Statistical Dynamical Models of Multivariate Financial Time Series

Nauman Shah
Exeter College
University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2013

Nauman Shah
Exeter College

Doctor of Philosophy
Trinity 2013

Statistical Dynamical Models of Multivariate Financial Time Series

---

Abstract

The last few years have witnessed an exponential increase in the availability and use of financial market data, which is sampled at increasingly high frequencies. Extracting useful information about the dependency structure of a system from these multivariate data streams has numerous practical applications and can aid in improving our understanding of the driving forces in the global financial markets. These large and noisy data sets are highly non-Gaussian in nature and require the use of efficient and accurate interaction measurement approaches for their analysis in a real-time environment. However, most frequently used measures of interaction have certain limitations to their practical use, such as the assumption of normality or computational complexity. This thesis has two major aims; firstly, to address this lack of availability of suitable methods by presenting a set of approaches to dynamically measure symmetric and asymmetric interactions, i.e. causality, in multivariate non-Gaussian signals in a computationally efficient (online) framework, and secondly, to make use of these approaches to analyse multivariate financial time series in order to extract interesting and practically useful information from financial data.

Most of our proposed approaches are primarily based on independent component analysis, a blind source separation method which makes use of higher-order statistics to capture information about the mixing process which gives rise to a set of observed signals. Knowledge about this information allows us to investigate the information coupling dynamics, as well as to study the asymmetric flow of information, in multivariate non-Gaussian data streams. We extend our multivariate interaction models, using a variety of statistical techniques, to study the scale-dependent nature of interactions and to analyse dependencies in high-dimensional systems using complex coupling networks. We carry out a detailed theoretical, analytical and empirical comparison of our proposed approaches with some other frequently used measures of interaction, and demonstrate their comparative utility, efficiency and accuracy using a set of practical financial case studies, focusing primarily on the foreign exchange spot market.

Contents

List of Notations

The following notations are adopted throughout the thesis, unless otherwise indicated.

General Notations

x	scalar value
$\mathbf{x}$	column vector; $\mathbf{x} = [x_1, x_2, ..., x_T]^\mathsf{T} = [x(t)]_{t=1}^{t=T}$
X	variable
$\mathbf{X}$	matrix
$\mathbf{A}^+$	pseudo-inverse of $\mathbf{A}$
$\det(\mathbf{W}), \mid \mathbf{W} \mid$	determinant of $\mathbf{W}$
$p(x)$	probability density over x
$\mathrm{Tr}(\mathbf{X})$	trace of $\mathbf{X}$
$\mathrm{vec}(\mathbf{W})$	columns of $\mathbf{W}$ stacked on top of each other
$\mid\mid \mathbf{W} \mid\mid_2$	2-norm of $\mathbf{W}$
$\mathbf{x}^\mathsf{T}, \mathbf{X}^\mathsf{T}$	transpose of vector or matrix
$x \in [0,1]$	x lies within the range $0 \leq x \leq 1$
$x \sim \mathcal{N}(\mu, \sigma^2)$	x is drawn from a normal distribution with mean μ and standard deviation σ
$X \rightarrow Y$	X and Y are causally linked and X causes Y
$\mathbf{X} \otimes \mathbf{Y}$	Kronecker product of matrices $\mathbf{X}$ and $\mathbf{Y}$
$\mathrm{Ga}(x; b, c)$	gamma distribution over x with scale parameters b and c
$\mathcal{MN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate normal distribution with vector of means $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and standard deviation σ
$\mathrm{Wi}(\boldsymbol{\Lambda}; a, \mathbf{B}_\Lambda)$	Wishart distribution over $\boldsymbol{\Lambda}$ with parameters a and $\mathbf{B}_\Lambda$

Specific Variables

$a(t)$	univariate recovered independent component
$\mathbf{a}(t)$	vector of recovered source signals at the instant t; $\mathbf{a}(t) = [a_1(t), a_2(t), ..., a_M(t)]^\mathsf{T}$
$\mathbf{A}$	ICA mixing matrix
$\alpha_{ij,p}$	weight parameters of a MAR (or AR) model at time lag of p
b	wavelet localisation parameter
$\mathbf{B}$	matrix of recovered ICA source signals at all time instances; $\mathbf{B} = [\mathbf{a}(t)]_{t=1}^{T}$
β	precision (of covariance matrix) of ICA observation noise
$\mathbf{c}(t)$	vector of wavelet coefficients
$C(\mathbf{x}, \varepsilon)$	correlation dimension, with length scale ε
$C_{u,b}$	wavelet coefficients obtained using CWT at scale u and localisation parameter b
$\mathbf{C}_{con}$	connectivity matrix
$\mathbf{C}_\mathbf{x}$	matrix of wavelet coefficients for multivariate time series $\mathbf{x}(t)$

$\chi^2(k)$	chi-square distribution with k degrees of freedom
d	dimension of data set
d_{ij}	pseudo-distance between currency pairs i and j, used for constructing a MST
$d(\cdot)$	distance (based on the 2-norm) between two matrices
D	data set
$\mathbf{D}, \hat{\mathbf{D}}$	diagonal matrices
δ_{jk}	Kronecker's delta function
Δ	sampling period of data
e_{TE}	tracking error
$e_{x,rms}$	root mean square error of predicted values of x
$\mathrm{E}[\cdot]$	expectation operator
$\mathbf{E}$	set of lagged regression error terms (as used in the GIC and ICA-AR models)
η	ICA-based information coupling
f	frequency
f_o, f_u	centre and pseudo frequencies of a wavelet respectively
F	F-statistic value
$F(p)$	negative variational free energy at time lag of p
$g(\cdot)$	negative log-likelihood loss function
$G_{c,ij}$	elements of VG causality matrix, representing causality between time series i and j, where $i \rightarrow j$
G_n	number of groups (structured priors) of different parameters
$\mathbf{G}$	Hessian matrix (obtained during the ICA log-likelihood analysis)
$\mathbf{G}_c$	Granger causality matrix
γ	skewness of a distribution
Γ	gamma function
h	Parzen window size (used to compute mutual information)
H	information entropy
H_0	null hypothesis
H_1	alternate hypothesis
$\mathbf{H}$	Hessian matrix (obtained during the BFGS optimisation process)
I	mutual information
$I_C(X,Y)$	average amount of information between X and Y, obtained using the generalised correlation integral
I_N	normalised mutual information
$\mathbf{I}$	identity matrix
$\mathbf{I}_M$	M dimensional identity matrix
J	negentropy
$\mathbf{J}$	skew-symmetric matrix used to parameterise $\mathbf{Q}$
JB	Jarque-Bera statistic
JB_{avg}, JB_{MV}	average and multivariate values of the Jarque-Bera statistic
JB_c	critical value of the Jarque-Bera test
K	Parzen window function
$\mathbf{K}$	unit matrix (a matrix of ones)
KL	Kullback-Leibler divergence

κ	kurtosis of a distribution
L	likelihood
L_{av}	average log-likelihood
L_τ	lag operator
ℓ	log-likelihood
λ_{max}	maximum eigenvalue
M	number of source signals
n_s, N_D	number of data points
N	number of observed signals
N_θ	number of model parameters
$\boldsymbol{\omega}_i$	i-th row of $\boldsymbol{\Omega}$ (which has been row-normalised and permutation adjusted)
$\boldsymbol{\Omega}$	set of MAR model parameters
$\mathcal{O}(\cdot)$	order of complexity
p_{IV}	variable sampled from a Pearson type IV distribution
$p_{ij}, p(j \mid i)$	HMM and HMICA model state transition probability from state i to state j
$P_B(t)$	mid-price of currency basket at time t
$P_I(t)$	mid-price of currency index at time t
$\hat{P}_{UC}$	triangulated USDCHF exchange rate
P_{sig}	significance level of VG causality
$P(t), P_t$	exchange rate (mid-price) at time t
$\mathbf{P}_{hmm}$	HMICA state transition probability matrix
$\mathbf{P}_{sig}$	significance values matrix of VG causality
$\psi(t)$	mother wavelet
$\psi_{u,b}(t)$	normalised wavelet function
$\mathbf{Q}$	real orthogonal matrix
Q_k	HMM auxiliary function for state k
Q	HMM auxiliary function (summed over all states)
$r(t), r_t$	log of returns at time t
$\hat{r}_{UC}(t)$	triangulated USDCHF returns at time t
$\mathbf{R_n}$	noise covariance of the ICA source model
$R(t), R_t$	returns at time t
ρ	linear correlation (Pearson's product-moment correlation coefficient)
ρ_R	Spearman's rank correlation
$s(t)$	univariate ICA source signal
$\mathbf{s}(t)$	vector of source signals at the instant t; $\mathbf{s}(t) = [s_1(t), s_2(t), ..., s_M(t)]^\mathsf{T}$
$\sigma_{SR}(t)$	survival ratio of a MST at time t
$\sigma_{SR}(t,k)$	multi-step survival ratio of a MST at time t, computed at k steps
ς	weight precision parameter (as used in the VB-MAR model)
$\boldsymbol{\Sigma}$	diagonal matrix of singular values of a set of observed signals
t	discrete time index
$t_{rebalance}$	rebalance period for a portfolio
Δt	length of sliding-window
T_{ij}	time in any given HMICA state
$T_{C,X\to Y}$	statistic used as part of a correlation integral based Granger causality test

$T_{E,X\to Y}$	transfer entropy
τ	time lag
θ_{MAP}	MAP estimation of model parameters
θ_{MLE}	MLE of model parameters
Θ	Heaviside function
u	wavelet scale
$\mathbf{U}$	orthogonal matrix with columns the same as the principal components of X
$\mathbf{V}$	orthogonal matrix obtained using singular value decomposition of observed signals
w_P	weight of any given instrument in a portfolio
$\mathbf{w}_P(t)$	vector of weights of a portfolio at time t
$\mathbf{W}$	ICA unmixing matrix
$\mathbf{x}_G$	Gaussian random variable
$\mathbf{x}_i$	i-th time series selected from a set of multivariate time series
$\mathbf{x}(t)$	vector of ICA observed signals at the instant t; $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_N(t)]^\mathsf{T}$
$\mathbf{X}$	matrix of ICA observed signals at all time instances; $\mathbf{X} = [\mathbf{x}(t)]_{t=1}^{T}$
$z(t)$	HMM latent states
$\mathbf{Z}$	set of lagged data (as used in the GIC and ICA-AR models)
ζ	significance level of GIC causality
ζ_c	critical value of significance level of GIC causality

List of Acronyms

AE	Absolute Error
AR	Autoregressive
BFGS	Broyden-Fletcher-Golfarb-Shanno
BIC	Bayesian Information Criterion
cdf	cumulative distribution function
CWT	Continuous Wavelet Transform
EBS	Electronic Broking Services
FX	Foreign Exchange
GIC	Granger Independent Component
GMV	Global Minimum Variance
gPDC	generalised Partial Directed Coherence
HMICA	Hidden Markov Independent Component Analysis
HMM	Hidden Markov Model
ICA	Independent Component Analysis
JB	Jarque-Bera
KL	Kullback-Leibler
MAE	Mean Absolute Error
MAP	Maximum a Posteriori
MAR	Multivariate Autoregressive
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate
MST	Minimum Spanning Tree
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
pdf	probability density function
RBF	Radial Basis Function
rms	root mean square
RTT	Real-Time Trading
SNR	Signal-to-Noise Ratio
TF	Trend-Following
VaR	Value-at-Risk
VB	Variational Bayes
VG	Variational Granger

Acknowledgments

Writing this thesis would not have been possible without the help and support of a number of people and organisations. Foremost, I thank my supervisor, Professor Stephen Roberts, for his invaluable help, support and guidance. I am also grateful to Dr Will Addison for some very useful discussions. Moreover, I thank members of the Machine Learning Research Group and the Oxford-Man Institute of Quantitative Finance for their help and advice. I am also thankful to my undergraduate tutors at Oxford, Professor Ian Reid and Dr Nik Petrinic, for instilling in me a passion for engineering research and for helping me to develop a strong foundation in the principles of engineering sciences. Most importantly, I thank my parents for providing unwavering support to all my pursuits since an early age; their constant encouragement has been instrumental in inspiring me to achieve my goals in life.

It will be fair to say that this project would not have materialised without the generous financial support provided by Exeter College, Oxford. Therefore, I thank the Rector and Fellows of Exeter College for awarding me the Amelia Jackson Senior Studentship for carrying out research leading to this thesis. I also gratefully acknowledge the funding and support provided by the Department of Engineering Science, Oxford-Man Institute of Quantitative Finance and the Man Group.

Chapter 1
Introduction

1.1 Motivation

The task of accurately inferring the statistical dependency structure in multivariate systems has been an area of active research for many years, with a wide range of practical applications. Many of these applications require real-time sequential analysis of interactions in multivariate data streams with dynamically changing properties. However, most existing measures of interaction have some serious limitations in terms of the type of data sets they are suitable for or their computational and analytical complexities. If the data being analysed is generated using a known stable process, with known marginal and multivariate distributions, the level of dependence can be relatively easily estimated. However, most real-world data sets have dynamically changing properties to which a single distribution cannot be assigned. Data generated in the global financial markets is an obvious example of such data sets. Financial data exhibits rapidly changing dynamics and is highly non-Gaussian (heavy-tailed) in nature. Over the last few years, financial markets have also witnessed the availability and widespread use of data sampled at high frequencies, which require the use of computationally efficient algorithms for analysis in an online environment. Due to these reasons, most commonly used measures of interaction are not suitable for accurate real-time analysis of multivariate financial time series.

Interactions can broadly be classified into two distinct groups, i.e. symmetric and asymmetric. Symmetric interaction measurement approaches aim to estimate the common instantaneous information content of a set of signals. In contrast, asymmetric measures aim to estimate the strength and direction of information flow between signals at non-zero time lags, i.e. they can be used to infer the presence of causality in a system. This thesis presents the develop-

ment and application of a set of symmetric and asymmetric measures of interaction which are suitable for use with multivariate financial time series. As interactions in financial systems show significant variations across time as well as scale, we extend our interaction measurement approaches to efficiently and accurately capture the dynamically evolving dependency structure in financial markets as well as to analyse scale-dependent variations in dependencies. To analyse interactions in high-dimensional multivariate systems, we make use of static and dynamic complex networks to extract the underlying hierarchical dependency structure. The interaction measurement approaches which we present in this thesis are not only suitable for modelling dependencies in multivariate data sets with non-Gaussian distributions, but are also computationally efficient, which makes it possible to use them in an online dynamic environment, even when dealing with data sampled at high-frequencies. The approaches make use of various statistical and signal processing techniques, and most are primarily based on a well-known blind-source separation method known as independent component analysis (ICA). The utility and practical application of these approaches is demonstrated by applying them to various practical financial problems, and the results obtained are compared with other standard interaction measurement approaches currently used in practise. The implications of accurately inferring the interaction structure is evident from the results of these applications, which include, among others, analysing financial portfolios, predicting exchange rates and tracking financial indices. All applications are simulated using suitable data sets, which include spot foreign exchange (FX) returns sampled at different frequencies and an equities data set. All approaches developed are suitable for use in a causal, online, environment and (where applicable) all results presented are obtained as such.

1.2 Overview of the thesis

The thesis is divided into seven chapters and is broadly focused on two major related topics, i.e. the development of efficient interaction (symmetric and asymmetric) measurement approaches which are suitable for real-time analysis of multivariate non-Gaussian data streams, and the application of these approaches to practical financial problems with the aim of extracting interesting and useful information from multivariate financial time series. This chapter serves as an introduction to the thesis and gives the motivation for the research undertaken.

Chapter 2 provides a critical review of current literature dealing with existing interaction measurement approaches and sets the foundation for developing new approaches in later chapters. Chapter 3 presents an ICA-based symmetric interaction measurement approach which can be used to accurately model interactions in non-Gaussian systems in a computationally efficient manner. For the purpose of this thesis, we refer to this measure as "information coupling" (or simply "coupling"). The chapter also proposes methods suitable for analysing the dynamics and scale dependence of information coupling. Moreover, it discusses suitable methods for analysing information coupling in high-dimensional multivariate systems using complex coupling networks. Chapter 4 starts by providing an introduction to the statistical properties of financial time series and describes the synthetic and financial data used to obtain the results presented in this thesis. It goes on to present a range of empirical case studies which demonstrate application of the information coupling measure (as well as other existing symmetric interaction measurement approaches) for analysing multivariate financial time series. Chapter 5 presents two approaches to estimate asymmetric interactions, i.e. causality, in multivariate systems across both time- and frequency-domains. One of these is based on a combination of the principles of ICA, multivariate autoregressive (MAR) models and Granger causality, and is well suited for efficiently analysing causality in non-Gaussian dynamic environments. For the purpose of this thesis, we call this the Granger independent component (GIC) causality detection approach (we also present a variant of the GIC approach for autoregressive modelling of univariate time series). The second approach makes use of variational Bayesian MAR (VB-MAR) models for time- and frequency-domain causal inference. This approach, which we call variational Granger (VG) causality, enables us to address some of the limitations of standard MAR models (such as model overfitting) and provides us with a simple framework to measure non-linear asymmetric dependencies. Chapter 6 focuses on making use of these (and other standard) causal inference approaches to investigate the presence of asymmetric dependencies in multivariate financial data streams. The thesis concludes with Chapter 7 which presents a summary of the research undertaken and provides directions for future work in this area.

Chapter 2
Previous work (with critique)

This chapter presents a literature review of existing concepts and approaches for dynamic multivariate interaction measurement. We critically analyse the merits and limitations of various methods in context of their potential utility for modelling dependencies in multivariate financial time series. The chapter goes on to present the theoretical background of the methods used to develop a set of symmetric and asymmetric interaction measurement approaches, and their extensions, in later chapters. We start by presenting a review of some fundamental concepts of statistical inference, which provide us with an overarching framework for the development of interaction measurement approaches presented later in the thesis.

2.1 Review of fundamentals of statistical inference

Empirical data obtained as a result of real-world processes is generally noisy and finite in size. Any useful information extracted from such data sets will have a degree of uncertainty associated with it, reflecting the noise and sparsity of the data. Probability theory provides us with an elegant framework to deal with this uncertainty in the observed data, as described below.

2.1.1 Probability theory

Probability theory is a branch of mathematics concerned with the analysis of random phenomena. It provides a consistent framework for the quantification and manipulation of uncertainty [41]. Using probability theory, optimal predictions about an event can be made using all the available information, which may be ambiguous or incomplete. Here we provide a brief overview of some basic concepts and principles of probability theory; a more detailed review is presented in [196]. Probability can be measured using either the frequentist or Bayesian

approach. The frequentist approach, also called the classical approach, interprets probabilities in terms of repeatable experiments. It assumes that data is randomly generated by a fixed set of parameters; these parameters can be estimated from the data using, for example, maximum likelihood estimation techniques (which we discuss later). On the contrary, Bayesian statistics makes use of probabilities to quantify the *degrees of belief* in different models. Bayesian statistics takes its name from the commonly used Bayes' theorem, which was derived from the work of Thomas Bayes and published in 1764 [30]. An interesting comparison of the frequentist and Bayesian approaches can be found in [183].

Bayesian probability is measured as a probability distribution over a given parameter. It treats the entity or parameter of interest as a random variable, which makes it possible to estimate the uncertainty associated with the estimation process using a single observed data set [183]. This makes the use of Bayesian approaches much more flexible for most practical analysis, as in most cases enough real-world data is not available. Due to their numerous advantages for computation in the presence of uncertainty, Bayesian approaches are finding increasing use in time series analysis in general and financial data modelling in particular [299]. Another advantage of using the Bayesian approach is the ability to incorporate prior knowledge about a system into the model, thus improving the posterior probability estimate. Given a parameter θ and data set D, the posterior probability, $p(\theta \mid D)$, can be calculated using Bayes' rule:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)} \tag{2.1}$$

where $p(D \mid \theta)$ is the likelihood, $p(\theta)$ is the prior probability and $p(D)$ is the marginal likelihood, given by:

$$p(D) = \int p(D \mid \theta)p(\theta)d\theta \tag{2.2}$$

Inference in Bayesian methods

Bayesian inference can be computationally expensive [326], mainly due to the cost associated with computing the intractable marginal integral, as given by (2.2). It is possible to estimate this integral using various stochastic Markov chain Monte Carlo (MCMC) techniques,

however, MCMC methods can be computationally demanding and have convergence problems [15]. Therefore, we need to consider the use of approximate Bayes methods. These methods are generally based on the principle of approximating the posterior distribution using optimisation techniques[1]. A commonly used approximate Bayes approach is the Laplace approximation which makes a local Gaussian approximation around a maximum a posteriori (MAP) estimate (as we discuss later, MAP estimation is synonymous to selecting the mode of the posterior distribution as the best estimate). However, this approach can be relatively inaccurate for small data sets [137]. Expectation propagation is another technique which relies on local optimisation of the cost function and can therefore be inaccurate in practise [31]. A few other, less frequently used, methods for obtaining an approximation for the posterior are discussed in [218].

The method of choice used in this thesis for Bayesian approximations is variational Bayes (VB) [169]. Using VB techniques, it is possible to get a parametric approximation for the true posterior density of an intractable distribution by using an approximate distribution for which the required inferences are tractable [218]. The approximate posterior distribution can be obtained by finding a distribution, such that it minimises the Kullback-Leibler (KL) divergence between this distribution and the actual posterior distribution. Denoting the true posterior distribution by $p(\theta \mid D)$ and its tractable approximate distribution by $q(\theta \mid D)$, the KL divergence (also called relative entropy) gives us a measure of difference, or non-metric distance, between two distributions over the same variables and can be written as [218]:

$$KL[q(\theta \mid D) \mid\mid p(\theta \mid D)] = \int q(\theta \mid D) \log \left[\frac{q(\theta \mid D)}{p(\theta \mid D)} \right] d\theta \tag{2.3}$$

where D is the observed data, and θ is the unknown parameter. All equations assume that log is to base e, unless otherwise stated. Multiplying the top and bottom part of the bracketed term by $p(D)$, (2.3) can be rewritten as:

$$KL[q(\theta \mid D) \mid\mid p(\theta \mid D)] = \int q(\theta \mid D) \log \left[\frac{q(\theta \mid D)}{p(D, \theta)} \right] dy + \log p(D) \tag{2.4}$$

[1]Optimisation is the process of maximising or minimising a given function or set of functions by estimating a set of decision variables. Most practical optimisation problems need to take into account multiple constraints and often require the use of advanced optimisation techniques [250]. A function can have global as well as local extrema. Many optimisation techniques can only accurately identify local extrema, although there are various global optimisation methods in common use as well [81].

In the above equation, the joint density, $p(D,\theta)$, is relatively simple to compute at any given point. However, in most practical cases, it is somewhat complicated to evaluate the distribution of the observed variable, $p(D)$, as well as to marginalise the posterior distribution, $q(\theta \mid D)$. Luckily, $p(D)$ is the same for all models, therefore it is not significant in our analysis and can be ignored by defining a cost function given by:

$$C_\theta(D) = \log p(D) - KL[q(\theta \mid D) \,||\, p(\theta \mid D)] = -\int q(\theta \mid D) \log \left[\frac{q(\theta \mid D)}{p(D,\theta)}\right] d\theta \qquad (2.5)$$

where $C_\theta(D)$ is the cost function for the unknown variable θ. This is the fundamental equation of the VB framework, and the main aim of VB-learning is to maximise the cost function by making the approximate distribution as close as possible to the true posterior distribution (we note that the cost function is maximised when $q(\theta \mid D) = p(\theta \mid D)$) [131]. This can be done by starting with a fixed tractable parametric form for $q(\theta \mid D)$ and then training the parameters of this distribution such that the cost function given in (2.5) is maximised (for practical purposes, $q(\theta \mid D)$ is often chosen as a product of simple terms [291]).

2.1.2 Parameter estimation of probabilistic models

As previously discussed, it is often the case that probabilistic models have a set of flexible parameters which need to be inferred in order to optimise a function or set of functions. This section provides an overview of some of the basic parameter estimation methods in common use.

Maximum likelihood estimation

The likelihood, $L(\theta \mid D)$, of a set of model parameters, θ, given some observed data, D, refers to the probability, $p(D \mid \theta)$, of obtaining that set of observed data given the set of model parameters. Many practical statistical problems relate to fitting a model, with various parameters, to some data set. The data set is usually fixed while the parameters can be fine-tuned to obtain the best possible results by optimising some statistical measure. The most commonly used measure is the log-likelihood of the data. This is because in many practical applications, the overall likelihood function is a product of a number of statistically independent likelihood functions, and as logarithms convert products into summations, it is often more convenient to use the log-

arithm of the likelihood function, i.e. the log-likelihood function, $\ell(\theta \mid D)$. Estimates for the model parameters can be obtained by varying the parameters with the aim of maximising the log-likelihood function. This process is referred to as maximum likelihood estimation (MLE) and can be represented as:

$$\theta_{MLE} = \arg\max_{\theta} \ell(\theta \mid D) \tag{2.6}$$

where θ_{MLE} is the MLE of the model parameters. Representing the first and second derivatives of the log-likelihood function by ℓ' and ℓ'' respectively, estimates for θ_{MLE} can be obtained by setting $\ell'(\theta_{MLE} \mid D) = 0$ and making sure that $\ell''(\theta_{MLE} \mid D) < 0$. A detailed analysis of the MLE procedure is presented in [262]. Using the MLE approach for parameter estimation has multiple advantages. The ML estimator is consistent, which implies that for large data sets the MLE will converge to the true parameter [351]. It also has the smallest variance of all estimators and can be used to obtain confidence bounds and hypothesis tests for the inferred parameters [96]. However, the MLE approach also has some limitations. The likelihood functions can be complex to work out and the numerical estimation of the ML can sometimes be difficult. The MLE approach is also sensitive to the choice of starting values and can give inaccurate results for small samples [96].

Maximum a posteriori estimation

In Bayesian statistics, estimates for the posterior probability, $p(\theta \mid D)$, are frequently required. Maximum a posteriori (MAP) estimation is a procedure through which the mode of the posterior distribution is selected as the best estimate. Noting that $p(D)$ in (2.1) serves as a normalising term, the θ_{MAP} estimate of model parameters can be written as:

$$\theta_{MAP} = \arg\max_{\theta} p(D \mid \theta) p(\theta) \tag{2.7}$$

MAP is closely related to the MLE approach for parameter estimation, with the difference that MAP can be used to include prior information, $p(\theta)$, about the model parameters. Hence, estimation of model parameters using the MAP approach can be computationally less demanding than the MLE approach. However, if the posterior distribution is multi-modal, MAP estimation will choose the highest mode, which may result in relatively inaccurate results as the highest

mode may not correctly reflect the characteristics of the actual posterior [142].

Model order selection

In statistics, model order selection refers to the task of selecting the optimal model from a set of competing models. The main idea behind model order selection can best be described using Occam's razor [44], which in principle states that given a set of models all of which explain data equally well, the simplest, i.e. the model with the fewest parameters, should be chosen. There are various model order selection approaches which are based on this trade-off between the increase in data likelihood and model overfitting when adding parameters to a model, e.g. the Akaike information criterion (AIC) [310], and minimum description length [157]. Our preferred approach in this thesis is the Bayesian information criterion (BIC), also known as the Schwarz information criterion. BIC provides a framework for estimating the optimal model order by penalising models with larger number of parameters more heavily than alternate models (such as AIC), hence, models with relatively lower complexity are selected. For a given data set D with parameters θ, the BIC is given by [41, 67]:

$$BIC = \ell(\theta \mid D) - \frac{1}{2} N_\theta \log N_D \tag{2.8}$$

where N_θ and N_D are the number of model parameters and number of data points respectively. An optimum model is one which maximises the BIC; hence, we note that BIC selects a model which maximises the log-likelihood of the data, $\ell(\theta \mid D)$, with respect to the model's complexity, $\frac{1}{2} N_\theta \log N_D$.

2.1.3 Information theory

Probability theory, as discussed so far, provides us with a framework to quantify and manipulate uncertainty in real-world settings. However, to gain knowledge about the information content of a variable (or set of variables) we need to rely on the principles of information theory, which is based on the foundations of probability theory. Information theory offers a unified framework for the quantification of the flow and "value" of information; it allows us, for example, to quantify the "value" of observed data. The foundations of information theory were laid in a 1948 paper published by Claude Shannon [319]. It has wide-ranging applications in various sectors, such as communications, cryptography, speech processing and (as

discussed later in this thesis) in multivariate financial time series analysis. A fundamental topic in the field of information theory is information flow, which, as the name implies, is the transfer, or flow, of information from one variable to another. Information flow can be measured by calculating changes in *entropy*.

Information entropy, also known as Shannon's entropy, is a measure of the uncertainty, or unpredictability, of a variable [319]. Entropy is measured in *nats*, if the calculation is based on natural logarithms; in *bans*, if the calculation is based on base-10 logarithms; or in *bits*, if base-2 logarithms are used. One *nat* corresponds to $\frac{1}{\log 10}$ *bans* or $\frac{1}{\log 2}$ *bits*. For a random vector $\mathbf{x}$, with a continuous probability distribution $p(\mathbf{x})$, entropy in *nats* is given by:

$$H[\mathbf{x}] = \mathrm{E}\left[\log \frac{1}{p(\mathbf{x})}\right]_{\mathbf{x}} = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \tag{2.9}$$

where $\mathrm{E}[\cdot]$ is the expectation operator. For multivariate systems, we generally consider the joint entropy as a measure of uncertainty of a set of random variables which are described by a joint distribution. The information entropy of $\mathbf{x}$ is given by (2.9). Using this equation, the joint entropy for two variables can be written as [41]:

$$H[\mathbf{x}, \mathbf{y}] = -\int \int p(\mathbf{x}, \mathbf{y}) \log p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \tag{2.10}$$

To quantify the amount of information which one random variable contributes about another random variable, we can measure the conditional entropy, also known as equivocation. For two variables, the conditional entropy can be written as [41]:

$$H[\mathbf{x} \mid \mathbf{y}] = -\int \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} d\mathbf{x} d\mathbf{y} \tag{2.11}$$

An information theoretic measure of symmetric interaction, based on information and conditional entropies, is mutual information. We present further details of this measure in the next section, in which we discuss the advantages and limitations of various approaches to symmetric interaction measurement.

