of some of the functions in the model, and replacing them with a constant: $X = x$. An operation $do(x_0)$ of setting X to be equal x_0 in the system from the Equations (2.4) is visually represented in the Figure (2.3 b). We note here that the Equations (2.4) represent a structural system of functions, which means that the the operation $do(x_0)$ affects only the functional representation f_X, but not f_Y or f_Z:

$$\begin{aligned} \text{Model } M_{x_0}: \quad & z = f_Z(u_Z) \\ & x = x_0 \\ & y = f_X(x, u_Y) \end{aligned} \tag{2.5}$$

With the pre-intervention joint distribution of the original model M denoted as $\mathbb{P}(X, Y, Z)$, the notation for post-intervention joint distribution of the modified model M_x is $\mathbb{P}(Y, Z \mid X = x)$. We define

the **post-intervention distribution** of the outcome Y in the model M_x, is the probability $\mathbb{P}(Y = y \mid do(x)) = \sum_z \mathbb{P}(Y = y, Z = z \mid do(x))$ that the model M_x assigns to each outcome level $Y = y$. Pearl gives two possible ways of assessing an effectiveness of a treatment: average difference and experimental Risk Ratio:

$$\text{average difference :} \qquad \mathbb{E}\left(Y \mid do(x_0')\right) - \mathbb{E}\left(Y \mid do(x_0)\right) \tag{2.6}$$

$$\text{experimental Risk Ratio :} \qquad \frac{\mathbb{E}\left(Y \mid do(x_0')\right)}{\mathbb{E}\left(Y \mid do(x_0)\right)}. \tag{2.7}$$

To assess the **causal effect of** X **on** Y one need to obtain the distribution $\mathbb{P}(Y = y \mid do(x))$, [Pearl, 2000, Pearl et al., 2009]. Let us introduce the notation for the causal quantity of interest: $Q(M_x) = \mathbb{P}(Y = y \mid do(x))$. Not in all cases will one be able to estimate the causal effect $Q(M_x)$ from the data and the pre-intervention distribution. This is the question of **identification**, which Pearl has formalised with the following definition:

Definition 1 *(Identifiability, [Pearl, 2000, Pearl et al., 2009]) A quantity $Q(M)$ is identifiable, given a set of assumptions A, if for any two models M_1 and M_2 that satisfy A, we have:*

$$\mathbb{P}(M_1) = \mathbb{P}(M_2) \Rightarrow Q(M_1) = Q(M_2). \tag{2.8}$$

.

An important criterion for the causal effects for model M to be identifiable, is that model M is **Markovian**, defined as being represented by an acyclic graph, and having all of the error terms jointly independent, [Pearl, 2000]. Identifiability for non-Markovian models is more complicated, but also can be established on the basis of the graph that represents the model. What if the causal effect cannot be identified by any method? In such a case it can only be approximated by deriving bounds.

We would like to emphasise, that causality in the sense of Pearl, does not require temporal ordering - which is one of the principal assumptions of statistical causality. It is however possible to build a time-ordered representation, if the variables in V are time indexed, and the causal assumption encoded in the model M are such that only the earlier variables can cause later ones.

The Pearl's general theory of causation has been seen as irreconcilable with Granger's approach to modelling causality. Eichler [2001] proposed using Granger causality graphs – a way to merge Granger causality with graphs. Later, Eichler and Didelez [2007] introduced the idea of defining causality in time series in the context of intervention, that would bring the causal graphs as defined by Pearl and Granger causality even closer. Notably, Eichler and his colleagues were not the only to consider Granger causal graphs, or link Granger causality with Pearl's causal model. Billio et al. [2012] and Fiedor [2015] analyse financial networks built on the basis of, respectively, linear Granger causality and transfer entropy as

similarity measures. These authors use complex networks tools, rather than graphical model, to inspect the networks, and are interested in properties of the network as a whole, and not in individual causal relationships. White et al. [2011] demonstrate how Pearl's Causal Model and Granger causality are linked, when expressed in terms of extension of Pearl's Causal Model with settable systems.

Eichler [2012] defines causality in the framework of Florens and Mouchart [1982], Florens and Fougere [1996], in terms of conditional independence, and for continuous time, as introduced in Chapter(1) Equation (1.18):

Definition 2 *Granger non-causality, [Eichler and Didelez, 2007]*
Let $\{X_t\}, \{Y_t\}, \{Z_t\}$ *be stationary time series, possibly multivariate.* X_t *is (strongly) Granger non-causal for* Y_t *up to horizon* $h, h \in \mathbb{N}$*, with respect to process* $[X_t, Y_t, Z_t]$ *if:*

$$Y_{t+k} \perp\!\!\!\perp X_{1:t} \mid [Y_{1:t}, Z_{1:t}], \tag{2.9}$$

for all $k = 1 : h$*, and* $t \in \mathbb{Z}$*. If the above holds only for* $h = 1$*, then it is simply said that* X_t *is (strongly) Granger non-causal for* Y_t *with respect to process* $[X_t, Y_t, Z_t]$ *and denoted* $X_t \nrightarrow Y_t [X_t, Y_t, Z_t]$*, and if it holds for all* $h \in \mathbb{N}$*, then it is said that* X_t *is (strongly) Granger non-causal for* Y_t *at all horizons, with respect to process* $[X_t, Y_t, Z_t]$ *and denoted* $X_t \nrightarrow Y_t [X_t, Y_t, Z_t]$*.*

Definition 3 *Contemporaneous independence, [Eichler and Didelez, 2007]*
Let $\{X_t\}, \{Y_t\}, \{Z_t\}$ *be stationary time series, possibly multivariate.* X_t *and* Y_t *are contemporaneously independent, with respect to process* $[X_t, Y_t, Z_t]$ *if:*

$$Y_{t+1} \perp\!\!\!\perp X_{t+1} \mid [X_{1:t}, Y_{1:t}, Z_{1:t}], \tag{2.10}$$

for all $t \in \mathbb{Z}$*. Contemporaneous independence is denoted* $X_t \nsim Y_t [X_t, Y_t, Z_t]$*.*

Eichler [2012] proposed building so called Granger causality graphs to represent Granger causal relationships. Granger causality graphs $G = (V, E)$ are mixed graphs, with time series represented as nodes, directed edges representing Granger causality, and undirected edges – contemporaneous dependence. A direct edge from $X \longrightarrow Y$ (or $X \longleftarrow Y$) will be drawn if the hypothesis of non-causality is rejected, an undirected edge $X - -Y$, if the hypothesis of contemporaneous independence is rejected, therefore three edges can be drawn between every two of the nodes. A Granger causal graph can be shown on a following

example

$$\begin{aligned} \text{Model } M: \quad & X_t = a_X X_{t-1} + \epsilon_t^X \\ & Y_t = a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_t^Y \\ & Z_t = a_Z Z_{t-1} + b_Z Y_{t-1} + c_Z X_{t-3} + \epsilon_t^Z \end{aligned} \tag{2.11}$$

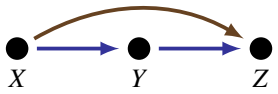


Figure 2.4: Diagram representing the model from the Equations (2.11).

Eichler and Didelez [2007] introduce interventions to the Granger causal framework in a way that is similar to Pearl's approach, albeit using a different notation. An intervention is signalled with an **intervention indicator** σ, where $\sigma_X(t) = x_0$ means that Y_t has been changed to y, with $\sigma_X(t) = \emptyset$ denoting no intervention. This can be seen as equivalent to Pearl's do-operator. Although we will not introduce them here, Eichler and Didelez [2007] defined also conditional, random and multiple interventions.

An assessment of a causal effect can be made by analysing the **average causal effect** (ACE) of X_t on $Y_{t+h}, h > 0$, following a strategy $\sigma_X(t) = x_0$:

$$ACE_{x_0} = \mathbb{E}_{sigma_X(t)=x_0} Y_{t+h}. \tag{2.12}$$

The difference $ACE_{x_1} - ACE_{x_2}$ will then be analogous to the Pearl's average difference from the Equation 2.6. Just like Pearl, Eichler and Didelez [2007] pose then a question about identifiability of the causal effect:

A priori there is no reason why data that is not collected under the regime of interest should allow estimation of the ACE. By identifiability we mean the possibility to express the ACE in terms of quantities that are known or estimable under the observational regime. [Eichler and Didelez, 2007]

The criteria for identification that Eichler and Didelez [2007] give are then based on the graphical model properties of the graph G.

One of the main conclusion from the work of Eichler and his colleagues is that Granger causality can be of use even if causality is understood in terms of interventions. In such a framework, however, Granger causality can be used as a **potentially causal relation**. The criteria for identification that Eichler and Didelez [2007] give can be seen as requiring that the set of multivariate time series and side information must be "rich" enough. They also point out that while not including any concepts of interventions or identifiability, Granger did suggest that the information available needs to include all "relevant" information.

2.2 Strengths and Weaknesses of Existing Methods, Based on Four Chosen Methods

A researcher who decided to use statistical causality in their work, has to choose how to formulate the hypothesis, which test statistic and test to use, how to estimate the test statistic, and finally how to interpret the results. In this section we consider four of the methods introduced in Chapter (1). We analyse their properties and performance in experiments on simulated data structures, and on real data. This section points out the strength of the four methods, but more importantly, also the weaknesses, that inspired the research presented in this thesis.

2.2.1 The Four Chosen Methods

The first method in this section is the classical Granger causality, which having many drawbacks is nevertheless still commonly used as a benchmark. It is compared to three nonlinear methods: transfer entropy (TE), kernel ridge regression (krr) and a nonparametric conditional dependence measure based on the normalised conditional cross-covariance operator (which I will refer to as HSNCIC, for Hilbert Schmidt Normalised Conditional Independence Criterion).

The test statistic for these four methods, which we remind below, were already defined in Equations (1.8, 1.26, 1.50, 1.62).

1. Classical Granger causality (GC); Test statistic defined as improvement in the prediction error variance.

$$L^{GC}_{X\to Y} = \log\left[\frac{V_Y\left[Y,Z;p\right]}{V_Y\left[X,Y,Z;p\right]}\right] \tag{2.13}$$

2. Transfer Entropy (TE); Test statistic defined as a difference in conditional entropies. Designed to measure departure from generalised Markov property.

$$L^{TE}_{X\to Y} = H(Y \mid Y_{t-k:t-1}) - H(Y \mid X_{t-k:t-1}, Y_{t-k:t-1}). \tag{2.14}$$

3. Kernel ridge regression (krr), in some literature called "kernelised Geweke" [Amblard et al., 2012b]; Test statistic quantifies improvement in the prediction error variance for kernel ridge regression.

$$L^{krr}_{X\to Y} = \log\frac{V\left(\hat{\mathbf{Y}}^{A}_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)}{V\left(\hat{\mathbf{Y}}^{B}_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)}. \tag{2.15}$$

4. Hilbert-Schmidt Normalised Conditional Independence Criterion (HSNCIC); Test statistic calcu-

lated as a squared Hilbert-Schmidt norm for a normalised cross-covariance operator.

$$L_{X\to Y}^{HSNCIC} = \|V_{(Y,Z)(X,Z)|Z}\|_{HS}^2 \tag{2.16}$$

The last method which has gained less popularity than the previous frameworks, is interesting from methodological point of view, and also is more strongly connected to Machine Learning tools. The name Hilbert-Schmidt Normalised Conditional Independence Criterion (HSNCIC) is used following Seth and Principe [2011]. We refer the reader to Chapter (3) Section (3.1.1) for an introduction to functional analysis including the Hilbert-Schmidt norm and the family of covariance operators.

In the Chapter (1), we have described several testing method, and we described that linear Granger causality has tests, whose asymptotic distributions are known analytically. The other methods, however, need to use some form of permutation test. When using the permutation test, we will work with the following hypotheses:

$$H_0: \qquad L_{X\to Y} = 0, \qquad \text{no causality from } \{X\} \text{ to } \{Y\} \tag{2.17}$$

$$H_1: \qquad L_{X\to Y} > 0, \qquad \text{causality from } \{X\} \text{ to } \{Y\} \tag{2.18}$$

The test statistic for permutation test is obtained by using time series, with permuted time order. Let $\mathrm{p}_i(t),\ t = t_1, ..., t_2$ denote a random permutation of the time index, and $\mathrm{p}_i(X)$ denote a time series, where the original time order has been reorganised according to the permutation $\mathrm{p}_i(\cdot)$. Then the null hypothesis is assessed by comparing the value of $L_{X\to Y}$ to a histogram of values of $L_{\mathrm{p}_i(X)\to Y}$, and a p-value:

$$\pi(L_{X\to Y} \mid H_0) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left(L_{\mathrm{p}_i(X)\to Y} > L_{X\to Y}\right) \tag{2.19}$$

where $\mathbf{1}(A)$ is a characteristic function for the set A.

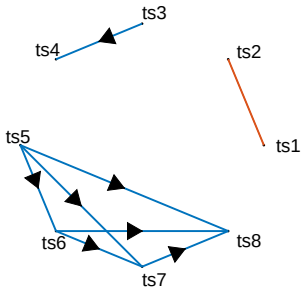
Depending on the number of permutations used, we suggest to accept the hypothesis of causality for the level of significance equal to 0.05 or 0.01. In our experiments, we report either single p-values or sets of p-values for overlapping moving windows. The latter is particularly useful when analysing noisy and non-stationary data. In the cases where not much data is available, we do not believe that using any kind of subsampling (as proposed by Sun [2008], Amblard et al. [2012b], Seth and Principe [2011]) will be beneficial, as far as the power of the tests is concerned.

2.2.2 Results

Linear multivariate data; causality at different lags and instantaneous coupling.

We performed an experiment on a set of eight time series with linear relationship: causal dependence at lags 1 to 3, as well as instantaneous coupling (illustrated in the Table (B.1). We expected all methods to

Table 2.1: Dependence structure of the simulated data.



	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	×	0						
ts2	0	×						
ts3			×	-1				
ts4			1	×				
ts5					×	-1	-2	-3
ts6					1	×	-1	-2
ts7					2	1	×	-1
ts8					3	2	1	×

Table 2.2: The directionality of causality between the eight simulated time series ts1 - ts8. Blue lines represent causality with the arrowheads indicating direction, red line indicates instantaneous coupling. The table shows lags at which true dependence occurs, with the interpretation that column variable causes row variable.

detect causality, and wanted to test the effect of causal relationships existing at different lags. The details of this experiment are in the Appendix (B.2).

All four methods performed well, although each of them had one false positive classification. Classical Granger causality and kernel ridge regression method performed similarly, correctly identified all lags and detecting causality and instantaneous coupling, but not rejecting spurious causality ts7→ts6, though. Those two methods were able to analyse all time series, and include all 3 lags. Transfer entropy, which typically analyses one lag at a time, successfully detected causality, however it failed to reject spurious causality ts1→ts7. Similar results were obtained for the HSNCIC, with false positive ts7→ts2. This method was, however,much slower slower.

Nonlinear multivariate data.

In this experiment, we analysed three time series $\{X_t\}, \{Y_t\}, \{Z_t\}$, with the following dynamic structure:

$$\begin{cases} X_t = aX_{t-1} + \epsilon_{X,t} \\ Y_t = bY_{t-1} + dX_{t-1}^2 + \epsilon_{Y,t} \\ Z_t = cZ_{t-1} + eY_{t-1} + \epsilon_{Z,t} \end{cases} \tag{2.20}$$

This data exhibits a direct linear causality $Y \to Z$, a direct nonlinear causality $X \to Y$, and a nondirect nonlinear causality $X \to Z$. The complete results are in the Appendix (B.3). A similar data structure is introduced in Section (3.2) and later used in experiments.

What this experiment allowed us to observe was that classical Granger causality was – as expected – able to detect the linear relationship, but not the nonlinear ones. The other three methods were all able to detect the nonlinear causal relationships $X \to Y$, but only kernel ridge regression and HSNCIC were also detecting the indirect causality $X \to Z$.

In this experiment we confirmed that the kernel ridge regressions method behaved like classical Granger causality when linear kernel, and also that the choice of kernel parameters play a role in the ability to detect causality. The choice of hyperparameters for kernel ridge regression was based on cross-validation.

Interest rates and inflation.

The first financial data application has been performed for the consumer price index for the United States (U.S. CPI) and the London Interbank Offered Rate (Libor) interest rate index. Both were monthly data from January 31, 1986, to October 31, 2013, obtained from Thomson Reuters. These data series are of vital importance from financial and microeconomic point of view, and have been studied extensively, also with tools of Granger causality [Eichler, 2007]. More details in Appendix (B.4.2).

Libor is often used as a base rate (benchmark) by banks and other financial institutions, and it is an important economic indicator. It is not a monetary measure associated with any country, and it does not reflect any institutional mandate in contrast to, for example, when the Federal Reserve sets interest rates. Instead, it reflects some level of assessment of risk by the banks who set the rate. Therefore, we ask whether we detect that one of these two economic indicators causes the other one in a statistical sense? [Zaremba and Aste, 2014].

The results shown that at lag of one month, there was a strong evidence for the direction CPI →Libor, at acceptance level of 1% , especially for the longest time window considered. In Figure (2.5) one can see that there was only one period in which the causal direction CPI→Libor would not be accepted at levels 1% and 5% (but would still be accepted at 10%). The direction from 1 month Libor to CPI, was exhibiting the opposite behaviour, with the hypothesis of lack of causality rejected only for one short period. As the scatter plot from Figure (2.5)shows, the two directions were clearly separated, meaning that there was strong evidence for causality but not feedback.

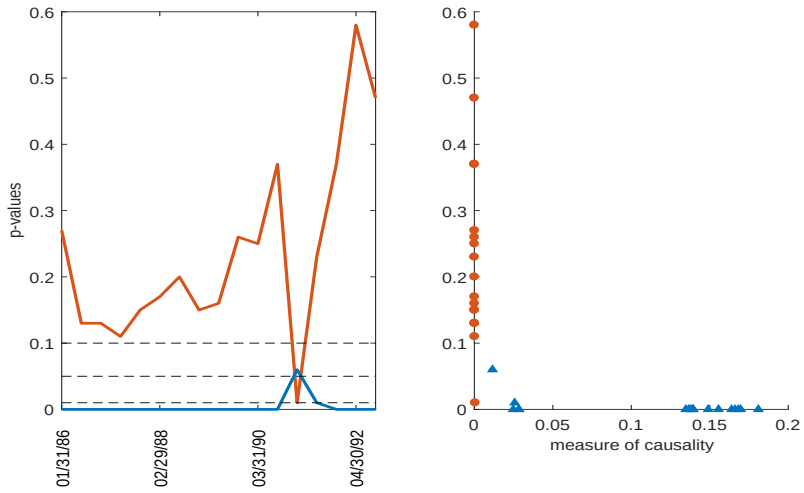


Figure 2.5: Kernelised Geweke's measure of causality. The left chart shows sets of p-values for the hypothesis that inflation statistically causes Libor (blue line) or the other way round (red line), when a model with one lag is considered. The right chart shows the scatter plot of p-values and the value of the causality measure. It represents the separation between the two causal directions – which in this case is substantial.

This causal effect was also detectable for larger lags, in particular lag 2 and 7, and decreased with shorter data windows. Moreover, the results from Figure (2.5) were nearly replicated by the results for linear Granger causality, which indicated that there was a strong linear effect. Transfer entropy and HSNCIC were in many cases unable to provide significant results, and were also not able to distinguish between the two directions.

As before, the choice of hyperparameters for kernel ridge regression was based on cross-validation, and possibly the most important conclusion was that parameter selection turned out to be critical for kernel ridge regression method. For some tests, like the simulated 8 time series data described earlier, size of the kernel did not play an important role, but in some cases the size of the kernel was crucial in allowing the detection of causality. However, there was no kernel size that worked for all of the types of the data.

Equity versus Carry Trade Currency Pairs

The last experiment was performed on the following exchange rates: AUDJPY, CADJPY, NZDJPY, AUDCHF, CADCHF, NZDCHF, together with S&P index, for daily data for the period July 18, 2008–October 18, 2013, from Thomson Reuters. The study aimed to investigate any patterns of the type "leader - follower". and had an expectation that S&P should be leading. More details in Appendix (B.4.3).

The analysis was performed with kernel ridge regression method using linear and Gaussian kernels, and it shown similar results for the two kernels. The study has shown several periods where S&P seemed to lead, however they were less prominent than expected. This has raised questions about how appropriate

Measures	Properties
	Linearity versus nonlinearity
Granger causality	assumes linearity; the best method for linear data, the worst for nonlinear
kernelised Geweke's	works for both linear and nonlinear data
transfer entropy	works for both linear and nonlinear data
HSNCIC	works for both linear and nonlinear data if low dimension
	Distinguishing direct from indirect causality
Granger causality	to some extent by comparing measure with and without side information
kernelised Geweke's	to some extent by comparing measure with and without side information
transfer entropy	not able to (consider partial transfer entropy)
HSNCIC	to some extent, as it is designed to condition on side information
	Spurious casuality
Granger causality	susceptible
kernelised Geweke's	susceptible
transfer entropy	susceptible
HSNCIC	susceptible
	Good numerical estimator
Granger causality	yes
kernelised Geweke's	yes
transfer entropy	no
HSNCIC	yes
	Nonstationarity
Granger causality	v. sensitive; test with ADF, KPSS, use windowing, differencing, large lag
kernelised Geweke's	somewhat sensitive; online learning is a promising approach
transfer entropy	somewhat sensitive
HSNCIC	somewhat sensitive
	Choice of parameters
Granger causality	lag
kernelised Geweke's	kernel, kernel size, regularisation parameter, lag; uses cross-validation
transfer entropy	lag, binning size (if histogram approach used)
HSNCIC	kernel, kernel size, regularisation parameter, lag

Table 2.3: The summary of main features of the different measures

was the choice of the models? How appropriate was the kernel choice, model selection, and parameter optimisation?

Chapter 3

Models

> *I am among those who think that science has great beauty.*
>
> Marie Skłodowska Curie, *Physics Book 1*

> *"You don't see, Genry, why we perfected and practice Foretelling?"*
> *"No"*
> *"To exhibit the perfect uselessness of knowing the answer to the wrong question."*
>
> Ursula K. Le Guin, *The Left Hand of Darkness*

This chapter provides an overview to models that form building blocks of our approach to modelling causality. The first section will introduce elements of functional analysis that will be foundations for kernels and kernel methods, second will concentrate on Gaussian processes (GPs), while the third section will be devoted to selected multivariate distributions: copulas, generalised hyperbolic distributions, skew-t distributions, tail behaviour.

3.1 Introduction to Gaussian Processes

A critical component for performing causality analysis, detection, and inference is to focus on the model component of a GP. GPs can be understood both from the probabilistic perspective and functional analysis perspective, and we will be adopting both to some extent. From the functional analysis perspective it suffices to analyse some basic properties of the space where sample paths – realisations – will be defined, and for this purpose it is useful to introduce Hilbert spaces. Quoting Hein et al.:

Positive definite kernels are extremely powerful and versatile tools. They allow to construct spaces of functions on an arbitrary set with the convenient structure of a Hilbert space. Methods based on such kernels are usually very tractable because of the particular structure (reproducing property) of the space of functions. This has a large number of applications, in particular for statistical learning, approximation or interpolation where one has to manipulate functions defined on various types of data (...)[Hein and

Bousquet, 2004]

3.1.1 Functional analysis and Hilbert spaces for positive definite kernels

The definitions and theorems below follow Steinwart and Christmann [2008], Gretton et al. [2005], Sun [2008]. All vector spaces will be over $\mathbb{R}$, rather than $\mathbb{C}$, however they can all be generalised for $\mathbb{C}$ with little modification.

Definition 4 *(Inner product) Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H}\times\mathcal{H}\to\mathbb{R}$ is said to be an inner product on $\mathcal{H}$ if:*

$$\begin{aligned} &i)\ \langle f_1+f_2, f\rangle = \langle f_1, f\rangle + \langle f_2, f\rangle, \quad \text{for all } f, f_1, f_2 \in \mathcal{H} \\ &ii)\ \langle \alpha f_1, f_2\rangle = \alpha\langle f_1, f_2\rangle \quad \text{for all } f_1, f_2 \in \mathcal{H}, \alpha\in\mathbb{R} \\ &iii)\ \langle f_1, f_2\rangle = \langle f_2, f_1\rangle \quad \text{for all } f_1, f_2 \in \mathcal{H} \\ &iv)\ \langle f, f\rangle \geq 0 \quad \text{and } \langle f, f\rangle = 0 \quad \text{if and only if } f = 0. \end{aligned} \tag{3.1}$$

Definition 5 *(Norm induced by inner product)*
If $\langle\cdot,\cdot\rangle$ is an inner product on $\mathcal{H}$ and $f\in\mathcal{H}$, then $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$ is a norm induced by the inner product.

Definition 6 *(Hilbert space)*
*If $\langle\cdot,\cdot\rangle$ is an inner product on $\mathcal{H}$, the pair $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ is called a **Hilbert space** if $\mathcal{H}$ with metric induced by the inner product is complete*[1].

Definition 7 *(Linear operator)*
*for two vector spaces $\mathcal{H}$ and $\mathcal{H}'$ over $\mathbb{R}$, a map $A : \mathcal{H}\to\mathcal{H}'$ is called **a linear operator** if it satisfies $A(\alpha f) = \alpha A(f)$ and $A(f_1+f_2) = A(f_1)+A(f_2)$ for all $\alpha\in\mathbb{R}, f_1, f_2\in\mathcal{H}$. Throughout the rest of the chapter we use standard notational convention $Af := A(f)$.*

The following three conditions can be proven to be equivalent:

i) linear operator A is continuous;

ii) A is continuous at 0;

iii) A is bounded[2].

This result along with the Riesz representation theorem given later, are fundamental for the theory of Reproducing Kernel Hilbert Spaces. It should be emphasised that while the operators we use, such as mean element and cross-covariance operator, are linear, the functions they operate on will not in general

[1] A metric space is complete if every Cauchy sequence converges in this space.
[2] A bounded linear operator is generally not a bounded function.

be linear. An important special case of linear operators are the linear functionals, which are operators $A : \mathcal{H} \to \mathbb{R}$.

Theorem 1 *(Riesz representation theorem)*

In a Hilbert space $\mathcal{H}$, all continuous linear functionals[3] *A can be written in the form $\langle \cdot, f_A \rangle$, for some $f_A \in \mathcal{H}$, so that:*

$$Af = \langle f, f_A \rangle_{\mathcal{H}} \tag{3.2}$$

One of the main mathematical concepts that will be used in this thesis, is a **Mercer kernel**, also known under the names: positive definite kernel, covariance function, reproducing kernel, admissible kernel, Support Vector kernel, nonnegative definite kernel. Mercer kernels were already referred to in Chapters (1 - 2), used in methods extending linear algorithms to nonlinear ones – called "kernelisation" or a "kernel trick", and in methods of studying independence based on Hilbert spaces called reproducing kernel Hilbert spaces (RKHSs). Below we will formally introduce Mercer kernels:

Definition 8 *(Mercer kernel)*

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a Mercer kernel if and only if it is symmetric, that is, $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$ and semi-positive definite, that is

$$\forall \mathbf{x}_1, ..., \mathbf{x}_n \in X \quad \forall c_1, ..., c_n \in \mathbb{R} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geqslant 0. \tag{3.3}$$

Remark 1 *The term kernel has been initially used in the context of integral operators. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which gives rise to an operator A_k according to:*

$$(A_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot, \mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \tag{3.4}$$

is called the kernel of A_k.

Lemma 1 *(Sums of kernels are kernels)*

Let k, k_1, k_2 are Mercer kernels on $\mathcal{X}$ and $\alpha > 0$, then αk, and $k_1 + k_2$ are kernels on $\mathcal{X}$.

Lemma 2 *(Products of kernels are kernels)*

Let Let k_1, k_2 be Mercer kernels on, respectively, $\mathcal{X}_1$ and $\mathcal{X}_2$, then $k_1 \times k_2$ is a Mercer kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_1$, then $k = k_1 \times k_2$ is a kernel on $\mathcal{X}$.

When defining Mercer kernels, we introduce the notion of a semi-positive definite function, even though in practice we will typically be interested in Mercer kernels that are strictly positive definite. It

[3] If H is a normed space, the space of all continuous linear functionals $A : \mathcal{H} \to \mathbb{R}$ is called a topological dual space of $\mathcal{H}$ and denoted as $\mathcal{H}'$.

will come as useful to introduce the following necessary condition for a function to be strictly positive definite.

Lemma 3 *Added*

Let ϕ, ψ be positive functions defined on an interval $I \in \mathbb{R}$ with ϕ/ψ strictly increasing. Set:

$$k(t,s) = \begin{cases} \phi(s)\psi(t) & s \leq t, \\ \phi(t)\psi(s) & t < s, \end{cases} \tag{3.5}$$

and assume that ϕ and ψ are such, that for all $s, t \in I$ $\phi(s)\psi(t) > 0$. Then $k(s,t)$ is a strictly positive definite function on $I \times I$.

There are several important properties of kernels. A centered Gaussian process is uniquely determined by its covariance function (semi-positive definite kernel). Conversely, any semi-positive kernel defines a covariance function and a unique centered Gaussian Process Hein and Bousquet [2004]. Moreover there exists a bijection between the set of all real-valued semi-positive kernels on some space $\mathcal{X}$ and the set of all centered Gaussian processes defined on $\mathcal{X}$. Kernels can also be seen as inner products Schölkopf et al. [2004]:

Theorem 2 *For any kernel k on space $\mathcal{X}$, there exists a Hilbert space F and a mapping $\phi : \mathcal{X} \to F$ such that:*

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle, \quad \text{for any } \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \tag{3.6}$$

where $\langle u, v \rangle, u, v \in \mathcal{H}$ represents an inner product in $\mathcal{H}$.

The above theorem shows how we can create a kernel provided we have a feature map. Because the simplest feature map is an identity map, this theorem proves that an inner product is a kernel.

Definition 9 *(Stationary kernel)*

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called stationary, if $k(\mathbf{x}, \mathbf{x}')$ is a function of $\mathbf{x} - \mathbf{x}'$.

Definition 10 *(Isotropic kernel)*

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called isotropic, if $k(\mathbf{x}, \mathbf{x}')$ is a function of $\|\mathbf{x} - \mathbf{x}'\|$.

Another important concept that will be broadly used in the context of kernel classes is the concept of **Automatic Relevance Determination** (ARD). It has been initially introduced by MacKay [1996], Neal [2012] as a Bayesian model where input relevance can be introduced and controlled with parameters. This has later become popular in a wider context of feature selection and sparse learning in Bayesian models Qi et al. [2004]. We use the same concept, but for a purpose of ensuring we have nested models for inference hypothesis design (see section 4.3), and it will be crucial when applying the Generalised Likelihood Ratio Test.

In the ARD model each input variable has an associated hyperparameter whose value can scale the effect of that input. In the Bayesian setting, this is achieved by setting a separate Gaussian prior for each of the inputs. In our case we treat each dimension as a separate input and define our mean and covariance functions in such manner that the effect of each of the univariate inputs can be separately changed with hyperparameters. In particular, by setting specific values of the hyperparameters we can practically eliminate some of the univariate variables from the mean/covariance. In the table below (Table 3.1) are two examples of popular kernels and their ARD versions. Rasmussen and Williams in their Matlab toolbox provide ARD versions of the squared exponential and Matern kernels, with one lengthscale parameter for each dimension of the input space $diag\left(\left[l_1^{-2}, ..., l_n^{-2}\right]\right)$, but our version from Table 3.1 allows us to choose $l_i = 0$ which removes the effect of the i-th dimension of input on the kernel. And so a covariance for lower dimensional space can be expressed as a covariance with a higher dimensional space $k^{SE}([\mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}], [\mathbf{Y}_{t'-1}, \mathbf{Z}_{t'-1}]) = k^{SE}([\mathbf{X}_{t-1}, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}], [\mathbf{X}_{t'-1}, \mathbf{Y}_{t'-1}, \mathbf{Z}_{t'-1}]; l_1 = 0)$.

Remark 2 *Popular kernels, such as squared exponential or Matern (see Table 3.1) use* ***lengthscale*** *parameter l and entering the covariance function as a multiplicative factor l^{-2}, or $diag\left(\left[l_1^{-2}, ..., l_n^{-2}\right]\right)$ for an ARD kernel. Such lengthscale parameter can be interpreted in terms of a smoothing effect, with large values l meaning that a variability is more likely to be attributed to noise, while in an ARD case, large value of l_i imply i-th factor or input explaining less of the variability than other factors or inputs. However, when we introduce definition of causality and of nested models in Chapter 5, we will require the ability for the hyperparameters responsible for scaling the input to be equal zero. When using the squared exponential ARD or Matern ARD kernel, this means defining the hyperparameter l as a reciprocal of the typical lengthscale.*

Given a set of input points $\{\mathbf{x}_i, i = 1, ..., n\}$ we can compute the Gram matrix K whose entries are $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. If k is a covariance function we call the matrix K the covariance matrix, or kernel matrix.

Definition 11 *(Reproducing kernel Hilbert space (RKHS)) [Scholkopf and Smola, 2001]*
Consider a Hilbert space $\mathcal{H}$ of real-valued functions on any set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:

1. *k has a reproducing property:*

$$\langle f, k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = f(\mathbf{x}), \text{ for all } f \in \mathcal{H}; \tag{3.7}$$

in particular

$$\langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot) \rangle_{\mathcal{H}} = k(\mathbf{x}, \mathbf{x}'). \tag{3.8}$$

covariance function	expression $k(x_p, x_q) =$	stationary
constant (noise)	$\sigma^2 \delta_{p,q}$	+
linear	$\mathbf{x}_p^T \mathbf{x}_q$	-
linear ARD	$\mathbf{x}_p^T \mathbf{A} \mathbf{x}_q$	-
polynomial	$\sigma_f^2 \left(const + \mathbf{x}_p^T \mathbf{x}_q \right)^m$	-
squared exponential	$\sigma_f^2 \exp\left(-\frac{(\mathbf{x}_p - \mathbf{x}_q)^T (\mathbf{x}_p - \mathbf{x}_q)}{2l^2} \right)$	+
squared exponential ARD	$\sigma_f^2 \exp\left(-\frac{1}{2} D \operatorname{diag}\left(\left[l_1^{-2}, ..., l_n^{-2} \right] \right) D^T \right)$	+
squared exponential* ARD	$\sigma_f^2 \exp\left(-\frac{1}{2} D \operatorname{diag}\left(\left[l_1^{2}, ..., l_n^{2} \right] \right) D^T \right)$	+
Matern	$\frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu} \frac{\mathrm{d}}{l} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{\mathrm{d}}{l} \right)$	+
Matern ARD	$\frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu} \mathbf{D} \left[l_1^{-1}, ..., l_n^{-1} \right]^T \right)^{\nu} K_\nu \left(\sqrt{2\nu} \mathbf{D} \left[l_1^{-1}, ..., l_n^{-1} \right]^T \right)$	+
Matern* ARD	$\frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu} \mathbf{D} \left[l_1, ..., l_n \right]^T \right)^{\nu} K_\nu \left(\sqrt{2\nu} \mathbf{D} \left[l_1, ..., l_n \right]^T \right)$	+
periodic	$\sigma^2 \exp\left(-\frac{2 \sin(\Pi(\mathbf{x}_q - \mathbf{x}_p)/a)}{l^2} \right)$	+
separable nonstationary	$k_1(\mathbf{x}_p, \mathbf{x}_p) k_2(\mathbf{x}_q, \mathbf{x}_q)$	-

Table 3.1: Summary of several popular kernel functions. In the linear - ARD function, $\mathbf{A}$ is a matrix, in the polynomial kernel m denotes a constant. In the Matern function d represents a distance $\mathrm{d} = \| \mathbf{x}_p - \mathbf{x}_q \|$, most often an Euclidean distance. In the modified Matern - ARD, we use the lengthscale to introduce weighting separately for each of the dimensions and therefore we express the distance differently, by defining D to be a vector of univariate "distances" $\mathbf{D} = \left[\| x_{p,1} - x_{q,1} \|, ..., \| x_{p,n} - x_{q,n} \| \right]$. In the periodic kernel a is a constant, and in the separable nonstationary kernel the functions k_1, k_2 are some stationary kernels.
(*) refers to alternative formula for the covariance function, where the typical lengthscale parameter is replaced by reciprocal lengthscale parameter.

2. *k spans $\mathcal{H}$, i.e. $\mathcal{H} = \overline{span\{k(\mathbf{x}, \cdot) \mid \mathbf{x} \in \mathcal{X}\}}$, where $\overline{X}$ denotes the completion of the set X.*

RKHS uniquely determines the kernel function k, and according to Moore-Aronszajn theorem the opposite direction is also true:

Theorem 3 *(Moore-Aronszajn)[Aronszajn, 1950]*

If k is a Mercer kernel then there exists $\mathcal{H}$ – a unique RKHS whose kernel is k.

The Moore-Aronszajn theorem proves existence of RKHS associated with a Mercer kernel without showing how to construct such an RKHS. The next theorem, proposed over hundred years ago by Mercer [1909], provides a series representation for continuous kernels on compact domains, now called Mercer kernels, and describes the corresponding RKHS.

Theorem 4 *(Mercer theorem) [Mercer, 1909, Scholkopf and Smola, 2001]*

Suppose $k \in L^\infty(\mathcal{X}^2)$ is a symmetric real-valued function such that an integral operator A_k

$$A_k : L^p(\mathcal{X}) \to L^p(\mathcal{X}), \qquad (A_k f)(\cdot) := \int_{\mathcal{X}} k(\cdot, \mathbf{x}) f(\mathbf{x}) dx \tag{3.9}$$

is positive definite, that is for all $f \in L^2(\mathcal{X})$

$$\int_{\mathcal{X}^2} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0. \tag{3.10}$$

Let $\psi_j \in L^2(\mathcal{X})$ *be the normalised orthogonal eigenfunctions of* A_k *associated with eigenvalues* $\lambda_j > 0$*, sorted in non-increasing order. Then:*

1. $(\lambda_j)_{j=1}^{\infty} \in l^1$*, meaning that the eigenvalues are absolutely summable;*
2. $k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \psi_j(\mathbf{x}) \psi_j(x')$ *holds for almost all* $(\mathbf{x}, \mathbf{x}')$*, where the series converges absolutely and uniformly for almost all* $(\mathbf{x}, \mathbf{x}')$*.*

Introducing kernel ridge regression method in the previous sections in the Equations (1.46 - 1.49), we used kernel trick without explaining why it was permissible. The explanation is given below as the representer theorem[4]. The theorem refers to a loss function $L(\mathbf{x}, \mathbf{y}, f(\mathbf{x}))$ that describes the cost of the discrepancy between the prediction $f(\mathbf{x})$ and the observation $\mathbf{y}$ at the point $\mathbf{x}$. The risk $\mathcal{R}_{L,S}$ associated with the loss L and data sample S is defined as the average future loss of the prediction function f.

Theorem 5 *(Representer theorem)[Steinwart and Christmann, 2008]*
Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ *be a convex loss,* $S := \{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ *be a set of observations and* $\mathcal{R}_{L,S}$ *denote associated risk. Furthermore, let* $\mathcal{H}$ *be an RKHS over* $\mathcal{X}$*. Then for all* $\lambda > 0$ *there exists a unique estimator which we denote by* $f_{S,\lambda} \in \mathcal{H}$ *satisfying the equality:*

$$f_{S,\lambda} = \arg\inf_{f \in \mathcal{H}} \lambda \|f\|_{\mathcal{H}}^2 + \mathcal{R}_{L,S}(f) \tag{3.11}$$

In addition, there exist $\alpha_1, \cdots \alpha_n \in \mathbb{R}$ *such that for k being the reproducing kernel associated with* $\mathcal{H}$

$$f_{S,\lambda}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i), \quad \text{for } \mathbf{x} \in \mathcal{X}. \tag{3.12}$$

The representation theorem is especially relevant, because it states that not only a solution to the minimisation problem exists and is unique, but it can be represented with Mercer kernels, as shown in Equation (3.12).

Below we present definitions which are building blocks of the Hilbert Schmidt normalized conditional independence criterion.

Definition 12 *(Hilbert-Schmidt norm)*
Let $\mathcal{H}$ *be a Reproducing Kernel Hilbert Space (RKHS) of functions from* $\mathcal{X}$ *to* $\mathbb{R}$*, induced by strictly positive kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*. Let* $\mathcal{H}'$ *be an RKHS of functions from* $\mathcal{Y}$ *to* $\mathbb{R}$*, induced by strictly positive*

[4] In some machine learning literature kernel trick is introduced via Mercers theorem.

kernel $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$[5]. Denote by $C : \mathcal{H}' \to \mathcal{H}$ a linear operator. The Hilbert-Schmidt norm of the operator C is defined as

$$\|C\|_{HS}^2 := \sum_{i,j} \langle Cv_i, u_j \rangle_{\mathcal{H}}^2, \tag{3.13}$$

given that the sum converges, where u_i and u_j are orthonormal bases of $\mathcal{H}$ and $\mathcal{H}'$ respectively; $\langle v, u \rangle_{\mathcal{H}}, u, v \in \mathcal{H}$ represents an inner product in $\mathcal{H}$

Following Gretton et al. [2005], Sun [2008], let $\mathcal{H}_{\mathcal{W}}$ denote the RKHS induced by a strictly positive kernel $k_{\mathcal{W}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$. Let $\mathbf{X}$ be random vector on $\mathcal{X}$, $\mathbf{Y}$ be random vector on $\mathcal{Y}$ and $(\mathbf{X}, \mathbf{Y})$ be random vector on $\mathcal{X} \times \mathcal{Y}$. We assume $\mathcal{X}$ and $\mathcal{Y}$ are topological spaces and the measurability is defined with respect to the relevant $\sigma-$fields. The marginal distributions are denoted by F_X, F_Y and the joint distribution of (X, Y) by F_{XY}. The expectations $\mathbf{E}_X$, $\mathbf{E}_Y$ and $\mathbf{E}_{XY}$ denote the expectations over F_X, F_Y and F_{XY}, respectively. To ensure $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}$ are included in, respectively $L^2(P_X)$ and $L^2(P_Y)$, we consider only random vectors (X, Y) such that the expectations $\mathbf{E}_X[k_{\mathcal{X}}(X, X)]$ and $\mathbf{E}_Y[k_{\mathcal{Y}}(Y, Y)]$ are finite.

Definition 13 *(Hilbert-Schmidt operator) A linear operator $C : \mathcal{H}' \to \mathcal{H}$ is Hilbert-Schmidt if its Hilbert-Schmidt norm exists.*

The set of Hilbert-Schmidt operators $HS(\mathcal{H}', \mathcal{H}) : \mathcal{H}' \to \mathcal{H}$ is a separable Hilbert space with the inner product:

$$\langle C, D \rangle_{HS} := \sum_{i,j} \langle Cv_i, u_j \rangle_{\mathcal{H}} \langle Dv_i, u_j \rangle_{\mathcal{H}}, \tag{3.14}$$

where $C, D \in HS(\mathcal{H}', \mathcal{H})$.

Definition 14 *(Tensor product) Let $f \in \mathcal{H}$ and $g \in \mathcal{H}'$, then the tensor product operator $f \otimes g : \mathcal{H}' \to \mathcal{H}$ is defined as follows:*

$$(f \otimes g)h := f\langle g, h \rangle_{\mathcal{H}'}, \quad \textit{for all } h \in \mathcal{H}'. \tag{3.15}$$

The definition above makes use of two standard notational abbreviations. The first one concerns omitting brackets when denoting application of an operator: $(f \otimes g)h$ instead of $(f \otimes g)(h)$. The second one relates to multiplication by a scalar and we write $f\langle g, h \rangle_{\mathcal{H}'}$ instead of $f \cdot \langle g, h \rangle_{\mathcal{H}'}$.

The Hilbert-Schmidt norm of the tensor product can be calculated as:

$$\begin{aligned} \|f \otimes g\|_{HS}^2 &= \langle f \otimes g, f \otimes g \rangle_{HS} = \langle f, (f \otimes g)g \rangle_{\mathcal{H}} \\ &= \langle f, f \rangle_{\mathcal{H}} \langle g, g \rangle_{\mathcal{H}'} = \|f\|_{\mathcal{H}}^2 \|g\|_{\mathcal{H}'}^2. \end{aligned} \tag{3.16}$$

When introducing the cross-covariance operator we will be using the following results for the tensor

[5] We will require that space $\mathcal{H}$ must be separable (to have a complete orthonormal system), but in practice we will use $\mathbb{R}^n$ and therefore that would not be an issue.

product. Given a Hilbert-Schmidt operator $L : \mathcal{H}' \to \mathcal{H}$ and $f \in \mathcal{H}$ and $g \in \mathcal{H}'$,

$$\langle L, f \otimes g \rangle_{HS} = \langle f, Lg \rangle_{\mathcal{H}}. \tag{3.17}$$

A special case of Equation (3.17) with the notation as earlier and $u \in \mathcal{H}$ and $v \in \mathcal{H}'$,

$$\langle f \otimes g, u \otimes v \rangle_{HS} = \langle f, u \rangle_{\mathcal{H}} \langle g, v \rangle_{\mathcal{H}'}. \tag{3.18}$$

Definition 15 *(The mean element)*
The mean element μ_X with respect to the probability measure P_X is defined as an element of the RKHS $\mathcal{H}_\mathcal{X}$ for which

$$\langle \mu_X, f \rangle_{\mathcal{H}_\mathcal{X}} := \boldsymbol{E}_X[\langle \phi(\mathbf{X}), f \rangle_{\mathcal{H}_\mathcal{X}}] = \boldsymbol{E}_X[f(\mathbf{X})], \tag{3.19}$$

where $\phi : \mathcal{X} \to \mathcal{H}_\mathcal{X}$ is a feature map and $f \in \mathcal{H}_\mathcal{X}$.

The mean elements exist as long as the respective norms are bounded, a condition that is met if the relevant kernels are bounded.

The cross-covariance operator is analogous to a covariance matrix, but is defined for feature maps.

Definition 16 *(Cross-covariance operator) The cross-covariance operator is a linear operator $\Sigma_{XY} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ associated with the joint measure P_{XY} defined as*

$$\Sigma_{XY} := \boldsymbol{E}_{XY}[(\phi(\mathbf{X}) - \mu_X) \otimes (\phi(\mathbf{Y}) - \mu_Y)] = \boldsymbol{E}_{XY}[\phi(\mathbf{X}) \otimes \phi(\mathbf{Y})] - \mu_X \otimes \mu_Y, \tag{3.20}$$

where we use symbol $\otimes$ for tensor product and μ for mean embedding. The cross-covariance operator applied to two elements $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$ gives the covariance:

$$\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_\mathcal{X}} = Cov(f(\mathbf{X}), g(\mathbf{Y})). \tag{3.21}$$

The notation and assumptions follow Gretton et al. [2005], Sun [2008]: $\mathcal{H}_\mathcal{X}$ denotes the Reproducing Kernel Hilbert Space (RKHS) induced by a strictly positive kernel $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, analogously for $\mathcal{H}_\mathcal{Y}$ and $k_\mathcal{Y}$. $\mathbf{X}$ is a random variable on $\mathcal{X}$, $\mathbf{Y}$ is a random variable on $\mathcal{Y}$ and $(\mathbf{X}, \mathbf{Y})$ is a random vector on $\mathcal{X} \times \mathcal{Y}$. We assume $\mathcal{X}$ and $\mathcal{Y}$ to be topological spaces and measurability is defined with respect to the relevant $\sigma-$fields. The marginal distributions are denoted by F_X, F_Y and the joint distribution of $(\mathbf{X}, \mathbf{Y})$ by F_{XY}. The expectations $\mathbf{E}_X$, $\mathbf{E}_Y$ and $\mathbf{E}_{XY}$ denote the expectations over P_X, P_Y and P_{XY}, respectively. To ensure $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}$ are included in, respectively, $L^2(P_X)$ and $L^2(P_Y)$, we consider only random vectors (X, Y) such that the expectations $\mathbf{E}_X[k_\mathcal{X}(X, X)]$ and $\mathbf{E}_Y[k_\mathcal{Y}(Y, Y)]$ are finite.

Definition 17 *(Hilbert-Schmidt Independence Criterion – HSIC)[Gretton et al., 2005]*

We define the Hilbert Schmidt independence criterion as the squared Hilbert Schmidt norm of the cross-covariance operator Σ_{XY}:

$$HSIC(P_{XY}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}) := \|\Sigma_{XY}\|^2_{HS}. \tag{3.22}$$

The HSIC can be written in the kernel notation:

$$\begin{aligned} HSIC(P_{XY}, \mathcal{F}_\mathcal{X}, \mathcal{F}_\mathcal{Y}) :=& \boldsymbol{E}_{X,X',Y,Y'}[k_\mathcal{X}(X,X')k_\mathcal{Y}(Y,Y')] + \boldsymbol{E}_{X,X'}[k_\mathcal{X}(X,X')]\boldsymbol{E}_{Y,Y'}[k_\mathcal{Y}(Y,Y')] \\ & -2\boldsymbol{E}_{X,Y}[\boldsymbol{E}_{X'}[k_\mathcal{X}(X,X')]\boldsymbol{E}_{Y'}[k_\mathcal{Y}(Y,Y')]]. \end{aligned} \tag{3.23}$$

(X, Y) and (X', Y') are i.i.d. from P_{XY}.

Just as the cross-covariance operator is related to the covariance, we can define an operator that is related to partial correlation:

Definition 18 *(Normalised conditional cross-covariance operator Fukumizu et al. [2008]) Using the cross-covariance operators we can define the normalised conditional cross-covariance operator in the following way:*

$$V_{XY|Z} = \Sigma_{XX}^{-1/2}(\Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1/2}\Sigma_{ZY})\Sigma_{YY}^{-1/2}. \tag{3.24}$$

Gretton et al. [2005] state that for rich enough RKHS [6], zero norm of the cross-covariance operator is equivalent to independence, which can be written as:

$$\mathbf{X} \perp\!\!\!\perp Y \iff \Sigma_{XY} = 0, \tag{3.25}$$

where 0 denotes a null operator. This equivalence is the premise from which follows the use of the Hilbert-Schmidt independence criterion (HSIC) as a measure of independence.

It was shown in Fukumizu et al. [2008] that there is a relationship similar to (3.25) between the normalised conditional cross-covariance operator and conditional independence, which can be written as:

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \iff V_{(XZ)(YZ)|Z} = 0, \tag{3.26}$$

where by (YZ) and (XZ) we denote extended variables. Therefore the Hilbert-Schmidt norm of the conditional cross-covariance operator has been suggested as a measure of conditional independence. Using the normalised version of the operator has the advantage that it is less influenced by the marginals than the non-normalised operator while retaining all the information about dependence. This is analogous to the difference between correlation and covariance.

Definition 19 *(Hilbert Schmidt normalised conditional independence criterion – HSNCIC) We define the*

[6] By "rich enough" we mean universal, i.e. dense in the sense of continuous functions on $\mathcal{X}$ with the supremum norm Hofmann et al. [2008].

HSNCIC as the squared Hilbert Schmidt norm of the normalised conditional cross-covariance operator $V_{(XZ)(YZ)|Z}$*:*

$$HSNCIC := \|V_{(XZ)(YZ)|Z}\|^2_{HS}, \tag{3.27}$$

where $\|\cdot\|_{HS}$ *denotes Hilbert-Schmidt norm of an operator.*

For the sample $S = \{(\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1), ..., (\mathbf{x}_n, \mathbf{y}_n, \mathbf{z}_n)\}$ HSNCIC has an estimator that is both straightforward and has good convergence behaviour [Fukumizu et al., 2008, Seth and Principe, 2011]. As shown in Appendix A.7, it can be obtained by defining empirical estimates of all of the components in following steps: first define mean elements $\hat{m}_X^{(n)}$ and $\hat{m}_Y^{(n)}$ and use them to define empirical cross-covariance operator $\hat{\Sigma}_{XY}^{(n)}$. Subsequently using $\hat{\Sigma}_{XY}^{(n)}$, together with $\hat{\Sigma}_{XX}^{(n)}$ and $\hat{\Sigma}_{YY}^{(n)}$ obtained in the same way, define $\hat{V}_{XY}^{(n)}$ for the empirical normalised cross-covariance operator. Note that V_{XY} requires inverting Σ_{YY} and Σ_{XX}, hence to ensure invertibility a regulariser $n\lambda I_n$ is added. The next step is to construct the estimator $\hat{V}_{XY|Z}^{(n)}$ from $\hat{V}_{XY}^{(n)}$, $\hat{V}_{XZ}^{(n)}$ and $\hat{V}_{ZY}^{(n)}$. Finally, construct the estimator of the Hilbert-Schmidt norm of $\hat{V}_{ZY}^{(n)}$ as follows:

$$HSNCIC_n := Tr[R_{(XZ)}R_{(YZ)} - 2R_{(XZ)}R_{(YZ)}R_Z + R_{(XZ)}R_ZR_{(YZ)}R_Z], \tag{3.28}$$

where Tr denotes a trace of a matrix, and $R_U = K_U(K_U + n\lambda I)^{-1}$ and $K_U(i, j) = k(u_i, u_j)$ is a Gram matrix. This estimator depends on the regularisation parameter λ which, in turn, depends on the sample size. Regularisation becomes necessary when inverting finite rank operators.

3.1.2 Defining Gaussian Processes

We begin with the definition popular in the field of Machine Learning, given by Williams and Rasmussen [2006].

Definition 20 *Gaussian Process (GP)[Williams and Rasmussen, 2006]*
A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Gaussian distribution is characterised by its mean and covariance, and in the Definition 21 we will show how a GP can be defined by analogous characterisation – by its mean function and covariance function. Below, we show how Gaussian process can be thought of as a Markov process, for which all finite dimensional distributions are Gaussian. Such a representation requires assumption on the covariance function, and without it one can have an infinite memory GP.

Lemma 4 *Gaussian Process (GP)*
Let $I \in \mathbb{R}$ *be an interval, open or closed, and let* $\{X_t\}$ *s.t.* $t \in I$ *be a mean zero GP with continuous strictly positive covariance k. Then* $\{X_t\}$ *is a Gaussian Markov process, i.e. for all increasing sequences*

$t_1, ..., t_n \in \mathbb{R}_+$, for all $n \in \mathbb{N}$

$$\boldsymbol{E}\left[X_{t_n} \mid X_{t_{n-1}}\right] = \boldsymbol{E}\left[X_{t_n} \mid X_{t_{n-1}}, ..., X_{t_1}\right]$$

if and only if k can be expressed as in Equation (3.5).

A third definition of a GP that we provide, shows how a GP can be defined through a characterisation.

Definition 21 *Characterisation of Gaussian Process (GP)*

Denote by $f(\mathbf{x}) : \mathcal{X} \mapsto \mathbb{R}$ a stochastic process parametrised by $\{\mathbf{x}\} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$. Then, the random function $f(\mathbf{x})$ is a Gaussian process if all its finite dimensional distributions are Gaussian, where for any $n \in \mathbb{N}$, the random vector $(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))$ is jointly normally distributed, see Williams and Rasmussen [2006].

We can therefore interpret a GP as formally defined by the following class of random functions:

$$\begin{aligned} f := \{ f(\cdot) : \mathcal{X} \mapsto \mathbb{R}, \text{ s.t. } f(\cdot) \sim \mathcal{GP}\left(\mu\left(\cdot;\boldsymbol{\theta}_\mu\right), k\left(\cdot,\cdot;\boldsymbol{\theta}_k\right)\right), \text{ with} \\ \mu\left(\cdot;\boldsymbol{\theta}_\mu\right) := \mathbb{E}\left[f(\cdot)\right] : \mathcal{X} \mapsto \mathbb{R}, \\ k\left(\cdot,\cdot;\boldsymbol{\theta}_k\right) := \mathbb{E}\left[\left(f(\cdot) - \mu\left(\cdot;\boldsymbol{\theta}_\mu\right)\right)\left(f(\cdot) - \mu\left(\cdot;\boldsymbol{\theta}_\mu\right)\right)\right] : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+ \Big\}. \end{aligned}$$

At each point the mean of the function is $\mu(\cdot;\boldsymbol{\theta}_\mu)$, parametrised by $\boldsymbol{\theta}_\mu$, and the spatial dependence between any two points is given by the covariance function (Mercer kernel) $k(\cdot,\cdot;\boldsymbol{\theta}_k)$, parametrised by $\boldsymbol{\theta}_k$, see detailed discussion in Williams and Rasmussen [2006]. We will later use notation $\theta = \theta_\mu \cup \theta_k$, and will refer to θ as hyperparameters of the Gaussian Process f.

3.1.2.1 Gaussian Processes Time Series

GPs are a flexible class of models, and can be successfully used for modelling distributions over functions with space or time domain. In the Definition (21) we introduced the notation $f(\mathbf{x}) : \mathcal{X} \mapsto \mathbb{R}$ a stochastic process parametrised by $\{\mathbf{x}\} \in \mathcal{X}$. There are different approaches to how they can be used for modelling time series, and in particular autoregressive time series, see [Roberts et al., 2012], [Requeima et al., 2019], [Candela et al., 2003].

A popular way of defining a GP time series is by using the index t as the input, and the time series observation as an output, in which case $\mathcal{X} \equiv \mathbb{R}$, $\mu : \mathbb{R} \to \mathbb{R}$, and $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

$$X_t = f(t) + \epsilon_t, \qquad f \sim \mathcal{GP}(\mu, k), \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{3.29}$$

In the context of statistical causality, we are typically interested in **autoregressive time series**, where what constitutes the input space is not time, but past observations. Let $\{X_t\}$ be two time series, and their past observations be denoted as $\mathbf{X}_{t-1}^{-k} = [X_{t-k}, \cdots, X_{t-1}]$. We nonparametericallly model the time series

$\{X_t\}$ as realizations from a GPs with additive Gaussian noise:

$$X_t = f\left(\mathbf{X}_{t-1}^{-k}\right) + \epsilon_t, \qquad f \sim \mathcal{GP}\left(\mu, k\right), \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{3.30}$$

Remark 3 *The autoregressive formulation from Equation 3.30 is more general than the Equation 3.29, to which it can be converted. The autoregressive formulation is, however,* ***not time reversible****, which is consistent with the fact that when modelling statistical causality we want to describe a property that is not symmetrical in terms of time.*

Remark 4 *The formulation from Equation 3.30 can be extended to include past values of an additional time series $\{Y_t\}$ as the input space:*

$$X_t = f\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}\right]\right) + \epsilon_t, \qquad f \sim \mathcal{GP}\left(\mu, k\right), \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{3.31}$$

Remark 5 *Please note that given the fact that sum of kernels is a kernel (Lemma 1) the additive Gaussian noise ϵ_t can be either explicitly stated as in the Equations 3.29 - 3.31 or represented as a GP itself and included in the formulation of the GP f.*

3.1.3 Multiple Output Gaussian Processes for Time Series

Gaussian Processes are typically used to model only a single output variable. When one wants to use Gaussian processes to cater for multiple outputs, the main difficulty that arises is how to define covariance functions that capture cross-covariances and still guarantees positive-definite covariance matrices. Some authors suggest modelling each output as a separate, independent Gaussian Process ("multi-kriging" [Williams and Rasmussen, 1996]). There are several approaches to achieving multiple output GP, such as "cokriging" (see Cressie 1993 [Cressie, 1993] and literature therein), modelling the outputs as linear combinations of latent channels [Teh et al., 2005], [Micchelli and Pontil, 2005], or modelling them as convolutions of the same underlying white noise process [Boyle and Frean, 2005], [Álvarez and Lawrence, 2011].

We can think of a multi-output GP in terms of describing marginal and joint distributions, and so on the most general level, a multi-output GP representation for two time series, $\{X_t\}, \{Y_t\}$, could be described as follows:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f_X\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) + \epsilon_t^X \\ f_Y\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) + \epsilon_t^Y \end{bmatrix} \quad \begin{matrix} \text{marginal } f_X \sim \mathcal{GP}(\mu^X, k^X);\ \epsilon_t^X \sim \mathcal{N}(0, \sigma_X^2) \\ \text{marginal } f_Y \sim \mathcal{GP}(\mu^Y, k^Y);\ \epsilon_t^Y \sim \mathcal{N}(0, \sigma_Y^2) \end{matrix} \tag{3.32}$$

where $X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}$ are the lagged observations of X_t, Y_t and the side information Z_t, with k, l, m being their respective lags.

Equation 3.32 can resemble the way how copulas are used for modelling joint distributions of multivariate random variables, and we indeed will introduce copulas as one way of expressing symmetrical dependence in out multivariate time series models (see Section 3.3 and Chapter 6 for more details).

We can write the covariance matrix of the Eqution 3.32 as follows:

$$\mathbf{\Sigma} = \begin{bmatrix} k^X & k^{XY} \\ k^{YX} & k^Y \end{bmatrix} + \begin{bmatrix} \delta_{t,s}\sigma_X^2 & 0 \\ 0 & \delta_{t,s}\sigma_Y^2 \end{bmatrix},$$

where $\delta_{t,s}$ is a Kronecker delta.

It is not enough for the functions k^{YX} and k^{XY} to be kernels, to ensure that the whole covariance matrix is positive definite, and it is also trivial to give (practical) conditions for the functions k^{YX} and k^{XY} to achieve positive definiteness. There are some special cases though, where this can be done easily. For example, if we choose a kernel $k(\cdot,\cdot)$ and set $k^X = k^{XY} = k^{YX} = k^Y = k$, then the covariance matrix $\mathbf{\Sigma}_t$ will automatically be positive-definite.

Alternatively, one can introduce the dependence structure through the structure of the noise component. The covariance between outputs of the multiple output GP is then equal to:

$$\mathbf{\Sigma} = \begin{bmatrix} k^X & 0 \\ 0 & k^Y \end{bmatrix} + \begin{bmatrix} \delta_{t,s}\sigma_X^2 & \rho \\ \rho & \delta_{t,s}\sigma_Y^2 \end{bmatrix}.$$

Below we talk in more details about two other approaches: using independent Gaussian processes, and already mentioned method of Boyle and Frean 2005 [Boyle and Frean, 2005] to incorporate convolutions of common white noise process.

3.1.3.1 Two Independent Gaussian Processes

The first approach to obtaining multiple outputs with GP models that we discuss is by using two independent Gaussian processes. The trivial solution to specifying cross-covariance is to set this structure to identity matrix, in which case the two GPs are decoupled. It should be pointed out, though, that failing to model the covariance structure can lead to the loss of information about the causal dependence, as this will only be encoded through the structure placed on the inputs.

Assume two (observed) time series $\{\mathbf{X}_t\}, \{\mathbf{X}_t\}$ whose dynamic can be described with two independent GP models analogously to how it was described in the Equations 3.30 - 3.31.

$$X_t = f_X\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}\right]\right) + \epsilon_t^X, \qquad f_X \sim \mathcal{GP}\left(\mu^X, k^X\right) \qquad \epsilon_t^X \sim \mathcal{N}(0, \sigma_X^2) \tag{3.33}$$

$$Y_t = f_Y\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}\right]\right) + \epsilon_t^Y, \qquad f_Y \sim \mathcal{GP}\left(\mu^Y, k^Y\right) \qquad \epsilon_t^Y \sim \mathcal{N}(0, \sigma_Y^2) \tag{3.34}$$

Remark 6 *One natural way how a source of symmetrical dependence can be incorporated in the multi-*

output GP model with independent GPs is via addition of another time series $\{\mathbf{Z}_t\}$

$$X_t = f_X\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right) + \epsilon_t^X, \qquad f_X \sim \mathcal{GP}\left(\mu^X, k^X\right) \qquad \epsilon_t^X \sim \mathcal{N}(0, \sigma_X^2) \tag{3.35}$$

$$Y_t = f_Y\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right) + \epsilon_t^Y, \qquad f_Y \sim \mathcal{GP}\left(\mu^Y, k^Y\right) \qquad \epsilon_t^Y \sim \mathcal{N}(0, \sigma_Y^2) \tag{3.36}$$

We will refer to $\{\mathbf{Z}_t\}$ *as* ***side information****, and we note that typically* $\{\mathbf{Z}_t\}$ *will be introduced as an observed time series.*

3.1.3.2 Convolved Multiple Output Gaussian Processes

A GP can be expressed as a convolution integral between a smoothing kernel and a latent function. According to Álvarez and Lawrence [2011], any type of latent process can be used, the smoothing kernel has to be absolutely integrable. Here we will describe the case, where as Boyle and Frean [2005] suggested, Gaussian process is constructed by stimulating a linear filter with noise.

A Gaussian Process u^X can be expressed as a convolution integral between a smoothing kernel h_X and an independent white Gaussian noise process u_0:

$$u^X(t) = h_X(t) * u_0(t) = \int_{-\infty}^{\infty} h_X(t-\tau)u_0(\tau)d\tau = \int_{-\infty}^{\infty} h_X(\tau)u_0(t-\tau)d\tau. \tag{3.37}$$

Two dependent time series $\{X_t\}, \{Y_t\}$ can be written as a sum of the independent GPs f_X, f_Y and dependent GPs u^X, u^X, (see Figure 7.3):

$$X(t) = f_X\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right) + u^X(t), \qquad Y(t) = f_Y\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right) + u^Y(t), \tag{3.38}$$

where $f_X \sim \mathcal{GP}\left(\mu^X, k^X\right)$, $f_Y \sim \mathcal{GP}\left(\mu^Y, k^Y\right)$. Additional additive Gaussian noise term can be included in f_X, f_Y to incorporate sources of noise that are not common for $\{X_t\}, \{Y_t\}$.

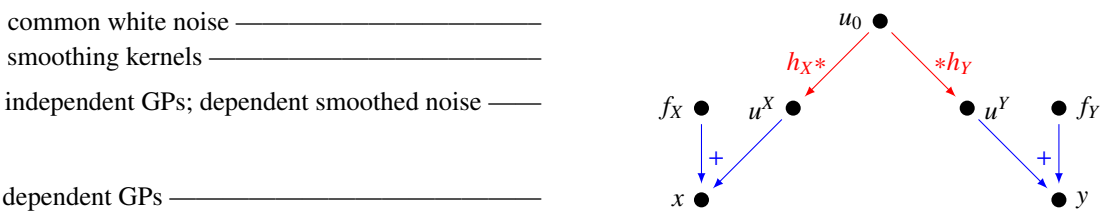


Figure 3.1: How to obtain dependent Gaussian Processes X, Y from independent f_X, f_Y and a common white noise u_0 smoothed by smoothing kernels (linear filters) h_X, h_Y.

If u_0 is a white Gaussian noise process with covariance $k_{u_0}(z, z') = \sigma_u^2 \delta_{z,z'}$, then the covariance

function of u^X is as follows (please note u^X will be zero mean), Boyle and Frean [2005]:

$$\begin{aligned} cov(u^X(t), u^X(t')) &= E\left\{u^X(t), u^X(t')\right\} = \\ &= E\left\{\int_{-\infty}^{\infty} h_X(\tau)u_0(t-\tau)d\tau \int_{-\infty}^{\infty} h_X(\lambda)u_0(t'-\lambda)d\lambda\right\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_X(\tau)h_X(\lambda)E\left[u_0(t-\tau)u_0(t'-\lambda)\right]d\tau d\lambda \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_X(\tau)h_X(\lambda)\sigma_u^2\delta(\lambda-(t'-t+\tau))d\tau d\lambda \\ &= \sigma_u^2\int_{-\infty}^{\infty} h_X(\tau)h_X(t'-t+\tau)d\tau, \end{aligned} \tag{3.39}$$

where we used the fact that the smoothing kernels are absolutely integrable, and so we can change the order of the integration.