2.2 Approaches to symmetric interaction measurement

Having reviewed some basic concepts of statistical inference which are of relevance to our work, we now proceed to discuss the advantages and limitations of various existing approaches to interaction measurement, which is the main topic of the research presented in this thesis. This section critically reviews some of the frequently used measures of symmetric interaction, while the next section presents a review of asymmetric measures of interaction. By far the most commonly used symmetric interaction measure is linear correlation. Rank correlation, copula functions and information theoretic measures such as mutual information are also widely used. However, all these measures have some serious drawbacks and limitations, as discussed below.

Linear correlation

Linear correlation, also known as Pearson's product-moment correlation, is a measure of similarity between two signals, and indicates the strength of linear relationship between them. Interactions between two signals tends to induce correlation between them; however, high correlation does not always occur due to interactions between the two signals under consideration, as the two signals may be causally (asymmetrically) driven by a third signal [302]. The three main limitations of using linear correlation measures are that they cannot accurately model interactions between signals with non-Gaussian distributions [171], they are restricted to measuring linear statistical dependencies, and they are very sensitive to outliers [102]. Financial time series have non-Gaussian distributions in the log-returns space [89, 98]. This is especially true for financial data recorded at high frequencies [244]. In fact, as the scale over which returns are calculated decreases, the distribution becomes increasingly non-Gaussian, a feature referred to as aggregational Gaussianity. Linear correlation analysis assumes that the bivariate time series, between which correlation is being estimated, has an elliptical joint distribution, of which the bivariate normal is a special case [116]. Therefore, second-order correlation functions alone are not suitable for capturing interactions in multivariate financial returns, which often have heavy-tailed, skewed, distributions [89]. Hence, any practical measure of interaction used for analysing financial returns needs to take into account the higher-order statistics of the data as well.

Rank correlation

A non-parametric measure of correlation which, unlike linear correlation, can be used to model both linear and non-linear (monotonic) interactions, is based on the concept of ranking of variables. This measure estimates the relationship (correlation) between different rankings of the same set of items. It assesses how well an arbitrary monotonic function can describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables [17]. There are two popular rank correlation measures, Spearman's coefficient and Kendall's coefficient, both of which usually produce similar results [87]. However, as Kendall's rank correlation measure is computationally more demanding [87], only Spearman's coefficient is used as a measure of rank correlation in this thesis. Spearman's rank correlation measure is only valid for monotonic functions and is not suitable for large data sets, as assigning ranks to a large number of observations can be computationally demanding. Financial returns often have a large fraction of zero values, which result in tied ranks [95]. Rank correlation measures cannot accurately deal with the presence of tied ranks and hence the results obtained can be misleading [141]. Moreover, use of ranks (instead of the actual values) can potentially cause a loss of information from the data being analysed, leading to inaccurate results [177]. An interesting study describing some other practical disadvantages of using the rank correlation measure is presented in [225].

Mutual information

Our discussion so far has pointed out various limitations of correlation measures in relation to the type of data sets they are valid for or the type of interactions they can accurately model. It is possible to address these limitations by making use of mutual information, an information theoretic measure of symmetric interaction based on information entropy. Mutual information, also known as transinformation, can be used to estimate the amount of common information content of a set of variables by measuring the reduction in uncertainty of one variable due to the knowledge of another variable [112]. Mutual information between two (or more) signals is always positive and is the canonical measure of symmetric interaction between the signals. It is a quantitative measurement of how much information the observation of one variable gives us regarding another. Thus, the higher the mutual information between $\mathbf{x}$ and $\mathbf{y}$, the less

uncertainty there is in **x** given **y** or **y** given **x**. The information entropy of a variable is given by (2.9) while (2.11) gives the conditional entropy for a set of variables. Using these two equations, the mutual information between **x** and **y** can be written as:

$$\begin{aligned} I[\mathbf{x},\mathbf{y}] &= H[\mathbf{x}] - H[\mathbf{x} \mid \mathbf{y}] \\ &= H[\mathbf{y}] - H[\mathbf{y} \mid \mathbf{x}] \end{aligned} \tag{2.12}$$

Using Bayes' theorem, (2.12) can be written as:

$$I[\mathbf{x},\mathbf{y}] = H[\mathbf{x}] + H[\mathbf{y}] - H[\mathbf{x},\mathbf{y}] \tag{2.13}$$

Substituting (2.9) and (2.10) into (2.13), the mutual information between **x** and **y** is given by [41]:

$$I[\mathbf{x},\mathbf{y}] = -\int\int p(\mathbf{x},\mathbf{y}) \log \left[\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})} \right] d\mathbf{x}d\mathbf{y} \tag{2.14}$$

Mutual information can theoretically be used to accurately measure the level of linear as well as non-linear interactions in multivariate systems, irrespective of the marginal distributions. However, a major problem associated with using information theoretic measures, such as mutual information, is the complexity of their computation, especially in high-dimensional spaces. Mutual information is exquisitely sensitive to the joint probability density function (pdf) over the variables of interest [212]. Various techniques for efficiently estimating the densities exist, however they all impose a trade-off between computational complexity and accuracy [278]. Therefore, in most practical cases, the direct use of mutual information is not feasible. The method we use in this thesis for computing mutual information (for comparative purposes) is based on a data-interpolation technique known as Parzen-window density estimation, also known as kernel density estimation [286, 346]. In order to calculate mutual information between two variables we need to obtain an estimate for three separate pdfs, namely $p(x)$, $p(y)$ and the joint pdf $p(x,y)$. For n samples of a variable x, the approximate pdf, $\hat{p}(x)$, can be written as [286]:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(x - x^{(i)}, h\right) \tag{2.15}$$

where $K(\cdot)$ is the Parzen window function (a Gaussian window is often chosen), $x^{(i)}$ is the i-th data sample, and h is the window size. Parzen proved that the estimated pdf, $\hat{p}(x)$, can converge to the true pdf, $p(x)$, when n approaches infinity, provided the function $K(\cdot)$ and the window size, h, are properly selected [278]. Similarly, the joint pdf, $p(x,y)$, can be approximated by using a multivariate Parzen window, for example a multivariate Gaussian window. By definition, mutual information estimates obtained using this, and other similar approaches, will not be normalised, which makes it difficult to ascertain the actual level of dependence between a set of variables[2]. Therefore, in order to compare mutual information with other commonly used measures of statistical dependence, we need to rescale the original mutual information values (I). This can be achieved by converting each value into a normalised mutual information coefficient (I_N), given by [108]:

$$I_N = \sqrt{1 - \exp(-2I)} \tag{2.16}$$

It is important to stress here that for many non-Gaussian signals, mutual information is very difficult to compute accurately, as there will always be some error in the approximated pdfs, especially the joint pdf in high-dimensional spaces. It is also important to remember that most real-world data sets, including financial returns, are finite in size, which can again effect the estimated value of the pdfs. To calculate mutual information, we also need to compute an integral over the pdfs, which can be computationally complex for large data sets. Hence, direct use of mutual information has very limited practical utility, especially when dynamically analysing data in an online environment.

Copula functions

Another widely used approach for multivariate symmetric interaction measurement, which overcomes some of the issues of computational complexity associated with estimating the joint distributions in high-dimensional spaces, is the use of copulas [115]. Copulas are functions

[2]It is a well-known fact that zero correlation does not imply independence [224]. In contrast, the mutual information between a set of variables is zero only if the variables are mutually independent. For completely dependent signals, I has no upper bound, i.e. mutual information varies in the range $0 \leq I \leq \infty$.

which join (or couple) multivariate distribution functions to their one dimensional marginal distribution functions [72, 269]. In recent years, there has been a significant interest in the application of copula functions for modelling statistical dependencies in multivariate financial time series [115], including FX spot returns [103, 236]. Copula functions can not only model dependencies between data sets with non-Gaussian distributions, but also capture all scale-free temporal dependencies in multivariate time series. However, in practise it can be very difficult to accurately compute statistical dependencies using copula functions. This is because computation using copula functions involves calculating multiple moments as well as integration of the joint distribution, which requires use of numerical methods and hence becomes computationally complex [201]. Copula-based methods suffer from other major limitations as well, namely the difficulties in the accurate estimation of the copula functions, the empirical choice of the type of copulas, as well as problems in the design and use of time-dependent copulas [126].

2.2.1 Summary: Symmetric interaction measurement approaches

Our discussion in this section has pointed out various limitations associated with the practical use of most standard symmetric interaction measurement approaches, which we summarise in Table 2.1. These limitations make most of these approaches unsuitable for real-time dynamic analysis of multivariate financial time series (especially those sampled at high frequencies).

As discussed, mutual information is the canonical measure of symmetric interaction. We also noted the practical difficulties (and high computational cost) associated with the accurate computation of mutual information. However, as we describe in the next chapter, it is possible to use a proxy measure, based on independent component analysis (ICA), to estimate mutual information in a computationally efficient framework. ICA is a blind source separation tool which makes use of higher-order statistics and is therefore able to capture information in the tails of multivariate distributions. It can be used to extract M mutually independent sources, $\mathbf{a}(t) = [a_1(t), a_2(t), ..., a_M(t)]^\mathsf{T}$, from a set of N observed signals, $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_N(t)]^\mathsf{T}$, such that the mutual information between the recovered source signals is minimised, i.e.:

$$\mathbf{a}(t) = \mathbf{W}\mathbf{x}(t) + \mathbf{n}(t) \tag{2.17}$$

Approaches	Limitations
-Linear correlation	Can only accurately measure linear symmetric interactions between signals with Gaussian distributions [171], and is very sensitive to outliers [102]. Therefore, second-order approaches, such as correlation functions, are not suitable for capturing interactions in multivariate financial returns, which often have heavy-tailed, skewed, distributions [89].
-Rank correlation	Only valid for monotonic functions and computationally demanding for large data sets. Not suitable for analysing data which may have tied ranks (such as financial returns) [141]. Use of ranks (instead of the actual values) can also potentially cause a loss of information from the data being analysed, leading to inaccurate results [177]. A detailed review of other practical limitations is presented in [225].
-Mutual information	Difficult to accurately estimate directly using finite data sets, especially in high-dimensional spaces; this is because computation of mutual information requires the estimation of multivariate joint distributions, a process which is unreliable (being exquisitely sensitive to the joint pdf over the variables of interest) as well as computationally expensive [212].
-Copula functions	Computation of copula functions involves calculating multiple moments as well as integration of the joint distribution, which can become computationally complex [201]. Results obtained can be very sensitive to the empirical choice of the type of copula. Design and use of time-dependent copulas can also be problematic [126].

Table 2.1: Summary of limitations of standard symmetric interaction measurement approaches.

where $\mathbf{W}$ is the ICA unmixing matrix which encodes dependencies and $\mathbf{n}(t)$ is the observation noise. ICA has wide-ranging applications, including audio signal analysis, image feature extraction and econometrics [189, 337]. Our choice of using ICA is based on our hypothesis that many real-world signals, including multivariate financial returns, are likely to be generated as a result of linear mixing of a set of mutually independent non-Gaussian source signals. ICA can be used to obtain information about this mixing process which gives rise to a set of observed signals [35]. We use this information to develop a set of symmetric and asymmetric interaction measurement approaches, which we present later in the thesis.

2.3 Approaches to asymmetric interaction measurement (causality)

The symmetric interaction measurement approaches discussed in the last section provide no direct information about the flow of information in a system at non-zero time lags. Knowledge about the asymmetric flow of information between a set of variables makes it possible to use one (or more) of these variables to improve the predicted future values of the others, i.e. it becomes possible to infer causation between the variables. Before going any further, let us first clarify the difference between symmetric and asymmetric dependencies. It is a well-known fact that correlation does not necessarily imply causation [174]. Although it is possible for a pair of correlated signals to be causal, this is not always the case. Commonly used measures of statistical dependence, such as linear correlation, mutual information, etc., are symmetric and therefore cannot be directly used for detecting causal relationships between variables [347]. In contrast, most causality detection (asymmetric interaction measurement) approaches are based on analysis of improvement in predictive accuracy of a variable by incorporating prior temporal information about other variables. Causality detection approaches have a wide variety of applications in financial data analysis, some of which we now discuss. Univariate financial time series, e.g. the spot rate of a currency pair, generally do not contain much extractable information that can be used independently for predictive purposes [113]. However, the way different financial instruments asymmetrically interact can be used to improve our understanding of what drives the dynamics of financial markets and to develop more accurate forecasting models. The speed at which new information is captured and reflected in the prices of different financial instruments also induces causal effects in financial data. Therefore, having knowledge about the strength and direction of causation between a set of signals can be very useful. There is a vast amount of work which has been done in this area. Here we provide a list of studies which give us a flavour of the type of financial applications which make use of these approaches. In [166], the authors examine the causal relationship between stock volume flow and price. A study examining the influence of financial development on economic growth in presented in [223]. An interesting study describing the use of wavelets for inferring frequency-dependent causality between energy consumption and economic growth is presented in [275]. Likewise, two independent studies focusing on causality detection be-

tween FX currency pairs are presented in [26, 127]. These references show the wide variety of financial applications, ranging from macro-economic forecasting to detecting causal links in high-frequency multivariate data, for which causality detection approaches can be used in practise.

For most practical applications, any causality analysis approach should not only be able to detect the presence of causality between a set of signals, but also infer its direction and strength, and how both these quantities dynamically change with time, possibly across different time-scales[3]. There are three commonly accepted conditions for inferring the presence of a causal structure in multivariate time series, namely time precedence, the presence of dependence or relationship, and non-spuriousness of the data [205]. These three conditions are briefly described below. The first property of time precedence implies that the "cause" must precede the "effect" in time. For example, if a hypothesis states that a variable X causes Y, i.e. $X \rightarrow Y$, then for all values of time, t, the relationship $X_t \rightarrow Y_{t+t'}$ must hold, where $t' > 0$. Therefore, unlike many other measures of statistical dependence, e.g. linear correlation, mutual information, etc., a causal measure is usually asymmetric, i.e. $X \rightarrow Y \neq Y \rightarrow X$. Secondly, the constituent variables of any causal system must be interdependent. The observed relationship has to be statistically significant in order to minimise the likelihood of the relationship being present as a result of random variations in the data sets. The third condition of non-spuriousness is one which is most challenging to accurately infer using only real-world empirical data [105]. It implies that for any two variables X and Y to be causal, there should not be a third (unobserved) variable Z which is driving both X and Y independently, hence, leading to the presence of a causal connection between X and Y. However, as presence of a spurious variable (in this case Z) can never be excluded empirically when using real data, therefore, in practise this condition is often ignored when testing for causation [105]. It is also important to distinguish between a spurious variable and an intervening variable [205]. An intervening variable is one which is being driven by one variable and at the same time is driving another. For example, in the relationship, $X \rightarrow Z \rightarrow Y$, Z is referred to as the intervening variable. In contrast to a spurious variable, an intervening variable does not necessarily break

[3]Analysis of causality in temporal data streams at multiple frequencies can lead to interesting results by extracting multiple hidden causal relationships. In this respect, we can make use of directed-coherence analysis approaches to extract frequency-domain information about the presence of causal links in multivariate time series. We further discuss these approaches later in the thesis.

a causal relationship, i.e. in the above example X and Y are still causally linked, albeit the link is indirect.

There is a vast amount of literature available which addresses the issue of inferring causation between a set of signals [150, 281]. However, most standard causality detection approaches have certain limitations with respect to their computational complexity or the type of data sets they are suitable for. These limitations make most of these approaches unsuitable for real-time dynamic analysis of financial data, especially data sampled at high-frequencies. The goal of this section is to critically analyse and compare the practical benefits and limitations of some commonly used causality detection approaches, with the aim of providing a foundation for developing new, computationally efficient, models for inferring causation between multivariate financial returns. We make use of the introductory material presented in this section as the basis for the development of causality analysis approaches later in the thesis.

2.3.1 Model-based approaches

Most standard causality detection approaches are primarily based on the multivariate autoregressive (MAR) model. An autoregressive (AR) model describes the data-generation process of a univariate time series as a weighted linear sum of a pre-defined number of its previous values. A MAR model extends the AR model by describing the data-generation process of multiple time series in a similar manner. For the bivariate case, a MAR model can be represented by the following set of equations:

$$x(t) = \sum_{j=1}^{p} \alpha_{11,j} x(t-j) + \sum_{j=1}^{p} \alpha_{12,j} y(t-j) + e_x(t) \tag{2.18}$$

$$y(t) = \sum_{j=1}^{p} \alpha_{21,j} x(t-j) + \sum_{j=1}^{p} \alpha_{22,j} y(t-j) + e_y(t) \tag{2.19}$$

where p is the model order, $x(t)$ and $y(t)$ are instances of variables X and Y respectively at time t, and α's represent the model parameters (weights), which can be estimated using standard least squares or other similar approaches. Using these equations, the MAR model parameters for a 2-dimensional system, at time lag j, can be presented in matrix form as:

$$\mathbf{W}_j = \begin{bmatrix} \alpha_{11,j} & \alpha_{12,j} \\ \alpha_{21,j} & \alpha_{22,j} \end{bmatrix} \tag{2.20}$$

The off-diagonal elements of this matrix contain information about the degree of asymmetric dependencies between the two variables at any given time lag, and hence, can be used for cross-variable causal inference. We now briefly describe some standard causality detection approaches based on the MAR model.

Linear Granger

The linear Granger causality model, that was originally proposed by Clive Granger [146], is one of the most commonly used causality detection approaches in practise, in part due to its simplicity and computational efficiency. Granger causality originated from the idea that two time series are causal if the linear predictability of one of them is improved by incorporating prior information about the second time series in the analysis. Numerous academic studies have demonstrated utility of this approach for financial applications [60, 80, 166]. A standard Granger causality regression model is based on the MAR model [43], and for the bivariate case can be represented by (2.18) and (2.19). Using these equations, the variable Y is said to Granger-cause X if the values of α_{12} in (2.18) are significantly different from zero; similarly, $X \rightarrow Y$ if values of α_{12} in (2.19) are significantly non-zero. The non-zero values of α_{12}, for example, indicate that the past values of Y have a positive impact on the predictability of X. We can also describe causality in terms of the prediction errors, $e_x(t)$ and $e_y(t)$. In other words, if the variance of the error terms for one of the variables is reduced by including past terms from the other variable in the regression equation, then the second variable is said to Granger-cause the first one. A commonly used method of inferring the presence of causality is to use the F-statistic, which can be used to compare the relative size of two population variances [37]. However, using Granger causality has certain limitations. Firstly, as it is based on a linear regression model, therefore it is only valid for linear systems. Secondly, a Granger causality model making use of standard parameter estimation approaches, such as ordinary least squares (OLS), suffers from model overfitting and is based on the assumption of normality of regression residuals, hence, it is unable to accurately analyse non-Gaussian signals [207, 238]. This is a major limitation, as many real data sets, including financial returns, have non-Gaussian distributions [98]. Moreover, the Granger causality model can only be accurately used for stationary time series. One possible way of overcoming this limitation is to use a windowing technique, whereby causation is inferred within short-time windows of

the data with the assumption that within each window the data can be considered to be locally stationary [107, 165].

Non-linear Granger

It is possible to extend the linear Granger causality model to analyse non-linear causal relationships in multivariate time series. Non-linear causality analysis can have important financial applications, e.g. in [18] the authors show that non-linear causality exists between various currency future returns, while in [166] the authors discover non-linear causal links between stock price and volume data. Although it is also possible to detect non-linear causal links using information theocratic approaches (which we discuss later), here we briefly describe a model-based approach which directly extends the linear Granger causality model by using a non-linear prediction model. This approach makes use of radial basis functions (RBF) for non-linear Granger causality analysis [11, 58][4], and is essentially based on a two-step algorithm. In the first step, RBFs are used to map the observed signals onto a kernel space (using e.g. Gaussian kernels), and in the second step a MAR model is used to analyse the mapped signals in order to test for the presence of non-linear causal links by comparing the variances of the regression residuals (using e.g. the F test). We describe this approach in more detail later in the thesis. Theoretically, the non-linear Granger causality analysis framework discussed above is straightforward. However, when analysing real-world data sets for practical applications, it has some limitations. Results obtained using the model can be very sensitive to the selection and tuning of various parameters, such as the number of RBFs or the value of RBF scaling parameter [193]. The algorithms can also result in high computational load [283], which limits their practical utility, especially when dealing with high-frequency data. Moreover, results obtained can be (at times) difficult to interpret, as discussed in [25].

Cointegration

In the log-returns space, financial time series can be considered to be locally stationary [107, 165], therefore the causality analysis approaches presented so far, which are based on the assumption of stationarity, are suitable for analysing financial log-returns. However, if a large

[4] A RBF is a function which satisfies the equality $\phi(\mathbf{x}) = \phi(||\,\mathbf{x}\,||)$, i.e. its value depends only on the distance from the origin or any other point. The Euclidean distance is usually used as the norm; however, other measures of distance can also be used.

data set needs to be analysed in order to infer causation or financial mid-prices are used, the assumption of stationarity may not always hold. This gives rise to the need for a non-stationary causality analysis method. One of the popular approaches to analyse non-stationary causal systems is referred to as cointegration [119], which won its developers Clive Granger and Robert Engle the Nobel prize in economics in 2003. A detailed description of cointegration is presented in [163], here we only provide a very brief introduction. A set of non-stationary signals are said to be cointegrated if their linear combination is stationary. Hence, the two non-stationary time series $x(t)$ and $y(t)$ in the following equation are cointegrated if the regression residuals $e(t)$ are stationary in the long-run:

$$y(t) - \alpha x(t) = e(t) \tag{2.21}$$

where α's represent the regression parameters. It is possible to test for cointegration using the two-step Engle-Granger approach [119]. In the first step, the OLS values of the regression parameters ($\hat{\alpha}$) and the associated values for the regression residuals ($\hat{e}$(t)) are estimated, and in the second step the Augmented Dickey–Fuller test (or any other similar test) is used to test for stationarity of the residuals; if the residuals are stationary, then the two variables are said to be cointegrated. For two (or more) cointegrated variables, the lagged values of the regression residuals, $e(t-1)$, can be used as an error correction term in a *vector error correction* model, hence, making it possible to study short-run dynamics of the system as well [119]. If cointegration exists, then at least one directional causality exists [146]. The concept of cointegration can best be explained using a simple example. Consider a set of three variables, X, Y and Z, all of which denote the share price of three different oil companies. All three variables will generally follow a common long-run path, effected by the global oil price [163]. Any of these three variables can be cointegrated if their linear combination is stationary, which is quite possible in this case. It is also possible that one of these stock prices changes before the other two, i.e. the variable is exogenous to the other two variables. This variable may actually be causally driving the other two variables. So, although cointegration can show the presence of causality between a set of variables, it cannot be straightforwardly used to infer the direction of causation [163]. Having discussed some advantages of the cointegration approach, we now describe some of its limitations. Cointegration is a long-run property, therefore it is best

applied to long spans of non-stationary data [154]; this makes it unsuitable for dynamically analysing multivariate financial log-returns (which are locally stationary). Moreover, although most cointegration algorithms can detect the presence of causality, they cannot be directly used to infer the direction of causation. A more in-depth discussion focusing on various limitations of the cointegration approach for causal inference in financial systems is presented in [152].

2.3.2 Model-free approaches

Even though the model-based causality detection approaches described so far are computationally efficient, they make implicit assumptions about the data generation process and most assume normality of regression residuals. They also have to be matched to the underlying dynamics of the system in order to avoid giving rise to spurious causality values due to model misspecification [170]. It is possible to address some of these limitations by developing causality detection approaches within an information theoretic framework [105, 170]. We briefly review some of these approaches here.

Transfer entropy

Conditional entropy, as given by (2.11), is an asymmetric measure. However, it is asymmetric only due to different individual entropies rather than actual information flow [315]. Based on conditional entropy, a better measure for computing the actual flow of information can be derived, which also fulfils the condition of time precedence. This measure is often referred to as the transfer entropy, and for discrete-valued processes X and Y is given by [315]:

$$T_{E,X\to Y} = \sum p(y(t), \mathbf{y}_n, \mathbf{x}_m) \log \left[\frac{p(y(t) \mid \mathbf{y}_n, \mathbf{x}_m)}{p(y(t) \mid \mathbf{y}_n)} \right] \tag{2.22}$$

where $\mathbf{x}_m = [x(t-1), \ldots, x(t-m)]$, $\mathbf{y}_n = [y(t-1), \ldots, y(t-n)]$ and m and n are the model orders (it is common practise to set $m = n$). Transfer entropy can calculate actual transported information rather than information which is produced due to the effects of a common driver or past history, which is the case with mutual information and other similar time-lagged information theoretic measures [315]. Although conceptually straightforward, information theoretic measures such as transfer entropy have some major limitations in their practical use. These include complexity of their computation, especially in high-dimensional spaces. This is due to the fact that these approaches are exquisitely sensitive to the joint pdf over the variables of

interest. Various techniques for efficiently estimating the densities exist, however they all impose a trade-off between computational complexity and accuracy [278]. These measures also require large data sets for accurate estimation of the pdfs. Moreover, transfer entropy based approaches are currently restricted to only bivariate systems [231].

Correlation integral

It is possible to express transfer entropies in terms of generalised correlation integrals. Correlation integral based entropies are a nonparametric measure and make minimal assumptions about the underlying dynamics of the system and the nature of the coupling [73]. Therefore, they do not make any assumption about the deterministic processes underlying the time series. Correlation integrals were originally designed for determining the correlation dimensions [149], and are often used for analysing systems with non-linear dynamics [179, 206], being suitable for almost any stationary and weakly-mixed system [297]. The generalised correlation integral (of order $q = 2$) can be expressed as [73]:

$$C(\mathbf{x}, \varepsilon) = \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \Theta(\varepsilon - ||x_j - x_i||) \tag{2.23}$$

where T is length of the time series, Θ is the Heaviside function (although a Gaussian or some other kernel can be used as well), $||\cdot||$ denotes the maximum norm, and ε is the length scale (denoting the radius centred on the point x_i). The expression $\frac{1}{T} \sum_{j=1}^{T} \Theta(\varepsilon - ||x_j - x_i||)$ is a form of kernel density estimator which gives us the fraction of data points within a radius ε of the point x_i. Further details about the properties of correlation integrals with different choices of q and ε can be found in [73, 297]. Denoting the amount of information about $y(t)$ contained in its past n values $\mathbf{y}_n = [y(t-1), ..., y(t-n)]$ by $I_C(\mathbf{y}_n; y(t))$, and likewise the amount of information about $y(t)$ contained in its past n values and the past m values of another variable $\mathbf{x}_m = [x(t-1), ..., x(t-m)]$ by $I_C(\mathbf{x}_m, \mathbf{y}_n; y(t))$, the information theoretic measure of Granger causality can be written as the gain in information about the current value of one variable by including information about the past values of another variable [117, 170]:

$$\begin{aligned} I_{C,X\to Y} &= I_C(\mathbf{x}_m, \mathbf{y}_n; y(t)) - I_C(\mathbf{y}_n; y(t)) \\ &= \log C(\mathbf{x}_m, \mathbf{y}_n, y(t)) - \log C(\mathbf{x}_m, \mathbf{y}_n) - \log C(\mathbf{y}_n, y(t)) + \log C(\mathbf{y}_n) \end{aligned} \tag{2.24}$$

where the variable X is said to Granger-causes Y if $I_{C,X\to Y} > 0$; to check if this statistic is significantly greater than zero, a bootstrapping procedure can be used, as discussed in [105, 173]. It is common practise to use the following statistic (which is derived using the same approach as previously described) as a non-parametric Granger causality test based on correlation integrals [24, 166]:

$$T_{C,X\to Y} = \frac{C(\mathbf{x}_m, \mathbf{y}_n, y(t))}{C(\mathbf{x}_m, \mathbf{y}_n)} - \frac{C(\mathbf{y}_n, y(t))}{C(\mathbf{y}_n)} \tag{2.25}$$

$T_{C,X\to Y}$ will have a high value if X contains information about the future values of Y; further information about using this test statistic for inferring causation in real data sets is given in [106, 166]. However, correlation integral based approaches have some disadvantages to their practical use. Firstly, they are very sensitive to the presence of noise [316]. Secondly, using data which is even slightly autocorrelated can considerably effect accuracy of the correlation integral estimates [273]. Thirdly, many correlation integral based algorithms in common use are computationally demanding [92], limiting their use for analysing data sampled at high-frequencies.

Causal conditional mutual information

It is possible to gain information about the predictability of one time series based on the past values of another using the principles of mutual information, an approach generally called predictive mutual information [292]. Predictive mutual information simply refers to the mutual information between one time series, $x(t)$, and the lagged values of another $y(t-\tau)$, i.e. $I[x(t), y(t-\tau)]$, where τ denotes the time lag. However, this approach makes use of only the instantaneous values of one time series and the past values of another to infer the presence of a causal relationship between them. Causal conditional mutual information is a more accurate and robust method, similar to predictive mutual information, as discussed in [292]. It infers the presence of a causal structure by making use of the conditional mutual information of $x(t)$ and $y(t-\tau)$ conditioned on $x(t-\tau)$ and $y(t)$. This can be written as $I[x(t), y(t-\tau) \mid x(t-\tau), y(t)]$, and is evident from the graphical model presented in Figure 2.1. However, as previously discussed, measuring interactions using information theoretic approaches can be computationally complex and often give unreliable results when using finite data sets.

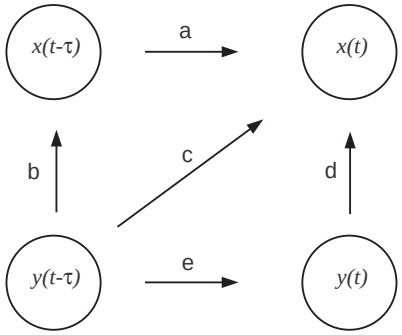

Fig. 2.1: Graphical representation of causal conditional mutual information. The arrows represent the direction of information flow. Arrows (a) and (e) represent predictive self-information, arrows (b) and (d) represent zero-lag cross-information, while arrow (c) represents the predictive cross-information [292].

Causal Bayesian network

Graphical models can provide a useful framework for analysing and understanding a causal network. They can not only facilitate efficient inferences from observations, but also enable us to represent joint probability functions in a straightforward way [280]. A Bayesian network is a graphical representation of probability distributions, often represented by a directed acyclic graph. The nodes of the network represent random variables, while the links represent the conditional probability of two variables. Standard Bayesian networks are only suitable for static data; for time series analysis we need to make use of dynamic Bayesian networks [260]. A major problem with the practical use of dynamic Bayesian network algorithms is their high computational cost [128], especially when analysing high-dimensional data sets [91]. They also require large data sets for increased accuracy, which makes them unsuitable for dynamic causal inference [162]. Moreover, as discrete Bayesian networks usually make use of combinatorial interaction models, therefore it can be difficult to accurately determine the relative magnitude and direction of causal interactions [358].