The form of the auto- and cross-covariances for the multiple output Gaussian Processes from the Equations 3.38 are as follows:

$$\begin{aligned} cov\left(X_t, X_s\right) &= cov\left(f_X\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right), f_X\left(\left[\mathbf{X}_{s-1}^{-k}, \mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\right]\right)\right) + cov\left(u^X(t), u^X(s)\right) \\ cov\left(Y_t, Y_s\right) &= cov\left(f_Y\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right), f_Y\left(\left[\mathbf{X}_{s-1}^{-k}, \mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\right]\right)\right) + cov\left(u^Y(t), u^Y(s)\right) \\ cov\left(X_t, Y_s\right) &= cov\left(u^X(t), u^Y(s)\right). \end{aligned}$$

The covariance term for $cov\left(u^X(t), u^X(s)\right)$ can be obtained from the Equation 3.39, and analogously for both $cov\left(u^Y(t), u^Y(s)\right)$ and $cov\left(u^X(t), u^Y(s)\right)$, where the Gaussian noise u_0 stays the same, but only the smoothing kernel changes:

$$cov\left(X_t, X_s\right) = k^X + \sigma_u^2\int_{-\infty}^{\infty} h_X(\tau)h_X(s-t+\tau)d\tau \tag{3.40}$$

$$cov\left(Y_t, Y_s\right) = k^Y + \sigma_u^2\int_{-\infty}^{\infty} h_Y(\tau)h_Y(s-t+\tau)d\tau \tag{3.41}$$

$$cov\left(X_T, Y_s\right) = \sigma_u^2\int_{-\infty}^{\infty} h_X(\tau)h_Y(s-t+\tau)d\tau. \tag{3.42}$$

The integrals in the Equations 3.39 and 3.40 do not have a closed form solution for all smoothing kernels. They do have closed form solutions for, for example, squared exponential smoothing kernels h_X, h_Y, defined as:

$$h_X(s) = \theta_X \exp\left(-\frac{1}{2}\frac{(s-\mu^X)^2}{l_X}\right) \qquad h_Y(s) = \theta_Y \exp\left(-\frac{1}{2}\frac{(s-\mu^Y)^2}{l_Y}\right)$$

Following the appendix of Boyle's thesis, the smoothing kernels above lead to:

$$cov\left(u^X(t), u^Y(s)\right) = \sigma_u^2 \frac{(2\Pi)^{\frac{1}{2}}}{\sqrt{|\frac{1}{l_X} + \frac{1}{l_Y}|}} \exp\left(-\frac{1}{2}\frac{\left(t-s-(\mu^X-\mu^Y)\right)^2}{l_X + l_Y}\right) \tag{3.43}$$

Analogously, the autocovariance function:

$$cov\left(u^X(t), u^X(s)\right) = \sigma_u^2 \frac{(2\Pi)^{\frac{1}{2}}}{\sqrt{|\frac{2}{l_Y}|}} \exp\left(-\frac{1}{2}\frac{(t-s)^2}{2l_X}\right). \tag{3.44}$$

In the sequel, we will be using shorthand notation for the covariances from Equations 3.43 and 3.44:

$$\rho^X_{t,s} = cov\left(u^X(t), u^X(s)\right) \qquad \rho^{XY}_{t,s} = cov\left(u^X(t), u^Y(s)\right)$$
$$\rho^Y_{t,s} = cov\left(u^Y(t), u^Y(s)\right) \qquad \rho^{YX}_{t,s} = cov\left(u^Y(t), u^X(s)\right).$$

With the symmetry of the covariance function implying equalities: $\rho^{XY}_{t,s} = \rho^{YX}_{s,t}$ and $\rho^{YX}_{t,s} = \rho^{XY}_{s,t}$.

The mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$ of the Multiple-Output GP, which refer to a joint distribution of random variables $[X_{t_1}, \ldots, X_{t_n}, Y_{t_1}, \ldots, Y_{t_n}]$, can be represented as:

$$\boldsymbol{\mu} := \left[\boldsymbol{\mu}^X, \boldsymbol{\mu}^Y\right] \tag{3.45}$$

$$\mathbf{K} := \begin{bmatrix} \mathbf{K}^{XX} & \mathbf{K}^{XY} \\ \mathbf{K}^{YX} & \mathbf{K}^{YY}, \end{bmatrix} \tag{3.46}$$

where $\mathbf{K} \in \mathcal{X}^+(\mathbb{R}^m)$ and

$$\left\{\mathbf{K}^{XX}\right\}_{ij} = k^X\left(\left[\mathbf{X}^{-k}_{t_i-1}, \mathbf{Y}^{-l}_{t_i-1}, \mathbf{Z}^{-m}_{t_i-1}\right], \left[\mathbf{X}^{-k}_{t_j-1}, \mathbf{Y}^{-l}_{t_j-1}, \mathbf{Z}^{-m}_{t_j-1}\right]\right) + \rho^X_{t_i,j} \tag{3.47}$$

$$\left\{\mathbf{K}^{YY}\right\}_{ij} = k^Y\left(\left[\mathbf{X}^{-k}_{t_i-1}, \mathbf{Y}^{-l}_{t_i-1}, \mathbf{Z}^{-m}_{t_i-1}\right], \left[\mathbf{X}^{-k}_{t_j-1}, \mathbf{Y}^{-l}_{t_j-1}, \mathbf{Z}^{-m}_{t_j-1}\right]\right) + \rho^Y_{t_i,j} \tag{3.48}$$

Note, that the convolved multiple output Gaussian process can be reduced to the case of independent Gaussian processes, by choosing a smoothing kernel that is equivalent to zero. If we choose equal smoothing kernel for both h_X and h_Y, and either choose f_X, f_Y to be always zero, or equivalent to additive noise – then the convolved case can be seen as equivalent to what we called "pseudo-multiple Gaussian process". In the sequel, when introducing the notion of statistical causality using GPs we will use use general notation that allows for multivariate time series and multi-output GPs.

3.2 Illustrative Non-Linear Multi-Variate Time Series Models

In order to motivate the causality studies in this thesis, we consider three illustrative nonlinear time series models that will be references that we will apply our causality testing framework to, throughout the synthetic studies undertaken in the results analysis for testing power, sensitivity and robustness of our proposed causality testing framework.

In particular the classes of model we have chosen as illustrations of data generating processes for the time series that will form inputs to our testing framework characterise a range of general model structures which allow for assessment of linear and non-linear causality structures in the trend or the volatility or both components of the resulting data generating models.

Example Time Series Model Class 1: Structural Trend Based Causality Consider an autoregressive non-linear model class comprised of structures incorporating time series with linear and non-linear polynomial causality in the trend, with Gaussian noise.

$$\begin{aligned} X_t &= a_X X_{t-1} + \epsilon_X & \epsilon_X &\sim \mathcal{N}\left(0, \sigma_X^2\right), \\ Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_Y & \epsilon_Y &\sim \mathcal{N}\left(0, \sigma_Y^2\right), \\ Z_t &= a_Z Z_{t-1} + b_Z Y_{t-1}^q + \epsilon_Z & \epsilon_Z &\sim \mathcal{N}\left(0, \sigma_Z^2\right), \end{aligned} \tag{3.49}$$

The examples that we will use will assume $q = 2$, which means that in the mean this time series will have a non-linear causality in the direction $Y \to Z$, aside from the linear causality $X \to Y$.

We will express the model from the Equations 3.49 in the form of three GPs, as in the Equations 3.50. When generating the data, as Equations 3.52 show, we will use Matern covariance functions for $k_1(t, t')$ and $k_2(t, t')$, we will also extend the model to allow causal relationship in covariance – relationships, that were not existing in the time series formulations from Equations 3.49.

A formulation of the time series from the Equation 3.49 explicitly as GPs can be done according to the following conditional distributions:

$$\begin{aligned} X_t &= f_X(X_{t-1}) & f_X &\sim \mathcal{GP}\left(\mu^X, k^X\right) \\ Y_t &= f_Y([Y_{t-1}, X_{t-1}]) & f_Y &\sim \mathcal{GP}\left(\mu^Y, k^Y\right) \\ Z_t &= f_Z([Z_{t-1}, Y_{t-1}]) & f_Z &\sim \mathcal{GP}\left(\mu^Z, k^Z\right) \end{aligned} \tag{3.50}$$

where the mean functions are linear:

$$\begin{aligned}
&\mu^X(X_{t-1}) = a_X X_{t-1} && \text{no causality} \\
&\mu^Y([Y_{t-1}, X_{t-1}]) = a_Y Y_{t-1} + b_Y X_{t-1} && \text{linear causality} \\
&\mu^Z([Z_{t-1}, Y_{t-1}]) = a_Z Z_{t-1} + b_Z Y_{t-1}^2 && \text{non-linear causality}
\end{aligned} \tag{3.51}$$

and covariance functions incorporate the noise which was already defined as a GP:

$$\begin{aligned}
k^X(X_{t-1}, X_{t'-1}) &= k^{Matern}_{l_a,\sigma_f}(X_{t-1}, X_{t'-1}) + \sigma_n^2 \delta_{t,t'} \\
k^Y([Y_{t-1}, X_{t-1}], [Y_{t'-1}, X_{t'-1}]) &= k^{Matern}_{l_a,l_b,\sigma_f}([Y_{t-1}, X_{t-1}], [Y_{t'-1}, X_{t'-1}]) + \sigma_n^2 \delta_{t,t'} \\
k^Z([Z_{t-1}, Y_{t-1}], [Z_{t'-1}, Y_{t'-1}]) &= k^{Matern}_{l_a,l_b,\sigma_f}([Z_{t-1}, Y_{t-1}], [Z_{t'-1}, Y_{t'-1}]) + \sigma_n^2 \delta_{t,t'}
\end{aligned} \tag{3.52}$$

Note that the main causality structure has been encoded in the mean functions, but the way the covariance functions are formulated allows some causality in the covariance in the directions $X \rightarrow Y$ and $Y \rightarrow Z$.

Example Time Series Model Class 2: Structural Causality Incorporated in Volatility The second causality structure has similar autoregressive and causal components to the Structure 1, but the error terms depend on past values of the other time series (so no autoregression in the covariance) via non-linear functions f_y, f_z:

$$\begin{aligned}
X_t &= a_X X_{t-1} + \epsilon_x \\
Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_y^*; \\
Z_t &= a_Z Z_{t-1} + b_Z Y_{t-1}^q + \epsilon_z^*; \\
&\text{where:} \qquad \epsilon_y^* = f_y(X_{t-1}, Z_{t-1})\,\epsilon_y = \left(g_y(t) + c_y X_{t-1}^p + d_y Z_{t-1}^r\right)^2 \epsilon_y \\
&\qquad\qquad\;\; \epsilon_z^* = f_z(X_{t-1}, Y_{t-1})\,\epsilon_z = \left(g_z(t) + c_z X_{t-1}^p + d_z Y_{t-1}^r\right)^2 \epsilon_z
\end{aligned} \tag{3.53}$$

The formulation above is general and the noise terms ϵ_y, ϵ_z can depend explicitly on time via the functions $g_y(t)$ and $g_z(t)$. We use c_y, c_z, d_y, d_z, p, q to denote constants. For this time series to be expressed in terms of GP we will have exactly the same general GP structure as for the time series 1 in the Equation 3.50, and exactly the same mean functions – the Equation 3.51. To construct the kernels that will match the covariance structure, we use the properties that summations and multiplications of kernels yields new

kernels, for example as follows:

$$k^X([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{t'-1}, Y_{t'-1}, Z_{t'-1}]) = \sigma_n^2 \delta_{t,t'} \tag{3.54}$$
$$k^Y([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{t'-1}, Y_{t'-1}, Z_{t'-1}]) = k^{ts2}_{g,p,r,c_Y,d_Y}([X_{t-1}, Z_{t-1}], [X_{t'-1}, Z_{t'-1}])\sigma_n^2 \delta_{t,t'}$$
$$k^Z([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{t'-1}, Y_{t'-1}, Z_{t'-1}]) = k^{ts2}_{g,p,r,c_Z,d_Z}([X_{t-1}, Y_{t-1}], [X_{t'-1}, Y_{t'-1}])\sigma_n^2 \delta_{t,t'}$$

where: $k^{ts2}_{g,p,r,c,d}([W_t, V_t], [W_{t'}, V_{t'}]) = (g + cW_t^p + dV_t^q)^2(g + cW_{t'}^p + dV_{t'}^q)^2$ is a kernel with the functions $g_y(t), g_z(t)$ simplified to a constant g. The notation $[W_t, V_t]$ should be understood as either $[X_{t-1}, Z_{t-1}]$ or $[X_{t-1}, Y_{t-1}]$.

Example Time Series Model Class 3: Causality Features in Presence of Long Memory The third data structure is a long memory process: ARFIMA(p,d,q), for $d \in [0, 0.5)$, with causality structure encoded in the form of external regressors:

$$X_t - a_X X_{t-1} = \epsilon_{x,t} \tag{3.55}$$
$$(Y_t - a_Y Y_{t-1} - b_Y X_{t-1})(1 - B)^d = \Theta_Y(B)\epsilon_{y,t} \tag{3.56}$$
$$(Z_t - a_Z Z_{t-1} - b_Z Y_{t-1}^q)(1 - B)^d = \Theta_Z(B)\epsilon_{z,t}, \tag{3.57}$$

where B is a backshift operator, the autoregressive coefficients for the time series Y_t, Z_t include external regressors, the moving average coefficient according to characteristic polynomial: $\Theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$, and the long memory operator has linear process series expansion given for $d \in (0, 0.5)$ as follows:

$$(1 - B)^{-d} = \sum_{k=0}^{\infty} \frac{\Gamma(k + d)}{\Gamma(k + 1)\Gamma(d)} B^k.$$

In this example, there is no natural way to trivially develop a GP representation, however it does not preclude fitting a misspecified model in order to screen for causality structures that may be present. We can fit such a model to partial observations of this reference example. This poses an interesting example to study the effect of model misspecification on the ability to detect linear and non-linear causality structures.

3.3 Selected Multivariate Distributions

3.3.1 Copulas

Copulas are well established methods for modelling dependence [Nelsen, 2007, Chiu et al., 2015, Durante, 2013, Trivedi et al., 2007, Oh and Patton, 2017]. There are many examples of copulas being used in finance and more specifically in modelling risk – the advances in the quantitative risk management beginning in the eighties lied behind the explosion of interest in copulas. Given the strong connection between modelling causality and dependence, employing copula for measuring causality seems natural,

but is not a popular approach. Several such applications were proposed recently: Bouezmarni et al. [2012], Hu and Liang [2014], Lee and Yang [2014], Póczos et al. [2012].

Below we present a few basic definitions and theorems from the copula theory and introduce notation that will be used later [Elidan, 2013, Lee and Yang, 2014, Nelsen, 2007].

3.3.1.1 Defining Copulas

Copulas are functions that link multivariate distribution functions one-dimensional marginal distribution functions. It can be seen as a way of modelling scale-free measures of dependence between variables. Formally:

Definition 22 *(Copula)*

Let $U_1, ..., U_d$ be d real random one-dimensional variables with uniform distribution on $[0, 1]$*. A copula function $C : [0, 1]^d \to [0, 1]$ is a joint distribution*

$$C(u_1, ..., u_d) = \mathbf{P}(U_1 \leq u_1, ..., U_d \leq u_d) \tag{3.58}$$

Embrechts et al. [2001] calls above definition, in which the copula is simply the original multivariate distribution function with transformed univariate margins, as "operational". A copula can be, however, constructed in various ways and not necessary involve multivariate distributions, or even be defined in probabilistic terms. Below, we introduce an analytic definition of a copula, which requires a definition of a C-volume and d-nondecreasing function. The last two concepts are introduced in the narrower sense - as needed for the copula.

Definition 23 *C-volume*

For a function $C : [0, 1]^d \to [0, 1]$ and a hyperrectangle $B = \Pi_{i=1}^{d} [a_i, b_i] \in [0, 1]^d$, C-volume, denoted V_C is defined as follows:

$$V_C(B) = \sum sgn(w)C(w), \tag{3.59}$$

where the sum is taken over all vertices of B and $sgn(w) = 1$ if $w_i = a_i$ for an even number of i's of $sgn(w) = -1$ if $w_i = a_i$ for an odd number of i's.

Definition 24 *d-nondecreasing function*

A function $C : [0, 1]^d \to [0, 1]$ is d-nondecreasing if for each hyperrectangle $B = \Pi_{i=1}^{d} [x_i, y_i] \in [0, 1]^d$ its C-volume in non-negative.

Definition 25 *(Copula)*

A function $C : [0, 1]^d \to [0, 1]$ is a d-dimensional copula, if:

- $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$*, if at least one arguments is zero, the copula is zero;*

- $C(1, \ldots, 1, u, 1, \ldots, 1) = u$, *if one argument is equal u and the rest are equal 1, the copula is equal to u;*

- *C is n-nondecreasing*

Following is the most important theorem of the copula theory, saying that any joint distribution can be represented as a copula function of univariate marginal distributions.

Theorem 6 *(Sklar's theorem)*

Let $(X) = (X_1, ..., X_d)$ *be a multivariate random variable with joint distribution* $F_{\mathbf{X}}(\mathbf{x})$ *and univariate marginal distributions* $F_1(x_1), ..., F_n(x_d)$*, and let* $C(\cdot)$ *be a copula function. Then the joint distribution can be represented as follows:*

$$F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), ..., F_n(x_d)), \tag{3.60}$$

Moreover, if the marginals are continuous, then the copula function $C(\cdot)$ *is uniquely defined.*

The converse is also true, meaning that and d-dimensional copula function and n marginal distributions will define a valid joint distribution with those marginals. Consequently copulas can be called "distribution generating".

Definition 26 *(Copula density)*

Let $C : [0, 1]^d \to [0, 1]$ *be a copula function that has d'th order partial derivative and let* $F_i(x_i)$ *denote marginal distribution functions as before. Then the joint density can be derived from the copula functions as follows:*

$$\pi(\mathbf{x}) = \frac{\partial^d (F_1(x_1), ..., F_d(x_d))}{\partial F_1(x_1) ... \partial F_d(x_d)} \prod_i \pi_i(x_i) \equiv c(F_1(x_1), ..., F_n(x_n)) \prod_i \pi_i(x_i), \tag{3.61}$$

where the introduced function $c(\cdot)$ *is called a* ***copula density***.

3.3.1.2 Examples of Copula Families

In this section we present several parametric copulas that will be used later in the thesis. We start from independence copula, which represent a joint distribution of independent variables.

Definition 27 *Independence copula*

The d-dimensional independence copula is given by:

$$C^d_{\Pi}(u_1, ..., u_d) = \prod_{i=1}^{d} u_i. \tag{3.62}$$

One of the most popular copulas is the Gaussian copula:

Definition 28 *Gaussian copula*

The d-dimensional Gaussian copula is given by:

$$C_{Gaussian}(u_1, ..., u_d) = \Phi_\rho\left(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)\right), \tag{3.63}$$

where Φ denotes a standard normal distribution, and ρ is a correlation coefficient. The density of Gaussian copula is:

$$c_{Gaussian}(u_1, ..., u_d) = \frac{\pi_{\mathcal{N}}\left(F_{\mathcal{N}}^{-1}(u_1), ..., F_{\mathcal{N}}^{-1}(u_d)\right)}{\Pi_{i=1}^{d}\pi_{\mathcal{N}}\left(F_{\mathcal{N}}^{-1}(u_i)\right)} \tag{3.64}$$

Gaussian copula appears naturally in Gaussian processes, and just like the Gaussian processes it is restrictive if one is interested in tail dependence. A popular copula that can be seen as generalisation of Gaussian copula that allows for tail dependence is a t-copula:

Definition 29 *Student t-copula*

The two dimensional student t-copula is given by:

$$C_t(u_1, u_2; \nu, \rho) = \int_{-\infty}^{t_\nu^{-1}(u_1} \int_{-\infty}^{t_\nu^{-1}(u_2} \frac{1}{2\pi\sqrt{1-\rho^2}}\left(1 + \frac{s^2 - 2\rho st + t^2}{\nu(1-\rho^2)}\right)^{-\frac{\nu+2}{2}} dsdt, \tag{3.65}$$

where t_ν is a cdf of a standard univariate t-distribution with ν degrees of freedom.

The student t distribution can also be obtained as a normal mixing distribution:

$$X \overset{d}{=} W + \sqrt{W}N, \qquad N \sim \mathcal{N}(0, \Sigma) \tag{3.66}$$
$$W \sim IG(-\nu/2, -\nu/2), \quad W \perp\!\!\!\perp N$$

3.3.2 Generalised Hyperbolic Distribution

Generalised hyperbolic distributions are a class of distributions, introduced by Barndorff-Nielsen [1977], that can be represented as a normal mean-variance mixture where the mixture variable is distributed according to Generalised Inverse Gaussian (GIG) distribution. The Generalised Inverse Gaussian distribution was first used by Good [1953] and became popular thanks to, among others, Barndorff-Nielsen [Barndorff-Nielsen and Halgreen, 1977]. For other resources on the two distribution consult [Barndorff-Nielsen, 1978, Blaesild and Jensen, 1981, Olbricht, 1991, Jorgensen, 2012]

This particular infinite mixture representation will be a key feature of such models that will be of relevance to the models developed and testing frameworks introduced for causality. We say that the

random variable X is obtained as a normal mean-variance mixture distribution if

$$X \overset{d}{=} \mu + \gamma W + \sqrt{W}N, \qquad N \sim \mathcal{N}(0, \Sigma) \qquad W \sim GIG(\lambda, \chi, \psi), \quad W \perp\!\!\!\perp N \tag{3.67}$$

where W is called a mixing variable, and is distributed according to Generalised Inverse Gaussian distribution (GIG), formally introduced in the Definition 31. The form of the transformation that defines the variable X implies the following conditional distribution:

$$X \mid W \sim \mathcal{N}(\mu + \gamma W, W\Sigma). \tag{3.68}$$

First, we introduce an integral presentation of the modified Bessel function of the third kind. This function is usually defined as one of the two solution to modified Bessel differential equation, but the integral form is useful when calculating the density of generalised hyperbolic distribution.

Definition 30 ***Modified Bessel Function of the Third Kind [Barndorff-Nielsen and Blaesild, 1981].***

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty y^{\lambda-1}\exp\left(-\frac{x}{2}(y + y^{-1})\right)dy, \qquad x > 0, \lambda \in \mathbb{R} \tag{3.69}$$

Definition 31 ***Generalised Inverse Gaussian (GIG).*** *The random variable W is distributed according to the generalised inverse Gaussian distribution with parameters λ, χ, ψ, denoted GIG(λ, χ, ψ), if its density function can be expressed as:*

$$\pi(x; \lambda, \chi, \psi) = \frac{\chi^{-\lambda}\left(\sqrt{\chi\psi}\right)^\lambda}{2K_\lambda\left(\sqrt{\chi\psi}\right)} x^{\lambda-1}\exp\left(-\frac{1}{2}\left(\chi x^{-1} + \psi x\right)\right), \qquad x > 0, \tag{3.70}$$

where $K_\lambda(x)$ is the modified Bessel function of the third kind, and the parameters λ, χ, ψ satisfy one of the three conditions:

$$\begin{cases} \chi > 0, \psi \geq 0, & \text{if } \lambda < 0, \\ \chi > 0, \psi > 0, & \text{if } \lambda = 0, \\ \chi \geq 0, \psi > 0, & \text{if } \lambda < 0. \end{cases} \tag{3.71}$$

The Generalised Inverse Gaussian (GIG) distribution is infinitely divisible [Barndorff-Nielsen and Halgreen, 1977] and its special cases include **Inverse Gaussian** ($\lambda = -0.5$), **Gamma** ($\chi = 0, \lambda > 0$) and **Inverse Gamma** distributions ($\psi = 0, \lambda < 0$). The Gamma and Inverse Gamma distributions are limiting cases, calculated using the limit: $K_\lambda(x) \sim \Gamma(\lambda)2^{\lambda-1}x^{-\lambda}$ as $x \downarrow 0$, and the property: $K_{-\lambda}(x) = K_\lambda(x)$.

Definition 32 ***Generalised Hyperbolic Distribution (GH).*** *Variable X is said to have a generalised hyperbolic distribution, denoted GH($\lambda, \chi, \psi, m, \Sigma, \gamma$), if the density is given by:*

$$\pi(x) = c \frac{K_{\lambda - \frac{d}{2}}\left(\sqrt{\left(\chi + (x - m_t)^T \Sigma^{-1} (x - m_t)\right)\left(\psi + \gamma^T \Sigma^{-1} \gamma\right)}\right) \exp\left((x - m_t)^T \Sigma^{-1} \gamma\right)}{\left(\sqrt{\left(\chi + (x - m_t)^T \Sigma^{-1} (x - m_t)\right)\left(\psi + \gamma^T \Sigma^{-1} \gamma\right)}\right)^{\frac{d}{2} - \lambda}} \tag{3.72}$$

where $x \in \mathbf{R}^d$, $K_\lambda(x)$ is a modified Bessel function of the third kind with λ degrees of freedom, and c is a constant:

$$c = \frac{(\sqrt{\chi\psi})^{-\lambda} \psi^\lambda \left(\psi + \gamma^T \Sigma^{-1} \gamma\right)^{\frac{d}{2} - \lambda}}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}} K_\lambda(\sqrt{\chi\psi})}, \tag{3.73}$$

where $\mid \cdot \mid$ denotes a determinant. The parameters λ, χ, ψ satisfy conditions from the Equation 3.71 [McNeil et al., 2015].

The first three parameters (λ, χ, ψ) are associated with the mixing GIG distribution, the scale parameter Σ comes from the normal variable, and the parameters μ, γ are the parameters of the mixing. In practice, there will often be an additional restriction on $\mid \Sigma \mid = 1$ (or other fixed value) to avoid identifiability problem. This is because $GH(\lambda, \chi, \psi, m, \Sigma, \gamma) = GH(\lambda, \chi/k, k\psi, m, k\Sigma, k\gamma)$, for any $k > 0$.

Theorem 7 ***Generalised Hyperbolic Distribution (GH) [McNeil et al., 2015].*** *If $N \sim \mathcal{N}(0, \Sigma)$ and Σ is full rank, with the mixing variable $W \sim GIG(\lambda, \chi, \psi)$, then the normal mean variance mixture $X \stackrel{d}{=} m + \gamma W + \sqrt{W} N$ will have a marginal distribution for X, having integrated out uncertainty attributed to W, given by a generalised hyperbolic distribution, denoted GH($\lambda, \chi, \psi, m, \Sigma, \gamma$), given by the formula 3.72.*

Proof: of the Theorem 7

We use the fact that the X will be normally distributed when conditioned on the mixing variable: $X \mid W \sim \mathcal{N}(m + \gamma W, W\Sigma)$. The unconditional distribution is therefore calculated as a following integral:

$$\pi(x) = \int_0^\infty f(x \mid w) \pi(w) dw \tag{3.74}$$

Using the density of a generalised inverse Gaussian (GIG) distribution $W \sim GIG(\lambda, \chi, \psi)$ from the Equation 31 we extend A.16 and write:

$$\begin{aligned}\pi(x) &= \int_0^\infty \frac{1}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}} w^{\frac{d}{2}}} \exp\left\{\frac{-(x-m-\gamma w)^T (w\Sigma)^{-1}(x-m-\gamma w)}{2}\right\}\pi(w)dw \\ &= \int_0^\infty \frac{e^{(x-m)^T\Sigma^{-1}\gamma}}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}} w^{\frac{d}{2}}} \exp\left\{-\frac{(x-m)^T\Sigma^{-1}(x-m)}{2w} - \frac{\gamma^T\Sigma^{-1}\gamma}{2/w}\right\}\pi(w)dw \\ &= \frac{\chi^{-\lambda}\left(\sqrt{\chi\psi}\right)^\lambda e^{(x-m)^T\Sigma^{-1}\gamma}}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}} K_\lambda\left(\sqrt{\chi\psi}\right)} \\ &\quad \frac{1}{2}\int_0^\infty w^{\lambda-\frac{d}{2}-1} \exp\left\{-\frac{(x-m)^T\Sigma^{-1}(x-m)+\chi}{2w} - \frac{\gamma^T\Sigma^{-1}\gamma+\psi}{2/w}\right\}\pi(w)dw\end{aligned}$$

To allow expressing the integral in terms of a modified Bessel function of the third kind, we perform a change of variable:

$$z = w\frac{\sqrt{(\psi+\gamma^T\Sigma^{-1}\gamma)}}{\sqrt{(\psi+(x-m)^T\Sigma^{-1}(x-m))}}$$

and as a result we obtain:

$$\begin{aligned}\pi(x) =& \frac{\chi^{-\lambda}\left(\sqrt{\chi\psi}\right)^\lambda e^{(x-m)^T\Sigma^{-1}\gamma}}{(2\pi)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}} K_\lambda\left(\sqrt{\chi\psi}\right)}\left(\frac{\sqrt{(\psi+(x-m)^T\Sigma^{-1}(x-m))}}{\sqrt{(\psi+\gamma^T\Sigma^{-1}\gamma)}}\right)^{\frac{d}{2}-\lambda} \\ & \underbrace{\frac{1}{2}\int_0^\infty z^{\lambda-\frac{d}{2}-1}\exp\left\{-\frac{1}{2}\sqrt{(\chi+(x-m)^T\Sigma^{-1}(x-m))(\psi+\gamma^T\Sigma^{-1}\gamma)}\left[\frac{1}{z}+z\right]\right\}f(z)dz}_{K_{\lambda-\frac{d}{2}}\left(\sqrt{(\chi+(x-m)^T\Sigma^{-1}(x-m))(\psi+\gamma^T\Sigma^{-1}\gamma)}\right)}\end{aligned}$$

which after reorganisation gives the requested density. □

An important observation is that the proof of the Theorem 7 holds for mean variance mixture where the normal random variable N has a normal distribution with zero mean.

3.3.3 Properties of the Generalised Hyperbolic Distribution

The class of generalised hyperbolic distributions have several very useful properties. The closed form expressions for characteristic function and moment generating function allow easy calculation of moments and allows to easily show that the class is closed under linear transformations, and therefore closed for marginalisation. The marginalisation property is then used when defining copula density.

Theorem 8 ***Moment Generating Function of Generalised Hyperbolic Distributions***. *Let* $X \sim GH(\lambda,\chi,\psi,\mu,\Sigma,\gamma)$*, then the moment generating function of* X *is:*

$$M_X(s) = e^{s^T\mu}M_{GIG(\lambda,\chi,\psi)}\left(s^T\gamma + \frac{1}{2}s^T\Sigma s\right),$$

this can also be expressed explicitly in terms of modified Bessel functions of the third kind:

$$M_X(s) = e^{s^T\mu}\left(\frac{\psi}{\psi - 2\left(s^T\gamma + \frac{1}{2}s^T\Sigma s\right)}\right)^{\frac{\lambda}{2}} \frac{K_\lambda\left(\sqrt{\chi + (\psi - 2\left(s^T\gamma + \frac{1}{2}s^T\Sigma s\right))}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}.$$

Proof: see Appendix, Section A.4

Theorem 9 ***Characteristic Function of Generalised Hyperbolic Distributions*** *Let $X \sim GH(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$, then the characteristic function of X is:*

$$\varphi_X(s) = e^{is^T\mu}\varphi_{GIG(\lambda,\chi,\psi)}\left(is^T\gamma - \frac{1}{2}s^T\Sigma s\right),$$

this can also be expressed explicitly in terms of modified Bessel functions of the third kind:

$$\varphi_X(s) = e^{s^T\mu}\left(\frac{\psi}{\psi - 2\left(is^T\gamma - \frac{1}{2}s^T\Sigma s\right)}\right)^{\frac{\lambda}{2}} \frac{K_\lambda\left(\sqrt{\chi + (\psi - 2\left(is^T\gamma - \frac{1}{2}s^T\Sigma s\right))}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}.$$

From the moment generating function we can obtain (using some properties of Bessel functions) the moments, for example the mean and the variance:

$$\mathbb{E}[X] = \mu + \gamma\left(\frac{\chi}{\psi}\right)^{\frac{1}{2}} \frac{K_{\lambda+1}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)},$$

$$\mathbb{C}\text{ov}[X] = \left(\frac{\chi}{\psi}\right)^{\frac{1}{2}} \frac{K_{\lambda+1}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}\Sigma + \gamma\gamma^T\left(\left(\frac{\chi}{\psi}\right)\frac{K_{\lambda+2}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)} - \left(\left(\frac{\chi}{\psi}\right)^{\frac{1}{2}} \frac{K_{\lambda+1}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}\right)^2\right).$$

The generalised hyperbolic is closed under linear transformations [Hu, 2005]:

Theorem 10 ***Linear Transformations of Generalised Hyperbolic Distributions, [McNeil et al., 2015].*** *If $X \sim GH_d(\lambda, \chi, \psi, m, \Sigma, \gamma)$ and $Y = BX + b$, where $B \in \mathbb{R}^{k\times d}, b \in \mathbb{R}^k$, then $Y \sim GH_d(\lambda, \chi, \psi, Bm + b, B\Sigma B^T, B\gamma)$.*

Theorem 11 ***Weighted Sum of Generalised Hyperbolic Distributions.*** *If $X \sim GH_d(\lambda, \chi, \psi, m, \Sigma, \gamma)$, $\omega^T = (\omega_1, ..., \omega_d)$ and $Y = \omega^T X$, then $Y \sim GH_1(\lambda, \chi, \psi, \omega^T m, \omega^T\Sigma\omega, \omega^T\gamma)$, which is a one dimensional distribution.*

As a consequence of the Theorem 11, we have that the generalised hyperbolic distribution is closed under marginalisations, [McNeil et al., 2015] if $X \sim GH(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$, then the marginal distribution of X_i is: $X_i \sim GH(\lambda, \chi, \psi, \mu_i, \Sigma_i i, \gamma_i)$.

A d-dimensional copula C is a d-dimensional distribution function on $[0,1]^d$ with standard uniform marginal distributions. Sklar's Theorem (see [Nelsen, 2007]) states that any multivariate joint distribution can be represented in terms of univariate marginal distribution functions and a function:

$$F(x_1, ..., x_d) = C\left(F_1(x_1), ..., F_d(x_d)\right),$$

where F is a joint distribution function, $F_1, ..., F_d$ are marginal distribution functions. If the marginal distribution functions are continuous and strictly increasing, we have the following formulation for the copula function:

$$C(u_1, ..., u_d) = F\left(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)\right).$$

Furthermore, if the multivariate distribution has a density function π, and the marginal distributions – $\pi_1, ..., \pi_d$, then the density of a copula can be expressed as:

$$\pi(x_1, ..., x_d) = c\left(F_1(x_1), ..., F_d(x_d)\right)\pi_1(x_1) \cdot ... \cdot \pi_d(x_d)$$

and conversely:

$$c(u_1, ..., u_d) = \frac{\pi\left(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)\right)}{\prod_{i=1}^{d}\pi_i\left(F_i^{-1}(u_i)\right)}.$$

Consequently, if we have a normal mean-variance mixture $X \overset{d}{=} \mu + \gamma W + \sqrt{W}N$, as defined in 3.67, then the following variable will be distributed from a Generalised Hyperbolic copula:

$$U = \left(F_{GH_1}(X_1; \lambda, \chi, \psi, \mu_1, \Sigma_{11}, \gamma_1), ..., F_{GH_1}(X_d; \lambda, \chi, \psi, \mu_d, \Sigma_{dd}, \gamma_d)\right)^T, \tag{3.75}$$

where $F_{GH_1}(\cdot)$ is the distribution of a one dimensional Generalised Hyperbolic variable.

Some of the most popular distributions are, in fact, special cases of the GH distribution. We are especially interested in the skew-t distribution, which is a generalisation of the t-student distribution. Those distributions are summarised in the Table 3.2.

3.3.4 Skew-t Distributions, Generalisations and Alternative Parametrizations

There are several different distributions that have been given the "skew-t" name and we mention three of those here. To differentiate between them we would use the name "GH skew-t" for our main skew-t distribution – one that is a special case of Generalised Hyperbolic distribution. The name "generalised skew-t" refers to a distribution very similar to the classical one, but with separate degrees of freedom for

name	$\lambda, \chi, \psi, m, \Sigma, \gamma$	properties
hyperbolic	$\lambda = 1$	the result is multivariate generalised hyperbolic distribution, with univariate hyperbolic distributions marginals
hyperbolic	$\lambda = \frac{d+1}{2}$	the result is d-dimensional hyperbolic distribution, with marginals which are not hyperbolic distributions
normal inverse Gaussian (NIG)	$\lambda = -\frac{1}{2}$	
Variance Gamma (VG)	$\lambda > 0, \chi = 0$	
skew-t	$\lambda > 0, \chi = \nu, \psi = 0$	normal mean-variance mixture using inverse gamma mixing distribution $IG(\frac{\nu}{2}, \frac{\nu}{2})$. Implicitly defined **skew t-copula** density is given (Sklar's theorem) as the ratio of the multivariate skew-t distribution over the product of the marginal skew-t densities.
student-t	$\lambda > 0, \chi = \nu, \psi, \gamma = 0$	
normal	$\lambda > 0, \chi, \nu = \infty, \psi, \gamma = 0$	t-student distribution in the limit $\nu = \infty$

Table 3.2: Special cases of the GH(λ, χ, ψ) distribution.

each dimension. The third, which we will call an "elliptical skew-t" is a special case of a skew-elliptical family of distributions, but it can also be obtained through a mean variance transformation.

3.3.4.1 The GH skew-t distribution

In this section we will recall the link between the GH model for the case of mixing random variable W being Inverse Gamma and the copula dependence structure implicitly defined for this multivariate GH sub-family which will correspond to what is known as a skew-t copula. The skew-t copula has been introduced by Demarta and McNeil as a generalisation of the t copula to provide " (. . .)more heterogeneity in the modelling of dependent observations" [Demarta and McNeil, 2005]. It has also been studied by Rachev [2003], see also: [McNeil et al., 2015]. In our research we devote special attention to this distribution and its tail properties. Being a special case of a generalised hyperbolic distribution, the GH skew-t can be obtained from the same normal mean-variance transformation from the Equation 3.67. Here, however, the mixing variable W has an Inverse Gamma (IG) distribution, which is a special case of GIG distribution:

$$X \stackrel{d}{=} m + \gamma W + \sqrt{W} N, \qquad \begin{aligned} &N \sim \mathcal{N}(0, \Sigma) \\ &W \sim IG(\frac{\nu}{2}, \frac{\nu}{2}), \quad W \perp\!\!\!\perp N \end{aligned} \tag{3.76}$$

The Inverse Gamma distribution used above is a limiting case of the Generalised Inverse Gaussian distribution (GIG), where the following limit is used: $K_\lambda(x) \sim \Gamma(-\lambda) 2^{-\lambda-1} x^\lambda$ as $x \downarrow 0$. At the same time we could say that the skew-t distribution as a limiting case of the Generalised Hyperbolic distribution, where the same limit is applied. The multivariate skew-t distribution can therefore be defined as follows:

Definition 33 ***Skew-t distribution*** *X has a multivariate skew-t distribution $St(\mu, \Sigma, \gamma, \nu)$, with location parameter μ, scale parameter Σ, skewness parameter γ and shape parameter ν, if it has density:*

$$\pi(x) = c \frac{K_{\frac{\nu+d}{2}}\left(\sqrt{\left(\nu + (x-m)^T \Sigma^{-1} (x-m)\right)\left(\gamma^T \Sigma^{-1} \gamma\right)}\right) \exp\left((x-m)^T \Sigma^{-1} \gamma\right)}{\left(\sqrt{\left(\nu + (x-m)^T \Sigma^{-1} (x-m)\right)\left(\gamma^T \Sigma^{-1} \gamma\right)}\right)^{-\frac{\nu+d}{2}} \left(1 + \frac{1}{\nu}(x-m)^T \Sigma^{-1} (x-m)\right)^{\frac{\nu+d}{2}}}, \tag{3.77}$$

where $K_\lambda(x)$ is a modified Bessel function of the third kind with λ degrees of freedom, and c is a constant: $c = \left\{2^{1-\frac{\nu+d}{2}}\right\}\left\{\Gamma(\frac{\nu}{2})(\pi\nu)^{\frac{d}{2}} \mid \Sigma \mid^{\frac{1}{2}}\right\}$.

Being a special case of the GH distribution, the skew-t distribution has the same conditional normal distribution as we have seen in the Equation 3.68:

$$X \mid W \sim \mathcal{N}(\mu + \gamma W, W\Sigma). \tag{3.78}$$

The most important property of this distribution is that it has an implicitly defined **skew t-copula** density that is given (Sklar's theorem) by the ratio of the multivariate skew-t distribution over the product of the marginal skew-t densities. Analogously to the Equation 3.75, we have that, for F_{St_1} representing distribution of a one dimensional skew-t variable,

$$U = (F_{St_1}(X_1; \mu_1, \Sigma_{11}, \gamma_1, \nu), ..., F_{St_1}(X_d; \mu_d, \Sigma_{dd}, \gamma_d, \nu))^T \tag{3.79}$$

is distributed from the skew-t copula.

The GH skew-t distribution also inherits many useful properties from the Generalised Hyperbolic family, including easy to calculate moment generating function $M_X t = e^{t^T \mu} M_W(t^T \gamma + \frac{1}{2} t^T \Sigma t)$, mean $\mathbf{E}(X) = m + \gamma \frac{\nu}{\nu-2}$ and covariance $COV(X) = \frac{\nu}{\nu-2}\Sigma + \gamma\gamma^T \frac{2\nu^2}{(\nu-2)^2(\nu-4)}$. Also, the normal mean-variance representation result in a convenient estimation with the Expectation Maximisation (EM) algorithm.

3.3.4.2 Generalised skew-t

In this section we provide the details of generalised skew-t distribution, where the generalisation pertains to allowing different degrees of freedom for each of the dimensions, see [Cruz et al., 2015, Luo and Shevchenko, 2010]. The construction of such a distribution will be analogous to the construction of classical skew-t distribution in Equation 3.76. The main difference is that the mixing variable W will be multivariate, with the marginals W_i having inverse gamma distributions, and being perfectly dependant

via a uniform variable U:

$$X \overset{d}{=} m + \gamma W + \sqrt{W}N, \qquad N \sim \mathcal{N}(0, \Sigma) \tag{3.80}$$
$$W = diag\left([W_1, W_2, \dots W_d]\right), \quad W \perp\!\!\!\perp N$$
$$W_i = F_{W_i}^{-1}(U) = IG(u; \frac{\nu_i}{2}, \frac{\nu_i}{2}), \text{ where } U \sim \mathcal{U}\left[0, 1\right].$$

Although distribution of W is slightly different, we have the same conditional normal distribution as for the skew-t in the Equation 3.68 and 3.78:

$$X \mid W \sim \mathcal{N}(\mu + \gamma W, W\Sigma). \tag{3.81}$$

We also have analogous form of the generalised skew-t copula (see Equations3.75 and 3.79), with F_{GSt_1} denoting a distribution of a variable distributed according to one dimensional generalised skew-t (which is equivalent to the one dimensional skew-t distribution):

$$U = \left(F_{GSt_1}(X_1; \mu_1, \Sigma_{11}, \gamma_1, \nu_1), ..., F_{GSt_1}(X_d; \mu_d, \Sigma_{dd}, \gamma_d, \nu_d)\right)^T. \tag{3.82}$$

The distribution of X is, however, no longer given analytically, for example in a bivariate case:

$$\pi(x) = \int \pi(x \mid w)\pi(w)dw = \int \pi(x \mid u)\pi(u) \mid \frac{\partial w}{\partial u} \mid^{-1} du \tag{3.83}$$
$$= \int \underbrace{\phi(\frac{x - \mu - \gamma w}{w\Sigma} \mid u) w_1^{-2}(u) w_2^{-2}(u)}_{\varphi(u)} du. \tag{3.84}$$

The last integral in the Equations 3.83 needs to be approximated. The simplest example of approximation would be: $\sum_{i=1}^{n}(u_{i+1} - u_i)(\varphi(u_{i+1}) + \varphi(u_i))$. However, as it will become clear later, we will not require calculation of the joint distribution for our test statistic, but only marginal, which has a skew-t distribution. This can be easily seen, for example, for bivariate generalised skew-t variable $X = [X_1, X_2]^T$

$$\pi(x_1) = \int_{-\infty}^{\infty} \pi(x)dx_2 = \int_{-\infty}^{\infty}\int_{0}^{\infty} \pi(x \mid w)\pi(w)dwdx_2$$
$$= \int_0^\infty \int_{-\infty}^{\infty} \underbrace{\pi(x \mid w)}_{\text{joint normal}} dx_2 \pi(w)dw = \underbrace{\int_0^\infty \pi(x_1 \mid w)\pi(w)dw}_{\text{univariate GH}}$$

The joint distribution will only be needed to estimate the model parameters.

3.3.4.3 The Alternative Skew-t Distribution

Fung and Seneta argued in their paper [Fung and Seneta, 2010b] that the extension of symmetric t distribution to the skew-t distribution proposed by Demarta and McNeil in [Demarta and McNeil, 2005]

is not appropriate. They suggested that an appropriate extension should have a non-trivial values of the tail dependence coefficients (discussed in the Section 3.3.5), given that these are non-trivial for symmetric t distribution.

The alternative definition of the skew-t distribution suggested by Fung and Seneta, is given following Azzalini and Capitanio [2003] (a comprehensive review in Azzalini [2005]), where it has been derived as a special case of a class of multivariate skew-elliptical distributions. Note, that Branco and Dey [2001] propose a different parametrisation of skew-elliptical distribution and a special case of a skew-t distribution being therefore obtained from a different parametrisation. Azzalini and Capitanio prove that their parametrisation of skew-elliptical distributions is closely connected to the one proposed by Branco and Dey [2001], while the two parametrisations of skew-t distributions actually coincide Azzalini and Capitanio [2003]. The class of skew-elliptical distributions can accommodate both the skewness and the heavy tails and can be generated from (symmetrical) elliptical distributions with the conditioning method.

Here we will follow with the derivation by Fung and Seneta [2010a], which we adapt to be consistent with the rest of the thesis. It is obtained as a mean variance mixture similar to the one used in the Equations 3.76 and 3.80 for defining the skew-t or generalised hyperbolic, but with one additional step and one additional mixing variable.

Definition 34 ***Skew normal distribution*** *An n dimensional variable X is said to have a skew normal distribution with skewness parameter γ, written $X \sim SN(\gamma)$ if its density function can be presented as:*

$$\pi(x) = 2\phi_n(x, \Omega)\Phi(\gamma x), \tag{3.85}$$

where $\phi_n(x, \Omega)$ denotes density function of n-variate normal variable with standardised marginals and correlation matrix Ω, and Φ is cumulative distribution function of the normal distribution $\mathcal{N}(0, 1)$.

The above definition, given by in Azzalini and Valle [1996], is sufficient for how we will use the skew normally distributed variables. For a more general definition – with nonzero mean and covariance rather than correlation matrix, we direct the reader to Azzalini and Capitanio [2003].

There are several ways how one might construct a skew normal distribution, we present the following transformation. Let $N \sim \mathcal{N}_n(0, \Omega)$, where Ψ is a correlation matrix (or in other words: N has standardised marginals), and let $N_0 \sim \mathcal{N}(0, 1)$ be independent of N. Therefore we have:

$$\begin{pmatrix} N_0 \\ N \end{pmatrix} \sim \mathcal{N}\left\{0, \begin{pmatrix} 1 & 0 \\ 0 & \Omega \end{pmatrix}\right\}$$

If we then choose $\theta_1, ..., \theta_n$, such that $\theta \in (-1, 1)$ for all i, then

$$Y_i = \theta_i \mid N_0 \mid + \sqrt{1 - \theta_i^2} N_i \qquad \text{for } i = 1, ..., n. \tag{3.86}$$

We then say that $Y_i \sim SN(\gamma_i)$ or that $Y \sim SN_n(\gamma, \Omega)$, where the skewness parameter $\gamma = (\gamma(\theta_1), ..., \gamma(\theta_n))^T$ is obtained as:

$$\gamma(\theta_i) = \frac{\theta_i}{\sqrt{1 - \theta_i^2}}.$$

After the construction of the skew-normal variable is complete, the skew-normal variable can be obtained as a mean-variance transformation:

$$X \overset{d}{=} \mu + \sqrt{W} Y, \qquad \begin{aligned} &Y \sim SN(\gamma, \Omega) \\ &W \sim IG(\frac{\nu}{2}, \frac{\nu}{2}), \quad W \perp\!\!\!\perp Y \end{aligned} \tag{3.87}$$

We formally define the elliptical skew-t distribution as follows:

Definition 35 ***Alternative skew-t distribution*** *A random vector X has an n - dimensional elliptical skew-t distribution, denoted as $X \sim AST_n(\mu, \Omega, \gamma, \nu)$, with location parameter ν, dispersion parameter Σ, skewness parameter γ and a shape parameter ν, if its density is of type:*

$$\pi(x) = 2p_{t_n}(x; \mu, \Sigma, \nu) F_{t_1}\left(\gamma^T \omega^{-1}(x - \mu) \sqrt{\frac{\nu + n}{\nu + Q_x}}; 0, 1, \nu + n\right) \tag{3.88}$$

where $Q_x = (x-\mu)\Omega^{-1}(x-\mu)$, ω is a diagonal matrix formed by standard deviations of Σ, $F_{t_1}(x; 0, 1, \nu+n)$ is a univariate, standardised distribution function of t distribution with $\nu + d$ degrees of freedom, and $f_{t_n}(x; \mu, \Sigma, \nu)$ is an n-variate density of t distribution defined:

$$p_{t_n}(x; \mu, \Omega, \nu) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\mid \Sigma \mid^{\frac{1}{2}} (\pi\nu)^{\frac{n}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{Q_x}{\nu}\right)^{-\frac{\nu+n}{2}} \tag{3.89}$$

The copula for the elliptical skew-t is, similarly to other distributions that we have seen (see: Equations 3.75, 3.79 and 3.82), we have that:

$$U = \left(F_{AST_1}(X_1; \mu_1, 1, \gamma_1, \nu), ..., F_{AST_1}(X_d; \mu_d, 1, \gamma_d, \nu)\right)^T, \tag{3.90}$$

with F_{AST_1} being the distribution of a one dimensional elliptical skew-t variable, is distributed from the elliptical skew-t copula, where we observe that $\Omega_{ii} = 1$.

We observe here, that when defining the mean-variance transformation to obtain the "classical"

skew-t, we were transforming a normal variable with zero mean, and when defining the elliptical skew-t, we were adding the restriction on covariance matrix as well. Also the skewness was introduced in a different way. We will later observe that those differences do seem to have effect on practical aspects of our models.

3.3.5 Tail Behaviour and Tail Dependence in the Skew-t distributions

Both generalised hyperbolic and skew-elliptical distributions are designed to allow modelling a wide range of shapes and in particular allow for heavy tailed features as well as asymmetric tail behaviours in each marginals left and right tails. Analysing the two skew-t distributions: GH and the elliptical we observe several things. We notice that both have a power tail for their marginal processes, but in the case of the elliptical skew-t the tails are symmetric. When we look at the tail dependence, we find out that the GH skew-t will have a trivial (0 or 1) behaviour whenever both of the skewness parameters are non-zero, while the elliptical skew-t variables will have non-trivial tail dependence in most cases.

Does it mean that one of the distributions is clearly better? We believe this means that in the first place the behaviour of the models need to be well understood so that an appropriate model can be matched with the problem or data to be modelled. We will return to this idea later in the context of testing for causality in such models.

3.3.5.1 Tail behaviour

Both GH skew-t model and the elliptical skew-t model display power tails, but there are some important differences.

First we start with the tail behaviour of univariate skew-t distribution. The variable $X \sim St(\nu, \mu, \sigma, \gamma)$ has a power tail, which can be described as follows:

$$p_X(x) \sim \text{ const } \mid x \mid^{-\frac{\nu}{2}-1} e^{-\alpha|x|+\beta x} \quad \text{as} \quad x \to \pm\infty,$$

where the constants are $\alpha = \frac{|\gamma|}{\sigma^2}$ and $\beta = \frac{\gamma}{\sigma^2}$. The sign of the skew parameter γ influences which tail will be heavier: the right for $\gamma > 0$, left for $\gamma < 0$. To reiterate, this means we either have symmetric tails, when the distribution is not skewed, or one tail heavier and one lighter.

As mentioned earlier, the elliptical skew-t distribution has been suggested as a way to improve on what Fung and Seneta [2010b] declared to be a non-satisfactory tail dependence. But the tail behaviour might be unexpected.

The tail behaviour of the elliptical skew-t is is also power law, however it behaves as the symmetrical of the classical skew-t (just with a different constant). Let $X \sim AS\,t_1(\mu, \sigma^2, \gamma, \nu)$, then:

$$p_X(x) \sim \text{ const } \mid x \mid^{-\nu} .$$

Arellano-Valle and Genton [2010] point out the elliptical skew-t distribution is therefore capable of modelling heavier tails, but not lighter tails. They also introduce a new class of distributions: multivariate extended skew-t distributions, which are supposed to help help relax this restriction.

3.3.5.2 Tail Dependence

Although we haven't yet formally defined causality, or explained how we propose to utilise skew-t distribution to model it, we would like to point out that we will be interested in analysing interaction between causality and dependence, in particular - tail dependence. We will be using the tail dependence coefficients, so below we first define it, and then report on it's behaviour for the GH and the elliptical skew-t distributions.

Definition 36 ***Bivariate Tail Dependence Coefficient. [Cruz et al., 2015]*** *Let X, Y be two random variables with distributions F, G respectively. Then the upper tail dependence coefficient λ_u and the lower tail dependence coefficient λ_l are defined by:*

$$\lambda_u = \lim_{u \uparrow 1} P\left(Y > G^{-1}(u) \mid X > F^{-1}(u)\right) \tag{3.91}$$

$$\lambda_l = \lim_{u \downarrow 0} P\left(Y \leq G^{-1}(u) \mid X \leq F^{-1}(u)\right). \tag{3.92}$$

Both upper and lower tails can also be expressed in terms of the copula:

$$\lambda_u = \lim_{u \uparrow 1} \frac{1 - 2u + C(u, u)}{1 - u}$$
$$\lambda_l = \lim_{u \downarrow 0} \frac{C(u, u)}{u}.$$

We observe that $\tilde{C}(1-u, 1-u) = 1 - 2u + C(u, u)$ is survival copula. Upper tail for copula C equals lower tail for the survival copula $\tilde{C}$, and other way round: $\lambda_u = \tilde{\lambda}_l, \lambda_l = \tilde{\lambda}_u$.

A result that we will later require is the following decomposition of the upper and lower tail coefficients:

Theorem 12 *Let X, Y be two random variables with continuous and strictly increasing distributions F, G respectively. Then λ_u and λ_l can be expressed as:*

$$\lambda_u = \lim_{u \uparrow 1} P\left(Y > G^{-1}(u) \mid X = F^{-1}(u)\right) + \lim_{u \uparrow 1} P\left(X > F^{-1}(u) \mid Y = G^{-1}(u)\right) \tag{3.93}$$

$$\lambda_l = \lim_{u \downarrow 0} P\left(Y \leq G^{-1}(u) \mid X = F^{-1}(u)\right) + \lim_{u \downarrow 0} P\left(X \leq F^{-1}(u) \mid Y = G^{-1}(u)\right). \tag{3.94}$$

Proof: The assumption of continuous distributions F, G implies the random variables $U := F(X)$ and $V := G(Y)$ to be uniformly distributed on $(0, 1)$. The joint distribution of $(U, V)^T$ is a copula, and the assumption of F, G being strictly increasing, then this copula is also a copula of of X and Y. Let's denote

it as C_{XY}. We are interested in the partial derivatives of the copula C_{XY}. To ensure the existence of the partial derivatives we refer to the Theorem 2.2.7 from [Nelsen, 2007], stating that: for any $v \in [0,1]$ partial derivative $\partial C(u,v)/\partial v$ exists for almost all u, and when it exists it is bounded: $0 \leq \frac{\partial}{\partial u} C(u,v) \leq 1$. Analogously for $\partial C(u,v)/\partial u$. As a consequence:

$$\begin{aligned}\frac{\partial C_{XY}}{\partial u}(u,v) &= \lim_{\epsilon \to 0} \frac{C_{XY}(u+\epsilon, v) - C_{XY}(u,v)}{\epsilon} = \lim_{\epsilon \to 0} \frac{P(u \leq U \leq u+\epsilon, V \leq v)}{P(u \leq U \leq u+\epsilon)} \\ &= P(V \leq v \mid U = u).\end{aligned}$$

Furthermore, we have that $P(V \leq v \mid U = u) = P(G(Y) \leq v \mid F(X) = u) = P\left(Y \leq G^{-1}(v) \mid X = F^{-1}(u)\right)$. And analogously, $\frac{\partial C_{XY}}{\partial v}(u,v) = P(U \leq u \mid V = v) = P\left(X \leq F^{-1}(u) \mid Y = G^{-1}(v)\right)$. We apply L'Hospital's rule and use the notation $\frac{\partial C_{XY}}{\partial u}$ and $\frac{\partial C_{XY}}{\partial v}$ to represent partial derivative with respect to the first variable of the multivariate function

$$\begin{aligned}\lambda_l &= \lim_{q \uparrow 1} \frac{C_{XY}(q,q)}{q} = \lim_{q \uparrow 1} \frac{dC_{XY}(q,q)}{dq} = \lim_{q \uparrow 1} \left(\frac{\partial C_{XY}}{\partial u}(q,q) + \frac{\partial C_{XY}}{\partial v}(q,q) \right) \\ &= \lim_{q \uparrow 1} \left(P(V \leq q \mid U = q) + P(U \leq q \mid V = q) \right) \\ &= \lim_{q \uparrow 1} \left(P\left(Y \leq G^{-1}(q) \mid X = F^{-1}(q)\right) + P\left(X \leq F^{-1}(q) \mid Y = G^{-1}(q)\right) \right)\end{aligned}$$

Likewise, the result $\lambda_u = \lim_{u \uparrow 1} P\left(Y > G^{-1}(u) \mid X > F^{-1}(u)\right)$, but with the survival copula.

□

In the sequel, we follow the results shown in Fung and Seneta [2010b] to describe the tail coefficients for the classical and elliptical skew-t distributions The result for generalised skew-t is very similar to the one for classical skew-t and can be found for example in Banachewicz and Van Der Vaart [2008], Banachewicz et al. [2009].

Theorem 13 ***Tail dependence for generalised skew-t distribution.*** *Let X_1, X_2 be a bivariate skew-t distributed variable, as defined in the Equations 3.88, with γ_1, γ_2 being the skewness parameters and ρ the correlation of the normal variables in the mixture. Then the upper and lower tail coefficients of X_1, X_2 are given by:*

1. *If $\gamma_1 = \gamma_2 = 0$ (i.e. bivariate symmetric t), then:*

$$\lambda_L = 2F_{t_1}\left(\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}} \right).$$

2. *If $\gamma_1 > 0, \gamma_2 > 0$, then $\lambda_L = 0$*

3. *If $\gamma_1 < 0, \gamma_2 < 0$, then $\lambda_L = 1$*

4. *If* $\gamma_1 < 0, \gamma_2 > 0$, *then* $\lambda_L = 0$

5. *If* $\gamma_1 = 0, \gamma_2 > 0$, *then* $\lambda_L = 0$

6. *If* $\gamma_1 = 0, \gamma_2 < 0$, *then*

$$\lambda_L = \int_0^1 \left(1 - \Phi\left(\left(\frac{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu+1}{2}\right)}{2\sqrt{\pi}}\right)^{1/\nu} u^{1/\nu_1}\right)\right) du.$$

.

For results about tail dependence coefficients in the generalised hyperbolic distribution we direct the reader to the book chapter by Hammerstein [2016]. Here we point out that those results are in line with the ones for the skew-t distribution in terms of showing full or no tail dependence for most of the cases, and non-trivial tail dependence only in a narrow range of cases.

Theorem 14 ***Tail dependence for an elliptical skew-t distribution [Fung and Seneta, 2010a].*** *Let* $[X_1, X_2]^T \sim AS\,t_2(\mu, \Sigma, \gamma, \nu)$ *be a bivariate elliptical skew-t distributed random vector. The asymptotic lower tail dependence coefficient for* $[X_1, X_2]^T$ *is given by:*

$$\lambda_L = \int_{-\infty}^{c_1} p f_{t_{\nu+1}}(z) \frac{F_{t_{\nu+2}}\left(\left(\theta_2\sqrt{\frac{1-\rho^2}{\nu+1}} - (\theta_1 + \rho\theta_2)\right)\sqrt{\frac{\nu+2}{1+\frac{z^2}{\nu+1}}}\right)}{F_{t_{\nu+1}}(-\lambda_1\sqrt{\nu+1})} dz$$
$$+ \int_{-\infty}^{c_2} p_{t_{\nu+1}}(z) \frac{F_{t_{\nu+2}}\left(\left(\theta_1\sqrt{\frac{1-\rho^2}{\nu+1}} - (\theta_2 + \rho\theta_1)\right)\sqrt{\frac{\nu+2}{1+\frac{z^2}{\nu+1}}}\right)}{F_{t_{\nu+1}}(-\lambda_2\sqrt{\nu+1})} dz,$$

where

$$c_1 = \left\{\left(\frac{F_{t_{\nu+1}}(-\lambda_2\sqrt{\nu+1})}{F_{t_{\nu+1}}(-\lambda_1\sqrt{\nu+1})}\right)^{1/\nu} - \rho\right\}\sqrt{\frac{\nu+1}{1-\rho^2}}$$
$$c_2 = \left\{\left(\frac{F_{t_{\nu+1}}(-\lambda_1\sqrt{\nu+1})}{F_{t_{\nu+1}}(-\lambda_2\sqrt{\nu+1})}\right)^{1/\nu} - \rho\right\}\sqrt{\frac{\nu+1}{1-\rho^2}}.$$

where the notation $p_{t_\nu}(.), F_{t_\nu}(.)$ *is used for, respectively, the p.d.f. and c.d.f. of a univariate symmetric t distributions with* ν *degrees of freedom.*

The elliptical skew-t distribution is not closed for marginalisations. But it's extension – the multivariate extended skew-t distributions proposed by Arellano-Valle and Genton [2010] does exhibit this property. One could obtain the tail dependence coefficient can from this conditional distribution and using the Equation 3.93. One more interesting result, is the one obtained by Bortot [2010], where a simple

formula is given for a tail in the case of the two gamma parameters having the same value:

$$\lambda_l = \frac{1 - F_{t_1}\left(2\gamma_{1,2}\sqrt{\frac{(\nu+2)(1+\rho)}{2}}\right)}{1 - T_{f_1}\left(\lambda_1\sqrt{\nu+1}\right)}\lambda_{t_1(\nu+1)} \tag{3.95}$$

$$\lambda_u = \frac{F_{t_1}\left(2\gamma_{1,2}\sqrt{\frac{(\nu+2)(1+\rho)}{2}}\right)}{T_{f_1}\left(\lambda_1\sqrt{\nu+1}\right)}\lambda_{t_1(\nu+1)}, \tag{3.96}$$

where $\gamma_{1,2} = \gamma_1 = \gamma_2$. Kollo et al. [2017] show that in the case of skewness parameters not being equal, the result from 3.95 can be used as a lower bound for the tail dependance coefficient, with $2\gamma_{1,2}$ replaced by $\gamma_1 + \gamma_2$.

A proper discussion about choosing an appropriate skew extension of the symmetric t distribution is beyond the scope of this work. But we would like to draw attention to a few points, without fully discussing them. Firstly, the tail behaviour of the elliptical skew-t distribution is symmetric, and this might be deemed to be undesirable. Secondly, in some applications the tail dependence coefficient might be too extreme as a measure of tail dependence. In fact, this idea was reflected also in Fung and Seneta [2010a]: "This sort of difference [different tail dependence coefficients obtained for the data under different model assumptions - author's comment] means tail dependence has to be interpreted as a property of the model rather than reflecting the real dependence structure in the data set. This leads one to question positivity of λ_L (...) as a measure of lower tail dependence. It is too extreme a measure, as substantial lower tail dependence may be present even when the limit is zero".

Chapter 4

Inference Procedures

> *"Nonsense," said the witcher. "And what's more, it doesn't rhyme. All decent predictions rhyme."*
>
> Andrzej Sapkowski, *The Last Wish.*

4.1 Assessing hypothesis tests

Let observations $\mathbf{x} = (x_1, ..., x_n)$ be realisations of random variables $\mathbf{X} = (X_1, ..., X_n)$, and lets assume that the **probability density function (pdf)** of $\mathbf{X}$, which we denote by $\pi(\mathbf{X}; \theta)$ belongs to a **family of distributions** $\mathcal{P}_\theta : \theta \in \Theta$, parametrised by θ which belongs to a **parameter space** Θ.

Definition 37 *Hypothesis Test*

A hypothesis test is a statement about a population parameter, [Casella and Berger, 2002].

Silvey adds to the definition, that a hypothesis test is a statement which implies that the true distribution $\pi(X; \theta)$ belongs to a subset of the family of distributions $\mathcal{P}_\theta : \theta \in \Theta$, [Silvey, 2017]. The hypothesis can subsequently be associated with that subset, and we can talk about the hypothesis of $\theta \in \omega$, where $\omega \subset \Theta$.

Definition 38 *Null and Alternative Hypotheses*

The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted H_0 and H_1 respectively, [Casella and Berger, 2002].

The general for of the null and alternative hypotheses will be, using the notation we introduced above: $H_0 : \theta \in \omega$ versus $H_1 : \theta \notin \omega$, or if we denote the complement of $\omega \subset \Theta$ by $\omega^c \subset \Theta$, then we can also write $H_1 : \theta \in \omega^c$.

Let $\pi^A(\mathbf{X}, \theta_A)$ and $\pi^B(\mathbf{X}, \theta_B)$ represent the density functions for the two models $\mathcal{M}_A$ and $\mathcal{M}_B$,

respectively. If we are interested in figuring out which model is the correct one, we will test for:

$$H_A : \left\{\pi^A\left(\mathbf{X}, \theta_A\right), \theta_A \in \Theta^A\right\} \tag{4.1}$$

$$H_B : \left\{\pi^B\left(\mathbf{X}, \theta_B\right), \theta_B \in \Theta^B\right\}. \tag{4.2}$$

In the case above, there is no natural null hypothesis. One could perform two tests, with each of the hypotheses treated as a null hypothesis and tested against the other. There are four potential outcomes: (i) H_A rejected against H_B, but not vice versa, (i) H_B rejected against H_A, but not vice versa, (i) H_A rejected against H_B and H_B rejected against H_A, (i) neither of the hypotheses rejected against the other. Only the first two outcomes are straightforward to interpret. This would, however, simplify, if we knew that the densities of the two models belonged to the same class of distributions, and if the parameter sets were complementing each other ($\Theta_A \cup \Theta_B = \Theta, \Theta_A = \Theta_B^c$), or containing one another ($\Theta_A \subset \Theta_B$). The latter case will be called nested models.

Definition 39 *Nested Models*

Two models, $\mathcal{M}_A$ parametrised by θ_A and $\mathcal{M}_B$ parametrised by θ_B, are said to be nested if it is possible to derive one from another by means of parametric restriction, [Clarke, 2001].

Example 1 *Nested Models for linear regression*

Let $\{Y_t\}$ be a time series that we want to explain by a linear model of multivariate $\{\mathbf{X}_t\}$, such that $\mathbf{X}_t = \left[X_{1,t}\, X_{2,t}\, X_{3,t}\right]^T$. Four models (hypotheses) are considered:

$$\mathcal{M}_1 : \quad Y_t = \beta_1 X_{1,t} + \epsilon_{1,t} \qquad \mathcal{M}_1 \text{ nested in } \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4 \tag{4.3}$$

$$\mathcal{M}_2 : \quad Y_t = \beta_1 X_{1,t} + \beta_2 X_{2,t} + \epsilon_{2,t} \qquad \mathcal{M}_2 \text{ nested in } \mathcal{M}_4 \tag{4.4}$$

$$\mathcal{M}_3 : \quad Y_t = \beta_1 X_{1,t} + \beta_3 X_{3,t} + \epsilon_{3,t} \qquad \mathcal{M}_3 \text{ nested in } \mathcal{M}_4 \tag{4.5}$$

$$\mathcal{M}_4 : \quad Y_t = \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + \epsilon_{4,t} \tag{4.6}$$

In the example above, there are five pairs of nested models, while models $\mathcal{M}_2$ and $\mathcal{M}_3$ could be seen as partially nested.