2.3.3 Summary: Asymmetric interaction measurement approaches

It is important to note that there is no universal causality analysis model. Each of the approaches presented so far has some limitations, as summarised in Table 2.2. These limitations make most of these approaches unsuitable for real-time dynamic analysis of multivariate financial data.

Our discussion in this section has highlighted the computational complexities associated

Approaches	Limitations
Model-based:	
-Linear Granger	Can only accurately measure linear causal relationships between data sets with Gaussian distributions and is very sensitive to outliers [28]. Suffers from the model overfitting problem [238].
-Non-Linear Granger	Various parameters, e.g. the number of RBFs or the value of the RBF scaling parameter, need to be estimated and tuned [193]. Algorithms can be computationally demanding [283]. Results obtained can be (at times) difficult to interpret [25].
-Cointegration	Cointegration is a long-run property, therefore it is best applied to long spans of data [154]. Moreover, most cointegration algorithms cannot be used to infer the direction of causation. An interesting study focusing on the limitations of the cointegration approach for inferring causation in bivariate financial time series is presented in [152].
Model-free:	
-Transfer entropy	Based on information theoretic measures which are computationally complex to analyse and require large data sets for accurate estimation. Moreover, transfer entropy based methods are currently restricted to bivariate time series [231].
-Correlation integral	Correlation integral based approaches are very sensitive to the presence of noise [316]. Also, even slightly autocorrelated data can considerably effect accuracy of the correlation integral estimates [273]. Moreover, correlation integral based algorithms can be computationally demanding [92].
-Causal conditional mutual information	Based on information theoretic measures which are computationally complex to analyse and require large data sets for accurate estimation.
-Causal Bayesian network	Require large data sets for best performance [162]. However, analysing large multivariate data sets with dynamic Bayesian networks can result in high computational load [91]. As discrete Bayesian networks usually make use of combinatorial interaction models, therefore it can be difficult to accurately determine the relative magnitude and direction of causal interactions [358].

Table 2.2: Summary of limitations of standard causality analysis approaches.

with the practical use of most model-free causality detection approaches. On the other hand, model-based approaches are computationally efficient and simple to use, which are some of the primary factors for their widespread use in practise. These approaches are often based on

the MAR model and have been in use for many years, the Granger causality model being one such example. However, a standard linear Granger model assumes Gaussian residuals and can suffer from the model overfitting problem. To address these limitations, we present a set of causality detection approaches later in the thesis which make use of the MAR model as their foundation. One of these, which we call Granger independent component (GIC) causality, is based on a combination of the principles of ICA, MAR and Granger causality. It allows us to efficiently analyse causal links in multivariate non-Gaussian data sets and is therefore suitable for dynamically analysing financial returns, which often have non-Gaussian distributions. The second method, which we call variational Granger (VG) causality, is based on variational Bayesian MAR models. It prevents model overfitting by estimating the MAR model parameters within a Bayesian setting, hence, making it possible to accurately infer causation in multivariate data sets.

2.4 Approaches to dynamic interaction measurement

Due to the continuously evolving state of interactions in financial markets, most practical financial applications require use of computationally efficient models which can be used in an online dynamic environment. In this section we present the use of some approaches to model, and to capture, the complex interaction dynamics of multivariate financial time series across both time and scale. Most of the statistical models presented in this thesis can be used effectively within a causal framework. Denoting the current time by t_c, a causal system is one in which the output value of a given function $f(t)$ only depends on the current and prior values of the input time series, i.e. $t \leq t_c$. Hence, a causal model cannot "look" into the future and only uses the data available up to time $t = t_c$.

Windowing techniques

Many practical statistical inference problems involve making use of real-time sequential data to infer a specific quantity, such as any given measure of interaction or for forecasting purposes. To get an idea of the temporal variations in the dependency structure of such data sets, windowing techniques have to be used. As financial log-returns are locally stationary [107, 165], therefore using windowed data also enables us to make use of statistical models meant solely for stationary data sets. Some popular windowing techniques, together with their

practical advantages and limitations, are discussed below.

It is common practise to use a sliding-window to dynamically model the changing properties of a single or multiple time series. The simplest method makes use of a window of fixed-length which slides through the data at fixed intervals, usually one datum at each step. This involves updating some or all of the model parameters at each time step, using a window in which a fixed number of past data points are used to estimate the parameters. We now describe some major criteria which need to be considered when selecting the appropriate window length for a fixed-length sliding-window. The window should be large enough so as to accurately capture variation of the signals within it. However, a large window also may not be able to capture rapid changes in the dynamics of the system and may result in computational times which lead to undesirable latency effects (an important consideration when high-frequency data is being analysed in real-time). Therefore, the window should be small enough so as to accurately compare disparity of the signals at corresponding points [272], without leading to "noisy" results. However, a very small window may not contain enough data points for accurate computation of the dependency measure. Hence, any potential high-frequency inference model carries this complexity-benefit trade-off with respect to the choice of window length.

In most cases, using a simple fixed-length sliding-window technique is ample; however, for certain applications (such as those making use of asynchronous data) it may be more useful to use a slightly more complex approach which offers a possible compromise between the two conflicting criteria (as discussed above) for selecting the window length. This approach is based on using an adaptive window whose size changes with time depending on some properties of the signals. Various algorithms have been proposed addressing this issue [40, 272]. A common approach to systematically update the window size is to use Kalman filtering techniques, as discussed in [39]. Kalman filters provide a computationally efficient framework for dynamically managing the window length for the purpose of learning from data streams. The windowing techniques discussed so far give equal importance to all data points within each window, irrespective of their temporal placement. We now describe a method which relaxes this condition. This windowing technique, known as exponentially weighted moving average (EWMA), resembles an infinite impulse response in which the weightings of each preceding time step decrease exponentially with time. Although EWMAs have been used for modelling

the dynamics of financial data [136], they have some limitations to their use. An EWMA based model uses all available data. This can be a disadvantage, as financial returns have evolving dynamics and a single major shock in the market can potentially continue to effect results for some time to come if an EWMA window is used for interaction analysis. Also, an EWMA window requires selection of a weight parameter (that represents the rate at which the weightings decrease) to be set beforehand, which can effect results [182]. It is also possible to capture dynamics of a systems using adaptive EWMAs, in which the weights of the EWMA change with time. However, this approach also suffers from problems associated with estimating the parameters of the EWMA [335].

Scale-based dynamics

Having reviewed suitable approaches for analysis of time-based dynamics of multivariate data, we now address the issue of analysis of scale-based dynamics. The dependency structure in multivariate financial time series changes with the frequency at which data is being analysed [56]. Studies focusing on analysis of financial time series using a multiple time-scale approach show promising results [56, 256, 274]. Knowledge gained regarding the properties of a time series at different time-scales can be used to build models that measure interactions at different frequencies (as discussed later in the thesis). There are a variety of methods that can potentially be used to determine the time-frequency representation of financial time series. The short-time Fourier transform (STFT) [68, 90], empirical mode decomposition (EMD) and the Hilbert-Huang transform [181, 282] are all popular time-frequency analysis techniques. However, our method of choice for time-scale decomposition of financial data is wavelet analysis [267, 331], due to the reasons discussed below.

Wavelet analysis is capable of revealing aspects of data that other time-scale analysis techniques miss, including trends, breakdown points, discontinuities, and self-similarity [3]. It can be used to analyse signals presenting fast local variations such as transients or abrupt changes; this makes them well-suited for analysing financial returns which exhibit similar properties (as we discuss in detail later in the thesis). Wavelet analysis is different from other related techniques as in wavelet analysis both the time window and the enclosed waveform are scaled together, whereas for other methods, e.g. STFT, the window length is kept constant and only the enclosed sinusoid is scaled. A wavelet can therefore localise itself in time for short du-

ration, i.e. high-frequency, fluctuations [6]. EMD, while similar to wavelet analysis in many aspects, has the disadvantage that it cannot separate some low-energy components from the signal being analysed, therefore these low-energy components may not appear in the time-frequency plane [287]. By using wavelets within a Bayesian framework, prior information about system dynamics can be included as a prior distribution on the wavelet coefficients [4]. Thus, any prior knowledge about the system dynamics can be easily incorporated into the model. In recent years, many studies have shown the advantages of using wavelets for decomposing univariate financial time series [42, 261]. In the next chapter, we return to discuss the utility of using wavelets for scale-based analysis of interactions in multivariate financial time series.

2.5 Concluding remarks

Having critically reviewed existing literature in this chapter, we are now in a position to describe the main objectives of this thesis. There is currently a lack of availability of suitable non-Gaussian measures of interaction which can be used to accurately model symmetric and asymmetric interactions in multivariate financial time series in a computationally efficient manner. The first major objective of this thesis is to address this problem to some extent by presenting a set of symmetric and asymmetric multivariate interaction measurement approaches, which can be used to dynamically analyse dependencies in financial data streams in a computationally efficient framework. Our second objective is to apply these (and other existing) approaches to a set of practical financial problems in order to extract interesting and useful information from multivariate financial time series.

We hypothesise that multivariate financial data may be generated as a result of linear mixing of some non-Gaussian latent variables. Therefore, we aim to rely on the data analysis power of some blind source separation tools which take into account the higher-order statistics of the data under analysis. To accurately and efficiently measure symmetric interactions, we aim to develop a statistical information coupling metric as a proxy for mutual information, which can be used to dynamically analyse dependencies in multivariate non-Gaussian data streams. Many real-world signals, including financial returns, exhibit time-scale behaviour. This can potentially result in noticeable changes in dependencies between signals at different

frequencies. To analyse this effect, we aim to extend our interaction models by making use of time-scale analysis methods that are best suited for use in financial applications. In high-dimensional spaces, analysing interactions between financial instruments can be problematic. We therefore aim to make use of network analysis approaches to develop static and dynamic complex coupling networks. Understanding the nature and strength of asymmetric interactions, i.e. causal relationships, in multivariate financial data streams is of utmost importance for gaining an insight into the dominant factors affecting complex financial systems and for developing improved forecasting models. We therefore aim to develop computationally efficient causality detection models that are suitable for analysing multivariate financial returns. Using a range of practical financial case studies, we aim to show the effectiveness, utility and relative accuracy of the interaction measurement approaches presented in this thesis by extracting interesting and useful information from multivariate financial time series. We now proceed to present an ICA-based information coupling model (and its extensions) in the next chapter.

Chapter 3
Information coupling: A new measure for symmetric interactions

Our discussion so far in this thesis has pointed out various limitations associated with the practical use of standard interaction measurement approaches for the purpose of dynamically analysing multivariate financial time series. In this chapter, we present a computationally efficient independent component analysis (ICA) based approach to dynamically measure information coupling in multivariate non-Gaussian data streams as a proxy measure for mutual information. The chapter is organised as follows. We first discuss the need for developing an ICA-based information coupling model and present the theoretical framework underlying the development of our approach. Next, we present a brief introduction to the principles of ICA and discuss our method of choice for accurately and efficiently inferring the ICA unmixing matrix in a dynamic environment. We then proceed to present the ICA-based information coupling metric and describe its properties. Later in the chapter we present suitable approaches for analysing both static and dynamic complex coupling networks and for dynamically measuring information coupling across both time and scale. We demonstrate the practical utility and accuracy of the information coupling model (and its extensions) using a range of financial case studies in the next chapter.

3.1 Measuring interactions using ICA: A conceptual overview

Let us first take a look at the conceptual basis on which we can use ICA as a tool for measuring interactions. Mutual information is the canonical measure of symmetric interaction (dependence) in multivariate systems [94]. Whilst the computation of mutual information is conceptually straightforward when the full pdfs (the marginal pdfs as well as the joint pdf)

of the variables under consideration are available, it is often difficult to accurately estimate mutual information directly using finite data sets. This is especially true in high-dimensional spaces, in which computation of mutual information requires the estimation of multivariate joint distributions, a process which is unreliable (being exquisitely sensitive to the joint pdf over the variables of interest) as well as computationally expensive [212]. The accuracy of existing approaches to compute mutual information (which we described earlier) is also highly sensitive to the choice of the model parameters, such as the number of kernels or neighbours. Therefore, for most practical applications, the direct use of mutual information is not feasible. However, as we discuss below, it is possible to make use of information encoded in the ICA unmixing matrix to calculate information coupling as a proxy measure for mutual information.

According to its classical definition, ICA estimates an unmixing matrix such that the mutual information between the independent source signals is minimised [10]. Hence, we can consider the unmixing matrix to contain information about the degree of mutual information between the observed signals. Although the direct computation of mutual information can be very expensive, alternative efficient approaches to ICA, which do not involve direct computation of mutual information, exist. Hence, it is possible to indirectly obtain an estimate for mutual information by using the ICA-based information coupling measure as a proxy. Now let us consider some properties of financial returns which make them well-suited to be analysed using ICA[1]. Financial markets are influenced by many independent factors, all of which have some finite effect on any specific financial time series. These factors can include, among others, news releases, price trends, macroeconomic indicators and order flows. We hypothesise that the observed multivariate financial data may hence be generated as a result of linear combination of some hidden (latent) variables [23, 271]. This process can be quantitatively described by using a linear generative model, such as principal component analysis (PCA), factor analysis (FA) or ICA. As financial returns have non-Gaussian distributions with heavy tails, PCA and FA are not suitable for modelling multivariate financial data, as both these second-order approaches are based on the assumption of Gaussianity [184]. ICA, in contrast, takes into account non-Gaussian nature of the data being analysed by making use of higher-

[1] As this thesis is focused on financial applications, therefore we consider the case of measuring dependencies in financial data; however, similar ideas can be applied to most real-world systems which give rise to non-Gaussian data.

order statistics. ICA has proven applicability for multivariate financial data analysis; some interesting applications are presented in [23, 229, 271]. These, and other similar studies, make use of ICA primarily to extract the underlying latent source signals. However, all relevant information about the source mixing process is contained in the ICA unmixing matrix, which hence encodes dependencies. Therefore, in our analysis we only make use of the ICA unmixing matrix (without extracting the independent components) to measure information coupling. The ICA-based information coupling model we present in this chapter can be used to directly measure statistical dependencies in high-dimensional spaces. This makes it particularly attractive for a range of practical applications in which relying solely on pair-wise analysis of dependencies is not feasible[2].

3.2 Independent components, unmixing and non-Gaussianity

Mixing two or more unique signals, to a set of mixed observations, results in an increase in the dependency of the pdfs of the mixed signals. The marginal pdfs of the observed mixed signals become more Gaussian due to the central limit theorem [270]. The mixing process also results in a reduction in the independence of the mixed signal distribution and hence increase in mutual information associated with it. Moreover, there is a rise in the stationarity of the mixed signals, which have flatter spectra as compared to the original sources [304]. The block diagram in Figure 3.1 shows the ICA mixing and unmixing processes. Given a set of N observed signals $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_N(t)]^\mathsf{T}$ at the time instant t, which are a mixture of M source signals $\mathbf{s}(t) = [s_1(t), s_2(t), ..., s_M(t)]^\mathsf{T}$, mixed linearly using a mixing matrix $\mathbf{A}$, with observation noise $\mathbf{n}(t)$, as per (3.1):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{3.1}$$

Independent component analysis (ICA) attempts to find an unmixing matrix $\mathbf{W}$, such that the M recovered source signals $\mathbf{a}(t) = [a_1(t), a_2(t), ..., a_M(t)]^\mathsf{T}$ are given by:

[2]There is surprisingly little work done towards addressing the important issue of estimating the dependency structure in high-dimensional multivariate systems, although there has been interest in this field for a long time [198]. High-dimensional analysis of information coupling has various important applications in the financial sector, including active portfolio management, multivariate financial risk analysis, statistical arbitrage, and pricing and hedging of various instruments [126].

$$\mathbf{a}(t) = \mathbf{W}\left(\mathbf{x}(t) - \mathbf{n}(t)\right) \tag{3.2}$$

For the case where observation noise $\mathbf{n}(t)$ is assumed to be normally distributed with a mean of zero, the least squares expected value of the recovered source signals is given by:

$$\hat{\mathbf{a}}(t) = \mathbf{W}\mathbf{x}(t) \tag{3.3}$$

where $\mathbf{W}$ is the pseudo-inverse of $\mathbf{A}$, i.e.:

$$\mathbf{W} = \mathbf{A}^{+} = \left(\mathbf{A}^{\mathsf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathsf{T}} \tag{3.4}$$

In the case of square mixing, $\mathbf{W} = \mathbf{A}^{-1}$.

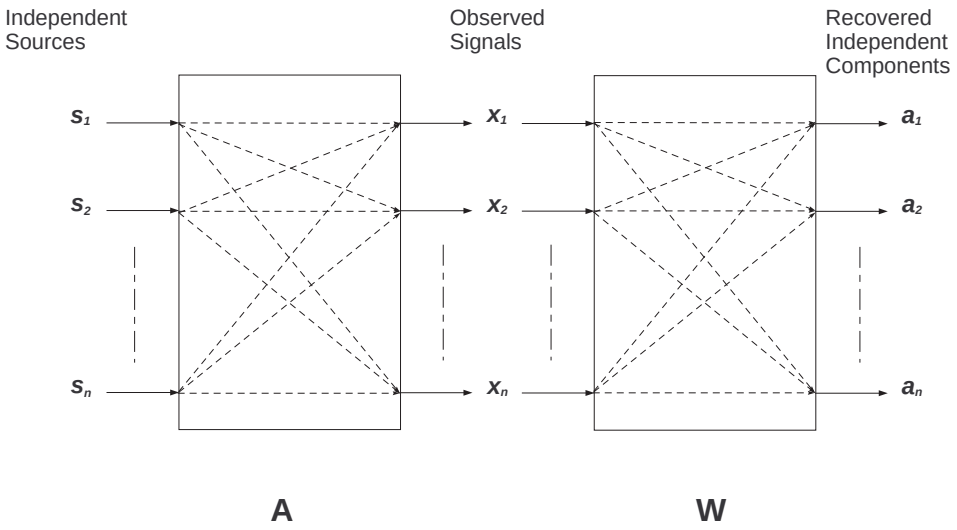

Fig. 3.1: ICA block diagram showing a set of observed signals, **x**, which are obtained by mixing the latent (unobservable) independent source signals, **s**, using an unknown mixing matrix, **A**. ICA obtains the unmixing matrix, **W**, using only the set of observed signals, **x**, such that the recovered independent components, **a**, are maximally statistically independent.

As an example, Figure 3.2 shows the results obtained when two non-Gaussian time series are randomly mixed, using a normally distributed random mixing matrix **A**, and then separated using ICA. The results clearly show the effectiveness of this blind source separation method in dealing with non-Gaussian data. This ability to handle non-Gaussian data is what distinguishes ICA from PCA, which is another well-known source separation approach. PCA, sometimes also known as the Karhunen-Loeve transform or the Hotelling transform, is one of the ear-

liest known factor analysis methods, and works by finding components which maximise the variance between a set of linearly transformed components [184]. It extracts principal components from a data set by choosing the eigenvectors associated with the highest eigenvalues of the covariance matrix of the data [325]. We note that PCA only uses second-order statistics for source separation, whereas ICA implicitly uses higher-order statistics; hence, PCA decomposes a set of observed signals into a set of decorrelated signals, whereas ICA extracts the maximally statistically independent source signals such that mutual information between the recovered source signals is minimised [330].

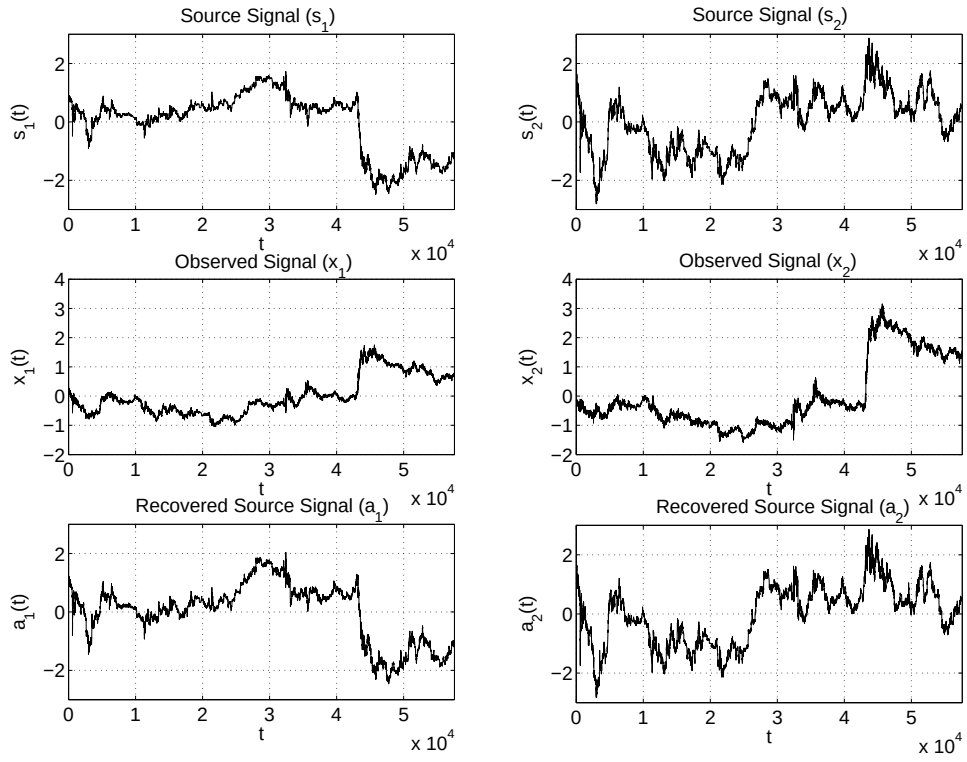

Fig. 3.2: Demonstration of effectiveness of ICA as a blind source separation tool. Two non-Gaussian sources (first row) are randomly mixed together to give the observed signals (second row). ICA is applied on these observed signals to extract the original independent sources (third row), without any prior knowledge of the original source signals or the mixing process.

There are various measures of independence which can be used to extract the independent source signals via an estimate for the ICA unmixing matrix, $\mathbf{W}$, some of which we briefly mention here. As previously discussed, mutual information is the canonical measure of dependence. For decoupled signals, mutual information is equal to zero, whereas for coupled signals it has a positive value. ICA can separate the underlying sources by finding an unmixing matrix,

W, which minimises the mutual information between the recovered source signals. However, as direct computation of mutual information can be computationally expensive and inaccurate for finite data sets, therefore this approach to ICA is rarely useful for practical applications. Another possible approach for estimating the unmixing matrix (and other parameters of an ICA model) is by directly maximising the log-likelihood of the data, which is equivalent to minimising the mutual information between the recovered sources [122]. The log-likelihood of a set of observations, $\mathbf{X} = [\mathbf{x}(t)]_{t=1}^{t=T}$, is given by [304]:

$$\ell(\mathbf{X} \mid \mathbf{A}, \boldsymbol{\theta}, \mathbf{R_n}) = \sum_{t=1}^{T} \log p(\mathbf{x}(t) \mid \mathbf{A}, \boldsymbol{\theta}, \mathbf{R_n}) \tag{3.5}$$

where T denotes the number of data points, $\boldsymbol{\theta}$ represents the parameters of the source models (if required) and $\mathbf{R_n}$ is the covariance of the white noise term, $\mathbf{n}(t)$. Thus, the variables $\mathbf{A}$, $\boldsymbol{\theta}$ and $\mathbf{R_n}$ are adjusted until the log-likelihood, ℓ, is maximised, giving us an estimate for the unmixing matrix $\mathbf{W} = \mathbf{A}^{+}$ [304]. Another approach to ICA, which directly makes use of the non-Gaussianity of source signals, is described below. ICA can only be accurately applied to those data sets which have at most one Gaussian source signal, while all other sources have non-Gaussian distributions. This property of ICA can be used to extract the independent sources from an observation set by finding a mixing matrix which minimises the Gaussianity of the recovered source signals. Negentropy is a measure of the non-Gaussianity of a signal, being zero for a Gaussian signal and always positive for a non-Gaussian signal. Thus, independent components can be obtained by maximising the negentropy, $J(\mathbf{x})$, of the recovered sources, given by [187]:

$$J(\mathbf{x}) = H(\mathbf{x}_G) - H(\mathbf{x}) \tag{3.6}$$

where $\mathbf{x}_G$ is a Gaussian random variable with the same covariance matrix as $\mathbf{x}$, and $H(\mathbf{x})$ is the entropy (as given by (2.9)). As computation of negentropy requires an estimation of the pdf of the data, it can be computationally expensive to compute directly. *FastICA* is a computationally efficient algorithm based on the concept of negentropy maximisation [189]. It is based on a fixed-point iteration scheme for maximising the cost function given by (3.6). However, *FastICA* also has some disadvantages. Firstly, it cannot guarantee global optimisation as it is sensitive to the initialising settings, which limits its usability in dynamic systems. Secondly, as it is not based on a probabilistic framework, therefore we cannot get a measure of confidence in

the estimation of the unmixing matrix (which is required for our information coupling model).

3.2.1 Inference

Having discussed the limitations of some commonly used approaches to ICA, we now present the approach which is best suited for developing our interaction models. For our analysis, we make use of the *icadec* algorithm to dynamically infer the unmixing matrix [122, 304]. Our choice of this approach is based on three primary reasons; it gives accurate results compared to other related ICA approaches [122], it offers rapid computation and guarantees the unmixing matrix to be linearly decorrelating by constraining it to the manifold of decorrelating matrices, and it provides us with a framework to obtain a confidence measure for the unmixing matrix. The benefits of these three points will become clear as we go through this chapter. We now present a brief overview of this algorithm; an in-depth description is presented in [122].

The independent source signals obtained using ICA, $\mathbf{a}(t)$, as given by (3.3), must be at least linearly decorrelated for them to be classed as independent [304]. This property of the independent components can be used to develop efficient ICA algorithms, which operate on, or close to, the manifold of decorrelating separation matrices [122]. Two signals (each with a mean of zero) are said to be linearly decorrelated if the expectation of their product is zero, i.e.:

$$\mathrm{E}[a_j a_k] = \delta_{jk} d_j^2 \tag{3.7}$$

where a_j is the j-th source, d_j is a scale factor corresponding to the j-th source, and δ_{jk} is the Kronecker's delta function, given by $\delta_{jk} = 1$ for $j = k$ and $\delta_{jk} = 0$ otherwise. For a set of observed signals, $\mathbf{X}$, where $\mathbf{X} = [\mathbf{x}(t)]_{t=1}^{t=T}$; the set of recovered independent components, $\mathbf{B} = [\mathbf{a}(t)]_{t=1}^{t=T}$, is given by $\mathbf{B} = \mathbf{W}\mathbf{X}$. Using (3.7), the independent components are linearly decorrelated if:

$$\mathbf{B}\mathbf{B}^{\mathsf{T}} = \mathbf{W}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}} = \mathbf{D}^2 \tag{3.8}$$

where $\mathbf{D}$ is a diagonal matrix of scaling factors. If none of the rows of $\mathbf{B}$ are identically zero, (3.8) can be written as:

$$\mathbf{D}^{-1}\mathbf{W}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{D}^{-1} = \mathbf{I} \tag{3.9}$$

where $\mathbf{I}$ is the identity matrix. If $\mathbf{Q}$ is a real orthogonal matrix, i.e. $\mathbf{Q}\mathbf{Q}^{\mathsf{T}} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}$; and $\hat{\mathbf{D}}$ is a second (arbitrary) diagonal matrix, then (3.9) can be written as:

$$\hat{\mathbf{D}}\mathbf{Q}\mathbf{D}^{-1}\mathbf{W}\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{Q}^{\mathsf{T}}\hat{\mathbf{D}} = \hat{\mathbf{D}}^2 \tag{3.10}$$

Now $\hat{\mathbf{D}}\mathbf{Q}\mathbf{D}^{-1}\mathbf{W}$ is a decorrelating matrix and $\mathbf{D}^{-1}\mathbf{W}$ makes the rows of $\mathbf{B}$ orthonormal. The singular value decomposition of the set of observed signals, $\mathbf{X}$, is given by:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}} \tag{3.11}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices (the columns of $\mathbf{U}$ are the principal components of $\mathbf{X}$), and $\mathbf{\Sigma}$ is the diagonal matrix of the singular values of $\mathbf{X}$. Let $\mathbf{W}_0 = \mathbf{\Sigma}^{-1}\mathbf{U}^{\mathsf{T}}$, then the rows of $\mathbf{W}_0\mathbf{X} = \mathbf{V}^{\mathsf{T}}$ are orthonormal, so the decorrelating matrix, $\mathbf{W}$, can be written as [122]:

$$\mathbf{W} = \mathbf{D}\mathbf{Q}\mathbf{W}_0 = \mathbf{D}\mathbf{Q}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathsf{T}} \tag{3.12}$$

It is interesting to note that if $\mathbf{Q} = \mathbf{I}$ and $\mathbf{D} = \mathbf{\Sigma}$, the decorrelating matrix $\mathbf{W} = \mathbf{U}^{\mathsf{T}}$, i.e. a representation of PCA. To obtain an estimate for the ICA unmixing matrix, we need to optimise a given contrast function (we use log-likelihood of the data, as described later). There are a variety of optimisation approaches which can be used; our approach of choice is the Broyden-Fletcher-Golfarb-Shanno (BFGS) quasi-Newton method, which gives the best estimate of the minimum of the negative log-likelihood in a computationally efficient manner and also provides us with an estimate for the Hessian matrix. However, parameterising the optimisation problem directly by the elements of $\mathbf{Q}$ makes it a constrained minimisation problem for which BFGS is not applicable. Therefore, to convert it into an unconstrained minimisation problem, we constrain $\mathbf{Q}$ to be orthonormal by parameterising its elements as the matrix exponential of a skew-symmetric matrix $\mathbf{J}$, i.e. $\mathbf{J}^{\mathsf{T}} = -\mathbf{J}$ (non-zero elements of this matrix are known as the

Cayley coordinates), whose above diagonal elements parameterise $\mathbf{Q}$ [304][3]:

$$\mathbf{Q} = \exp(\mathbf{J}) \tag{3.13}$$

Using the parameterisation given by (3.13) makes it possible to apply BFGS to any contrast function; the contrast function used as part of the *icadec* algorithm is an expression for the log-likelihood of the data (as described later). To efficiently optimise this contrast function, we need to select a suitable source model. Most ICA approaches make implicit or explicit assumptions regarding the parametric model of the pdfs of the independent sources [304], e.g. Gaussian mixture distributions are used as source models in [82, 276], while [122] makes use of a flexible source density model given by the generalised exponential distribution. In our analysis, we use a reciprocal cosh source model as a canonical heavy-tailed distribution, namely [304]:

$$p(s_i) = \frac{1}{\pi \cosh(s_i)} \tag{3.14}$$

where s_i is the i-th source. Our choice of this source model is based on two major factors. Firstly, as this analytical fixed source model has no adjustable parameters, therefore it has considerable computational advantages over alternative source models; and secondly, as this source model is heavier in the tails, it is able to accurately model heavy-tailed unimodal non-Gaussian distributions, such as financial returns.