Garthwaite et al. [2002] gives a definition of nested models that occurs naturally in the case of linear regression, as in the case of the Example (1).

Definition 40 *Nested Models for linear regression*

Model $\mathcal{M}_2$ has q_2 parameters and mode $\mathcal{M}_1$ has q_1 parameters, where $q_2 > q_1$. Let $\mathbf{X}_1$ denote model matrix for $\mathcal{M}_1$ and $\mathbf{X}_2$ the model matrix for $\mathcal{M}_2$. Model $\mathcal{M}_1$ is nested in the model $\mathcal{M}_2$ if the columns of $\mathbf{X}_1$ are contained within the linear span of the columns of $\mathbf{X}_2$, [Garthwaite et al., 2002].

The concept of nested models will be crucial for the GPC framework. The GPC approach, introduced in the Part (II) Chapter (5), will describe and test causality by comparing two nested GP models. The nesting will be important for two reasons: firstly, it will allow to either include or exclude the potential causal effect of one time series on another, secondly, it will allow to use GLRT which requires nested models. Below is an example of nested GP models.

Intuitively, we could say that a GP model A is nested in model B, as the input space of model A is embedded in input space of model B, but the definition 39 is formulated in terms of embedding of model parameter spaces, not input spaces. Formulating our Gaussian Process models A and B in such a way that they are nested according to the above definition is not always possible. This is because for the above definition of nested models we require the mean and covariance function to have parameters that correspond to the dimensionality of the input space, or that correspond to the inclusion or not of the input X.

In practice, when we talk about nested models, we talk about mean and kernel functions allowing the nested model representation. The simplest example of how mean and kernel functions can allow nested models are for linear mean and kernel functions.

Example 2 *Nested GP models*
Define $\mu_t([X_{t-1}, Y_{t-1}, Z_{t-1}]) = a_1X_{t-1} + a_2Y_{t-1} + a_3Z_{t-1}$, *which under restriction* $a_1 = 0$ *will become equivalent to a mean* $\mu_t([Y_{t-1}, Z_{t-1}]) = a_2Y_{t-1} + a_3Z_{t-1}$, *defined on the parameter space* $[Y_{t-1}, Z_{t-1}]$. *Analogously, for the linear kernel:*

$$k_{t,t'}([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{t'-1}, Y_{t'-1}, Z_{t'-1}]) = [X_{t-1}, Y_{t-1}, Z_{t-1}] \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{bmatrix}$$

restriction $A_{1,1}, A_{1,2}, A_{1,3}, A_{2,1}, A_{2,2}, A_{2,3}, A_{3,1} = 0$ *will make this kernel equivalent to a linear kernel defined on* $[Y_{t-1}, Z_{t-1}]$ *with parameters* $A_{2,2}, A_{2,3}, A_{3,2}, A_{3,3}$.

A popular kernel function that does not allow nested models is squared exponential kernel $k(w, w') = \sigma_f^2 \exp\left(-\frac{(w-w')^T(w-w')}{2l^2}\right)$, which however can be represented in an ARD form that does allow nested models (see Table 3.1)

$$k(w, w') = \sigma_f^2 \exp\left(-\frac{1}{2}(w - w')^T diag\left(\left[l_1^2, ..., l_n^2\right]\right)(w - w')\right),$$

Or more general, instead of $(w - w')$ we can use $D(w, w')$, where $D_i(w, w')$ is any function that is a distance between w_i and w_i'.

The Theorem 2 in Subsection 3.1.1 shows that each kernel has a basis function representation, so one could attempt to express the notion of nesting in terms of embedding of the basis function spaces.

However that representation can be infinite dimensional (as is the case for the squared exponential kernel) and might not be practical. So although in theory we're guaranteed to always have a nested model representation, we might not always be able to use it.

Definition 41 *(Type I error, Type II error)*
*Rejection of the true hypothesis H_0 is called a **Type I error**, and accepting false hypothesis H_0 is called **Type II error**, [Garthwaite et al., 2002].*

Definition 42 *(Significance level, Power of the test)*
***Significance level**, is the probability of a Type I error. **Power of the test** equals* 1 *minus probability of the Type II error, or equivalently: the probability of not accepting the null hypothesis, when it is false.*

If we look at the hypothesis testing from the point of view of (binary) classification problems, then the hypotheses can be understood as statement about participation in one of the two classes. We will then have two true classes, say $\{1, -1\}$, and two possible hypothesised classes. The result of the hypothesis assessment can then give on of the four possible outcomes: **true positive (TP), true negative (TN), false positive (FP) and false negative (FN)**. Below are popular ratios that can be used to asses the classifier: true positive rate (recall, sensitivity), false positive rate, specificity, precision

$$\text{true positive rate} = \frac{\text{true positives}}{\text{total positives}} = \frac{\text{true positives}}{\text{true positives + false negatives}} \tag{4.7}$$

$$\text{false positive rate} = \frac{\text{false negatives}}{\text{total negatives}} = \frac{\text{false negatives}}{\text{true negatives + false positives}} \tag{4.8}$$

$$\text{specificity} = \frac{\text{true negatives}}{\text{total negatives}} = \frac{\text{true negatives}}{\text{true negatives + false positives}} \tag{4.9}$$

$$\text{precision} = \frac{\text{true positives}}{\text{predicted positive}} = \frac{\text{true positives}}{\text{true positives + false positive}} \tag{4.10}$$

If we know the distributions of the two classes, then the TPR and FPR can be written in terms of cdfs. Let us assume that $\pi_1(x)$ is the density function of X, if it belongs to the positive class, and $\pi_0(x)$ is if it does not.

$$\text{true positive rate:} \qquad TPR(T) = \int_T^\infty \pi_1(x)dx \tag{4.11}$$

$$\text{false positive rate:} \qquad FPR(T) = \int_T^\infty \pi_0(x)dx \tag{4.12}$$

Receiver operating characteristics (ROC) graphs are two-dimensional graphs, commonly used in classification models to quantify the accuracy with which a model can discriminate between two classes, but also a trade-off between the benefits (TP) and costs (FP) for a classifier. The TPR is plotted on the Y axis, as a function of the threshold level T for a classifier, and the FPR is plotted on X axis, also as a function of threshold level T.

An example of four sets of ROC curves is given in Figure (4.1). When interpreting the results shown by the ROC curve, there are several points in the graph that have very clear interpretation. The point (0,0) represents a strategy in which all points are labelled as negative, and therefore both TPR and FPR are zero regardless of the threshold. When all points are always classified as positive, the result would be a point (1,1). A perfect classification, one with all true positives and no false positives, will be represented by point (0,1). A diagonal line from (0,0) to (1,1) would represent a random classification strategy of labelling a point according to $\mathbf{1}_{U>T}$, for threshold T and $U \sim \text{Uniform}(0, 1)$.

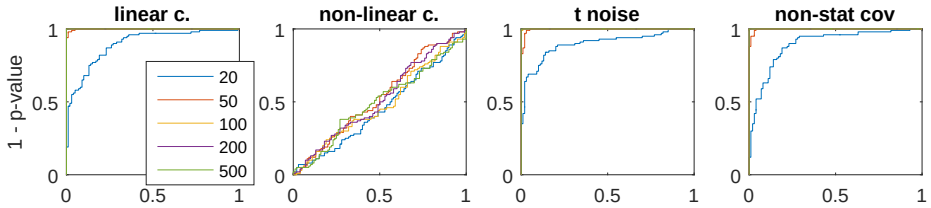


Figure 4.1: ROC curves for the data sets 1-4 from the table, calculated with (linear) Granger causality, tested with the GCCA toolbox.

See Zou et al. [2007], Hillis and Metz [2012].

4.2 Likelihood Ratio Test

The likelihood ratio test (LRT) compares the goodness of fit of two nested models based on the ratio of their likelihoods. This ratio is found by maximisation over the entire parameter space – for one of the models, and constrained parameter set – for the other model.If the null hypothesis of the restricted model being true is supported by the observation, then the maximum likelihood for this model over all available parameters will not differ much from the maximum likelihood for the unrestricted model.

Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N$ be a random sample from a distribution with pdf $f(\mathbf{x}; \theta)$, and suppose that we wish to test

$$H_0 : \theta \in \omega \qquad \text{vs} \qquad H_1 : \theta \in \Omega - \omega \tag{4.13}$$

Then define:

$$\lambda(\mathbf{x}) = \left\{ \frac{\max_{\theta \in \omega} L(\theta; \mathbf{x})}{\max_{\theta \in \Omega} L(\theta; \mathbf{x})} \right\}, \tag{4.14}$$

where $L(\theta; x) = \pi(x; \theta)$ is likelihood function. For some constant A, we can use a test with critical region $\lambda \leq A$.

Theorem 15 *Neyman–Pearson lemma*

- *Existence. Given a null hypothesis $H_0 : \theta \in \omega$ and the alternative hypothesis $H_1 : \theta \in \Omega - \omega$, there exist a test statistic L and a constant k such that:*

  1. $\mathbb{E}\lambda(\mathbf{x}) = \alpha$;
  2. *L has a form:*

$$L(x) = \begin{cases} 1 \text{ if } \left\{ \frac{\max_{\theta \in \omega} L(\theta;\mathbf{x})}{\max_{\theta \in \Omega} L(\theta;\mathbf{x})} \right\} > k \\ 0 \text{ if } \left\{ \frac{\max_{\theta \in \omega} L(\theta;\mathbf{x})}{\max_{\theta \in \Omega} L(\theta;\mathbf{x})} \right\} < k \end{cases} \tag{4.15}$$

- *Sufficiency. If L satisfies (1) and (2) for some constant k, then L is the most powerful at level α.*
- *Necessity. If a test $L*$ is most powerful at a level α, then it satisfies (2) for some level k, and it also satisfies (1) unless there exists a test strictly less than α with power 1.*

By Wilks' theorem, LRT has an asymptotic χ2-distribution under the null hypothesis.

4.3 Generalized Likelihood Ratio Test (GLRT)

The GLRT is a composite hypothesis test that can be used in the case of nested hypothesis if the parameters are unknown and need to be estimated. Below we describe the test, using notation from Garthwaite et al. [2002]. The GLRT gives us asymptotic distribution of the test statistics, but it requires that the hypotheses are nested – what can be expressed in terms of restriction on mean and covariance formulations.

Let $X_1, X_2, ..., X_N$ be a random sample from a distribution with pdf $f(x; \theta)$, and suppose that we wish to test

$$H_0 : \theta \in \omega \qquad \text{vs} \qquad H_1 : \theta \in \Omega - \omega \tag{4.16}$$

Then define:

$$\lambda = \left\{ \max_{\theta \in \omega} L(\theta; x) / \max_{\theta \in \Omega} L(\theta; x) \right\}, \tag{4.17}$$

where $L(\theta; x) = p(x; \theta)$ is likelihood function. For some constant A, we can use a test with critical region $\lambda \leq A$.

If we define d as the difference in dimensionality of H_0 and $H_0 \cup H_1$, then we have that under the null the asymptotic distribution of the test statistic is distributed according to:

$$-2 \log \lambda \sim \chi^2_d. \tag{4.18}$$

If nested model representation is not practical, then GLRT test should not be used.

4.4 Non-nested models and alternatives to GLRT

There are several approaches for non-nested models: modified (centered) loglikelihood ratio procedure – Cox procedure, "comprehensive model approach", "encompassing procedure", Vuong closeness test: likelihood-ratio-based test for model selection using the Kullback-Leibler information criterion.

We refer the reader to the following papers (and references therein): Vuong [1989], MacKinnon [1983], Pesaran and Weeks [2001] and Wilson [2015].

4.5 Permutation tests

Let us, first of all, emphasise that in the general case the causality measures introduced before should not be used as absolute values but rather serve the purpose of comparison. While we observe that, on average, increasing the strength of coupling increases the value of causality, there is a large deviation in results unless the data has been generated with linear dependence and small noise. Consequently, we need a way of assessing the significance of the measure as a way of assessing significance of the causal relationship itself. To achieve this goal we shall use permutation tests, following the approach in Amblard et al. [2012b], Sun [2008], Seth and Principe [2011].

By permutation test we mean a type of statistical significance test in which we use random permutations to obtain the distribution of the test statistic under the null hypothesis. We would like to compare the value of our causality measure on the analysed data and on "random" data and conclude that the former is significantly higher. We expect that destroying the time ordering should also destroy any potential causal effect, since statistical causality relies on the notion of time. Therefore we create the distribution of H_0 by reshuffling y, while keeping the order of x and z intact. More precisely, let $\pi_1, ..., \pi_{n_r}$ be a set of random permutations. Then instead of y_t we consider $y_{\pi_j(t)}$, obtaining a set of measurements $G_{Y_{\pi_j}\to X\|Z}$ that can be used as an estimator of the null hypothesis $G^0_{Y\to X\|Z}$. We will accept the hypothesis of causality only if, for most of the permutations, the value of the causality measure obtained on the shuffled (surrogate) data is smaller than the value of causality measure of original data. This is quantified with a p-value defined as follows:

$$p = \frac{1}{n_r}\sum_{j=1}^{n_r} \mathbf{1}(G_{Y_{\pi_j}\to X\|Z} > G_{Y\to X\|Z}). \tag{4.19}$$

Depending on the number of permutations used we suggest to accept the hypothesis of causality for the level of significance equal to 0.05 or 0.01. In our experiments we report either single p-values or sets of p-values for overlapping moving windows. The latter is particularly useful when analysing noisy and non-stationary data. In the cases where not much data is available we do not believe that using any kind of subsampling (as proposed by Sun [2008], Amblard et al. [2012b], Seth and Principe [2011]) will be beneficial as far as the power of the tests is concerned.

Part II

New Perspectives on Causality

Representation and Inference

Chapter 5

Characterising Causality With Gaussian Process Models

> *"But what am I going to see?"*
> *"I don't know. In a certain sense, it depends on you."*
>
> Stanislaw Lem, *Solaris.*

In this chapter we describe the representative model used throughout the rest of the thesis. Building on the Gaussian Process (GP) autoregressive time series representation introduced in the Section 3.1.2.1, we show how to use such representation to formulate causal hypotheses and test for statistical causality.

5.1 Semi Parametric Non-Linear Time Series Models

When performing inferential tests for statistical causality one will typically compare two alternative model hypotheses. We have already seen in the Section (1), that such hypotheses can be formulated in multiple ways, see Equations (1.5 - 1.6, 1.20, 1.22,1.23, 1.24, B.6 - B.7). We have also observed how with different modelling approaches the choice between the two hypotheses is seen as a choice between two models, which for nested models means a parametric restriction (see: Definition 39).

We begin by defining a GP representation for autoregressive time series, as this will serve as our base distribution (process) to characterise different examples of causality model structures (please refer to Equations 3.31, 3.31 and 3.32). We consider three multivariate time series denoted generically by $\{\mathbf{X}_t\} \in \mathbb{R}^p, \{\mathbf{Y}_t\} \in \mathbb{R}^{p'}$ and $\{\mathbf{Z}_t\} \in \mathbb{R}^{\bar{p}}$, which will be treated as column vectors. Below, we introduce notation that will allow convenient matrix operations, and that will be used throughout also in the later sections.

Below, we introduce notation that will allow convenient matrix operations, and that will be used

throughout also in the later sections:

$$
\begin{aligned}
&\mathbf{Y}_t \in \mathbb{R}^{p'}, && p' \times 1 \text{ column vector}\\
&\mathbf{Y}_{1:T} := [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_T]^{\mathrm{T}}, && T \times p'\\
&\mathbf{Y}_t^{-l} := \left[\mathbf{Y}_{t-l+1}^{\mathrm{T}}, \mathbf{Y}_{t-l+2}^{\mathrm{T}}, ..., \mathbf{Y}_t^{\mathrm{T}}\right], && 1 \times (lp')\\
&\mathbf{Y}^{-l} := \mathbf{Y}_{1:T}^{-l} = [\mathbf{Y}_{1-l+1:T-l+1}, \mathbf{Y}_{1-l+2:T-l+2}, ..., \mathbf{Y}_{1:T}], && T \times (lp')\\
&\mathbb{Q}_t := \left[\mathbf{X}_t^{\mathrm{T}}, \mathbf{Y}_t^{\mathrm{T}}, \mathbf{Z}_t^{\mathrm{T}}\right] \text{ for model B}, && 1 \times (p + p' + \bar{p})\\
&\mathbb{Q} := \left[\mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}\right] \text{ for model B}, && T \times (kp + lp' + m\bar{p}))
\end{aligned}
$$

We nonparametrically model the time series $\{\mathbf{Y}_t\}$ as realizations from a Gaussian Process with additive Gaussian noise:

$$
\begin{aligned}
\mathbf{Y}_t = f(\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) + \epsilon_t, \qquad & f(\cdot) \sim \mathcal{GP}(\mu, k; \boldsymbol{\theta}),\\
& \epsilon_t \sim \mathcal{N}(0, \sigma_t^2),
\end{aligned}
\tag{5.1}
$$

with the mean function $\mu : \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$ and the covariance function $k : \mathbb{R}^{kp+lp'+m\bar{p}} \times \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$. and associated mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$.

Please note, that the mean and covariance functions will often not depend explicitly on time. Rather, they will depend on time implicitly, through the temporal structure on the inputs. In the case of mean and covariance functions depending explicitly on time, it would be more appropriate to write, for example, $\mu\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]; t\right)$ and $k^{XY}\left(\left[X_{t_1-1}^{-k}, Y_{t_1-1}^{-l}, Z_{t_1-1}^{-m}\right], \left[X_{t_2-1}^{-k}, Y_{t_2-1}^{-l}, Z_{t_2-1}^{-m}\right]; t_1, t_2\right)$.

5.2 Testing for Causality - Introducing Two Models

The two model structures are generically represented in general as multidimensional GP time series models observed with additive Gaussian noise and denoted by Model A and Model B below in Equations 5.2 and the Equations 5.3 below as Gaussian Process models with White Noise:

$$
\begin{aligned}
\text{Model A: } \mathbf{Y}_t =& f_A(\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) + \epsilon_t^A, && f_A \sim \mathcal{GP}(\mu_A, k_A; \boldsymbol{\theta}_A) \qquad (5.2)\\
& && \epsilon_t^A \sim \mathcal{N}(0, \sigma_{A,t}^2)\\
\text{Model B: } \mathbf{Y}_t =& f_B(\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) + \epsilon_t^B, && f_B \sim \mathcal{GP}(\mu_B, k_B; \boldsymbol{\theta}_B) \qquad (5.3)\\
& && \epsilon_t^B \sim \mathcal{N}(0, \sigma_{B,t}^2).
\end{aligned}
$$

where the different dimensionality of the input spaces result in the following dimensionality of the domains of the mean and covariance functions: $\mu_A : \mathbb{R}^{lp'+m\bar{p}} \to \mathbb{R}$, $\mu_B : \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$ and covariance

functions $k_A : \mathbb{R}^{lp'+m\bar{p}} \times \mathbb{R}^{lp'+m\bar{p}} \to \mathbb{R}$, $k_B : \mathbb{R}^{kp+lp'+m\bar{p}} \times \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$.

We assume the mean and covariance functions, μ_A, k_A and respectively μ_B, k_B, have similar function forms and only differ in dimensionality and hyperparameters. Having defined these two models we may now state the form of the hypotheses for testing for non-causality (lack of causality) in non-linear times series.

5.3 Testing for Causality – Distributional Test

The test that allows as to compare two models from the Equations 5.2 and 5.3 is fundamentally a test comparing two distributions – conditional distribution of the time series Y conditioned on inputs from either of the two models. As it was already mentioned, we never actually confirm the statistical causality, but rather reject lack of causality (test for non-causality).

Under such a test, the null hypothesis is that there is no causal relationship from time series X to Y, and including the past of X does not improve the prediction of Y. Given the model formulations, this means equality of conditional distribution of Y, conditioning on either set of explanatory variables (analogously to Equations 1.20):

$$H_0 : \quad \pi(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) = \pi(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) \tag{5.4}$$

$$H_1 : \quad \pi(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) \neq \pi(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}). \tag{5.5}$$

The distributions above can be obtained in closed form only in the case of additive Gaussian noise, or in cases where there is no assumed additive noise in Model A or model B.

Since a Gaussian process is also specified by its sufficient mean and covariance functions, therefore testing for equality of distributions will be equivalent to testing for equality of the mean functions and the covariance functions. Hence, the convenient feature of the causality testing framework developed from the Gaussian process framework we propose is that these general distributional statements about population quantities in the null and alternative hypotheses are equivalent to the following population statements on mean and covariance functions.

$$\begin{aligned} H_0 :& \exists\, k_A(.,.) \in \mathcal{M},\, \mu_A \in \mathbf{C}(\mathbb{R}^d)\ \forall\, t_1, t_2 \in \{l+1, ..., T\} \\ & k_B\left(\left[\mathbf{X}_{t_1-1}^{-k}, \mathbf{Y}_{t_1-1}^{-l}, \mathbf{Z}_{t_1-1}^{-m}\right], \left[\mathbf{X}_{t_2-1}^{-k}, \mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right) = k_A\left(\left[\mathbf{Y}_{t_1-1}^{-l}, \mathbf{Z}_{t_1-1}^{-m}\right], \left[\mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right) \\ & \mu_B\left(\left[\mathbf{X}_{t_2-1}^{-k}, \mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right) = \mu_A\left(\left[\mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right) \end{aligned} \tag{5.6}$$

$$\begin{aligned} H_1 :& \neg\exists\, k(.,.) \in \mathcal{M}, \mu \in \mathbf{C}(\mathbb{R}^d)\ \forall\, t_1, t_2 \in \{l+1, ..., T\} \\ & k_B\left(\left[\mathbf{X}_{t_1-1}^{-k}, \mathbf{Y}_{t_1-1}^{-l}, \mathbf{Z}_{t_1-1}^{-m}\right], \left[\mathbf{X}_{t_2-1}^{-k}, \mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right) = k_A\left(\left[\mathbf{Y}_{t_1-1}^{-l}, \mathbf{Z}_{t_1-1}^{-m}\right], \left[\mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right) \\ & \mu_B\left(\left[\mathbf{X}_{t_2-1}^{-k}, \mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right) = \mu_A\left(\left[\mathbf{Y}_{t_2-1}^{-l}, \mathbf{Z}_{t_2-1}^{-m}\right]\right). \end{aligned} \tag{5.7}$$

If we restrict ourselves to certain classes of mean and covariance function so that the Model A is

nested in the Model B (Definition 39), then the above hypotheses can be tested with the Generalized Likelihood Ratio Test.

5.3.1 Generalised Likelihood Ratio Test for Testing Causality

One of the main advantages of Gaussian Process models is that many of the required probability distributions for evaluating the test statistic in the GLRT are easily accessible: in many cases, such as those described in this thesis, they can be calculated in closed form, in other cases there are effective approximation methods.

Lets refer to the null hypothesis of non-causality as it was formed in the Equation 5.4. Non-causality was expressed in terms of equality of two conditional distributions: $\pi(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) = \pi(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m})$. This is equivalent to equality of the two marginal log-likelihoods:

$$\log \pi(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) = \log \pi(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A), \tag{5.8}$$

and it leads to the definition of a causality test statistic $L_{X \to Y|Z}$:

$$L_{X \to Y|Z} = \max_{\boldsymbol{\theta}_B} \log \pi(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) - \max_{\boldsymbol{\theta}_A} \log \pi(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A) \tag{5.9}$$

This test statistic was first proposed by Amblard et al. [2012a].

We assume here additive Gaussian errors, which allows us to calculate the marginal likelihoods analytically. For the calculations please refer to the appendix A.3. The resulting distributions are:

$$\pi(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A) = \mathcal{N}(\mathbf{Y}; \boldsymbol{\mu}_A, K_A + \boldsymbol{\Sigma}^A) \tag{5.10}$$

$$\pi(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) = \mathcal{N}(\mathbf{Y}; \boldsymbol{\mu}_B, K_B + \boldsymbol{\Sigma}^B). \tag{5.11}$$

If we use the hat notation for MLE estimators, then the test statistic is given by:

$$\begin{aligned}
\hat{L}_{X \to Y|Z} = &- (Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_B))^{\mathrm{T}} \left(\oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{B,t}} + \hat{\sigma}_B^2 \mathbf{I}_{Tp' \times Tp'} \right)^{-1} (Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_B)) \\
&+ (Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_A))^{\mathrm{T}} \left(\oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{A,t}} + \hat{\sigma}_A^2 \mathbf{I}_{Tp' \times Tp'} \right)^{-1} (Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_A)) \\
&- \ln \left| \oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{B,t}} + \hat{\sigma}_B^2 \mathbf{I}_{Tp' \times Tp'} \right| + \ln \left| \oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{A,t}} + \hat{\sigma}_A^2 \mathbf{I}_{Tp' \times Tp'} \right|,
\end{aligned} \tag{5.12}$$

where $\mathbb{Q}_{B,t} = \left[\mathbf{X}_t^{\mathrm{T}}, \mathbf{Y}_t^{\mathrm{T}}, \mathbf{Z}_t^{\mathrm{T}} \right]$, $\mathbb{Q}_{A,t} = \left[\mathbf{Y}_t^{\mathrm{T}}, \mathbf{Z}_t^{\mathrm{T}} \right]$, and $Vec(.)$ denotes conversion of a matrix into a vector.

In the Equation 5.12 we present a general form of the test statistic for multivariate time series, and in the special case of a univariate time series $\mathbf{Y}$ this simplifies to a form from the Equation 5.13. Distinguishing between the two definitions can also be seen as a distinction between joint causality and

marginal causality.

$$\hat{L}_{X\to Y|Z} = -(\mathbf{Y}-\hat{\boldsymbol{\mu}}_B)^T\left(\hat{\mathbf{K}}_B+\hat{\sigma}_B^2\mathbf{I}\right)^{-1}(\mathbf{Y}-\hat{\boldsymbol{\mu}}_B) - \log|\,\hat{\mathbf{K}}_B+\hat{\sigma}_B^2\mathbf{I}\,| \\ +(\mathbf{Y}-\hat{\boldsymbol{\mu}}_A)^T\left(\hat{\mathbf{K}}_A+\hat{\sigma}_A^2\mathbf{I}\right)^{-1}(\mathbf{Y}-\hat{\boldsymbol{\mu}}_A) + \log|\,\hat{\mathbf{K}}_A+\hat{\sigma}_A^2\mathbf{I}\,|\,. \tag{5.13}$$

Under certain regularity conditions, with the assumptions of conditional independence of $\mathbf{Y}_t \mid \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}$ for all t, and with the assumption that models A and B are nested (Definition (39) and Section (4.3)) we can treat $L_{X\to Y|Z}$ as a GLRT and use the asymptotic results:

$$H_0: \qquad 2\hat{L}_{X\to Y|Z} \sim \chi_q^2 \quad \text{as } T\to\infty, \tag{5.14}$$

where q is the difference in dimensionality between the parameter space for $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$.

5.3.2 Statistical Causality and the Model Choice

As we have already discussed before, analysing causal structure with Gaussian processes hasn't been done in the likelihood ratio framework, and vice versa: existing approaches to casuality testing with likelihood ratio type of tests have not been utilising GPs. What we propose, is a way to construct model nesting that allows for application of the likelihood ratio test and thus formulation of the test statistic that can be written in a closed form, can be computed point-wise, and is efficient to compute.

There are numerous advantages of using GPs, beginning with: ease of optimisation and interpretability of hyperparameters, flexibility, richness of covariance functions, allowing for various model structures. Using a likelihood ratio type test with a GP is a very natural choice, as estimating GP model parameters is often done on the basis of maximising likelihood, and therefore this estimation can be incorporated into the compound version of the likelihood ratio test (Generalised Likelihood Ratio Test, GLRT). From Gaussian variables, GPs inherited the property of being fully specified by the mean and the covariance, and so testing for model equivalence inherently means testing for equivalence of the mean and covariance functions. But many popular kernels (for example squared exponential, see Table 3.1) do not have the ARD property, and using them for a likelihood ratio test settings gives no easy way to account for causal structures in covariance. Consequently, it is using GLRT with an ARD-GP that gives a uniformly most powerful test with an unparalleled flexibility: known asymptotic distribution under the null, explicit evaluation and in a closed form, and usefulness also for misspecified models.

When using GPs for modelling statistical causality, we do not assume that the data is truly generated by a Gaussian process, and we make a distinction between the knowledge of the true model, and formulating a model that is useful for testing a causal relationship. It is crucial that GLRT is a test for model selection, and therefore it can be employed to test for a model that is most useful, rather than one that is well specified. As a result, the choice of a mean and covariance function has the biggest impact on

the presence or absence of specific statistical structures that they introduce, and thus the interpretation in terms of statistical causality.

In Chapter 8, we demonstrate the ability to detect and identify causal structures in the mean and covariance, even in the presence of different types of model misspecifications. We give examples of three time series models, one of which does not have a natural GP representation (ARFIMA, example time series model class 3, Equations: 3.55 - 3.57), and we discuss results of using our framework for testing causality in those models. For example, for the time series with long memory (ARFIMA model), our framework is still able to identify causality, and the interpretability of the parameters mean we can conclude that the introduction of long memory is reflected by the expected increase in estimates of the serial correlation. We also note how a causal effect in the trend might overshadow a causal effect in the covariance. Additionally, in Chapter 9 we show on real data that the effect on the recognition of causality of choosing different degrees of freedom in the Matern kernel are negligible, unlike the effect of incorporating different statistical structures, such as serial correlation and causality in covariance.

5.4 Irregularly Sampled Time Series

Causal analysis based on statistic causality will, typically, be applied to time series. Consequently it will often be assumed that the data is regularly sampled, and therefore the mean and covariance functions will often not depend explicitly on time. Rather, they might depend on time implicitly, through the temporal structure on the inputs.

The most popular kernel is the squared exponential kernel k^{SE}, which is stationary (Definition 9) and does not depend on the absolute value of inputs, but rather on their difference.

$$k^{SE}(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{(\mathbf{x}_p - \mathbf{x}_q)^T(\mathbf{x}_p - \mathbf{x}_q)}{2l^2}\right) \tag{5.15}$$

Matern kernel, which we we used in majority of experiments is isotropic (Definition 10)

$$k^{Matern}(d) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}\left(\sqrt{2\nu}\frac{\mathrm{d}}{l}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{\mathrm{d}}{l}\right), \tag{5.16}$$

where $\mathrm{d} = \| \mathbf{x}_p - \mathbf{x}_q \|$. These, and other example of kernels are shown in the Table (3.1).

One could however change the definitions of mean and kernel functions in to include time explicitly:

$$\mu\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]; t\right) \tag{5.17}$$

$$k\left(\left[X_{t_1-1}^{-k}, Y_{t_1-1}^{-l}, Z_{t_1-1}^{-m}\right], \left[X_{t_2-1}^{-k}, Y_{t_2-1}^{-l}, Z_{t_2-1}^{-m}\right]; t\right). \tag{5.18}$$

Mean and covariance functions that depend on time explicitly can easily be incorporated into the GPC framework.

When would the above approach naturally appear as a way of reflecting the fact that data is not measured at fixed time intervals? For example, in the case of a phenomenon that has a known time-dependent pattern. Instances of such phenomena occur in everyday life: the average amount of precipitation varies with the season; the change in child's weight over time follows an empirically measured distribution. In the financial domain, zero-coupon bond price converges to its face value as the product approaches maturity; the risk of an investment changing in time can be modelled by a GP with a covariance function depending on time.

To deal with time series which are both irregularly sampled and sparse in a context of classification, Li and Marlin [2015] propose to first re-represent the time series through a GP posterior induced under a GP regression model, and then to define kernels over the space of GP posteriors and apply standard kernel-based classification.

Cunningham et al. [2012] propose a method to deal with multivariate time series which are observed with different time markers. In the case of multiple time markers, it might not be meaningful to temporarily align observations, so they treat each of the (originally univariate) time series as a multidimensional time series where each input dimension is time with respect to a particular marker. Lets say a collection of N time series is observed $\left\{y^{(n)}(t)\right\}_{n=1:N}$ with time markers $\left\{m_k^{(n)}\right\}_{n=1:N,k=1:K}$. Subsequently, Cunningham et al. [2012] suggest to define kernels as follows:

$$k_{TM}\left(t_i^{(p)}, t_j^{(q)}\right) = k\left(t_i^{(p)} - \begin{bmatrix} m_1^{(p)} \\ \vdots \\ m_K^{(p)} \end{bmatrix}, t_j^{(q)} - \begin{bmatrix} m_1^{(q)} \\ \vdots \\ m_K^{(q)} \end{bmatrix}\right).$$

Chapter 6

Characterising Causality With Warped Gaussian Process Models

> *Come, let us hasten to a higher plane,*
> *Where dyads tread the fairy fields of Venn,*
> *Their indices bedecked from one to n,*
> *Commingled in an endless Markov chain!*
>
> Stanislaw Lem, *The Cyberiad.*

Building on from the Multi-Output GP model first introduced in Section 3.1, we take one step further to generalise the Multi-Output GP framework by widening the class of joint distributions considered. Based on the warping transformation from Section 3.3 we first construct warped Multi-Output GPs, and then show how they can be used as a framework for modelling statistical causality. We finish the chapter by discussing properties of tail dependence, and how this can interrelate with causal dependence.

Assume we have two (unobserved) univariate time series of interest, for which $\{X_t\}, \{Y_t\}$ will denote the marginal and unobserved distributions (modelled with Gaussian processes), while $\left\{\tilde{X}_t\right\}, \left\{\tilde{Y}_t\right\}$ will refer to observed joint distribution that can be described with a warped Gaussian process model related to the skew-t distribution. We also have side information – observed time series $\{Z_t\}$, that contains any additional information about the "state of the world" and is described with a Gaussian process model.

We want to model the time series $\{X_t\}, \{Y_t\}$ as an autoregressive process, depending on lagged values of $\mathbf{X}_{t-1}^{-k} = [X_{t-k}, \cdots, X_{t-1}], \mathbf{Y}_{t-1}^{-l} = [Y_{t-l}, \cdots, Y_{t-1}]$ and lagged values of a side information time series $\mathbf{Z}_{t-1}^{-m} = [Z_{t-m}, \cdots, Z_{t-1}]$.

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f_X\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) \\ f_Y\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) \end{bmatrix} + \begin{bmatrix} \epsilon_{X,t} \\ \epsilon_{Y,t} \end{bmatrix}. \tag{6.1}$$

We will be assuming that the Gaussian processes have zero mean.

$$f^X \sim \mathcal{GP}\left(0, k^X\right) \qquad\qquad f^Y \sim \mathcal{GP}\left(0, k^Y\right) \tag{6.2}$$

We refer the reader to the Section 3.3.2 and in particular to the Theorem 7 and to its proof in the Appendix A.4, for an explanation of why the mean functions need to be equal to zero. The noise is uncorrelated and normally distributed.

$$\epsilon_t^X \sim \mathcal{N}(0, \sigma_X^2), \qquad\qquad \epsilon_t^Y \sim \mathcal{N}(0, \sigma_Y^2).$$

We assume that the Gaussian processes f_X, f_Y might be correlated. In line the discussion about multiple output GPs from the Section (3.1.3), the fact that f_X, f_Y are correlated can be expressed in many ways. One way is to introduce the kernel functions, $k^{XY}, k^{YX} : \mathbb{R}^{k+l+m} \times \mathbb{R}^{k+l+m} \to \mathbb{R}$ to describe the joint process of f_X, f_Y.

6.1 Warped Gaussian Processes: Normal Mean-Variance Mixture of Gaussian Processes

We have already introduced the three mean-variance mixtures and resulting three skew-t distributions. We have announced that these will be used to connect the Gaussian processes to warped Gaussian processes and in this section we present details.

We will take a GP time series model as defined in Equation (6.1) for $\{X_t\}, \{Y_t\}$ and we will modify them to produce warped GP processes $\left\{\tilde{X}_t\right\}, \left\{\tilde{Y}_t\right\}$ which will admit joint dependence structures given by the skew-t copula structures defined in Section (3.3.4). We assume that we are observing partially a time series for $\tilde{X}_t$ and $\tilde{Y}_t$ represented by the warped GP processes of the GH type. These time series are implicitly defined by latent GP processes X and Y and the specified model transform for a particular class of GH and skew-t copula 3.76. Here, variable W_t is inverse gamma distributed.

$$W_t \perp\!\!\!\perp [X_t, Y_t]^T \qquad\qquad W_t \sim IG\left(\frac{\nu}{2}, \frac{\nu}{2}\right).$$

Then the transformation from $[X_t, Y_t]^T$ to $\left[\tilde{X}_t, \tilde{Y}_t\right]^T$ is :

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \stackrel{d}{=} \mathbf{m}_t + \boldsymbol{\gamma} W_t + \sqrt{W_t}\,[X_t, Y_t]^T, \tag{6.3}$$

where: $\mathbf{m}_t, \boldsymbol{\gamma} \in \mathbb{R}^2$ and $\mathbf{m}_t$ represents the mean and $\boldsymbol{\gamma}$ – the skewness. Just like in the case of the mean, the skewness $\boldsymbol{\gamma}$ can be interpreted as a marginal skewness. The relationship between the variables is represented in Figure (7.2).

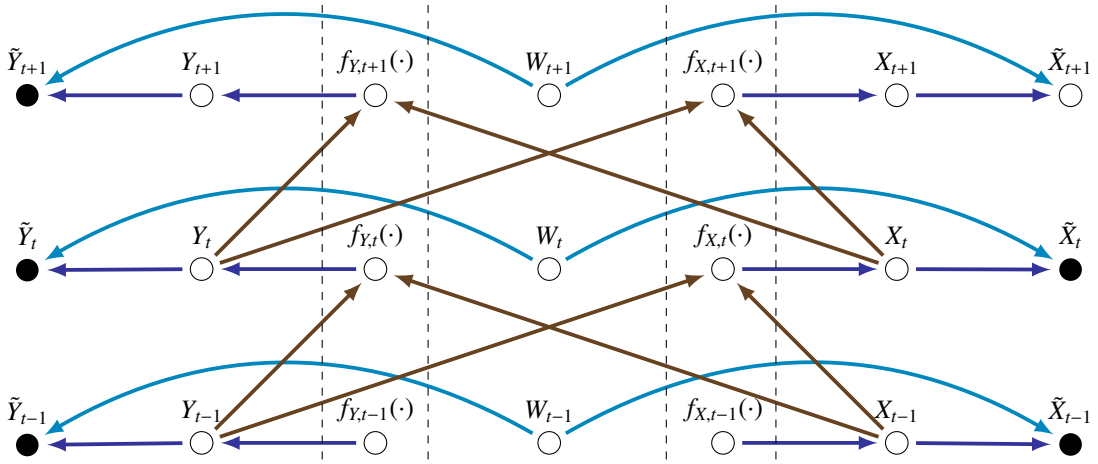


Figure 6.1: Direct Acyclic Graph (DAG) representation of the time series $\{X_t\}, \{Y_t\}$ and $\left\{\tilde{X}_t\right\}, \left\{\tilde{Y}_t\right\}$.

According to the properties of Gaussian processes, we know that conditioned on the past time series values, the $[X_t, Y_t]^T$ has a normal distribution:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} \mid \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_t), \qquad \text{where:} \quad \boldsymbol{\Sigma}_t = \begin{bmatrix} k^X + \delta_{t_1 t_2}\sigma_X^2 & k^{XY} \\ k^{YX} & k^Y + \delta_{t_1 t_2}\sigma_Y^2 \end{bmatrix} \tag{6.4}$$

We would like to point out that the conditioning on $\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}$ will often be omitted in later sections. The reasoning is that if we look at $\{X_t\}, \{Y_t\}$ as stochastic processes together with their natural filtrations, say $\mathcal{F}_t^X, \mathcal{F}_t^Y$, then it's only natural to consider a filtration $\mathcal{F}_t := \mathcal{F}_t^X \cup \mathcal{F}_t^Y$. The variables X_t and Y_t are measurable with respect to the filtration $\mathcal{F}_t$.

As a direct result of how we defined the mixture, we know that the conditional distribution of the transformed variable conditioned on the mixing variable is normally distributed:

$$\begin{bmatrix} \tilde{X}_t \\ \tilde{Y}_t \end{bmatrix} \mid W_t, \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{m}_t + \boldsymbol{\gamma} W_t, W_t \boldsymbol{\Sigma}_t). \tag{6.5}$$

We also know that the unconditional distribution of the transformed variables will be bivariate skew-t of the GH type, with location $\mathbf{m}_t$ and skewness $\boldsymbol{\gamma}$ parameters coming from the transformation, shape parameter ν coming from the mixing variable W_t, and scale parameter Σ_t coming from the Gaussian variable (Gaussian process):

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \sim St_2(\mathbf{m}_t, \boldsymbol{\Sigma}_t, \boldsymbol{\gamma}, \nu).$$

We are not making any model assumption about the time series $\{Z_t\}$, which represents the side information. Also we are not including the side information in the transformation, although this can be

done.

6.2 Alternative Normal Mean-Variance Mixture

Using generalised mean-variance mixture requires no additional steps or explanations – after all it's the same transformation as in the Subsection 6.1, but with the Inverse Gamma variable having two degrees of freedom 3.76.

The elliptical skew-t does however have more complicated transformation, as was already shown in 3.87. As a first step, we have to transform the variables X_t, Y_t into one that has a skew-normal distribution, lets call them X_t^{SN}, Y_t^{SN}:

$$\begin{bmatrix} X_t^{SN} \\ Y_t^{SN} \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} |N| + \begin{bmatrix} \sqrt{1-\theta_1^2} & 0 \\ 0 & \sqrt{1-\theta_2^2} \end{bmatrix} \begin{bmatrix} \frac{X_t - mean(X_t)}{Var(X_t)} \\ \frac{Y_t - mean(Y_t)}{Var(Y_t)} \end{bmatrix}, \tag{6.6}$$

where $N \sim \mathcal{N}(0,1)$. The transformed variable is skew-normally distributed with skewness parameters $\boldsymbol{\gamma} = [\gamma_1, \gamma_2]^T$ related to θ_1, θ_2 as follows: $\gamma_i = \theta_i / \sqrt{1-\theta_i^2}$, and with the correlation matrix $\boldsymbol{\Omega}_t$.

The second step of the transformation is more similar to what we had before:

$$\begin{aligned} &\left[\tilde{X}_t, \tilde{Y}_t\right]^T \stackrel{d}{=} \mathbf{m}_t + \sqrt{W_t}\left[X_t^{SN}, Y_t^{SN}\right]^T, && \left[X_t^{SN}, Y_t^{SN}\right]^T \sim SN(\boldsymbol{\gamma}, \boldsymbol{\Omega}_t) \\ &W_t \perp\!\!\!\perp \left[X_t^{SN}, Y_t^{SN}\right]^T, && W_t \sim IG\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \end{aligned} \tag{6.7}$$

The conditional and unconditional distributions obtained with this transformation are as follows:

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \mid W_t, N, \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{m}_t + \boldsymbol{\gamma} N W_t, W_t \boldsymbol{\Omega}_t), \tag{6.8}$$

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \sim AS\,t_2(\mathbf{m}_t, \boldsymbol{\Omega}_t, \boldsymbol{\gamma}, \nu). \tag{6.9}$$

6.3 Testing for Causality - Two Alternative Models

Once again, we look at statistical causality as a comparison of two alternative models, and testing for causality – as comparing conditional distributions in those two models. In this section we develop we show how such a test can be expressed for warped GP models. Formulation of the two alternative tests is based on a similar rationale, regardless of the form of the test that is later used: the time series whose effect is investigated needs to be absent from one of the models (lets call it model A) and present in the other (lets call it model B).

Model B - the unrestricted model. So far we have described the model:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f_X^B\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) \\ f_Y^B\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) \end{bmatrix} + \begin{bmatrix} \epsilon_{X,t} \\ \epsilon_{Y,t} \end{bmatrix} \tag{6.10}$$

$$\begin{bmatrix} \tilde{X}_t \\ \tilde{Y}_t \end{bmatrix} = \begin{bmatrix} m_{B;t}^X \\ m_{B;t}^Y \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} W_t + \sqrt{W_t} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} \tag{6.11}$$

the covariance between outputs of the multiple output GP is then equal to:

$$\mathbf{\Sigma}_{A;t} = cov([X_t, Y_t], [X_t, Y_t]) = \begin{bmatrix} k^{B;X} + \delta_{t,s}\sigma_{B;X}^2 & \rho_{t,s}^{B;XY} \\ \rho_{t,s}^{B;XY} & k^{B;Y} + \delta_{t,s}\sigma_{B;Y}^2 \end{bmatrix}.$$

And our goal is the conditional probability:

$$\pi\left(\tilde{Y}_t \mid \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}; \mathcal{M}_B\right).$$

Model A - the restricted model. We want to compare it with the model A which would be correct if X and Y where not causally dependant:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f_X^A\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) \\ f_Y^A\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) \end{bmatrix} + \begin{bmatrix} \epsilon_{X,t} \\ \epsilon_{Y,t} \end{bmatrix} \tag{6.12}$$

$$\begin{bmatrix} \tilde{X}_t \\ \tilde{Y}_t \end{bmatrix} = \begin{bmatrix} m_{B;t}^X \\ m_{B;t}^Y \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} W_t + \sqrt{W_t} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} \tag{6.13}$$

the covariance between outputs of the multiple output GP is then equal to:

$$\mathbf{\Sigma}_{A;t} = cov\left([X_t, Y_t], [X_t, Y_t]\right) = \begin{bmatrix} k^{A;X}\delta_{t,s}\sigma_{A;X}^2 & \rho_{t,s}^{A;XY} \\ \rho_{t,s}^{A;XY} & k^{A;Y}\delta_{t,s}\sigma_{A;Y}^2 \end{bmatrix}.$$

And our goal is the conditional probability:

$$\pi\left(\tilde{Y}_t \mid \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}; \mathcal{M}_A\right).$$

We say that X has no causal effect on Y if model A and B are equivalent:

$$\pi(\tilde{Y}_t \mid \tilde{X}_{t-1}^{-k}, \tilde{Y}_{t-1}^{-l}, Z_{t-1}^{-m}; \mathcal{M}_B) = \pi(\tilde{Y}_t \mid \tilde{Y}_{t-1}^{-l}, Z_{t-1}^{-m}; \mathcal{M}_A).$$

6.3.1 Obtaining the Correct Conditional Distribution

For the test above we require a conditional probability which is conditioned on the observed (transformed) $\tilde{X}_{t-1}^{-k}, \tilde{Y}_{t-1}^{-l}$. The theorem 7 gives the distribution in the case of unobserved $X_{t-1}^{-k}, Y_{t-1}^{-l}$. In this section we show that variables $\tilde{Y}_t \mid \tilde{X}_{t-1}^{-k}, \tilde{Y}_{t-1}^{-l}, Z_{t-1}^{-m}$ and $\tilde{Y}_t \mid X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}$ are equal in distribution.

According to the transformation formula from Equation 6.67, and according to how the lags were define, we have that:

$$\tilde{X}_{t-1}^{-k} = \left[\tilde{X}_{t-k}, ..., \tilde{X}_{t-1}\right] = \left[m_{t-k(1)} + \gamma_{(1)} W_{t-k} + \sqrt{W_{t-k}} X_{t-k}, ..., m_{t-1(1)} + \gamma_{(1)} W_{t-1} + \sqrt{W_{t-1}} X_{t-1}\right]$$
$$\tilde{Y}_{t-1}^{-l} = \left[\tilde{Y}_{t-l}, ..., \tilde{Y}_{t-1}\right] = \left[m_{t-l(2)} + \gamma_{(2)} W_{t-l} + \sqrt{W_{t-l}} Y_{t-l}, ..., m_{t-1(2)} + \gamma_{(2)} W_{t-1} + \sqrt{W_{t-1}} Y_{t-1}\right]$$

and so the conditional distribution:

$$\pi(\tilde{Y}_t \mid \tilde{X}_{t-1}^{-k}, \tilde{Y}_{t-1}^{-l}, Z_{t-1}^{-m}) = \pi\Big(\tilde{Y}_t \mid \left[m_{t-k(1)} + \gamma_{(1)} W_{t-k} + \sqrt{W_{t-k}} X_{t-k}, ..., m_{t-1(1)} + \gamma_{(1)} W_{t-1} + \sqrt{W_{t-1}} X_{t-1}\right],$$
$$\left[m_{t-l(2)} + \gamma_{(2)} W_{t-l} + \sqrt{W_{t-l}} Y_{t-l}, ..., m_{t-1(2)} + \gamma_{(2)} W_{t-1} + \sqrt{W_{t-1}} Y_{t-1}\right], Z_{t-1}^{-m}\Big).$$

Furthermore, we observe that:

$$Y_t = \frac{\tilde{Y}_t - m_{t(2)} + \gamma_{(2)} W_t}{\sqrt{W_t}} \qquad \Rightarrow \qquad Y_t \mid \tilde{Y}_t, W_t \sim const,$$
$$X_t = \frac{\tilde{X}_t - m_{t(2)} + \gamma_{(2)} W_t}{\sqrt{W_t}} \qquad \Rightarrow \qquad X_t \mid \tilde{X}_t, W_t \sim const.$$

Therefore if we condition on the whole history of the mixing variable W_t, and all of the relevant lags of $\tilde{X}_t, \tilde{Y}_t$, we obtain the following equivalence:

$$\pi(\tilde{Y}_t \mid \tilde{X}_{t-1}^{-k}, \tilde{Y}_{t-1}^{-l}, Z_{t-1}^{-m}, W_{t-k:t-1}) = \pi(\tilde{Y}_t \mid X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}, W_{t-k:t-1})$$

But this implies

$$\int \pi(\tilde{Y}_t \mid \tilde{X}_{t-1}^{-k}, \tilde{Y}_{t-1}^{-l}, Z_{t-1}^{-m}, W_{t-k:t-1}) \pi(W_{t-k:t-1}) dW_{t-k:t-1} =$$
$$\int \pi(\tilde{Y}_t \mid X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}, W_{t-k:t-1}) \pi(W_{t-k:t-1}) dW_{t-k:t-1}$$

which means equality on all non-zero sets, and therefore:

$$\mathbb{P}\left(\tilde{Y}_t \mid \tilde{X}_{t-1}^{-k}, \tilde{Y}_{t-1}^{-l}, Z_{t-1}^{-m} \neq \tilde{Y}_t \mid X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right) = 0.$$

6.3.2 Generalised Likelihood Ratio Test

The GLRT is a composite hypothesis test that can be used in the case of nested hypothesis if the parameters are unknown and need to be estimated. Below we describe the test, using notation from Garthwaite et al. [2002]. We are reminding the Section (4.3):

Let $X_1, X_2, ..., X_N$ be a random sample from a distribution with pdf $\pi(x; \theta)$, and suppose that we wish to test

$$H_0 : \theta \in \omega \qquad \text{vs} \qquad H_1 : \theta \in \Omega - \omega. \tag{6.14}$$

Then define:

$$\lambda = \left\{ \max_{\theta \in \omega} L(\theta; x) / \max_{\theta \in \Omega} L(\theta; x) \right\}, \tag{6.15}$$

where $L(\theta; x) = p(x; \theta)$ is likelihood function. For some constant A, we can use a test with critical region $\lambda \leq A$.

If we define d as the difference in dimensionality of H_0 and $H_0 \cup H_1$, then we have that:

$$-2 \log \lambda \sim \chi^2_d. \tag{6.16}$$

6.4 Evaluating the test statistic

We observe the data $\left\{\tilde{X}_t\right\}_{t=1}^{T}, \left\{\tilde{Y}_t\right\}_{t=1}^{T}$ and we want to asses causal dependence in the direction $\tilde{X} \rightarrow \tilde{Y}$. Causality will be tested by comparing the marginal likelihoods from the two models, which we will call model A and model B, and which will be defined in the next sections. To use a GLRT type causal test, we need to compare (an estimate) two marginal likelihoods, that arise from the models A and B:

$$H_0 : \qquad \pi(\tilde{Y}_t \mid \tilde{Y}_{t-1}, \tilde{X}_{t-1}; \mathcal{M}_B) = \pi(\tilde{Y}_t \mid \tilde{Y}_{t-1}; \mathcal{M}_A), \quad \forall t = 1, ..., T. \tag{6.17}$$

We will be estimating the likelihoods above jointly for all t:

$$H_0 : \qquad \pi(\tilde{\mathbf{Y}}_{2:T} \mid \tilde{\mathbf{X}}_{1:T-1}; \mathcal{M}_B) = \pi(\tilde{\mathbf{Y}}_{2:T}; \mathcal{M}_A). \tag{6.18}$$

For the sake of estimation, we define all of the models with two simplifications: firstly not looking at the side information $\left\{\tilde{Z}_t\right\}_{t=1}^{T}$, secondly using only one lag, i.e. in the general formulations of $X_{t-1}^{-k}, Y_{t-1}^{-l}$ we will take $k = l = 1$

6.4.1 Model A

We define two univariate Gaussian Processes:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f_X^A(X_{t-1}) \\ f_Y^A(Y_{t-1}) \end{bmatrix} + \begin{bmatrix} \epsilon_{X,t}^A \\ \epsilon_{Y,t}^A \end{bmatrix}, \qquad \begin{aligned} f_X^A &\sim \mathcal{GP}\left(0, k_{X,t_1,t_2}^A\right), \epsilon_{X,t}^A \sim \mathcal{N}(0, \sigma_{X,A}^2), \\ f_Y^A &\sim \mathcal{GP}\left(0, k_{Y,t_1,t_2}^A\right), \epsilon_{Y,t}^B \sim \mathcal{N}(0, \sigma_{Y,A}^2). \end{aligned} \tag{6.19}$$

Then the transformation from $[X_t, Y_t]^T$ to $\left[\tilde{X}_t, \tilde{Y}_t\right]^T$ is :

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \overset{d}{=} \mathbf{m}_t^A + \boldsymbol{\gamma}_t^A W_t + \sqrt{W_t}\,[X_t, Y_t]^T, \qquad \begin{aligned} &W_t \perp\!\!\!\perp [X_t, Y_t]^T \\ &W_t \sim IG\left(\frac{\nu^A}{2}, \frac{\nu^A}{2}\right), \end{aligned} \tag{6.20}$$

The marginal likelihood of interest can be written as follows:

$$\begin{aligned} &\pi\left(\tilde{Y}_{1:T}; \mathcal{M}_A\right) \\ &\overset{1.}{=} \int\int \pi(\tilde{Y}_{1:T}, W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A) dW_{1:T} df_Y^A(\cdot) \\ &\overset{2.}{=} \int\int \pi(\tilde{Y}_{1:T} \mid W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A) \pi(W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A) dW_{1:T} df_Y^A(\cdot) \\ &\overset{3.}{=} \int\int \pi(\tilde{Y}_{1:T} \mid W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A) \pi(W_{1:T}; \mathcal{M}_A) \pi(f_Y^A(\cdot); \mathcal{M}_A) dW_{1:T} df_Y^A(\cdot) \\ &\overset{4.}{=} \mathbb{E}_{f_Y^A(\cdot); \mathcal{M}_A}\left[\mathbb{E}_{W_{1:T}; \mathcal{M}_A}\left[\pi\left(\tilde{Y}_{1:T} \mid W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right)\right]\right] \\ &\overset{5.}{=} \mathbb{E}_{f_Y^A(\cdot); \mathcal{M}_A}\left[\mathbb{E}_{W_{1:T}; \mathcal{M}_A}\left[\prod_{t=1}^{T} \pi\left(\tilde{Y}_t \mid \tilde{Y}_{1:s-1} W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right)\right]\right] \\ &\overset{6.}{=} \mathbb{E}_{f_Y^A(\cdot); \mathcal{M}_A}\left[\mathbb{E}_{W_{1:T}; \mathcal{M}_A}\left[\prod_{t=1}^{T} \pi\left(\tilde{Y}_t \mid \tilde{Y}_{t-1}, W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right)\right]\right] \\ &\overset{7.}{=} \mathbb{E}_{f_Y^A(\cdot); \mathcal{M}_A}\left[\mathbb{E}_{W_{1:T}; \mathcal{M}_A}\left[\prod_{t=1}^{T} \pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right)\right]\right]. \end{aligned} \tag{6.21, 6.22}$$

Figure 6.2 shows the graphical model representation of the Model A. This is followed by Graph 6.3, which shows visually the variables that we condition on (in green).

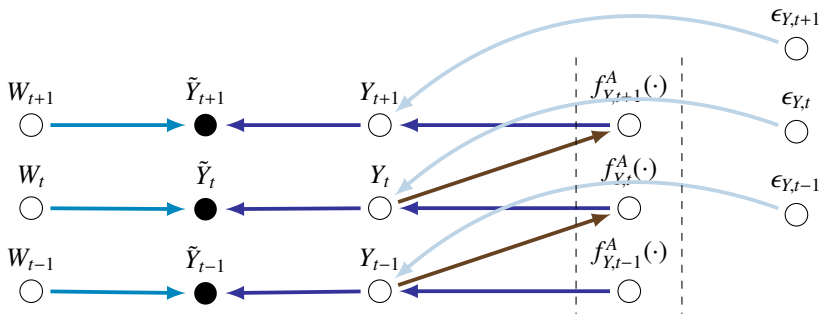


Figure 6.2: Direct Acyclic Graph (DAG) representation of the Model A random variables.

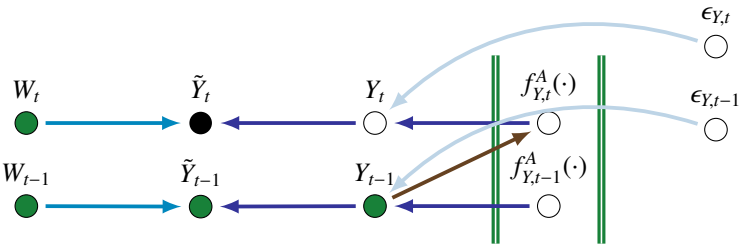


Figure 6.3: Direct Acyclic Graph (DAG) representation of the time series Model A variables, visualising the conditional probability of $\pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f^A_Y(\cdot); \mathcal{M}_A\right)$, with the conditioning variables marked in green.

The fact that the time series $\tilde{Y}_t$ is defined as a mean-variance mixture (Equation 7.2), results in the following distributions:

$$\tilde{Y}_t \mid W_t \sim \mathcal{N}\left(\mathbf{m}^A_t + \boldsymbol{\gamma}^A_t W_t, W_t\left(k^A_{Y,t,t} + \sigma^2_{Y,A}\right)\right), \tag{6.23}$$

$$\tilde{Y}_t \mid W_t, Y_{t-1}, f^A_Y(\cdot) \sim \mathcal{N}\left(\mathbf{m}^A_t + \boldsymbol{\gamma}^A_t W_t + \sqrt{W_t} f^A_Y(Y_{t-1}), W_t \sigma^2_{Y,A}\right). \tag{6.24}$$

Now, what is $f^A_Y(Y_{t-1})$? We are not able to simulate realisation of this GP, but we can draw a finite values of such a realisation, and approximate in between. To do this, we first draw points from a multivariate normal distribution, with zero mean and $k^A_{Y,t,t'}$ covariance function, and then calculate the approximation for the latent variables $\{Y_t\}$ using the conditional distribution from the equation A.14, from the Appendix.

Take a grid of N points $\mathbf{y} = [\mathsf{y}_1, ..., \mathsf{y}_N]$, for which we will obtain the values for the covariance matrix $\mathbf{K}^A$ and an N-dimensional random vector $\mathbf{g}^A_Y$ that represents the noisy realisation of the GP.

$$\mathbf{y} = [\mathsf{y}_1, ..., \mathsf{y}_N]$$
$$\mathbf{K}^A : \ \mathsf{k}^A_{l,m} = k^A_Y(\mathsf{y}_l, \mathsf{y}_m), \ l, m \in [1, T]$$
$$\mathbf{g}^A_Y = \left[\mathsf{g}^A_{Y,1}, ..., \mathsf{g}^A_{Y,N}\right] \sim \mathcal{N}(0, \mathbf{K}^A + \sigma^2_{A,Y}\mathbf{I})$$

For a point Y_{t-1}, the conditional distribution for $f^A_Y(Y_{t-1})$ will be as follows:

$$f^A_Y(Y_{t-1}) \mid \mathbf{y}, \mathbf{g}^A_Y, Y_{t-1} \sim \left(f^A_Y(Y_{t-1}); \bar{f}^A_Y(Y_{t-1}), cov(f^A_Y(Y_{t-1}))\right), \tag{6.25}$$

$$\bar{f}^A_Y(Y_{t-1}) = K^A(Y_{t-1}, \mathbf{y})\left[\mathbf{K}^A + \sigma^2_{Y,A}\mathbf{I}\right]^{-1}\mathbf{g}^A_Y \tag{6.26}$$

$$cov(f^A_Y(Y_{t-1})) = K^A(Y_{t-1}, Y_{t-1}) - K^A(Y_{t-1}, \mathbf{y})\left[\mathbf{K}^A + \sigma^2_{A,Y}\mathbf{I}\right]^{-1} K^A(\mathbf{y}, Y_{t-1}). \tag{6.27}$$

To account for the fact, that we do not simulate whole $f^A_Y(\cdot)$, but only N points, and therefore need to

include the approximation detailed in the Equations 6.25 - 6.27:

$$\begin{aligned}&\tilde{Y}_t \mid W_t, Y_{t-1}, f_Y^A(\cdot) \sim \\ &\mathcal{N}\Big(\tilde{Y}_t; \mathbf{m}_t^A + \boldsymbol{\gamma}_t^A W_t + \sqrt{W_t} K^A(Y_{t-1}, \mathbf{y}) \left[\mathbf{K}^A + \sigma_{Y,A}^2 \mathbf{I}\right]^{-1} \mathbf{g}_Y^A, \\ &\qquad W_t(\sigma_{Y,A}^2 + K^A(Y_{t-1}, Y_{t-1}) - K^A(Y_{t-1}, \mathbf{y}) \left[\mathbf{K}^A + \sigma_{Y,A}^2 \mathbf{I}\right]^{-1} K^A(\mathbf{y}, Y_{t-1})\Big).\end{aligned} \tag{6.28}$$

6.4.1.1 Approximation algorithm

This algorithm shows how to calculate the likelihood $\pi\left(\tilde{Y}_{1:T}; \mathcal{M}_A\right)$ from Eqution 6.21. We will introduce the following notation:

$$\hat{\pi}_{Y,A,t} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(\cdot)\right) \approx \pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right),$$

which approximates the distribution from the Equation 7.4,

$$\begin{aligned}&\hat{\pi}_{Y,A,T} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_{1:T} \mid W_{1:T}, f_Y^A(\cdot)\right) \approx \pi\left(\tilde{Y}_{1:T} \mid W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right) = \prod_{t=1}^{T} \pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right) \\ &\mathsf{E}_{W,A} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_{1:T} \mid f_Y^A(\cdot)\right) \approx \pi\left(\tilde{Y}_{1:T} \mid f_Y^A(\cdot); \mathcal{M}_A\right) \\ &\mathsf{E}_{f,A} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_{1:T}\right) \approx \pi\left(\tilde{Y}_{1:T}; \mathcal{M}_A\right),\end{aligned}$$

where the main part of the approximation will consist of approximating the expected values:

$$\begin{aligned}&\mathsf{E}_{W,A} \approx \mathbb{E}_{W_{1:T}; \mathcal{M}_A}\left[\prod_{t=1}^{T} \hat{\pi}_{Y,A,t}\right] \approx \mathbb{E}_{W_{1:T}; \mathcal{M}_A}\left[\prod_{t=1}^{T} \pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right)\right] \\ &\mathsf{E}_{f,A} \approx \mathbb{E}_{f_Y^A(\cdot); \mathcal{M}_A}\left[\mathsf{E}_{W,A}\right] \approx \mathbb{E}_{f_Y^A(\cdot); \mathcal{M}_A}\left[\mathbb{E}_{W_{1:T}; \mathcal{M}_A}\left[\prod_{t=1}^{T} \pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(\cdot); \mathcal{M}_A\right)\right]\right]\end{aligned}$$

6.4.2 Model B

We define two univariate Gaussian Processes:

$$\begin{aligned}&\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f_X^B([X_{t-1}, Y_{t-1}]) \\ f_Y^B([X_{t-1}, Y_{t-1}]) \end{bmatrix} + \begin{bmatrix} \epsilon_{X,t}^B \\ \epsilon_{Y,t}^B \end{bmatrix}, \qquad \begin{aligned} f_X^B &\sim \mathcal{GP}\left(0, k_{X,t_1,t_2}^B\right), \epsilon_{X,t}^B \sim \mathcal{N}(0, \sigma_{X,B}^2), \\ f_Y^B &\sim \mathcal{GP}\left(0, k_{Y,t_1,t_2}^B\right), \epsilon_{Y,t}^B \sim \mathcal{N}(0, \sigma_{Y,B}^2). \end{aligned}\end{aligned} \tag{6.29}$$

Then the transformation from $[X_t, Y_t]^T$ to $\left[\tilde{X}_t, \tilde{Y}_t\right]^T$ is :

$$\begin{aligned}&\left[\tilde{X}_t, \tilde{Y}_t\right]^T \overset{d}{=} \mathbf{m}_t^B + \boldsymbol{\gamma}_t^B W_t + \sqrt{W_t}\,[X_t, Y_t]^T, \qquad W_t \perp\!\!\!\perp [X_t, Y_t]^T \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad W_t \sim IG\left(\frac{\nu^B}{2}, \frac{\nu^B}{2}\right),\end{aligned} \tag{6.30}$$

Figure 7.2 shows the graphical model representation of the Model B.

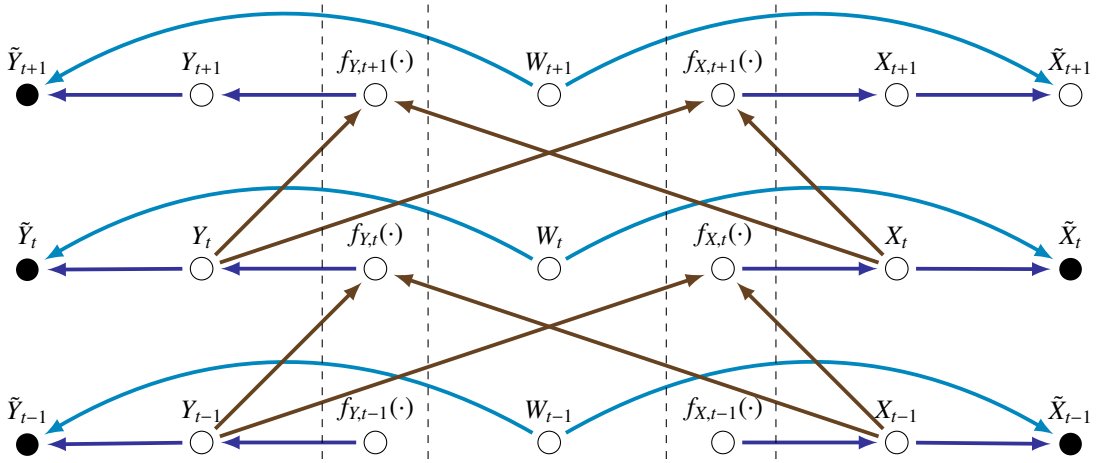


Figure 6.4: Direct Acyclic Graph (DAG) representation of the time series X, Y, Z from the point of view of generating a realization of Y_t.

When calculating the joint density of interest, we have a more complicated form, with additional elements:

$$\pi\left(\tilde{Y}_{2:T} \mid \tilde{X}_{1:T-1}; \mathcal{M}_B\right) = \prod_{t=2}^{T} \pi\left(\tilde{Y}_t \mid \tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1}\right) = \tag{6.31}$$

$$\prod_{t=2}^{T} \int\int\int \pi\left(\tilde{Y}_t, W_{1:t}, f_Y^B(\cdot), f_X^B(\cdot) \mid \tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1}\right) dW_{1:t} df_Y^B(\cdot) df_X^B(\cdot) =$$

$$\prod_{t=2}^{T} \int\int\int \pi\left(\tilde{Y}_t \mid W_{1:t}, f_Y^B(\cdot), f_X^B(\cdot), \tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1}\right) \pi\left(W_{1:t}, f_Y^B(\cdot), f_X^B(\cdot) \mid \tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1}\right) dW_{1:t} df_Y^B(\cdot) df_X^B(\cdot) =$$

$$\prod_{t=2}^{T} \int\int\int \left[\pi\left(\tilde{Y}_t \mid Y_{t-1}, X_{t-1}, W_{1:t}, f_Y^B(\cdot)\right) \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\right) \frac{\pi(f_X^B(\cdot), f_Y^B(\cdot))\pi(W_{1:t-1})}{\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1})}\right]$$

$$dW_{1:t} df_Y^B(\cdot) df_X^B(\cdot) =$$

$$\prod_{t=2}^{T} \mathbb{E}_{f_Y^B(\cdot);\mathcal{M}_B} \mathbb{E}_{f_X^B(\cdot);\mathcal{M}_B} \mathbb{E}_{W_{1:t-1};\mathcal{M}_B} \left[\pi\left(\tilde{Y}_t \mid Y_{t-1}, X_{t-1}, W_{1:t}, f_Y^B(\cdot)\right) \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\right) \right.$$

$$\left. \frac{1}{\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1})}\right] \tag{6.32}$$

where the integrand can be broken down into three components:

First component

$$\pi\left(\tilde{Y}_t \mid Y_{t-1}, X_{t-1}, W_{1:t}, f_Y^B(\cdot)\right) \stackrel{d}{=} \pi\left(\tilde{Y}_t \mid W_{t-1}, f_Y^B(\cdot), Y_{t-1}, X_{t-1}\right) \tag{6.33}$$

$$\sim \mathcal{N}\left(\tilde{Y}_t; \mathbf{m}_t^B + \boldsymbol{\gamma}_t^B W_t + \sqrt{W_t} f_Y^B([X_{t-1}, Y_{t-1}]^T), W_t \sigma_{Y,B}^2\right) \tag{6.34}$$

As in the case of the Model A, when simulating the realisation of the GP, we will be only drawing a finite

number of points for f_Y^B, and based on those values, we will be using conditional distribution to calculate the $f_Y^B([X_{t-1}, Y_{t-1}]^T)$

Take two grids of N points each $\mathbf{x} = [\mathsf{x}_1, ..., \mathsf{x}_N]$ and $\mathbf{y} = [\mathsf{y}_1, ..., \mathsf{y}_N]$, for which we will obtain the values for the covariance matrix $\mathbf{K}^A$ and an N-dimensional random vector $\mathbf{g}_Y^B$ that represents the noisy realisation of the GP.

$$\begin{aligned}
&\mathbf{x} = [\mathsf{x}_1, ..., \mathsf{x}_N] \;\; \mathbf{y} = [\mathsf{y}_1, ..., \mathsf{y}_N] \\
&\mathbf{K}^B: \;\; \mathsf{k}_{l,m}^B = k_Y^B([\mathsf{x}_l, \mathsf{y}_l]^T, [\mathsf{x}_m, \mathsf{y}_m]^T), \; l, m \in [1, T] \\
&\mathbf{g}_Y^B = \left[\mathsf{g}_{Y,1}^B, ..., \mathsf{g}_{Y,N}^B\right] \sim \mathcal{N}(0, \mathbf{K}^B + \sigma_{B,Y}^2 \mathbf{I})
\end{aligned}$$

For a point $[X_{t-1}, Y_{t-1}]^T$, the conditional distribution for $f_Y^B([X_{t-1}, Y_{t-1}]^T)$ will be as follows:

$$\begin{aligned}
&f_Y^B([X_{t-1}, Y_{t-1}]^T) \mid \mathbf{x}, \mathbf{y}, \mathbf{g}_Y^B, X_{t-1}, Y_{t-1} && (6.35)\\
&\qquad \sim \left(f_Y^B([X_{t-1}, Y_{t-1}]^T); \bar{f}_Y^B([X_{t-1}, Y_{t-1}]^T), cov(f_Y^B([X_{t-1}, Y_{t-1}]^T))\right), \\
&\bar{f}_Y^B([X_{t-1}, Y_{t-1}]^T) = K^B([X_{t-1}, Y_{t-1}]^T, [\mathbf{x}, \mathbf{y}]^T)\left[\mathbf{K}^B + \sigma_{Y,B}^2 \mathbf{I}\right]^{-1} \mathbf{g}_Y^B && (6.36)\\
&cov(f_Y^B([X_{t-1}, Y_{t-1}]^T) = K^B([X_{t-1}, Y_{t-1}]^T, [X_{t-1}, Y_{t-1}]^T) && (6.37)\\
&\qquad - K^B([X_{t-1}, Y_{t-1}]^T, [\mathbf{x}, \mathbf{y}]^T)\left[\mathbf{K}^B + \sigma_{B,Y}^2 \mathbf{I}\right]^{-1} K^B([\mathbf{x}, \mathbf{y}]^T, [X_{t-1}, Y_{t-1}]^T).
\end{aligned}$$

To account for the fact, that we do not simulate whole $f_Y^B(\cdot)$, but only N points, and therefore in the distribution 6.33 we need to include the approximation detailed in the Equations 6.35 - 6.37:

$$\begin{aligned}
&\tilde{Y}_t \mid W_t, X_{t-1}, Y_{t-1}, f_Y^B(\cdot) \sim && (6.38)\\
&\mathcal{N}\Big(\tilde{Y}_t; \mathbf{m}_{Y,t}^B + \boldsymbol{\gamma}_{Y,t}^B W_t + \sqrt{W_t} K^B([X_{t-1}, Y_{t-1}]^T, [\mathbf{x}, \mathbf{y}]^T)\left[\mathbf{K}^B + \sigma_{Y,B}^2 \mathbf{I}\right]^{-1} \mathbf{g}_Y^B, && (6.39)\\
&\qquad W_t\Big(\sigma_{Y,B}^2 + K^B\left([X_{t-1}, Y_{t-1}]^T, [X_{t-1}, Y_{t-1}]^T\right) && (6.40)\\
&\qquad\qquad - K^B\left([X_{t-1}, Y_{t-1}]^T, [\mathbf{x}, \mathbf{y}]^T\right)\left[\mathbf{K}^B + \sigma_{B,Y}^2 \mathbf{I}\right]^{-1} K^B\left([\mathbf{x}, \mathbf{y}]^T, [X_{t-1}, Y_{t-1}]^T\right)\Big)\Big). && (6.41)
\end{aligned}$$

Second component The second element will be calculated in a similar way as the first one: with the multivariate version of the distribution 6.33 including the approximation analogous to the Equations 6.35 -

6.37:

$$\pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\right) = \tag{6.42}$$

$$\pi\Big(\mathbf{m}^B_{X,1:t-1} + \left[\boldsymbol{\gamma}^B_{X,1} W_1, ..., \boldsymbol{\gamma}^B_{X,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1} X_1, ..., \sqrt{W_{t-1}} X_{t-1}\right]^T, \tag{6.43}$$

$$\mathbf{m}^B_{Y,1:t-1} + \left[\boldsymbol{\gamma}^B_{Y,1} W_1, ..., \boldsymbol{\gamma}^B_{Y,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1} Y_1, ..., \sqrt{W_{t-1}} Y_{t-1}\right]^T \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\Big) = \tag{6.44}$$

$$\pi\Big(\mathbf{m}^B_{X,1:t-1} + \left[\boldsymbol{\gamma}^B_{X,1} W_1, ..., \boldsymbol{\gamma}^B_{X,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1}\left(f_X^B\left([X_0, Y_0]^T\right) + \epsilon^B_{X,1}\right), ..., \sqrt{W_{t-1}}\left(f_X^B\left([X_{t-2}, Y_{t-2}]^T\right) + \epsilon^B_{X,t-1}\right)\right]^T, \tag{6.45}$$

$$\mathbf{m}^B_{Y,1:t-1} + \left[\boldsymbol{\gamma}^B_{Y,1} W_1, ..., \boldsymbol{\gamma}^B_{Y,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1}\left(f_Y^B\left([X_0, Y_0]^T\right) + \epsilon^B_{Y,1}\right), ..., \sqrt{W_{t-1}}\left(f_Y^B\left([X_{t-2}, Y_{t-2}]^T\right) + \epsilon^B_{Y,t-1}\right)\right]^T \tag{6.46}$$

$$\mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\Big) = \tag{6.47}$$

$$\mathcal{N}\Big(\tilde{X}_{1:t-1}; \mathbf{m}^B_{X,1:t-1} + \left[\boldsymbol{\gamma}^B_{X,1} W_1, ..., \boldsymbol{\gamma}^B_{X,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1} f_X^B\left([X_0, Y_0]^T\right), ..., \sqrt{W_{t-1}} f_X^B\left([X_{t-2}, Y_{t-2}]^T\right)\right]^T, \sigma^2_{X,B} W_{1:t-1}\mathbb{I}\Big) \tag{6.48}$$

$$\mathcal{N}\Big(\tilde{Y}_{1:t-1}; \mathbf{m}^B_{Y,1:t-1} + \left[\boldsymbol{\gamma}^B_{Y,1} W_1, ..., \boldsymbol{\gamma}^B_{Y,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1} f_Y^B\left([X_0, Y_0]^T\right), ..., \sqrt{W_{t-1}} f_Y^B\left([X_{t-2}, Y_{t-2}]^T\right)\right]^T, \sigma^2_{Y,B} W_{1:t-1}\mathbb{I}\Big) = \tag{6.49}$$

using the Equations 6.35, 6.36, 6.37:

$$\mathcal{N}\Big(\tilde{X}_{1:t-1}; \mathbf{m}^B_{X,1:t-1} + \left[\boldsymbol{\gamma}^B_{X,1} W_1, ..., \boldsymbol{\gamma}^B_{X,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1}, ..., \sqrt{W_{t-1}}\right]^T \mathbb{I} K_X^B\left([X_{0:t-2}, Y_{0:t-2}]^T, [\mathbf{x}, \mathbf{y}]^T\right)\left[\mathbf{K}_X^B + \sigma^2_{X,B}\mathbf{I}\right]^{-1} \mathbf{g}_X^B, \tag{6.50}$$

$$\sqrt{W_{1:t-1}}\mathbb{I}\Big[\sigma^2_{X,B}\mathbb{I} + K_X^B([X_{0:t-2}, Y_{0:t-2}]^T, [X_{0:t-2}, Y_{0:t-2}]^T) \tag{6.51}$$

$$- K_X^B([X_{0:t-2}, Y_{0:t-2}]^T, [\mathbf{x}, \mathbf{y}]^T)\left[\mathbf{K}_X^B + \sigma^2_{B,X}\mathbf{I}\right]^{-1} K_X^B([\mathbf{x}, \mathbf{y}]^T, [X_{0:t-2}, Y_{0:t-2}]^T)\Big)\mathbb{I}\sqrt{W_{1:t-1}}^T\Big] \tag{6.52}$$

$$\mathcal{N}\Big(\tilde{Y}_{1:t-1}; \mathbf{m}^B_{Y,1:t-1} + \left[\boldsymbol{\gamma}^B_{Y,1} W_1, ..., \boldsymbol{\gamma}^B_{Y,t-1} W_{t-1}\right]^T + \left[\sqrt{W_1}, ..., \sqrt{W_{t-1}}\right]^T \mathbb{I} K_Y^B\left([X_{0:t-2}, Y_{0:t-2}]^T, [\mathbf{x}, \mathbf{y}]^T\right)\left[\mathbf{K}_Y^B + \sigma^2_{Y,B}\mathbf{I}\right]^{-1} \mathbf{g}_Y^B, \tag{6.53}$$

$$\sqrt{W_{1:t-1}}\mathbb{I}\Big[\sigma^2_{Y,B}\mathbb{I} + K_Y^B([X_{0:t-2}, Y_{0:t-2}]^T, [X_{0:t-2}, Y_{0:t-2}]^T) \tag{6.54}$$

$$- K_Y^B([X_{0:t-2}, Y_{0:t-2}]^T, [\mathbf{x}, \mathbf{y}]^T)\left[\mathbf{K}_Y^B + \sigma^2_{B,Y}\mathbf{I}\right]^{-1} K_Y^B([\mathbf{x}, \mathbf{y}]^T, [X_{0:t-2}, Y_{0:t-2}]^T)\Big)\mathbb{I}\sqrt{W_{1:t-1}}^T\Big] \tag{6.55}$$

Third component And finally, $\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1})$ is a normalising constant whom we don't know, and therefore we need to approach it as a self normalising constant.

$$\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1}) \overset{d}{=} \mathbb{E}_{f_Y^B(\cdot);\mathcal{M}_B}\mathbb{E}_{f_X^B(\cdot);\mathcal{M}_B}\mathbb{E}_{W_{1:t-1};\mathcal{M}_B}\left[\pi\left(\tilde{X}_{1:t-1}\tilde{Y}_{1:t-1} \mid W_{1:t-1}, f_X^B(\cdot), f_Y^B(\cdot)\right)\right] \tag{6.56}$$

This can be estimated by summing the distribution from 6.50. We can't calculate such a sum separately, but to do this, we will use the approach of importance sampling with self normalising weights and

calculate the normalising constant 6.56 together when calculating the whole density from Equation 6.32.

Imagine, that one needs to calculate an expectation $\mu = \mathbb{E}_\pi(h(x)) = \int_D h(x)\pi(x)dx$, but cannot draw from the distribution p. Assume it is possible to draw from a distribution q (which is nonzero, when $h(x)\pi(x)$ is non-zero), such that:

$$\mu = \int_D h(x)\pi(x)dx = \int_D \frac{h(x)\pi(x)}{q(x)}q(x)dx = \mathbb{E}_q\left[\frac{h(x)\pi(x)}{q(x)}dx\right]$$

Lets define $w(x)$ as a ratio of the two densities: $w(x) = \frac{\pi(x)}{q(x)}$, then the estimator of μ can be written as follows:

$$\hat{\mu} = \frac{1}{n}\sum_i^n \frac{h(x_i)\pi(x_i)}{q(x_i)} = \frac{1}{n}\sum_i^n h(x_i)w(x_i).$$

This estimation can be calculated, if we're able to evaluate both $h(x)$ and $w(x)$ using importance sampling [Kloek and Van Dijk, 1978], [Tokdar and Kass, 2010]. It can also be used in the case that will interest us the most – in the case where we know $w(x)$ up to a normalising constant. We will denote $\pi_u(x)$ as the unnormalised version of the density $\pi(x)$, $q_u(x)$ as the unnormalised version of the density $q(x)$, which means that there are constants c, d such that $\pi_u(x) = c\pi(x), q_u(x) = dq(x)$, and we will introduce an unnormalised ratio $u(x) = \frac{\pi_u(x)}{q_u(x)} = \frac{c\pi(x)}{dq(x)}$. The constants c, d being normalising constatnts mean that $c = \sum_i^n \pi_u(x_i)$, $d = \sum_i^n q_u(x_i)$, and $w(x) = u(x)\frac{d}{c} = \frac{u(x)}{\sum_i^n u(x_i)}$. And so:

$$\hat{\mu} = \frac{1}{n}\sum_i^n \frac{h(x_i)\pi(x_i)}{q(x_i)} = \frac{1}{n}\sum_i^n h(x_i)w(x_i) = \frac{1}{n}\sum_i^n h(x_i)u(x_i)\frac{d}{c} = \frac{1}{n}\frac{\sum_i^n h(x_i)u(x_i)}{\sum_i^n u(x_i)}$$

When translating the notation to our problem, I will use the notation as follows:

$$h_B^{(t,i,k,j)} = \pi\left(\tilde{Y}_t \mid Y_{t-1}, X_{t-1}, W_{1:t}, f_Y^B(\cdot)\right) \tag{6.57}$$

$$w_{B,f_Y,f_X,W}^{t,i,k,j} = \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\right)\frac{1}{\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1})} \tag{6.58}$$

$$u_{B,f_Y,f_X,W}^{t,i,k,j} = \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\right) \tag{6.59}$$

the expected value that we want to calculate is as follows:

$$\mathsf{E}_B = \prod_{t=2}^{T} \mathbb{E}_{f_Y^B(\cdot);\mathcal{M}_B}\left[\mathbb{E}_{f_X^B(\cdot);\mathcal{M}_B}\left[\mathbb{E}_{W_{1:t};\mathcal{M}_B}\left[\pi\left(\tilde{Y}_t \mid Y_{t-1}, X_{t-1}, W_{1:t}, f_Y^B(\cdot)\right)\frac{\pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\right)}{\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1})}\right]\right]\right] \tag{6.60}$$

and with the estimator:

$$\hat{\mathsf{E}}_B \approx \prod_{t=2}^{T} \frac{1}{N_i} \sum_{i=1}^{N_i} \left[\frac{1}{N_k} \sum_{k=1}^{N_k} \left[\frac{1}{N_j} \sum_{j=1}^{N_j} \left[h_B^{(t,i,k,j)} w_{B,f_Y,f_X,W}^{(t,i,k,j)} \right] \right] \right] \tag{6.61}$$

$$= \prod_{t=2}^{T} \frac{1}{N_i} \sum_{i=1}^{N_i} \left[\frac{1}{N_k} \sum_{k=1}^{N_k} \left[\frac{1}{N_j} \sum_{j=1}^{N_j} \left[h_B^{(t,i,k,j)} \frac{u_{B,f_Y,f_X,W}^{(t,i,k,j)}}{\sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} u_{B,f_Y,f_X,W}^{(t,i,k,j)}} \right] \right] \right] \tag{6.62}$$

$$= \prod_{t=2}^{T} \frac{1}{N_i N_k N_j} \frac{\sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} h_B^{(t,i,k,j)} u_{B,f_Y,f_X,W}^{(t,i,k,j)}}{\sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} u_{B,f_Y,f_X,W}^{(t,i,k,j)}} \tag{6.63}$$

Note: $f_Y, f_X, W_{1:t}$ are independent, so the sums can be seen as three dimensional.

6.4.2.1 Approximation algorithm B

This algorithm shows how to calculate the likelihood $\mathsf{E}_B \approx \pi\left(\tilde{Y}_{2:T} \mid \tilde{X}_{1:T-1}; \mathcal{M}_B\right)$ from Eqution 6.31, which is reminded below:

$$\begin{aligned}
&\pi\left(\tilde{Y}_{2:T} \mid \tilde{X}_{1:T-1}; \mathcal{M}_B\right) \\
&= \prod_{t=2}^{T} \mathbb{E}_{f_Y^B(\cdot);\mathcal{M}_B} \mathbb{E}_{f_X^B(\cdot);\mathcal{M}_B} \mathbb{E}_{W_{1:t-1};\mathcal{M}_B} \Big[\pi\left(\tilde{Y}_t \mid Y_{t-1}, X_{t-1}, W_{1:t}, f_Y^B(\cdot)\right) \\
&\qquad \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(\cdot), f_X^B(\cdot)\right) \frac{1}{\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1})} \Big]
\end{aligned}$$

Just like for the model A, we will simulate the realisation of Gaussian process (in finite number of points), and the mixing variable. These will be encapsulated, and that encapsulation is represented by the superscripts referring to each realisation: i-th realisation of f_Y, k-th relisation of f_X, j-th realisation of $W_{1:t}$. We will introduce the following notation, to approximate components from the above equation (Equation 6.32), continuing with the notation from the Equation 6.63:

$$\begin{aligned}
h_B^{(t,i,k,j)} &= \pi\left(\tilde{Y}_t \mid Y_{t-1}^{(t,i,k,j)}, X_{t-1}^{(t,i,k,j)}, W_{1:t}^{(t,i,k,j)}, f_Y^{B,(t,i)}(\cdot)\right) \\
u_{B,f_Y,f_X,W}^{(t,i,k,j)} &= \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, f_Y^{B,(t,i)}(\cdot), f_X^{B,(t,i,k)}(\cdot)\right) \\
u_{B,f_Y,f_X}^{(t,i,k)} &= \sum_{j=1}^{N_j} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, f_Y^{B,(t,i)}(\cdot), f_X^{B,(t,i,k)}(\cdot)\right) \\
u_{B,f_Y}^{(t,i)} &= \sum_{k=1}^{N_k} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid f_Y^{B,(t,i)}(\cdot), f_X^{B,(t,i,k)}(\cdot)\right) \\
u_B^{(t)} &= \sum_{i=1}^{N_i} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid f_Y^{B,(t,i)}(\cdot)\right)
\end{aligned}$$

The next step of the notation is introduced to remind that we're not simulating the whole Gaussian processes, but only their value on a grid (and the points between the grid will be calculated as conditional

distributions)

$$
\begin{aligned}
\check{h}_B^{(t,i,k,j)} &= \pi\left(\tilde{Y}_t \mid Y_{t-1}^{(t,i,k,j)}, X_{t-1}^{(t,i,k,j)}, W_{1:t}^{(t,i,k,j)}, \check{f}_Y^{B,(t,i)}(\cdot)\right) \\
\check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)} &= \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, \check{f}_Y^{B,(t,i)}(\cdot), \check{f}_X^{B,(t,i,k)}(\cdot)\right) \\
\check{u}_{B,f_Y,f_X}^{(t,i,k)} &= \sum_{j=1}^{N_j} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, \check{f}_Y^{B,(t,i)}(\cdot), \check{f}_X^{B,(t,i,k)}(\cdot)\right) \\
\check{u}_{B,f_Y}^{(t,i)} &= \sum_{k=1}^{N_k} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid \check{f}_Y^{B,(t,i)}(\cdot), \check{f}_X^{B,(t,i,k)}(\cdot)\right) \\
\check{u}_B^{(t)} &= \sum_{i=1}^{N_i} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid \check{f}_Y^{B,(t,i)}(\cdot)\right)
\end{aligned}
$$

All of the summations will be calculated in loops, hence notation for the partial sums (estimation of the unnormalised expectations):

$$
\begin{aligned}
\mathsf{F}_{B,W}^{(t,i,k)} &= \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)} \\
\mathsf{F}_{B,f_X}^{(t,i)} &= \sum_{k=1}^{N_k} \mathsf{F}_{B,W}^{(t,i,k)} = \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)} \\
\mathsf{F}_{B,f_Y}^{(t)} &= \sum_{i=1}^{N_i} \mathsf{F}_{B,f_X}^{(t,i)} = \sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)}
\end{aligned}
$$

and finally:

$$
\hat{\mathsf{E}}_B = \prod_{t=2}^{T} \frac{1}{N_i N_k N_j} \frac{1}{\check{u}_B^{(t)}} \mathsf{F}_{B,f_Y}^{(t)} = \prod_{t=2}^{T} \frac{1}{N_i N_k N_j} \frac{1}{\check{u}_B^{(t)}} \sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)}
$$

Note! In the case that we're considering here, the two Gaussian processes are independent, and both are independent from the mixing variables. This means that we are allowed to treat the integrals as one multidimensional integrals, and the sum, as one sum, as follows:

$$
\hat{\mathsf{E}}_B = \prod_{t=2}^{T} \frac{1}{N_i} \frac{1}{\check{u}_B^{(t)}} \mathsf{F}_B^{(t)} = \prod_{t=2}^{T} \frac{1}{N_i} \frac{1}{\check{u}_B^{(t)}} \sum_{i=1}^{N_i} \check{h}_B^{(t,i)} \check{u}_B^{(t,i)}
$$

6.4.3 Introducing autoregression.

Autoregression can already be introduced via covariance structure. This, however, will not suffice for how we want to be able to model autoregression and causality.

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \overset{d}{=} \mathbf{m}_t + \boldsymbol{\gamma}_t W_t + \sqrt{W_t}\,[X_t, Y_t]^T, \text{ where:} \qquad (6.64)$$
$$\mathbf{m}_t = \mathbf{m}_t\left(\left[\tilde{X}_{t-1}, \tilde{Y}_{t-1}\right]^T\right)$$

We'd like to point out, that conditioning on the past values of the time series, the mean function $\mathbf{m}_t\left(\left[\tilde{X}_{t-1}, \tilde{Y}_{t-1}\right]^T\right)$ will be deterministic. Also because $\left\{\tilde{X}_t\right\}_{t=1}^T, \left\{\tilde{Y}_t\right\}_{t=1}^T$ are observed time series, then the conditioning does not introduce a state space equation.

The algorithms for the models A and B will be the same, up to a different formulation for the mean function, which will no longer be a constant. When performing optimisation, we are going to have more parameters to optimise though.

6.4.4 Interplay between different dependence structures

Referring back to the way we have introduced the multi-output GPs using convolution (see Section 3.1.3.2 and Figure 7.3), we can present the relationships between different components of warped multi-output GP as in Figure 6.5. Like before, we assume that $\tilde{X}_t, \tilde{Y}_t$ are observable variables, while all other structures are latent.

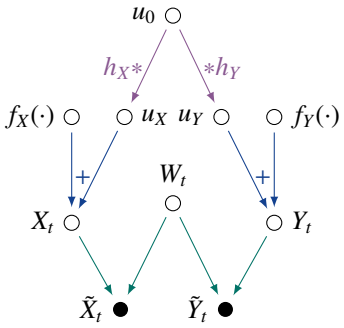


Figure 6.5: How to obtain dependent Gaussian Processes X, Y from independent f_X, f_Y and a common white noise U_0 smoothed by smoothing kernels (linear filters) h_X, h_Y.

Firstly, at the level of marginal distributions of X_t, Y_t, described in the Equations 6.65, we can study the number of lags, and the effect of structural properties encoded via the covariance functions. We can also introduce side information, or multimodality at this level. The autoregressive GPs f_X, f_Y are generally not time-reversible (see Section 3.1.2.1), which agrees with the conceptual notion of statistical causality that can be encoded via these type of structures.

$$X_t = f_X\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) + u_{X,t} \qquad (6.65)$$
$$Y_t = f_Y\left(\left[X_{t-1}^{-k}, Y_{t-1}^{-l}, Z_{t-1}^{-m}\right]\right) + u_{Y,t}$$

Convolved GP structure allows to introduce contemporaneous dependence, with $u_{X,t}, u_{Y,t}$ expressed

as convolution integral between a smoothing kernels h_X, h_Y and u_0:

$$u_{\bullet,t} = h_\bullet(t) * u_0(t) = \int_{-\infty}^{\infty} h_\bullet(t-\tau)u_0(\tau)d\tau. \tag{6.66}$$

The warping with mean-variance transformation is the source of encoding temporal dependence, in particular tail dependence, and a wide class of transformations can be used for this purpose:

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \stackrel{d}{=} \mathbf{m}_t + \boldsymbol{\gamma}_t W_t + \sqrt{W_t}\,[X_t, Y_t]^T, \tag{6.67}$$

The use of skew-t copula has the advantage of marginalisation property, meaning the marginal distribution of $\tilde{Y}$ is described by the same law as the joint distribution of $\left[\tilde{X}_t, \tilde{Y}_t\right]$. But it is not the only choice that can be made here. In fact, we also briefly look at what the use of alternative skew-t distributions could mean for the ability to model tail dependence in the data.

6.4.4.1 Causality versus Tail Dependence

The warped GP setting has been designed in a such a way, that the marginal distributions can exhibit skewness and leptokurtic tails, while the joint distribution can exhibit tail dependence.

Theorem 16 ***Tail dependence for generalised skew-t distribution.*** *Let X_1, X_2 be a bivariate skew-t distributed variable, as defined in the Equations 3.88, with γ_1, γ_2 being the skewness parameters and ρ the correlation of the normal variables in the mixture. Then the upper and lower tail coefficients of X_1, X_2 are given by:*

1. *If $\gamma_1 = \gamma_2 = 0$ (i.e. bivariate symmetric t), then:*

$$\lambda_L = 2F_{t_1}\left(\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right).$$

2. *If $\gamma_1 > 0, \gamma_2 > 0$, then $\lambda_L = 0$*

3. *If $\gamma_1 < 0, \gamma_2 < 0$, then $\lambda_L = 1$*

4. *If $\gamma_1 < 0, \gamma_2 > 0$, then $\lambda_L = 0$*

5. *If $\gamma_1 = 0, \gamma_2 > 0$, then $\lambda_L = 0$*

6. *If $\gamma_1 = 0, \gamma_2 < 0$, then*

$$\lambda_L = \int_0^1 \left(1 - \Phi\left(\left(\frac{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu+1}{2}\right)}{2\sqrt{\pi}}\right)^{1/\nu} u^{1/\nu_1}\right)\right) du.$$

.

Interesting points: the tail is increasing function of ρ and ν in point 1, not sure about point 6.

Theorem 17 ***Tail dependence for an elliptical skew-t distribution [Fung and Seneta, 2010a].*** *Let* $[X_1, X_2]^T \sim AS\,t_2(\mu, \Sigma, \gamma, \nu)$ *be a bivariate elliptical skew-t distributed random vector. The asymptotic lower tail dependence coefficient for* $[X_1, X_2]^T$ *is given by:*

$$\lambda_L = \int_{-\infty}^{c_1} p f_{t_{\nu+1}}(z) \frac{F_{t_{\nu+2}}\left(\left(\theta_2\sqrt{\frac{1-\rho^2}{\nu+1}} - (\theta_1 + \rho\theta_2)\right)\sqrt{\frac{\nu+2}{1+\frac{z^2}{\nu+1}}}\right)}{F_{t_{\nu+1}}(-\lambda_1\sqrt{\nu+1})} dz$$
$$+ \int_{-\infty}^{c_2} p_{t_{\nu+1}}(z) \frac{F_{t_{\nu+2}}\left(\left(\theta_1\sqrt{\frac{1-\rho^2}{\nu+1}} - (\theta_2 + \rho\theta_1)\right)\sqrt{\frac{\nu+2}{1+\frac{z^2}{\nu+1}}}\right)}{F_{t_{\nu+1}}(-\lambda_2\sqrt{\nu+1})} dz,$$

where

$$c_1 = \left\{\left(\frac{F_{t_{\nu+1}}(-\lambda_2\sqrt{\nu+1})}{F_{t_{\nu+1}}(-\lambda_1\sqrt{\nu+1})}\right)^{1/\nu} - \rho\right\}\sqrt{\frac{\nu+1}{1-\rho^2}}$$
$$c_2 = \left\{\left(\frac{F_{t_{\nu+1}}(-\lambda_1\sqrt{\nu+1})}{F_{t_{\nu+1}}(-\lambda_2\sqrt{\nu+1})}\right)^{1/\nu} - \rho\right\}\sqrt{\frac{\nu+1}{1-\rho^2}}.$$

where the notation $p_{t_\nu}(.), F_{t_\nu}(.)$ *is used for, respectively, the p.d.f. and c.d.f. of a univariate symmetric t distributions with* ν *degrees of freedom.*

Chapter 7

Algorithms

> *It's a rare gift, to know where you need to be, before you've been to all the places you don't need to be.*
>
> Ursula K. Le Guin, *Tales from Earthsea*

7.1 Estimating the test statistic for wGP

This algorithm is an implementation of the estimation from Section 6.4

For observed data $\left\{\tilde{X}_t\right\}_{t=1}^{T}, \left\{\tilde{Y}_t\right\}_{t=1}^{T}$, which are modelled as a wGP, we want to asses causal dependence in the direction $\tilde{X} \to \tilde{Y}$. The null hypothesis is:

$$H_0 : \qquad \pi(\tilde{Y}_t \mid \tilde{Y}_{t-1}, \tilde{X}_{t_1}; \mathcal{M}_B) = \pi(\tilde{Y}_t \mid \tilde{Y}_{t-1}; \mathcal{M}_A), \quad \forall t = 1, ..., T. \tag{7.1}$$

For the sake of estimation, we define all of the models with two simplifications: firstly not looking at the side information $\left\{\tilde{Z}_t\right\}_{t=1}^{T}$, secondly using only one lag, i.e. in the general formulations of $X_{t-1}^{-k}, Y_{t-1}^{-l}$ we will take $k = l = 1$

7.1.1 Model A

First of all, we assume that the observed data $\left\{\tilde{X}_t\right\}_{t=1}^{T}, \left\{\tilde{Y}_t\right\}_{t=1}^{T}$ can be described byt a following warped GP:

$$\left[\tilde{X}_t, \tilde{Y}_t\right]^T \stackrel{d}{=} \mathbf{m}_t^A + \boldsymbol{\gamma}_t^A W_t + \sqrt{W_t}\,[X_t, Y_t]^T, \qquad W_t \perp\!\!\!\perp [X_t, Y_t]^T \tag{7.2}$$
$$W_t \sim IG\left(\frac{\nu^A}{2}, \frac{\nu^A}{2}\right),$$

where

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f_X^A(X_{t-1}) \\ f_Y^A(Y_{t-1}) \end{bmatrix} + \begin{bmatrix} \epsilon_{X,t}^A \\ \epsilon_{Y,t}^A \end{bmatrix}, \qquad \begin{aligned} f_X^A &\sim \mathcal{GP}\left(0, k_{X,t_1,t_2}^A\right), \epsilon_{X,t}^A \sim \mathcal{N}(0, \sigma_{X,A}^2), \\ f_Y^A &\sim \mathcal{GP}\left(0, k_{Y,t_1,t_2}^A\right), \epsilon_{Y,t}^B \sim \mathcal{N}(0, \sigma_{Y,A}^2). \end{aligned} \tag{7.3}$$

This model can be represented by Diagram (7.1).

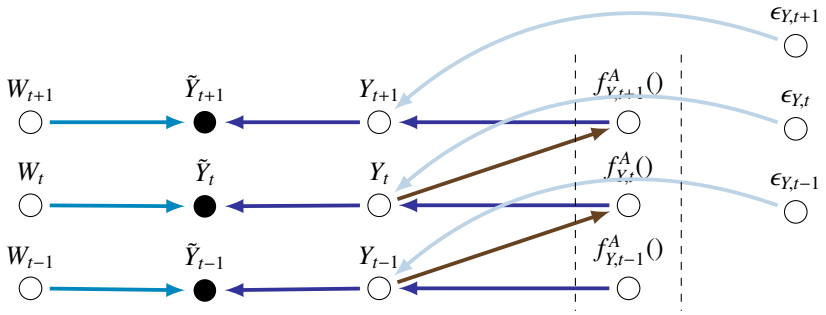


Figure 7.1: Direct Acyclic Graph (DAG) representation of the Model A random variables.

The conditional likelihood of interest is as follows:

$$\begin{aligned} &\tilde{Y}_t \mid W_t, Y_{t-1}, f_Y^A() \sim \\ &\mathcal{N}\Big(\tilde{Y}_t; \mathbf{m}_t^A + \boldsymbol{\gamma}_t^A W_t + \sqrt{W_t} K^A(Y_{t-1}, \mathbf{y}) \left[\mathbf{K}^A + \sigma_{Y,A}^2 \mathbf{I}\right]^{-1} \mathbf{g}_Y^A, \\ &\qquad W_t(\sigma_{Y,A}^2 + K^A(Y_{t-1}, Y_{t-1}) - K^A(Y_{t-1}, \mathbf{y}) \left[\mathbf{K}^A + \sigma_{Y,A}^2 \mathbf{I}\right]^{-1} K^A(\mathbf{y}, Y_{t-1})\Big). \end{aligned} \tag{7.4}$$

7.1.1.1 Approximation algorithm

This algorithm shows how to calculate the likelihood $\pi\left(\tilde{Y}_{1:T}; \mathcal{M}_A\right)$ from Eqution 6.21. We will introduce the following notation:

$$\hat{\pi}_{Y,A,t} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A()\right) \approx \pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(); \mathcal{M}_A\right),$$

which approximates the distribution from the Equation 7.4,

$$\hat{\pi}_{Y,A,T} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_{1:T} \mid W_{1:T}, f_Y^A()\right) \approx \pi\left(\tilde{Y}_{1:T} \mid W_{1:T}, f_Y^A(); \mathcal{M}_A\right) = \prod_{t=1}^{T} \pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(); \mathcal{M}_A\right)$$

$$\mathsf{E}_{W,A} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_{1:T} \mid f_Y^A()\right) \approx \pi\left(\tilde{Y}_{1:T} \mid f_Y^A(); \mathcal{M}_A\right)$$

$$\mathsf{E}_{f,A} \overset{def.}{=} \hat{\pi}\left(\tilde{Y}_{1:T}\right) \approx \pi\left(\tilde{Y}_{1:T}; \mathcal{M}_A\right),$$

where the main part of the approximation will consist of approximating the expected values:

$$\mathsf{E}_{W,A} \approx \mathbb{E}_{W_{1:T};\mathcal{M}_A}\left[\prod_{t=1}^{T}\hat{\pi}_{Y,A,t}\right] \approx \mathbb{E}_{W_{1:T};\mathcal{M}_A}\left[\prod_{t=1}^{T}\pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(); \mathcal{M}_A\right)\right]$$

$$\mathsf{E}_{f,A} \approx \mathbb{E}_{f_Y^A();\mathcal{M}_A}\left[\mathsf{E}_{W,A}\right] \approx \mathbb{E}_{f_Y^A();\mathcal{M}_A}\left[\mathbb{E}_{W_{1:T};\mathcal{M}_A}\left[\prod_{t=1}^{T}\pi\left(\tilde{Y}_t \mid Y_{t-1}, W_{1:T}, f_Y^A(); \mathcal{M}_A\right)\right]\right]$$

Initialise:

$\mathsf{E}_{f,A} = 0$.

Choose parameters:

$\mathbf{m} = [m_1, ..., m_T]$, $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_T]$, $\nu, \sigma^2_{Y,A}$, hyperparameters for $k^A_{Y,s,t}$.

Note: start with fixed m_t, γ_t.

Set a grid of N points $\mathbf{y} = [\mathsf{y}_1, ..., \mathsf{y}_N]$

For this grid we will obtain the values for the covariance matrix $\mathbf{K}$ with elements $\mathsf{k}_{l,m} = k^A_Y(\mathsf{y}_l, \mathsf{y}_m)$, $l, m \in [1, T]$.

Note: Start with the same grid for all $i = 1..N_i$. Which is why the grid before the loops!

Note: We can use a random grid! It's worth comparing to fixed grid.

Calculate the conditioning number of $\mathbf{K}^A$. If $R_{con}(\mathbf{K}^A < 10^{-3}$, perform a regularisation (for example with Nystrom approximation).

Precalculate $\left[\mathbf{K}^A + \sigma^2_{Y,A}\mathbf{I}\right]^{-1}$.

For $i = 1..N_i$, repeat (1) - (6), to obtain value $\mathsf{E}_{f,A}$

⟼⟶**Loop (i)**

1. Draw an N-dimensional noisy random vector $\mathbf{g}_Y^{A,(i)} = \left[\mathsf{g}_{Y,1}^{A,(i)}, ..., \mathsf{g}_{Y,N}^{A,(i)}\right] \sim \mathcal{N}(0, \mathbf{K}^A + \sigma^2_{Y,A}\mathbf{I})$.
2. Precalculate: $\left[\mathbf{K}^A + \sigma^2_{Y,A}\mathbf{I}\right]^{-1}\mathbf{g}_Y^{A,(i)}$.
3. Initialise: $\mathsf{E}_{W,A} = 0$.

   For $j = 1..N_j$, repeat (a) - (c), to estimate the expected value $\mathsf{E}_{W,A}$

   ⟼⟶**Inner Loop (j)**

   (a) Draw i.i.d. $W_t^{(i,j)} \sim IG(\nu/2, \nu/2)$, $t \in [1, T]$.

   (b) Initialise: $\hat{\pi}_{Y,A,T} = 1$.

For $t = 1..T$, repeat (i) - (iv), to estimate $\hat{\pi}_{Y,A,t}$

$\longmapsto\longrightarrow$**Inner Loop (t)**

i. Evaluate points $Y_t^{(i,j)}, t \in [1,T]$, as a function of the observations $\tilde{Y}_t$ and the draws of $W_t^{(i,j)}$: $Y_t^{(i,j)}(\tilde{Y}_t, W_t^{(i,j)}) = \frac{1}{\sqrt{W_t}}(\tilde{Y}_t - m_t - \gamma_t W_t^{(i,j)})$.

ii. Evaluate the conditional distribution $f_Y^A(Y_{t-1}^{(i,j)}) \mid \mathbf{y}, \mathbf{g}_Y^{A,(i)}, Y_{t-1}^{(i,j)} \sim \left(f_Y^A(Y_{t-1}^{(i,j)}); \bar{f}_Y^A(Y_{t-1}^{(i,j)}), cov(f_Y^A(Y_{t-1}^{(i,j)}))\right)$, from the Equation 6.25, with the mean and the covariance given by Equations 6.26 and 6.27:
$\bar{f}_Y^A(Y_{t-1}^{(i,j)}) = K^A(Y_{t-1}^{(i,j)}, \mathbf{y}) \left[\mathbf{K}^A + \sigma_{Y,A}^2 \mathbf{I}\right]^{-1} \mathbf{g}_Y^{A,(i)}$,
$cov(f_Y^A(Y_{t-1}^{(i,j)})) = K^A(Y_{t-1}^{(i,j)}, Y_{t-1}^{(i,j)}) - K^A(Y_{t-1}^{(i,j)}, \mathbf{y}) \left[\mathbf{K}^A + \sigma_{A,Y}^2 \mathbf{I}\right]^{-1} K^A(\mathbf{y}, Y_{t-1}^{(i,j)})$

iii. Use normal distribution from 7.4:
$\mathcal{N}\left(m_t + \gamma_t W_t^{(i,j)} + \sqrt{W_t^{(i,j)}} \bar{f}_Y^A(Y_{t-1}^{(i,j)}), W_t^{(i,j)} \left(\sigma_{Y,A}^2 + cov(f_Y^A(Y_{t-1}^{(i,j)}))\right)\right)$ to evaluate $\hat{\pi}_{Y,A,t} = \hat{\pi}\left(\tilde{Y}_{1:T} \mid W_{1:T}^{(i,j)}, f_Y^A()\right)$.

iv. Update the product: $\hat{\pi}_{Y,A,T} := \hat{\pi}_{Y,A,T} * \hat{\pi}_{Y,A,t}$

End **Inner Loop (t)**

(c) Update the sum: $\mathsf{E}_{W,A} := \mathsf{E}_{W,A} + \hat{\pi}_{Y,A,T}$

End **Inner Loop (j)**

4. Obtain the final value of the estimation: $\mathsf{E}_{W,A} := \mathsf{E}_{W,A}/N_j$

5. Update the sum: $\mathsf{E}_{f,A} := \mathsf{E}_{f,A} + \mathsf{E}_{W,A}$

End **Inner Loop (i)**

Obtain the final value of the estimation: $\mathsf{E}_{f,A} := \mathsf{E}_{f,A}/N_i$

End of the algorithm.

7.1.2 Model B

Figure 7.2 shows the graphical model representation of the Model B.

This algorithm shows how to calculate the likelihood $\mathsf{E}_B \approx \pi\left(\tilde{Y}_{2:T} \mid \tilde{X}_{1:T-1}; \mathcal{M}_B\right)$ from Eqution 6.31, which is reminded below:

$$\begin{aligned}
&\pi\left(\tilde{Y}_{2:T} \mid \tilde{X}_{1:T-1}; \mathcal{M}_B\right) \\
&= \prod_{t=2}^{T} \mathbb{E}_{f_Y^B();\mathcal{M}_B} \mathbb{E}_{f_X^B();\mathcal{M}_B} \mathbb{E}_{W_{1:t-1};\mathcal{M}_B} \Big[\pi\left(\tilde{Y}_t \mid Y_{t-1}, X_{t-1}, W_{1:t}, f_Y^B()\right) \\
&\qquad \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}, f_Y^B(), f_X^B()\right) \frac{1}{\pi(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1})}\Big]
\end{aligned}$$

Just like for the model A, we will simulate the realisation of Gaussian process (in finite number of points), and the mixing variable. These will be encapsulated, and that encapsulation is represented by

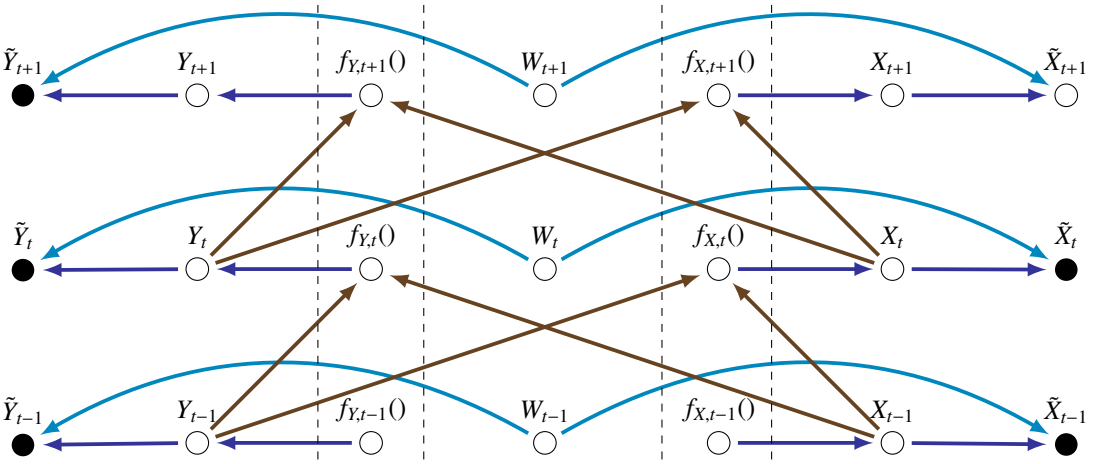


Figure 7.2: Direct Acyclic Graph (DAG) representation of the time series X, Y, Z from the point of view of generating a realization of Y_t.

the superscripts referring to each realisation: i-th realisation of f_Y, k-th relisation of f_X, j-th realisation of $W_{1:t}$. We will introduce the following notation, to approximate components from the above equation (Equation 6.32), continuing with the notation from the Equation 6.63:

$$
\begin{aligned}
h_B^{(t,i,k,j)} &= \pi\left(\tilde{Y}_t \mid Y_{t-1}^{(t,i,k,j)}, X_{t-1}^{(t,i,k,j)}, W_{1:t}^{(t,i,k,j)}, f_Y^{B,(t,i)}()\right) \\
u_{B,f_Y,f_X,W}^{(t,i,k,j)} &= \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, f_Y^{B,(t,i)}(), f_X^{B,(t,i,k)}()\right) \\
u_{B,f_Y,f_X}^{(t,i,k)} &= \sum_{j=1}^{N_j} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, f_Y^{B,(t,i)}(), f_X^{B,(t,i,k)}()\right) \\
u_{B,f_Y}^{(t,i)} &= \sum_{k=1}^{N_k} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid f_Y^{B,(t,i)}(), f_X^{B,(t,i,k)}()\right) \\
u_B^{(t)} &= \sum_{i=1}^{N_i} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid f_Y^{B,(t,i)}()\right)
\end{aligned}
$$

The next step of the notation is introduced to remind that we're not simulating the whole Gaussian processes, but only their value on a grid (and the points between the grid will be calculated as conditional

distributions)

$$
\begin{aligned}
\check{h}_B^{(t,i,k,j)} &= \pi\left(\tilde{Y}_t \mid Y_{t-1}^{(t,i,k,j)}, X_{t-1}^{(t,i,k,j)}, W_{1:t}^{(t,i,k,j)}, \check{f}_Y^{B,(t,i)}()\right) \\
\check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)} &= \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, \check{f}_Y^{B,(t,i)}(), \check{f}_X^{B,(t,i,k)}()\right) \\
\check{u}_{B,f_Y,f_X}^{(t,i,k)} &= \sum_{j=1}^{N_j} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid W_{1:t-1}^{(t,i,k,j)}, \check{f}_Y^{B,(t,i)}(), \check{f}_X^{B,(t,i,k)}()\right) \\
\check{u}_{B,f_Y}^{(t,i)} &= \sum_{k=1}^{N_k} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid \check{f}_Y^{B,(t,i)}(), \check{f}_X^{B,(t,i,k)}()\right) \\
\check{u}_B^{(t)} &= \sum_{i=1}^{N_i} \pi\left(\tilde{Y}_{1:t-1}, \tilde{X}_{1:t-1} \mid \check{f}_Y^{B,(t,i)}()\right)
\end{aligned}
$$

All of the summations will be calculated in loops, hence notation for the partial sums (estimation of the unnormalised expectations):

$$
\begin{aligned}
\mathsf{F}_{B,W}^{(t,i,k)} &= \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)} \\
\mathsf{F}_{B,f_X}^{(t,i)} &= \sum_{k=1}^{N_k} \mathsf{F}_{B,W}^{(t,i,k)} = \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)} \\
\mathsf{F}_{B,f_Y}^{(t)} &= \sum_{i=1}^{N_i} \mathsf{F}_{B,f_X}^{(t,i)} = \sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)}
\end{aligned}
$$

and finally:

$$
\hat{\mathsf{E}}_B = \prod_{t=2}^{T} \frac{1}{N_i N_k N_j} \frac{1}{\check{u}_B^{(t)}} \mathsf{F}_{B,f_Y}^{(t)} = \prod_{t=2}^{T} \frac{1}{N_i N_k N_j} \frac{1}{\check{u}_B^{(t)}} \sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \sum_{j=1}^{N_j} \check{h}_B^{(t,i,k,j)} \check{u}_{B,f_Y,f_X,W}^{(t,i,k,j)}
$$

Note! In the case that we're considering here, the two Gaussian processes are independent, and both are independent from the mixing variables. This means that we are allowed to treat the integrals as one multidimensional integrals, and the sum, as one sum, as follows:

$$
\hat{\mathsf{E}}_B = \prod_{t=2}^{T} \frac{1}{N_i} \frac{1}{\check{u}_B^{(t)}} \mathsf{F}_B^{(t)} = \prod_{t=2}^{T} \frac{1}{N_i} \frac{1}{\check{u}_B^{(t)}} \sum_{i=1}^{N_i} \check{h}_B^{(t,i)} \check{u}_B^{(t,i)}
$$

Choose parameters:

$\mathbf{m}^B = \left[m_1^B, ..., m_T^B\right]$, $\boldsymbol{\gamma}^B = \left[\gamma_1^B, ..., \gamma_T^B\right]$, $\nu^B, \sigma_{Y,B}^2, \sigma_{X,B}^2$, hyperparameters for k_Y^B.

Note: start with fixed m_t^B, γ_t^B.

Set two grids of N points $\mathbf{y} = [\mathsf{y}_1, ..., \mathsf{y}_N]$, $\mathbf{x} = [\mathsf{x}_1, ..., \mathsf{x}_N]$

For this grid we will obtain the values for the covariance matrices:

$\mathbf{K}_Y^B$ with elements $\mathsf{k}_{Y,l,m}^B = k_Y^B([\mathsf{x}_l, \mathsf{y}_l]^T, [\mathsf{x}_m, \mathsf{y}_m]^T)$, $l, m \in [1, T]$,

$\mathbf{K}_X^B$ with elements $\mathsf{k}_{X,l,m}^B = k_X^B([\mathsf{x}_l, \mathsf{y}_l]^T, [\mathsf{x}_m, \mathsf{y}_m]^T)$, $l, m \in [1, T]$.

Note: Start with the same grid for all $i = 1..N_i$. Which is why the grid before the loops!

Note: We can use a random grid! It's worth comparing to fixed grid.

Calculate the conditioning number of $\mathbf{K}_Y^B$ and $\mathbf{K}_X^B$. If either $R_{con}(\mathbf{K}_Y^B) < 10^{-3}$ or $R_{con}(\mathbf{K}_X^B) < 10^{-3}$, perform a regularisation (for example with Nystrom approximation).

Precalculate $\left[\mathbf{K}_Y^B + \sigma_{Y,B}^2 \mathbf{I}\right]^{-1}$ and $\left[\mathbf{K}_X^B + \sigma_{X,B}^2 \mathbf{I}\right]^{-1}$.

Initialise: $\hat{\mathsf{E}}_B = 1$.

For $t = 2..T$, repeat (1) - (6), to obtain value $\hat{\mathsf{E}}_B$

$\longmapsto\longrightarrow$**Loop (t)**

Initialise: $\mathsf{F}_B^{(t)} = 0$, $\breve{u}_B^{(t)} = 0$.

For $i = 1..N_i$, repeat (1) - (6), to obtain value $\mathsf{F}_B^{(t)}$ and $\breve{u}_B^{(t)}$.

$\longmapsto\longrightarrow$**Loop (i)**

1. Draw an N-dimensional noisy random vector $\mathbf{g}_Y^{B,(t,i)} = \left[\mathsf{g}_{Y,1}^{B,(t,i)}, ..., \mathsf{g}_{Y,N}^{B,(t,i)}\right] \sim \mathcal{N}(0, \mathbf{K}_Y^B + \sigma_{Y,B}^2 \mathbf{I})$.

2. Precalculate: $\left[\mathbf{K}_Y^B + \sigma_{Y,B}^2 \mathbf{I}\right]^{-1} \mathbf{g}_Y^{B,(t,i)}$.

3. Draw an N-dimensional noisy random vector $\mathbf{g}_X^{B,(t,i)} = \left[\mathsf{g}_{X,1}^{B,(t,i)}, ..., \mathsf{g}_{X,N}^{B,(t,i)}\right] \sim \mathcal{N}(0, \mathbf{K}_X^B + \sigma_{X,B}^2 \mathbf{I})$.

4. Precalculate: $\left[\mathbf{K}_X^B + \sigma_{X,B}^2 \mathbf{I}\right]^{-1} \mathbf{g}_X^{B,(t,i)}$.

5. Draw i.i.d. $W_s^{(t,i)} \sim IG\left(\nu^B/2, \nu^B/2\right)$, $s \in [1, t]$.

6. Assume $Y_0^{(t,i)} = 0$. Evaluate the vector of points $Y_{1:t-1}^{(t,i)}$, as a function of the observations $\tilde{Y}_{1:t-1}$ and the draws of $W_{1:t-1}^{(t,i)}$:

$$\begin{bmatrix} Y_1^{(t,i)} \\ .. \\ Y_{t-1}^{(t,i)} \end{bmatrix} = \begin{bmatrix} \sqrt{W}_1^{(t,i)} & 0 & 0 \\ 0 & .. & 0 \\ 0 & 0 & \sqrt{W}_{t-1}^{(t,i)} \end{bmatrix}^{-1} \left(\begin{bmatrix} \tilde{Y}_1 \\ .. \\ \tilde{Y}_{t-1} \end{bmatrix} - \begin{bmatrix} m_{Y,1}^B \\ .. \\ m_{Y,t-1}^B \end{bmatrix} - \begin{bmatrix} W_1^{(t,i)} & 0 & 0 \\ 0 & .. & 0 \\ 0 & 0 & W_{t-1}^{(t,i)} \end{bmatrix} \begin{bmatrix} \gamma_{Y,1}^B \\ .. \\ \gamma_{Y,t-1}^B \end{bmatrix} \right)$$

7. Assume $X_0^{(t,i)} = 0$. Evaluate the vector of points $X_{1:t-1}^{(t,i)}$, as a function of the observations $\tilde{X}_{1:t-1}$ and the draws of $W_{1:t-1}^{(t,i)}$:

$$\begin{bmatrix} X_1^{(t,i)} \\ .. \\ X_{t-1}^{(t,i)} \end{bmatrix} = \begin{bmatrix} \sqrt{W}_1^{(t,i)} & 0 & 0 \\ 0 & .. & 0 \\ 0 & 0 & \sqrt{W}_{t-1}^{(t,i)} \end{bmatrix}^{-1} \left(\begin{bmatrix} \tilde{X}_1 \\ .. \\ \tilde{X}_{t-1} \end{bmatrix} - \begin{bmatrix} m_{X,1}^B \\ .. \\ m_{X,t-1}^B \end{bmatrix} - \begin{bmatrix} W_1^{(t,i)} & 0 & 0 \\ 0 & .. & 0 \\ 0 & 0 & W_{t-1}^{(t,i)} \end{bmatrix} \begin{bmatrix} \gamma_{X,1}^B \\ .. \\ \gamma_{X,t-1}^B \end{bmatrix} \right)$$

8. Evaluate the conditional distribution $f_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) \mid \mathbf{y}, \mathbf{x}, \mathbf{g}_Y^{B,(t,i)}, Y_{0:t-1}^{(t,i)}, X_{0:t-1}^{(t,i)}$ $\sim \left(f_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right); \bar{f}_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right), cov(f_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right)\right)$, from the Equation 6.35, with the mean and the covariance given by Equations 6.36 and 6.37:

$$\begin{aligned} \bar{f}_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) =& K_Y^B(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T, [\mathbf{x}, \mathbf{y}]^T)\left[\mathbf{K}_Y^B + \sigma_{Y,B}^2\mathbf{I}\right]^{-1}\mathbf{g}_Y^{B,(t,i)}, \\ cov\left(f_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right)\right) =& K_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T, \left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) \\ & - K_Y^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T, [\mathbf{x}, \mathbf{y}]^T\right)\left[\mathbf{K}_Y^B + \sigma_{B,Y}^2\mathbf{I}\right]^{-1} K_Y^B\left([\mathbf{x}, \mathbf{y}]^T, \left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) \end{aligned}$$

9. Evaluate the conditional distribution $f_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) \mid \mathbf{y}, \mathbf{x}, \mathbf{g}_X^{B,(t,i)}, Y_{0:t-1}^{(t,i)}, X_{0:t-1}^{(t,i)}$ $\sim \left(f_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right); \bar{f}_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right), cov(f_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right)\right)$, analogously to the Equation 6.35, with the mean and the covariance given analogously to the Equations 6.36 and 6.37:

$$\begin{aligned} \bar{f}_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) =& K_X^B(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T, [\mathbf{x}, \mathbf{y}]^T)\left[\mathbf{K}_X^B + \sigma_{X,B}^2\mathbf{I}\right]^{-1}\mathbf{g}_X^{B,(t,i)}, \\ cov\left(f_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right)\right) =& K_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T, \left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) \\ & - K_X^B\left(\left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T, [\mathbf{x}, \mathbf{y}]^T\right)\left[\mathbf{K}_X^B + \sigma_{B,X}^2\mathbf{I}\right]^{-1} K_X^B\left([\mathbf{x}, \mathbf{y}]^T, \left[X_{0:t-1}^{(t,i)}, Y_{0:t-1}^{(t,i)}\right]^T\right) \end{aligned}$$

10. Evaluate $\check{h}_B^{(t,i)} \approx \pi\left(\tilde{Y}_t \mid W_t, X_{t-1}, Y_{t-1}, f_Y^B()\right)$, using the normal distribution from 6.38:

$$\mathcal{N}\left(m_t + \gamma_t W_t^{(t,i)} + \sqrt{W_t^{(t,i)}}\bar{f}_Y^B(\left[X_{t-1}^{(t,i)}, Y_{t-1}^{(t,i)}\right]^T), W_t^{(t,i)}\left(\sigma_{Y,B}^2 + cov(f_Y^B(\left[X_{t-1}^{(t,i)}, Y_{t-1}^{(t,i)}\right]^T))\right)\right).$$

11. Evaluate $\breve{u}_B^{(t,i)} \approx \pi\left(\tilde{Y}_{1:t-1}, \tilde{Y}_{1:t-1} \mid W_{1:t-1}, f_Y^B(), f_X^B()\right)$, using distribution from 6.50:

$$\breve{u}_B^{(t,i)} \sim \mathcal{N}\Big(\tilde{X}_{1:t-1}; \mathbf{m}^B_{X,1:t-1} + \left[\boldsymbol{\gamma}^B_{X,1} W_1^{(t,i)}, ..., \boldsymbol{\gamma}^B_{X,t-1} W_{t-1}^{(t,i)}\right]^T + \left[\sqrt{W_1^{(t,i)}}, ..., \sqrt{W_{t-1}^{(t,i)}}\right]^T \mathbb{I} \bar{f}_X^B\left(\left[X_{0:t-2}^{(t,i)}, Y_{0:t-2}^{(t,i)}\right]^T\right),$$
$$\sqrt{W_{1:t-1}^{(t,i)}}^T \mathbb{I}\left[\sigma^2_{X,B}\mathbb{I} + cov\left(f_X^B\left(\left[X_{0:t-2}^{(t,i)}, Y_{0:t-2}^{(t,i)}\right]^T\right)\right)\right]\mathbb{I}\sqrt{W_{1:t-1}^{(t,i)}}^T\Big)$$
$$\mathcal{N}\Big(\tilde{Y}_{1:t-1}; \mathbf{m}^B_{Y,1:t-1} + \left[\boldsymbol{\gamma}^B_{Y,1} W_1^{(t,i)}, ..., \boldsymbol{\gamma}^B_{Y,t-1} W_{t-1}^{(t,i)}\right]^T + \left[\sqrt{W_1^{(t,i)}}, ..., \sqrt{W_{t-1}^{(t,i)}}\right]^T \mathbb{I} \bar{f}_Y^B\left(\left[X_{0:t-2}^{(t,i)}, Y_{0:t-2}^{(t,i)}\right]^T\right),$$
$$\sqrt{W_{1:t-1}^{(t,i)}}^T \mathbb{I}\left[\sigma^2_{Y,B}\mathbb{I} + cov\left(f_Y^B\left(\left[X_{0:t-2}^{(t,i)}, Y_{0:t-2}^{(t,i)}\right]^T\right)\right)\right]\mathbb{I}\sqrt{W_{1:t-1}^{(t,i)}}^T\Big).$$

12. Update the sums: $\mathsf{F}_B^{(t)} := \mathsf{F}_B^{(t)} + \check{h}_B^{(t,i)}\breve{u}_B^{(t,i)}$ $\breve{u}_B^{(t)} := \breve{u}_B^{(t)} + \breve{u}_B^{(t,i)}$.

End **Inner Loop (i)**

Update the multiplication: $\hat{\mathsf{E}}_B := \hat{\mathsf{E}}_B \frac{\mathsf{F}_B^{(t)}}{\breve{u}_B^{(t)}}$

End **Loop (t)**

Obtain the final value of the estimation: $\hat{\mathsf{E}}_B := \frac{1}{N_i N_k N_j}\hat{\mathsf{E}}_B$.

End of the algorithm.

7.2 On simulating autoregressive time series data with GP

Causality structure 1. Autoregressive time series with linear and power law causality.

$$\begin{aligned} X_t &= f_X(X_{t-1}) & f_X &\sim \mathcal{GP}\left(\mu_{X,t}, k_{X,t,t'}\right) \\ Y_t &= f_Y([Y_{t-1}, X_{t-1}]) & f_Y &\sim \mathcal{GP}\left(\mu_{Y,t}, k_{Y,t,t'}\right) \\ Z_t &= f_Z([Z_{t-1}, Y_{t-1}]) & f_Z &\sim \mathcal{GP}\left(\mu_{Z,t}, k_{Z,t,t'}\right) \end{aligned} \tag{7.5}$$

This data is introduced as a time series, with the means as below:

$$\begin{aligned} \mu_{X,t} &= \mu_{X,t}(X_{t-1}) = a_X X_{t-1} \\ \mu_{Y,t} &= \mu_{Y,t}([Y_{t-1}, X_{t-1}]) = a_Y Y_{t-1} + b_Y X_{t-1} \\ \mu_{Z,t} &= \mu_{Z,t}([Z_{t-1}, Y_{t-1}]) = a_Z Z_{t-1} + b_Z Y_{t-1}^2 \end{aligned} \tag{7.6}$$

and noise is already incorporated in the Gaussian processes:

$$
\begin{aligned}
k_{X,t,t'} &= k_{X,t,t'}(X_{t-1}, X_{t'-1}) = k^{Matern}_{l_a,s_f}(X_{t-1}, X_{t'-1}) + s_n^2\delta_{t,t'} \\
k_{Y,t,t'} &= k_{Y,t,t'}([Y_{t-1}, X_{t-1}], [Y_{t'-1}, X_{t'-1}]) = k^{Matern}_{l_a,l_b,s_f}([Y_{t-1}, X_{t-1}], [Y_{t'-1}, X_{t'-1}]) + s_n^2\delta_{t,t'} \\
k_{Z,t,t'} &= k_{Z,t,t'}([Z_{t-1}, Y_{t-1}], [Z_{t'-1}, Y_{t'-1}]) = k^{Matern}_{l_a,s_f}([Z_{t-1}, Y_{t-1}], [Z_{t'-1}, Y_{t'-1}]) + s_n^2\delta_{t,t'}
\end{aligned} \tag{7.7}
$$

The Matern covariance:

- d denote distance between two points w and w'
- $C(d)$ Matern covariance as a function of d; $C(d) = C(d(x, y))$
- ν is a non-negative parameter of the covariance, referred to as a "smoothness parameter" or as degrees of freedom
- l is a non-negative parameters of the covariance, referred to as a lenghtscale; we define it here as a vector that introduce Automatic Relevance Determination (ARD), if it's zero for a particular dimention, then this dimension is removed; please note that if it's zero for all dimensions, then the value fo the kernel function is equal to σ^2
- σ^2 is a a parameter of the covariance, it's also the limit of the Matern covariance function in $d = 0$
- $K_\nu(z)$ denotes the modified Bessel function of the second kind
- Γ is the gamma function

$$
C(d) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu}dl\right)^\nu K_\nu\left(\sqrt{2\nu}dl\right) \tag{7.8}
$$

In the kernel definitions above, we use $l = [l_a]$ for k_X, k_Z and $l = [l_a, l_b]^T$ for k_Y. The l_a is the reciprocal lengthscale corresponding to the autocorrelation, while l_b – causality. If we choose $l_b = 0$, then there's no causality in the covariance, while $l_a = 0$ affects the autocorrelation in covariance. If we choose $\sigma^2 = 0$, then the Matern kernel component is removed altogether with any autocovariance or causality in the dependence, and we end up with a simple AR(1):

$$
\begin{aligned}
X_t &= a_X X_{t-1} + \epsilon_X \\
Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_Y \\
Z_t &= a_Z Z_{t-1} + b_Z Y^q_{t-1} + \epsilon_Z
\end{aligned} \tag{7.9}
$$

This data is introduced as a time series, but the errors are generated as Gaussian processes, which depend on previous value of the time series:

$$\epsilon_Y \sim \mathcal{N}\left(0, \sigma^2\right),$$
$$\epsilon_Z \sim \mathcal{N}\left(0, \sigma^2\right),$$

How to simulate the data The fact that we have data $Y_1, ..., Y_n$ from a Gaussian process means that this data has joint normal distribution with appropriate mean function and kernel. When no autoregression is present, the data generation is simple: get all of the input points $W_1, ..., W_n$, apply mean function $\mu(\cdot)$ and covariance function $k_{t,t'}(\cdot)$ to obtain mean vector m and covariance matrix K and generate n points normal distribution $\mathcal{N}(m, K)$.

But when autoregression is present (and even worse: causality as well), then everything needs to be simulated recursively, with every new step agreeing with the all previous steps. This means generating from the posterior distribution. Quoting Ebden [Ebden, 2015] "Since the key assumption in GP modelling is that our data can be represented as a sample from a multivariate Gaussian distribution, we have that

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K & K^{*T} \\ K_* & K^{**} \end{bmatrix}\right),$$

"

The equation above was given for a GP with mean zero, and with the notation: y^* for a value of y at a new point (for convenience lets call it a testing point, although in our case that won't be meaningful), $\mathbf{y}$ for all past values of y (lets call them training points). The kernel notation is as follows: K as a Gramm matrix for all training points, K^{**} – for the training point, and K^* for a cross-covariance matrix of all training points with the testing point.