Let us now describe the contrast function which is optimised to compute the ICA unmixing matrix. Using (3.1) and assuming that the observation noise ($\mathbf{n}$) is normally distributed with a mean of zero and having an isotropic covariance matrix with precision β, the distribution of the observations (as a preprocessing step, we normalise each observed signal to have a mean of zero and unit variance) conditioned on $\mathbf{A}$ and $\mathbf{s}$ (where we drop the time index t for ease of presentation) is given by:

$$p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}) = \mathcal{N}\left(\mathbf{x}; \mathbf{A}\mathbf{s}, \beta^{-1}\mathbf{I}\right) \tag{3.15}$$

[3]For M sources and N observed signals, the ICA unmixing matrix may be optimised in the $\frac{1}{2}M(M+1)$ dimensional space of decorrelating matrices rather than in the full MN dimensional space, as $\frac{1}{2}M(M-1)$ and M parameters are required to specify $\mathbf{Q}$ and $\mathbf{D}$ respectively. This feature offers considerable computational benefits (especially in high-dimensional spaces) and the resulting matrix hence obtained is guaranteed to be decorrelating [122].

where $\mathbf{As}$ is the mean of the normal distribution and $\beta^{-1}\mathbf{I}$ is its covariance. The likelihood of an observation occurring is given by:

$$p(\mathbf{x} \mid \mathbf{A}) = \int p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}) p(\mathbf{s}) d\mathbf{s} \tag{3.16}$$

Assuming that the source distribution has a single dominant peak, $\hat{\mathbf{s}}$, the integral in (3.16) can be analysed by using Laplace's method, as shown in [342]:

$$\int p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}) p(\mathbf{s}) d\mathbf{s} \approx p(\mathbf{x} \mid \mathbf{A}, \hat{\mathbf{s}}) p(\hat{\mathbf{s}}) (2\pi)^{\frac{M}{2}} \det(\mathbf{F})^{-\frac{1}{2}} \tag{3.17}$$

where:

$$\mathbf{F} = -\left[\frac{\partial^2 \log p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}) p(\mathbf{s})}{\partial \mathbf{s}_i \partial \mathbf{s}_j} \right]_{\mathbf{s}=\hat{\mathbf{s}}} \tag{3.18}$$

For ease of computation, it is common practise to make use of alternate forms of Laplace's method; most of these methods are based around the process of omitting part (often the comparatively less informative) of the integrand from the exponent when performing the Taylor expansion, as described in detail in [204]. For our analysis, we use a simplified (computationally efficient) variant of the Laplace's method which enables us to replace the matrix $\mathbf{F}$ in (3.17) by the Hessian ($\mathbf{G}$) of the log-likelihood (which is evaluated at the MLE of the source distributions, i.e. $\hat{\mathbf{s}} = \left(\mathbf{A}^\mathsf{T}\mathbf{A}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{x}$) [204, 304]:

$$\mathbf{G} = -\left[\frac{\partial^2 \log p(\mathbf{x} \mid \mathbf{A}, \mathbf{s})}{\partial \mathbf{s}_i \partial \mathbf{s}_j} \right]_{\mathbf{s}=\hat{\mathbf{s}}} \tag{3.19}$$

Taking log of the expanded form of (3.15) gives:

$$\log p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}) = \frac{N}{2} \log\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}(\mathbf{x} - \mathbf{As})^\mathsf{T}(\mathbf{x} - \mathbf{As}) \tag{3.20}$$

which, via (3.19), results in $\mathbf{G} = \beta \mathbf{A}^\mathsf{T}\mathbf{A}$. The log-likelihood, $\ell \equiv \log p(\mathbf{x} \mid \mathbf{A})$, is therefore:

$$\ell = \frac{N-M}{2} \log\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}(\mathbf{x} - \mathbf{A}\hat{\mathbf{s}})^\mathsf{T}(\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}) + \log p(\hat{\mathbf{s}}) - \frac{1}{2}\log\det(\mathbf{A}^\mathsf{T}\mathbf{A}) \tag{3.21}$$

By using (3.12), we obtain $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{Q}^\mathsf{T}\mathbf{D}^{-1}$. Hence, the log-likelihood becomes [304]:

$$\ell = \frac{N-M}{2} \log\left(\frac{\beta}{2\pi e}\right) + \log p(\hat{\mathbf{s}}) + \log\det(\mathbf{\Sigma}^{-1}\mathbf{D}) \tag{3.22}$$

Noting that we use a reciprocal cosh source model (as given by (3.14)), it can be shown that taking the derivative of this log-likelihood expression with respect to **D** and **J** (which parameterises **Q**), and by following the resulting likelihood gradient using a BFGS optimiser, makes it possible to efficiently compute an optimum value for the ICA unmixing matrix; details of this procedure are presented in [122]. The same approach can be used to estimate the optimum number of ICA source signals if the mixing is non-square; with the optimum model order (M) being one which maximises this log-likelihood term.

Dynamic mixing

The standard (offline) ICA model uses all available data samples at times $t = 1, 2, ..., T$ of the observed signals, $\mathbf{x}(t)$, to estimate a single static unmixing matrix, $\mathbf{W}$. The unmixing matrix obtained provides a good estimate of the mixing process for the complete time series and is well suited for offline data analysis. However, many time series, such as financial data streams, are highly dynamic in nature with rapidly changing properties and therefore require a source separation method that can be used in a sequential manner. This issue is addressed here by using a sliding-window ICA model [7]. This model makes use of a sliding-window approach to sequentially update the current unmixing matrix using information contained in the previous window and can easily handle non-stationary data. The unmixing matrix for the current window, $\mathbf{W}(t)$, is used as a prior for computing the unmixing matrix for the next window, $\mathbf{W}(t+1)$. This results in significant computational efficiency as fewer iterations are required to obtain an optimum value for $\mathbf{W}(t+1)$. The algorithm also results in an improvement in the source separation results obtained when the mixing process is drifting and addresses the ICA permutation and sign ambiguity issues, by maintaining a fixed (but of course arbitrary) ordering of recovered sources through time.

3.3 Information coupling

We now proceed to derive the ICA-based information coupling metric. Later in this section we discuss the practical advantages this metric offers when used to analyse multivariate financial time series.

3.3.1 Coupling metric

Let $\mathbf{W}$ be any arbitrary square ICA unmixing matrix[4]:

$$\mathbf{W} \in \mathbb{R}^{N \times N} \tag{3.23}$$

Since multiplication of $\mathbf{W}$ by a diagonal matrix does not affect the mutual information of the recovered sources, therefore, we row-normalise the unmixing matrix in order to address the ICA scale indeterminacy problem[5]. Row-normalisation implies that the elements w_{ij} of the unmixing matrix $\mathbf{W}$ are constrained, such that each row of the matrix is of unit length, i.e.:

$$\sum_{j=1}^{N} w_{ij}^2 = 1 \tag{3.24}$$

for all rows i. For a set of observed signals to be completely decoupled, their latent independent components must be the same as the observed signals, therefore, the row-normalised unmixing matrix for decoupled signals ($\mathbf{W}_0$) must be a permutation of the identity matrix ($\mathbf{I}$):

$$\mathbf{W}_0 = \mathbf{PI} \in \mathbb{R}^{N \times N} \tag{3.25}$$

where $\mathbf{P}$ is a permutation matrix. For the case where the observed signals are completely coupled, all the latent independent components must be the same, therefore, the row-normalised unmixing matrix for completely coupled signals ($\mathbf{W}_1$) is given by:

$$\mathbf{W}_1 = \frac{1}{\sqrt{N}}\mathbf{K} \in \mathbb{R}^{N \times N} \tag{3.26}$$

where $\mathbf{K}$ is the unit matrix (a matrix of ones).

To calculate coupling, we need to consider the *distance* between any arbitrary unmixing matrix ($\mathbf{W}$) and the zero coupling matrix ($\mathbf{W}_0$). The distance measure we use is the generalised 2-norm, also called the spectral norm, of the difference between the two matrices [167],

[4]For the purpose of brevity and clarity, we only consider the case of square mixing while deriving the metric here. However, the metric derived in this section is valid for non-square mixing as well, and the corresponding derivation can be undertaken using a similar approach as presented here but converting the non-square ICA unmixing matrices in each instance into square matrices by padding them with zeros.

[5]ICA algorithms suffer from the scale indeterminacy problem, i.e. the variances of the independent components cannot be determined; this is because both the unmixing matrix and the source signals are unknown and any scalar multiplication on either will be lost in the mixing process.

although we can use some other norms as well to get similar results. The spectral norm of a matrix corresponds to its largest singular value and is the matrix equivalent of the vector Euclidean norm. Hence, the distance, $d(\mathbf{W},\mathbf{W}_0)$, between the two matrices can be written as:

$$d(\mathbf{W},\mathbf{W}_0) = ||\mathbf{W}-\mathbf{W}_0||_2 \tag{3.27}$$

where $||\cdot||_2$ is the spectral norm of the matrix. As $\mathbf{W}_0$ is a permutation of the identity matrix, therefore:

$$d(\mathbf{W},\mathbf{W}_0) = ||\mathbf{W}-\mathbf{PI}||_2 \tag{3.28}$$

As the spectral norm of a matrix is independent of its permutations, therefore, we may define another permutation matrix ($\hat{\mathbf{P}}$) such that:

$$d(\mathbf{W},\mathbf{W}_0) = \left|\left|\hat{\mathbf{P}}\mathbf{W}-\mathbf{I}\right|\right|_2 \tag{3.29}$$

For this equation, the following equality holds:

$$d(\mathbf{W},\mathbf{W}_0) = \left|\left|\hat{\mathbf{P}}\mathbf{W}\right|\right|_2 - 1 \tag{3.30}$$

Again, noting that $\left|\left|\hat{\mathbf{P}}\mathbf{W}\right|\right|_2 = ||\mathbf{W}||_2$, we have:

$$d(\mathbf{W},\mathbf{W}_0) = ||\mathbf{W}||_2 - 1 \tag{3.31}$$

We normalise this measure with respect to the range over which the distance measure can vary, i.e. the distance between matrices representing completely coupled ($\mathbf{W}_1$) and decoupled ($\mathbf{W}_0$) signals. From (3.26) we have that $\mathbf{W}_1 = \frac{1}{\sqrt{N}}\mathbf{K}$, therefore:

$$d(\mathbf{W}_1,\mathbf{W}_0) = ||\mathbf{W}_1-\mathbf{W}_0||_2 = \left|\left|\frac{1}{\sqrt{N}}\mathbf{K}-\mathbf{PI}\right|\right|_2 \tag{3.32}$$

Using the same analysis as presented previously, this equation can be simplified to:

$$d(\mathbf{W}_1,\mathbf{W}_0) = \frac{1}{\sqrt{N}}||\mathbf{K}||_2 - 1 \tag{3.33}$$

For a N-dimensional square unit matrix, the spectral norm is given by $||\mathbf{K}||_2 = N$. Therefore,

for a row-normalised unit matrix, the spectral norm is $\frac{1}{\sqrt{N}}||\mathbf{K}||_2 = \sqrt{N}$. Hence, if $\mathbf{W}$ is row-normalised, (3.33) can be written as:

$$d(\mathbf{W}_1, \mathbf{W}_0) = \sqrt{N} - 1 \tag{3.34}$$

The normalised information coupling metric (η) is then defined as:

$$\eta = \frac{d(\mathbf{W}, \mathbf{W}_0)}{d(\mathbf{W}_1, \mathbf{W}_0)} \tag{3.35}$$

Substituting (3.31) and (3.34) into (3.35), the normalised information coupling between N observed signals is given by:

$$\eta = \frac{||\mathbf{W}||_2 - 1}{\sqrt{N} - 1} \tag{3.36}$$

We can consider the bounds of η as described below. Suppose $\mathbf{M}$ is an arbitrary real matrix. We can look upon the spectral norm of this matrix ($||\mathbf{M}||_2 = ||\mathbf{M} - \mathbf{0}||_2$) as a measure of departure (distance) of $\mathbf{M}$ from a null matrix ($\mathbf{0}$) [99]. The bounds on this norm are given by:

$$0 \leq ||\mathbf{M}||_2 \leq \infty \tag{3.37}$$

If $\mathbf{M}$ is a row-normalised ICA unmixing matrix ($\mathbf{W}$), then (as discussed earlier) it lies between $\mathbf{W}_0$ and $\mathbf{W}_1$. Hence, the bounds on $\mathbf{W}$ are:

$$||\mathbf{W}_0||_2 \leq ||\mathbf{W}||_2 \leq ||\mathbf{W}_1||_2 \tag{3.38}$$

Using (3.25) and (3.26), we can write:

$$||\mathbf{PI}||_2 \leq ||\mathbf{W}||_2 \leq \left|\left|\frac{1}{\sqrt{N}}\mathbf{K}\right|\right|_2 \tag{3.39}$$

which can be simplified to:

$$1 \leq ||\mathbf{W}||_2 \leq \sqrt{N} \tag{3.40}$$

Rearranging terms in this inequality gives:

$$0 \leq \frac{||\mathbf{W}||_2 - 1}{\sqrt{N} - 1} \leq 1 \tag{3.41}$$

which gives us the same coupling metric as in (3.36) and shows that the metric is normalised, i.e. $0 \leq \eta \leq 1$. For real-valued $\mathbf{W}$, $||\mathbf{W}||_2$ can be written as:

$$||\mathbf{W}||_2 = \sqrt{\lambda_{max}(\mathbf{W}^\mathsf{T}\mathbf{W})} \tag{3.42}$$

where $\lambda_{max}(\mathbf{W}^\mathsf{T}\mathbf{W})$ is the maximum eigenvalue of $\mathbf{W}^\mathsf{T}\mathbf{W}$. The unmixing matrix obtained using ICA suffers from row permutation and sign ambiguity problems, i.e. the rows are arranged in a random order and the sign of elements in each row is unknown [122, 304]. We note that as $||\mathbf{W}||_2$ is independent of the sign and permutations of the rows of $\mathbf{W}$, therefore our measure of information coupling straightaway addresses the problems of ICA sign and permutation ambiguities. Also, as the metric's value is independent of the row permutations of $\mathbf{W}$, it provides symmetric results. The information coupling metric is valid for all dimensions of the unmixing matrix, $\mathbf{W}$. This implies that information coupling can be easily measured in high-dimensional spaces.

It is possible to obtain a measure of uncertainty in our estimation of the information coupling measure. We make use of the BFGS quasi-Newton optimisation approach over the most probable skew-symmetric matrix, $\mathbf{J}$, of (3.13), from which estimates for the unmixing matrix, $\mathbf{W}$, can be obtained, and thence the coupling measure η calculated. We also estimate the Hessian (inverse covariance) matrix, $\mathbf{H}$, for $\mathbf{J}$, as part of this process. Hence, it is possible to draw samples, $\mathbf{J}'$, say from the distribution over $\mathbf{J}$ as a multivariate normal:

$$\mathbf{J}' \sim \mathcal{MN}\left(\mathbf{J}, \mathbf{H}^{-1}\right) \tag{3.43}$$

These samples can be readily transformed to samples in η using (3.13), (3.12) and (3.36) respectively. Confidence bounds (and here we use the 95% bounds) may then be easily obtained from the set of samples for η (in our analysis we use 100 samples).

In multivariate systems, it is quite likely that a given set of N observed signals may have a different number (M) of underlying source signals, i.e. the mixing process is non-square. The ICA-based information coupling model can be used even if the source mixing process

is non-square (as we describe below). There are three possible mixing cases, i.e. square, undercomplete and overcomplete. Undercomplete ICA mixing refers to the case when the number of latent source signals is less than the number of observed signals, i.e. $M < N$. For the undercomplete case, information coupling can dynamically be computed in three steps. In the first step, the ICA log-likelihood based model order estimation algorithm is used to estimate the optimum number of latent source signals within each sliding-window at each time step; this is achieved by calculated the log-likelihood for different number of source signals, ranging from $M = 2$ to $M = N$, and the value of M associated with the maximum log-likelihood is selected as the optimum model order. In the second step, the non-square ICA unmixing matrix at time t, $\mathbf{W}(t)$, is estimated which has dimensions $M \times N$. Finally, in the third step, the metric presented by (3.36) can be used to compute information coupling, $\eta(t)$. In some instances, the mixing process may be overcomplete, i.e. the optimum number of source signals may be greater than the number of observed signals ($M > N$). Although there is no simple solution for the overcomplete ICA problem (as no unique solutions exist), many studies have focused on finding an optimum estimate for the unmixing matrix and thus the source signals [221, 340]. For the purpose of results presented in this thesis, the overcomplete mixing case is not considered. This is because (as we show later in the thesis) the estimated number of optimum ICA source signals for multivariate financial data has a clear peak at a value much lower than the number of observed signals, i.e. the mixing process is often undercomplete; also, the lack of any robust (and computationally efficient) algorithms for overcomplete ICA means that results obtained may be misleading and computationally complex to obtain (especially in high-dimensional spaces).

Computational complexity

The information coupling algorithm achieves computational efficiency by making use of the sliding-window based decorrelating manifold approach to ICA. Making use of the reciprocal cosh based ICA source model also results in significant computational advantages. We now take a look at the comparative computational complexity of the information coupling measure and three frequently used measures of statistical dependence, i.e. linear correlation, rank correlation and mutual information. For bivariate data (n_s data points long), for which these four measures are directly comparable, linear correlation and rank correlation have time

complexities of order $\mathcal{O}(n_s)$ and $\mathcal{O}(n_s \log n_s)$ respectively [50], while mutual information and information coupling scale as $\mathcal{O}(n_s^2)$ and $\mathcal{O}(n_s)$ respectively[6] [345]. Hence, even though the time complexity of the information coupling measure is of the same order as linear correlation, it can still accurately capture statistical dependencies in non-Gaussian data streams and is a computationally efficient proxy for mutual information.

For N-dimensional multivariate data, direct computation of mutual information has time complexity of order $\mathcal{O}(n_s^N)$ compared to $\mathcal{O}(n_s N^3)$ for the information coupling measure. In high-dimensions, even an approximation for mutual information can be computationally very costly. For example, using a Parzen-window density estimator, the mutual information computational complexity can be reduced to $\mathcal{O}(n_s n_b^N)$, where n_b is the number of bins used for estimation [264], which will incur a very high computational cost even for relatively small values of N, n_b and n_s. As a simple example, Table 3.1 shows a comparison of computation time (in seconds) taken by mutual information and information coupling measures for analysing bivariate data sets of varying lengths. As expected, mutual information estimation using the Parzen window based approach (which is considered to be a relatively efficient approach to compute mutual information) becomes computationally very demanding with an increase in the number of samples of the bivariate data set. In contrast, the information coupling measure is computationally efficient, even when used to analyse very large high-dimensional multivariate data sets.

Computation time (sec)	$n_s = 10^2$	$n_s = 10^3$	$n_s = 10^4$	$n_s = 10^5$
Mutual information	0.0214	2.8289	17.0101	119.5088
Information coupling	0.0073	0.0213	0.0561	0.5543

Table 3.1: Example showing comparison of average computation time (in seconds) of mutual information and information coupling, when these measures are used to analyse bivariate data sets containing different number of samples (n_s). The approach used to estimate mutual information is based on a Parzen window based algorithm, as described in [346]. The computational cost of this algorithm is dependent on the window-size (h). The values of h used for the simulations are: $h = 20$ for $n_s = 10^2$, and $h = 100$ for all other simulations. Results are obtained using a 2.66 GHz processor as an average of 100 simulations.

[6]There have been various estimation algorithms proposed for efficient computation of mutual information, however, they all result in increased estimation errors and require careful selection of various user-defined parameters [120].

Discussion

The information coupling model offers us with multiple advantages when used to analyse multivariate financial data. Here we summarise some of the main properties the model, while the empirical results presented in the next chapter demonstrate its accuracy and practical benefits.

- The information coupling measure, a proxy for mutual information, is able to accurately pick up statistical dependencies in data sets with non-Gaussian distributions (such as financial returns).

- The information coupling algorithm is computationally efficient, which makes it particularly suitable for use in an online dynamic environment. This makes the algorithm especially attractive when dealing with data sampled at high frequencies. This is because with the ever-increasing use of high-frequency data, overcoming sources of latency is of utmost importance in a variety of applications in modern financial markets.

- It gives confidence levels on the information coupling measure. This allows us to estimate the uncertainty associated with the measurements.

- The metric provides normalised results, i.e. information coupling ranges from 0 for decoupled systems to 1 for completely coupled systems. This makes it easier to analyse results obtained using the metric and to compare its performance with other similar measures of association. The metric also gives symmetric results[7].

- The metric is valid for any number of signals in high-dimensional spaces, i.e. it consistently gives accurate results irrespective of the number of time series between which information coupling is being computed. This makes it suitable for a range of financial applications.

- It is not data intensive, i.e. it gives relatively accurate results even when a small sample size is used. This allows the metric to model the complex and rapidly changing dynamics of financial markets.

[7] Although a symmetric measure, the information coupling metric can give us an indication of the presence of asymmetric interactions in a set of signals by making use of lead-lag relationships between them; we refer to this approach as *predictive information coupling*.

- It does not depend on user-defined parameters which can restrict its practical utility, as the evolving market conditions may require the parameters to be constantly updated, which may not be practical.

3.3.2 Complex coupling networks

The ICA-based information coupling model can be used to measure the overall mutual symmetric interactions in multivariate systems. However, to extract the hierarchical interaction structure in multi-dimensional systems, we need to make use of complex coupling networks. This is because, unlike in low-dimensional systems, in higher dimensions analysing and understanding the nature of dependencies between variables can be a complex undertaking, as the number of pair-wise relationships varies as $\frac{N}{2}(N-1)$ for N variables. For example, for a 100 dimensional system, the number of pair-wise coupling terms will be $4,950$. Clearly analysing the overall structure of interdependencies within such a system using only the numerical coupling values is not practical. Therefore, we make use of complex coupling networks to extract the hierarchical interaction structure of a system in order to study the characteristics of relationships in high-dimensional multivariate financial systems. The primary method used in this thesis for analysing financial networks is the minimum spanning tree (MST) approach. The MST approach to complex network analysis has numerous advantages. MSTs are attractive because they exhibit only the most relevant connections for each node, thus simplifying the presentation and comprehension of results. Also, the MST approach is deterministic and straightforward to implement while most other network analysis methods are not [74]. Moreover, MST algorithms are computationally efficient [133], and hence allow us to develop efficient dynamical coupling networks (which we discuss in detail later). MSTs have diverse applications and proven applicability in the financial sector. Analysis of multivariate financial returns using a MST can be useful for constructing a portfolio of financial instruments, such as currencies [251] or equities [48]. MSTs can also reveal information about the equities or currencies which are driving the market, i.e. are “in play”, at any given time [199, 243]. Examining the properties of a dynamical MST can inform us about the degree of stability of a financial market [199, 248]. Using a MST, we can also extract a topological influence map of different currencies [268].

Let us now describe our approach for constructing a MST based on the information cou-

pling measure. A spanning tree of a graph with N vertices (nodes) is a subset of $N-1$ edges (links) that form a tree. The MST of a weighted graph with N vertices is a set of $N-1$ edges of minimum total weight which form a spanning tree of the graph [352]. There are two standard methods for constructing a MST, namely Kruskal's algorithm and Prim's algorithm. The method we use in this thesis is based on Prim's algorithm due to its relatively low computational cost [195]. Prim's algorithm grows the MST one edge at a time by adding a minimal edge that connects a node in the growing MST with any other node. Let us denote the set of vertices of a graph by V and the set of edges by E. Then, for a given fully-connected, undirected graph $G=(V,E)$, with each edge (u,v) having positive weights $w(u,v)$, a spanning tree is a tree composed of the edges of G that touches every vertex in G [330]; a MST is a spanning tree which minimises the sum of its edge weights. Some recent studies have analysed FX currency interactions using the MST approach, with the edges representing the linear correlation between the currency pairs [243, 251]. However, (as previously discussed) linear correlation is not a suitable approach for measuring interactions between signals with non-Gaussian distributions, such as financial returns. Therefore, we combine the ICA-based information coupling model with techniques used for building a MST to produce coupling networks which are better suited to analyse the dependency structure in multivariate financial systems. A MST provides information about the subdominant ultrametric hierarchical organisation of each node in a network; the subdominant ultrametric distance between nodes i and j is the maximum value of a distance metric (d_{ij}) detected when moving in single steps from i to j through the shortest path connecting the two nodes in the network [139]. We calculate this distance (representing weights of the edges in a MST) between nodes i and j using the following non-linear distance metric [237]:

$$d_{ij} = 2\sqrt{1-\eta_{ij}} \tag{3.44}$$

where η_{ij} is the information coupling obtained using the ICA-based coupling model. We use this non-linear distance metric in order to remain consistent with some other studies focusing on building correlation networks using MSTs [237, 243, 251]. As $0 \leq \eta_{ij} \leq 1$, therefore the pseudo-distance is bound in the range $0 \leq d_{ij} \leq 2$. A higher value of information coupling, η_{ij}, between any two nodes in a network translates into a smaller pseudo-distance, d_{ij}, between

them; hence, the metric d_{ij} can be viewed as a measure of the "statistical similarity" of any two nodes. Also, as $\eta_{ij} = \eta_{ji}$, therefore, $d_{ij} = d_{ji}$, and $d_{ij} = 0$ iff $i = j$. Moreover, this distance metric meets the triangular inequality, i.e. $d_{ik} \leq d_{ij} + d_{jk}$.

Dynamic coupling networks

We have so far described the development and use of static coupling networks. However, studying the properties of a MST with a dynamically evolving structure can be useful for finding temporal dependencies in multivariate financial time series in high-dimensional spaces. The dynamically changing structure of a MST can be used to monitor the characteristics of a financial network in an online environment and to learn about the dominant and dependent nodes at any given time. However, a high-dimensional network can contain more than 100 nodes. Observing and analysing the dynamically changing structure of such a network can prove to be difficult. Therefore, we make use of the concept of survival ratio of a MST, which gives us information about the sequentially evolving structure of a MST in a concise and clear way. The survival ratio of the edges of a MST can be used as a measure of the temporal stability of its structure [158]. A single-step survival ratio coefficient, $\sigma_{SR}(t)$, at time t, is defined as:

$$\sigma_{SR}(t) = \frac{1}{N-1} \mid \mathrm{E}(t) \cap \mathrm{E}(t-1) \mid \tag{3.45}$$

where N is the number of nodes in the MST, and $\mathrm{E}(t)$ represents the set of edges (or links) of the MST at time t. The survival ratio ranges between 0 and 1, giving $\sigma_{SR}(t) = 1$ if all links stay exactly the same between time steps $t-1$ and t. Similarly, $\sigma_{SR}(t) = 0$ occurs if all links change during this time interval. Likewise, the multi-step survival ratio coefficient is defined as:

$$\sigma_{SR}(t,k) = \frac{1}{N-1} \mid \mathrm{E}(t) \cap \mathrm{E}(t-1) \ldots \mathrm{E}(t-k+1) \cap \mathrm{E}(t-k) \mid \tag{3.46}$$

where k is the number of time-steps over which the ratio is calculated. Survival ratios can be helpful to predict the likelihood of a market's dependency structure remaining stable over a given time period. An interesting study demonstrating use of survival ratio of MSTs to analyse the interdependencies and stability in the global equity markets is presented in [86]. Later in

this thesis we make use of information obtained using static and dynamic coupling-MSTs in a variety of financial applications, which showcase some of their numerous practical uses.

3.4 Dynamic information coupling

Most practical financial applications require the use of interaction analysis models in an online dynamic environment, often across different time-scales. In Chapter 2 we had presented an overview of various approaches to time- and scale-based dynamic interaction measurement and discussed their respective merits and drawbacks. We now expand on that work to present approaches which are best suited for modelling dynamics of interactions in multivariate financial time series and which we use to obtain results presented later in the thesis.

3.4.1 Time-based dynamics

We start by discussing our preferred windowing technique for sequentially capturing the fast local variations in information coupling in multivariate financial data streams. Later in this section we present use of the hidden Markov ICA (HMICA) model as a suitable discrete state-based approach for modelling the temporal variations in information coupling.