In our model the "training" point is Y_t, the "testing" points are all of the earlier ones: $Y_1, ..., Y_{t-1}$. Let's introduce the short notation of K for Gramm matrix of all "training points", $K_{t,t}$ for $cov(Y_t, Y_t)$ and K_t for a vector of cross covariances $cov(Y_s, Y_t)$, $s = 1, ..., t-1$. The joint distribution of $Y_1, ..., Y_{t-1}, Y_t$ is then:

$$\begin{bmatrix} Y_1 \\ ... \\ Y_{t-1} \\ Y_t \end{bmatrix} \sim \mathcal{N}\left(\mu_Y \begin{bmatrix} Y_1 \\ ... \\ Y_{t-1} \\ Y_t \end{bmatrix}, \begin{bmatrix} K & K_t^T \\ K_t & K_{t,t} \end{bmatrix}\right), \tag{7.10}$$

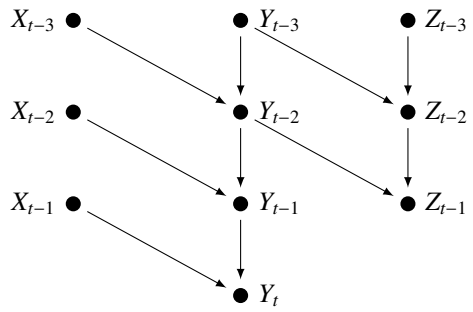


Figure 7.3: Direct Acyclic Graph (DAG) representation of the time series X, Y, Z from the point of view of generating a realization of Y_t. The graph helps to visually explain why $Y_t \perp\!\!\!\perp Y_{t-n} \mid Y_{t-1}$ for any $n >= 2$.

In our case the data is generated recurrently, so when adding the Y_t point, it needs to consistend with the distribution of previous points, and hence it should be drawn from a conditional distribution of $p(Y_t \mid Y_1, ..., Y_{t-1})$. Based on equation 7.10 and properties of normal distribution will be as follows, and with $\mu \equiv \mu([Y_1, ..., Y_{t-1}]^T)$ and $\mu_t = \mu(Y_t)$ the conditional distribution will be:

$$p(Y_t \mid Y_1, ..., Y_{t-1}) \sim \mathcal{N}(\mu_t + K_t K^{-1}([Y_1, ..., Y_{t-1}]^T - \mu), K_{tt} - K_t K^{-1} K_t^t). \tag{7.11}$$

In general simulating data in this way will be either computationally expensive (having to inverse a matrix K with grows with each iteration) or would need approximation techniques. However in case of the time series structure as introduced by 7.5, we don't need to include all history of Y. If we look at the graph associated with our model:

As the graph 7.3 explains, we have the conditional independence of $Y_t \perp\!\!\!\perp Y_{t-n} \mid Y_{t-1}$ for any $n >= 2$. This means, that $p(Y_t \mid Y_1, ..., Y_{t-1}) = p(Y_t \mid Y_{t-1})$ and instead of requiring consistence with all of the previous points, it's enough to ensure the consistence of new point Y_t with the previous point Y_{t-1}. Hence, we can generate the point Y_t from the conditional distribution:

$$p(Y_t \mid Y_{t-1}) \sim \mathcal{N}(\mu_{Y,t} + \frac{k_{Y,t,t-1}}{k_{Y,t-1,t-1}}(Y_{t-1} - \mu_{Y,t-1}), k_{Y,t,t} - \frac{k_{Y,t}^2}{k_{Y,t-1,t-1}}). \tag{7.12}$$

We've been concentrating on generating time series Y, but analogously we will generate Z – from the conditional distribution:

$$p(Z_t \mid Z_{t-1}) \sim \mathcal{N}(\mu_{Z,t} + \frac{k_{Z,t,t-1}}{k_{Z,t-1,t-1}}(Z_{t-1} - \mu_{Z,t-1}), k_{Z,t,t} - \frac{k_{Z,t}^2}{k_{Z,t-1,t-1}}). \tag{7.13}$$

7.3 Efficient testing procedures

With the simplifying assumptions we were able to achieve closed form solutions for our causality metric $L_{X\to Y|Z}$ and approximate distribution. But for some data, we will find that the computations are theoretically achievable, but very impractical.

Let's recall the log marginal-likelihood used for the causality metric $L_{X\to Y|Z}$:

$$\begin{aligned}&\log \pi(\mathbf{Y} \mid \ X^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m})\\ &= \frac{1}{2}(\mathbf{Y}-\mu_B)^T(K_B+\sigma^2_{B,t})^{-1}(\mathbf{Y}-\mu_B) - \frac{1}{2}\log \mid K_B+\Sigma^B \mid -\frac{N}{2}\log 2\Pi.\end{aligned}$$

Here the inversion $(K_B+\Sigma^B)^{-1}$, and similarly $(K_A+\Sigma^A)^{-1}$ for the model A, can become computationally very expensive if the matrix becomes high dimensional. The second place, where the same matrix inversion appears is when calculating the derivatives of the marginal log-likelihood needed to optimise the hyperparameters:

$$\begin{aligned}&\frac{\partial}{\partial\theta^B_j}\log\pi(Y \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \theta_B,) =\\ &\frac{1}{2}\left(\left(\alpha\alpha^T - (K_B+\Sigma^B)^{-1}\right)\frac{\partial(K_B+\Sigma^B)}{\partial\theta^B_j}\right), \text{ where } \alpha = (K_B+\Sigma^B)^{-1}(\mathbf{Y}-\mu_B)\end{aligned}$$

Referring to Chapter (5), here is a reminder of what $(K_B+\Sigma^B)$ is:

$$\begin{aligned}\left\{K_B+\Sigma^B\right\}_{t_1,t_2} &= k_{B,t_1,t_2} + \sigma^2_{B,t}\delta_{t_1,t_2}\\ &= k_B\left(\left[\mathbf{X}^{-k}_{t_1-1}, \mathbf{Y}^{-l}_{t_1-1}, \mathbb{Z}^{-m}_{t_1-1}\right], \left[\mathbf{X}^{-k}_{t_2-1}, \mathbf{Y}^{-l}_{t_2-1}, \mathbf{Z}^{-m}_{t_2-1}\right]\right) + \sigma^2_{B,t}\delta_{t_1,t_2}\\ &= cov(\mathbf{Y}_{t_1}, \mathbf{Y}_{t_2}).\end{aligned}$$

There are some suggestions how to achieve an efficient implementation (for example mentioned by Garthwaite et al. [2002]) including: the use of Cholesky decomposition instead of straightforward inversion, pre-computing elements involving the matrix inverse that are used repetitively for example in the calculation of derivatives. But for very high dimensional problems, that may not suffice. Might be necessary to introduce some sparse approximations.

7.4 Software for Causality

For the purpose of the experiments described in the section 2.2, we have used code from several sources: own Matlab code, open access Matlab toolbox for Granger causality GCCA[1] [Seth, 2010] and open access Matlab code provided by Sohan Seth [Seth and Principe, 2011][2].

[1]The code can be requested from the Author's website: `http://www.sussex.ac.uk/Users/anils/aks_code.htm`
[2]Code available at `http://www.sohanseth.com/Home/publication/causmci`

Our code allows the calculation of: a) kernelised Geweke's measures of Granger causality with 3 kernels: Gaussian, linear, polynomial; choice of the linear kernel gives linear Geweke's measure (the effect of regularisation negligible), b) transfer entropy based on naive histogram. Parameters for the ridge regression can be chosen with n-fold cross-validation [3], and the significance of the measurement can be calculated with a permutation test.

The GCCA tooldbox has been very popular for Granger causality, including causality in the frequency domain and analysing Granger causality network. Also the GCCA toolbox is employed in the code written by Seth. For many examples we incorporated the framework from Seth's code as this proved to be better optimised to run large quantities of tests. From the original Seth's framework we have employed the tests for HSNCIC and Granger causality (using GCCA toolbox, results in line with the ones were our own code was used, but faster), we have added our own tests for kernelised Granger causality and transfer entropy. In the Seth's framework we have also changed the implementation for permutation tests from using rotation to using actual permutation, which we believe is more suitable given that we never run fewer tests per experiment that Seth performed.

The GCCA toolbox[4] for calculating Granger causality provides some tools for detecting non-stationarity and to a limited degree also for managing it [Seth, 2010]. In the VAR setting of Granger causality it is possible to run parametric tests to detect nonstationarity: ADF test (Augmented Dickey Fuller) and KPSS test (Kwiatkowski, Phillips, Schmidt, Shin). For managing non-stationarity Seth suggests analysing shorter time series (windowing) and differencing, although both approaches can introduce new problems. It also advisable to detrend and demean the data, and in the case of for example economic data it might also be possible to perform seasonal adjustment.

7.5 Software for Gaussian Processes

7.5.0.1 Errors in the GPML Matern covariance

Below, I'm presenting original code for `covMaterniso.m` from the GPML toolbox. The function `covMaterniso.m` uses the same lengthscale parameter for each dimension, the toolbox also contains a function `covMaternard.m`, which uses different lengthscale for each dimension. Both functions are very similar (and contains the same mistakes).

```
function K = covMaterniso(d, hyp, x, z, i)


% Matern covariance function with nu = d/2 and isotropic distance
% measure. For d=1 the function is also known as the exponential
```

[3] Validation is a process of confirming that a model or parameters are acceptable to describe the given data. In machine learning validation is often performed by splitting the data into training and validation sets and analysing the error of modelling the validation set with parameters optimised on the training set. Cross-validation is a type of validation when the whole procedure is performed several time with the same data being randomly selected to either the training or in the validation set. For more information on model selection and cross-validation please refer to Friedman et al. [2001] or Grasa [2013]

[4] Code can be requested at: `http://www.sussex.ac.uk/Users/anils/aks_code.htm`

```
% covariance function for the Ornstein-Uhlenbeck covariance in 1d.
% The covariance function is:
%
%   k(x^p,x^q) = sf^2 * f( sqrt(d)*r ) * exp(-sqrt(d)*r)
%
% with f(t)=1 for d=1, f(t)=1+t for d=3 and f(t)=1+t+tA^2/3 for d=5.
% Here r is the distance sqrt((x^p-x^q)'*inv(P)*(x^p-x^q)), P is ell
% times the unit matrix and sf2 is the signal variance. The
% hyperparameters are:
%
% hyp = [ log(ell)
%         log(sf) ]
%
% Copyright (c) by Carl Edward Rasmussen and Hannes Nickisch,
% 2010-09-10.
% See also COVFUNCTIONS.M.

if nargin<3, K = '2'; return; end       % report number of parameters
if nargin<4, z = []; end                          % make sure, z exists
xeqz = isempty(z); dg = strcmp(z,'diag');            % determine mode

ell = exp(hyp(1));
sf2 = exp(2*hyp(2));
if all(d~=[1,3,5]), error('only 1, 3 and 5 allowed for d'), end
% degree

switch d
                                                  % df(t) = f(t)-f'(t)
  case 1, f = @(t) 1;                  df = @(t) 1;
  case 3, f = @(t) 1 + t;              df = @(t) t;
  case 5, f = @(t) 1 + t.*(1+t/3);     df = @(t) t.*(1+t)/3;
end
          m = @(t,f) f(t).*exp(-t); dm = @(t,f) df(t).*exp(-t).*t;

% precompute distances
if dg                                                      % vector kxx
  K = zeros(size(x,1),1);
else
  if xeqz                                          % symmetric matrix Kxx
    K = sq_dist(sqrt(d)/ell*x');
  else                                            % cross covariances Kxz
    K = sq_dist(sqrt(d)/ell*x',sqrt(d)/ell*z');
  end
```

```
end

if nargin<5                                                    % covariances
  K = sf2*m(sqrt(K),f);
else                                                           % derivatives
  if i==1
    K = sf2*dm(sqrt(K),f);
  elseif i==2
    K = 2*sf2*m(sqrt(K),f);
  else
    error('Unknown hyperparameter')
  end
end
```

The function above uses simplified forms of the Matern covariance for three degrees of freedom: 0.5, 1.3, 2.5. While values of the covariance function are correct, the derivatives are not. Below I'm presenting the analytical derivation of the derivatives for $\nu = 3/2$ and show why ones in the function above are incorrect.

$$C_{\nu=3/2}(d) = \sigma^2\left(1 + \frac{\sqrt{3}d}{l}\right)\exp\left(-\frac{\sqrt{3}d}{l}\right) \tag{7.14}$$

7.5.0.2 Correct derivatives:

With respect to σ:

$$\frac{\partial C_{\nu=3/2}(d)}{\partial \sigma} = 2\sigma\left(1 + \frac{\sqrt{3}d}{l}\right)\exp\left(-\frac{\sqrt{3}d}{l}\right), \tag{7.15}$$

with respect to lengthscale:

$$\begin{aligned}
&\frac{\partial C_{\nu=3/2}(d)}{\partial l} \\
&= \sigma^2\frac{\partial\left(1+\frac{\sqrt{3}d}{l}\right)}{\partial l}\exp\left(-\frac{\sqrt{3}d}{l}\right) + \sigma^2\left(1+\frac{\sqrt{3}d}{l}\right)\frac{\partial\exp\left(-\frac{\sqrt{3}d}{l}\right)}{\partial l} \\
&= \sigma^2\left(-\frac{\sqrt{3}d}{l^2}\right)\exp\left(-\frac{\sqrt{3}d}{l}\right) + \sigma^2\left(1+\frac{\sqrt{3}d}{l}\right)\exp\left(-\frac{\sqrt{3}d}{l}\right)\left(\frac{\sqrt{3}d}{l^2}\right) \\
&= \frac{-\sigma^2\sqrt{3}d}{l^2}\exp\left(-\frac{\sqrt{3}d}{l}\right) + \frac{\sigma^2\sqrt{3}d}{l^2}\exp\left(-\frac{\sqrt{3}d}{l}\right) + \frac{\sigma^2\sqrt{3}d}{l}\left(\frac{\sqrt{3}d}{l^2}\right)\exp\left(-\frac{\sqrt{3}d}{l}\right) \\
&= \frac{3\sigma^2 d^2}{l^3}\exp\left(-\frac{\sqrt{3}d}{l}\right)
\end{aligned} \tag{7.16}$$

Derivatives in the GPML code:

Note that the hyperparameters are read as follows:

```
ell = exp(hyp(1));
sf2 = exp(2*hyp(2));
```

which means that the second parameter always appears as σ^2.

The derivative with respect to σ is therefore as follows:

$$\frac{\partial C_{\nu=3/2}(d)}{\partial \sigma} = 2\sigma^2\left(1+\frac{\sqrt{3}d}{l}\right)\exp\left(-\frac{\sqrt{3}d}{l}\right), \tag{7.17}$$

which is incorrect, and would be incorrect even if the derivative was taken with respect to σ^2 rather than σ.

For the derivative with respect to lengthscale, some explanations as to the details of the code are needed. `sq_dist` is a function that computes a matrix of all pairwise squared distances between two sets of vectors and it is homogenous of degree 2 and so:

$$\texttt{sq\_dist}\left(\frac{\sqrt{3}}{l}\underline{x}', \frac{\sqrt{3}}{l}\underline{x}'\right) = \frac{3}{l^2}\texttt{sq\_dist}\left(\underline{x}', \underline{x}'\right) \equiv \frac{3}{l^2}\mathbf{d}^2. \tag{7.18}$$

For clarity when comparing with the code, I've used above the notation of $\underline{x}$ for the vector and $\mathbf{d}$ for the matrix. Below I drop this notation as it's not neccessary, because many operations in the code are performed element-wise anyway.

To compute the derivative with respect to the lengthscale, the code defines the following functions (I fix the degrees of freedom to be equal $3/2$):

$$\begin{aligned} f(t) &\equiv 1+t \\ df(t) &\equiv 1 \\ m(t,f) &\equiv f(t) * \exp(-t) = (1+t) * \exp(-t) \\ dm(t,f) &\equiv df(t) * \exp(-t) = exp(-t) \end{aligned} \tag{7.19}$$

then the derivative:

$$\frac{\partial C_{\nu=3/2}(d)}{\partial \sigma} = \sigma^2 dm\left(\frac{\sqrt{3}}{l}d, f\right)\frac{3}{l^2}d^2 = \frac{3\sigma^2 d^2}{l^2}\exp(-\frac{\sqrt{3}}{l}d) \tag{7.20}$$

The difference is that 7.20 contains a division by l^2, while 7.16 has a division by l^3. After introducing that correction, I'm obtaining same results when using either my own code for Matern function or the code from GPML toolbox. The same fix works for all three options of degrees of freedom that exist in the GPML toolbox.

Chapter 8

Experiments

> *Nature behaves in ways that look mathematical, but nature is not the same as mathematics. Every mathematical model makes simplifying assumptions; its conclusions are only as valid as those assumptions.*
>
> Ian Stewart and Martin Golubitsky, *Fearful Symmetry: Is God a Geometer?*

The framework for modelling and testing causality proposed in this article is novel, and thus it is important to demonstrate that it behaves appropriately. Therefore, we have designed a series of tests of performance and robustness of our framework. The experiments in this chapter are all based on synthetic data, for the results on real data we direct the reader to Chapter (9). Firstly, we explain how the synthetic data sets are generated. The actual experiments begin with sensitivity and misspecification tests, which are followed with experiments on the power of the test (probability of rejecting null hypothesis if it's not true, also equal to 1 - type II error rate) for simple and compound tests. We refer to the 3 types of data with different structural and causal features, that were introduced in the subsection 3.2.

The sensitivity analysis shows how the test reacts to varying the parameter values used to generate the time series data in Example model 1 (Section 3.2). Here, we know the exact model so that test is a simple test where we asses its power over the parameter space.

The model misspecification tests show how the test reacts to discrepancy between the parameter values used to generate the time series data for Examples in Section 3.2 and the parameters used in the test statistic. This is a structured form of compound test analysis, since in practical settings in general the parameters will be estimated from data and then used in a compound testing procedure, and therefore even in a synthetic study with known parameters, they will still not correspond to the "true" values used to generate the synthetic data time series.

Note: throughout these tests the 50% change in the parameters relates to the model parameters, and the covariance parameters are all used as logarithm, so the actual decrease/increase is much bigger than for the mean.

The analysis of the power of the hypothesis tests shows that the framework not only behaves as

expected, but also has properties that make it practical. An important result in this research is obtaining a test statistic with known asymptotic distribution, but what is even more important is that we don't need a very large sample to to be able to use that result in practice. For simple tests – ones that use exact hyperparameters, and compound tests – where the hyperparameters are estimated, we look at popular tools for assessing quality of a testing procedure: test statistic distribution, power of the test and the ROC curves.

Before going into specific result, an illustration of the type of outputs that we have when running our simulations / analysis. Below two examples showing the values of the test statistics from the Equation 5.9 change for different data samples, and what values of the χ^2 cdf they would obtain. The rejection level of 0.9 (significance value of $\alpha = 0.1$) is a value that we will often use, but that has been chosen arbitrarily. The Figure C.1 illustrates a compound test with optimised parameters – showing the values of test statistics $L_{X\to Y}$ vs $L_{Y\to X}$ and the distribution $\chi^2_2(2L_{X\to Y})$ vs $\chi^2_2(2L_{Y\to X})$. The data has been generated from causality structure 1 with strong causal effect $X \to Y$, with each of the 50 data sample being of length 500.

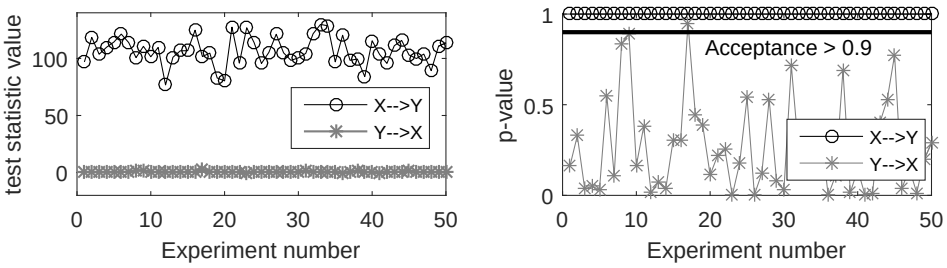


Figure 8.1: Test statistics and corresponding cumulative density function evaluations. Causality structure 1, true parameters: $a_X = a_Y = a_Z = 0.3, b_Y = b_Z = 0.7, q = 2, l_a = l_b = e^{-6}, \sigma_f = e^{-10}, \sigma_n = 0.01$. The horizontal axis represents 50 separate trials, each with a time series of length 500.

The interpretation of the Figure C.1 is the following. From the left plot we can see that the test statistics $L_{X\to Y}$ has values which are separated from and considerably larger than the test statistics $L_{Y\to X}$. This by itself is an indication that the causal effect $X \to Y$ should be stronger than $Y \to X$. From the plot of cdf evaluations we observe that all of the values of $L_{X\to Y}$ are in the tail (with cdf values of exactly 1) and therefore the null hypothesis is strongly rejected at any confidence level, for each of the trials. This means that the estimator of the power of the test (probability of rejecting null hypothesis if it's not true, also equal to 1 - type II error rate) is equal to 1 at any confidence level. If we set up confidence level at 0.1, then one trial will lead to rejecting the null hypothesis in the $Y \to X$ direction, which corresponds to type I error rate of 0.02.

8.1 Power of the Hypothesis Tests

Summary of the section: *analysing power of the test (1-rate of type II error) is popular for assessing quality of a test or a testing procedure. It is expected that the power of the test will increase with the increasing sample size, and showing that this is indeed the case for our testing procedure will be the focus of this and the following sections. We start by analysing the results of simple tests, where exact parameters are used, and there is no effect of parameter misspecification. Strictly speaking, the simple test can be performed only for the first two data structures, as the third has been defined as an econometric model with no Gaussian Process representation. However for the third data structure we perform a few tests with chosen parameters – to show reaction of the test to some properties of the data.*

8.1.1 Simple Tests

Example Time Series Model Structure 1: When using the exact parameters, as in a simple test, typically the behaviour for the Example model data 1 (Equations 3.51 and 3.52) is as expected: power of the test increases with the sample size, and even in case of short time series the classification rule works well. The notable exceptions observed are as detailed below. The Figure 8.2 shows evolution of receiver operating characteristic (ROC) curves with increasing sample length, for three sets of parameters. The left chart represents typical behaviour for most of the parameters: positives and negatives almost always properly classified, even for short time series. The middle and the right figures coincide with large value of $\sigma^2 = e^{-2} \simeq 0.1353$. The right chart shows an extreme case, where the power of the test degrades with length of the time series to a random coin flip on the hypothesis, although it improves if we consider exceptionally long samples of 5000 data points.

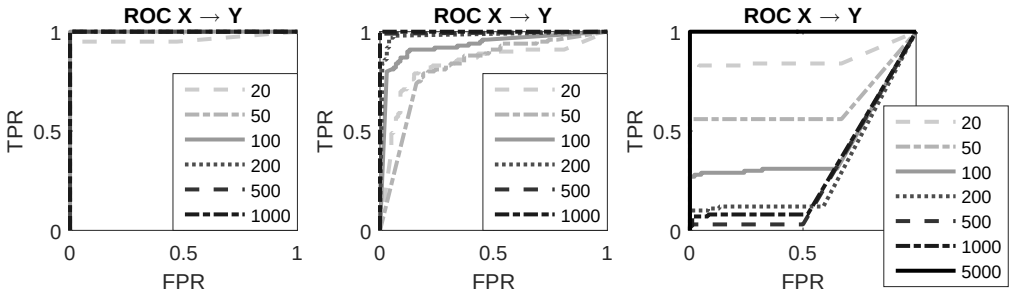


Figure 8.2: Examples of parameter combinations for which the ROC curve shows different behaviour with longer sample (time series). True parameters: $a_X = 0.3, b_Y = 0.7$ in all 3 charts, the kernel parameters respectively: (left) $l_a = e^{-3}, l_b = e^{-1}, \sigma^2 = e^{-10}$, (middle) $l_a = e^{-3}, l_b = e^{-1}, \sigma^2 = e^{-2}$ and (right) $l_a = e^{-1}, l_b = e^{-3}, \sigma^2 = e^{-2}$.

The Figure 8.3 shows box plots representing the distributions of the test statistics for data corresponding to that from the Figure 8.2. In line with the ROCs, distributions of the test statistic in the first set of data converges to 1 very quickly (sample of size 100). For the second set of data we see much slower convergence and increased number of outliers for the data of middle sizes. The third set of data

sees distribution that would appear to converge to 0 if only the standard samples of 20 - 1000 sizes were considered, but that eventually rebounds to value 1 for data length 5000.

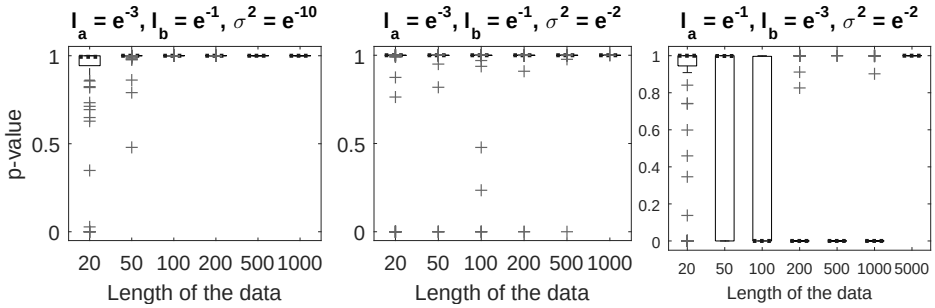


Figure 8.3: Examples of parameter combinations that lead to different evolution of the test statistics distribution. True parameters: $a_X = 0.3, b_Y = 0.7$ in all 3 charts, the kernel parameters respectively: (left) $l_a = e^{-3}, l_b = e^{-1}, \sigma^2 = e^{-10}$, (middle) $l_a = e^{-3}, l_b = e^{-1}, \sigma^2 = e^{-2}$ and (right) $l_a = e^{-1}, l_b = e^{-3}, \sigma^2 = e^{-2}$.

The parameters that cause such behaviour is primarily the signal noise σ^2, and to a smaller extent l_a – the coefficient of autoregression in covariance function. The hyperparameter σ^2 increases the value of the covariance proportionately, while l_a - inversely and less than proportionately. Higher values of the covariance function means higher volatility clustering, an effect which could compete with causality, but that could be less visible in short time series. We won't elaborate on this point here, but additional dependence structure can complicate explanation of causality structure. Therefore longer time series appears necessary to correctly recognise causality in this case. The Figure 8.4 shows the effect of length of a time series on the value of the test statistics $L_{X\to Y}$ for a particular combination of parameters. A single data set of length 5000 has been simulated and subsequently tests statistics have been calculated on the first $100, 200, 300, ...5000$ data points. The chosen data set has a general trend of test statistics increasing for longer data lengths (as for all other data sets generated with the same parameters) but it shows to major dips of test statistics temporarily worsening.

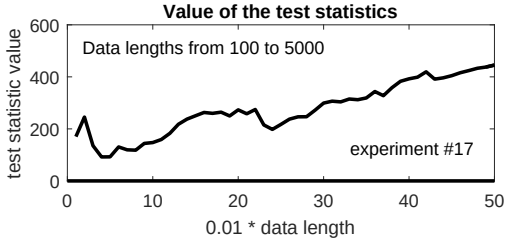


Figure 8.4: Evolution of $L_{X\to Y}$ when (overlapping) data of different length is used. True parameters: $a_X = 0.3, b_Y = 0.7, l_a = e^{-3}, l_b = e^{-3}, \sigma^2 = e^{-2}$.

The causal effect in the covariance function is difficult to observe. This is because on one hand it seems to have a much subtler effect than the causality in mean, but also because it is entwined with other effects that can be observed for different parameter combinations. Figure 8.5 shows that for following parameters $b_Y = 0, a_Y = 0, l_a = e, l_b = e, \sigma^2 = e^4$ the causality in covariance is unambiguously observed

already for sample size of 50. Reminder, according to the Equations 3.51, $b_Y = 0$ means no causality in the mean and $a_Y = 0$ means no autoregression in the mean.

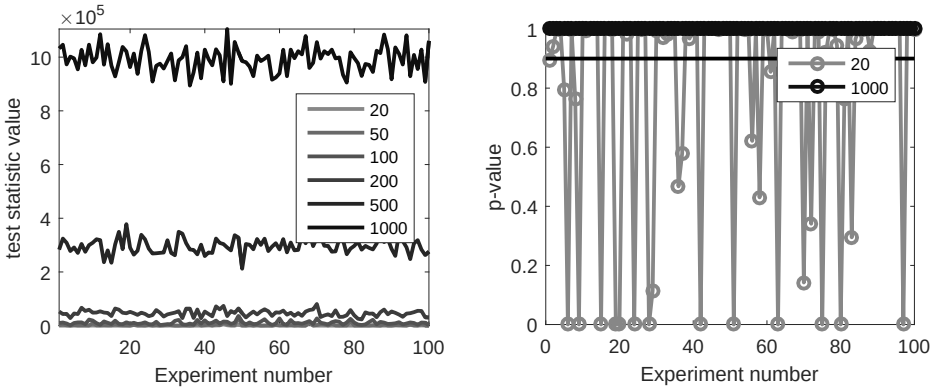


Figure 8.5: Test statistics and the distribution evaluation: no causality in mean ($b_Y = 0$), no autocorrelation in mean ($a_Y = 0$), very large covariance parameters $l_a = e, l_b = e, \sigma^2 = e^4$. The right subplot doesn't explicitly show distribution evaluations for sample sizes from 50 to 500, because they are all equal 1 (just like for sample size 1000).

Example Time Series Model Structure 2: The results for Example model data 2 (Section 3.2) are just commented on here since in the simple testing framework, they don't show anything unexpected. In particular, the power of the test does increase with increasing length of the time series. Arguably, there is much less opportunity for problematic behaviour. This is firstly because the range of parameters which are available for the Example structure 2 is much narrower than for the Example structure 1 (i.e. parameters for which the series does not explode to infinity). Secondly, we assumed $cov(\epsilon_{Y,t}, \epsilon_{Y,t'}) = 0$, but if we didn't we could have had again the problem with volatility clustering masquerading as causality.

Example Time Series Model Structure 3: We do however report a few observations on the data 3. Firstly, data 3 does not have a Gaussian Process representation, so when reporting on the results of the "simple test" in this case we don't mean a test with "true" parameters, but a test with fixed, rather than optimised parameters. These observations become particularly interesting when compared with the results of the compound test for the data 3. The main property of interest in the data 3 is the long memory and this is what we concentrate on here. When analysing results for the data 3 (simple or compound test), on one hand we expect that existence of the long memory will make recognition of causality more difficult, but on the other hand we would like to see that causality can still be reasonably detected. Figure 8.6 shows how the power of the test is affected by increasing the long memory (values of the parameter $d = 0.1$ vs $d = 0.45$), and how this effect can be increased by changing other parameters (the degree of moving average from MA(1) to MA(4), noise covariance from $\sigma^2 = 0.1$ to $\sigma^2 = 10$, strength of linear causality from $b_Y = 0.7$ to $b_Y = 0.2$). It's worth emphasizing that decreasing strength of causality has the biggest influence, and is the only factor that affects the power of the test for long time series (length = 1000).

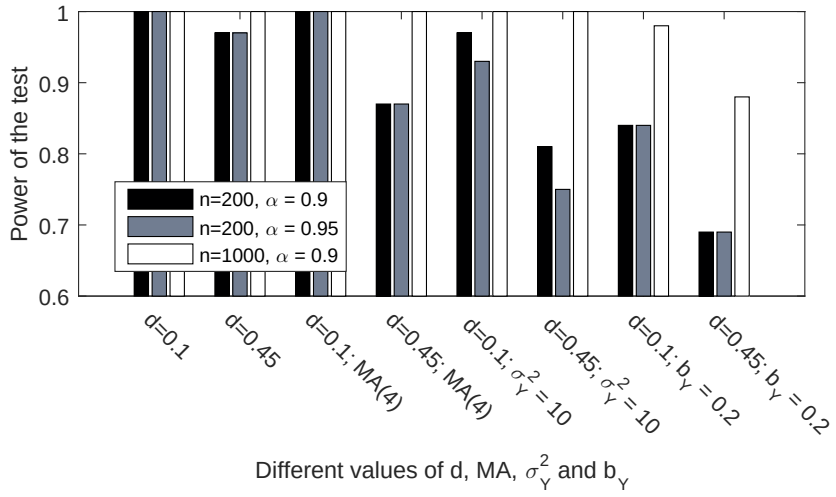


Figure 8.6: The effect of longer memory on the power of the test in data 3 varies strongly with different parameters.

8.1.2 Compound Tests

Summary of the section: *Compound tests are two stage tests where both the likelihood as well as the model parameters are estimated. Robust estimation of parameters while possibly costly, is one of the most important pillars of robust testing with compound tests. In this section we want to draw attention of the reader to a few important phenomena: firstly, that the framework is much better in picking up causality than accepting the lack of causality; and secondly, that even with strong model misspecification - which we will see for the data 3, it is possible to identify causality.*

One of the biggest factors influencing quality of the compound test is the efficiency of the optimisation algorithm. Likelihood is, in general case, not a convex optimisation problem, which means that existence of local optima is likely. Using multiple starting points is highly recommended, but can potentially make the calculations very time consuming (our implementation involves a random grid of starting points). Using Gaussian Processes with the assumptions we have in this research (mainly: additive Gaussian noise) offers the advantage of being able to calculate the likelihood analytically. However it is still possible that the data set can be so large, that this calculation would be prohibitive. A popular approach in the literature is to decrease the dimensionality of the input data (Snelson and Ghahramani [2007]), or strive for efficient implementation (Williams and Rasmussen [2006]). An interesting an little known approach is to choose covariance function that promotes sparsity of the covariance matrix (mel [2009]). Ensuring an approach is suitable to time series potentially adds a level of complication.

Example Time Series Model Structure 1: An observation that holds for arguably all data – not only the Model Structure 1, is that when causality does exist in the data, the distribution of the test statistics estimator is much narrower than when there is no causality. An example is shown in the Figure 8.7: the first plot shows that the causal signal can be picked up even for the shortest data, and the distribution of

the tests statistics converges to value 1 already for length 100. When causality is not present (subplots 2 to 4) even for the longest used samples the distributions of test statistics are wide with median at zero, but 75th percentile often reaching close to 1.

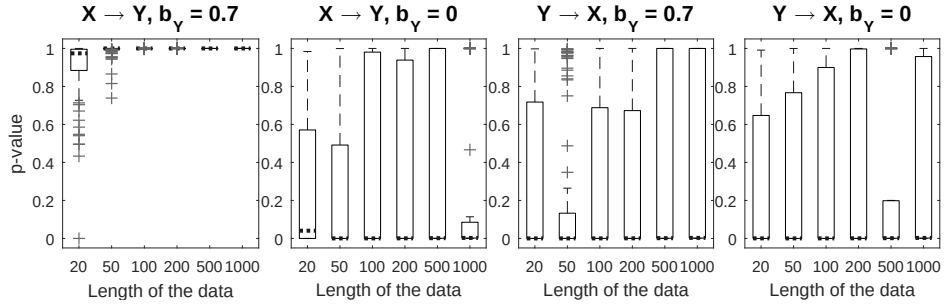


Figure 8.7: Boxplots showing how the sample size affects distributions of the test statistics, in the case of existing causal effect (first subplot $X \to Y$ and $b_Y = 0.7$) and in the case where causal effect disappears due to causal coefficient equal to zero (second subplot $X \to Y$ and $b_Y = 0$), construction (third subplot $Y \to X$) or both (fourth subplot).

Example Time Series Model Structure 2: The results for Example model data 2 shows some very interesting behaviours. When fitting the model, we introduced some model misspecification, because we allowed the structures to be the same for both directions. The first misspecification is in using polynomial means of second degree for $Y \to Z \mid X$ as well as $Z \to Y \mid X$. The second misspecification is in using the same volatility structure for both $X \to Y \mid \mathbb{Z}$ and $Y \to X \mid Z$. As a result the estimated parameters in mean are often correctly estimated to be near zero, but the parameters in variance are strongly misspecified. The results still have reasonable power of the test: the existence of causality is always correctly picked, however in same cases we have results which could be interpreted as spurious causality. Also, like with data 1, there are cases where we seem to be spotting the causal effect in the covariance function when there's no causality in the mean, shown in the Figure 8.8.

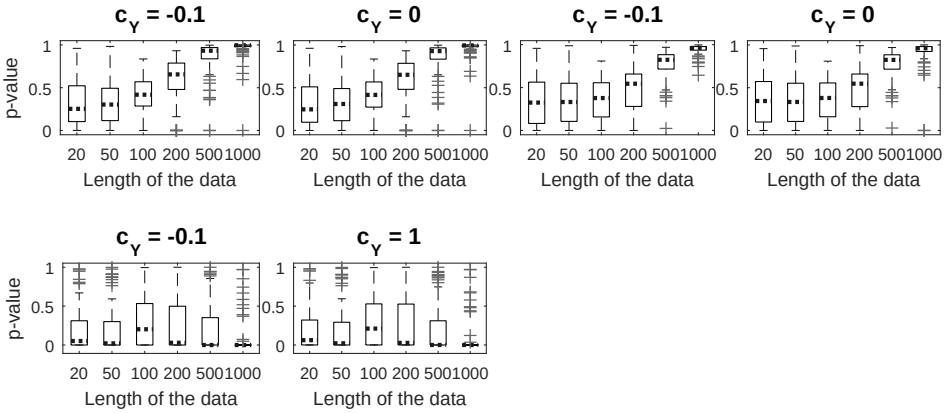


Figure 8.8: Data 2, $X \to Y \mid Z$. Changes in recognition of causality in covariance with increases sample size: different parameter settings. The top row shows the parameter settings where causal effect in covariance can be expected ($c_Y \neq 0$), while the bottom row shows cases where causality in covariance is not expected ($c_Y = 0$). In those cases there was no causality in the mean ($b_Y = 0$).

At the same time we see that we detect some spurious causality signals for the opposite direction: $Y \rightarrow X \mid Z$. Figure 8.9 show how in the presence of causality $X \rightarrow Y \mid Z$ ($b_Y = 0.7$), the opposite direction also starts displaying causality with growing sample size. Explaining spurious causality is often complicated. In this case we want to emphasize the following observations. First of all, the value of the test statistics is much bigger for the side where true causality exist, and much smaller sample is needed to start indicating that causality with a high confidence. Secondly, we run misspecified model for the $Y \rightarrow X \mid Z$ direction (the misspecification is in the covariance function, with the multiplicative parameter σ_f having to equal zero to achieve properly specified function consisting of the multiplicative noise only), and even with multiple starting points, the optimised parameters are not as close to the true parameters as we would like.

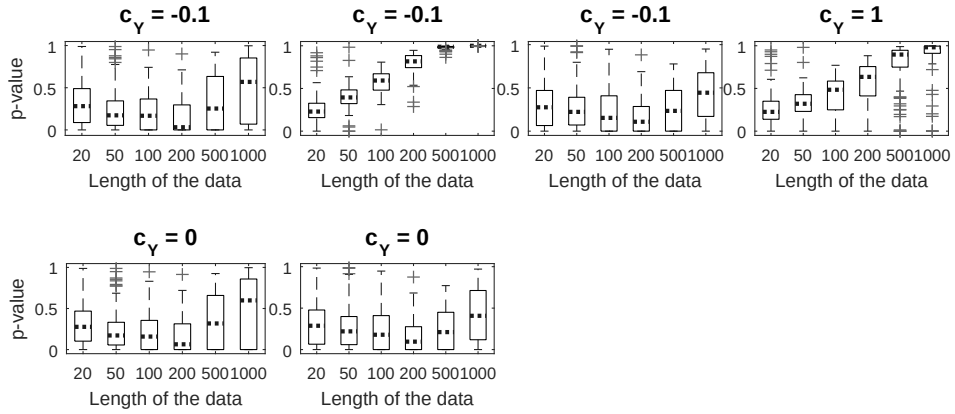


Figure 8.9: Data 2, $Y \rightarrow X \mid Z$. Changes in recognition of causality in covariance with increases sample size: different parameter settings. The top row shows the parameter settings where causal effect in covariance can be expected ($c_Y \neq 0$), while the bottom row shows cases where casuality in covariance is not expected ($c_Y = 0$). In those cases there was no causality in the mean ($b_Y = 0.7$).

Example Time Series Model Structure 3: The results for the third data set follow similar trend in the aspect that when a strong causal signal is present, it is correctly recognised. In case of lack of causality, or with very weak causal component, the distribution of the test statistics can be wide, but no spurious causality was detected. The data 3 set has a long memory component, controlled by the parameter $d \in [0, 0.5)$, and one of the most interesting aspects is understanding is analysing the effect of long memory.

The First of all, when in case of standard parameters, long memory hardly influences recognition of causality. Here, standard parameters are: strong causal component present ($b_Y = 0.7, b_Z = 0.7$), and the noise variance isn't substantial ($\sigma_n^2 = 0.01$).

Figure 8.10 shows the distribution of test statistics for no long memory ($d = 0$) and strong long memory $d = 0.45$) for different data lengths. The effect on the data 3 of changing parameters in particular of changing the memory parameter d is not significant. This seems unexpected at first, compared to the

results of the simple test.

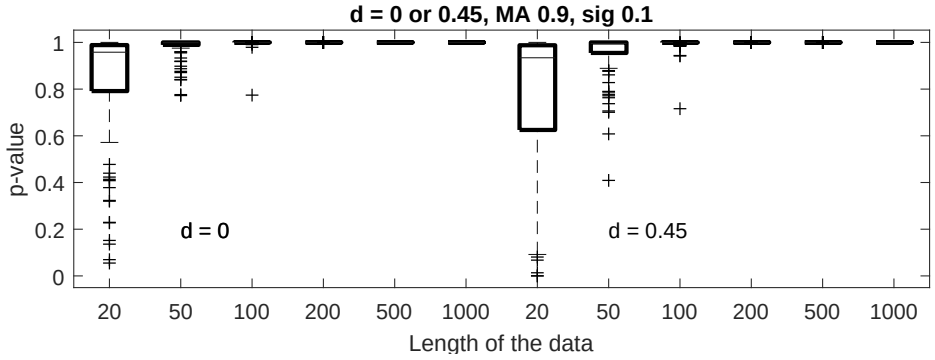


Figure 8.10: Long memory barely affects the distribution of test statistics. This figure shows the distribution for the test statistics for $X \to Y$ for increasing length of the time series, first with no long memory $d = 0$, then with strong long memory $d = 0.45$.

However the explanation lies in how the parameter estimation works, illustrated in the Figure 8.11. The model is strongly misspecified and several properties of the data are not well described by the model. But lets remember that the long memory component has an infinite sum moving average representation, and the moving average model has an autoregressive representation. So the primary effect of increasing moving average part and the long memory part is the increase of parameters responsible for autoregression.

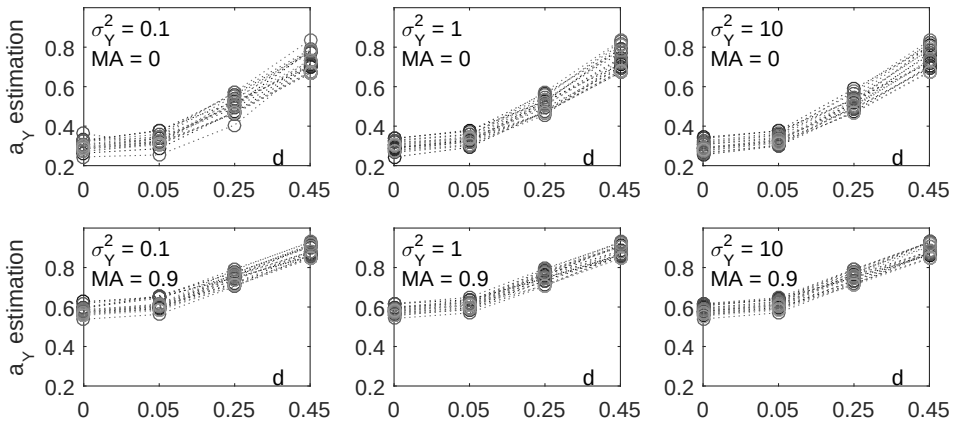


Figure 8.11: How estimation of the autoregressive a_Y parameter "compensates" long memory or moving average effects. This figure shows the estimates of $\hat{a}_Y$ for different values of d, MA, σ^2_Y and for different experiments, all of length 1000. It can be seen that the estimates strongly increase with increasing d and MA, and that this pattern appears for all values of the noise variance.

8.1.3 Warped Gaussian Process Models

This section presents results of the experiments that use the GH skew-t distribution. We establish the quality of the test by assessing the power of the test (1-rate of type II error), and how it reflects changes in parameters, sample size, and how it reacts to tail dependence. It is expected that the power of the test will increase with the increasing sample size, and that is what we observe. We also observe that existence of

heavier tails affects the ability to recognise causal effect, and that therefore using the Gaussian Process model that does not take tails to account makes it impossible to detect causality.

In the Figure 8.12 we show the effect of increasing length of the time series, and how it is affected by changing skewness parameter. As expected, the power of the test increases whenever the length of the time series increases, regardless of other parameters. Skewness parameter has a very pronounced effect: when the skewness parameter moves away from zero, it translates into widening of the test statistic distribution and the need to have larger samples, with the direction of skewness not making much difference. Figure 8.12 shows that only the skewness parameter corresponding to the Y time series makes a difference, because the parameter γ_1 does not have any effect. This is in line with the property of GH distributions discussed in 3.3.3, stating that if $X \sim GH(\lambda, \chi, \psi, \mu, \Sigma, \gamma)$, then the marginal distribution of X_i is: $X_i \sim GH(\lambda, \chi, \psi, \mu_i, \Sigma_i i, \gamma_i)$ with $\gamma = [\gamma_1, \gamma_2]^T$ being the skewness parameter [McNeil et al., 2015].

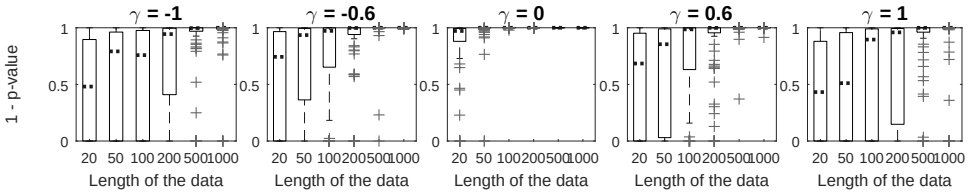


Figure 8.12: Boxplots showing how the sample size affects distributions of the test statistics for the GH skew-t distribution, for different skewness parameters for the $X \to Y$ direction. The other parameters are: strength of causality $b_Y = 0.7$, parameter of autoregression $a_Y = 0.3$, kernel parameters: $l_a = l_b = e^{-3}, \sigma_f^2 = e^{-10}$, degrees of freedom $\nu = 5$.

In the Figure 8.13 we analyse the effect of changing the degrees of freedom parameter. For time series of length 100 we change both the skewness parameter $\gamma_2 \in \{-1, -0.6, 0, 0.6, 1\}$ and the shape parameter $\nu \in \{1, 3, 10, 20\}$. Both parameters have a considerable effect. It is worth noting, however, that the covariance for the GH skew-t distribution exists only for $\nu > 4$.

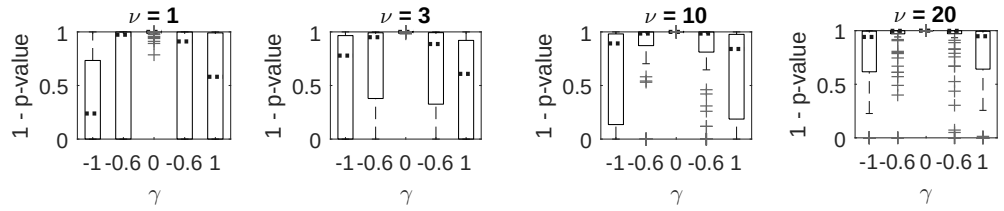


Figure 8.13: Boxplots showing how the sample size affects distributions of the test statistics for the GH skew-t distribution, for skewness parameters $\gamma_2 \in \{-1, -0.6, 0, 0.6, 1\}$, and for different $\nu \in \{1, 3, 10, 20\}$ parameters. The other parameters are: strength of causality $b_Y = 0.7$, parameter of autoregression $a_Y = 0.3$, kernel parameters: $l_a = l_b = e^{-3}, \sigma_f^2 = e^{-10}$. The length of the time series is 100 for each of the samples.

To better illustrate the effect of the parameters ν, γ on the tail dependence, the Figure 8.14 shows numerically estimated upper and lower tails of the GH skew-t distribution (for the estimation method, see Ames et al [Ames et al., 2015]). The results that we are numerically obtaining vary from the theoretical in

that a considerable lower and upper tail dependence is observed not only when both skewness parameters are negative, but also when they are both positive. However, the numerical estimation of the tail dependence coefficient is, in practice, estimating less extreme dependence.

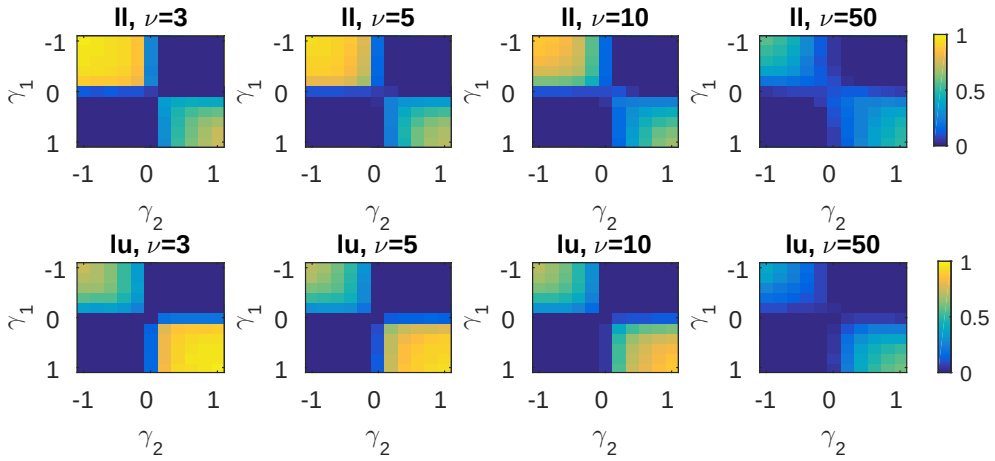


Figure 8.14: Numerically estimated upper and lower tail with a GH skew-t distribution. Causality exists in the $X \to Y$ direction. The parameters are: strength of causality $b_Y = 0.7$, parameter of autoregression $a_Y = 0.3$, kernel parameters: $l_a = l_b = e^{-3}, \sigma_f^2 = e^{-10}$, degrees of freedom $\nu = 3, 5, 10, 50$, the skewness parameter is $\rho = \{-1, -0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8, 1\}$.

8.1.4 Experiments with the "Alternative" Skew-t Distribution

In the case of the alternative skew-t distribution, both the skewness parameter and the degrees of freedom parameters will have different effect on the underlying Gaussian Process, as this enters the transformation after its marginals were standardised. This is clearly visible when we use skewness parameters similar to the ones from the previous example, the effect of the gamma parameter is not what we would intuitively expect – see Figure 8.15. The distribution for zero γ is actually marginally worse than for other non-zero skewness parameters.

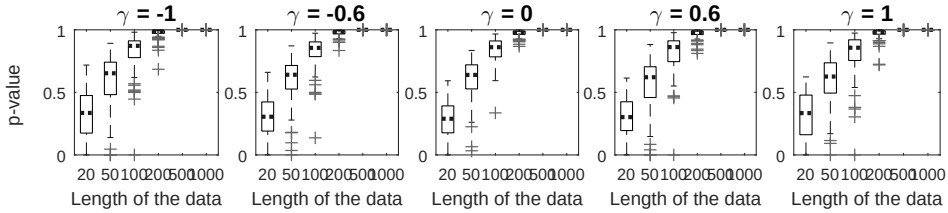


Figure 8.15: Boxplots showing how the sample size affects distributions of the test statistics, for different skewness parameters for the $X \to Y$ direction, for the alternative skew-t distribution. The other parameters are: strength of causality $b_Y = 0.7$, parameter of autoregression $a_Y = 0.3$, kernel parameters: $l_a = l_b = e^{-3}, \sigma_f^2 = e^{-10}$, degrees of freedom $\nu = 5$.

Dividing by the standard deviations has a complex effect. Most importantly, it scales the variable compared to the mean vector, and in the case of a Matern kernel, the scaling is inversely proportional to the standard deviation of the noise. We know that (symmetrical) dependence of copula is invariant

under affine transformation, but the ability to detect causality does seem to be affected. Unlike in the GH skew-t distribution, in the alternative skew-t distribution the skewness parameter is not the sole lever for changing the amount of skewness in the model. Figure 8.16 shows an effect of choosing big skewness parameters: $-10, -3, 3, 10$. We can see now the expected effect of skewness: an increase in skewness parameter worsening the test statistics.

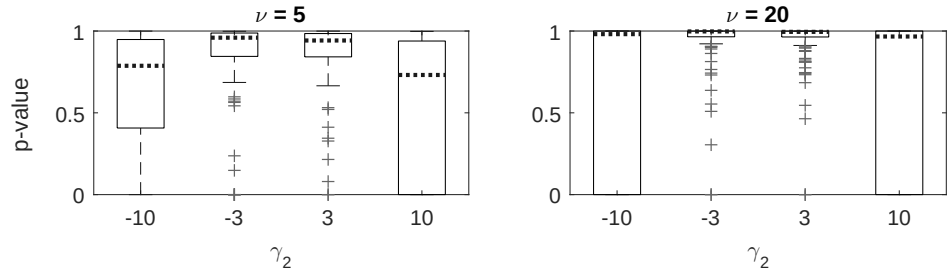


Figure 8.16: Boxplots showing how the sample size affects distributions of the test statistics, for different skewness parameters for the $X \to Y$ direction, for the alternative skew-t distribution. The other parameters are: strength of causality $b_Y = 0.7$, parameter of autoregression $a_Y = 0.3$, kernel parameters: $l_a = l_b = e^{-3}, \sigma_f^2 = e^{-10}$. Data length equals 200.

We also note that γ_1 could participate in the dynamics of the variable $\tilde{Y}_t$, if the two variables X_t, Y_t were correlated, but in our models we were using the assumption of no correlation, so again γ_1 does not influence the test statistics for the $X \to Y$.

Compared to the GH skew-t distribution, the alternative skew-t shows much lower tail dependence with comparable range of γ parameter - which is in line with the previous observations. We recall here that one of the arguments for studying the "alternative" skew-t model was the fact that the theoretical tail dependence coefficient for the skew-t distribution of the GH type has trivial values. But analysing the empirical tail dependence coefficients we see, that this might not be a problem in practical applications. Ultimately, it is the role of the researcher/analyst to decide which features of the data are most important to be reflected by the model.

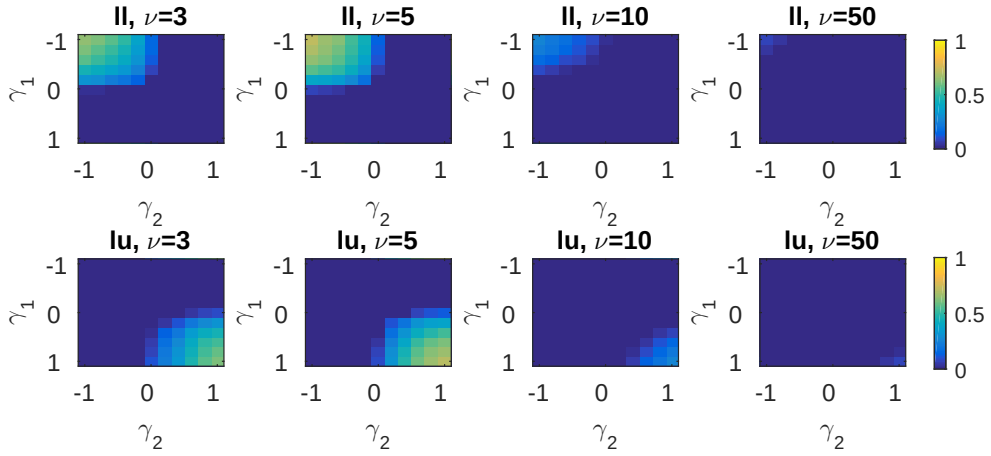


Figure 8.17: Numerically estimated upper and lower tail with an alternative skew-t distribution. Causality exists in the $X \rightarrow Y$ direction. The parameters are: strength of causality $b_Y = 0.7$, parameter of autoregression $a_Y = 0.3$, kernel parameters: $l_a = l_b = e^{-3}, \sigma_f^2 = e^{-10}$, degrees of freedom $\nu = 3, 5, 10, 50$, the skewness parameter is $\rho = \{-1, -0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8, 1\}$.

When comparing the empirical and theoretical lower tails for the alternative skew-t distribution (Figures 8.17 and 8.18) we see that the the pattern is similar to the one obtained for GH skew-t distribution, although the theoretical has higher values.

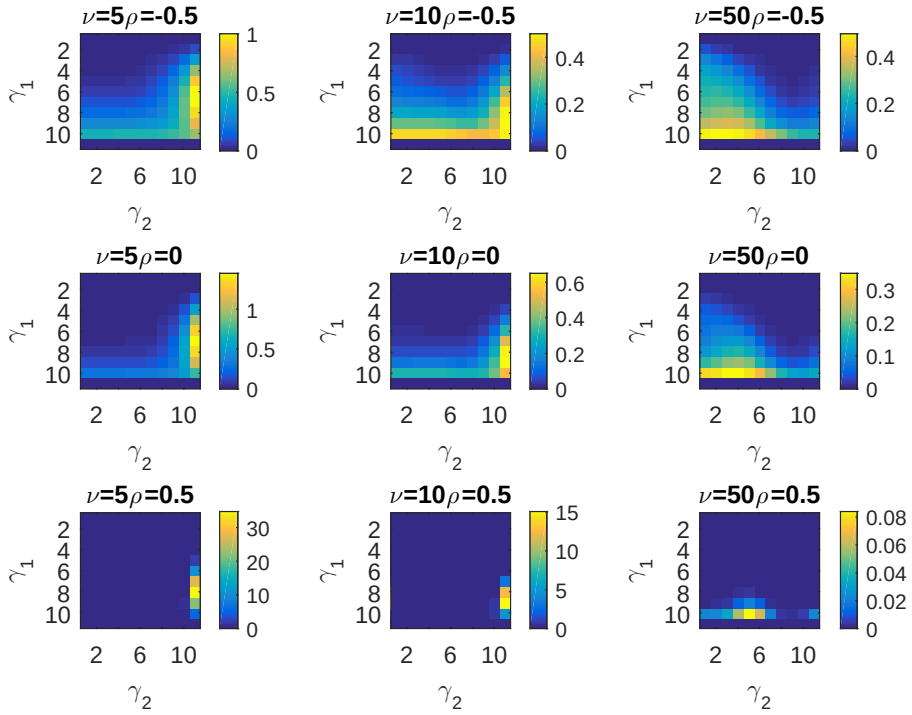


Figure 8.18: Theoretical lower tail for an alternative skew-t distribution. Degrees of freedom $\nu = 5, 10, 50$, correlation $\rho = -0.5, 0, 0.5$, the skewness parameter is $\rho = \{-1, -0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8, 1\}$ Results for $\nu = 100$ omitted from the figure, as they were all zero.

8.2 Comparison to Other Models

This section is provided to substantiate some of the claims we make about how our methods compare to existing methods. We provide three case studies, two of them compare our method to benchmark methods for causality: Granger causality and transfer entropy. The third case study compares our method to using generalised likelihood ratio test on a well specified econometric model (ARFIMA, example time series model class 3, Equations: 3.55 - 3.57). What we show in our experiments is that our model has good results for all types of data, but in all cases, except for applying linear Granger causality test to linear causality, our method has superior asymptotic properties, as it reaches good power of the test for small samples.

Please note that in these case studies we concentrate on the ability to detect causality, and not on the speed of the algorithm.

Case Study 1: Granger Causality Granger causality can be seen as the original, but also the simplest method of assessing causality. For Gaussian noise and linear causal relationship, Granger causality is arguably the best method, given that the test statistics have known asymptotic distributions, and estimators have excellent numerical properties. What is more, Granger causality can perform well for a range of data that departs from the model assumptions.

In this, and in the next case study, we will use four data sets, designed to show the effect of the departure from the assumption of data with linear dependence, stationary distributions, and Gaussian noise (as introduced earlier in the Equations 3.49), replicated below with slight modifications:

$$
\begin{aligned}
X_t &= a_X X_{t-1} + \epsilon_X, \\
Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_Y, \\
Z_t &= a_Z Z_{t-1} + b_Z Y_{t-1}^2 + \epsilon_Z, \qquad \epsilon_X, \epsilon_Y, \epsilon_Z \sim i.i.d \text{ white noise},
\end{aligned}
\tag{8.1}
$$

The data model from Equations 8.1 exhibits two causal relationships. The causal relationship $X \to Y$ is – if we assume Gaussian white noise – of the type that Granger causality has been designed to model: linear, stationary, with Gaussian distributions. We will call this a base case (set one), and we will consider three other cases, each presenting a departure from one of those three properties. The causal relationship $Y \to Z$ is not linear, and it forms the set 2. We will also consider what happens to the ability to detect relationship $X \to Y$, if we changed Gaussian noise to t-student noise (set 3), and if we changed stationary to non-stationary marginal distributions (set 4); in this case we use polynomial covariance, please refer to the Table 3.1). These four set and their properties are summarised in the Table 8.1.

set nr.	1	2	3	4
direction	$X \to Y$	$Y \to Z$	$X \to Y$	$X \to Y$
linearity	linear	non-linear (square)	linear	linear
noise	Gaussian	Gaussian	t-student, 5 df	Gaussian
stationarity	stationary	stationary	stationary	non-stationary

Table 8.1: Data used for Case Study 2 and 3. Causal relationship number 1 is the base case: linear, with stationary marginal distributions and Gaussian noise. The three other causal relationships show three types of departure from the base case.

We present the results for the Granger causality method, using the GCCA toolbox. The test statistic used in the toolbox is the one that has been introduced by [Geweke, 1982], as in the Equation 1.8. The corresponding test used for testing the null hypothesis of lack of causality is the F-test. The results are presented graphically in the Figures 8.19 - 8.20. The results for using Grangercausality can be

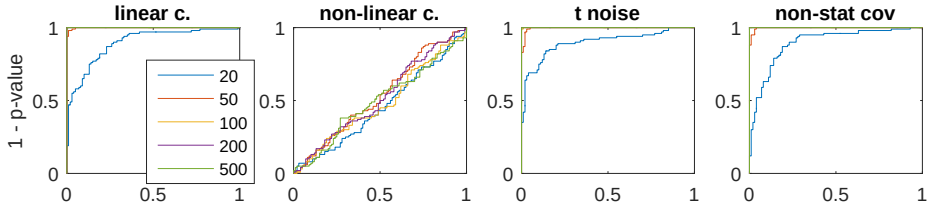


Figure 8.19: ROC curves for the data sets 1-4 from the table, calculated with (linear) Granger causality, tested with the GCCA toolbox.

summarised by two main observation. Firstly, for strong linear causality relationship, using the linear Granger causality test is very robust and practical even if we do not observe Gaussian noise or stationary covariance. Secondly, for non-linear causality, the linear Grange causality method behaves no better than a random guess, regardless of the data size. How does that compare to our method? The figure 8.20 shows that for strong, linear causality, our method is not as robust as linear Granger causality, and requires a bigger sample. However, our method can successfully detect non-linear causality. For the data with t-distributed noise, we present results for the test statistic calculated by assuming the correctly specified model, and using an approximate method. Assuming a misspecified model with Gaussian likelihood, and then using the exact method to optimise parameters bring comparable results in this case.

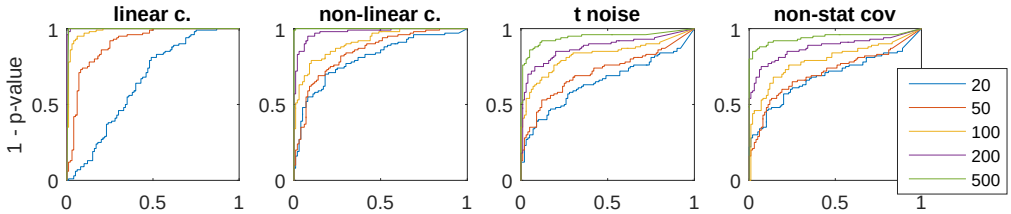


Figure 8.20: ROC curves for the data sets 1-4 from the table, tested with our method.

Case Study 2: Transfer Entropy This is a continuation of the Case Study number 1. We have used the same data structures as described in the Equations 8.1 together with the Table 8.1. The results are graphically shown in the Figure 8.21.

Transfer entropy is a popular method used as a non-linear extension of the linear Granger causality (for Gaussian distributions these two methods are equivalent). It is able to consider wider range of data types and relationships, however it is much more difficult to estimate. Compared to our method, transfer entropy requires much larger data samples, and at the same time it is not be able to deal with model structures like long memory, non-stationarity, etc. Comparing Figures 8.19 – 8.21 shows inferior performance of transfer entropy to our method in each of the four cases, and inferior to (linear) Granger causality in three cases. Transfer entropy is better than Granger causality in recognising non-linear causality, however only for the sample of size 500 is transfer entropy performing recognisably better than a random choice.

What is not shown in the results, but for the sake of fairness needs to be mentioned, is the fact that transfer entropy is much faster than our method, with the current implementation.

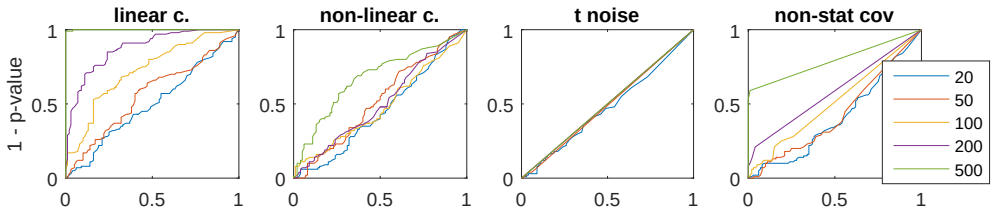


Figure 8.21: ROC curves for the data sets 1-4 from the table, calculated with transfer entropy based on the binning algorithm.

Case Study 3: ARFIMA model The data that was used for this example has been generated according to an ARFIMA (1,d,1) model with external regressors, Equations 3.55 - 3.57, can be represented in a form emphasising the autoregressive part (this is possible because we restricted the choice of d to $(0, 0.5)$):

$$
\begin{aligned}
X_t &= a_X X_{t-1} + \epsilon_X \\
Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon^*_{y,t}, & \epsilon^*_{Y,t} &= (1-B)^{-d}\Theta_Y(B)\epsilon_{Y,t} \\
Z_t &= a_Z Z_{t-1} + b_Z Y^q_{t-1} + \epsilon^*_{z,t}, & \epsilon^*_{Z,t} &= (1-B)^{-d}\Theta_Z(B)\epsilon_{Z,t}.
\end{aligned} \tag{8.2}
$$

We estimate data using modified Matlab code ARFIMA-SIM by Fatichi [2009]. For fitting the ARFIMA with external regressors we use the rugarch R library. We present results for nine parameter settings, which are listed in the Table 8.2 .

Set nr	a_Y	b_Y	MA	d	
1	0	0	0	0	pure noise
2	0.3	0	0	0	ARFIMA(1,0,0)
3	0.3	0.7	0	0	ARFIMA(1,0,0) and causality
4	0	0	0.9	0	ARFIMA(0,0,1)
5	0	0	0	0.49	ARFIMA(0,d,0)
6	0.3	0.7	0	0.25	ARFIMA(1,d,0) and causality
7	0.3	0.7	0.9	0	ARFIMA(1,0,1) and causality
8	0.3	0.7	0.9	0.25	ARFIMA(1,d,1) and causality
9	0.3	0.7	0.9	0.49	ARFIMA(1,d,1) and causality

Table 8.2: Nine sets of parameters for the ARFIMA model, that were used in our analysis, in the Case Study 3.

We present the results of using our causality method to estimate causality in Figure 8.22, while the results of using a fully specified likelihood of the ARFIMA model is shown in the Figure 8.23.

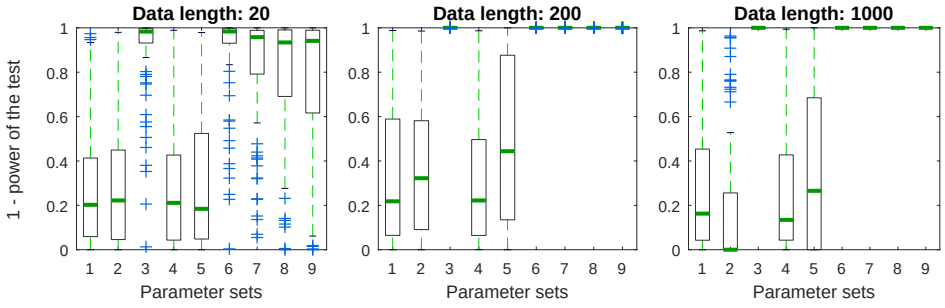


Figure 8.22: Distributions of test statistic for GPC method, shown for three lengths of the time series, and for 9 data sets.

Our method is operating on the GP model representation, which is clearly misspecified. However, that does not prevent our model from detecting causality even for the smallest samples of length 20. That is not the case for using the well specified ARFIMA model and estimated likelihood – in this case a very large sample is needed for the estimation to even be successful – data of length 1000 is needed to be able to present results for all 9 data sets.

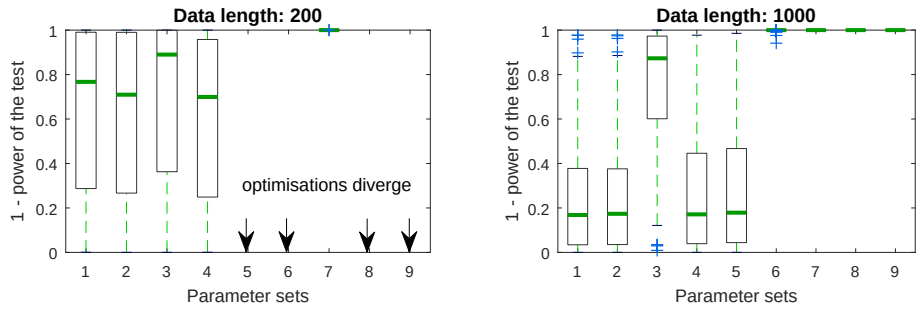


Figure 8.23: Distributions of test statistic for the AFRIMA likelihood method, shown for two lengths of the time series, and for 9 data sets.

8.3 Warped Gaussian Process Models - Causality vs Tail Dependence

8.3.1 Tail ordered warping

In this paragraph we present the finite sample effect of warping Gaussian processes when applying the asymptotic distribution of the test-statistic to make decisions on existence of causality. If we use the median power of the test for the finite sample as an estimator of the power of the test, we can observe (Figure 8.24) that it is a decreasing function of the tail ordering for all cases, except of the shortest time series presented. As we will also observe later, for the shortest time series, the number of observation might be too short for the heavy tail to be noticeable. We do not show it in the Figure, but the effect on the power of the tail is further amplified by skewness.

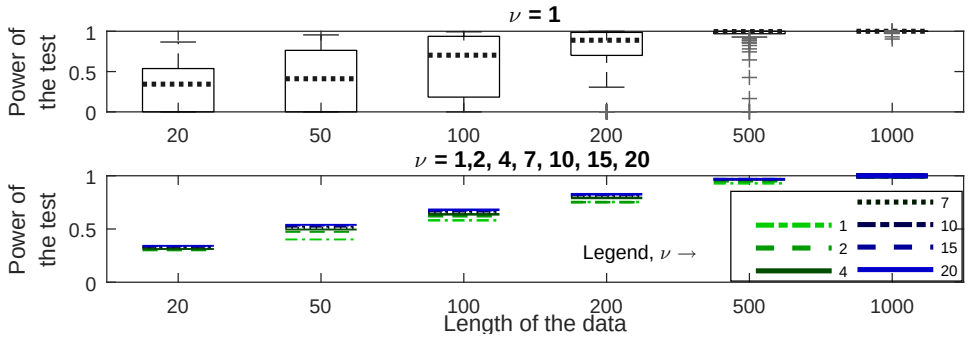


Figure 8.24: Effect of the shape parameter ν on the recognition of causality in the symmetric case. The upper figure shows the effect on distribution of the power of the test with increasing length of the time series for $\nu = 1$, the lower chart shows only medians, but for different tails $\nu = 1, 2, 4, 7, 10, 15, 20$.

8.3.2 Sensitivity to misspecification in the mean

We present the effect on performance of incorrectly specifying the mean function for the direction $Y \to X$, by adding the causal term: $b_X Y_{t-1}$ to the mean $\mu_{X,t}$ (compare: Equation 3.51). We set $b_X \in \{-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3\}$. We present results for skewed $\{Y_t\}$ and symmetrical $\{X_t\}$, by setting $\gamma = [0, -1]$, for tails close to normal with $\nu = 20$, or heavier than normal, with $\nu = 5$, and for data lengths

of 20 and 100. We conclude that the effect of misspecification in the mean is amplified by skewness (the effect is different for γ_X and γ_Y) and to a smaller degree by heavy tails, nevertheless, it starts to disappear for time series of length 200 and longer (not shown). The decrease in performance resulting from heavier tails is bigger with longer time series, which is to be expected, as for short time series rare events might not be observed. For short time series (Figure 8.25) the effect of skewing the $\{Y_t\}$ time series is more prominent – the performance decrease nearly monotonically with increasing absolute deviation in the b_X, but for longer time series (Figure 8.26) that effect is offset by lack of skewness in $\{X_t\}$. This agrees with the property that for exact parameters γ_X does not influence $L_{\tilde{X}\to\tilde{Y}|Z}$ (Sections 3.3.4.2 and 3.3.5).

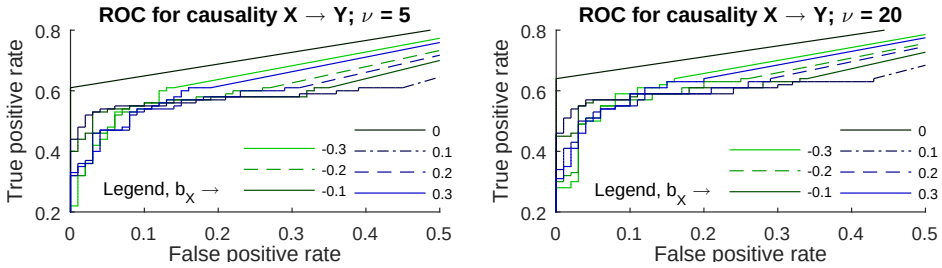


Figure 8.25: Sensitivity to misspecification in $\mu_{X,t}$. Data length = 20, Skewness = $[0, -1]$.

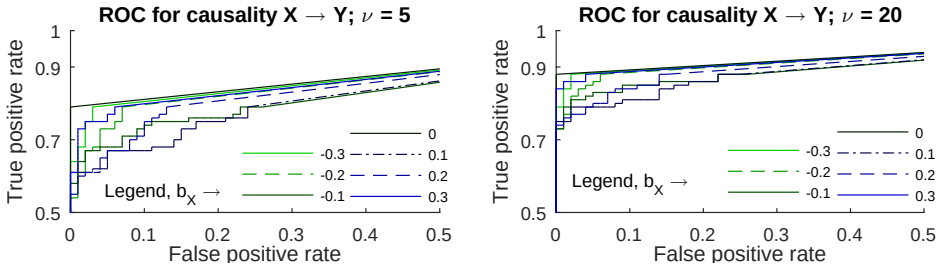


Figure 8.26: Sensitivity to misspecification in $\mu_{X,t}$. Data length = 100, Skewness = $[0, -1]$.

8.3.3 Tail dependence decreases power of the test

The effect of the changing skewness parameter γ_Y on the recognition of causality, as measured by the power of the test, depends on the absolute value of γ_Y and diminishes with the size of the sample. As mentioned, the power of the test does not depend on the value of γ_X $L_{\tilde{X}\to\tilde{Y}|Z}$ (Section 3.3.4.2). is shown in the Figure 8.24. The larger the skewness (regardless of the sign), the lower average power of the test, and wider range of values the power of the test take. This effect diminishes with the size of the sample.

The effect of the tail dependence is more complicated. While the expectation is that the larger the skewness, the less likely we should be to detect causality, that is not always the case, because unlike causality, dependence coefficient is affected by γ_X. In the Figure 8.27, we present the effect of numerically estimated λ_u (estimation from [Ames et al., 2015]) on the power of the test, with $\gamma_X = 0$. The distribution of the power of the test is shown as corresponding to the tail dependence coefficient binned in the intervals $[0, 0.1], \cdots, (0.9, 1]$. The plots present results for data lengths 50, 100, 200, 1000 and for a) 100

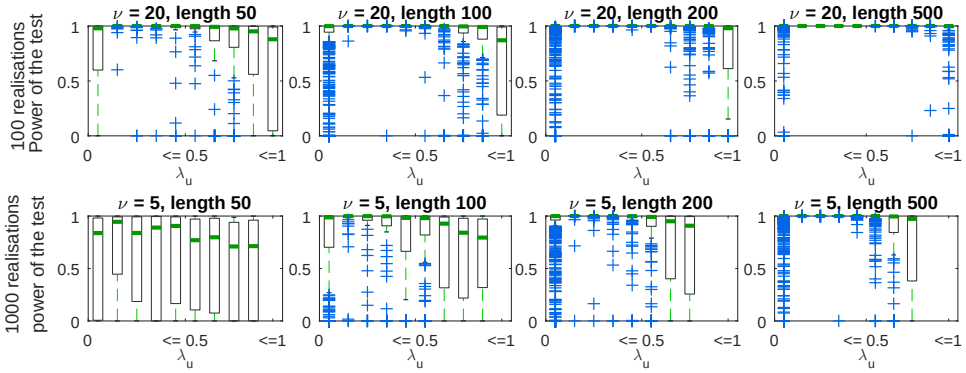


Figure 8.27: The effect of λ_u on the power of the test, numerically estimated λ_u is binned in intervals $[0, 0.1], (0.1, 0.2], \cdots, (0.9, 1]$. Analytical value of the λ_u is in the range $[0, 0.1)$.

realisations, $\nu = 20$, b) 1000 realisations, $\nu = 5$. Figure 8.27 shows that the increasing tail dependence indeed makes the causality more difficult to detect. The numerical estimation of λ_u converges very slowly, but for for 1000 realisations is closer to the analytical value, calculated according to the Theorem 13, which belongs to $[0, 0.1)$.

Chapter 9

Real Data

> *Notice that this approach again changes the meaning of "solve." First that word meant "find a formula." Then its meaning changed to "find approximate numbers." Finally, it has in effect become "tell me what the solutions look like."*
>
> Ian Stewart, *Nature's Number: The Unreal Reality of Mathematics*

This chapter demonstrates how our framework for modelling statistical causality can be applied in practice to help to understand the characteristics of real data. We benefit from the flexibility of the framework that allows to test for statistical causality under different model assumptions, as it aid with the data exploration. We show the effect of inclusion of a range of statistical properties on recognition of causality, and we provide interpretation of the results.

We apply our framework to commodity futures data, and show how understanding of causal relations can be a part of to analysing risk factors that investors should consider when building a portfolio of oil futures, currencies and physicals.

9.1 Commodity Futures Data

In this section we apply the testing procedures to analyze commodity futures data.

In our analysis we use the following data: 1 and 36 month expiry oil futures contracts, obtained from futures curves built on the basis of West Texas Intermediate (WTI) Crude oil futures prices traded on the New York Mercantile Exchange, as described by Ames et al. [2016]. The affect of the currency level, captured by the US Dollar Index DXY, is constructed as index of USD relative to EUR, JPY, GBP, CAD, SEK, CHF. Thirdly, we also use a widely considered proxy for convenience yield based on a component related to transportation expense, given by the cost of freighting and short term storage, measured by the Baltic Dry Index (BDI), see Ames et al. [2016]. There is a stochastic functional relationship between commodity futures contracts of different maturities (term structure) based on: spot price, convenience

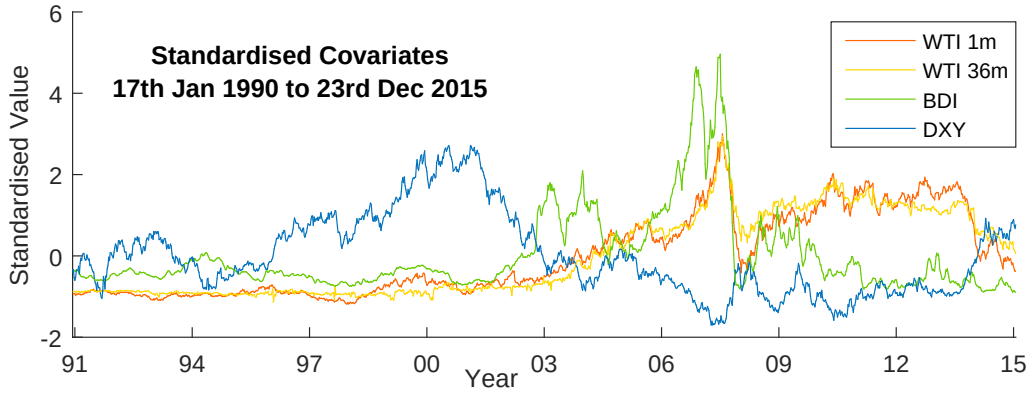


Figure 9.1: 1 and 36 month oil futures (WTI), Baltic Dry Index (BDI), Dollar index (DXY), all standardised.

yield, interest rate, and dollar value. Convenience yield is very hard to model, but can be captured to some extent by BDI, and the interest rate can be proxied by the time value of money expressed by the futures contracts. Hence the choice of both long and short dated futures contracts for our analysis. The Figure 9.1 shows the four covariates from 17th Jan 1990 to 23rd Dec 2015. For literature studying classical relationships between these data, we refer to: Ames et al. [2016], Bakshi et al. [2010] and Dempster et al. [2012].

9.1.1 Interpreting causal relationships

The study performed here uses causality testing to demonstrate the risk factors that investors should consider in their decision process. It also shows how speculators in currency markets and futures markets have a propensity to respond to information observed at different lags and the time it takes them to re-adjust the expectations for futures market hedging or speculation in light of this information.

Figures 9.2 - 9.5 present the changing significance of causal relationships between the dates 17th Jan 1990 to 23rd Dec 2015. The four pairs that we look at, and the abbreviations that we will use are as follows: 1 month oil futures (1m WTI) and freighting/ storage index (BDI), 36 months oil futures (36m WTI) and freighting/ storage index, 1 month oil futures and dollar index (DXY), 36 months oil futures and dollar index. We are presenting causal reactions at two lags: one week, which can be seen as instantaneous, and eight weeks. Figures 9.2 - 9.5 show charts smoothed with cubic spline smoothing, which makes it easier to observe the main trends, in particular in the case of lags of 8 weeks.

Markets learn from the news and facilitate them into the price, according to the efficient market hypothesis [1], to which we subscribe (Malkiel and Fama [1970], Fama and French [1988], Malkiel [1973], Campbell and Shiller [1988], Campbell et al. [1997], Malkiel [2003]). We want to learn which variables

[1] **Efficient market** can be defined as " [a market that] (...) *do not allow investors to earn above-average returns without accepting above-average risks.* (...) *Markets can be efficient in this sense even if they sometimes make errors in valuation*", Malkiel [2003]. Market efficiency, as it is understood nowadays, is the belief that new information is reflected in price quickly and accurately, but not necessarily instantaneously. See Malkiel [2003] and sources therein.

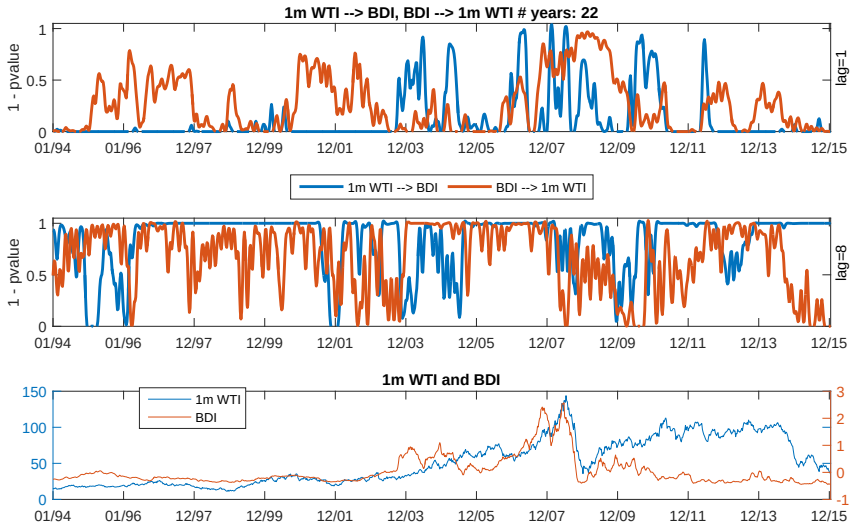


Figure 9.2: Evolution of the causal influence: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of BDI index.

have effect on price formation, and at what time horizon. We also want to relate to the fact that the three different classes of investments (oil futures, currencies, physicals) have different investor profiles, and thus we expect difference in the type and speed of reaction. The last question that interests us, is whether the results confirm the intuition that regimes affect the direction and significance of causal influence.

The interplay between WTI oil futures and the cost of freighting (BDI). Market participants investing in freighting are likely to be interested in the ownership of the physical asset, therefore BDI can be used as a proxy for convenience yield. It is expected that the WTI oil futures will not have instantaneous effect on the BDI, which is confirmed by our analysis showing that the causal direction from WTI to BDI is generally not statistically significant at 1 lag (Figures 9.2 - 9.3, top subplots).

The effect to which the WTI futures incorporate the BDI movements varies across maturities. Short contracts have not been reacting to BDI changes in 1 week, with the exception of 2008/2009, which was a reaction to crisis. Similar response can be seen for longer maturities, however for longer maturities we observe the BDI→36m WTI to be significant through late nineties.

At 8 lags, we observe that the causal effects are significant in both directions, majority of the time. This can be seen as markets being able to absorb the information and adjust the expectation. For the times when this relationship breaks, investors use other sources, to inform their long term perception of risk and expectations: for example as a result of the 2008 crisis investors across many markets were decreasing their exposure to risk. In late nineties, as well as in 2014, we can observe a divergence of reactions of BDI to short and long term oil futures at 8 week lags: this could be seen as investors using outside information to decide on their long term expectations: for example about advancement in methodology or legislation pertaining renewable energy.

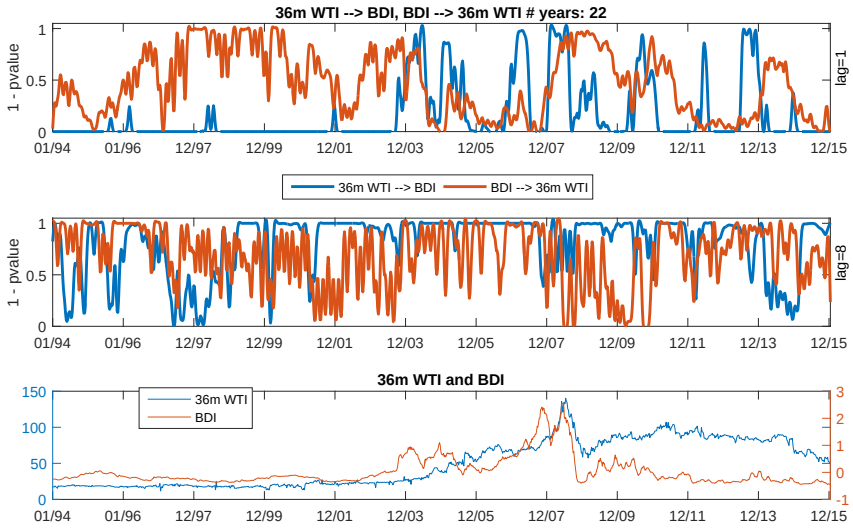


Figure 9.3: Evolution of the causal influence: 1-pvalues of the test statistic for 36 months WTI and BDI, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of BDI index.

The interplay between WTI oil futures and the dollar index (DXY). The dollar index is a weighted geometric mean of the dollar's value relative to a basket of foreign currencies: Euro (EUR) 57.6% weight, Japanese yen (JPY) 13.6% weight, Pound sterling (GBP) 11.9% weight, Canadian dollar (CAD) 9.1% weight, Swedish krona (SEK) 4.2% weight, Swiss franc (CHF) 3.6% weight. Canadian dollar is considered a commodity currency, while Japanese yen is particularly sensitive to changes in oil prices due to Japan importing almost all of its oil. Therefore market expectations towards dollar index will incorporate to a large degree the expectations that arise from the oil market.

Following the results from the Figure 9.5, there is evidence to suggest that DXY drives longer dated futures more strongly. At the same time, when comparing top charts from Figures 9.5 and 9.3, we notice similarity in causal pattern between DXY → 36m WTI and BDI → 36m WTI, in particular during the nineties. This could suggest another direct or indirect factor, common for the two causal direction, for example general attitude to risk.

We look at Markov Switching Model, to analyse if DXY and BDI will have similar patterns of states for volatility, when explained with VIX. We use the following models:

$$D_t = \alpha_{1,S_t} + \alpha_2 V_t + \epsilon_t^D \qquad \epsilon_t^D \sim \mathcal{N}(0, \sigma^2_{D,S_t}), \tag{9.1}$$

$$B_t = \beta_{1,S'_t} + \beta_2 V_t + \epsilon_t^B \qquad \epsilon_t^B \sim \mathcal{N}(0, \sigma^2_{B,S'_t}), \tag{9.2}$$

where: S_t and S'_t, which we assume to only take values 1 and 2, are the states at time t for DXY and BDI respectively, $\sigma^2_{D,S_t}, \sigma^2_{B,S'_t}$ are the variances of the innovation at state S_t, S'_t, $\alpha_{1,S_t}, \beta_{1,S'_t}$ are the mean coefficients at state S_t, S'_t, and $\epsilon_t^D, \epsilon_t^B$ are innovations.

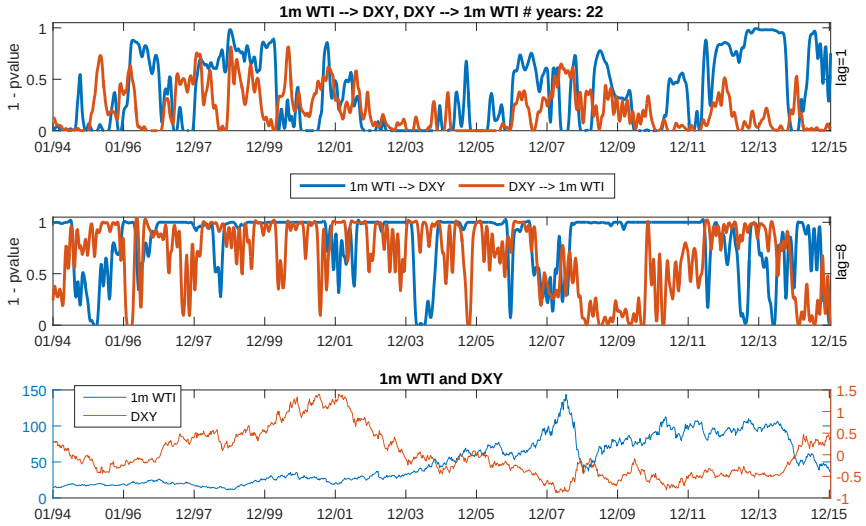


Figure 9.4: Evolution of the causal influence: 1-pvalues of the test statistic for 1 months WTI and DXY, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of dollar index.

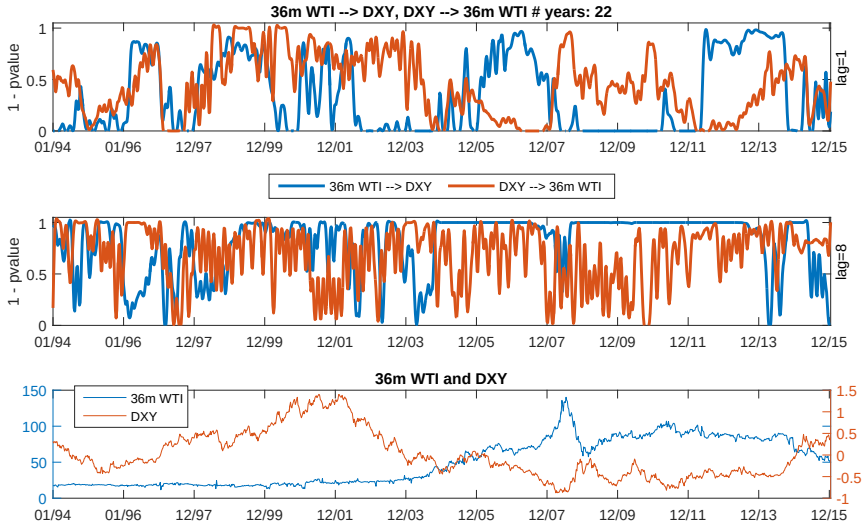


Figure 9.5: Evolution of the causal influence: 1-pvalues of the test statistic for 36 months WTI and DXY, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of dollar index.

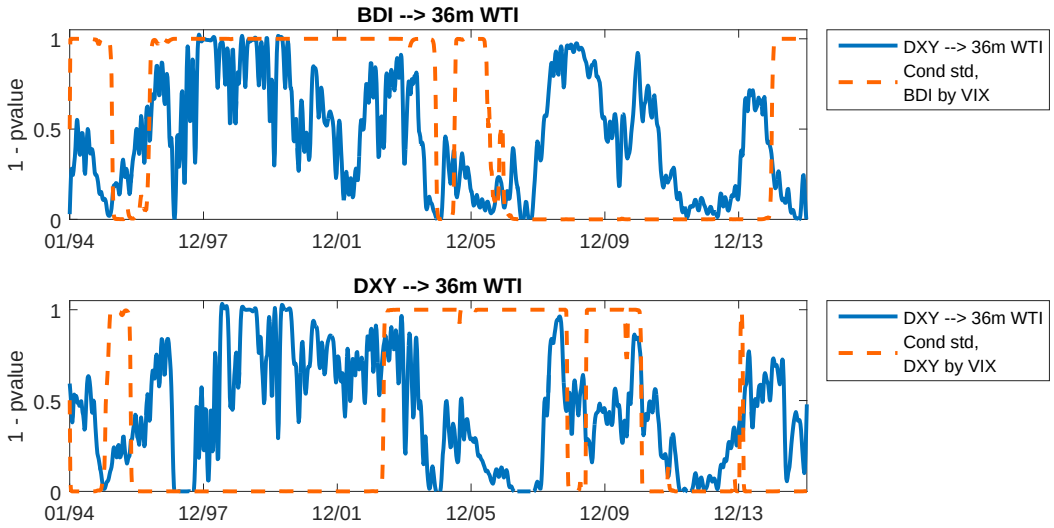


Figure 9.6: Conditional standard deviation of error of the regime switching model explaining DXY or BDI with constant and VIX, scaled to [0, 1], compared to the 1-pvalue of the BDI → 36m WTI and DXY → 36m WTI, for 1 lag.

Figure 9.6 presents the conditional standard deviation of error term for regime switching models from Equations 9.1 - 9.2, scaled for clarity to [0, 1], and superimposed on the power of the tests of BDI → 36m WTI and DXY → 36m WTI, for 1 lag. First of all, for BDI it is the decreased conditional volatility that coincides with higher evidence of causality, while for DXY it is the increased volatility. However the persistence of high evidence for causality from 1996 to 2002 for both DXY → 36m WTI and BDI → 36m WTI, coincides with the persistence of one state for conditional standard deviation of respective covariates over that period of time. This suggests that the perception of market risk as seen via VIX is a common driving factor for during the nineties, a factor which can supersede other dependencies.

9.1.2 Influence of the absolute value of the oil prices on the causal structure.

During the times when world oil prices are seen as high, it is more reasonable to expect investments in oil infrastructure as well as storage and transport. Therefore, we would expect that the absolute level of the oil price affects the behaviour (direction, strength, persistence) of causality. To test this, we compare the causal structure, as well as the fitted models, during the period of low prices: 17.01.1990 – 11.08.1999 (bellow $40), and period of high prices: 26.05.2004 – 11.03.2009 (above $90). We will be interested in the relative difference between the fitted mean values, as well as the relative difference between hyperparameters (coefficients of the mean): autoregressive and causal. For that we will be using two sample mean test. Please note, that while we are particularly interested in the change of regime in the fitted models, we also check the regime change of the causal test statistics – this is because we were earlier making a point of being able to detect causality even in misspecified models!

Lets assume that for each of the pairs: 1 month oil futures and freighting/ storage index, 36 months

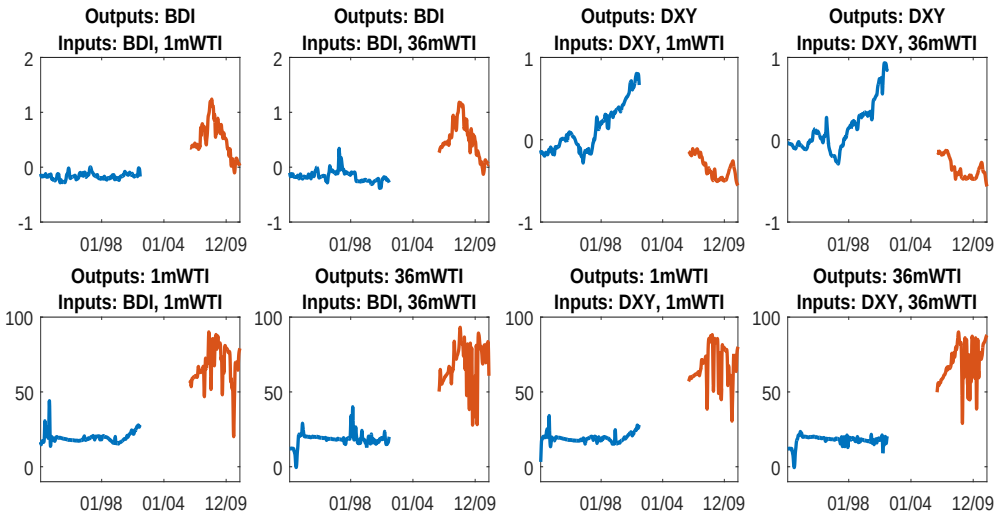


Figure 9.7: Mean function estimations for each of the pairs of time series. The two colours represent the two different segments: 17.01.1990 – 11.08.1999 (oil prices bellow $40), and period of high prices: 26.05.2004 – 11.03.2009 (oil prices above $90)

oil futures and freighting/ storage index, 1 month oil futures and dollar index, 36 months oil futures and dollar index, we take X_t to denote one of the time series from the pair, and Y_t - the other:

$$X_t = f_X([X_{t-1}, Y_{t-1}]) \qquad f_X \sim \mathcal{GP}(\mu_X, k_X)$$
$$Y_t = f_Y([X_{t-1}, Y_{t-1}]) \qquad f_Y \sim \mathcal{GP}(\mu_Y, k_Y),$$

with the usual notation. We denote M_t^X and M_t^Y as time series of values of the mean functions fitted by the models used for causality testing on rolling windows. Figure 9.7 shows the two segments of the fitted means: segment corresponding to prices below $40 and above $90. The mean function estimations are calculated on moving windows, with one mean function estimation equal to a mean of fitted values for the respective window.

For each of the pairs, we performed a two means test:

$$H_0: \quad mean(M^X_{01.90-08.99}) = mean(M^Y_{05.04-03.09})$$
$$H_1: \quad mean(M^X_{01.90-08.99}) \neq mean(M^Y_{05.04-03.09})$$

We have run the popular student-t distribution two means test, as well as a two means test using sieve bootstrap to correct for serial dependence. The results are unanimously rejecting the hypotheses of equal means.

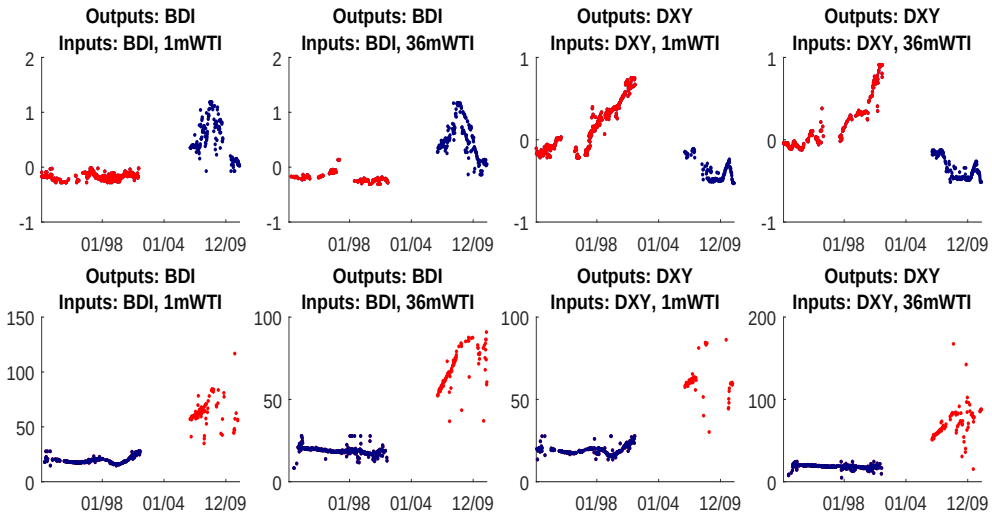


Figure 9.8: Mean function estimations for each of the pairs of time series, shown only for the time points for which the hypothesis of lack of 8 week lag causality has been rejected at the level of $\alpha = 5\%$. The two colours represent the two different segments: 17.01.1990 – 11.08.1999 (oil prices bellow $40), and period of high prices: 26.05.2004 – 11.03.2009 (oil prices above $90)

9.2 Effect of model assumptions on recognition and explanation of causality.

In this section we look at the commodity data and results presented earlier from a different perspective. We use a series of models with different model assumptions as an exploratory tool to help us understand better the various structures and statistical properties of the data. We want to understand better causal effects in the mean, covariance, and both, and how are they affected by: serial correlation, skewness, kurtosis, and tail dependence.

We use the models built in the previous section as a reference point, and we start with a comparison to a linear regression model. For the purpose of clarity and compactness we focus our attention on the causal relationships between the 1 month future contracts (1m WTI), and Baltic Dry Index (BDI). To ensure comparability, we continue using the same settings: weekly data, lags 1 and 8, window length of 104, Matern covariance with 3 degrees of freedom. The code we use for Granger causality is based on the GCCA toolbox Geweke [1982].

As we have already seen on the synthetic examples, Section 8.2, using linear regression / Granger causality has a comparably high, or higher power of the test (and ROC ratio) for data with linear causal structure, but it can perform no better than a random classifier when nonlinear causality is present. Comparing the result of testing for causality with linear regression (Figure 9.9) with our ARD-GP model framework (Figure 9.10) we see a difference in confidence for some, but not all, data windows. We conjecture, that linear regression model is overconfident due to not being able to recognise nonlinear effects, in particular to remove excess serial correlation that would subsequently invalidate the assumptions

of the hypothesis test resulting in excess kurtosis in the test statistic distribution and overly confident decision outcomes as a result.

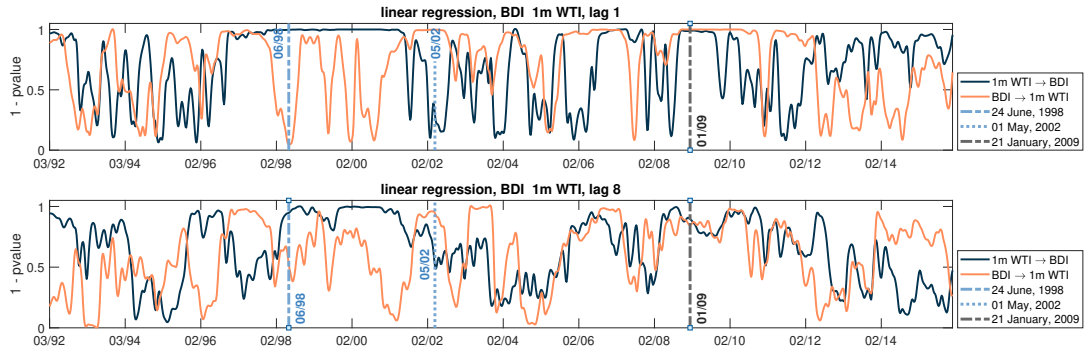


Figure 9.9: Evolution of the causal influence tested with the linear regression (GCCA toolbox): 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing.

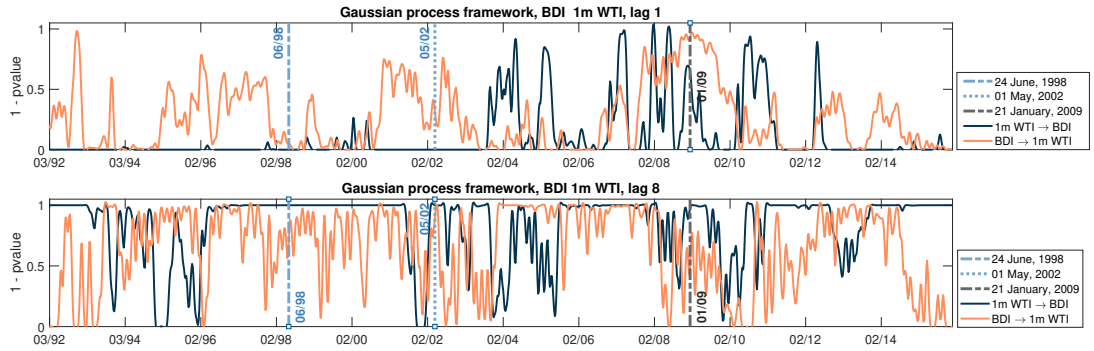


Figure 9.10: Evolution of the causal influence tested with the framework based on GPs: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing.

To argue our interpretation of the difference in the model confidence, we begin by analysing residuals of the linear regression fits. Earlier, we came to a conclusion that the lag of one week was not generally sufficient for the information to be incorporated, which is the interpretation from 1-pvalues of the GPC model rarely reaching close to 1 for either direction. But the result from the GCCA model advocate for rejection of non-causality throughout much of the data history. In Figures 9.9 and 9.10 we marked three specific point in time to show three scenarios where either one, or both of the directions show a high confidence for the linear model, that we observe with our framework. These are summarised in Table 9.1.

Table 9.1: Direction of causality implied for lag 1 by GCCA and GCP models, for three windows highlighted in Figures 9.9 and 9.10.

window	GCCA	GCP
27 Jan 1998 – 24 Jun 1998	$1mWTI \rightarrow BDI$	
30 Nov 2001 – 1 May 2002	$BDI \rightarrow 1mWTI$	
22 Aug 2008 – 21 Jan 2009	$1mWTI \rightarrow BDI$, $BDI \rightarrow 1mWTI$	$BDI \rightarrow 1mWTI$

Figure 9.11 presents a series Quantile-Quantile (QQ) plots of empirical residual quantiles versus

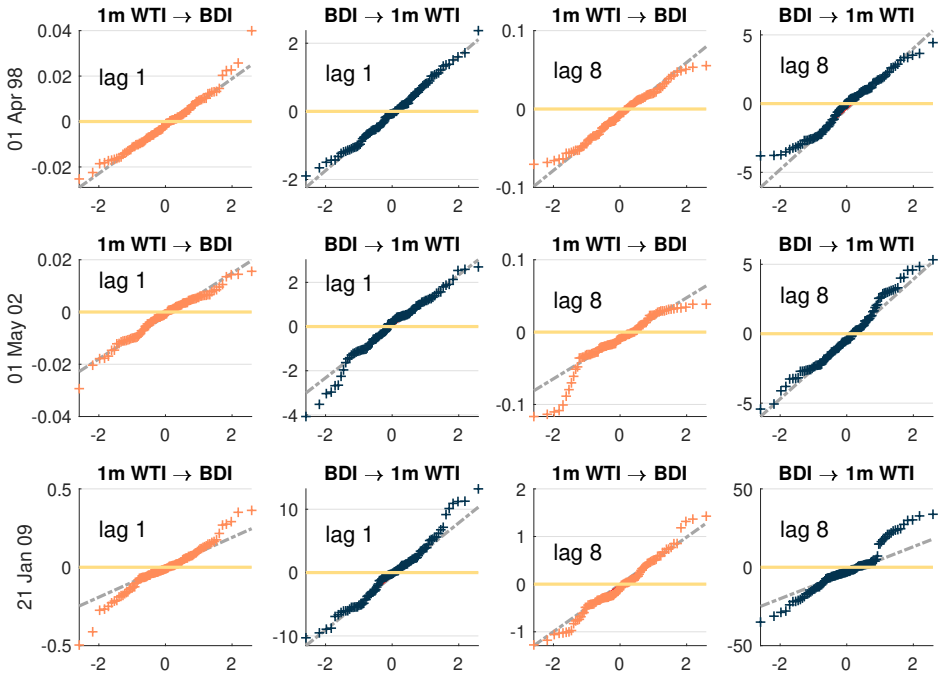


Figure 9.11: QQ plots of the residuals for the linear regression models for testing causality, for data windows ending on: data windows ending on 24th January 1998, 1st May 2002, and 21st January 2009 (rows). Each of the four columns of qq plots represent a combination of lag and direction of the causality.

normal quantiles of the residuals for the linear regression models for testing causality, and relate to the three dates marked on the evolution of causal influence in Figure 9.10. Linear regression for the window ending on 24th June 1998 (first row in Figure 9.11), strongly suggests a causal direction from 1 month futures contract to the Baltic Dry Index for 1 lag, a relationship which our framework strongly rejects. But when we look at the residuals of the linear model, we see evidence of serial correlation and skewness, and this is arguably stronger than for the opposite direction for which linear regression model does not support the existence of causality. For window ending on 1st May 2002, linear regression results with residuals that exhibit very strong leptokurtic tails in both directions – and again our framework does not support the hypothesis of lack of causality here. Finally for the window ending on 21st January 2009 linear regression again does not sufficiently account for serial correlation, but in this case our framework rejects the hypothesis of lack of causality for the direction of BDI to 1 month WTI.

Our flexible causal framework allows to further the analysis with the step-wise correction for various nonlinear effect, as summarised in Table 9.2.

Our conjecture of serial correlation in residuals leading to overconfidence of the linear model is supported by results that correct for such serial correlation. Figure 9.12 presents the result of testing for causality with the model denotes as M2 in Table 9.2: GP framework that a) incorporates linear trend from linear regression, and b) does not incorporate causal structure in the covariance. The GP framework from

Table 9.2: Models designed for step-wise correction for various nonlinear effects.

no.	description	properties
GCCA	linear model (Granger causality)	linear causality
M2	GP with: linear trend from GCCA, no causality in covariance	corrects for: serial correlation only
M3	GP with: linear trend from GCCA, covariance function allowing for causality	corrects for: serial correlation, causality in covariance
GCP	GP with: linear mean function, covariance function allows for causality	corrects for: serial correlation, causality in covariance

Figure 9.13, named M3 in the Table 9.2, incorporates a) linear trend from the linear regression, b) allows for causality in covariance. Correcting for serial correlation removes some of the overconfidence of the linear regression model, which is then further reduced by also correcting for potential dependence in the covariance. To be more precise, allowing for causality in covariance decreases the test statistic (1-pvalue) in all but one cases, the direction 1m WTI $\rightarrow$ BDI for the 22 Aug 2008 – 21 Jan 2009 which suggests causal effect which at lag 1 could be better captured by the causality in covariance than causality in the mean.

We conclude that while using linear regression models for testing causality can have higher power, this could be misleading, as the model could be overconfident due to incorrect statistical assumptions. Using GPs can not only help with these specific structural properties that we mentioned: serial correlation and causality in covariance, but it goes even further, by allowing to test for causality under a range of model assumptions without penalising model misspecification.

9.2.1 Skewness, kurtosis and tail dependence.

It is natural to ask, what else can we discover about our data, if we are able to control for additional statistical structures? Looking at the partial autocorelation of the GPC models (Figure 9.19) we clearly see evidence of additional stochastic structures present for at least some of the windows, and that structure is more present when analysing the lag 8. To figure out which properties of the data might be of interest, we begin by looking at skewness, kurtosis, empirical lower and upper tail estimates, as well as lower and upper extremograms. The values of skewness, kurtosis, empirical lower and upper tail estimates are presented in two forms: in Table 9.3 for the whole history, as well as for the three previously identified time windows, and in Figures 9.14 - 9.15 for the two year moving windows.

As is common in financial data, we observe some type of departure from normality throughout the history. The most apparent effect is the skewness, which for BDI increases and decreases with a cyclical pattern, while the excess kurtosis appears infrequently. When analysing the properties of returns of the time series instead, however, we can see that excess kurtosis is the norm, which suggests that we might particularly benefit from using the warped GP framework for returns rather than for the levels.

In terms of properties related to symmetrical dependence, the data expresses empirical tail depen-

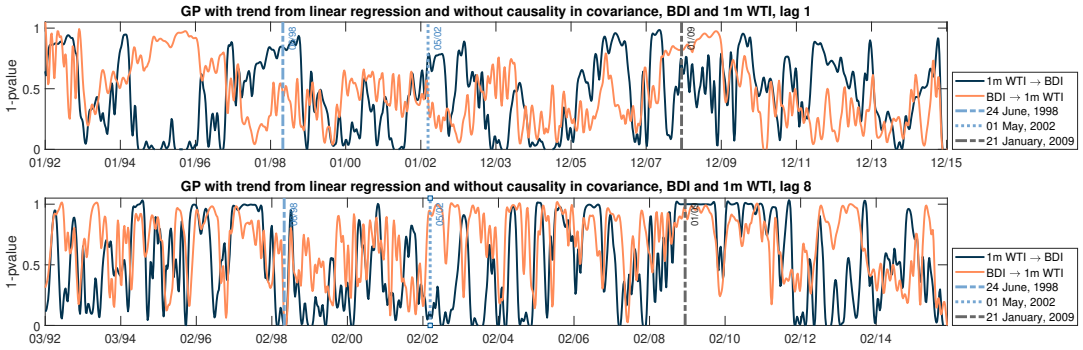


Figure 9.12: Evolution of the causal influence tested with the model M2, framework based on GPs with trend from linear regression and no causality in covariance: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing.

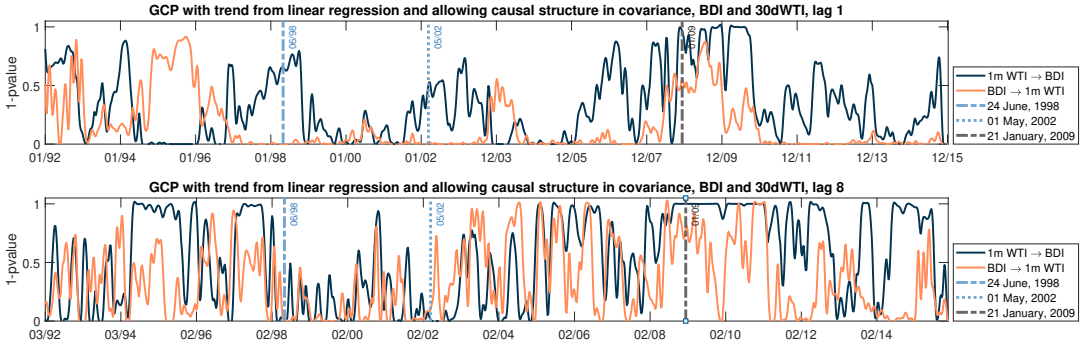


Figure 9.13: Evolution of the causal influence tested with the model M3, framework based on GPs with trend from linear regression and allowing for causality in covariance: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing.

Table 9.3: Statistical properties of the data that refer to the windows ending on 24 Jun 1998, 1 May 2002, and 21 Jan 2009.

data	property	27.01.98 - - 24.06.98	30.11.01 - - 01.05.02	22.08.08 - - 21.01.09	17.01.90 - - 23.12.15
WTI 1m	skewness	-0.17	-0.24	0.43	0.76
BDI	skewness	-0.21	-0.39	-0.38	2.56
WTI 1m	kurtosis	2.41	2.53	2.23	2.24
BDI	kurtosis	1.97	1.51	2.44	10.19
	empirical upper tail	0.37	0.61	0.28	0.17
	empirical lower tail	0.54	0.59	0.64	0.09
WTI 1m, returns	skewness	0.21	-0.8	-0.02	-0.22
BDI, returns	skewness	0.73	-0.45	-0.62	-1.14
WTI 1m, returns	kurtosis	2.95	4.42	3.3	8
BDI, returns	kurtosis	4.56	3.62	5.01	23.36
	empirical upper tail	0.08	0	0.12	0.11
	empirical lower tail	0	0.02	0.32	0.15

dence intermittently, with lower and upper tail dependence following similar pattern, especially for the levels. However, when we look at the extremograms in Figures 9.16 and 9.17, what we can observe is a strikingly asymmetrical effect, with high level of lower extremogram, which can be interpreted as high correlation for negative returns.

We run the warped GP model with similar settings as previously when using GP models: rolling window of 104 length, lags of 1 and 8. We utilise linear mean function and Matern covariance function with 3 degrees of freedom. We focus our attention on the standardised returns rather than levels, to form an insight about how the returns, and how the risk connected to those returns is affected by causal and non-causal dependence.

Previously, we have pointed out that in the case of GP framework, the existence of causality in the trend can overshadow the ease of detecting causality in covariance. But in the generalised setting of warped GPs however, we have additional interactions that can be incorporated, which affect the ability to detect causality. In the Section 8.3 we have shown on simulated data how tail dependence decreases the ability to detect causal dependence. And so we more consciously interpret the results as testing for relationships under certain structural assumptions.

The results of the warped GP framework with presented in Figure 9.18 points to the hypothesis of lack of causality being rejected for a majority of time windows. Given that the data, and in particular the returns time series, show clear evidence that kurtosis was required, once the ability to incorporate kurtosis has been included in the model, we find that the model gives preference to strong kurtosis. Subsequently, performing causal test with such models lead to more evidence for causal structures.

Restricting the warped GP model from incorporating skewness, as presented in Figure 9.20, does not have any substantial effect on the power of the model though. Which allows to conclude that it is the kurtosis structure, not the skewness structure, that is important for this case.

When we look at three windows that we have chosen for our analysis in Section 9.2, we see that allowing to incorporate skewness and kurtosis has led to two interesting effects. Firstly, as we have expected, there is a higher detection of causal dependence in volatility than in trend. Secondly, on the example of causal direction from WTI 1m to BDI in the second window (30.11.01 - 01.05.02), there is more evidence for causality in covariance rather than in the trend, but the hypothesis of lack of causality is most strongly reject when both of these effects are taken into consideration. When we look more closely at the parameters chosen for these models, the shape parameter ν that maximises the model B in each case is, respectively $7, 5$ and 3, which indicates that when causal effect in the covariance was not allowed, the model was compensating by incorporating higher multiplicative scaling covariance parameter.

Table 9.4: Power of the test for analysing causal dependence in trend, volatility and both, on three chosen time windows.

WTI 1m → BDI	27.01.98 - - 24.06.98	30.11.01 - - 01.05.02	22.08.08 - - 21.01.09
trend and volatility	1	1	0.95
volatility only	0	0.98	0.87
trend only	1	0.85	0
BDI → WTI 1m			
trend and volatility	1	0.61	1
volatility only	1	0	1
trend only	0.99	0.57	0

9.3 Commodity Futures Experiment Conclusions

We summarise the results of the real data experiment, by revisiting our questions and remarks from the Section 9.1.1. Firstly, analyse the causal structure using GP framework, and we observe that 8 weeks is generally enough for each of the markets to price in associated causal impacts in both oil futures markets and currency markets, which supports the literature that relates to efficient market hypothesis. We conclude that the different classes of investments affect the type and speed of reaction. We also observe, that the direction and significance of causal influence is affected by regimes, as shown on the example of the period of low prices: 17.01.1990 – 11.08.1999 (bellow $40), and period of high prices: 26.05.2004 – 11.03.2009 (above $90).

Our analysis involved only three investment classes, and therefore in no way sufficient to understand all important risk factors. We do however point out, that useful information can be obtained from analysing similarity of causal effects of two different factors. Such similarity can suggest that both factors are affected by a common factor (market volatility in our case). Increasing similarity of causal dependence can be understood in terms of systemic risk, see Billio et al. [2012].

We follow by employing a range of models to study the effect of including various structural properties of the data, and we concentrate on the Baltic Dry Index and 1 month oil futures pair. We carefully consider how the difference in the power of the model between using GP framework and linear regression (Granger causality) can be explained by correcting for serial correlation and causality in covariance. Subsequently, we study the returns of the time series, for which the covariance structure becomes more prominent, and studying them with the warped GP framework we are able to consistently detect causality.

We propose, that if one wants to test with misspecified models for causality structure in the mean, one might be better off with using the GP framework, and to test for causality in covariance with a misspecified model, one might consider choosing warped GP.

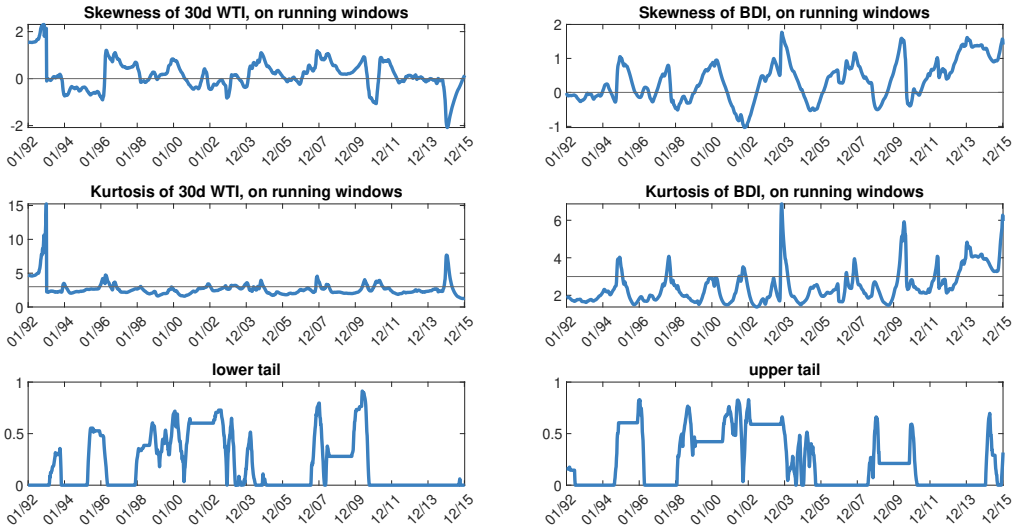


Figure 9.14: Skewness, kurtosis and tail dependence for running windows of length 104, for 30d WTI and BDI.

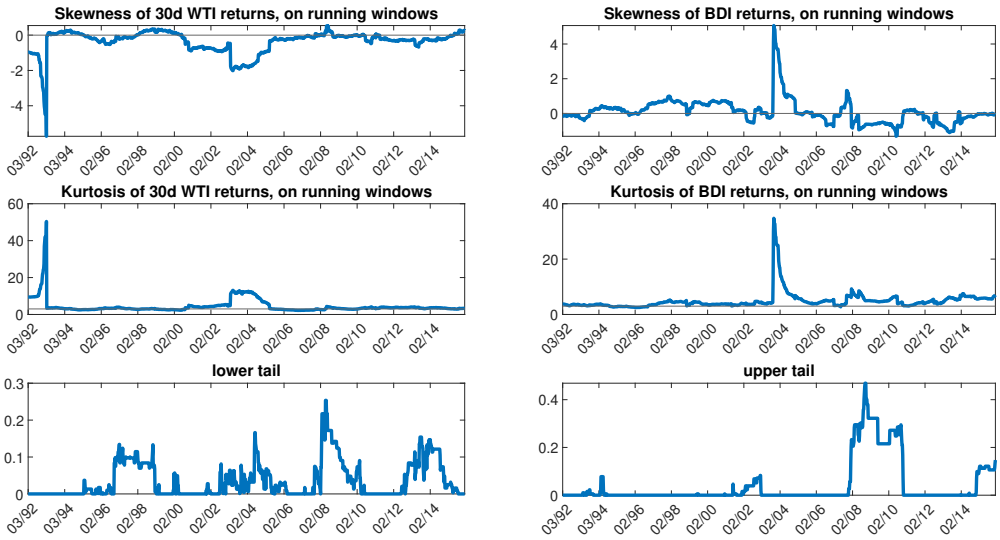


Figure 9.15: Skewness, kurtosis and tail dependence for running windows of length 104, for returns time series of 30d WTI and BDI.

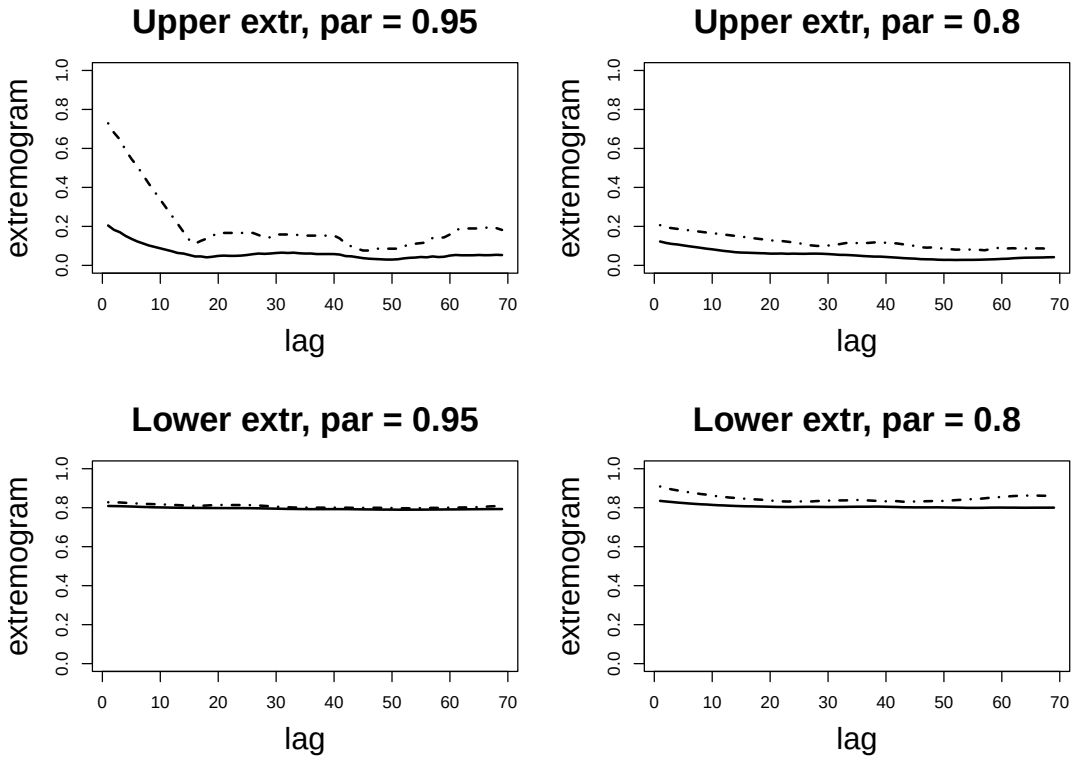


Figure 9.16: Lower and upper extremogram for the whole history, with par 0.95 and 0.8

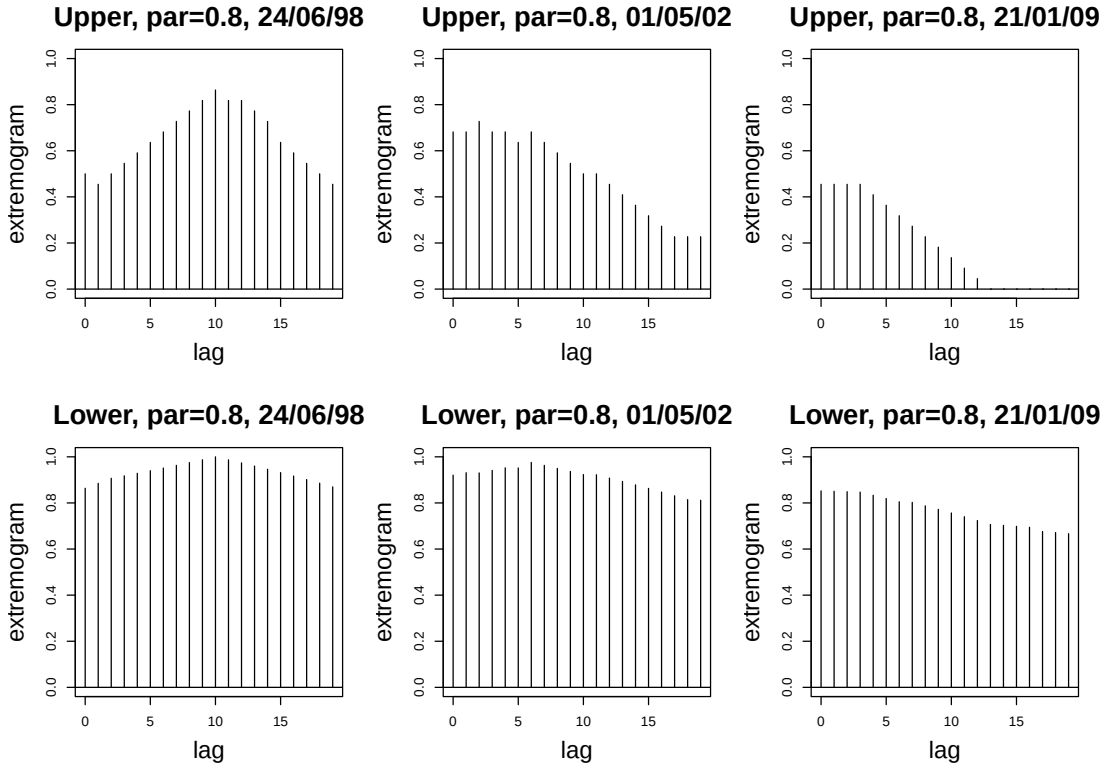


Figure 9.17: Lower and upper extremogram for the three analised dates, with par 0.8

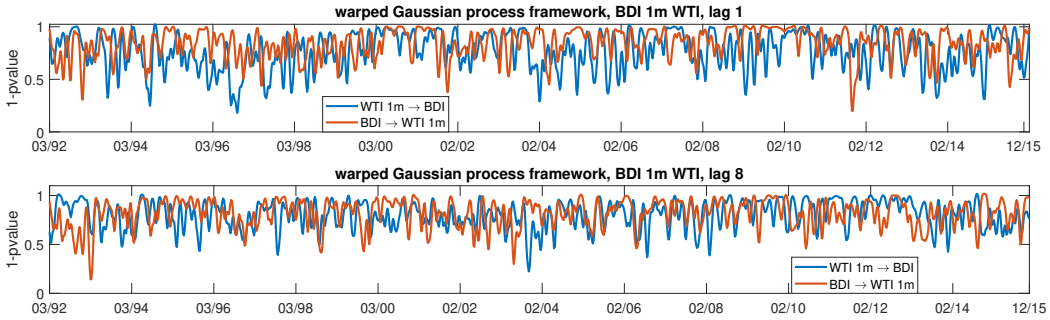


Figure 9.18: Evolution of the causal influence as modelled with warped GPC model: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing.

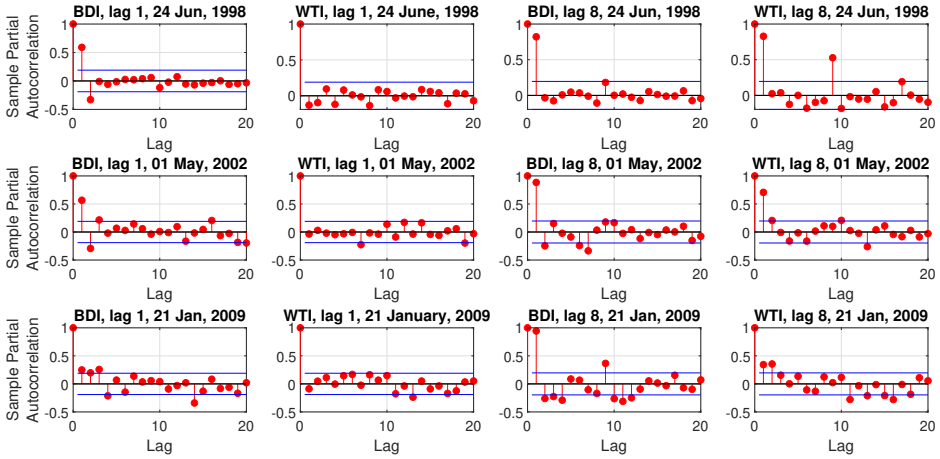


Figure 9.19: Sample partial autocorrelation of the residuals of the GPC model for the three chosen dates, and for lags 1 and 8.

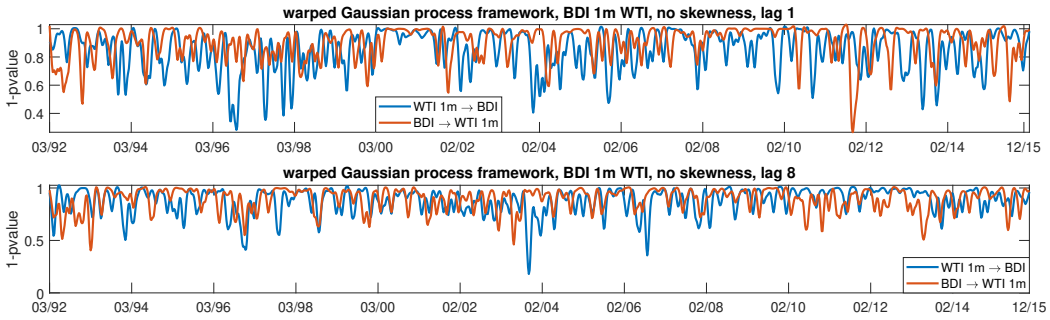


Figure 9.20: Evolution of the causal influence as modelled with wGPC model, but without skewness: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing.

Chapter 10

Conclusions

10.1 Summary

In the thesis we present a novel testing framework for statistical causality in general classes of multivariate nonlinear time series models: Gaussian process models for testing causality (GPC) framework, and its generalisation, warped Gaussian process models for testing causality (wGPC). They accommodate flexible features where causality may be present in linear or nonlinear forms either in the trend, volatility or both structural components of the general multivariate Markov processes under study. We also develop new classes of nonparametric multivariate time-series models based on warped Multiple Output Gaussian Processes. This allows to encode a serial dependency structure through a covariance function and introduce a more complex dependence structure using copulas to couple each warped marginal Gaussian process. In addition, we accommodate the added possibilities of flexible structural features such as long memory and persistence in the multivariate processes when applying our semiparametric approach to causality detection.

Proposed are calibration and formal testing procedures to detect these relationships through semiparametric models. We provide a generic framework which can be applied to a wide range of problems, including partially observed generalised diffusions or general multivariate linear or nonlinear time series models.

We develop several illustrative examples of features that are easily testable under our framework, to study the properties of the inference procedure developed including power of the test, sensitivity and robustness. We then illustrate our method on several real data examples from commodity modelling and interest rates and inflation.

10.2 Findings

We were interested in methods that had wide range of applicability, but in particular ones that would be practical for financial time series. Real data, and financial data specifically, exhibits a range of stochastic features including temporal trends, heteroskedasticity, nonstationarity, long memory, tail dependence. The

earlier methods we analysed did not accommodate most of these features. They either did not perform well on nonlinear, noisy time series, or had problems with estimation, model and/or parameter selection, or lacked good asymptotic properties of the test. Furthermore, there were no works claiming to be applicable for data with more than one of the following structural properties: nonstationarity, long memory, tail dependence. Researching existing methodology inspired to create a framework that will address the main shortcomings, while keeping many of the strengths of established approaches. While statistical causality is only one of the conceptual representations of causality, it is the one that is, arguably, the most suitable to the task of understanding causal dependence in financial time series.

In the Motivation section in Chapter (1) we identified the main research questions and the directions that we wanted to explore in our research. Listed below are the main results we have managed to achieve:

1. We defined models that allow parametrisation of different forms of statistical causality, in particular second order aspects of the process such as linear and nonlinear causality in covariance. The use of Gaussian process (GP) models meant that such parametrisations can be done in a convenient and easy to interpret way, that allow to form tractable causal tests.