Choice of sliding-window

Some inference problems involve analysing data which has already been collected. This means that data points from the future may be used to improve the performance of an algorithm, a process known as offline inference. In contrast, many practical algorithms make use of sequential data which is being obtained in real-time, a process known as online inference. In most practical applications, parameters of a statistical model need to be updated as soon as new data is available. This requirement means that only models that can carry out *rolling regression* for online inference are useful. For computational efficiency and higher precision, it is useful to use a sliding-window technique for this purpose. In such a model, only data within the window is used to calculate the model parameters (e.g. to measure interactions) at each time step, hence dynamically capturing evolving structure of the signals and doing away with the need to use large amounts of data. Our choice of the type of sliding-window is based on some properties of financial markets, which we now discuss. Financial markets give rise to well defined events, such as orders, trades and quote revisions. These events

are irregularly spaced in *clock-time*, i.e. they are asynchronous. Statistical models in *clock-time* make use of data aggregated over fixed intervals of time [51]. The time at which these events are indexed is called the *event-time*. Hence, for dynamic modelling, in *event-time* the number of data points can be regarded as fixed while time varies, while in *clock-time* the time period is considered to be fixed with variable number of data points. Although we may need adaptive windows in *clock-time*, we can use sliding-windows of fixed-length in *event-time*. Using fixed-length sliding-windows in *event-time* can be useful for obtaining consistent (and unbiased) results when developing and testing different statistical models and algorithms. Also, statistical models deployed for online analysis of financial data operate best in *event-time* as they often need to make decisions as soon as some new market information (such as quote update etc.) becomes available. Consider an online trading model making use of an adaptive window in *event-time*. At specific times of the day, e.g. at times of major news announcements, trading volume can significantly increase. Hence, more data will be available to the algorithm and thus results obtained can be misleading [284]. Using a sliding-window of fixed-length in *event-time* can overcome this problem. Another possible approach we discussed earlier was use of the EWMA. However, using the EWMA approach can give us biased results, as a single major market shock can potentially affect results obtained well into the future; in comparison, a fixed-length sliding-window is well-suited for capturing the rapidly changing dynamics of financial markets. Also, unlike the EWMA, results obtained using a fixed-length sliding-window are not dependent on any user-defined parameters. The length of the sliding-window needs to be selected appropriately. The financial application for which the model is being used is one of the factors which drives the choice of window length. As a general rule, for trading models a window of approximately the same size as the average time period between placing trades (inverse of trading frequency) is often used. This makes it possible to accurately capture the rapidly evolving dynamics of the markets over the corresponding period, without being too long so as to only capture major trends or too short to capture noise in the data. The interaction approach being used to model dependencies also influences the choice of window length, as some approaches, such as mutual information and transfer entropy, require large data sets for accurate estimation.

Capturing discrete state-based coupling dynamics

By using sliding-windows we can accurately and precisely capture fast local variations in information coupling in an online dynamic environment. However, for certain applications (some of which we later discuss), it may be more useful to analyse the temporal variations in information coupling over a large section of data in order to observe different regimes of interaction, i.e. discrete state-based information about the level of coupling may be required. For this purpose, we need to make use of alternate approaches to dynamic interaction measurement. Here we discuss one such approach which is well-suited for analysing financial returns. Many real-world signals, including financial returns, exhibit rapidly changing dynamics, often characterised by regions of quasi-stability punctuated by abrupt changes. We hypothesise that regions of persistence in underlying information coupling may hence be captured using a Markov process model with switching states. Therefore, we make use of the hidden Markov ICA (HMICA) model, which is a hidden Markov model (HMM) with an ICA observation model, to capture changes in information coupling dynamics [317]. As we show later in this section, this is possible because a HMICA model extracts latent states based on the determinant of the ICA unmixing matrix, which in turn encodes dependencies.

Let us first take a brief look at the foundations of a HMICA model. A Markov model is a statistical process in which future probabilities are determined by only its most recent values [41]. A hidden Markov model (HMM) is a statistical model consisting of a set of observations which are produced by an unobservable set of latent states, the transitions between which are a Markov process. The typical goal of a HMM is to infer the hidden states from a set of observations. It is widely used within the speech recognition sector [202, 298], and is finding increasing use in a range of financial applications [38], such as finding regions of financial market volatility clustering and persistence [308]. The mathematical details of a HMM can be found in Appendix A and in [41], here we focus on the conceptual basis of the model. A HMM is represented in graphical form in Figure 3.3, showing the hidden (latent) layer of Markov states and the layer of observed data. As the figure shows, an observed variable, $x(t)$, in a HMM depends *only* on the current state, $z(t)$, and the current state depends *only* on the previous state, $z(t-1)$ [288]. Also shown in the figure is the state transition probability $p(z(t+1) \mid z(t))$ and the emission model probability $p(x(t) \mid z(t))$. The HMM state transition

probability matrix, $\mathbf{P}_{hmm}$, with elements p_{ij}, gives the probability of change of state from state i to state j, i.e.:

$$p_{ij} = p(z(t+1) = j \mid z(t) = i) \tag{3.47}$$

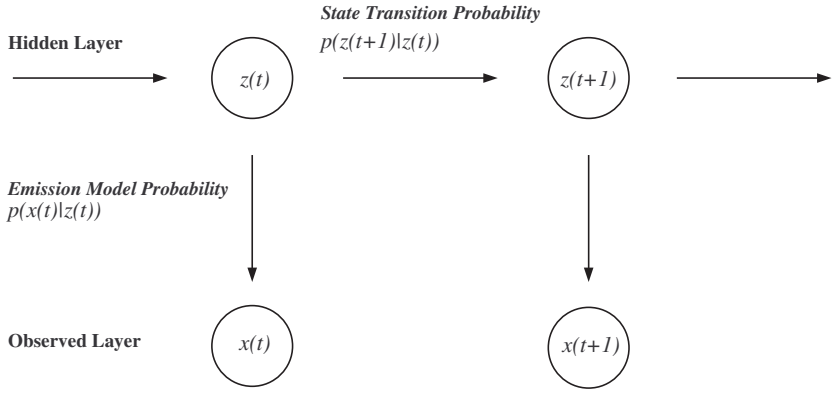

Fig. 3.3: Hidden Markov model (HMM) graphical representation.

We may combine ICA and HMM to form the hidden Markov ICA (HMICA) model [288], which can be seen as a HMM with an ICA observation model. The HMICA model is well-suited for analysis of non-stationary multivariate time series and can provide information about discrete state-based changes in either the ICA source dynamics or the mixing process or both [288]. Detailed mathematical framework of the HMICA model is presented in Appendix A, here we only outline the main steps which are relevant to our work. The HMICA auxiliary cost function for state k is given by [288]:

$$Q_k = \log|\det(\mathbf{W}_k)| + \frac{\sum_t \gamma_k[t] \sum_i \log p(a_i[t])}{\sum_t \gamma_k[t]} \tag{3.48}$$

where $\mathbf{W}_k$ is the ICA unmixing matrix for state k, $a_i[t]$ is the i-th ICA source and $\gamma_k[t]$ is the probability of being in state k at time t. The auxiliary function, summed over all states, is hence:

$$Q = \sum_k Q_k \tag{3.49}$$

The HMICA model finds the unmixing matrix $\mathbf{W}_k$, for state k, by minimising the cost function given by (3.49) over all underlying parameters using a set of iterated update equations, as

described in detail in [288]. The optimal number of hidden HMM states can be estimated using the approach described in [309]. As the cost function shows, the HMICA states are influenced by the ICA unmixing matrix ($\mathbf{W}_k$), which encodes information about the source mixing process, and hence can be used to capture changes in information coupling dynamics in multivariate systems (as previously discussed). The HMICA model can be useful for detecting regions of information coupling persistence and to identify sections of data which exhibit abrupt changes in mixing dynamics (the Viterbi algorithm can aid in this process, using which we can infer the most likely sequence of hidden states). It also provides us with an estimate for the state transition probability matrix which contains the state transition probabilities (p_{ij}) as its elements. It is possible to use this matrix to infer the time (T_{ij}) in any given state:

$$T_{ij} = \frac{1}{1 - p_{ij}} \tag{3.50}$$

Lower values of T_{ij} indicate a higher frequency of state transitions and (as we empirically demonstrate later in the thesis) are indicative of rapid dynamic variability of the ICA-based information coupling measure. To simultaneously capture changes in dynamics across both time and scale, we develop the wavelet-HMICA model, as discussed in the next section.

3.4.2 Scale-based dynamics

Earlier we discussed the suitability of using wavelets for time-scale decomposition of financial data. It is possible to combine wavelets with ICA to form the wavelet-ICA model, which can be used to infer the ICA unmixing matrices using the wavelet coefficients of a set of observed signals; hence, information about a frequency-dependent mixing process can be gained. It will therefore be possible to compute information coupling at different time-scales. Knowledge gained about the scale dependence of information coupling using the wavelet-ICA model can have numerous practical applications in the financial markets. Studies have shown the possibility of using wavelet-correlation models for optimising pairs trading strategies [70, 360][8]; the wavelet-ICA model can be used to estimate the optimum time-scale for executing trades

[8] Pairs trading algorithms continuously monitor the market prices of two or more closely coupled instruments, usually belonging to the same sector of the economy, and place trades as soon as these instruments become decoupled for short time periods. This is a well-known market-neutral trading strategy and is frequently used in the equities and to some extent the FX markets. Market-neutral strategies are typically not affected by the overall market direction as they employ some sort of hedging mechanism, in the case of pairs trading, simultaneously going long and short two coupled stocks.

using such strategies. Another possible area of application of approaches to scale-dependent analysis of interactions is financial risk management using portfolio optimisation techniques [88]. To minimise risk, it is often the case that assets are maintained as a portfolio. The return-to-risk characteristics of such a portfolio are dependent on the coupling of the instruments in the portfolio. As coupling varies with scale, the wavelet-ICA model can be useful for estimating the time-scale which is best suited for obtaining a portfolio with the required return-to-risk profile. It is also possible to make use of wavelets for the purpose of portfolio allocation, as discussed in detail in [246].

Choice of wavelet function

Let us now provide a brief description of some basic properties of wavelets. Wavelets are continuous functions, usually represented by $\psi(t)$, which meet the following two conditions [6]. They have a mean of zero and a finite energy, i.e.:

$$\int \psi(t)dt = 0, \quad \int \mid \psi(t) \mid^2 dt < \infty \tag{3.51}$$

The normalised wavelet function is given by:

$$\psi_{u,b}(t) = \frac{1}{\sqrt{u}}\psi\left(\frac{t-b}{u}\right) \tag{3.52}$$

where u is the scale, also known as the dilation parameter, and b is the localisation parameter. The function $\psi(t)$, from which different dilated and translated versions of the wavelets are derived, is called the mother wavelet. In this thesis we use the continuous wavelet transform (CWT) instead of the discrete wavelet transform (DWT) because the CWT can more efficiently handle high-frequency data with abrupt changes [265, 321]. The CWT is a powerful signal processing tool that can be used to analyse properties of a financial time series at different time-scales. Results obtained by using the CWT are often easier to interpret, since its redundancy tends to reinforce the traits and makes all information more visible [253]. Also, for analysis purposes, the main concern is not numerical or transmission efficiency or representation compactness, but rather the accuracy and adaptive properties of the analysing tool, leading to the CWT being mostly used for the purpose of analysis [331], whereas the DWT is commonly used for coding purposes or for data compression and transmission. Using CWT,

the wavelet coefficients of a function $x(t)$, at scale u and position b, are given by:

$$C_{u,b} = \frac{1}{\sqrt{u}} \int x(t) \psi \left(\frac{t-b}{u} \right) dt \tag{3.53}$$

We now proceed to discuss our choice of the wavelet function, $\psi(t)$. This choice depends on the application for which wavelets are meant to be used for. For high-frequency financial time series analysis for example, a wavelet which has good localisation properties in both time and frequency is required. There are many types of wavelet functions in common use, e.g. Haar, Mexican hat, Shannon, Morlet to name but a few. The Haar wavelet has poor decay in frequency, whereas the Shannon wavelet has compact support in frequency with poor decay in time [334]. Other wavelets typically fall in the middle of these two extremes. Having exponential decay in both time and frequency domains, the Morlet wavelet has optimal joint time-frequency properties [334]. It represents a modulated Gaussian function with exponential decay and has proven applicability for financial data analysis [71, 136, 334]. Therefore, we use the Morlet wavelet for our analysis. Morlet wavelet is a non-orthogonal wavelet which has both a real and a complex part, such wavelets are also referred to as analytical wavelets. Due to the complex component, Morlet wavelets can be used to separate both the phase and amplitude parts of a signal. The Morlet wavelet (with a centre frequency of f_o) is represented by (3.54); the first part of the equation represents a normalisation factor, the second part is a complex sinusoid, while the third part represents a Gaussian bell curve [6].

$$\psi(t) = \pi^{-\frac{1}{4}} \exp(i2\pi f_0 t) \exp\left(-\frac{t^2}{2} \right) \tag{3.54}$$

Using (3.52), we can convert the Morlet mother wavelet (given by (3.54)) to a normalised Morlet wavelet function:

$$\psi_{u,b}(t) = \frac{\pi^{-\frac{1}{4}}}{\sqrt{u}} \exp\left[i2\pi f_0 \frac{(t-b)}{u} \right] \exp\left[-\frac{(t-b)^2}{2u^2} \right] \tag{3.55}$$

A Morlet wavelet, with a scale of $u = 1$, is plotted in Figure 3.4. The exponentially decaying sinusoidal shape of the wavelet is evident from the figure. Each wavelet is characterised by a particular scale. For ease of data analysis, it is important to have an idea of what time length each scale represents. A wavelet scale u can be converted to a pseudo-frequency f_u in Hz as follows:

$$f_u = \frac{f_o}{u\Delta} \tag{3.56}$$

where Δ is the sampling period. The centre frequency (f_o) of a Morlet wavelet is equivalent to that of a periodic sine wave fitted to the wavelet, as shown in Figure 3.4. This method gives the period of the Morlet wavelet as 1.2308 seconds, which corresponds to a centre frequency of 0.8125 Hz. As an example, if using high-frequency financial data sampled at 2 samples per second, the sampling period is $\Delta = \frac{1}{2}$, and the pseudo-frequency (in Hz) at scale u is given by $f_u = \frac{1.6250}{u}$ (calculated using (3.56)). The time period corresponding to scale u is simply the reciprocal of this term. Thus, for a scale of $u = 1$, the pseudo-frequency is 1.6250 Hz, while the associated time period is 0.6154 seconds.

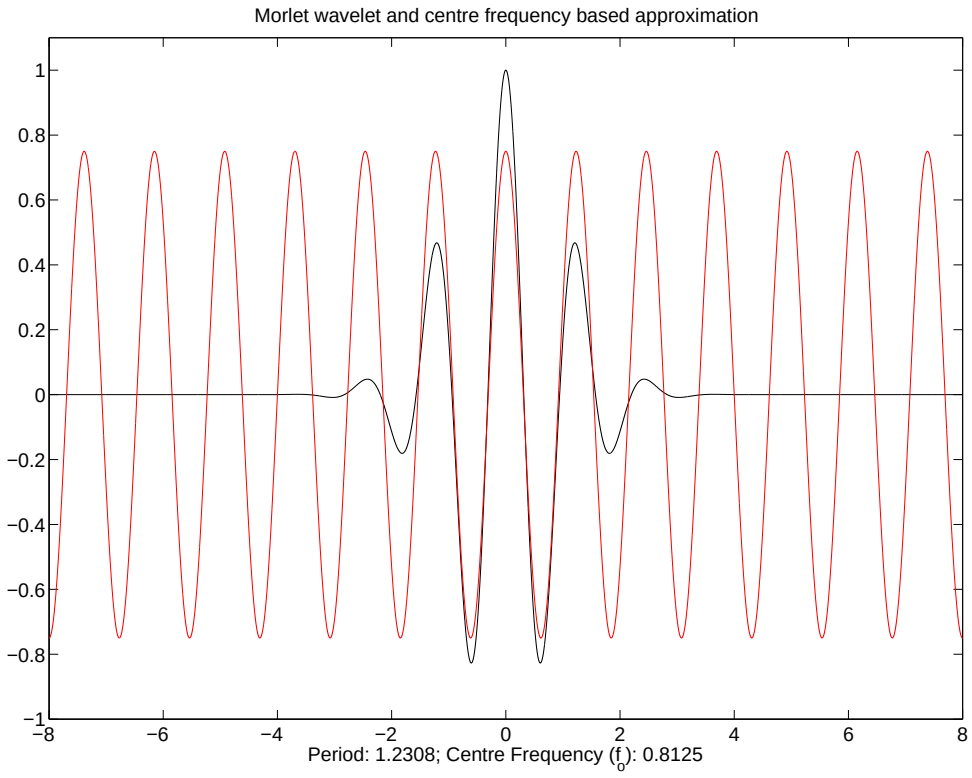

Fig. 3.4: Scale to frequency conversion of a Morlet wavelet with a scale of $u = 1$, and a localisation parameter of $b = 0$. The Morlet wavelet (real part) is represented by the black line, and the red line represents a periodic sine wave.

Wavelet-ICA model

We are now in a position to develop the wavelet-ICA model, which can be used to analyse coupling at different time-scales. For N observed signals, $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_N(t)]^\mathsf{T}$, analysed using the CWT at a scale of u, the wavelet coefficients can be combined into a single vector, $\mathbf{c}_{\mathbf{x},u}(t) = [c_{1,u}(t), c_{2,u}(t), ..., c_{N,u}(t)]^\mathsf{T}$, where $c_i(t)$ represents the wavelet coefficient for the i-th time series at time t. The location parameter b is dropped in this analysis for clarity. Using (3.53), we can represent the multivariate set of wavelet coefficients at a scale of u as:

$$\mathbf{C}_{\mathbf{x},u} = \int \mathbf{x}(t)\psi_u(t)dt \tag{3.57}$$

Substituting $\mathbf{x}(t) = \mathbf{A}_u\mathbf{s}(t)$ (where $\mathbf{A}_u$ denotes the mixing matrix at scale u, in this case corresponding to sampling frequency of the data):

$$\mathbf{C}_{\mathbf{x},u} = \int \mathbf{A}_u\mathbf{s}(t)\psi_u(t)dt \tag{3.58}$$

As the mixing matrix, $\mathbf{A}_u$, is time-independent and both ICA and CWT are based on linear transformations, therefore:

$$\mathbf{C}_{\mathbf{x},u} = \mathbf{A}_u \int \mathbf{s}(t)\psi_u(t)dt \tag{3.59}$$

The term within the integral represents the CWT of the set of source signals ($\mathbf{C}_{\mathbf{s},u}$), hence:

$$\mathbf{C}_{\mathbf{x},u} = \mathbf{A}_u\mathbf{C}_{\mathbf{s},u} \tag{3.60}$$

As CWT represents a linear transformation, therefore, the constituent signals of $\mathbf{C}_{\mathbf{s},u}$ are mutually independent. Thus, (3.60) represents an ICA model with $\mathbf{C}_{\mathbf{x},u}$ representing the set of observed signals, $\mathbf{C}_{\mathbf{s},u}$ representing the latent independent components, and $\mathbf{A}_u$ representing the mixing matrix which contains information about the frequency-dependent mixing process. The ICA unmixing matrix, $\mathbf{W}_u = \mathbf{A}_u^+$, can now be obtained using the *icadec* algorithm. This matrix contains information about the frequency-dependent mixing process, hence, it can be used to calculate information coupling between a set of signals at different time-scales (using the analysis presented earlier).

Wavelet-HMICA model

Combining the HMICA model with a wavelet basis allows us to model changes in the HMICA latent states at different frequencies, hence, simultaneously capturing both time- and scale-based dynamics of the system. This can be achieved by modelling the wavelet coefficients at different time-scales as being generated by a HMICA model with switching states. The resulting wavelet-HMICA model can best be described using the graphical model shown in Figure 3.5. The graphical model shows a series of observations, $\mathbf{x}(t)$, which are generated via a CWT with a set of wavelet coefficients $\mathbf{c}(t)$. The HMICA model infers the most probable set of latent states using the wavelet coefficients as the "observed" data. It achieves this by estimating an unmixing matrix $\mathbf{W}_k$ (for state k) by minimising the cost function given by (3.49). Hence, discrete state-based, frequency-dependent, dynamics of information coupling may be captured using the wavelet-HMICA model. We empirically demonstrate this process later in the thesis, where we show that the Viterbi state sequence obtained using the wavelet-HMICA model is indicative of temporal persistence in scale-based information coupling dynamics in multivariate financial data streams.

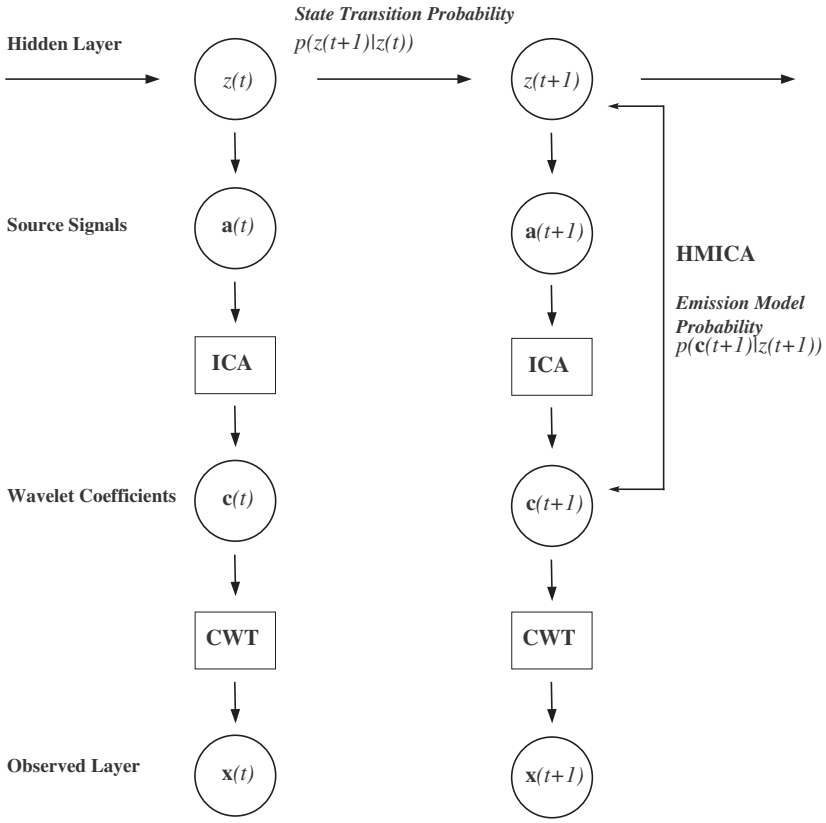

Fig. 3.5: Graphical representation of the wavelet-HMICA model. With each of the k latent states of the model (with state at time t denoted by $z(t)$), there is an associated ICA mixing matrix $\mathbf{A}_k = \mathbf{W}_k^+$, and a set of generalised AR (GAR) model coefficients $\boldsymbol{\alpha}_k^i$ (for each source i) which are used to generate the source estimates $\mathbf{a}(t)$ at time t using a GAR model with non-Gaussian noise. The wavelet coefficients (at any given scale) are then generated as $\mathbf{c}(t) = \mathbf{A}_k\mathbf{a}(t)$ and the observed data $\mathbf{x}(t)$ can be regarded as being generated as a result of application of the inverse CWT to $\mathbf{c}(t)$. Further details are presented in Appendix A and in [288].

Chapter 4
Analysis of information coupling

We start this chapter by providing an overview of properties of financial time series, focusing on the spot foreign exchange (FX) market. We then describe the synthetic and financial data used for analysis presented in this thesis. The rest of the chapter is focused on the analysis of symmetric interactions in multivariate synthetic and financial data sets, and includes a set of practical financial case studies using which we demonstrate utility of our proposed approaches for extracting interesting and useful information from multivariate financial data streams. We end the chapter by providing concluding remarks focused on the merits and limitations of our proposed approaches.

4.1 Properties of financial time series

Financial markets are highly complex and dynamic systems which play a pivotal role in the globalised economy. Due to the vast scale of the global financial markets and their constant evolution, research in this sector presents unique challenges and opportunities. This section provides an overview of the basic stylised statistical facts of financial time series, in particular FX log-returns (as most of the analysis presented in this thesis makes use of FX data). We start by providing a brief description of the global FX market. The FX market is by far the largest financial market in the world, accounting for over $4 trillion in average daily turnover, which includes $1.5 trillion worth of spot transactions [1]. The FX market operates on a 24 hour basis, and spans all time zones. It is active for five days a week and each day is generally considered to comprise of three 8-hour trading sessions. There are over 33 currencies that are actively traded [1]. However, a large bulk of the global FX trading volume is accounted for by four major currencies, the United States dollar (USD), Euro (EUR), Japanese yen (JPY) and the British pound (GBP). These four currencies together account for some 77% of all global FX

trading volume [1]. FX currencies are traded in pairs. As an example, the EURUSD rate refers to the number of EUR one USD can buy. The FX rate is adjusted according to the strength or weakness of each component of a currency pair. So the EURUSD value will increase if the EUR is strengthened with respect to the USD or the USD is weakened with respect to the EUR. The three most liquid currency pairs are EURUSD, USDJPY and GBPUSD, which together account for close to 51% of all trades placed. FX data is electronically provided by financial data providers such as Reuters and Electronic Broking Services (EBS). The FX market is a highly dynamic and liquid entity, which has over time become increasingly interlinked with the wider economy, due to which even a slight variation in any financial index can influence the exchange rates of various dependent currencies and vice versa[1]. This has resulted in significant interest, amongst practitioners and academics alike, into investigating the structural properties of the FX market, in particular the nature of symmetric and asymmetric interactions between various currency pairs [69, 257, 284]. Interest in this field also stems from the fact that the exchange rate of a single currency pair generally does not contain much (practically useful) extractable information [113]. However, knowledge about the way in which exchange rates of different currency pairs interact can be used to improve our understanding of the driving dynamics of the FX market.

One of the most prominent changes in the FX market (and across many other asset classes) in recent years has been the rapid growth of algorithmic trading[2]. Until recently, algorithmic trading strategies usually made use of low- or mid-frequency data. However, due to the easy and relatively cheap availability of high-frequency market data, some of the latest algorithmic trading engines trade on a sub-second or even tick by tick (shortest time interval between quote updates) basis. The dominance of algorithmic trading making use of high-frequency data can be judged from the fact that it currently accounts for over 70% of all trading volume in the U.S.

[1] As an example, U.S. consumer price index (a measure of inflation in the U.S.) can affect the relative exchange rates of numerous currency pairs across the globe. Likewise, there are many other underlying factors which can effect the global FX markets, e.g. the interbank interest rates, interest rate differential, relative liquidity of the currencies, overall market sentiment, inflation, variation in the gross domestic product of a country, U.S. non-farm payroll data, house price indices, political stability of a region, among many others [118, 233].

[2] Algorithmic trading refers to an automated trading platform which relies on statistical signal processing algorithms to make online trading decisions. Since the introduction of electronic trading in 1971 [329], the proportion of trades that can be attributed to algorithmic trading has steadily increased. Initially algorithmic trading strategies were deployed primarily in the equities markets, however, recently they are being increasingly used in the FX market as well [69]. Many of the algorithmic trading engines currently in use harness market inefficiencies in order to generate positive returns.

capital markets [361]. High-frequency financial data incorporates the rapidly changing dynamics of financial markets, which allows practitioners to develop more robust trading algorithms based on the micro-structure of the markets [135]. Most of these algorithms, often deployed as part of "real-time trading" (RTT) models, use real-time price information, electronically provided by brokerage firms such as Reuters and EBS, to make online trading decisions. A comprehensive analysis of the use of RTT models in the FX market is presented in [135], empirical results (obtained using seven years of high-frequency data) presented in this study show that RTT models can consistently generate positive returns while deployed in the FX market. Many of these RTT models capture statistical inefficiencies in the financial markets to generate a risk-free return, a process commonly known as statistical arbitrage [294][3]. Use of interaction measurement approaches can aid in identifying statistical arbitrage opportunities, for example, by estimating the coupling or causal links between the real-time bid and ask quotes being provided by various brokers and selecting a broker which continuously provides a favourable rate; thus, information obtained using interaction measurement approaches can be used to predict the availability of statistical arbitrage opportunities and to capture these opportunities as soon as they arise.

4.1.1 FX market terminology

In this thesis, all currencies are referred by their standardised international three-letter codes, as described by the ISO-4217 standard. For the currencies mentioned in this thesis, the three-letter codes are: USD (U.S. dollar), EUR (Euro), JPY (Japanese yen), GBP (British pound), CHF (Swiss franc), AUD (Australian dollar), NZD (New Zealand dollar), CAD (Canadian dollar), NOK (Norwegian krone) and SEK (Swedish krona). Together, these ten currencies are often referred to as the G10 (Group of Ten) currencies. Some commonly used terms which are associated with FX data (and which we use in this thesis) are as follows:

- *Spot price*: The price which is actually quoted for a transaction to take place is called the spot price or spot rate.

[3] A specific type of statistical arbitrage is commonly known as triangular arbitrage. Triangular arbitrage trading is a process by which profit is made by exploiting statistical differences between exchange rates of three FX currency pairs [8]. The currency pair used as the third, and final, leg of a set of triangular arbitrage trades is often referred to as the arbitrage-leg currency pair.

- *Bid price*: The highest price that a buyer, e.g. a market maker, is willing to pay to buy a currency is called the bid price or buy price.

- *Ask price*: The lowest price at which a seller, e.g. a market maker, is willing to sell a currency is called the ask price or sell (offer) price. Ask price is almost always higher than the bid price.

- *Mid-price*: Mid-price is the average of the bid price and ask price of a currency pair at any given time. It is a non-stationary process.

- *Spread*: Spread is the difference between the ask price and bid price of a currency pair at any given time. It is generally known as the bid/ask, bid/offer or buy/sell spread. It is also a pseudo-measure of the liquidity of the market for any given currency pair.

- *Volatility*: Volatility is defined as the standard deviation of a financial time series. In financial markets, volatility is often referred to as the *beta coefficient* and is commonly used to calculate the risk associated with the underlying asset. In practise, volatility is often predicted using a generalised AR conditional heteroskedasticity (GARCH) model [178].

- *Liquidity*: Liquidity refers to general interest in the market for buying or selling any given currency pair. Liquidity is generally inversely proportional to the spread of a currency pair, i.e. more liquid currency pairs generally have lower spreads (on average) than less liquid ones.

- *Pip*: FX spot prices are typically quoted to the fourth decimal point, e.g. EURUSD bid/ask rate is generally quoted as 1.3500/1.3501. A major exception to this rule are JPY crosses which are quoted to the second decimal point. However, recently prices are being quoted to the fifth decimal place for some liquid currency pairs. The smallest price change in the exchange rate, by convention the fourth decimal place (second for JPY crosses) for most currency pairs, is defined as a pip, e.g. in the EURUSD example the spread is 1 pip. FX pairs are generally traded in lots of 1 million. So a ± 1 pip move in one lot of EURUSD will translate into a profit and loss (PnL) of ± 100 USD.