2. Our framework is defined in multivariate time series context, by models which allow encoding both linear and nonlinear causality. These models are also capable of accounting for the following properties:

   (a) inheriting properties of GPs, they are flexible and can capture linear / nonlinear causality, while admitting Markov structure and knowledge of the conditional distribution of the model;

   (b) we describe how to use automatic relevance determination (ARD) models so that noncausality can be tested with nested hypothesis, allowing the use of generalised likelihood ratio test (GLRT);

   (c) use of GLRT with our frameworks results in test statistics that can be evaluated in closed form. In the case of GPC framework, these are given in analytic form, in the case of wGPC, these are approximated;

   (d) use of GLRT results in known asymptotic behaviour of the test statistic under the null;

   (e) In the GPC framework, parameter optimisation is statistically unbiased, efficient, consistent and computationally efficient, and parameters can be interpreted with respect to the structural properties of the model;

   (f) The framework can detect causality in the mean, covariance function, or higher order moments.

3. Our use of the mean-variance transformation (warping) and skew-t copula in the wGPC framework

extends the marginal models to incorporate range of special structures, for example nonstationarity, heteroscedasticity, leptokurtic tails, long memory.

4. Warping allows the joint model to have a wider range of dependence structures, for example: asymmetry, leptokurtic tails, tail dependence.

5. The experiments performed show that both skewness and kurtosis can affect the ability to perform inference, detect causality, and accuracy and power of the test.

6. With our testing framework it is possible to asses the power of the test for the models that meet the requirements from the points 1 through 5.

We would like to reiterate that the methods of Gaussian Process for causality (GPC) and warped Gaussian process for causality (wGPC) are easily customised. The choice of mean and kernel functions, noise structure, autoregressive structure – all these allow to capture a vast range of properties in the data. Warpings allow extra layer of flexibility, and the way that the warping is defined can also be altered to further generalise the entire framework.

10.3 Applications and results of Experiments

We provided an extensive set of experiments illustrating many of the above properties. The experiments with synthetic data demonstrate model sensitivity analysis, model misspecification analysis and power of the hypothesis test for simple and compound tests. The GPC framework, in particular, is shown to have good power of the test even for relatively small samples, and to not be sensitive to parameter changes for wide ranges of parameters. Following the properties of GP models, GPC framework has efficient algorithms for parameter optimisation. Most importantly, GPC is able to detect causality even when the model is misspecified. A particularly interesting illustration of that is given in Section (8.1), where we show that GPC can detect causality in data generated from a model with long memory property (ARFIMA).

The real data section explained how the framework can be used in practice, and how it can be combined with, or enhance more typical approaches to analysing financial time series. We provide illustrative examples for two sets of real data, firstly, commodity futures, dollar index, and Baltic Dry Index, and, secondly, 1 and 10 year US treasury bonds, US inflation linked swaps and US Consumer Price Index. Our results agree with economic interpretations, but also allow more insight into the dynamic relationships between the time series.

10.3.1 Commodity Futures Experiment Conclusions

In our analysis we used 1 and 36 month expiry oil futures contracts, US Dollar Index DXY and Baltic Dry Index which can be seen as a proxy to convenience yield, based on a component related to transportation

expense, given by the cost of freighting and short term storage [Ames et al., 2016]. Firstly, we concluded that 8 weeks was generally enough for each of the markets to price in associated causal impacts in both oil futures markets and currency markets, which supports the literature that relates to efficient market hypothesis. Different classes of investments, however, affect the type and speed of reaction. We also observed evidence of causal influence being affected by regimes, which is a commonly agreed on property of commodity markets.

Our analysis involved only three investment classes, and therefore in no way sufficient to understand all important risk factors. We do demonstrate, however, that useful information can be obtained from analysing similarity of causal effects of two different factors. Such similarity can suggest that both factors are affected by a common factor (market volatility in our case), and can be interpreted in terms of systemic risk [Billio et al., 2012].

The second study was based on US inflation proxied by 10 US year inflation based swaps, the interest rates based on 1 year and 10 year US treasury bonds, and US Consumer Price Index (CPI) as a side information.

10.4 Future Research and Directions

We believe that one of the greatest strengths of the model framework that we are proposing is its flexibility, making it easy to extend or alter. During our research, we came across a multitude of topics, models, special cases or extensions that were closely related to our work, and deciding what not to pursue was sometimes a bigger problem than deciding what to pursue. There are some areas that we did not have the opportunity to fully explore, and which will form the foundations of future publications.

In our research we have included some work on employing multiple output Gaussian processes, but more could be done to describe and test how the cross-correlation can affect recognition of causality. The use of GPC with multiple output GPs allows a convenient definition of instantaneous causality (instantaneous coupling), which could be of interest in a wide range of applications. A property of Gaussian processes that we have not explored, but that makes them particularly useful for time series data, is feasibility to treat data with different time markers.

We mentioned use of different warpings – to obtain different skew-t copulas. This work is of particular interest for modelling causality in the presence of tail dependence. We looked at three types of skew-t distributions, with different types of tail dependence, but one might find different structures that are preferable. Various structures could be introduced to the framework – especially via the warping representation. For example, warping with the generalised inverse Gaussian, which results in generalised hyperbolic distribution, can be catered for in the wGPC very similarly to the skew-t, with the main difference being the increased computational complexity of estimating additional hyperparameters.

In the Gaussian process literature methods of sparse approximations are seen as important for

enabling the processing of bigger data sets. Some of the natural approaches here would be to alter the input data (choosing subset of points, or using pseudo-inputs), choose or design appropriate covariance matrix (low rank approximations for the covariance matrix, covariance function that generates sparse covariance matrix), alter inference methods (perform the inference only on points in the neighbourhood of a query point, use Nyström approximation).

A direction which we did not intend to explore, but which could be found useful for some applications, is the extension of GLRT to non-nested models, or replacement of GLRT with a different test for non-nested models.

There are several applications of great interest – that had to be sacrificed due to limited time and resources. A particularly interesting application, to which our framework would have been very suitable, is the analysis of the relationship between different types of bonds (corporate, municipal and green bonds) and their term structure.

Part III

Additional Material

Appendices

Appendix A

Appendix

A.1 Solving Ridge Regression

We have introduced ridge regression and kernel ridge regression in Chapter (1), here we describe how the solutions are obtained. The time series of interest $\{Y_t\}$ is explained by a regression model with time series $\{Q_t\}$ as covariates, and p lags taken into consideration:

$$Y_t = \alpha^T \mathbf{Q}_{t-p:t-1} + \epsilon_t. \tag{A.1}$$

Or equivalently:

$$\mathbf{Y}_{t_1:t_2} = \alpha^T \mathbb{Q}_{t_1,t_2;p} + \boldsymbol{\epsilon}. \tag{A.2}$$

Ridge regression aims to find coefficients α that minimise the squared error $\sum_{t=t_1}^{t_2} \left(Y_t - \boldsymbol{\alpha}^T \mathbf{Q}_{t-p:t-1}\right)^2$ subject to an L2 regularisation – added penalty, which equals the square of the magnitude of coefficients $\boldsymbol{\alpha}^T \boldsymbol{\alpha}$ scaled by a chosen constant λ:

$$\boldsymbol{\alpha}^* = \underset{\alpha}{\operatorname{argmin}} \left\{ \sum_{t=t_1}^{t_2} \left(Y_t - \boldsymbol{\alpha}^T \mathbf{Q}_{t-p:t-1}\right)^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} \right\}. \tag{A.3}$$

The minimisation problem from Equation (A.3) is typically solved using the technique of Lagrange multipliers:

$$\begin{aligned} \mathcal{L} :=& \sum_{t=t_1}^{t_2} \left(Y_t - \boldsymbol{\alpha}^T \mathbf{Q}_{t-p:t-1}\right)^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} \\ =& \left(\mathbb{Q}_{t_1,t_2;p} \boldsymbol{\alpha} - \mathbf{Y}_{t_1:t_2}\right)^T \left(\mathbb{Q}_{t_1,t_2;p} \boldsymbol{\alpha} - \mathbf{Y}_{t_1:t_2}\right) + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} \\ =& \boldsymbol{\alpha}^T \mathbb{Q}_{t_1,t_2;p}^T \mathbb{Q}_{t_1,t_2;p} \boldsymbol{\alpha} - 2 \mathbf{Y}_{t_1:t_2}^T \mathbb{Q}_{t_1,t_2;p} \boldsymbol{\alpha} + \mathbf{Y}_{t_1:t_2}^T \mathbf{Y}_{t_1:t_2} + \lambda \boldsymbol{\alpha}^T \boldsymbol{\alpha} \end{aligned} \tag{A.4}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = 2\mathbb{Q}^T_{t_1,t_2;p}\mathbb{Q}_{t_1,t_2;p}\boldsymbol{\alpha} - 2\mathbf{Y}^T_{t_1:t_2}\mathbb{Q}_{t_1,t_2;p} + 2\lambda\boldsymbol{\alpha}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = 0 \Leftrightarrow$$

$$\mathbb{Q}^T_{t_1,t_2;p}\mathbb{Q}_{t_1,t_2;p}\boldsymbol{\alpha}^* + \lambda I_{t_2-t_1}\boldsymbol{\alpha}^* = \mathbb{Q}^T_{t_1,t_2;p}\mathbf{Y}_{t_1:t_2} \Leftrightarrow \tag{A.5}$$

primal solution: $$\boldsymbol{\alpha}^* = \left(\mathbb{Q}^T_{t_1,t_2;p}\mathbb{Q}_{t_1,t_2;p} + \lambda I_{t_2-t_1}\right)^{-1}\mathbb{Q}^T_{t_1,t_2;p}\mathbf{Y}_{t_1:t_2} \tag{A.6}$$

We notice however that using equality A.5 $\boldsymbol{\alpha}^*$ can also be expressed as follows:

$$\mathbb{Q}^T_{t_1,t_2;p}\mathbb{Q}_{t_1,t_2;p}\boldsymbol{\alpha}^* + \lambda I_{t_2-t_1}\boldsymbol{\alpha}^* = \mathbb{Q}^T_{t_1,t_2;p}\mathbf{Y}_{t_1:t_2} \Leftrightarrow$$

$$\boldsymbol{\alpha}^* = \frac{1}{\lambda}\mathbb{Q}^T_{t_1,t_2;p}\left(\mathbf{Y}_{t_1:t_2} - \mathbb{Q}_{t_1,t_2;p}\boldsymbol{\alpha}^*\right) \Leftrightarrow \tag{A.7}$$

$$\boldsymbol{\alpha}^* = \mathbb{Q}^T_{t_1,t_2;p}\underbrace{\frac{1}{\lambda}\left(\mathbf{Y}_{t_1:t_2} - \mathbb{Q}_{t_1,t_2;p}\boldsymbol{\alpha}^*\right)}_{\text{denote by } \boldsymbol{\beta}^*} \Leftrightarrow \tag{A.8}$$

$$\boldsymbol{\alpha}^* = \mathbb{Q}^T_{t_1,t_2;p}\boldsymbol{\beta}^*. \tag{A.9}$$

As a next step, several transformations are performed that would allow us to express the new weights in terms covariance matrix (inner product) of the covariates:

$$\boldsymbol{\beta}^* = \frac{1}{\lambda}\left(\mathbf{Y}_{t_1:t_2} - \mathbb{Q}_{t_1,t_2;p}\boldsymbol{\alpha}^*\right) \Leftrightarrow \tag{A.10}$$

$$\boldsymbol{\beta}^* = \frac{1}{\lambda}\left(\mathbf{Y}_{t_1:t_2} - \mathbb{Q}_{t_1,t_2;p}\mathbb{Q}^T_{t_1,t_2;p}\boldsymbol{\beta}^*\right) \Leftrightarrow \tag{A.11}$$

dual solution: $$\boldsymbol{\beta}^* = \left(\mathbb{Q}_{t_1,t_2;p}\mathbb{Q}^T_{t_1,t_2;p} + \lambda I_{t_2-t_1}\right)^{-1}\mathbf{Y}_{t_1:t_2}. \tag{A.12}$$

The dual solution of ridge regression, is in such a form that covariates appear only in the form of covariance matrix (inner product) $\mathbb{Q}_{t_1,t_2;p}\mathbb{Q}^T_{t_1,t_2;p}$. This means that it can be "kernelised" by substituting the existing covariance matrix (inner product) by one that is inferred from a Mercer kernel $k(\cdot,\cdot)$, and denoted $\mathbf{K}_\mathbb{Q}$, where $\{\mathbf{K}_\mathbb{Q}\}_{i,j} = k(\mathbf{Q}_{i-p:i-1}, \mathbf{Q}_{j-p:j-1})$. Consequently, the optimal weights for kernel ridge regression will take the form:

$$\boldsymbol{\beta}^{krr} = \left(\mathbf{K}_\mathbb{Q} + \lambda\mathbf{1}_{t_2-t_1}\right)^{-1}\mathbf{Y}_{t_1:t_2} \tag{A.13}$$

A.2 Predictive / conditional distribution

Here we remind a formula that is typically used in the case of predictive distribution, but which can be simply seen as a conditional distribution.

Lets say that we have data $\mathbf{y} = [\mathsf{y}_1, ..., \mathsf{y}_N]$, $\mathbf{g} = [\mathsf{g}_1, ..., \mathsf{g}_N]$ and we assume that the data come from a

following GP model:

$$g_i = f(y_i) + \epsilon_i, \qquad f \sim \mathcal{GP}(0, k_{t,t'}),\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

So if we use denote the covariance matrix $\mathbf{K}$: $\mathsf{k}_{l,m} = k_Y(\mathsf{y}_l, \mathsf{y}_m)$, $l, m \in [1, T]$, then we know that $\mathbf{g} \sim \mathcal{N}(0, \mathbf{K} + \sigma_Y^2 \mathbf{I})$.

Lets take a new point y_*, for which we are interested in the noiseless value of $\mathsf{f}_* = f(\mathsf{y}_*)$, that would be consistent with the model and observations so far. If we treat y_* as a new observation, then f_* will be seen as predictive distribution, however if y_* is not a new observation, but for example part of a more dense grid of observations we had so far, then we are simply talking about a conditional distribution, given by equation A.14:

$$\begin{aligned} &\mathsf{f}_* \mid \mathbf{y}, \mathbf{g}, \mathsf{y}_* \sim \left(\mathsf{f}_*; \bar{\mathsf{f}}_*, cov(\mathsf{f}_*)\right), \\ &\bar{\mathsf{f}}_* = K(\mathsf{y}_*, \mathbf{y}) \left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1} \mathbf{g} \\ &cov(\mathsf{f}_*) = K(\mathsf{y}_*, \mathsf{y}_*) - K(\mathsf{y}_*, \mathbf{y}) \left[\mathbf{K} + \sigma^2 \mathbf{I}\right]^{-1} K(\mathbf{y}, \mathsf{y}_*). \end{aligned} \tag{A.14}$$

A.3 Obtaining marginal likelihood

We want to show:

$$p(Y_t \mid X_t) = \mathcal{N}(\mu(X_t), k(X_t, X_t) + \sigma_t^2). \tag{A.15}$$

Introduce shorthand notation: $f_t \equiv f(X_t), \mu_t \equiv \mu(X_t), k_t \equiv k(, X_t)$.

$$
\begin{aligned}
p(Y_t \mid X_t) &= \int p(Y_t \mid X_t, f_t) p(f_t \mid X_t) df_t = \\
&= \int \mathcal{N}(Y_t; f_t, \sigma_t^2) \mathcal{N}(f_t; \mu_t, k_t) df_t = \\
&= \int \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_t^2}} \exp\left(-\frac{1}{2}\frac{(Y_t - f_t)^2}{\sigma_t^2}\right) \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{k_t}} \exp\left(-\frac{1}{2}\frac{(f_t - \mu_t)^2}{k_t}\right) df_t = \\
&= \int \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma_t^2}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{k_t}} \exp\left(-\frac{1}{2}\frac{Y_t^2}{\sigma_t^2} + \frac{Y_t f_t}{\sigma_t^2} - \frac{1}{2}\frac{f_t^2}{\sigma_t^2} - \frac{1}{2}\frac{f_t^2}{k} + \frac{f_t \mu_t}{k} - \frac{1}{2}\frac{\mu_t^2}{k}\right) df_t = \\
&= \int \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{k_t + \sigma_t^2}} \exp\left(-\frac{1}{2}\frac{(Y_t - \mu_t)^2}{k_t + \sigma_t^2}\right) \\
&\frac{1}{\sqrt{2\pi}} \sqrt{\frac{1}{k_t} + \frac{1}{\sigma_t^2}} \exp\left(-\frac{1}{2}\left(f_t - \left(\frac{Y_t}{\sigma_t^2} + \frac{\mu_t}{k_t}\right)\left(\frac{1}{k_t} + \frac{1}{\sigma_t^2}\right)^{-1}\right)^2 \left(\frac{1}{k_t} + \frac{1}{\sigma_t^2}\right)\right) df_t = \\
&= \mathcal{N}(Y_t; \mu_t, k_t + \sigma_t^2) \int \mathcal{N}\left(f_t; \left(\frac{Y_t}{\sigma_t^2} + \frac{\mu_t}{k_t}\right)\left(\frac{1}{k_t} + \frac{1}{\sigma_t^2}\right)^{-1}, \left(\frac{1}{k_t} + \frac{1}{\sigma_t^2}\right)^{-1}\right) df_t \\
&= \mathcal{N}(Y_t; \mu_t, k_t + \sigma_t^2)
\end{aligned}
$$

A.4 Proof of Theorem (7)

Proof:

We use the fact that the $\tilde{V}$ will be normally distributed when conditioned on the mixing variable: $\tilde{V} \mid W \sim \mathcal{N}(m + \gamma W, W\mathbf{\Sigma})$. The unconditional distribution is therefore calculated as a following integral:

$$
f(\tilde{v}) = \int_0^\infty f(\tilde{v} \mid w) p(w) dw \tag{A.16}
$$

Using the density of a generalised inverse Gaussian (GIG) distribution $W \sim GIG(\lambda, \chi, \psi)$ from the Equation 31 we extend A.16 and write:

$$
\begin{aligned}
p(\tilde{v}) &= \int_0^\infty \frac{1}{(2\pi)^{\frac{d}{2}} \mid \mathbf{\Sigma} \mid^{\frac{1}{2}} w^{\frac{d}{2}}} \exp\left\{\frac{-(\tilde{v} - m - \gamma w)^T (w\mathbf{\Sigma})^{-1} (\tilde{v} - m - \gamma w)}{2}\right\} p(w) dw \\
&= \int_0^\infty \frac{e^{(x-m)^T \mathbf{\Sigma}^{-1} \gamma}}{(2\pi)^{\frac{d}{2}} \mid \mathbf{\Sigma} \mid^{\frac{1}{2}} w^{\frac{d}{2}}} \exp\left\{-\frac{(\tilde{v} - m)^T \mathbf{\Sigma}^{-1} (\tilde{v} - m)}{2w} - \frac{\gamma^T \mathbf{\Sigma}^{-1} \gamma}{2/w}\right\} p(w) dw \\
&= \frac{\chi^{-\lambda} \left(\sqrt{\chi\psi}\right)^\lambda e^{(x-m)^T \mathbf{\Sigma}^{-1} \gamma}}{(2\pi)^{\frac{d}{2}} \mid \mathbf{\Sigma} \mid^{\frac{1}{2}} K_\lambda\left(\sqrt{\chi\psi}\right)} \\
&\quad \frac{1}{2} \int_0^\infty w^{\lambda - \frac{d}{2} - 1} \exp\left\{-\frac{(\tilde{v} - m)^T \mathbf{\Sigma}^{-1} (\tilde{v} - m) + \chi}{2w} - \frac{\gamma^T \mathbf{\Sigma}^{-1} \gamma + \psi}{2/w}\right\} p(w) dw
\end{aligned}
$$

To allow expressing the integral in terms of a modified Bessel function of the third kind, we perform a change of variable:

$$z = w\frac{\sqrt{\left(\psi + \gamma^T\mathbf{\Sigma}^{-1}\gamma\right)}}{\sqrt{\left(\psi + (\tilde{v}-m)^T\mathbf{\Sigma}^{-1}(\tilde{v}-m)\right)}}$$

and as a result we obtain:

$$p(\tilde{v}) = \frac{\chi^{-\lambda}\left(\sqrt{\chi\psi}\right)^{\lambda} e^{(x-m)^T\mathbf{\Sigma}^{-1}\gamma}}{(2\pi)^{\frac{d}{2}} \mid \mathbf{\Sigma} \mid^{\frac{1}{2}} K_{\lambda}\left(\sqrt{\chi\psi}\right)} \left(\frac{\sqrt{\left(\psi + (\tilde{v}-m)^T\mathbf{\Sigma}^{-1}(\tilde{v}-m)\right)}}{\sqrt{\left(\psi + \gamma^T\mathbf{\Sigma}^{-1}\gamma\right)}}\right)^{\frac{d}{2}-\lambda}$$
$$\underbrace{\frac{1}{2}\int_0^{\infty} z^{\lambda-\frac{d}{2}-1} \exp\left\{-\frac{1}{2}\sqrt{\left(\chi + (\tilde{v}-m)^T\mathbf{\Sigma}^{-1}(\tilde{v}-m)\right)\left(\psi + \gamma^T\mathbf{\Sigma}^{-1}\gamma\right)}\left[\frac{1}{z} + z\right]\right\} f(z)dz}_{K_{\lambda-\frac{d}{2}}\left(\sqrt{\left(\chi+(\tilde{v}-m)^T\mathbf{\Sigma}^{-1}(\tilde{v}-m)\right)\left(\psi+\gamma^T\mathbf{\Sigma}^{-1}\gamma\right)}\right)}$$

which after reorganisation gives the requested density. □

A.5 Proof of (Theorem 8)

Proof:

The moment generating function of GIG is:

$$M_X(s) = \left(\frac{\psi}{\psi - 2s}\right)^{\frac{\lambda}{2}} \frac{K_{\lambda}\left(\sqrt{\chi + (\psi - 2s)}\right)}{K_{\lambda}\left(\sqrt{\chi\psi}\right)},$$

and the fact that moment generating function of variable $X \sim \mu, \mathbf{\Sigma}$ is equal $\exp(s^T\mu + \frac{1}{2}s^T\mathbf{\Sigma}s)$. Then for the variable $X \sim GH(\lambda, \chi, \psi, \mu, \mathbf{\Sigma}, \gamma)$, with $W \mid GIG(\lambda, \chi, \psi)$ being the mixing variable:

$$\begin{aligned} M_X(s) =& \mathbf{E}\left(\mathbf{E}\left(e^{s^TX}\right) \mid W\right) = \mathbf{E}\left(e^{s^T(\mu+\gamma W)+\frac{1}{2}s^TW\mathbf{\Sigma}s}\right) = e^{s^T\mu}\mathbf{E}\left(e^{\left(s^T\gamma+\frac{1}{2}s^T\mathbf{\Sigma}s\right)W}\right) \\ =& e^{s^T\mu}M_W\left(s^T\gamma + \frac{1}{2}s^T\mathbf{\Sigma}s\right). \end{aligned}$$

Expressing the moment generating function of GH distribution in terms of GIG distribution is practical. But we can also expand it to obtain:

$$M_X(s) = e^{s^T\mu}\left(\frac{\psi}{\psi - 2\left(s^T\gamma + \frac{1}{2}s^T\mathbf{\Sigma}s\right)}\right)^{\frac{\lambda}{2}} \frac{K_{\lambda}\left(\sqrt{\chi + (\psi - 2\left(s^T\gamma + \frac{1}{2}s^T\mathbf{\Sigma}s\right))}\right)}{K_{\lambda}\left(\sqrt{\chi\psi}\right)}.$$

□

A.6 Hilbert-Schmidt Independence Criterion (HSIC)

Let $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}$ denote the RKHS induced by strictly positive kernels $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Let X be random variable on $\mathcal{X}$, Y be random variable on $\mathcal{Y}$ and (X, Y) be random vector on $\mathcal{X} \times \mathcal{Y}$. The marginal distributions are denoted by P_X, P_Y and the joint distribution of (X, Y) by P_{XY}.

Definition 43 *(Hilbert-Schmidt Independence Criterion – HSIC) With the notation for $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}, P_X, P_Y$ as introduced earlier, we define the Hilbert Schmidt independence criterion as the squared Hilbert Schmidt norm of the cross-covariance operator Σ_{XY}:*

$$HSIC(P_{XY}, \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) := \|\Sigma_{XY}\|^2_{HS}. \tag{A.17}$$

We cite without proof the following lemma from Gretton et al. [2005]:

Lemma 5 ***(HSIC in kernel notation)***

$$\begin{aligned} HSIC(P_{XY}, \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) :=& \boldsymbol{E}_{X,X',Y,Y'}[k_\mathcal{X}(X, X')k_\mathcal{Y}(Y, Y')] + \boldsymbol{E}_{X,X'}[k_\mathcal{X}(X, X')]\boldsymbol{E}_{Y,Y'}[k_\mathcal{Y}(Y, Y')] \\ & -2\boldsymbol{E}_{X,Y}[\boldsymbol{E}_{X'}[k_\mathcal{X}(X, X')]\boldsymbol{E}_{Y'}[k_\mathcal{Y}(Y, Y')]]. \end{aligned} \tag{A.18}$$

where X, X' and Y, Y' are independent copies of the same random variable.

A.7 Estimator of HSNCIC

Empirical mean elements:

$$\begin{aligned} \hat{m}^{(n)}_X &= \frac{1}{n}\sum_{i=1}^{n} k_\mathcal{X}(\cdot, X_i), \\ \hat{m}^{(n)}_Y &= \frac{1}{n}\sum_{i=1}^{n} k_\mathcal{Y}(\cdot, Y_i) \end{aligned} \tag{A.19}$$

Empirical cross-covariance operator:

$$\begin{aligned} \hat{\Sigma}^{(n)}_{XY} &= \frac{1}{n}\sum_{i=1}^{n}(k_\mathcal{Y}(\cdot, Y_i) - \hat{m}^{(n)}_Y)\langle k_\mathcal{X}(\cdot, X_i) - \hat{m}^{(n)}_X, \cdot\rangle_{\mathcal{H}_\mathcal{X}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\{k_\mathcal{Y}(\cdot, Y_i) - \hat{m}^{(n)}_Y\} \otimes \{k_\mathcal{X}(\cdot, X_i) - \hat{m}^{(n)}_X\}. \end{aligned} \tag{A.20}$$

Empirical normalised cross-covariance operator

$$\hat{V}^{(n)}_{XY} = (\hat{\Sigma}^{(n)}_{XX} + n\lambda I_n)^{-1/2}\hat{\Sigma}^{(n)}_{XY}(\hat{\Sigma}^{(n)}_{YY} + n\lambda I_n)^{-1/2}, \tag{A.21}$$

where $n\lambda I_n$ is added to ensure invertibility.

Empirical normalised conditional cross-covariance operator

$$\hat{V}^{(n)}_{XY|Z} = \hat{V}^{(n)}_{XY} - \hat{V}^{(n)}_{XZ}\hat{V}^{(n)}_{ZY}. \tag{A.22}$$

For U symbolising any of the variables (XZ), (YZ) or Z, we denote by K_U a centred Gram matrix, such that each elements equal to: $K_{U,ij} = \langle k_{\mathcal{U}}(\cdot, U_i) - \hat{m}^{(n)}_U, k_{\mathcal{U}}(\cdot, U_j) - \hat{m}^{(n)}_U\rangle_{\mathcal{H}_{\mathcal{U}}}$, let $R_U = K_U(K_U + n\lambda I)^{-1}$. With this notation the empirical estimation of HSNCIC can be written as:

$$HSNCIC_n := Tr[R_{(XZ)}R_{(YZ)} - 2R_{(XZ)}R_{(YZ)}R_Z + R_{(XZ)}R_ZR_{(YZ)}R_Z]. \tag{A.23}$$

A.8 Sieve bootstrap two-sample t-test

Following Chen and Gel [2011], this section describes a sieve bootstrap two-sample t-test that corrects for serial correlation.

A.8.1 Classical t-test for two sample problem

Random samples: $\mathbf{X}_1 = (X_{1,1}, ..., X_{1,n_1})$ and $\mathbf{X}_2 = (X_{2,1}, ..., X_{2,n_2})$, with joint continuous distributions: F_1, F_2, with n_1 and n_2 observations. Let μ_1, μ_2 be corresponding population means. We want to test if the means are equal or not:

$$H_0: \quad \mu_1 = \mu_2 \qquad\qquad H_1: \quad \mu_1 \neq \mu_2.$$

Let $X_{1,j} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_{2,j} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$. The t-test statistic is defined as:

$$T = \frac{\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \tag{A.24}$$

where: for $i = 1, 2$ we have sample means $\bar{X}_i = n_i^{-1}\Sigma_{j=1}^{n_i} X_{i,j}$ and sample variances $s_i^2 = (n_i - 1)^{-2}\Sigma_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2$.

Under H_0, the test statistic T follows approximately a t-distribution with ν degrees of freedom:

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2(n_1 - 1) + \left(s_2^2/n_2\right)^2(n_2 - 1)}.$$

A.8.2 Model Assumptions

Random samples: $\mathbf{X}_1 = (X_{1,1}, ..., X_{1,n_1})$ and $\mathbf{X}_2 = (X_{2,1}, ..., X_{2,n_2})$, with joint continuous distributions: F_1, F_2, with n_1 and n_2 observations. Let μ_1, μ_2 be corresponding population means. We want to test if the

means are equal or not:

$$H_0: \quad \mu_1 = \mu_2 \qquad\qquad H_1: \quad \mu_1 \neq \mu_2.$$

We assume that the samples are no longer independent, but instead we assume that $\{X_{1,t}\}_{t\in\mathbb{Z}}$, $\{X_{2,t}\}_{t\in\mathbb{Z}}$ can be described by models:

$$\Sigma_{j=0}^{\infty} c_j \left(X_{1,t-j} - \mu_1\right) = \epsilon_{1,t} \qquad \Sigma_{j=0}^{\infty} c_j < \infty \tag{A.25}$$

$$\Sigma_{j=0}^{\infty} d_j \left(X_{2,t-j} - \mu_2\right) = \epsilon_{2,t} \qquad \Sigma_{j=0}^{\infty} d_j < \infty \tag{A.26}$$

A.8.3 Algorithm

Step 0. Prepare the time series, by differencing if the time series is not stationary.

Step 1. Select AR **orders** p_1, p_2 with Akaike information criterion for $\mathbf{X}_1, \mathbf{X}_2$.

Step 2. Estimate AR **coefficients** for $\mathbf{X}_1, \mathbf{X}_2$: $(\hat{c}_1, ...\hat{c}_{p_1}), (\hat{d}_1, ...\hat{d}_{p_2})$, using Yule Walker equations.

Step 3. Estimate **residuals** $\{\hat{\epsilon}_{1,t}\}_{t=p_1+1}$ and $\{\hat{\epsilon}_{2,t}\}_{t=p_2+1}$ for the models A.25 - A.26, with:

$$\hat{\epsilon}_{1,t} = \Sigma_{j=0}^{p_1} \hat{c}_j \left(X_{1,t-j} - \bar{X}_1\right) \qquad \hat{c}_0 = 1, t = p_1 + 1, ..., n_1$$

$$\hat{\epsilon}_{2,t} = \Sigma_{j=0}^{p_2} \hat{d}_j \left(X_{2,t-j} - \bar{X}_2\right) \qquad \hat{d}_0 = 1, t = p_2 + 1, ..., n_2$$

Step 4. Center estimated residuals:

$$\tilde{\epsilon}_{i,t} = \left(\hat{\epsilon}_{i,t} - \frac{1}{T - p_i} \Sigma_{t=p_i+1}^{n_i} \hat{\epsilon}_{i,t}\right).$$

The **empirical distributions of the centered residuals** $\{\bar{\epsilon}_{i,t}\}_{t=p_i+1}$ are:

$$\hat{F}_{\epsilon_{i,t},T}(y) = \Sigma_{t=p_i+1}^{n_i} \mathbf{1}_{\{\bar{\epsilon}_{i,t} \leq y\}}.$$

Step 5. Sample with replacement **bootstrap error processes** $\left\{\epsilon_{1,t}^*\right\}_{t=1}^{T}$ and $\left\{\epsilon_{2,t}^*\right\}_{t=2}^{T}$ from the empirical distributions $\hat{F}_{\epsilon_{1,t},T}(y)$, $\hat{F}_{\epsilon_{2,t},T}(y)$.

Step 6. Construct bootstrap samples $\mathbf{X}_1^* = (X_{1,1}^*, ..., X_{1,n_1}^*)$ and $\mathbf{X}_2^* = (X_{2,1}^*, ..., X_{2,n_1}^*)$, by recursion:

$$\Sigma_{j=0}^{\infty} \hat{c}_j \left(X_{1,t}^* - \hat{X}_1\right) = \epsilon_{1,t}^* \qquad \hat{c}_0 = 1, t = p_1 + 1, ..., n_1$$

$$\Sigma_{j=0}^{\infty} d_j \left(X_{2,t}^* - \hat{X}_2\right) = \epsilon_{2,t}^* \qquad \hat{d}_0 = 1, t = p_2 + 1, ..., n_2$$

Step 7. Calculate the bootstrap t-statisic T_b^*:

$$T_b^* = \frac{\bar{X}_1^* - \bar{X}_2^*}{\sqrt{s_1^{2*}/n_1 + s_2^{2*}/n_2}},$$

where: for $i = 1, 2$ we have sample means $\bar{X}_i^* = n_i^{-1}\Sigma_{j=1}^{n_i} X_{i,j}^*$ and sample variances $s_i^{2*} = (n_i - 1)^{-2}\Sigma_{j=1}^{n_i}(X_{i,j}^* - \bar{X}_i^*)^2$.

Step 8. Generate B values of the bootstrap t-statistic $\left\{T_1^*, ..., T_B^*\right\}$.

Under the null hypothesis, the distribution of the test statistic T from the equation A.24 is approximated by the bootstrap distribution of T^*:

$$\hat{F}_{T^*}(x) = \Sigma_{b=1}^{B} \mathbf{1}_{\{T_b^* < x\}}.$$

Reject the null hypothesis if $T \leq Q_T^*(\alpha)$ or $T \geq Q_T^*(1-\alpha)$, if α is significance level, $Q_T^*(\alpha)$ is a lower α-quantile and $Q_T^*(1-\alpha)$ is an upper α-quantile of the distribution $\hat{F}_{T^*}(x)$.

Appendix B

Experiments and real data applications from Zaremba and Aste 2014.

This chapter Experiments and Real Data applications from "Measures of Causality in Complex Datasets with Application to Financial Data", [Zaremba and Aste, 2014].

B.1 The Four Chosen Methods

Four methods of measuring and testing causality used in this section, defined in Chapter (1) in Equations (1.8, 1.26, 1.50, 1.62).

1. Classical Granger causality (GC)

$$L^{GC}_{X\to Y} = \log\left[\frac{V_Y\left[Y,Z;p\right]}{V_Y\left[X,Y,Z;p\right]}\right] \tag{B.1}$$

2. Transfer entropy (TE)

$$L^{TE}_{X\to Y} = H(X \mid X_{t-k:t-1}) - H(X \mid X_{t-k:t-1}, Y_{t-k:t-1}). \tag{B.2}$$

3. Kernel ridge regression (krr), in some literature called "kernelised Geweke" [Amblard et al., 2012b]

$$L^{krr}_{X\to Y} = \log\frac{V\left(\hat{\mathbf{Y}}^{A}_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)}{V\left(\hat{\mathbf{Y}}^{B}_{t_1:t_2} - \mathbf{Y}_{t_1:t_2}\right)}. \tag{B.3}$$

4. Hilbert-Schmidt Normalised Conditional Independence Criterion (HSNCIC)

$$L^{HSNCIC}_{X\to Y} = \|V_{(Y,Z)(X,Z)|Z}\|^2_{HS} \tag{B.4}$$

The estimator of the test statistic $L^{GC}_{X\to Y}$, denoted $\hat{L}^{GC}_{X\to Y}$, is based on a approximations of prediction

error variances ($\hat{V}(\cdot)$ is used for finite approximation), with N - sample size, p - maximum number of lags, d - the dimensionality of the data:

$$\hat{L}^{GC}_{X\to Y} = (N - d - p)\log\left[\frac{\hat{V}_Y[Y, Z; p]}{\hat{V}_Y[X, Y, Z; p]}\right] \sim \chi^2_p \qquad \text{if } \{\mathbf{X}_t\} \nRightarrow \{\mathbf{Y}_t\}. \tag{B.5}$$

The other methods need to use some form of permutation test. When using the permutation test, we will work with the following hypotheses:

$$H_0: \quad L_{X\to Y} = 0, \qquad \text{no causality from } \{X\} \text{ to } \{Y\} \tag{B.6}$$

$$H_1: \quad L_{X\to Y} > 0, \qquad \text{causality from } \{X\} \text{ to } \{Y\} \tag{B.7}$$

If there is no explicitly known distribution for the test statistic, we need to obtain them numerically using a permutation test. Let $\mathrm{p}_i(t)$, $t = t_1, ..., t_2$ denote a random permutation of the time index, and $\mathrm{p}_i(X)$ denote a time series, where the original time order has been reorganised according to the permutation $\mathrm{p}_i(\cdot)$. Then the null hypothesis is assessed by comparing the value of $L_{X\to Y}$ to a histogram of values of $L_{\mathrm{p}_i(X)\to Y}$, and a p-value:

$$\pi(L_{X\to Y} \mid H_0) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left(L_{\mathrm{p}_i(X)\to Y} > L_{X\to Y}\right) \tag{B.8}$$

where $\mathbf{1}(A)$ is a characteristic function for the set A.

Depending on the number of permutations used, we suggest to accept the hypothesis of causality for the level of significance equal to 0.05 or 0.01. In our experiments, we report either single p-values or sets of p-values for overlapping moving windows. The latter is particularly useful when analysing noisy and non-stationary data. In the cases where not much data is available, we do not believe that using any kind of subsampling (as proposed by Sun [2008], Amblard et al. [2012b], Seth and Principe [2011]) will be beneficial, as far as the power of the tests is concerned.

B.2 Testing on simulated data - detecting lag in a linear example

Before applying the methods to real-world data it is prudent to verify whether they work for data with known and simple dependence structure. We tested the methods on a data set containing eight time series with a relatively simple causal structure at different lags and some instantaneous coupling. We used the four methods to try to capture the dependence structure as well as to figure out which lags show dependence. The data was simulated by first generating a set of eight time series from a Gaussian distribution with correlation matrix represented in Table B.1a. Subsequently, some of the series were shifted by one, two or three time steps to obtain the following "causal" relations: $x_1 \longleftrightarrow x_2$ at lag 0 i.e. instantaneous coupling of the two variables, $x_3 \to x_4$ at lag 1, $x_5 \to x_6$ at lag 1, $x_5 \to x_7$ at lag 2, $x_5 \to x_8$

Table B.1: Dependence structure of the simulated data.

(a) Correlation matrix that has been used to generate the testing data.

	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	1	0.7	0.1	0.1	0.1	0.1	0.1	0.1
ts2	0.7	1	0.1	0.1	0.1	0.1	0.1	0.1
ts3	0.1	0.1	1	0.7	0.1	0.1	0.1	0.1
ts4	0.1	0.1	0.7	1	0.1	0.1	0.1	0.1
ts5	0.1	0.1	0.1	0.1	1	0.7	0.7	0.7
ts6	0.1	0.1	0.1	0.1	0.7	1	0.7	0.7
ts7	0.1	0.1	0.1	0.1	0.7	0.7	1	0.7
ts8	0.1	0.1	0.1	0.1	0.7	0.7	0.7	1

(b) Lags at which true dependence occurs, with the interpretation that column variable causes row variable.

	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	×	0						
ts2	0	×						
ts3			×	-1				
ts4			1	×				
ts5					×	-1	-2	-3
ts6					1	×	-1	-2
ts7					2	1	×	-1
ts8					3	2	1	×

at lag 3, $x_6 \to x_7$ at lag 1, $x_6 \to x_8$ at lag 2, $x_7 \to x_8$ at lag 1. The network structure is shown in Figure B.1, while the lags at which the causality occurs are given in the Table B.1b. The length of the data is 250.

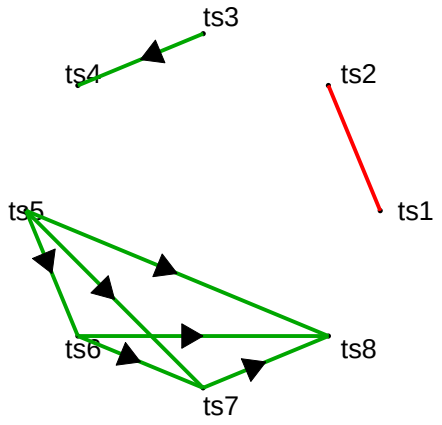


Figure B.1: The directionality of causality between the eight simulated time series. Green lines represent causality with the arrowheads indicating direction; red line indicates instantaneous coupling.

For the purpose of the experiments described in this paper, we used code from several sources: Matlab code that we developed for kernelised Geweke's measure and transfer entropy, open access Matlab toolbox for Granger causality GCCA[1] [Seth, 2010] and open access Matlab code provided by Sohan Seth [Seth and Principe, 2011][2].

To calculate Geweke's measure and kernelised Geweke's measure we used the same code, with a linear kernel in the former case and a Gaussian kernel in the latter. The effect of regularisation on the (linear) Geweke's measure is negligible, and the results are comparable to the ones obtained with GCCA code with the main difference being more flexibility on the choice of lag ranges allowed by our code. Parameters for the ridge regression were either calculated with n-fold cross-validation for the grid of regulariser values in the range of $[2^{-40}, \cdots, 2^{-26}]$ and kernel sizes in the range of $[2^{7}, \cdots, 2^{13}]$, or fixed

[1] The code can be requested from the author's website: `http://www.sussex.ac.uk/Users/anils/aks_code.htm`
[2] Code available at `http://www.sohanseth.com/Home/publication/causmci`

Table B.2: P-values for four measures for lag 1. From top left to bottom right: Geweke's measure (Gc), kernelised Geweke's measure (kG), transfer entropy (TE), HSNCIC (HS). All lag 1 causalities were correctly retrieved by all methods.

Gc	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8	kG	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	×	0.97	0.35	0.26	0.24	0.68	0.11	0.23	ts1	×	0.92	0.30	0.25	0.20	0.74	0.16	0.16
ts2	0.42	×	0.88	0.52	0.37	0.69	0.14	0.46	ts2	0.50	×	0.93	0.54	0.52	0.71	0.19	0.46
ts3	0.26	0.86	×	0.75	0.45	0.19	0.43	0.72	ts3	0.29	0.88	×	0.68	0.48	0.11	0.38	0.62
ts4	0.14	0.11	**0**	×	0.24	0.49	0.41	0.64	ts4	0.12	0.14	**0**	×	0.22	0.47	0.41	0.65
ts5	0.78	0.94	0.10	0.02	×	0.40	0.96	0.91	ts5	0.73	0.93	0.11	0.04	×	0.47	0.99	0.93
ts6	0.96	0.31	0.62	0.22	**0**	×	0.04	1.00	ts6	0.94	0.38	0.55	0.18	**0**	×	0.07	0.99
ts7	0.74	0.98	0.10	0.53	0.35	**0**	×	0.96	ts7	0.81	0.92	0.04	0.55	0.36	**0**	×	0.95
ts8	0.86	0.70	0.05	0.63	0.68	0.87	**0**	×	ts8	0.83	0.67	0.06	0.63	0.62	0.86	**0**	×

TE	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8	HS	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	×	0.59	0.53	0.60	0.34	0.91	0.38	0.66	ts1	×	1.00	0.81	0.35	0.19	0.48	0.71	0.82
ts2	0.48	×	0.86	0.30	0.87	0.96	0.49	0.70	ts2	1.00	×	0.80	0.61	0.85	0.34	0.02	0.72
ts3	0.45	0.17	×	0.33	0.34	0.57	0.81	0.81	ts3	0.90	0.95	×	0.18	0.59	0.47	0.21	0.19
ts4	0.04	0.31	**0**	×	0.12	0.09	0.76	0.08	ts4	0.90	0.29	**0**	×	0.31	0.81	0.26	0.31
ts5	0.21	0.52	0.86	0.05	×	0.68	0.60	0.30	ts5	0.75	0.59	0.77	0.14	×	0.71	0.85	0.46
ts6	0.53	0.89	0.65	0.30	**0**	×	0.77	0.09	ts6	0.64	0.88	0.75	0.79	**0**	×	0.71	0.79
ts7	0.01	0.42	0.59	0.37	0.95	**0**	×	0.77	ts7	0.38	0.13	0.75	0.24	0.75	**0**	×	0.60
ts8	0.85	0.46	0.07	0.48	0.85	0.13	**0**	×	ts8	0.90	0.55	0.46	0.73	0.78	0.78	**0**	×

at a preset level, with no noticeable impact on the result. Transfer entropy utilises a naive histogram to estimate distributions. The code for calculating HSNCIC and for performing p-value tests incorporates a framework written by Seth Seth and Principe [2011]. The framework has been altered to accommodate some new functionalities; the implementation of permutation tests has also been changed from rotation to actual permutation[3]. In the choice of parameters for the HSNCIC we followed Seth and Principe [2011], where the size of the kernel is set up as the median inter-sample distance and regularisation is set to 10^{-3}.

Our goal was to uncover the causal structure without prior information, and detect the lags at which causality occurred. This was performed by applying all three measures of causality with following sets of lags: $\{[1-10]\}, \{[1-20]\}, \{[1-5], [6-10], [11-15]\}, \{[1-3], [4-6], [7-9]\}$ and finally with all four measures to single lags $\{0, 1, 2, 3, 4\}$. Those ranges were used for linear and kernelised Geweke's measures and HSNCIC but not for transfer entropy, for which only single lags are available with the current framework. Using five sets of lags allowed us to analyse the effects of using ranges of lags that are different from lags corresponding to the "true" dynamic of the variables. Table B.2 presents part of the results: p-values for the four measures of interest for lag 1. Below we present the conclusions for each of the methods separately, with two Geweke's measures presented together:

Geweke's measures. Both Geweke's measures performed similarly, which was expected as the data was simulated with linear dependencies. Causalities were correctly identified for all ranges of lags, for which the causal direction existed, including the biggest range [1-20]. For the shorter ranges $\{[1-5], [1-3]\}$ as well as for the single lags $\{0, 1, 2, 3\}$ the two measures reported p-values of 0 for all of the existing causal directions. This means that the measures were able to detect precisely the lags at which causal directions existed, including the lag 0, i.e. instantaneous coupling. However, with number of permutations equal 200 and at acceptance level of 0.01, the two measures detected only the required causalities, but would fail to reject some spurious causalities at level of 0.05.

[3]Use of permutation, which is more general than rotation, is helpful when data is short or the analysed time windows are short.

Transfer entropy. By design, this measure can only analyse one lag at a time. It is also inherently slow, and for these two reasons it will be inefficient when a wide range of lags needs to be considered. Furthermore, it cannot be used for instantaneous coupling. In order to detect this we applied the mutual information method instead. For the lags $\{1, 2, 3\}$ transfer entropy reported 0 p-values for all the relevant causal directions. However, it failed to reject spurious direction $1 \rightarrow 7$ with p-value of 0.01. For lag $\{0\}$ where mutual information has been applied, the instantaneous coupling $x_1 \longleftrightarrow x_2$ was recognised correctly with p-value 0.

HSNCIC. Due to slowness, HSNCIC is impractical for the largest ranges of lags. More importantly, HSNCIC performs unsatisfactorily for any of the ranges of lags that contained more than a single lag. This is deeply disappointing, as the design suggests HSNCIC should be able to handle both side information and higher dimensional variables. Even for a small range $[1 - 3]$ HSNCIC correctly recognised only the $x_5 \rightarrow x_8$ causality. Nevertheless, it did recognise all of the causalities correctly when analysing one lag at a time, reporting p-values of 0. This suggests that HSNCIC is unreliable for data that has more than one lag or more than two time series. HSNCIC is also not designed to detect instantaneous coupling.

From this experiment we conclude that Geweke's measures with linear and Gaussian kernels provide the best performance, are not vulnerable to lag misspecification and seem the most practical. The other two measures, transfer entropy and HSNCIC, provide good performance when analysing one lag at a time. In Section B.3 we show the results of one of the tests from Amblard et al. [2012b], which investigates ability to distinguish between direct and non-direct causality in data where both linear and non-linear dependence have been introduced. We refer to Seth and Principe [2011] for results of a wide range of tests applied to linear Granger causality and HSNCIC. We tested all four methods and managed to reproduced the results from Seth and Principe [2011] to a large degree, however we used smaller number of permutations and realisations and we obtained somewhat lower acceptance rates for true causal directions, particularly for HSNCIC. From all of those tests we conclude that linear causality can be detected by all measures in most cases, with the exception of HSNCIC when more lags or dimensions are present. Granger causality can detect some nonlinear causalities, especially if they can be approximated by linear functions. Transfer entropy will flag more spurious causalities in the case where causal effects exist for different lags. There is no maximum dimensionality that HSNCIC can accept; in some experiments this measure performed well for three and four dimensional problems, in others three dimensions proved to be too many.

Possibly the most important conclusion is that parameter selection turned out to be critical for kernelised Geweke's measure. For some tests, like the simulated 8 time series data described earlier, size of the kernel did not play an important role, but in some cases the size of the kernel was crucial in allowing the detection of causality. However, there was no kernel size that worked for all of the types of the data.

B.3 Testing on simulated data - nonlinear multivariate example

Our second example follows one presented by Amblard [Amblard et al., 2012b] and involves a system with both linear and non-linear causality. Apart from presenting the benefits of generalising Granger causality, this example demonstrates the potential effect of considering side information on distinguishing direct and indirect cause. The true dynamic of the time series is as follows:

$$\begin{cases} X_t = aX_{t-1} + \epsilon_{X,t} \\ Y_t = bY_{t-1} + dX_{t-1}^2 + \epsilon_{Y,t} \\ Z_t = cZ_{t-1} + eY_{t-1} + \epsilon_{Z,t} \end{cases} \tag{B.9}$$

where the parameters were chosen in the following way: $a = 0.2, b = 0.5, c = 0.8, d = 0.8, e = 0.7$, the variables $\epsilon_{x,t}, \epsilon_{y,t}, \epsilon_{z,t}$ are i.i.d. Gaussian with zero mean and unit variance. From the setup we know that we have the following causal chain $x \to y \to z$ (with nonlinear effect of x on y) and therefore there is an indirect causality $x \to z$. We calculate kernelised Geweke measures $L^{krr}_{X\to Z}$ and $L^{krr}_{X\to Z|Y}$ to assess the causality.

We repeat the experiment 500 times, each time generating a time series of length 500. We choose an embedding of 2, i.e. we consider the lag range $[1 - 2]$. To evaluate the effect of using kernelised rather than linear Granger causality, we run each experiment for the Gaussian kernel and for linear kernel $k(x, y) = x^T y$. Using the linear kernel is nearly equivalent to use the linear Geweke measures. We obtain a set of 500 measurements for $L^{krr}_{X\to Z}$ and $L^{krr}_{X\to Z|Y}$, each run with a Gaussian and with a linear kernel. The results are shown in Figures B.2, B.3 and B.4. As expected, $L^{krr}_{X\to Z|Y}$ does not detect any causality regardless of the kernel chosen. When no side information is taken into consideration we should see the indirect causality $x \to z$ being picked up, however this is the case only for kernelised Geweke with Gaussian kernel and for HSNCIC. As the dependence was nonlinear, the linear Geweke's measure did not detect it.

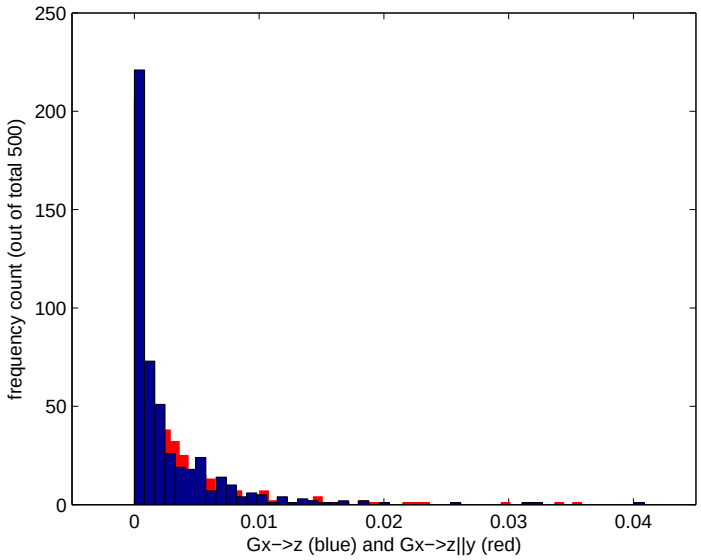


Figure B.2: Histogram of the measurements $L^{krr}_{X\to Z}$ (red face),$L^{krr}_{X\to Z|Y}$ (blue face) calculated with the kernelised Geweke's using the linear kernel (i.e. equivalent of Granger causality).

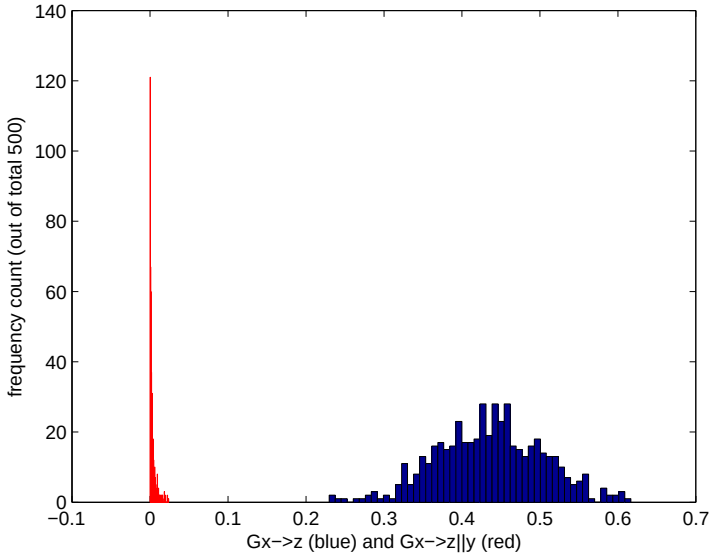


Figure B.3: Histogram of the measurements $L^{krr}_{X\to Z}$ (red face),$L^{krr}_{X\to Z|Y}$ (blue face) calculated with the kernelised Geweke's using the Gaussian kernel.

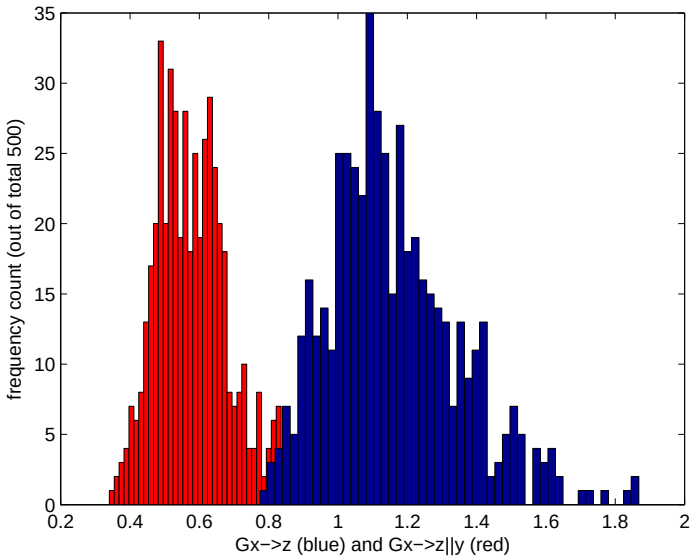


Figure B.4: Histogram of the measurements $L^{HSNCIC}_{X\to Z}$ (red face),$L^{HSNCIC}_{X\to Z|Y}$ (blue face) calculated with the HSNCIC.

Transfer entropy, as defined in this paper, does not allow side information and therefore the result we achieve is a distribution that appear significantly different from zero (fig B.5).

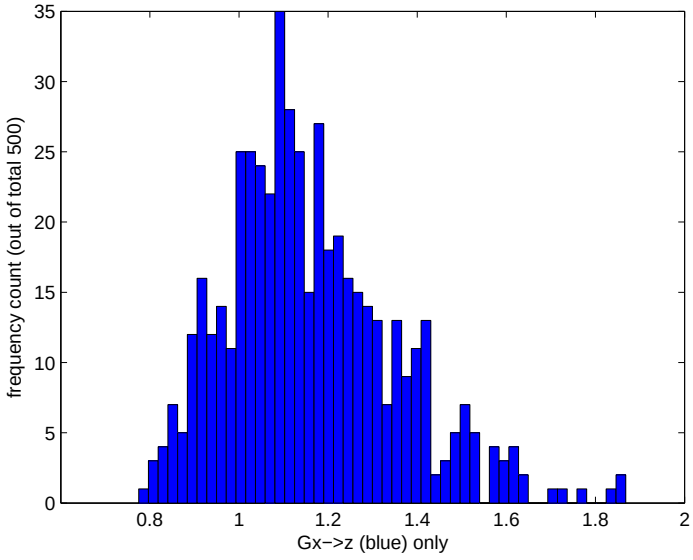


Figure B.5: Histogram of the measurements $L^{TE}_{X\to Z}$ (red face),$L^{TE}_{X\to Z|Y}$ (blue face) calculated with the transfer entropy.

B.4 Applications

Granger causality was introduced as an econometrics concept, and for many years, it was mainly used in economic applications. After around 30 years of relatively little acknowledgement, the concept of causality

started to gain significance in a number of scientific disciplines. Granger causality and its generalisations and alternative formulations became popular, particularly in the field of neuroscience, but also climatology and physiology Hlaváčková-Schindler et al. [2007], Amblard and Michel [2011], Chávez et al. [2003], Knuth et al. [2005], Gourévitch and Eggermont [2007], Vicente et al. [2011]. The methodology was successfully applied in those fields, particularly in neuroscience, due to the characteristics of the data common in those fields and the fact that the assumptions of Gaussian distribution and/or linear dependence are often reasonable [Bressler and Seth, 2011]. This is generally not the case for financial time series.

B.4.1 Applications to Finance and Economics

In finance and economics, there are many tools devoted to modelling dependence, mostly for symmetrical dependence, such as correlation/covariance, cointegration, copula and, to a lesser degree, mutual information Alexander and Wyeth [1994], Cont [2005], Patton [2009], Durante [2013]. However, in various situations where we would like to reduce the dimensionality of a problem (e.g., choose a subset of instruments to invest in, choose a subset of variable for a factor model, *etc*.), knowledge of the causality structure can help in choosing the most relevant dimensions. Furthermore, forecasting using the causal time series (or Bayesian priors in Bayesian models or parents in graphical models [Pearl, 2000, Barber, 2012]) helps to forecast "future rather than the past".

Financial data often have different characteristics than data most commonly analysed in biology, physics, *etc*. In finance, the typical situation is that the researcher has only one long, multivariate time series at her disposal, while in biology, even though the experiments might be expensive, most likely, there will be a number of them, and usually, they can be reasonably assumed to be independent identically distributed (i.i.d.). The assumption of linear dependencies or Gaussian distributions, often argued to be reasonable in disciplines, such as neuroscience, are commonly thought to be invalid for financial time series. Furthermore, many researchers point out that stationarity usually does not apply to this kind of data. As causality methods in most cases assume stationarity, the relaxation of this requirement is clearly an important direction for future research. In the sections below, we describe the results of applying causal methods to two sets of financial data.

B.4.2 Interest Rates and Inflation

Interest rates and inflation have been investigated by economists for a long time. There is considerable research concerning the relationship between inflation and nominal or real interest rates for the same country or region, some utilising tools of Granger causality (for example, Eichler [2007]).

In this experiment, we analyse related values, namely the consumer price index for the United States (U.S. CPI) and the London Interbank Offered Rate (Libor) interest rate index. Libor is often used as a base rate (benchmark) by banks and other financial institutions, and it is an important economic indicator. It is not a monetary measure associated with any country, and it does not reflect any institutional mandate

in contrast to, for example, when the Federal Reserve sets interest rates. Instead, it reflects some level of assessment of risk by the banks who set the rate. Therefore, we ask whether we detect that one of these two economic indicators causes the other one in a statistical sense?

We ran our analysis for monthly data from January 31, 1986, to October 31, 2013, obtained from Thomson Reuters. The implementation and parameter values used for this analysis were similar to those in the simulated example (Section 2.2.2). We used kernelised Geweke's measure with linear and Gaussian kernels. Parameters for the ridge regression were at a preset level in the range of $[2^7, \cdots, 2^{13}]$ or as a median.

We investigated time-windows of size 25, 50, 100 and 250. The most statistically significant and interpretable results were observed for the longer windows (250 points), where Geweke's measure and kernelised Geweke's measure show a clear indication of the direction U.S. CPI $\rightarrow$ Libor. For shorter windows of time, significant p-values were obtained considerably less often, but the results were consistent with the results for the longer time window. The dependence for the 250 day window were seen most strongly for Lag 1 (i.e., one month) and less strongly for Lags 2, 7, 8, 9, but there is no clear direction for the interim lags. In Figures B.6–B.9, we report p-values for the assessment of causality for Lags 1, 2 and 7 alongside the scatter plot showing p-values and the values of Geweke's measure. All of the charts have been scaled to show p-values in the same range [0,1]. We can clearly see the general trend that the higher the values of causality, the lower their corresponding p-values.

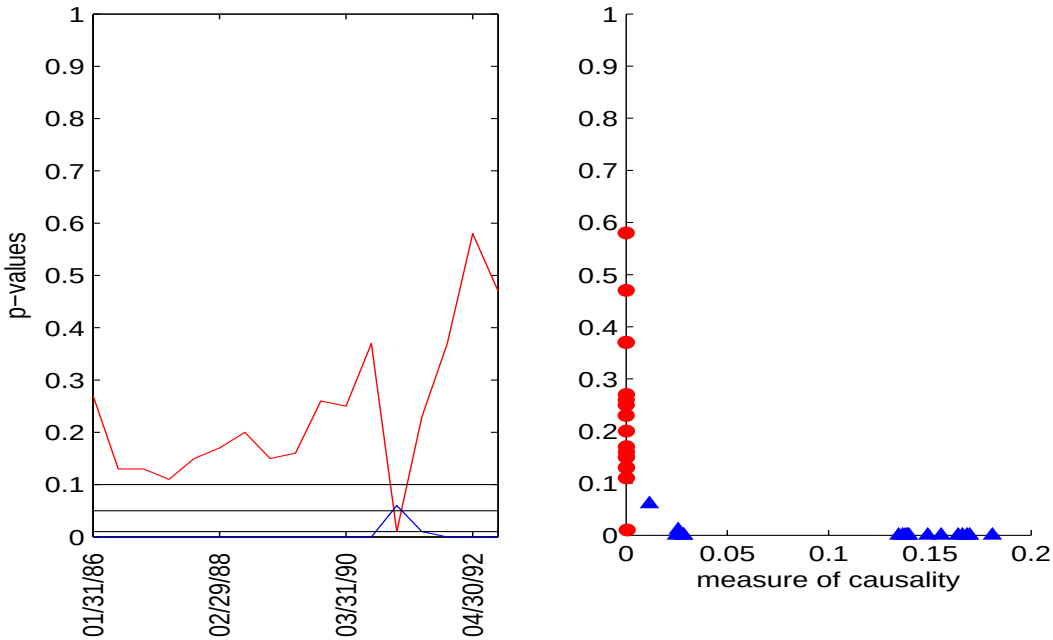


Figure B.6: Kernelised Geweke's measure of causality. The left chart shows sets of p-values for the hypothesis that inflation statistically causes Libor (blue line) or the other way round (red line), when a model with one lag is considered. The right chart shows the scatter plot of p-values and the value of the causality measure.

In Figure B.6, we observe that the U.S. CPI time series lagged by one month causes one-month Libor in a statistical sense, when assessed with kernelised Geweke's measure with Gaussian kernel. The p-values for the hypothesis of causality in this direction allow us to accept (not reject) this hypothesis at a significance level of 0.01 in most cases, with the p-values nearly zero most of the time. We can also observe that several of the causality measurements are as high as 0.2, which can be translated to an improvement of roughly 0.18 in the explanatory power of the model citenote6. Applying the linear kernel (Figure B.7) resulted in somewhat similar patterns of measures of causality and p-values, but the two directions were less separated. Interest rates causing Libor still have p-values at zero most of the time, but the other direction has p-values that fall below the 0.1 level for several consecutive windows at the beginning, with the improvement in the explanatory power of the model at a maximum 0.07 level; our interpretation is that the causality is nonlinear.

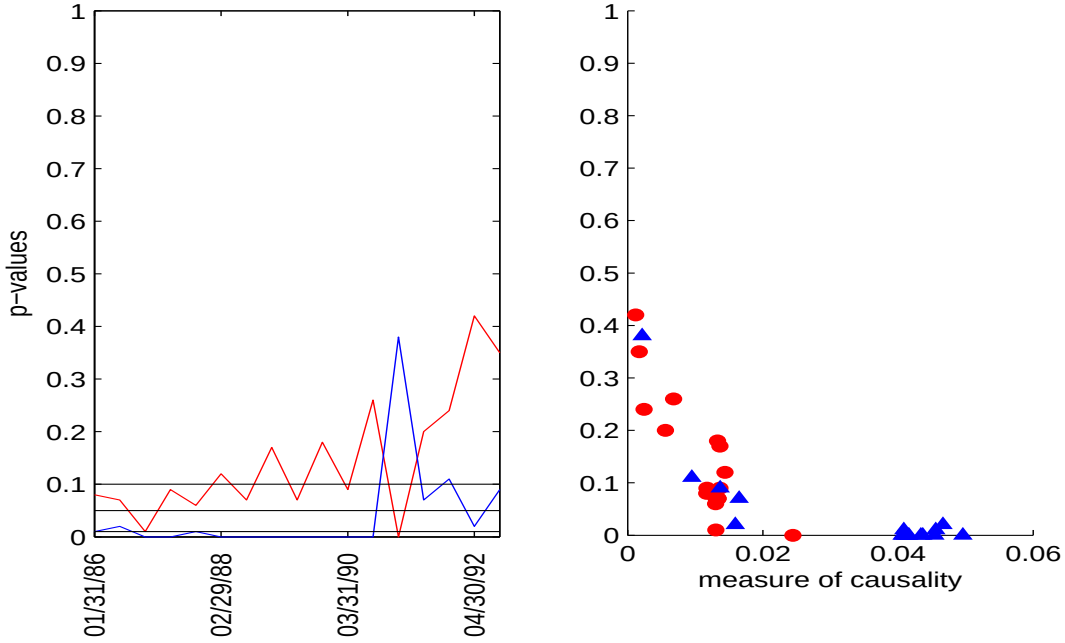


Figure B.7: Linear Geweke's measure of causality. (**Left**) Sets of p-values for the hypothesis of statistical causality in the direction U.S. consumer price index → one-month Libor (blue line) or the other way round (red line), when a model with a linear kernel and Lag 1 is considered. (**Right**) Scatter plot of p-value and value of the causality measure.

The results for the second lag, given in Figure B.8, are no longer as clear as for Lag 1 in Figure B.6 (Gaussian kernel in both cases). The hypothesis of inflation causing interest rates still has p-values close to zero most of the time, but the p-values for the other direction are also small. This time, the values of causality are lower and reach up to just below 0.08. Using a linear kernel, we obtain less clear results, and our interpretation is that the causal direction CPI → Libor is stronger, but there might be some feedback, as well.

Figure B.9 presents the results of using a linear kernel, which shows a much better separation of

the two directions, applied to the model with Lag 7. Very similar results can be seen for models with Lags 8 and 9. There is no obvious reason why the linear kernel performed much better than the Gaussian kernel for these large lags. We offer the interpretation that no nonlinear causality was strong enough and consistent enough and that this was further obscured by using a nonlinear kernel. The conclusion here is that model selection is an important aspect of detecting causality and needs further research.