- *Implied price*: Data for two "direct-leg" FX currency pairs, i.e. each currency pair contains a common currency, can be used to generate data for a third pair, giving an "implied price". For example, if we have data for the direct EURUSD and EURCHF exchange rates, then we can obtain implied price for USDCHF simply by dividing values of one data set by the corresponding values of the other, a process known as *triangulation*, i.e.:

$$P_{USDCHF}(t) = \frac{P_{EURCHF}(t)}{P_{EURUSD}(t)} \tag{4.1}$$

where $P(t)$ refers to the mid-price at time t. Similarly, if we have data for a particular currency pair, e.g. USDCHF, we can obtain data for CHFUSD simply by taking the reciprocal of the exchange rates, i.e.:

$$P_{CHFUSD}(t) = \frac{1}{P_{USDCHF}(t)} \tag{4.2}$$

The analysis presented above can be used to obtain data for multiple combinations of various currency pairs. For multiple currency pairs containing a total of N unique currencies, the number of possible combinations is $\frac{N}{2}(N-1)$. In the example above, there are 3 unique currencies, EUR, USD and CHF, therefore the total number of possible currency pairs which we can obtain using this data is 3. For financial applications, such as analysing financial networks, different permutations of the currency pairs are not important to consider, as the information coupling between currency pairs is independent of their permutations. We make use of implied FX prices in some of the examples presented later in the thesis.

4.1.2 Properties of FX log-returns

As already stated, the price which is actually quoted for a currency transaction to take place is called the spot price or spot rate. Return is the fractional change in the exchange rate of a currency pair at any given time. For an exchange rate of $P(t)$ at time t, the arithmetic spot return is given by:

$$R(t) = \frac{P(t) - P(t-1)}{P(t-1)} \tag{4.3}$$

It is common practise to use the log of returns in most calculations. Using log-returns makes it possible to convert exponential calculations into linear ones, thus significantly simplifying relevant analysis. A normalised log-returns data set, with a mean of zero and unit variance, can generally be regarded as a locally stationary [114, 245] and locally linear [300] process[4]. Therefore, many signal processing techniques meant solely for stationary and linear processes can be successfully applied to the normalised log-returns time series in an adaptive environment. FX spot returns in the log-returns space can be written as:

$$r(t) = \log\left[\frac{P(t)}{P(t-1)}\right] \tag{4.4}$$

FX log-returns have many inherent properties which can be used to extract valuable information from within FX data sets. They show time-scale behaviour in the log-returns space and exhibit rapidly changing dynamics. They often have highly non-Gaussian distributions, as discussed below.

Many statistical models are based on a prior assumption about the shape of the distribution of the data being analysed. As an example, linear correlation analysis assumes that the bivariate time series between which correlation is being computed have elliptical distributions, of which the multivariate Gaussian is a special case [197]. The linear correlation measure will give misleading results if the multivariate Gaussian condition is not met, even if individual distributions of the signals are Gaussian [116]. For the purpose of analysis, it is safe to assume that a set of signals, each of which have individual non-Gaussian distributions, also have multivariate non-Gaussian (non-elliptical) distributions, and hence are not suitable to be analysed using the linear correlation measure. Therefore, it is important to briefly look at properties of the pdfs of FX time series. FX log-returns have unimodal non-Gaussian distributions with heavy-tails and generally tend to be leptokurtic [89, 258]; log-returns of data belonging to other asset classes (such as equities) also exhibit similar properties. The distributions tend to be slightly skewed and become increasingly non-Gaussian as the frequency at which the data

[4]Financial time series representing the mid-price of an asset's value are usually non-stationary [249]. However, in the log-returns space financial time series are considered to be locally stationary, i.e. they display quasi-stationarity [85]. As the models presented in this thesis make use of normalised log-returns data sets within an adaptive environment to dynamically measure interactions, therefore within each window the data can be considered to be stationary [132]. Moreover, the sliding-window ICA algorithm used in our models is good at handling non-stationary data, allowing the ICA-based information coupling model to deal with any non-stationary dynamics.

is being analysed increases [244]. Therefore, to accurately capture information in the tails of these non-Gaussian data sets, statistical approaches implicitly making use of higher-order statistics (such as those proposed in this thesis) need to be used.

To test the hypothesis that a set of samples from a FX log-returns data set come from a non-Gaussian distribution (and to measure their "degree" of non-Gaussianity), we can make use of the Jarque-Bera (JB) statistic [338]. The JB test uses the skewness and kurtosis of a data set to calculate a value for the degree of non-Gaussianity; as most commonly used tests for normality make use of a sample's skewness and kurtosis [209], therefore the JB statistic is a useful measure of a sample's non-Gaussianity. The JB statistic is calculated using the following equation:

$$JB = \frac{n_s}{6}\left[\gamma^2 + \frac{(\kappa-3)^2}{4}\right] \tag{4.5}$$

where n_s is the number of samples under consideration, γ is the skewness of the distribution and κ is its kurtosis. For normal distributions $\gamma = 0$ and $\kappa = 3$, therefore $JB = 0$. For large data sets, the Jarque-Bera test uses the chi-squared distribution to estimate a critical value (JB_c) at a particular significance level. If value of the JB statistic is equal to, or greater than, this critical value, i.e. $JB \geq JB_c$, then the null hypothesis that the sample comes from a normal distribution is rejected. A larger value of JB implies a higher degree of non-Gaussianity. As an example, Figure 4.1 shows log-returns and their associated pdfs for EURUSD (the most liquid currency pair) at two different sampling frequencies. The pdf plots clearly show the non-Gaussian nature of the data at both sampling frequencies. For this example, we obtained JB statistic values of 7.73×10^5 and 2.06×10^5 respectively for the 0.5 second sampled and 0.5 hour sampled data sets, while the associated JB_c values were 13.92 and 14.06 respectively (at a significance level of 0.1%); hence, the null hypothesis of the samples coming from a normal distribution is rejected in both cases. The very high values of the JB statistic are indicative of the highly non-Gaussian nature of the data, and as the JB value obtained for the 0.5 second sampled data is significantly higher than the 0.5 hour sampled data, we can conclude that the high-frequency sampled data is more non-Gaussian (this is a general property of financial returns, as we discuss in detail later). We describe statistical properties of financial data (which we use in this thesis) in more detail in the next section.

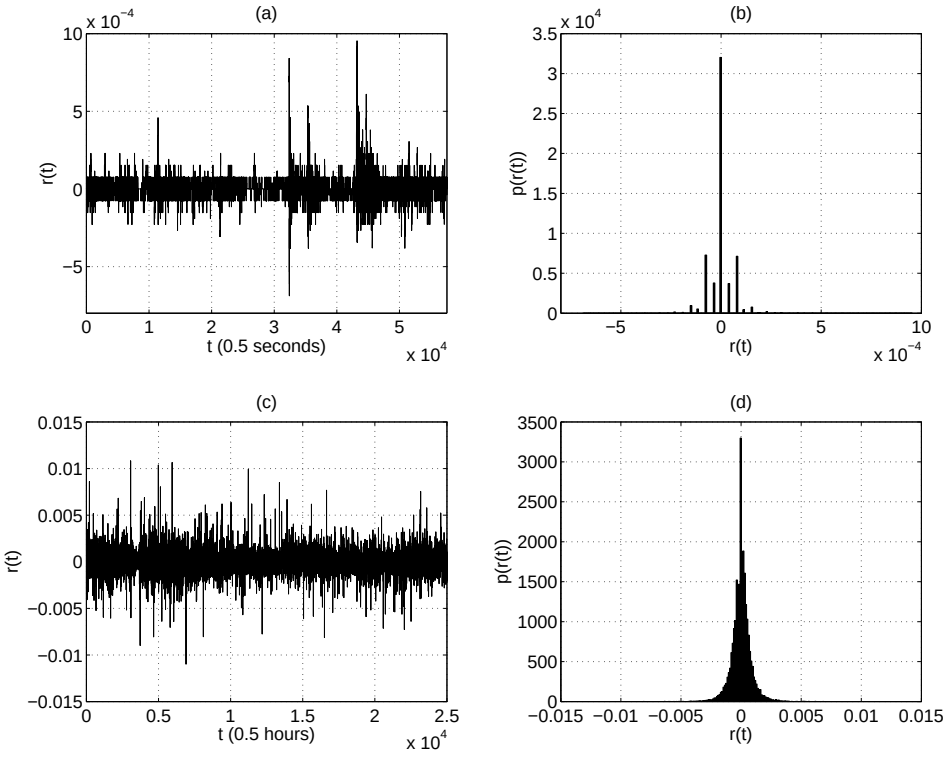

Fig. 4.1: (a,b). Log-returns of EURUSD (0.5 second sampled) over a period of eight hours and its associated pdf plot. (c,d). Log-returns of EURUSD (0.5 hour sampled) over a period of two years and its associated pdf plot.

4.2 Description of data analysed

Having reviewed some basic stylised facts of financial log-returns, we now proceed to provide a detailed description of the data sets used in this thesis together with their statistical properties. The first part of this section presents details of the financial data, while the second part describes the synthetic data used in this thesis. In most cases, we normalise the raw data before analysis. Normalisation is achieved by converting the data to a form such that it has a mean of zero and unit variance. This is easily achieved by removing the mean and dividing by the standard deviation of the time series.

4.2.1 Financial data

To obtain the results included in this thesis, we make use of five financial data sets. Four of these are spot FX data sets sampled at varying frequencies, covering all major (G10) currency pairs. The fifth is a daily sampled equities data set. The primary reason for using different

data sets is based on the practical financial applications for which these data sets are relevant, as a single data set cannot be used to show the efficiency, practical utility and the effectiveness of the different interaction models presented in this thesis. Using data sampled at different frequencies also allows us to test for robustness of the interaction models in dealing with financial data at a range of sampling frequencies. The data sets, together with their salient features, are described in Table 4.1.

Asset class	Sampling period	Number of samples[a]	Length of dataset
FX (spot)	0.25 sec	130,000	9 hours[b]
FX (spot)	0.50 sec	58,000	8 hours
FX (spot)	0.50 hour	25,000	2 years
FX (spot)	1 day	2,600	10 years
Equities	1 day	2,600	10 years

Table 4.1: Salient features of the data sets used to obtain the results presented in this thesis. [a]Approximate number of samples available; depending on the application, not the whole data set is always used. [b]For this data set, data for five trading sessions (over five days) is used in analysis (total of $5 \times 130,000$ data points).

Financial returns have non-Gaussian (fat-tailed) distributions, as is evident from the summary statistics presented in Table 4.2. The table shows the average kurtosis (κ), a measure of the tail distribution, as well as the average JB statistic values for four major liquid currency pairs from each of the four spot FX data sets analysed in this thesis. Also included in the table are the 25th and 75th percentile kurtosis values as well as the p-values associated with the JB statistic estimation. Noting that a standard normal distribution has a kurtosis of 3, it is clear that all the kurtosis values in the table show the presence of fat-tailed distributions. Likewise, all the JB statistic values at higher frequencies point to the highly non-Gaussian nature of the data, with non-Gaussianity generally increasing with the frequency at which the data is sampled, a well-known stylised fact as described in [98]. All values are obtained using averages over 50 data point long samples, in order to reflect properties of the data analysed dynamically using various interaction models later in the thesis[5]. We now take a more in-depth look into the distribution of FX data which we analyse. Figure 4.2 shows the cumulative distribution plots

[5]Similarly, for the daily sampled equities data set analysed in this thesis, the average kurtosis for a sample size of 50 data points is 3.72 with $\kappa_{25\%} - \kappa_{75\%}$ range of 2.93-4.37. The average JB statistic value is 7.21 (p-value of 0.2160) which is higher than the JB critical value (JB_c) of 4.95 (at 5% significance level). Hence, the equities data analysed is also non-Gaussian in nature with heavy-tailed distributions.

of normalised log-returns for the four FX data sets sampled at different frequencies; for ease of presentation, the plots only show distribution of EURUSD as a representative example. The plots clearly show non-Gaussian nature of the data, especially at higher frequencies. The results presented in Table 4.2 and Figure 4.2 show that the 0.25 second and 0.5 second sampled data sets have broadly similar summary statistics and so do the 0.5 hour and daily sampled data sets. Therefore, for most of the general results presented in this thesis, we make use of the 0.5 second and 0.5 hour sampled data sets as representative examples of FX data sampled at high and medium frequencies respectively; however, as previously mentioned, when presenting specific financial case studies, we make use of data sets which are practically relevant for that application domain.

	κ ($\kappa_{25\%} - \kappa_{75\%}$)				JB_{avg} ($JB_c = 4.95$) (p-value)			
Data	EURUSD	GBPUSD	USDJPY	EURCHF	EURUSD	GBPUSD	USDJPY	EURCHF
FX: 0.25 sec	13.2	10.1	14.8	13.2	338.7	155.5	459.9	363.1
	(7.6-15.8)	(5.3-10.8)	(8.3-18.0)	(6.9-14.6)	(0.0070)	(0.0180)	(0.0073)	(0.0147)
FX: 0.5 sec	15.3	15.6	16.1	21.0	630.0	598.6	658.3	1200.7
	(6.3-21.8)	(7.7-18.0)	(7.1-21.8)	(6.6-27.5)	(0.0086)	(0.0070)	(0.0054)	(0.0022)
FX: 0.5 hour	7.3	6.3	6.7	5.6	102.4	60.3	77.8	35.7
	(3.9-8.3)	(3.9-7.1)	(3.7-6.5)	(3.6-6.5)	(0.1077)	(0.1056)	(0.1457)	(0.1532)
FX: 1 day	3.2	3.7	4.8	4.9	2.6	2.4	46.9	32.1
	(2.6-3.3)	(2.7-3.7)	(3.3-4.5)	(2.7-5.7)	(0.3386)	(0.3425)	(0.2453)	(0.1877)

Table 4.2: Table showing summary statistics of the data sets used in this thesis. The average kurtosis (κ) together with its 25th ($\kappa_{25\%}$) and 75th ($\kappa_{75\%}$) percentile values show the fat-tailed nature of the distributions. The average JB statistic (JB_{avg}) values, which take into account the skewness and kurtosis of the data, show the highly non-Gaussian nature of the high-frequency sampled data sets. The critical values of the JB statistic (JB_c) are calculated at a significance level of 0.05 (5%). The results show average values for 50 data points long samples, in order to reflect general properties of the data analysed dynamically using various interaction models later in the thesis.

It is important to note that the degree of non-Gaussianity of FX spot returns varies dynamically with time (i.e. the distributions are not stable across time), as shown by the four representative examples presented in Figure 4.3. The plots show significant variations in the temporal value of the JB statistic. The critical value of the JB statistic (JB_c) is 5.88 (at 5% significance level) for the results presented, which is much lower than the value of the JB statistic at all times for all four plots. These results once again show the non-Gaussian, dynamically changing, properties of FX spot returns. Similarly, in Figure 4.4 we show the distribution of two higher-order moments, i.e. skewness (γ) and kurtosis (κ) for data sets sampled at three different frequencies. The results are obtained using a sliding-window of length 50 data points,

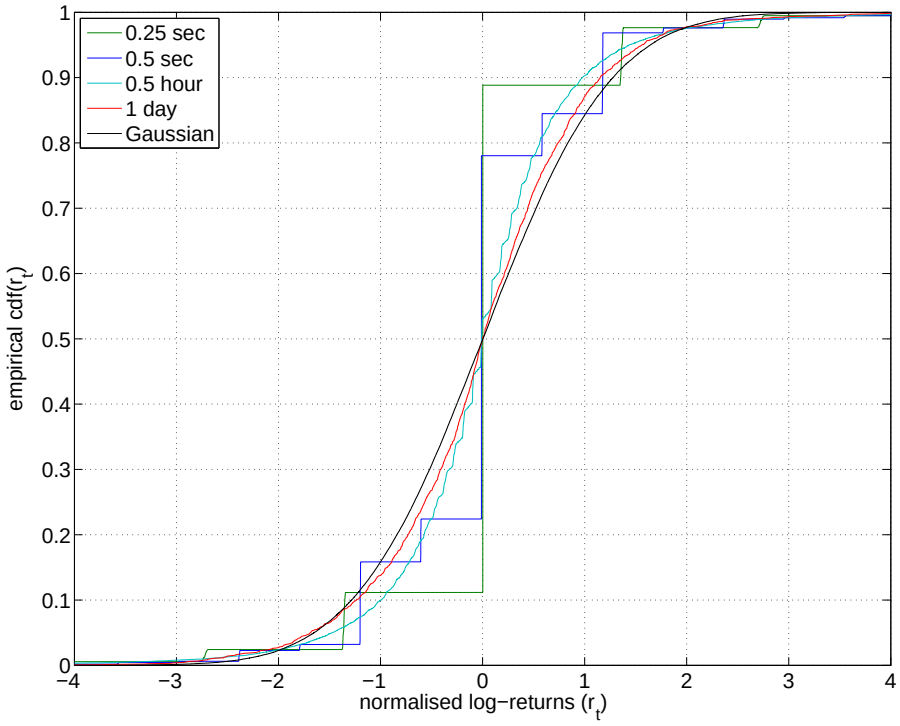

Fig. 4.2: Plots showing the empirical cumulative density function (cdf) of normalised log-returns for EURUSD at different frequencies. Also included is a plot showing the cumulative distribution for a Gaussian distribution. The non-Gaussian nature of the data is clearly visible, with non-Gaussianity generally increasing with the frequency at which the data is sampled.

as an average of all G10 currency pairs, covering a period of 8 hours in the case of 0.25 second and 0.5 second sampled data and 2 years in case of 0.5 hour sampled data. Once again, the non-Gaussian (heavy-tailed, skewed) nature of the data is clearly visible. It is interesting to note that the kurtosis value almost never goes below three for any of the data sets, signifying the temporal persistence of non-Gaussianity for medium and high frequency sampled FX log-returns. These results once again signify the need for developing and using interaction measurement approaches which take into account higher-order statistics of the data being analysed. With the widespread availability and use of high-frequency sampled financial data over the last few years, standard second-order approaches which previously may have been reasonably accurate when used for analysing low-frequency sampled data, will give misleading results when used for analysing high-frequency data (especially within a dynamic environment).

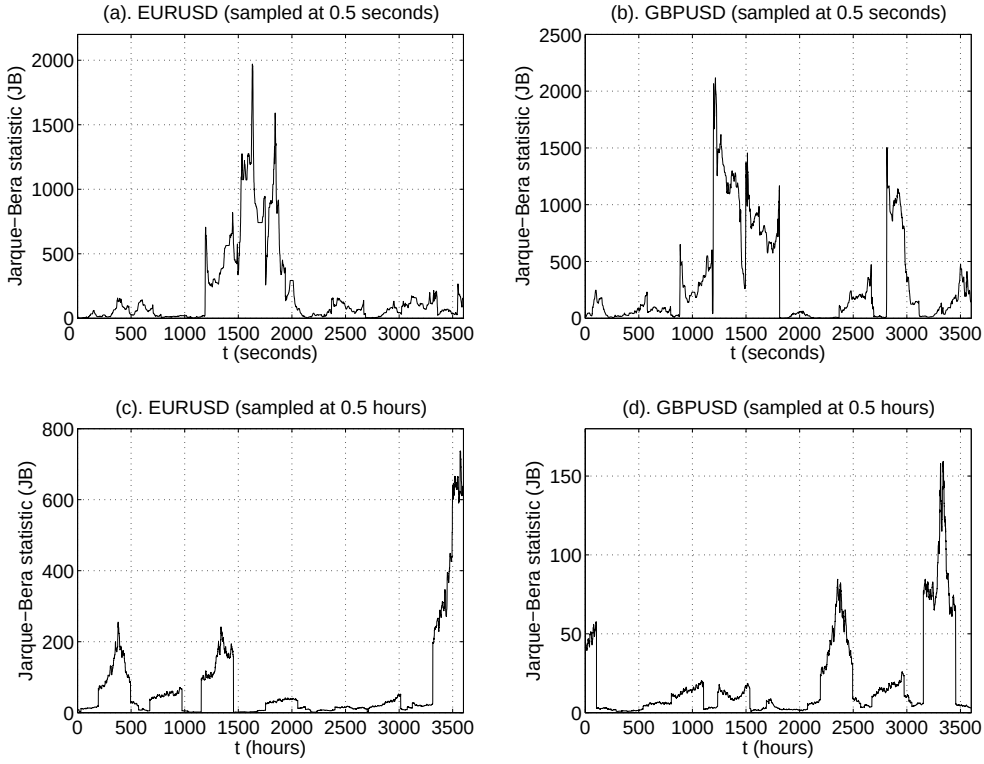

Fig. 4.3: Snap-shots of plots representing temporal variation of the JB statistic for FX log-returns. Plots (a) and (b) represent results obtained using 0.5 second sampled EURUSD and GBPUSD data sets respectively. Plots (c) and (d) represent results obtained using 0.5 hour sampled EURUSD and GBPUSD data sets respectively. All the plots were obtained using a sliding-window 600 data points in length. $JB_c = 5.88$ (at 5% significance level) for all four plots. Note the different y-axes for all plots.

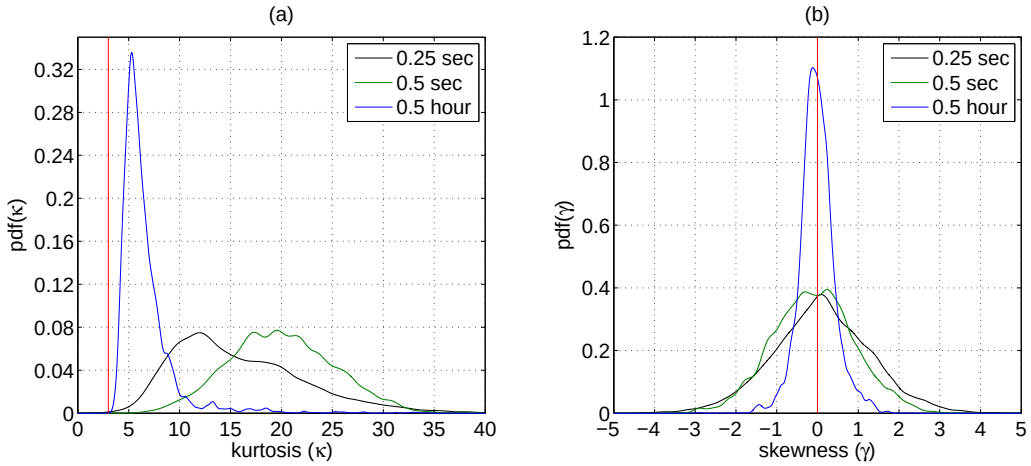

Fig. 4.4: Plots showing the normalised pdfs of: (a). kurtosis (κ) and (b). skewness (γ), obtained using a sliding-window of length 50 data points, as an average of all G10 currency pairs, covering a period of 8 hours in case of 0.25 second and 0.5 second sampled data and 2 years in case of 0.5 hour sampled data. The red vertical lines show values for a Gaussian distribution. The plots clearly show persistence in non-Gaussianity (heavy-tailed) of FX log-returns at all three sampling frequencies.

4.2.2 Synthetic data

Data generated from a Gaussian distribution can be uniquely described by the mean and variance of the distribution. However, for non-Gaussian data there can be a range of different distributions which approximately fit the data with various parameters. For such data sets, the higher-order moments also need to be taken into account. There is a vast amount of work which has been done with the aim of explaining the behaviour of financial returns using a variety of parametric distributions. However, it is important to note that there is no global distribution which fits all types of financial returns [27]. Generally, different asset classes and individual instruments within each asset class exhibit different distributions. These distributions are not stable and can rapidly change with time depending on the market conditions (as we empirically demonstrated earlier). Moreover, there is a great deal of variation in the properties of these distributions conditioned on the frequency at which the data is being analysed. From Table 4.2 and Figure 4.4, we can see that the average kurtosis of FX data sampled at higher frequencies (0.25 sec/0.5 sec/0.5 hour), which we use for majority of the examples presented in this thesis, is 12.1. Our analysis also shows that the average skewness for this data is -0.24; the slight negative skewness of financial returns has also been noticed previously in [285]. These skewness and kurtosis values give us an average JB statistic value of 173.0 for the data analysed, which is much higher than the critical value of 4.95 (at 5% significance level), showing the highly non-Gaussian nature of the data (as we use normalised data, therefore the first (mean) and second (variance) moments stay constant at 0 and 1 respectively). We use these estimates for the higher-order moments as a guide to generate synthetic data to test our models (as described below).

As already mentioned, there is no single distribution which fits financial returns, especially those sampled at higher frequencies which tend to be highly non-Gaussian [61], although there have been attempts to model returns using a variety of distributions [151]. In this thesis, we aim to capture the heavy-tailed, skewed, properties of financial returns using a Pearson type IV distribution [76, 263], which can be used to generate data with the desired mean, variance, skewness and kurtosis values, and thus is useful for representing distributions of financial returns [322]. The first four moments of this distribution can be uniquely determined by setting four parameters which characterise the distribution, which is analytically given by

[332, 348]:

$$p_{IV}(x) = \frac{\left|\frac{\Gamma\left(m+\frac{\nu}{2}i\right)}{\Gamma(m)}\right|^2}{\alpha\,\mathrm{B}\left(m-\frac{1}{2},\frac{1}{2}\right)}\left[1+\left(\frac{x-\lambda}{\alpha}\right)^2\right]^{-m}\exp\left[-\nu\arctan\left(\frac{x-\lambda}{\alpha}\right)\right] \tag{4.6}$$

where λ and α are the *location* (mean) and *scale* (variance) parameters respectively, m and ν are the *shape* parameters (skewness and kurtosis), while Γ and B denote the Gamma and Beta functions respectively. Until recently, due to its mathematical and computational complexity, this distribution has not been widely used in financial literature [327], although this is rapidly changing with advances in computational power and proposal of new, improved analytical methods and related algorithms [76, 200, 354]. Unless otherwise stated, we use data sampled from a Pearson type IV distribution (denoted by p_{IV}), with properties described earlier in this section, for all the synthetic data examples presented later in the thesis.

4.3 Analysis of synthetic data

Earlier we presented a theoretical overview of some commonly used approaches to symmetric interaction measurement, and compared their relative advantages and limitations. We now delve further into this topic by empirically comparing these approaches with the ICA-based information coupling measure. Unless otherwise indicated, the following notations are used for different measures of symmetric interaction in this thesis: ICA-based information coupling (η), linear correlation (ρ), Spearman's rank correlation (ρ_R) and normalised mutual information (I_N).

Comparative analysis

To test accuracy of the symmetric interaction models, we need to generate correlated non-Gaussian data with known, pre-defined, correlation values. There is no straightforward way to simulate correlated random variables when their joint distribution is not known [160], as is the case with multivariate financial returns. One possible method that can be used to induce any desired pre-defined correlation between independent (randomly distributed) variables, irrespective of their distributions, is commonly known as the Iman-Conover method, as presented in [190]. This method is based on inducing a known dependency structure in samples taken

from the input independent marginal distributions using reordering techniques. The multivariate coupled structure obtained as the output can thus be used as the input data in various interaction models to test their relative accuracies. We use the Iman-Conover method to induce varying levels of correlation between 1000 data points long samples taken from an independent (randomly distributed) bivariate Pearson type IV distribution. A 1000 data points long sample makes it easier to accurately induce pre-defined correlations in the system as well as makes it possible to generate data using a Pearson type IV distribution with relatively accurate average kurtosis and skewness values. Figure 4.5 shows representative scatter plots for the coupled data for different levels of "true correlation" (ρ_{TC}).

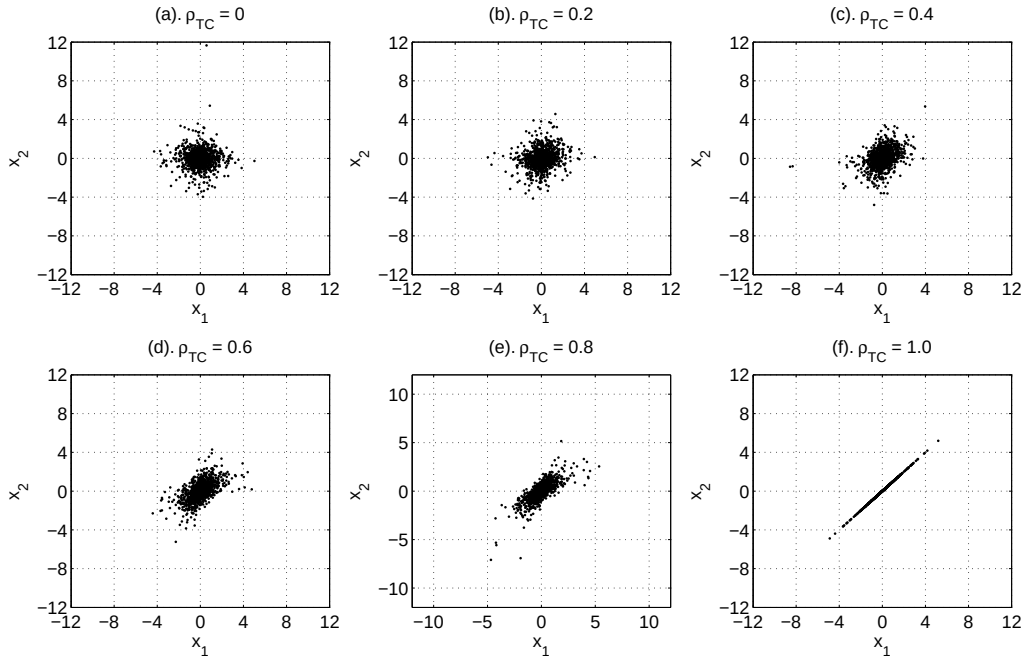

Fig. 4.5: Scatter plots showing a representative sample used to test the accuracy of the different interaction measures. ρ_{TC} is the true correlation induced into the system. The underlying marginal distributions are sampled from an independent (randomly distribution) bivariate Pearson type IV distribution.

Four different approaches are now used to estimate the level of dependence between the output coupled data. The process is repeated 1000 times for each level of ρ_{TC}. The average kurtosis values for the 1000 simulations of the two coupled variables were 12.00 and 12.17 and mean skewness values were -0.0540 and -0.0966 which closely match properties of the financial data sets (as presented earlier). Figure 4.6 shows distribution of the kurtosis and skewness of the two variables for different simulations. We note the similarity of these plots

with the (average of) corresponding distributions of higher-order moments for financial data, as presented in Figure 4.4. Once again, this shows the effectiveness of using synthetic data sampled from a Pearson type IV distribution for capturing higher-order moments of financial returns. Results of the comparative analysis are presented in Table 4.3. The results show accuracy of the information coupling measure when used to analyse non-Gaussian data. For this synthetic data example, on average, the information coupling measure was 53.7% more accurate than the linear correlation measure and 25.6% more accurate with respect to the rank correlation measure. The normalised mutual information provided the least accurate results.

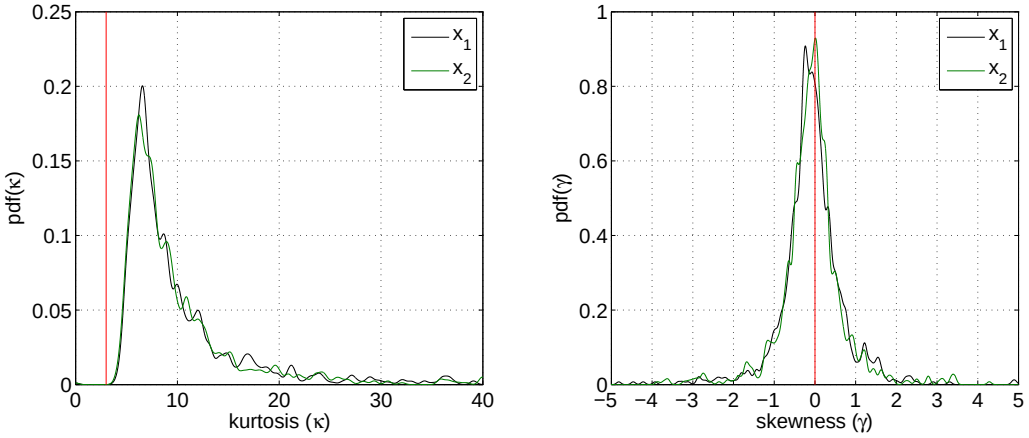

Fig. 4.6: Normalised pdf plots showing the average kurtosis (κ) and skewness (γ) values for the data used to test accuracy of the dependency measures. The red vertical lines show values for a Gaussian distribution. We note that the data has properties similar to (average of) the distributions of financial returns sampled at higher-frequencies, as presented in Figure 4.4. The results of the comparative analysis are presented in Table 4.3.

We now extend this example by incorporating data dynamics. The same data generation process (as described above) is now used to construct a 32000 data points long bivariate data set in which the induced true correlation changes every 8000 time steps, i.e. $\rho_{TC} = 0.2$ when t=1:8000, $\rho_{TC} = 0.4$ when t=8001:16000, $\rho_{TC} = 0.6$ when t=16001:24000 and $\rho_{TC} = 0.8$ when t=24001:32000. A 1000 data points wide sliding-window is used to dynamically measure dependencies in the data set. The resulting temporal information coupling plot, together with the 95% confidence bounds, is presented in Figure 4.7(a). The four different coupling regions are clearly visible, together with the step changes in coupling after every 8000 time steps; showing ability of the algorithm to detect abrupt changes in coupling. The normalised empirical probability distributions over η for the four coupling regions are shown in Figure

ρ_{TC}	η	ρ	ρ_R	I_N	$\lvert\rho_{TC}-\eta\rvert$	$\lvert\rho_{TC}-\rho\rvert$	$\lvert\rho_{TC}-\rho_R\rvert$	$\lvert\rho_{TC}-I_N\rvert$
0	0.0046±0.0026	0.0038±0.0022	0.0147±0.0115	0.1851±0.0401	0.0046	0.0038	0.0147	0.1851
0.1	0.1079±0.0073	0.0917±0.0058	0.0942±0.0184	0.1687±0.0412	0.0079	0.0083	0.0058	0.0687
0.2	0.2145±0.0102	0.1862±0.0093	0.1899±0.0177	0.1131±0.0464	0.0145	0.0138	0.0111	0.0869
0.3	0.3176±0.0138	0.2812±0.0130	0.2863±0.0167	0.1644±0.0538	0.0176	0.0188	0.0137	0.1356
0.4	0.4166±0.0210	0.3764±0.0165	0.3835±0.0153	0.2787±0.0500	0.0166	0.0236	0.0165	0.1213
0.5	0.5135±0.0196	0.4720±0.0197	0.4818±0.0136	0.3819±0.0475	0.0135	0.0280	0.0182	0.1181
0.6	0.6070±0.0347	0.5688±0.0226	0.5815±0.0117	0.4788±0.0458	0.0070	0.0312	0.0185	0.1212
0.7	0.7011±0.0232	0.6670±0.0242	0.6830±0.0093	0.5728±0.0483	0.0011	0.0330	0.0170	0.1272
0.8	0.7936±0.0239	0.7676±0.0249	0.7864±0.0066	0.6652±0.0490	0.0064	0.0324	0.0136	0.1348
0.9	0.8864±0.0252	0.8713±0.0249	0.8919±0.0034	0.7595±0.0501	0.0136	0.0287	0.0081	0.1405
1.0	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000	0.9601±0.0185	0.0000	0.0000	0.0000	0.0399
MAE					0.0093	0.0201	0.0125	0.1163

Table 4.3: Table showing accuracy of four measures of dependence, i.e. information coupling (η), linear correlation (ρ), rank correlation (ρ_R) and normalised mutual information (I_N), when used to estimate the level of dependence in a coupled system with varying levels of true correlation (ρ_{TC}). ρ_{TC} is induced in an independent (randomly distributed) bivariate system using the Iman-Conover method as described in the text. The dependence estimates, together with their standard deviation confidence intervals, shown in the table are obtained using 1000 independent simulations using 1000 data points long data sets for each simulation. The last row of the table gives values for the mean absolute error (MAE).

4.7(b). Also plotted in the same figure are the normalised empirical pdfs for linear correlation (ρ) and rank correlation (ρ_R); the mutual information pdf is omitted for clarity as it gives relatively less accurate results (as presented in Table 4.3). It is interesting to see how the peaks of the η distribution correspond very closely to ρ_{TC} values, showing ability of the information coupling model to accurately capture statistical dependencies in a dynamic environment. The least accurate measure in this example is the linear correlation.

Let us now consider another comparative empirical example. We know that the ranking order of a variable under a monotonic transformation is preserved [156], implying that the rank correlation of two variables, which are both transformed using the same monotonic transformation, will be the same as the rank correlation of the original variables. This example makes use of this property of rank correlations to compare the relative accuracy of different measures of symmetric interaction. Consider a normally distributed random variable, $x_1(t) \sim \mathcal{N}(0,1)$, which is linked to another variable, $x_2(t)$, as follows:

$$x_2(t) = \alpha x_1(t) + (1-\alpha)n(t) \tag{4.7}$$

where $n(t)$ is white noise. For the analysis which follows, we use 1000 data points long samples of the two variables for each simulation, denoted by $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. For each

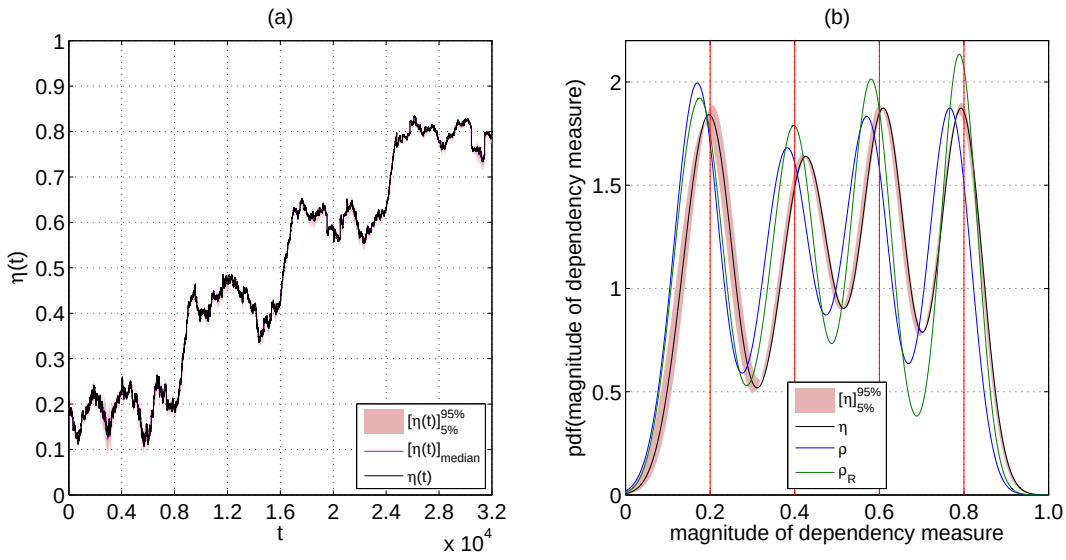

Fig. 4.7: (a). Temporal variation of information coupling, $\eta(t)$, plotted as a function of time. Step changes in coupling are visible after every 8000 time steps, i.e. $\rho_{TC} = 0.2$ when t=1:8000, $\rho_{TC} = 0.4$ when t=8001:16000, $\rho_{TC} = 0.6$ when t=16001:24000 and $\rho_{TC} = 0.8$ when t=24001:32000. Also plotted are the median, $[\eta(t)]_{\text{median}}$, and the 95% confidence interval contours, $[\eta(t)]^{95\%}_{5\%}$. (b). Normalised empirical pdf plots for information coupling (η), linear correlation (ρ), and rank correlation (ρ_R). The vertical lines represent the true correlation (ρ_{TC}) values. The relative accuracy of the information coupling measure is evident from these results.

value of α from 0 to 1, in steps of 0.001, we calculate the rank correlation (ρ_R) between normalised values of two different transformations of the variables, i.e. between $\exp(\mathbf{x}_1)$ and $\exp(\mathbf{x}_2)$, and between $\mathbf{x}_1^3$ and $\mathbf{x}_2^3$; we repeat this process 100 times for each value of α. As these are monotonic transformations, therefore, for each value of α, rank correlation will be the same for all three set of variables, i.e. $\rho_R(\mathbf{x}_{1,2}) = \rho_R(\exp(\mathbf{x}_{1,2})) = \rho_R(\mathbf{x}_{1,2}^3)$. Figures 4.8(a) and 4.8(c) present a representative example (at $\alpha = 0.5$) of the pdf plots for the variables after the transformations, which clearly show the non-Gaussian nature of the data. We also estimate the information coupling, linear correlation and mutual information at each value of α, and calculate the absolute error (AE) between the results obtained and the corresponding rank correlation value. Figures 4.8(b) and 4.8(d) show the normalised pdf plots for the AEs obtained, while Table 4.4 gives values for the mean absolute error (MAE) obtained using different measures of interaction. Once again, we note accuracy of the information coupling measure compared to standard dependency measures.

It is clear from our discussion so far that any practical measure of interaction for analysing financial data needs to be able to handle non-Gaussian data in a computationally efficient

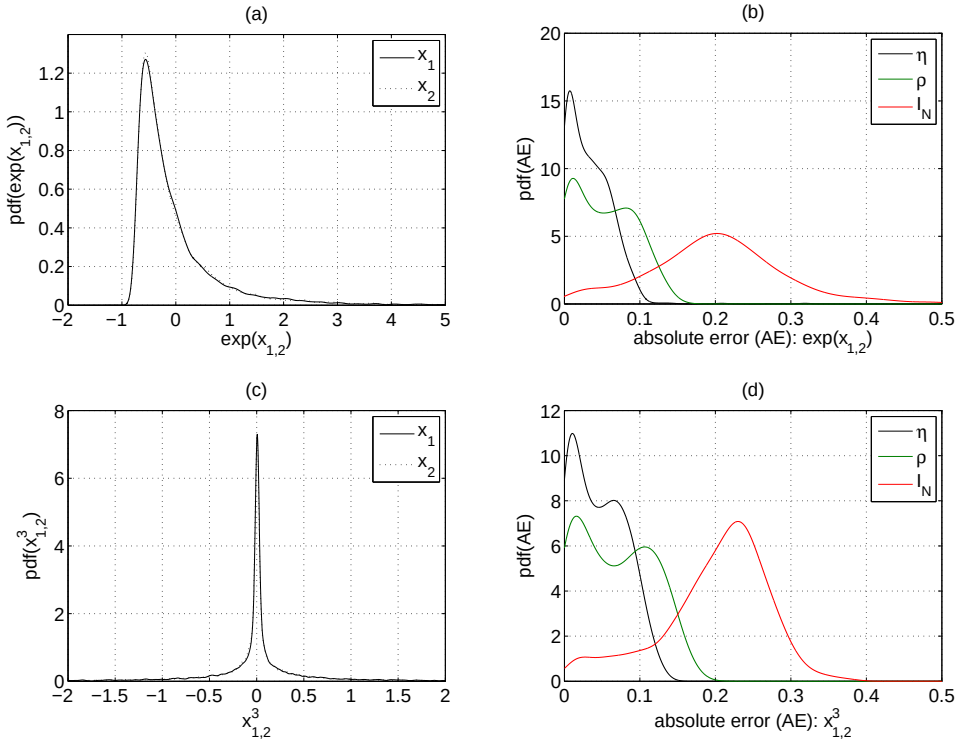

Fig. 4.8: (a,c): Normalised pdf plots showing distributions of a representative sample (obtained for $\alpha = 0.5$) of the data analysed in this example. Notice the non-Gaussian nature of all distributions. (b,d): Normalised distribution of absolute errors (AE) obtained using different measures of interaction. The accuracy of the information coupling (η) measure is evident, as it gives the lowest average AE values as compared to linear correlation (ρ) and normalised mutual information (I_N).

MAE	η	ρ	I_N
$\exp(\mathbf{x}_{1,2})$	0.0335	0.0522	0.1998
$\mathbf{x}^3_{1,2}$	0.0454	0.0667	0.1985

Table 4.4: Table showing accuracy of three measures of dependence, i.e. information coupling (η), linear correlation (ρ) and normalised mutual information (I_N), when used to estimate the level of dependence in a correlated system (as given by (4.7)) with varying values of α. The results show the mean absolute errors (MAE) obtained over 1000 different values of α ranging from 0 to 1, using 100 independent simulations at each value of α.

framework. The comparative study presented in this section shows that of the four measures of symmetric interaction considered, the only one which fulfils these requirements is information coupling. Other measures have certain limitations (as previously discussed) which make them less suitable for analysing financial data. Linear correlation, although computationally efficient, is not suitable for analysing non-Gaussian data, while both rank correlation and mutual information can be computationally complex and expensive to compute accurately.

Mutual information also requires large data sets for accurate estimation, which are often not available in practise.

Non-square source mixing

The synthetic data examples presented so far make use of bivariate data. However, the information coupling measure is also well-suited for measuring interactions in higher-dimensional spaces. When analysing data in high-dimensions, we need to estimate the optimum number of latent sources, M, given a set of N observed signals. Therefore, we now empirically demonstrate accuracy of the ICA model order estimation algorithm and the effect of non-square mixing ($M \neq N$) of ICA source signals on the information coupling metric. As an example, consider a synthetic non-Gaussian data set sampled from a Pearson type IV distribution with the same properties as those described in the previous section. Eight of these independent 1000 data points long data sets are used to mimic a set of source signals, giving us $\mathbf{S}$ (a 8×1000 dimensional matrix). We now generate a 20×8 dimensional normally distributed random mixing matrix $\mathbf{A}$. The 20×1000 dimensional set of observed signals can then be computed as $\mathbf{X} = \mathbf{AS}$. Non-Gaussian noise (at 10% amplitude), sampled from a separate independent Pearson type IV distribution, is also added to the observed signals. The log-likelihood of the observed data is calculated for different number of source signals, i.e. for $M = \{2 : 20\}$, and the average results over 100 independent simulations plotted in Figure 4.9(a) together with the standard deviation contours. As expected, the log-likelihood (on average) is maximum for $M = 8$, the number of predefined source signals. We repeat this analysis using 12 sources and present the results in Figure 4.9(d), once again showing accuracy of the model order estimation algorithm for estimating the correct number of ICA source signals.

We now extend this example to show the need for accurate estimation of the number of sources. Figures 4.9(b) and 4.9(e) show the variation of information coupling for different number of source signals for the same data sets which were used to obtain the log-likelihood plots in Figures 4.9(a) and 4.9(d) respectively. Average results for 100 independent simulations are plotted together with the standard deviation contours at each value of M. It is evident that information coupling varies significantly for different number of sources. Therefore, using the correct number of source signals for computing information coupling in higher-dimensional systems is very important. We also note that coupling increases gradually as M approaches

the correct number of sources, i.e. eight for Figure 4.9(b) and twelve for Figure 4.9(e), before experiencing a sudden jump in magnitude. Figures 4.9(c) and 4.9(f) show the difference in information coupling ($\Delta\eta_M$) for serially increasing values of the number of source signals (M). $\Delta\eta_M$ clearly peaks at the correct model order in both cases; we have observed this general property of the metric for different number of actual source signals and it can potentially be used in itself for model order estimation purposes, as $\Delta\eta_M$ may have a clear peak even when the log-likelihood plot does not. These results indicate that when the estimated number of source signals is less than (or equal to) the optimum number, information coupling gradually increases as information contained in more sources is included in the analysis; however, as soon as the estimated number of sources becomes greater than the optimum number, information coupling experiences a sudden jump in magnitude due to inclusion of information contained in "redundant sources".

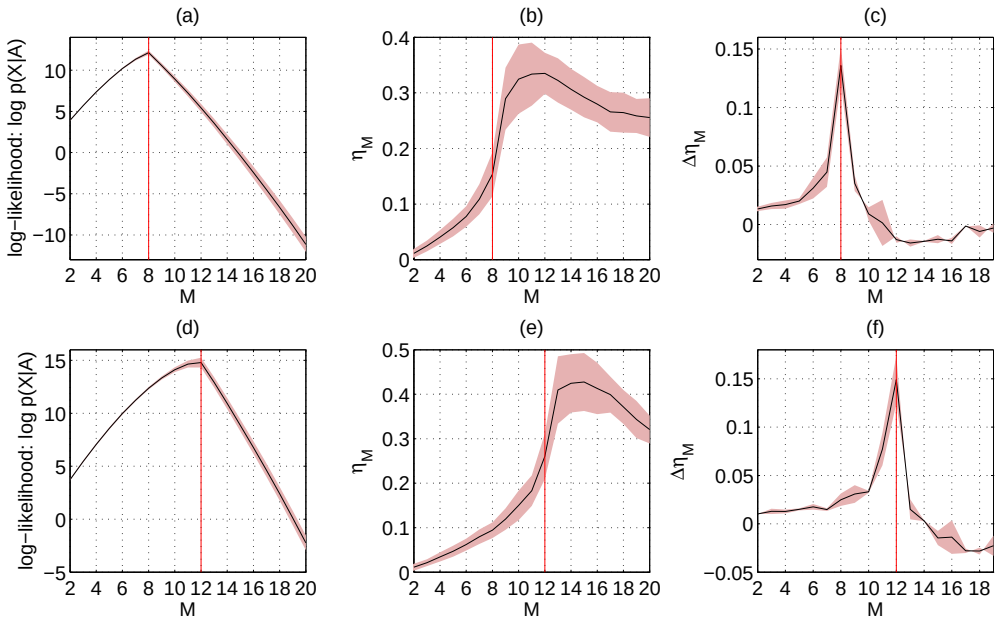

Fig. 4.9: (a,d): ICA log-likelihood (ℓ) plotted as a function of the number of source signals (M) for the purpose of model order estimation. The true number of source signals for the two data sets are 8 and 12 respectively. As expected, the log-likelihood is maximum at $M = 8$ and $M = 12$ respectively for the two examples. The standard deviation contours for 100 independent simulations are also plotted. (b,e): Effect of number of ICA sources on information coupling (η_M). The large variation of η_M with M shows the need for accurate model order estimation. Once again, the contours reflect the standard deviation of η_M for 100 independent simulations. (c,f): Difference in information coupling ($\Delta\eta_M$) for serially increasing values of the number of source signals (M), e.g. $\Delta\eta_M$ at $M = 8$ implies the difference in η_M as M goes from 8 to 9, i.e. $\Delta\eta_8 = \eta_9 - \eta_8$.

Scale-dependent information coupling

So far we have considered cases of scale-independent variations of information coupling. We now proceed to extend our analysis to the scale-dependent case by presenting some time-scale analysis results obtained using synthetic data. We first present a simple example of the continuous wavelet transform (CWT), which exhibits its ability to analyse data at different frequencies. The example presented in Figure 4.10 shows use of the CWT, with a Morlet basis function, for the time-scale analysis of two combined sinusoidal signals. The two sinusoidal signals, $x_1(t) = \cos(t)$ and $x_2(t) = \cos\left(\frac{t}{5}\right)$, are mixed together and analysed using the CWT at scales of $u = 1$ to $u = 400$. The data is sampled at 10 samples per second, i.e. $\Delta = \frac{1}{10}$. The pseudo-frequencies, in Hz, for the two mixed signals are $f_{x_1} = \frac{1}{2\pi}$ and $f_{x_2} = \frac{1}{10\pi}$ respectively, with a centre frequency of $f_o = 0.8125$ Hz. As given by (3.56), we may write the scale as $u = \frac{f_o}{f_u \Delta}$. Therefore, for the two mixed signals, the scales should be $u_{x_1} = 51.05$ and $u_{x_2} = 255.25$ respectively. The scalogram in Figure 4.10 and the corresponding plot in Figure 4.11 clearly show two distinct regions of high intensity at scales of 51.05 and 255.25, thus validating accuracy of the CWT model for the time-scale analysis of this particular set of mixed sinusoidal signals.

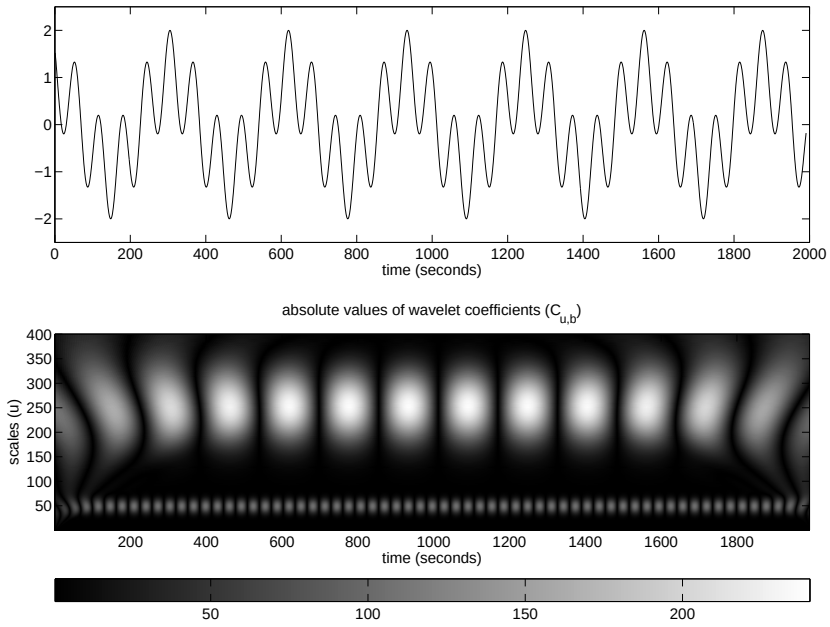

Fig. 4.10: Top: Two combined sinusoids of periods p_1 and p_2, where $p_2 = \frac{p_1}{5}$. Bottom: Scalogram of the above signal, with high intensity regions at scales of 51.05 and 255.25 visible.

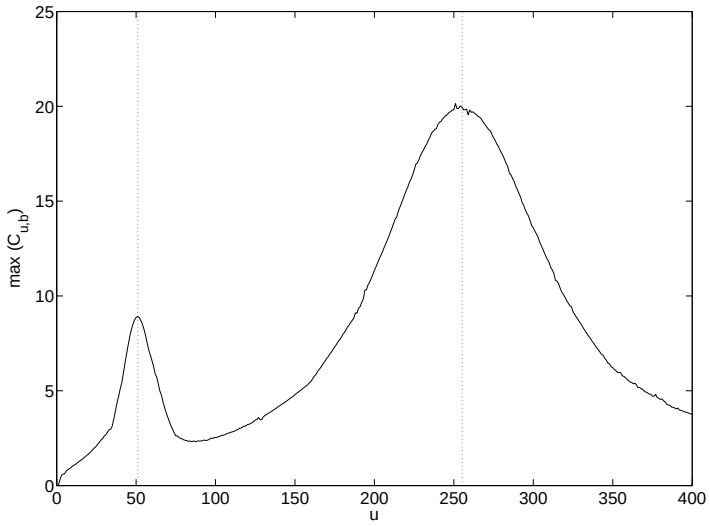

Fig. 4.11: Plot showing maximum temporal values of the CWT coefficients plotted with respect to scale (u). As expected, the largest values of coefficients occur at scales of 51.05 and 255.25.

It is important to know the general level of information coupling between random non-Gaussian data sets at different time-scales in order to better judge the significance of scale dependent information coupling in financial systems (which we look at later in this chapter). The two plots in Figure 4.12 show information coupling across scale (u) and time lags (τ) between two non-Gaussian random variables, sampled from the same bivariate Pearson type IV distribution as previously used in this section. All results are obtained as an average over 100 independent simulations. We note the relatively low coupling magnitudes (generally less than 0.08) for all plots, which is expected for randomly sampled data. We can use these results as a benchmark for the significance of scale-dependent information coupling when analysing financial data later in this chapter.

Capturing discrete state-based coupling dynamics

We now present a set of synthetic data examples which demonstrate use of the HMICA model for identifying regimes of low and high coupling in multivariate data sets [317]. This is made possible because of the ability of the HMICA model to accurately compute the independent components (by estimating the state-based unmixing matrices) from a set of observed signals with discrete changes in mixing dynamics, i.e. from "partitioned" data. To demonstrate this, we compare accuracy of HMICA to a standard ICA algorithm trained on each separate parti-

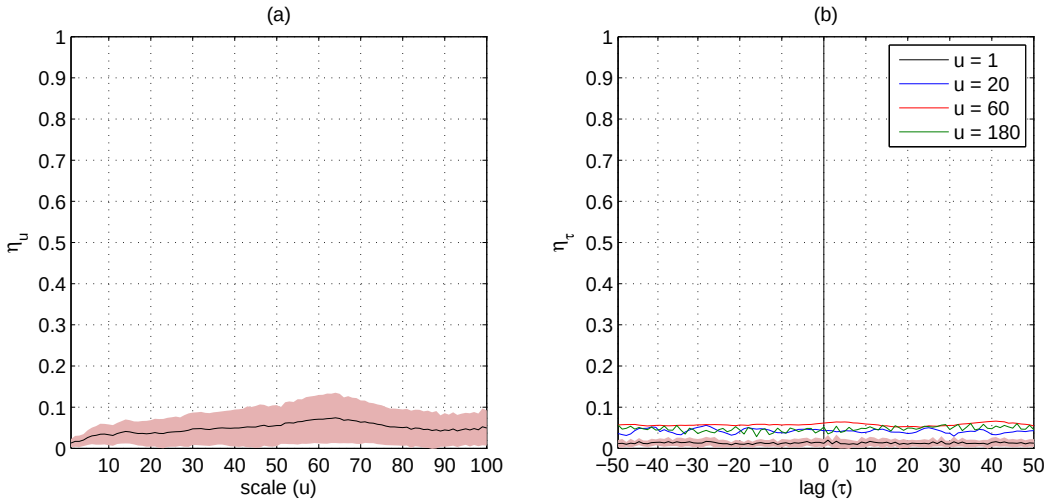

Fig. 4.12: (a). Average information coupling (η_u) between bivariate non-Gaussian random time series (sampled from a Pearson type IV distribution) at different time scales (u). The contour represents standard deviation values for 100 independent simulations. (b). Information coupling (η_τ) at different time lags (τ) for data analysed at scales of 1, 20, 60 and 180, obtained as an average of 100 independent simulations. The standard deviation contour for only scale of 1 is plotted for clarity; standard deviation values at other scales are of a comparable magnitude. We note the low coupling values across scale and time lags in both plots, as expected for random data. We use these plots as a benchmark for judging the significance of results obtained when measuring the scale-dependence of information coupling in financial systems later in this chapter.

tion. A standard ICA model trained using the entire signal will fail to accurately extract the independent components, as each partition has a different mixing matrix associated with it. For this example, we use two 800 data points long non-Gaussian signals (sampled from a Pearson type IV distribution) obtained using mixing matrices $\mathbf{A}_1$ in the interval $t = \{1:400\}$ and $\mathbf{A}_2$ in the interval $t = \{401:800\}$, where:

$$\mathbf{A}_1 = \begin{bmatrix} 0.8352 & 0.5500 \\ 0.5268 & 0.8500 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 0.8000 & -0.6000 \\ -0.4500 & 0.8930 \end{bmatrix} \tag{4.8}$$

We now estimate direction of the independent components (obtained from the basis vectors) using a standard ICA algorithm trained separately on each partition as well as by using a 2-state HMICA model on the entire length of the signals; the results obtained are presented in Figure 4.13. We note high accuracy of the HMICA model in directly extracting the independent components from the mixed signals, by accurately estimating the state-based unmixing matrices.