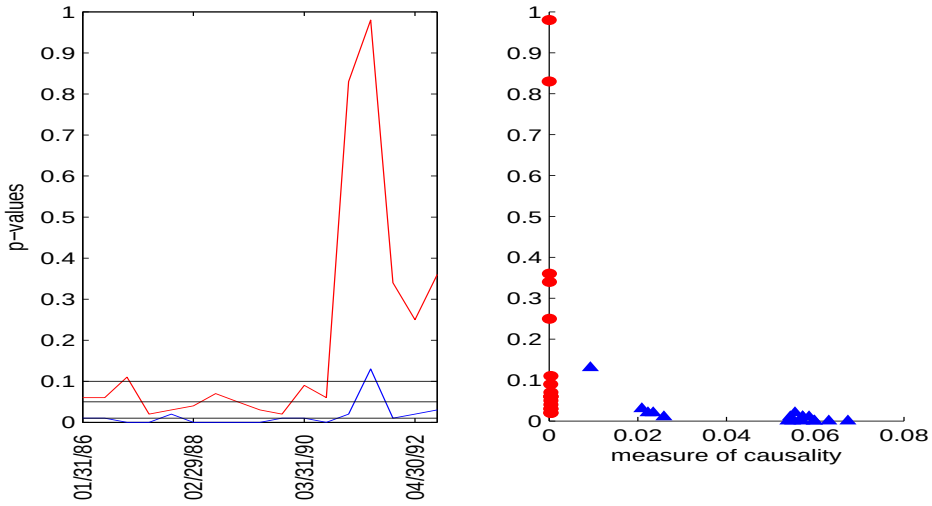


Figure B.8: Kernelised Geweke's measure of causality. (**Left**) Sets of p-values for the hypothesis of statistical causality in the direction U.S. CPI → one-month Libor (blue line) or the other way round (red line), when the model with the Gaussian kernel and Lag 2 is considered. (**Right**) Scatter plot of the p-value and the value of the causality measure.

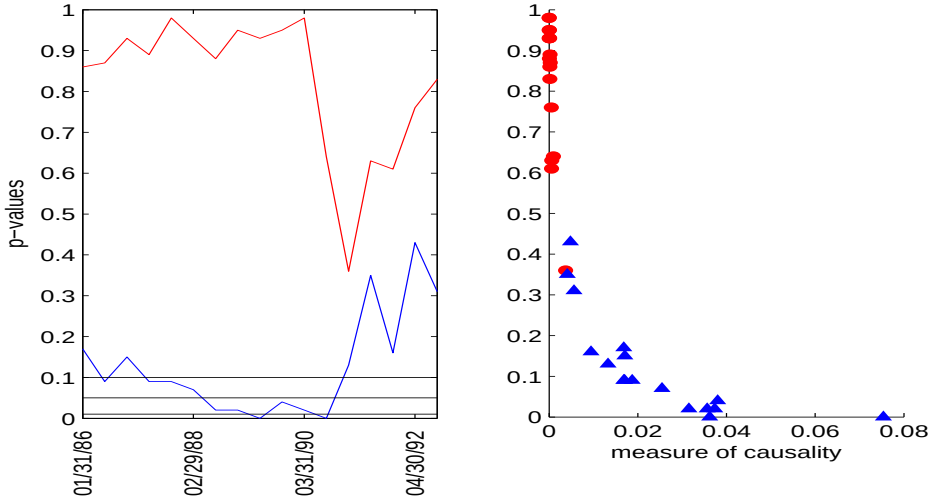


Figure B.9: Linear Geweke's measure of causality. (**Left**) Sets of p-values for the hypothesis of statistical causality in the direction U.S. CPI → one-month Libor (blue line) or the other way round (red line), when model with a linear kernel and Lag 7 is considered. (**Right**) Scatter plot of the p-value and the value of the causality measure.

In our analysis, we did not obtain significant results for transfer entropy or HSNCIC. The results for Lag 1 for transfer entropy and HSNCIC are shown in Figures B.10 and B.11, respectively. For Lag

1, there was a significant statistical causality in the direction U.S. CPI → one-month Libor supported by both Geweke's measures. This is barely seen for transfer entropy and HSNCIC. *p*-values for transfer entropy are at a level that only slightly departs from a random effect, and for HSNCIC, they are often significant; however, the two directions are not well separated. The results for higher lags were often even more difficult to interpret. We must stress that the different implementation of transfer entropy and parameter choice for HSNCIC might result in better performance (please refer to Sections B.5.0.1 and B.5.1).

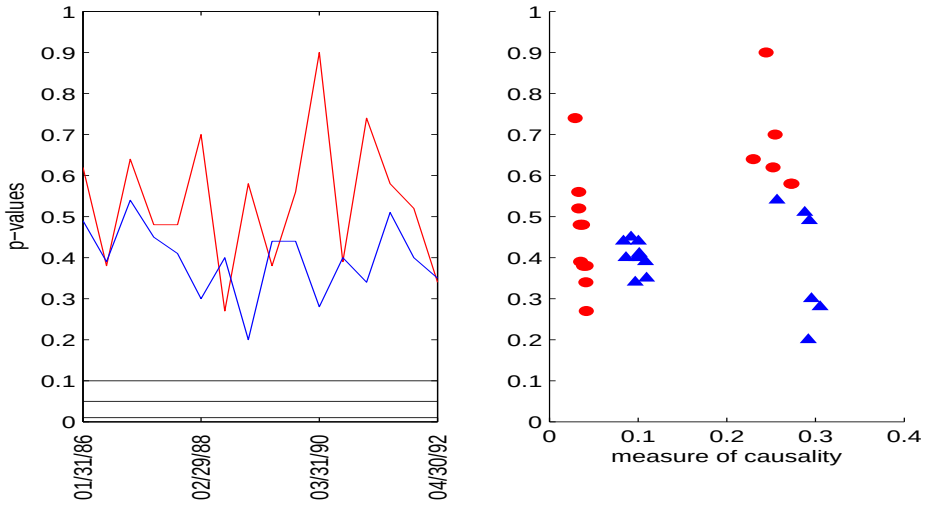


Figure B.10: Transfer entropy. (**Left**) sets of *p*-values for the hypothesis of statistical causality in the direction U.S. CPI → one-month Libor (blue line) or the other way round (red line), when Lag 1 is considered. (**Right**) Scatter plot of the *p*-value and the value of the causality measure.

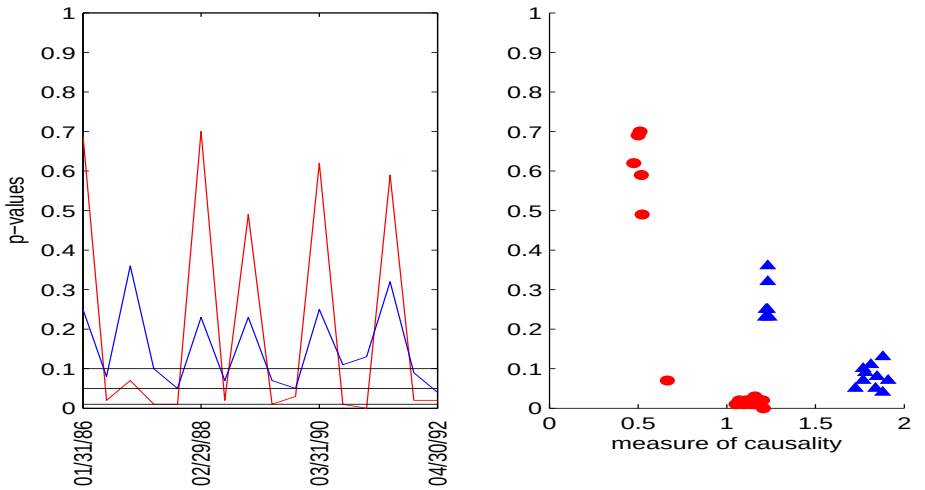


Figure B.11: HSNCIC. (**Left**) sets of *p*-values for the hypothesis of statistical causality in the direction U.S. CPI → one-month Libor (blue line) or the other way round (red line), when Lag 1 is considered. (**Right**) Scatter plot of the *p*-value and the value of the causality measure.

B.4.3 Equity versus Carry Trade Currency Pairs

We analysed six exchange rates (AUDJPY, CADJPY, NZDJPY, AUDCHF, CADCHF, NZDCHF and the index S&P) and investigated any patterns of the type "leader - follower". Our expectation was that S&P should be leading. We used daily data for the period July 18, 2008–October 18, 2013, from Thomson Reuters. We studied the pairwise dependence between the currencies and S&P, and we also analysed the results of adding the Chicago Board Options Exchange Market Volatility Index (VIX) as side information. In all of the cases, we used logarithmic returns.

Figure B.12 presents the results of applying kernelised Geweke's measure with a Gaussian kernel. The plots show series of p-values for a moving window of a length of 250 data points (days), moving each window by 25 points. Unlike in the previous case of interest rates and inflation, there is little actual difference between the linear and Gaussian kernel methods. However, in a few cases, employing a Gaussian kernel results in better separation of the two directions, especially CADCHF → S&P and S&P → CADCHF given VIX.

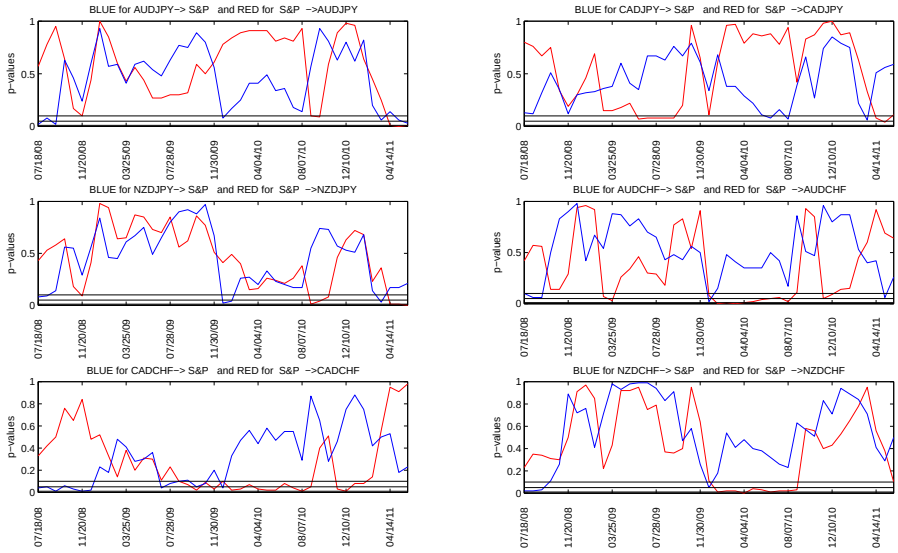


Figure B.12: Sets of p-values for the hypothesis that an exchange rate causes the equity index, S&P (blue), or the other way round (red).

Excepting CADCHF, all currency pairs exhibit similar behaviour when analysed for the causal effect on the S&P. This behaviour consists of a small number of windows for which a causal relationship is significant at a p-value below 0.1, but that does not persist. CADCHF is the only currency with a consistently significant causal effect on S&P, which is indicated for periods starting in 2008 and 2009. As for the other direction, for AUDCHF, CADCHF and NZDCHF, there are periods where S&P has a significant effect on them as measured by p-values.

Figure B.13 shows similar information as in Figure B.12, but taking into consideration VIX as side

information. The rationale is that the causal effect of S&P on the carry trade currencies is likely to be connected to the level of perceived market risk. However, the charts do not show the disappearance of a causal effect after including VIX. While the patterns do not change considerably, we observe that exchange rates have lost most of their explanatory power for S&P, with the biggest differences for CADCHF. There is little difference for the p-values for the other direction; hence, the distinction between the two directions became more significant.

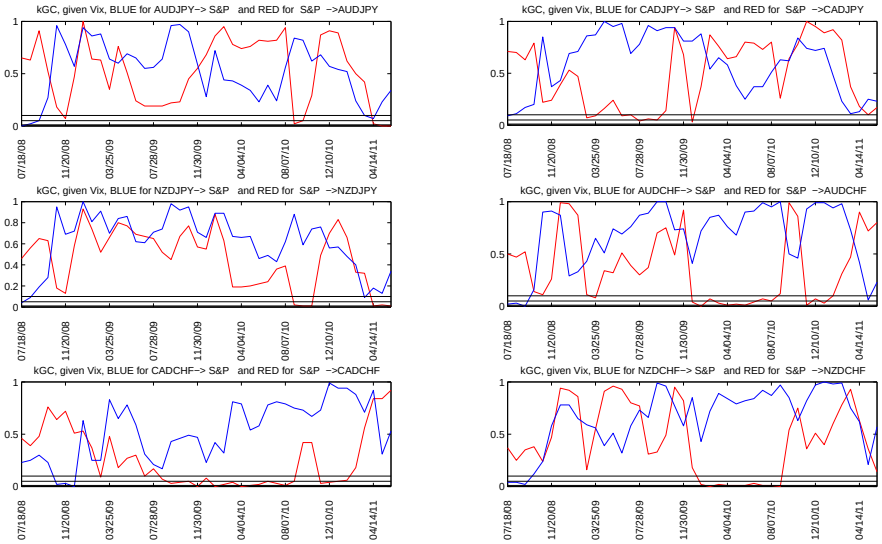


Figure B.13: Sets of p-values for the hypothesis that an exchange rate causes the equity index, S&P, given the Volatility Index (VIX) as side information (blue) or the other way round (red).

We obtained all of the main "regimes": periods when either one of the exchange rates or S&P had more explanatory power (p-values for one direction were much lower than for the other) and periods when both exhibited low or both exhibited high p-values. p-values close to one did not necessarily mean purely a lack of causality: in such cases, the random permutations of the time series tested for causality at a specific lag appear to have higher explanatory power than the time series at this lag itself. There are a few possible explanations related to the data, the measures and to the nature of the permutation test itself. We observed on the simulated data that when no causality is present, autocorrelation introduces biases to the permutation test: higher p-values than we would expect from a randomised sample, but also the higher likelihood of interpreting correlation as causality. Furthermore, both of these biases can result from assuming a model with a lag different from that of the data. Correspondingly, if the data has been simulated with instantaneous coupling and no causality, this again can result in high p-values. Out of all four methods, transfer entropy appeared to be most prone to these biases.

B.5 Discussion

While questions about causal relations are asked often in science in general, the appropriate methods of quantifying causality for different contexts are not well developed. Firstly, often answers are formulated with methods not intended specifically for this purpose. There are fields of science, for example nutritional epidemiology, where causation is commonly inferred from correlation. A classical example from economics, known as "Milton Friedman's thermostat", describes how lack of correlation is often confused with lack of causation in the context of the evaluation of the Federal Reserve Friedman [2003]. Secondly, often questions are formulated in terms of (symmetrical) dependence because it involves established methods and allows clear interpretation. This could be a case in many risk management applications where the question of what causes losses should be central but is not commonly addressed with causal methods. The tools for quantifying causality that are currently being developed can help to better quantify causal inference and better understand the results.

In this section we provide a critique of the methods to help understand their weaknesses and enable the reader to choose the most appropriate method for intended use. This will also set out possible directions of future research. The first part of this section describes the main differences between the methods, followed by a few comments on model selection and problems related to permutation testing. Suggestions of future research directions conclude the section.

B.5.0.1 Theoretical differences

Linearity versus nonlinearity. The original Granger causality and its Geweke's measure formulation were developed to assess linear causality and they are very robust and efficient in doing so. For data with linear dependence using linear Granger causality is most likely to be the best choice. The measure can work well also in cases where the dependence is not linear but has a strong linear component.

As financial data does not normally exhibit stationarity, linearity or Gaussianity, linear methods should arguably not be used to analyse them. In practice, requirements on the size of the data sets and difficulties with model selection take precedence and mean that linear methods should still be considered.

Direct and indirect causality. Granger causality is not transitive, which might be unintuitive. Although transitivity would bring the causality measure closer to the common understanding of the term, it could also make it impossible to distinguish between direct and indirect cause. As a consequence it could make the measure useless for the purpose of reduction of dimensionality and repeated information. However, differentiation between direct and indirect causality is not necessarily well defined. This is because adding a conditioning variable can both introduce as well as remove dependence between variables Hsiao [1982]. Hence the notion of direct and indirect causality is relative to the whole information system and can change if we add new variables to the system. Using methods from graphical modelling [Pearl, 2000] could facilitate defining the concepts of direct and indirect causality, as these two terms are well

defined for causal networks.

Geweke's and kernelised Geweke's measures can distinguish direct and indirect cause in some cases. Following the example of Amblard [Amblard et al., 2012b] we suggest comparing the conditional and non-conditional causality measurements as means of distinguishing between direct and indirect cause for both linear and kernel Granger causality. Measures like HSNCIC are explicitly built in such a way that they condition on side information and therefore are geared towards picking up only the direct cause; however, this does not work as intended as we noticed that HSNCIC is extremely sensitive to the dimensionality of the data. Transfer entropy – in the form we are using – does not consider side information at all. A new measure, called partial transfer entropy Papana et al. [2013], Kugiumtzis [2013] has been proposed to distinguish between direct and indirect cause.

Spurious causality. Partially covered in the previous point about direct and indirect cause, the problem of spurious causality is a wider one. As already indicated, causality is inferred in relation to given data and introducing more data can both add and remove (spurious) causalities. The additional problem is that data can exhibit many types of dependency. None of the methods we discuss in this paper is capable of managing several simultaneous types of dependency, be it instantaneous coupling, linear or nonlinear causality. We refer the interested reader to literature on modelling Granger causality and transfer entropy in the frequency domain or using filters Seth [2010], Lungarella et al. [2007b], Dhamala et al. [2008].

Numerical estimator. It was already mentioned that Granger causality and kernel Granger causality are robust for small samples and high dimensionality. Both of those measures optimise quadratic cost, which means they can be sensitive to outliers, but kernelised Geweke's measure can somewhat mitigate this with parameter selection. Granger causality for bivariate data has good statistical tests for significance, while the others do not and need permutation tests which are computationally expensive. Also, in the case of ridge regression, there is another layer of optimising parameters which is also computationally expensive. Calculating kernels is also relatively computationally expensive (unless the data is high-dimensional), but they are robust for small samples.

The HSNCIC is shown to have a good estimator which in the limit of infinite data does not depend on the type of kernel. Transfer entropy, on the other hand, suffers from issues connected to estimating a distribution: problems with small sample size and high dimensionality. Choosing the right estimator can help reduce the problem. A detailed overview of possible methods of estimation of entropy can be found in Hlaváčková-Schindler et al. [2007]. Trentool, one of more popular open access toolboxes for calculating transfer entropy, uses a nearest neighbour technique to estimate joint and marginal probabilities, that has been first proposed by Kraskov et.al. Kraskov et al. [2004], Lindner et al. [2011], Vicente et al. [2011]. The nearest neighbour technique is data efficient, adaptive and has minimal bias Hlaváčková-Schindler et al. [2007]. The important aspect of this approach is that it depends on a correct choice of embedding

parameter and therefore does not allow analysing the information transfer for arbitrary lags. It also involves additional computational cost and might be slower for low dimensional data. We tested Trentool on several data sets and found that the demands on the size of the sample were higher than for the naive histogram and the calculations were slower, with comparable results. The naive histogram however does not have good performance for higher dimensions Hlaváčková-Schindler et al. [2007], in which case the nearest neighbour approach would be advised.

Non-stationarity. This is one of the most important areas for future research. All of the described measures suffer to some degree from an inability to deal with non-stationary data. Granger causality in the original, linear formulation, is the only measure that explicitly assumes stationarity (more precisely, covariance stationarity Granger [1969], Geweke [1984b]) and the asymptotic theory is developed for that case. Geweke describes in Geweke [1984a] special cases of non-stationary processes that can still be analysed within the standard framework and corresponding literature on adapting the linear Granger causality framework to the case of integrated or cointegrated processes Toda and Yamamoto [1995]. In all of those cases the type of non-stationarity needs to be known and that is a potential source of new biases Toda and Yamamoto [1995]. The GCCA toolbox[4] for calculating Granger causality provides some tools for detecting nonstationarity and to a limited degree also for managing it [29]. In the vector autoregressive setting of Granger causality it is possible to run parametric tests to detect nonstationarity: ADF test (Augmented Dickey Fuller) and KPSS test (Kwiatkowski, Phillips, Schmidt, Shin). For managing non-stationarity GCCA toolbox manual Seth [2010] suggests analysing shorter time series (windowing) and differencing, although both approaches can introduce new problems. It is also advisable to detrend and demean the data, and in the case of economic data it might also be possible to perform seasonal adjustment.

The other measures described in this article do not explicitly assume stationarity, however some assumptions about stationarity are necessary for the methods to work correctly. Schreiber developed transfer entropy under the assumption that analysed system can be approximated by stationary Markov processes [Schreiber, 2000]. Transfer entropy in practice can be affected if the time series is highly nonstationary, as the reliability of the estimation of probability densities will be biased Vicente et al. [2011], but non-stationarity due to slow change of parameters does not have to be a problem Gómez-Herrero et al. [2015]. For the other two methods, kernelised Geweke's measure and HSNCIC, the results for estimator convergence are available only for stationary data, according to our knowledge. However, the asymptotic results for HSNCIC have been developed for the too restrictive case of i.i.d. data Fukumizu et al. [2008][5]. The results for kernel ridge reression given by Hang and Steinwart [2014] have been developed for alpha-mixing data.

[4]Code can be requested at: `http://www.sussex.ac.uk/Users/anils/aks_code.htm`

[5]We believe that the generalisation from i.i.d. data to alpha-mixing can be done similarly as for the HSICChwialkowski and Gretton [2014]

Choice of parameters. Each of the methods requires parameter selection – an issue related to model selection described in Section B.5.1. All of the methods need a choice of the number of lags (lag order), while kernel methods additionally require choice of kernel, kernel parameter (kernel size) and regularisation parameter.

In the case of the Gaussian kernel, the effect of the kernel size on the smoothing of the data can be understood as follows Fukumizu [2007], Shawe-Taylor et al. [2004]. The Gaussian kernel $k(x, y) = exp(-\|x - y\|^2/\sigma^2)$ corresponds to an infinite dimensional feature map consisting of all possible monomials of input features. If we express a kernel as a Taylor series expansions, using the basis $1, u, u^2, u^3, ...$ the random variables X and Y can be expressed in RKHS by:

$$\begin{aligned} \Phi(X) = k(X, \cdot) &\sim (1, c_1 X, c_2 X^2, c_3 X^3, ...)^T \\ \Phi(Y) = k(Y, \cdot) &\sim (1, c_1 Y, c_2 Y^2, c_3 Y^3, ...)^T, \end{aligned} \tag{B.10}$$

therefore the kernel function can be expressed as follows:

$$k(x, y) = 1 + c_1 xy + c_2 x^2 y^2 + c_3 x^3 y^3 + ... \tag{B.11}$$

and the cross-covariance matrix will contain information on all of the higher-order covariances:

$$\Sigma_{XY} \sim \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & c_1^2 Cov[X, Y] & c_1 c_2 Cov[X, Y^2] & c_1 c_3 Cov[X, Y^3] & ... \\ 0 & c_2 c_1 Cov[X^2, Y] & c_2^2 Cov[X^2, Y^2] & c_2 c_3 Cov[X^2, Y^3] & ... \\ 0 & c_3 c_1 Cov[X^3, Y] & c_3 c_2 Cov[X^3, Y^2] & c_3^2 Cov[X^3, Y^3] & ... \\ 0 & ... & ... & ... & ... \end{pmatrix} \tag{B.12}$$

According to Fukumizu et al. [2008] the HSNCIC measure does not depend on the kernel in the limit of infinite data. However, the other parameters still need to be chosen, which is clearly a drawback. Kernelised Geweke's measure optimises parameters explicitly with the cross-validation while HSNCIC focuses on embedding the distribution in RKHS with any characteristic kernel. Additionally, transfer entropy requires the choice of method for estimating densities and the binning size in the case of naive histogram approach.

Another important aspect is the choice of lag order and the number of lags. We observed in Section B.2 that the two Geweke's measures were not sensitive to the choice of lags and we were able to correctly recognise causality both in the case of smaller and bigger lag ranges used. The two other measures however behaved differently. HSNCIC is often not able to observe causality in the case of more lags analysed at the time, but performed well for single lags. Transfer entropy flagged spurious causality in

Measures	Properties
	Linearity versus nonlinearity
Granger causality	assumes linearity; the best method for linear data, the worst for nonlinear
kernelised Geweke's	works for both linear and nonlinear data
transfer entropy	works for both linear and nonlinear data
HSNCIC	works for both linear and nonlinear data if low dimension
	Distinguishing direct from indirect causality
Granger causality	to some extent by comparing measure with and without side information
kernelised Geweke's	to some extent by comparing measure with and without side information
transfer entropy	not able to (consider partial transfer entropy)
HSNCIC	to some extent, as it is designed to condition on side information
	Spurious casuality
Granger causality	susceptible
kernelised Geweke's	susceptible
transfer entropy	susceptible
HSNCIC	susceptible
	Good numerical estimator
Granger causality	yes
kernelised Geweke's	yes
transfer entropy	no
HSNCIC	yes
	Nonstationarity
Granger causality	v. sensitive; test with ADF, KPSS, use windowing, differencing, large lag
kernelised Geweke's	somewhat sensitive; online learning is a promising approach
transfer entropy	somewhat sensitive
HSNCIC	somewhat sensitive
	Choice of parameters
Granger causality	lag
kernelised Geweke's	kernel, kernel size, regularisation parameter, lag; uses cross-validation
transfer entropy	lag, binning size (if histogram approach used)
HSNCIC	kernel, kernel size, regularisation parameter, lag

Table B.3: The summary of main features of the different measures

one case where lag was far from the "true" one. However for real data, with more complex structure, the choice of lag is likely to be important for all measures (see Section B.5.1).

B.5.1 Model selection

For the kernel measures we observed that model selection was an important issue. In general, the choice of kernel influences the smoothness of the class of functions considered, while the choice of regulariser controls the trade-off between smoothness of the function and the error of the fit. Underfitting can be a consequence of a too large regulariser and a too large kernel size (in case of a Gaussian kernel); conversely, overfitting can be a consequence of a too small regulariser and a too small kernel size. One of the methods suggested to help with model selection is cross-validation [Amblard et al., 2012b]. This method is particularly popular and convenient for the selection of kernel size and regulariser in the ridge regression. Given nonstationary data it would seem reasonable to fit the parameters; however, we concluded that cross-validation was too expensive in the computational sense and did not provide the expected benefits.

Another aspect of model selection (and choice of parameters) is the determination of an appropriate lag order. For kernel methods increasing the number of lags does not increase the dimensionality of the problem as could be expected in case of the methods representing the data explicitly. In the case of kernelised Geweke's measure, increasing the number of lags decreases the dimensionality of the problem, due to the fact that the data is represented in terms of $(n-p)\times(n-p)$ pairwise comparisons, where n is the number of observations and p – the number of lags. On the other hand, increasing the number of lags will decrease the number of degrees of freedom. This decrease will be less pronounced for kernel methods which allocate smaller weights to higher lags (as is the case in Gaussian kernel, but not for the linear kernel). Apart from cross-validation the other approaches to choosing the lag order suggested in the literature are based on the analysis of the autocorrelation function or partial autocorrelation Hamilton [1994], Lindner et al. [2011].

We feel that more research is needed on model selection.

B.5.2 Testing

Indications of spurious causality can be generated not only when applying measures of causality but also when testing those measures. The permutation test that was described in the Section 4.5 involves the destruction of all types of dependency, not just causal dependence. In practice it means that for example the existence of instantaneous coupling can result in incorrect deduction of causal inference, if the improvement in prediction due to existence of causality is confused with improvement due to instantaneous coupling. Nevertheless, simplicity is the deciding factor in favour of permutation tests over other approaches.

Several authors Seth and Principe [2011], Amblard et al. [2012b], Sun [2008] propose repeating the permutation test on subsamples to achieve acceptance rates, an approach we do not favour in practical applications. The rationale for using acceptance rates is that the loss in significance from decreasing the size of the sample will be more than made up by calculating many permutation tests for many subsamples. We believe this might be reasonable in the case where the initial sample is big and the assumption of stationarity is reasonable, but that was not the case for our data. We instead decided to report p-values for an overlapping running window. This allows us to additionally assess consistency of results and does not require us to choose the same significance rate for all of the windows.

Appendix C

Experiments: testing sensitivity and misspecification

Before going into specific result, an illustration of the type of outputs that we have when running our simulations / analysis. Below two examples showing the values of the test statistics from the Equation 5.9 change for different data samples, and what values of the χ^2 cdf they would obtain. The rejection level of 0.9 (significance value of $\alpha = 0.1$) is a value that we will often use, but that has been chosen arbitrarily. The Figure C.1 illustrates a compound test with optimised parameters – showing the values of test statistics $L_{X\to Y}$ vs $L_{Y\to X}$ and the distribution $\chi^2_2(2L_{X\to Y})$ vs $\chi^2_2(2L_{Y\to X})$. The data has been generated from causality structure 1 with strong causal effect $X \to Y$, with each of the 50 data sample being of length 500.

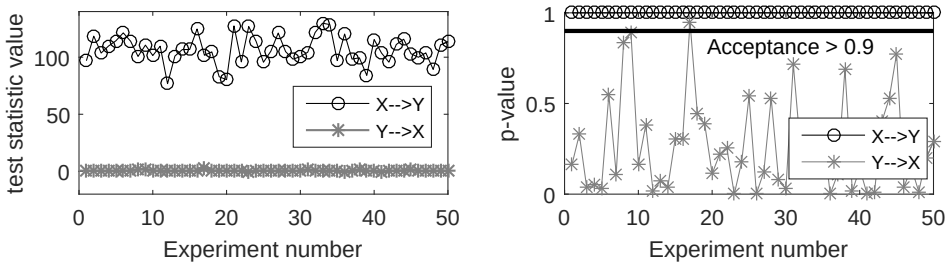


Figure C.1: Test statistics and corresponding cumulative density function evaluations. Causality structure 1, true parameters: $a_X = a_Y = a_Z = 0.3, b_Y = b_Z = 0.7, q = 2, l_a = l_b = e^{-6}, \sigma_f = e^{-10}, \sigma_n = 0.01$. The horizontal axis represents 50 separate trials, each with a time series of length 500.

The interpretation of the Figure C.1 is the following. From the left plot we can see that the test statistics $L_{X\to Y}$ has values which are separated from and considerably larger than the test statistics $L_{Y\to X}$. This by itself is an indication that the causal effect $X \to Y$ should be stronger than $Y \to X$. From the plot of cdf evaluations we observe that all of the values of $L_{X\to Y}$ are in the tail (with cdf values of exactly 1) and therefore the null hypothesis is strongly rejected at any confidence level, for each of the trials. This

means that the estimator of the power of the test (probability of rejecting null hypothesis if it's not true, also equal to 1 - type II error rate) is equal to 1 at any confidence level. If we set up confidence level at 0.1, then one trial will lead to rejecting the null hypothesis in the $Y \to X$ direction, which corresponds to type I error rate of 0.02.

C.1 Model Sensitivity Analysis

It is important to ensure that on one hand the tests behave in a stable way when parameters are changed – at least in some non-extreme region, and on the other hand that the tests are not heavily penalizing misspecifications.

This test is performed for the first data structure (Section 3.2). We use the following settings: Matern kernel, additive noise with variance of $\sigma_n^2 = 0.01$, grid of 21 different parameter values for each parameter. We use each time 100 trials and the length of the simulated time series of $20, 50, 100, 200, 500, 1000$. We report rejection or lack of rejection of the test with the significance of $\alpha = 0.1$ (so rejecting null hypothesis above 0.9). The starting point is the parameter set: $a_X = a_Y = a_Z = 0.3$ and $b_Y = b_Z = 0.7$ (parameters of, respectively, autoregression and causality in the mean, as per Equations 3.51), $l_a = l_b = e^{-1}, \sigma_f = e^{-3}, \sigma_n = 0.1$ (covariance parameters: autoregression, causality, multiplicative scaling, noise covariance, Equations 3.52). Parameters are changed one at a time, and new set of data is generated for each set of parameters.

We don't report results of the sensitivity test for the directions without causality: $Y \to X$ or $Z \to X$, as the test statistics in those cases will always be zero. When performing simple test (i.e. with the true parameters) the direction with no causality present will always show test statistics equal to zero - by definition lack of causality means that test statistics is quantifying difference between equivalent models. When changing parameters in both models at the same time, we no longer use the true parameters, but we still compare models that are equivalent.

In the direction with causality $X \to Y$ we see that the behaviour of the test is very stable, with the changes in the frequency of rejection / non-rejection (here presented as estimated power of the test) influenced mostly by the sample size. The power of the test is the probability $P(H_0 \text{ rejected} | H_1 \text{ true})$, which in our case is estimated as $0.01 \cdot \sum_i^{100} F(2L_{X \to Y})$, where we have 100 trials, F denotes the cdf of χ_2^2 and 0.9 is 1 - confidence level.

When compared to the $X \to Y$ direction, the results for $Y \to Z$ are less uniform, as shown in Table C.1. The Table C.1 demonstrates the power of the test for minimum and maximum of the parameter range, which is enough to portray the behaviour of the test for all parameters except σ_f for the $Y \to Z$ direction, for which local minimum can be seen in the Figure C.2. Based on the Table C.1, and corresponding Figure C.2, we can also observe that the results for $Y \to Z$ are more sensitive to the change in parameters than the results for $X \to Y$, in particular the causal coefficient b_Z.

XY	b_Y	a_Y	a_X	l_b	l_a	σ_f	σ_n^Y	σ_n^X
uniform?	+	+	+	+	+	+-	–	–
	min, max	min, max	min, max	min, max	min, max	min, max	min, max	min, max
n=20	0.45, 0.98	0.84, 0.84	0.83, 0.88	0.84, 0.84	0.84, 0.84	0.80, 0.84	1.00, 0.09	0.40, 1.00
n=50	0.76, 1.00	0.98, 0.98	0.97, 1.00	0.98, 0.98	0.98, 0.98	0.97, 0.92	1.00, 0.46	0.80, 1.00
n=100	0.92, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 0.96	1.00, 0.70	0.91, 1.00
n=200	0.99, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 0.87	0.99, 1.00
n=500	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 0.99	1.00, 1.00
n=1000	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00
YZ	b_Z	a_Z	a_Y	l_b	l_a	σ_f	σ_n^Z	σ_n^Y
uniform?	–	+-	+	+	+	–	–	–
	min, max	min, max	min, max	max, min	max, min	max, min	max, min	max, min
n=20	0.02, 0.87	0.55, 0.22	0.33, 0.39	0.35, 0.35	0.35, 0.35	0.38, 0.63	0.35, 0.38	0.30, 0.96
n=50	0.02, 1.00	0.72, 0.40	0.53, 0.61	0.55, 0.55	0.56, 0.56	0.79, 0.80	0.26, 0.77	0.50, 0.98
n=100	0.05, 1.00	0.87, 0.52	0.69, 0.79	0.70, 0.71	0.71, 0.72	0.99, 0.85	0.21, 0.97	0.64, 1.00
n=200	0.13, 1.00	0.98, 0.70	0.81, 0.93	0.86, 0.87	0.85, 0.89	1.00, 0.98	0.25, 1.00	0.77, 1.00
n=500	0.31, 1.00	1.00, 0.90	0.94, 0.99	0.98, 0.98	0.97, 0.99	1.00, 1.00	0.80, 1.00	0.85, 1.00
n=1000	0.50, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.97, 1.00

Table C.1: How power of the test changes with length of the time series (n) and changes of single parameters. The values of the power of the test are given at the boundary parameter values (corresponding to the values in Fig C.2 and for time series of length $n = 20, 50, 100, 200, 500, 1000$).

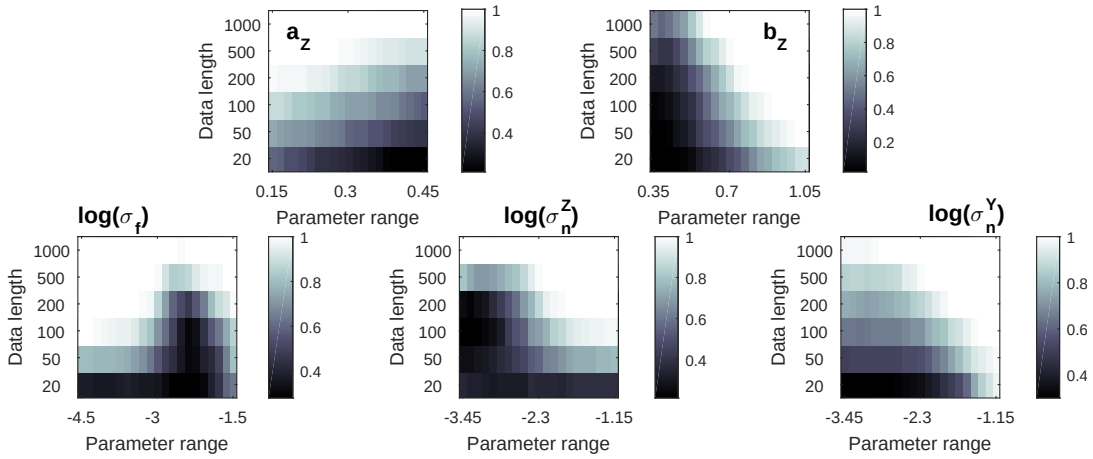


Figure C.2: Causality structure 1, direction $Y \to Z$ original parameters: $a_X = a_Y = a_Z = 0.3, b_Y = b_Z = 0.7, q = 2, l_a = l_b = e^{-1}, \sigma_f = e^{-3}, \sigma_n = 0.1$. Heatmaps show power of the test (hypothesis of no-causality rejected for cdf above 0.9) for different lengths of the time series and for one of the mean or covariance parameters changing $+ - 50\%$ in simulation and model as well.

C.1.1 Model Sensitivity of wGPC

To test sensitivity of the wGPC model, we have chosen a starting point of parameters similar to those in the sensitivity experiment for GPC model, results in Table (C.1). Those parameters are: $a_Y = 0.3, a_X = 0.3, l_a = -1, l_b = -1, sf = -3, \sigma_n = 0.1$. We have chose, however, to decrease the main parameter responsible for causality: $b_Y = 0.3$ to be able to see the results more clearly. The parameters specific for the warpings were chosen to have starting point of $\gamma = 0.3, \nu = 5$

For most of the parameters, the wGPC model behaves in a robust way – increasing power of the test with increasing length of the data, and not exhibiting numerical instability. The only exception is the shape parameter ν, responsible for the leptokurtic tails. It is something that we have observed repeatedly, that for small values of ν, become less stable. The theoretical explanation for that is the fact that for $\nu = 2$, skew-t distribution does not have a finite second moment!

length	a_Y		a_X		b_Y		γ			
	min	max	min	max	min	max	min	max	min	max
	0.15	0.45	0.15	0.45	0.15	0.45	0.15	0.45		
20	0.67	0.63	0.61	0.73	0.3	0.87	0.35	0.89		
50	0.83	0.63	0.62	0.92	0.05	0.98	0.41	0.99		
100	0.99	0.99	0.98	1	0.81	1	0.94	1		
200	1	1	1	1	1	1	1	1		
length	ν		l_a		l_b		sf		σ_n	
	min	max	min	max	min	max	min	max	min	max
	2.5	7.5	-0.5	-1.5	-0.5	-1.5	-1.5	-4.5	0.05	0.15
20	0.98	0.52	0.65	0.65	0.65	0.65	0.24	0.68	0.98	0.97
50	1	0.72	0.76	0.76	0.76	0.76	0	0.83	1	1
100	1	0.99	0.99	0.99	0.99	0.99	1	0.99	1	1
200	**NaN**	1	1	1	1	1	1	1	1	1

Table C.2: How power of the test changes with length of the time series (n) and changes of single parameters. The values of the power of the test are given at the boundary parameter values and for time series of length $n = 20, 50, 100, 200$).

C.2 Model Misspecification Analysis

For the misclassification test we've chosen different starting settings for the covariance function $l_a = l_b = e^{-3}, \sigma_f = e^1$, which result in higher covariance, and much more pronounced effects of misclassification of covariance functions parameters. Starting from the base set of parameters we alter one parameter at a time when calculating the test statistic, but we use data generated for the base parameters: hence that

altered parameter is misspecified. It has to be emphasized that in the misspecification test a parameter will be altered for model A or model B, but not both.

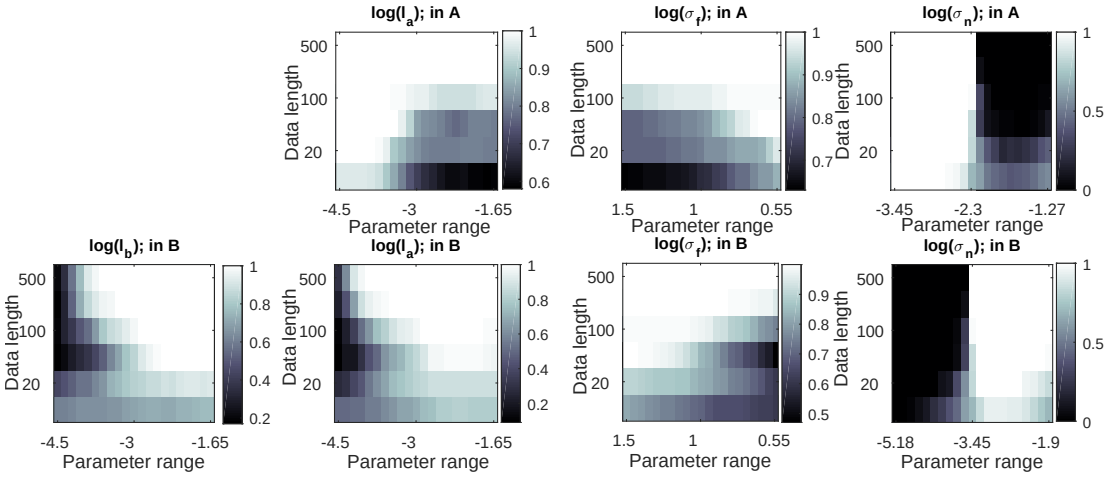


Figure C.3: Power of the test of the hypothesis of non-causality in the direction $X \to Y$ changes with the sample size and misspecification of a single hyperparameter (here – covariance parameters).

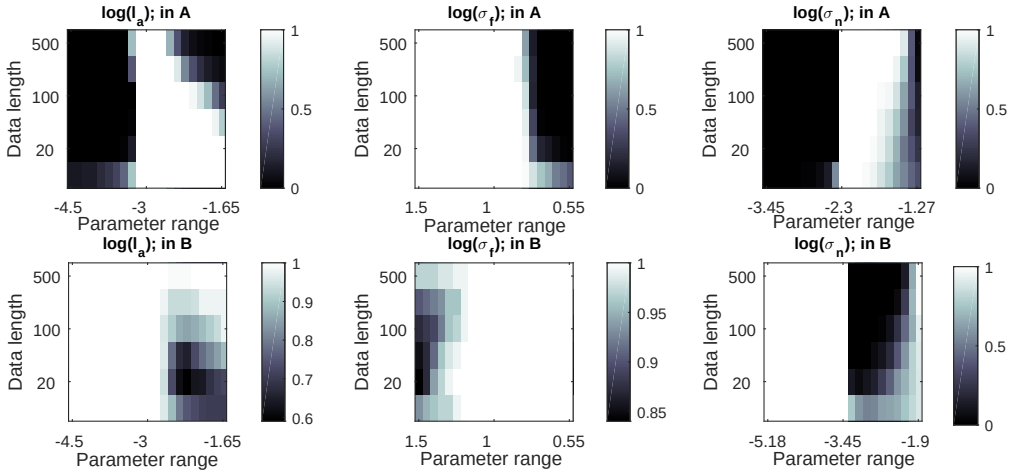


Figure C.4: How 1-rejection rate of the hypothesis of non-causality in the direction $Y \to X$ changes with the sample size and misspecification of a single hyperparameter (here – covariance parameters).

Results of misclassification in the mean, which we don't report, are straightforward to understand and interpret. The power of the test depends mostly on the size of the sample and, to a smaller degree, on the deviation from the true mean. For the direction where causality exists, the power of the test changes almost uniformly with the misclassification of the mean parameter. This is in line with observations that we will see repeatedly – that the power of the test is more robust to any parameter changes in the presence of causality in mean.

Results of misclassification in the covariance, Figures C.3 and C.4, are not so straightforward to understand and interpret though. In particular, the performance of the tests seems to be more sensitive to the misclassification of the noise – this is not observed when parameters of the covariance (mainly σ_f)

are smaller.

Note: the 50% change in the parameters relates to the model parameters, and the covariance parameters are all used as logarithm, so the actual decrease/increase is much bigger than for the mean.

Chapter 11

Bibliography

Bibliography

A sparse covariance function for exact gaussian process inference in large datasets. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, pages 1936–1942. Morgan Kaufmann Publishers Inc., 2009.

Ryan Prescott Adams and Oliver Stegle. Gaussian process product models for nonparametric nonstationarity. In *Proceedings of the 25th international conference on Machine learning*, pages 1–8. ACM, 2008.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Carol Alexander and John Wyeth. Cointegration and market integration: An application to the Indonesian rice market. *The Journal of Development Studies*, 30(2):303–334, 1994.

Mauricio A Álvarez and Neil D Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12(May):1459–1500, 2011.

Pierre-Olivier Amblard and Olivier JJ Michel. On directed information theory and granger causality graphs. 30(1):7–16, 2011.

Pierre-Olivier Amblard, Olivier JJ Michel, Cédric Richard, and Paul Honeine. A Gaussian process regression approach for testing granger causality between time series data. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3360. IEEE, 2012a.

Pierre-Olivier Amblard, Rémy Vincent, Olivier JJ Michel, and Cédric Richard. Kernelizing geweke's measures of granger causality. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012b.

Matthew Ames, Gareth W Peters, Guillaume Bagnarosa, and Ioannis Kosmidis. Upside and downside risk exposures of currency carry trades via tail dependence, 2015.

Matthew Ames, Guillaume Bagnarosa, Tomoko Matsui, Gareth Peters, and Pavel V Shevchenko. Which risk factors drive oil futures price curves?, 2016.

Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear granger causality of time series. *Physical Review E*, 70(5):056221, 2004.

Ethan B Anderes and Michael L Stein. Estimating deformations of isotropic gaussian random fields on the plane. *The Annals of Statistics*, 36(2):719–741, 2008.

Reinaldo B Arellano-Valle and Marc G Genton. Multivariate extended skew-t distributions and related families. *Metron*, 68(3):201–234, 2010.

Aristotle and Hugh Lawson-Tancred. *The Metaphysics*. Penguin Classics, 1998.

Aristotle and Richard McKeon. *The Basic Works of Aristotle*. Random House, 1941.

Aristotle, R. Waterfield, and D. Bostock. *Physics*. Oxford World's Classics, 2008.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Adelchi Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188, 2005.

Adelchi Azzalini and Antonella Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):367–389, 2003.

Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4): 715–726, 1996.

Mohammad Taha Bahadori and Yan Liu. An examination of practical granger causality inference. In *Proceedings of the 2013 SIAM International Conference on data Mining*, pages 467–475. 2013.

Gurdip Bakshi, George Panayotov, and Georgios Skoulakis. The baltic dry index as a predictor of global stock returns, commodity returns, and global economic activity, 2010.

Konrad Banachewicz and Aad Van Der Vaart. Tail dependence of skewed grouped t-distributions. *Statistics & Probability Letters*, 78(15):2388–2399, 2008.

Konrad Banachewicz, Aad van der Vaart, et al. Corrigendum to:" tail dependence of skewed grouped t-distributions"[statist. probab. lett. 78 (2008) 2388-2399]. *Statistics & Probability Letters*, 79(15): 1731–1731, 2009.

David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

Ole Barndorff-Nielsen. Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 353(1674):401–419, 1977.

Ole Barndorff-Nielsen. Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics*, pages 151–157, 1978.

Ole Barndorff-Nielsen and Preben Blaesild. Hyperbolic distributions and ramifications: Contributions to theory and application. pages 19–44, 1981.

Ole Barndorff-Nielsen and Christian Halgreen. Infinite divisibility of the hyperbolic and generalized inverse gaussian distributions. *Probability Theory and Related Fields*, 38(4):309–311, 1977.

Lionel Barnett and Terry Bossomaier. Transfer entropy as a log-likelihood ratio. *Physical review letters*, 109(13):138105, 2012.

Lionel Barnett and Anil K Seth. The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference. 223:50–68, 2014.

Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. 104(3):535–559, 2012.

Preben Blaesild and J Ledet Jensen. Multivariate distributions of hyperbolic type. In *Statistical distributions in scientific work*, pages 45–66. 1981.

Luke Bornn, Gavin Shaddick, and James V Zidek. Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, 107(497):281–289, 2012.

Paola Bortot. Tail dependence in bivariate skew-normal and skew-t distributions. Technical report, 2010.

Taoufik Bouezmarni, Jeroen VK Rombouts, and Abderrahim Taamouti. Nonparametric copula-based test for conditional independence with applications to granger causality. *Journal of Business & Economic Statistics*, 30(2):275–287, 2012.

Phillip Boyle and Marcus Frean. Dependent gaussian processes. In *Advances in neural information processing systems*, pages 217–224, 2005.

Sofiane Brahim-Belhouari and Amine Bermak. Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis*, 47(4):705–712, 2004.

Márcia D Branco and Dipak K Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79(1):99–113, 2001.

Steven L Bressler and Anil K Seth. Wiener–granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, 2011.

John Y Campbell and Robert J Shiller. Stock prices, earnings, and expected dividends. *The Journal of Finance*, 43(3):661–676, 1988.

John Y Campbell, Andrew Lo, and A. Craig MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

Joaquin Quinonero Candela, Agathe Girard, Jan Larsen, and Carl Edward Rasmussen. Propagation of uncertainty in bayesian kernel models-application to multiple-step ahead forecasting. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–701. IEEE, 2003.

Bertrand Candelon, Marc Joëts, and Sessi Tokpavi. Testing for granger causality in distribution tails: An application to oil markets integration. *Economic Modelling*, 31:276–285, 2013.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Mario Chávez, Jacques Martinerie, and Michel Le Van Quyen. Statistical assessment of nonlinear causality: application to epileptic eeg signals. *Journal of neuroscience methods*, 124(2):113–128, 2003.

Bei Chen and Yulia R Gel. A sieve bootstrap two-sample t-test under serial correlation. *Journal of biopharmaceutical statistics*, 21(6):1100–1112, 2011.

Wen-Den Chen. Estimating the long memory granger causality effect with a spectrum estimator. *Journal of Forecasting*, 25(3):193–200, 2006.

Daniel Chicharro. Parametric and non-parametric criteria for causal inference from time-series. In *Directed Information Measures in Neuroscience*, pages 195–219. 2014.

Wan-Chien Chiu, Juan Ignacio Peña, and Chih-Wei Wang. Measuring systemic risk: Common factor exposures and tail dependence effects. *European Financial Management*, 21(5):833–866, 2015.

Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. pages 1422–1430, 2014.

Kevin A Clarke. *Testing Nonnested Models of International Relations: Reevaluating Realism*. 2001.

Rama Cont. Long range dependence in financial markets. In *Fractals in engineering*, pages 159–179, 2005.

Noel Cressie. *Statistics for Spatial Data*. Wiley-Interscience, 1993.

Marcelo G Cruz, Gareth W Peters, and Pavel V Shevchenko. *Fundamental aspects of operational risk and insurance analytics: A handbook of operational risk*. John Wiley & Sons, 2015.

John Cunningham, Zoubin Ghahramani, and Carl Rasmussen. *Gaussian Processes for time-marked time-series data*. 2012.

Rainer Dahlhaus and Michael Eichler. Causality and graphical models in time series analysis. pages 115–137. 2003.

Stefano Demarta and Alexander J McNeil. The t copula and related copulas. *International statistical review*, 73(1):111–129, 2005.

M. A. H. Dempster, Elena Medova, and Ke Tang. Determinants of oil futures prices and convenience yields. *Quantitative Finance*, 12(12):1795–1809, 2012.

Mukeshwar Dhamala, Govindan Rangarajan, and Mingzhou Ding. Estimating granger causality from fourier and wavelet transforms of time series data. *Physical review letters*, 100(1):018701, 2008.

Cees Diks and Valentyn Panchenko. A note on the hiemstra-jones test for granger non-causality. *Studies in nonlinear dynamics & econometrics*, 9(2), 2005.

Stavros Dimitriadis, Yu Sun, Nikolaos Laskaris, Nitish Thakor, and Anastasios Bezerianos. Revealing cross-frequency causal interactions during a mental arithmetic task through symbolic transfer entropy: A novel vector-quantization approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(10):1017–1028, 2016.

Fabrizio Durante. Copulas, tail dependence and applications to the analysis of financial time series. In *Aggregation Functions in Theory and in Practise*, pages 17–22. 2013.

Mark Ebden. Gaussian processes: A quick introduction. *arXiv preprint arXiv:1505.02965*, 2015.

Michael Eichler. Granger causality graphs for multivariate time series, 2001.

Michael Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, 2007.

Michael Eichler. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153(1-2):233–268, 2012.

Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press, 2007.

Michael Eichler and Vanessa Didelez. On granger causality and the effect of interventions in time series. *Lifetime data analysis*, 16(1):3–32, 2010.

Gal Elidan. Copulas in machine learning. In *Copulae in mathematical and quantitative finance*, pages 39–60. 2013.

Paul Embrechts, Filip Lindskog, and Alexander McNeil. Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*, 2001.

Luca Faes, Giandomenico Nollo, and Alberto Porta. Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique. *Physical Review E*, 83(5): 051112, 2011.

Eugene F Fama and Kenneth R French. Permanent and temporary components of stock prices. *Journal of political Economy*, 96(2):246–273, 1988.

Simone Fatichi. ARFIMA simulations, 2009.

Paweł Fiedor. Causal non-linear financial networks. *arXiv preprint arXiv:1407.5020*, 2014.

Pawel Fiedor. Granger-causal nonlinear financial networks. *Journal of Network Theory in Finance*, 1(2): 53–82, 2015.

Jean-Pierre Florens and Denis Fougere. Noncausality in continuous time. *Econometrica: Journal of the Econometric Society*, pages 1195–1212, 1996.

Jean-Pierre Florens and Michel Mouchart. A note on noncausality. *Econometrica: Journal of the Econometric Society*, pages 583–591, 1982.

Jean-Pierre Florens and Michel Mouchart. A linear theory for noncausality. 53(1):157–75, 1985.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Milton Friedman. The fed's thermostat. *The Wall Street Journal*, pages 8–8, 2003.

Caren A Frosch and Philip N Johnson-Laird. Is everyday causation deterministic or probabilistic? *Acta psychologica*, 137(3):280–291, 2011.

Kenji Fukumizu. Kernel methods for dependence and causality. *Machine Learning Summer School*, 2007.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. pages 489–496, 2008.

Thomas Fung and Eugene Seneta. Modelling and estimation for bivariate financial returns. *International statistical review*, 78(1):117–133, 2010a.

Thomas Fung and Eugene Seneta. Tail dependence for two skew t distributions. *Statistics & probability letters*, 80(9-10):784–791, 2010b.

Paul H Garthwaite, Ian T Jolliffe, IT Jolliffe, and Byron Jones. *Statistical inference*. Oxford University Press on Demand, 2002.

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77(378):304–313, 1982.

John Geweke. Inference and causality in economic time series models. *Handbook of econometrics*, 2: 1101–1144, 1984a.

John F Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915, 1984b.

Sid Ghoshal and Stephen Roberts. Extracting predictive information from heterogeneous data streams using gaussian processes. 5(1-2):21–30, 2016.

Agathe Girard, Carl Edward Rasmussen, Joaquin Quinonero Candela, and Roderick Murray-Smith. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. In *Advances in neural information processing systems*, pages 545–552, 2003.

Germán Gómez-Herrero, Wei Wu, Kalle Rutanen, Miguel Soriano, Gordon Pipa, and Raul Vicente. Assessing coupling dynamics from an ensemble of time series. *Entropy*, 17(4):1958–1970, 2015.

Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

Boris Gourévitch and Jos J Eggermont. Evaluating information transfer between auditory cortical neurons. *Journal of neurophysiology*, 97(3):2533–2543, 2007.

Clive William John Granger. Economic processes involving feedback. *Information and control*, 6(1): 28–48, 1963.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2(1):329–352, 1980.

Clive WJ Granger, Bwo-Nung Huangb, and Chin-Wei Yang. A bivariate causality between stock prices and exchange rates: evidence from recent asianflu. *The Quarterly Review of Economics and Finance*, 40(3):337–354, 2000.

Antonio Aznar Grasa. *Econometric model selection: A new approach*, volume 16. Springer Science & Business Media, 2013.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

James D Hamilton. Oil and the macroeconomy since world war ii. *Journal of political economy*, 91(2): 228–248, 1983.

James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, NJ, 1994.

Ernst August v Hammerstein. Tail behaviour and tail dependence of generalized hyperbolic distributions. In *Advanced modelling in mathematical finance*, pages 3–40. Springer, 2016.

Hanyuan Hang and Ingo Steinwart. Fast learning from α-mixing observations. *Journal of Multivariate Analysis*, 127:184–199, 2014.

A. C. Harvey and James H. Stock. The estimation of higher-order continuous time autoregressive models. *Econometric Theory*, 1(1):97–117, 1985.

Matthias Hein and Olivier Bousquet. Kernels, associated structures and generalizations. 2004.

José Miguel Hernández-Lobato, James R Lloyd, and Daniel Hernández-Lobato. Gaussian process conditional copulas with applications to financial time series. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2013.

Craig Hiemstra and Jonathan D Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

Stephen L Hillis and Charles E Metz. An analytic expression for the binormal partial area under the ROC curve. *Academic radiology*, 19(12):1491–1498, 2012.

Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1): 1–46, 2007.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, (3):1171–1220, 2008.

Cheng Hsiao. Autoregressive modeling and causal ordering of economic variables. *Journal of Economic Dynamics and Control*, 4:243–259, 1982.

Meng Hu and Hualou Liang. A copula approach to assessing granger causality. *NeuroImage*, 100: 125–134, 2014.

Meng Hu, Wu Li, and Hualou Liang. A copula-based granger causality measure for the analysis of neural spike train data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2): 562–569, 2015.

Meng Hu, Mingyao Li, Wu Li, and Hualou Liang. Joint analysis of spikes and local field potentials using copula. *NeuroImage*, 133:457–467, 2016.

Wenbo Hu. Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk. 2005.

David Hume. *An Enquiry concerning Human Understanding*. Oxford World's Classics, 2008.

David Hume. *A Treatise of Human Nature*. The Project Gutenberg, 2010.

Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.

Rodney L Jacobs, Edward E Leamer, and Michael P Ward. Difficulties with testing for causation. *Economic Inquiry*, 17(3):401–413, 1979.

Wayne Joerding. Economic growth and defense spending: Granger causality. *Journal of Development Economics*, 21(1):35–40, 1986.

Bent Jorgensen. *Statistical properties of the generalized inverse Gaussian distribution*, volume 9. Springer Science & Business Media, 2012.

Andreas Kaiser and Thomas Schreiber. Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1-2):43–62, 2002.

Immanuel Kant. *Critique of Pure Reason*. Penguin Modern Classics, 2007.

Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90 (430):773–795, 1995.

Tuen Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19, 1978.

KH Knuth, A Gotera, CT Curry, KA Huyser, KR Wheeler, and WB Rossow. Revealing relationships among relevant climate variables with information theory. *Earth Science Technology Office, NASA, Adelphi, Md*, 2005.

Tõnu Kollo, Gaida Pettere, and Marju Valge. Tail dependence of skew t-copulas. *Communications in Statistics-Simulation and Computation*, 46(2):1024–1034, 2017.

Andrei N Kolmogorov. *Foundations of the Theory of Probability, 2nd English edition*. 1933.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Dimitris Kugiumtzis. Partial transfer entropy on rank vectors. *The European Physical Journal Special Topics*, 222(2):401–420, 2013.

Miguel Lázaro-Gredilla. Bayesian warped gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1619–1627, 2012.

Bong-Soo Lee. Causal relations among stock returns, interest rates, real activity, and inflation. *The Journal of Finance*, 47(4):1591–1603, 1992.

Tae-Hwy Lee and Weiping Yang. Granger-causality in quantiles between financial markets: Using copula approach. *International Review of Financial Analysis*, 33:70–78, 2014.

David Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974.

Steven Cheng-Xian Li and Benjamin M Marlin. Classification of sparse and irregularly sampled time series with mixtures of expected gaussian kernels and random features. In *UAI*, pages 484–493, 2015.

Michael Lindner, Raul Vicente, Viola Priesemann, and Michael Wibral. TRENTOOL: A matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC neuroscience*, 12(1):119, 2011.

Max Lungarella, Katsuhiko Ishiguro, Yasuo Kuniyoshi, and Nobuyuki Otsu. Methods for quantifying the causal structure of bivariate time series. *International journal of bifurcation and chaos*, 17(03): 903–921, 2007a.

Max Lungarella, Alex Pitti, and Yasuo Kuniyoshi. Information transfer at multiple scales. *Physical Review E*, 76(5):056117, 2007b.

Xiaolin Luo and Pavel V Shevchenko. The t copula with multiple parameters of degrees of freedom: bivariate characteristics and application to risk management. *Quantitative Finance*, 10(9):1039–1054, 2010.

David JC MacKay. Bayesian non-linear modeling for the prediction competition. In *Maximum Entropy and Bayesian Methods*, number 62, pages 221–234. Springer, 1996.

James G MacKinnon. Model specification tests against non-nested alternatives. *Econometric Reviews*, 2 (1):85–110, 1983.

Burton G. Malkiel. *A Random Walk Down Wall Street*. W. W. Norton & Company, 1973.

Burton G Malkiel. The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17 (1):59–82, 2003.

Burton G Malkiel and Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel-granger causality and the analysis of dynamical networks. *Physical review E*, 77(5):056215, 2008a.

Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel method for nonlinear granger causality. *Physical review letters*, 100(14):144103, 2008b.

Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools-revised edition*. Princeton university press, 2015.

J Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909.

Charles A Micchelli and Massimiliano Pontil. Kernels for multi–task learning. In *Advances in neural information processing systems*, pages 921–928, 2005.

John Stuart Mill. *A System of Logic: Ratiocinative and Inductive annotated*. CreateSpace, 2015.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

Dong Hwan Oh and Andrew J Patton. Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics*, 35(1):139–154, 2017.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

W Olbricht. On mergers of distributions and distributions with exponential tails. *Computational statistics & data analysis*, 12(3):315–326, 1991.

Angeliki Papana, Catherine Kyrtsou, Dimitris Kugiumtzis, and Cees Diks. Simulation study of direct causality measures in multivariate time series. *Entropy*, 15(7):2635–2661, 2013.

Angeliki Papana, Catherine Kyrtsou, Dimitris Kugiumtzis, and Cees Diks. Detecting causality in non-stationary time series using partial symbolic transfer entropy: evidence in financial data. *Computational economics*, 47(3):341–365, 2016.

Andrew J Patton. Copula–based models for financial time series. In *Handbook of financial time series*, pages 767–785. Springer, 2009.

Judea Pearl. [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2), 2010.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

M Hashem Pesaran. Non-nested hypotheses. In *Econometrics*, pages 167–173. Springer, 1990.

M Hashem Pesaran and Melvyn Weeks. Non-nested hypothesis testing: an overview. *A companion to theoretical econometrics*, pages 279–309, 2001.

Fortunato Pesarin. *Multivariate permutation tests: with applications in biostatistics*, volume 240. Wiley Chichester, 2001.

Barnabás Póczos, Zoubin Ghahramani, and Jeff Schneider. Copula-based kernel dependency measures. *arXiv preprint arXiv:1206.4682*, 2012.

Alberto Porta and Luca Faes. Wiener–granger causality in network physiology with applications to cardiovascular control and neuroscience. 104(2):282–309, 2015.

Yuan Alan Qi, Thomas P Minka, Rosalind W Picard, and Zoubin Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of the twenty-first international conference on Machine learning*, page 85. ACM, 2004.

Svetlozar Todorov Rachev. *Handbook of heavy tailed distributions in finance: Handbooks in finance*, volume 1. Elsevier, 2003.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

James Requeima, William Tebbutt, Wessel Bruinsma, and Richard E Turner. The Gaussian process autoregressive regression model (gpar). In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1860–1869. PMLR, 2019.

S Roberts, M Osborne, M Ebden, S Reece, N Gibson, and S Aigrain. Gaussian processes for timeseries modelling. 2012.

Bertrand Russell. On the notion of cause. In *Proceedings of the Aristotelian society*, volume 13, pages 1–26. JSTOR, 1912.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Support vector machine applications in computational biology*. MIT press, 2004.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Anil K Seth. A matlab toolbox for granger causal connectivity analysis. *Journal of neuroscience methods*, 186(2):262–273, 2010.

Anil K Seth, Adam B Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

Sohan Seth and José C Principe. Assessing granger non-causality using nonparametric measure of conditional independence. *IEEE transactions on neural networks and learning systems*, 23(1):47–59, 2011.

Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3): 379–423, 1948.

John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Steven E Shreve. *Stochastic calculus for finance II: Continuous-time models*, volume 11. Springer Science & Business Media, 2004.

Samuel David Silvey. *Statistical inference*. Routledge, 2017.

Herbert A Simon. Causal ordering and identifiability. In *Models of Discovery*, pages 53–80. Springer, 1977.

Christopher A Sims. Money, income, and causality. *The American economic review*, 62(4):540–552, 1972.

Christopher A. Sims. Comparison of interwar and postwar business cycles: Monetarism reconsidered. 70 (2):250–257, 1980a.

Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980b.

Edward Snelson and Zoubin Ghahramani. Local and global sparse gaussian process approximations. In *Artificial Intelligence and Statistics*, pages 524–531, 2007.

Edward Snelson, Zoubin Ghahramani, and Carl E Rasmussen. Warped gaussian processes. In *Advances in neural information processing systems*, pages 337–344, 2004.

Jasper Snoek, Kevin Swersky, Rich Zemel, and Ryan Adams. Input warping for bayesian optimization of non-stationary functions. In *International Conference on Machine Learning*, pages 1674–1682, 2014.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Xiaohai Sun. Assessing nonlinear granger causality from multivariate time series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 440–455. Springer, 2008.

Yee Whye Teh, Matthias Seeger, and Michael I. Jordan. Semiparametric latent factor models. In *Workshop on Artificial Intelligence and Statistics 10*, 2005.

Daniel L Thornton and Dallas S Batten. Lag-length selection and tests of granger causality between money and income. *Journal of Money, credit and Banking*, 17(2):164–178, 1985.

Hiro Y Toda and Taku Yamamoto. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, 66(1-2):225–250, 1995.

Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Pravin K Trivedi, David M Zimmer, et al. Copula modeling: an introduction for practitioners. *Foundations and Trends® in Econometrics*, 1(1):1–111, 2007.

Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1): 45–67, 2011.

Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

Graham White. Medieval theories of causation, 2018.

Halbert White, Karim Chalak, and Xun Lu. Linking granger causality and the pearl causal model with settable systems. In *NIPS Mini-Symposium on Causality in Time Series*, pages 1–29, 2011.

Peter Whittle. The analysis of multiple stationary time series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(1):125–139, 1953.

Norbert Wiener. The theory of prediction. In Edwin F. Beckenbach, editor, *Modern Mathematics for Engineers*, volume 1. New York: McGraw-Hill, 1956.

Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520, 1996.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Paul Wilson. The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*, 127:51–53, 2015.

Patricia Wollstadt, Mario Martínez-Zarzuela, Raul Vicente, Francisco J Díaz-Pernas, and Michael Wibral. Efficient transfer entropy analysis of non-stationary neural time series. *PloS one*, 9(7):e102833, 2014.

Anna Zaremba and Tomaso Aste. Measures of causality in complex datasets with application to financial data. *Entropy*, 16(4):2309–2349, 2014.

Kelly H Zou, A James O'Malley, and Laura Mauri. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–657, 2007.