The example presented above demonstrates utility of the HMICA model to accurately extract latent states from multivariate data sets. As we observed, these states correspond to

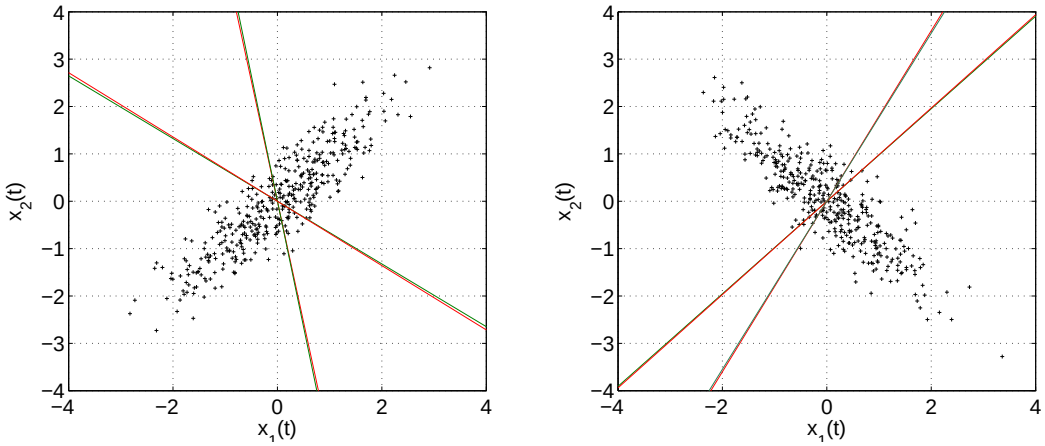

Fig. 4.13: Scatter plots of observed signals $x_1(t)$ and $x_2(t)$ for the interval $t = \{1:400\}$ (left) and $t = \{401:800\}$ (right). Also shown are the directions of the two independent components obtained using individual ICA models (green lines) trained on each partition of the data, and the directions estimated by the HMICA model (red lines). The plots verify accuracy of the HMICA model for estimating independent components in data sets with discrete changes in mixing dynamics.

different mixing dynamics. We now present a synthetic data example showing ability of the HMICA model to capture changes in information coupling dynamics. Figures 4.14(a) and 4.14(b) show two time series, $x_1(t)$ and $x_2(t)$, which are generated using mixing matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ in successive 400 data points wide intervals (as elaborated in the caption of the figures), where:

$$\mathbf{A}_1 = \begin{bmatrix} 0.5200 & 0.4100 \\ -0.1200 & 0.9100 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 0.4300 & -0.6000 \\ -0.1900 & 0.7400 \end{bmatrix} \tag{4.9}$$

Figure 4.14(c) shows temporal variation of information coupling values (obtained using a 200 data points wide sliding-window) together with the Viterbi state sequence (the most likely sequence of states) obtained using the HMICA model. Also plotted are the 95% confidence interval contours on the coupling measure. It is clear that (in this example) the hidden states are indicative of discrete (abrupt) state-based changes in information coupling; hence, it may be possible to use the HMICA Viterbi state sequence as a binary indicator of the coupling magnitude, with each latent state corresponding to regions of either low or high information coupling.

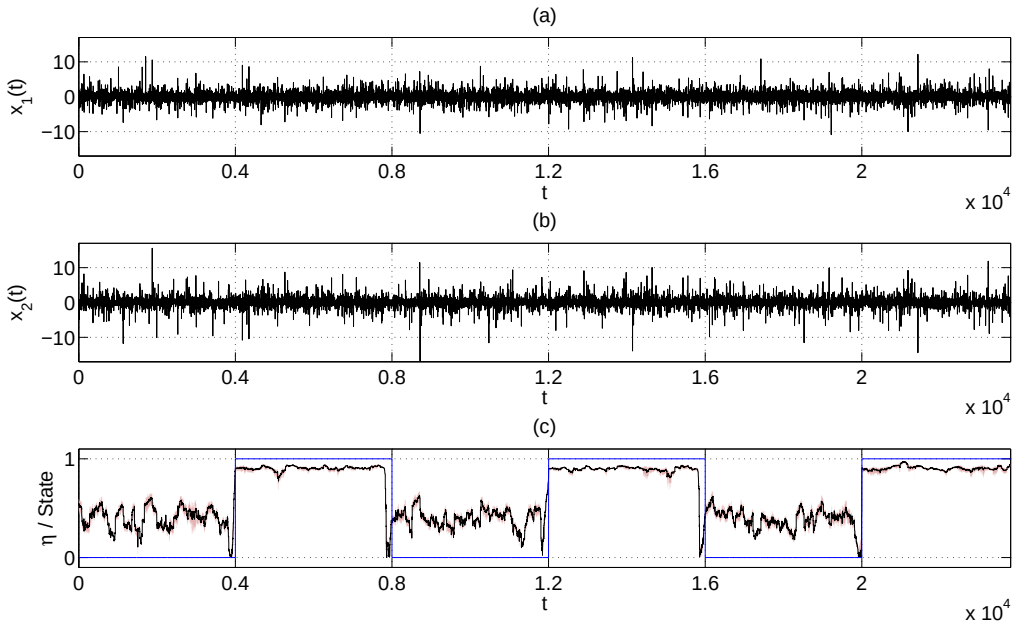

Fig. 4.14: Plots (a) and (b) show time series $x_1(t)$ and $x_2(t)$ respectively, which are generated using mixing matrix $\mathbf{A}_1$ in the interval $t = \{1:4000, 8001:12000, 16001:20000\}$, and mixing matrix $\mathbf{A}_2$ in the interval $t = \{4001:8000, 12001:16000, 20001:24000\}$. Plot (c) shows the information coupling (black) between $x_1(t)$ and $x_2(t)$, together with the Viterbi state sequence obtained using the HMICA model (blue). It is clear from plot (c) that the HMICA states are indicative of discrete state-based changes in information coupling dynamics.

4.4 Analysis of financial data

We now proceed to demonstrate the accuracy and practical utility of the ICA-based information coupling model (and its extensions) for analysing multivariate financial returns. For most examples presented in this section, results for various other symmetric interaction measures are also presented for comparative purposes. We start this section by presenting a set of general results obtained using the information coupling model and its extensions. These results lead us to the main part of this section, which presents a set of financial case studies using which we demonstrate the practical utility, efficiency and accuracy of using the information coupling model for extracting interesting and useful information from financial data streams.

Let us now present a set of examples showing some general results relating to bivariate as well as multivariate analysis of information coupling in financial systems. We refer back to these results later in this chapter when presenting various practical financial case studies. As previously discussed, financial returns have different properties at different sampling frequen-

cies. For most of the general results presented here, we make use of the 0.5 second and 0.5 hour sampled data sets as representative examples of FX data sampled at high and medium frequencies respectively. We first present a set of simple examples of analysis of bivariate information coupling between different FX currency pairs; these examples give us a flavour for some of the properties of interactions (both static and dynamic) present in FX markets. Figures 4.15(a) and 4.15(b) show information coupling between 10 liquid spot FX currency pairs, sampled every 0.5 second and 0.5 hour respectively. The results are obtained as an average of 50 data points long samples over different parts of the data sets representing an 8-hour period for the 0.5 second sampled data and 2 years for the 0.5 hour sampled data. From the figures it is evident that information coupling between currency pairs containing the U.S. dollar is generally much higher than non-USD pairs. This is observable at different frequencies and points to dominance of the USD in the global FX market (we discuss this further later in this chapter). We also notice the relatively high coupling values for the data sampled every 0.5 hour as compared to the 0.5 second sampled data; this points to the scale dependence of information coupling (which we also look into in detail later in this chapter).

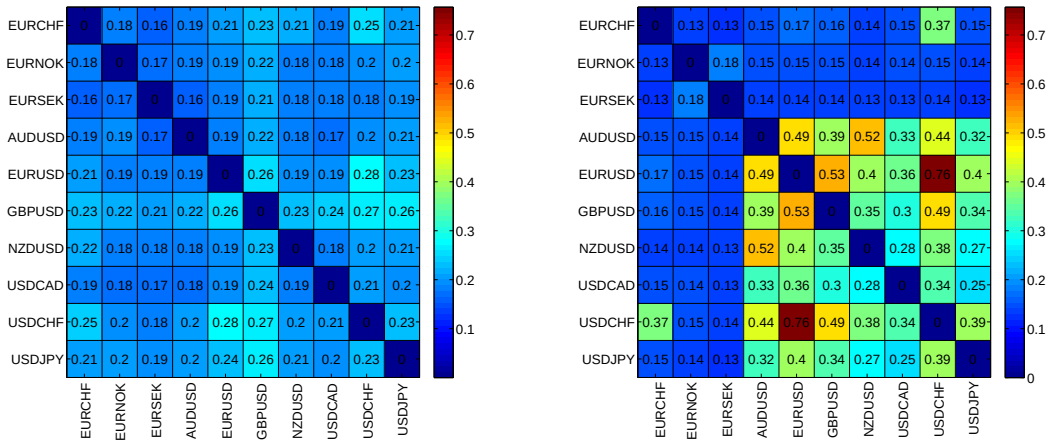

Fig. 4.15: Average information coupling (η) between log-returns of 10 liquid spot FX currency pairs, for: (Left). 0.5 second sampled data over a period of 8 hours, (Right). 0.5 hour sampled data over a period of 2 years. η between the same currency pairs is set to zero for ease of visualisation. Results are obtained as an average of 50 data points long samples over different parts of the data set. We notice the relatively high coupling between USD containing currency pairs.

We now present a simple example of application of the information coupling algorithm to a section of 0.5 second sampled FX spot log-returns data set in order to "observe" the general dynamics of coupling in bivariate financial time series. Figure 4.16 shows the variation

of information coupling and linear correlation with time for EURUSD and GBPUSD. The results are obtained using a 5 minute wide sliding-window. We note that the two measures of dependence frequently give different results, which reflects on the inability of linear correlation to capture dependencies in non-Gaussian data streams. We also note that dependencies in FX log-returns exhibit rapidly changing dynamics, often characterised by regions of quasi-stability punctuated by abrupt changes; these regions of persistence in statistical dependence in financial time series may be captured using a HMICA model (as we demonstrate later).

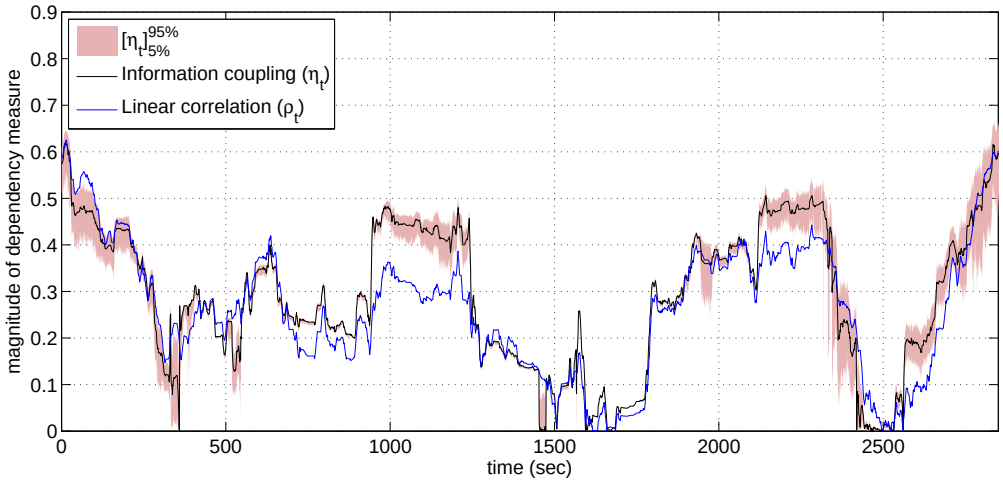

Fig. 4.16: Information coupling (η_t) and linear correlation (ρ_t) plotted as a function of time for a section of 0.5 second sampled EURUSD and GBPUSD log-returns data set. A 5 minute wide sliding-window is used to obtain the results.

As another comparative example, Figure 4.17 displays plots showing a representative example of the temporal variation of linear correlation, rank correlation, mutual information and information coupling between two liquid currency pairs, i.e. EURUSD and USDJPY sampled every 0.5 seconds; the results are obtained using a 100 data points long sliding-window. The plots for linear correlation, rank correlation and information coupling are somewhat similar for significant periods of time. However, it is interesting to note that there are regions where the information coupling plot significantly deviates from the other plots. We believe the dynamically changing level of non-Gaussianity of the data (an example of which was presented in Figure 4.3) most likely explains this deviation of the information coupling measure at certain time periods, a point which is reinforced by results we present later in this chapter. This is because unlike the other interaction measurement approaches considered in this example, information coupling gives reliable results when dynamically analysing non-Gaussian data

streams (as previously discussed and empirically demonstrated). Hence, we can be fairly certain that any changes in the information coupling measure are indicative of genuine changes in symmetric interactions rather than being an artefact of some properties of the data (as might be the case when using some other interaction measurement approaches). As mutual information requires large data sets for accurate estimation (and has normalisation issues), therefore it gives relatively inaccurate results in dynamic environments, as shown by this simple example.

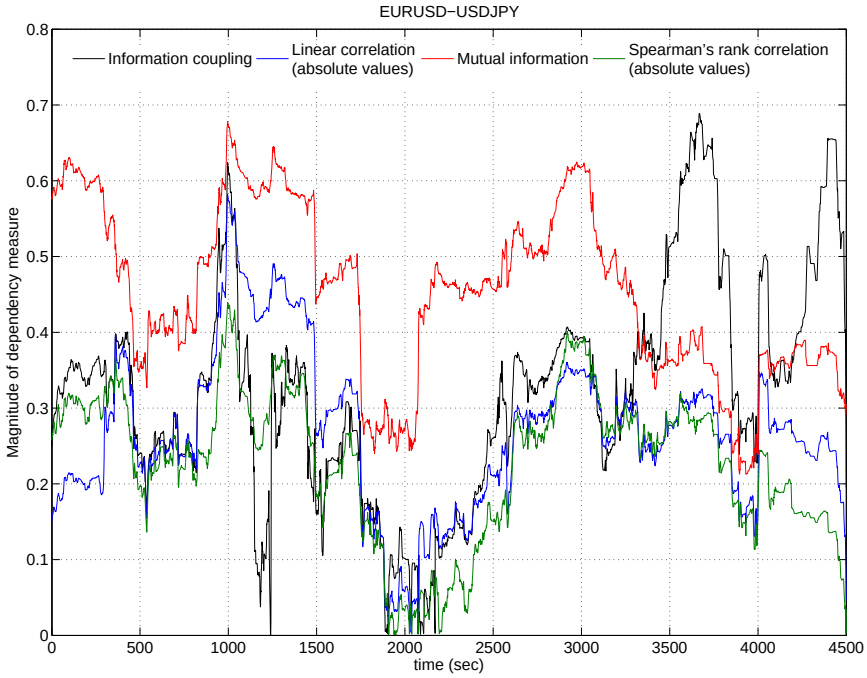

Fig. 4.17: A snap-shot of data showing four different measures of symmetric interaction used to dynamically measure temporal interactions between 0.5 second sampled EURUSD and USDJPY spot log-returns. Results are obtained using a 100 data points long sliding-window. The information coupling (η) measure significantly deviates from linear and rank correlation measures at certain time periods, most likely due to the dynamically changing level of non-Gaussianity of the underlying data. See text for details.

Earlier we discussed the possibility of using information about the lead-lag variations in coupling to measure *predictive information coupling*, which can give us an insight into the presence and direction of asymmetric interactions and can be potentially useful for developing improved prediction models. We now consider a few examples to ascertain whether *predictive information coupling* exists in the FX markets. Figure 4.18 shows the variation of information coupling (η_τ) with time lag (τ) between four combinations of 0.5 second sampled currency

pairs. The plots are obtained using buckets of 15 minutes of data, with the solid lines representing the average coupling-lag values while the contours represent the standard deviation estimates obtained over an 8-hour trading session (obtained using the 32 sections of data analysed). Plots (a) and (b) show results for the EURUSD-USDCHF and EURUSD-USDJPY currency pairs respectively. Both plots exhibit significantly high coupling values at non-zero time lags, pointing to the presence of *predictive information coupling* between these currency pairs. We notice that coupling gradually decays with time lag and becomes negligible at a lag of approximately 8 seconds for both plots. These results show that for the data sets analysed, EURUSD seems to have some causal effect on both USDCHF and USDJPY, i.e. EURUSD $\rightarrow$ USDCHF and EURUSD $\rightarrow$ USDJPY. This possibly indicates that EURUSD is "in play" during this trading session and (being the most liquid currency pair) incorporates any new market information about USD before the other two currency pairs. This information can be useful for predictive purposes and later in this chapter we present a practical financial case study showing the use of information thus obtained for developing an improved forecasting model for exchange rates. Similarly, plots (c) and (d) show results for the EURUSD-EURCHF and EURUSD-EURGBP currency pairs respectively. Both these plots show no sign of significant coupling at any time lags. This is possibly because both EURCHF and EURGBP are non-USD currency pairs and hence do not exhibit a significant causal link with EURUSD during this trading session (indicating that EURUSD dynamics are being driven by USD instead of EUR).

The examples we have presented so far made use of bivariate data. When analysing multivariate data, an important step in computing information coupling is estimating the optimum number of ICA source signals (i.e. we need to infer the ICA model order). Due to the rapidly changing dynamics of multivariate FX spot returns, it is very likely that the number of optimum ICA source signals also changes with time; here we present a set of examples to study this effect. Figure 4.19(a) shows the normalised distribution of the number of estimated ICA source signals (M) for nine liquid currency pairs, covering all G10 currencies; results are obtained for 0.5 second sampled data over an eight hour period using a 50 data points long sliding-window. Likewise, Figure 4.19(c) shows results for the 0.5 hour sampled data over a 2 year period, again using a 50 data points long sliding-window. The results obtained show that the mixing

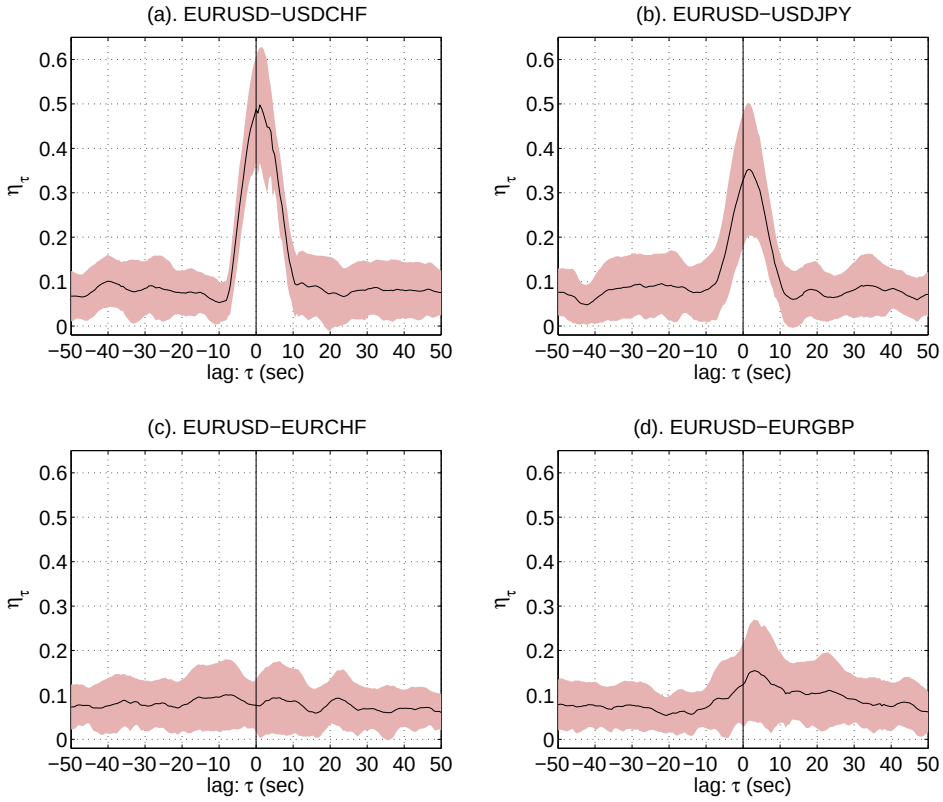

Fig. 4.18: Coupling-lag plots showing the variation of information coupling (η_τ) with time lag (τ) between four combinations of 0.5 second sampled currency pairs. The plots are obtained using buckets of 15 minutes of data, with the solid lines representing the average coupling-lag values while the contours represent the standard deviation estimates obtained over an 8-hour trading session.

process is usually undercomplete, i.e. the optimum number of estimated sources is often less than the number of observed signals (the pdf(M) plot has a clear peak at $M < N$). Therefore, in practise information coupling can be estimated without the need for considering the computationally complex overcomplete mixing case. We also notice that (on average) M is lower for the high-frequency 0.5 second sampled data as compared to the 0.5 hour sampled data. Figures 4.19(b) and 4.19(d) show variation of the average information coupling (η_M) with the optimal number of source signals (M); also plotted are the standard deviation contours. These plots show that more closely coupled multivariate data streams generally have more latent source signals which are giving rise to the data; we also observed this property earlier when analysing synthetic data. The optimal value of M rapidly changes with time for both data sets. For the 0.5 second sampled data, the probability of change in the value of M at any given time-step is 0.63, while for the 0.5 hour sampled data it is 0.56. These results signify the importance

of correctly estimating the number of ICA sources while analysing information coupling in high-dimensional dynamic systems. The ICA log-likelihood based model order estimation approach we presented earlier is well-suited for this purpose.

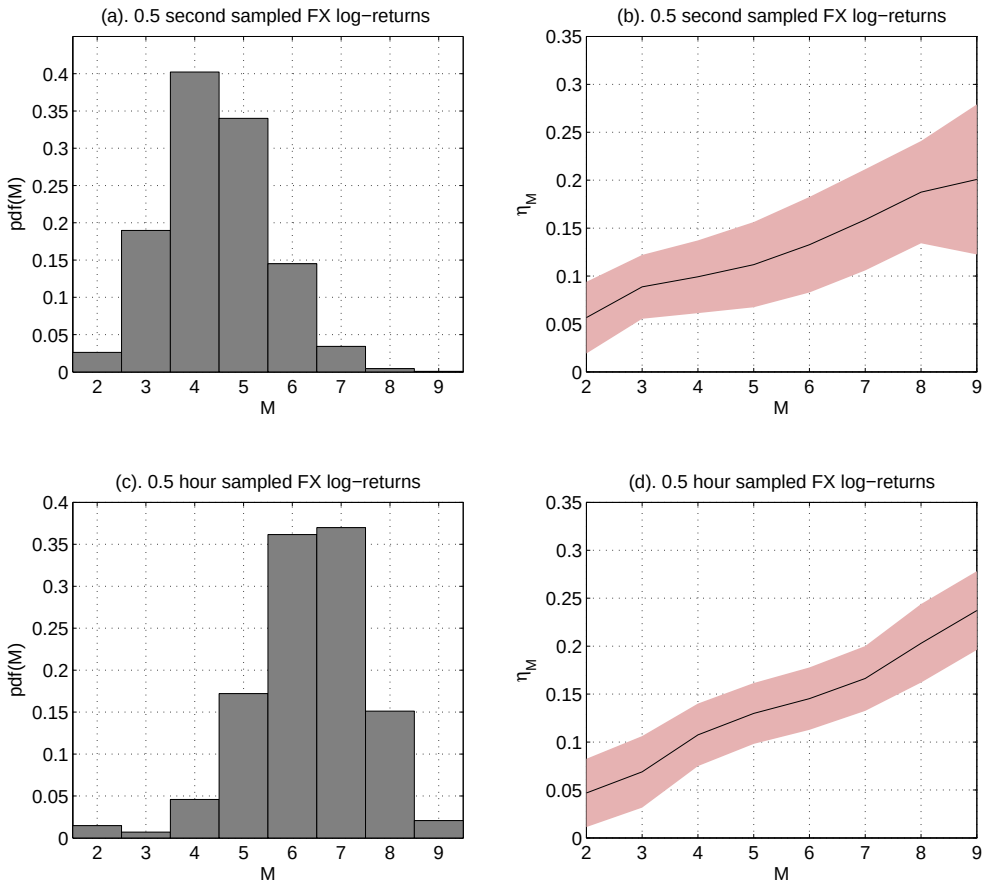

Fig. 4.19: (a,c): Normalised distributions of the number of ICA source signals (M) for nine liquid currency pairs for 0.5 second and 0.5 hour sampled data respectively. Plot (a) was obtained using an 8-hour long data set while a 2 year long data set was used to obtain plot (c). In both cases, a 50 data points wide sliding-window was used. (b,d): Variation of mean information coupling (η_M) between the nine currency pairs for the optimal number of ICA source signals (M), obtained using: (b) 8 hours of 0.5 second sampled log-returns, (d) 2 years of 0.5 hours sampled log-returns. The contours show the standard deviations of η_M at each value of M.

So far in this section we have focused on direct analysis of information coupling in financial systems. We now consider some examples which make use of the information coupling measure to extract the hierarchical coupling structure in high-dimensional financial networks. We come back to results presented here later in this chapter when presenting some financial case studies. Figure 4.20 shows an example of a network obtained using the coupling-MST

approach for 12 currency pairs, each sampled at 2 samples per second and covering an 8-hour period. Each node, or vertex, represents a currency pair and each link, or edge, represents a pseudo-distance, which is calculated using the distance metric given by (3.44) and is dependent on the ICA-based information coupling between any two currency pairs. From this currency network it is evident that there are two distinct groups; one is centred around EURUSD, and the other one around USDCHF. This shows that for the 8 hour period for which coupling is calculated, these two currency pairs are dominant, i.e. they are "in play". Later on we present financial case studies which make use of this stable, "pivotal", role of USDCHF for developing robust exchange rate forecasting models.

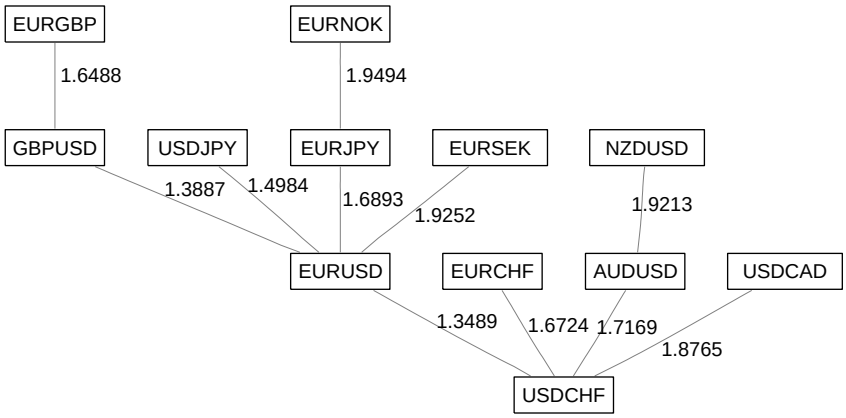

Fig. 4.20: MST showing the hierarchical interaction structure in a network of 12 FX currency pairs, obtained using 8 hours of 0.5 second sampled data. The 12 vertices (nodes) represent the currency pairs, while the 11 edges (links) represent the pseudo-distance (d_{ij}) between the vertices, calculated using the distance measure given by (3.44).

We now consider a set of examples of FX coupling-MSTs in even higher dimensional spaces. Figures 4.21 and 4.22 present two static MSTs for 45 currency pairs, which are obtained using 0.5 second (over an 8-hour trading session) and 0.5 hour (over a 6 month period) sampled data respectively. As different permutations of a currency pair result in the same value of information coupling, the nodes of the MSTs present permutation-independent currency pair names. It is interesting to note that currency pairs containing a common currency are grouped together, with each group being coloured differently for ease of identification. Knowledge about currency pairs at the centre of these groups can be useful for numerous practical

financial applications, some of which we present later in this chapter. Looking at the three MSTs presented in Figures 4.20, 4.21 and 4.22, it quickly becomes clear that at the centre of the majority of major (as well as minor) groups is a currency pair containing the USD as one of the currencies. This shows that the USD is driving other currency pairs, which is not surprising given the dominance of USD in global FX trading, with currency pairs containing the USD accounting for over 86% of the global FX transaction volume [143]. Status of the USD as the premium reserve currency [296], and its widespread use in global trade and financing [144], are some of the other primary reasons for the pivotal role the USD plays in the global currency markets.

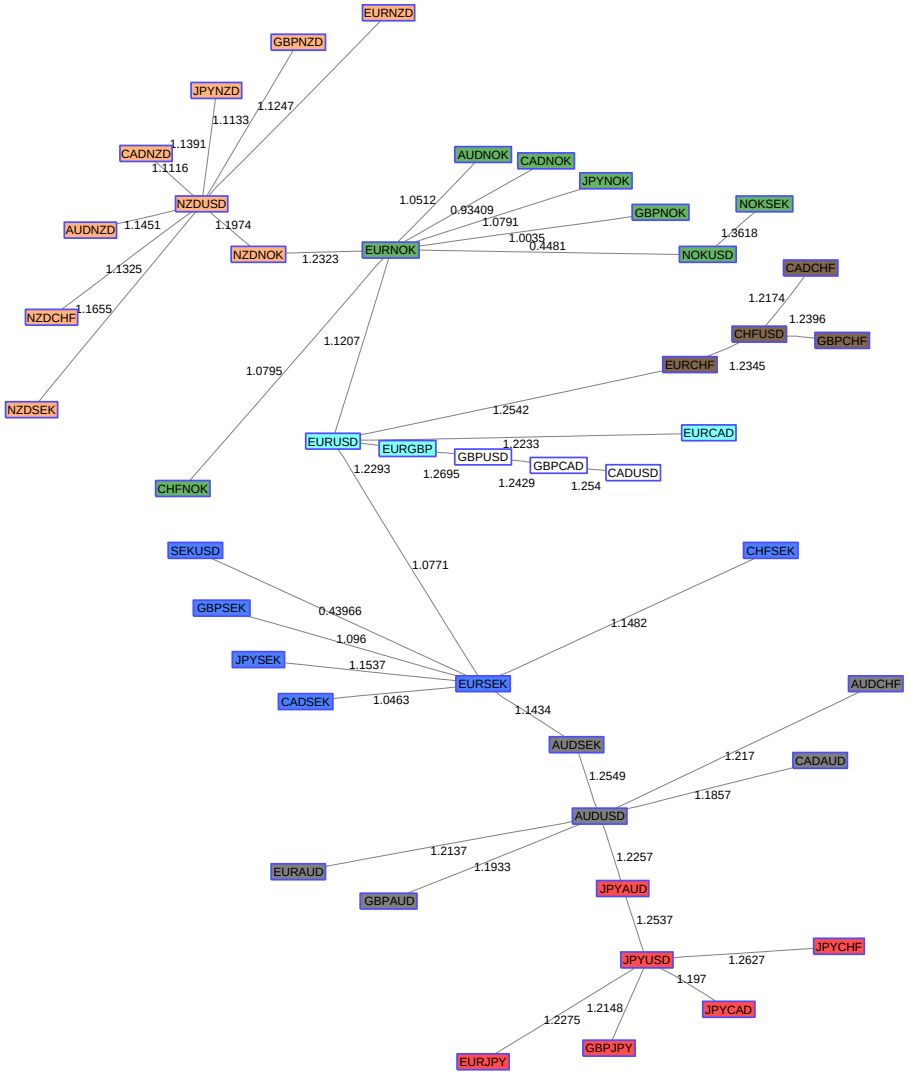

Fig. 4.21: MST showing the structure of dependencies between 45 currency pairs. The data is sampled at 2 samples per second and covers a period of eight hours. Groups of currency pairs which contain a common currency are represented by the same colour.

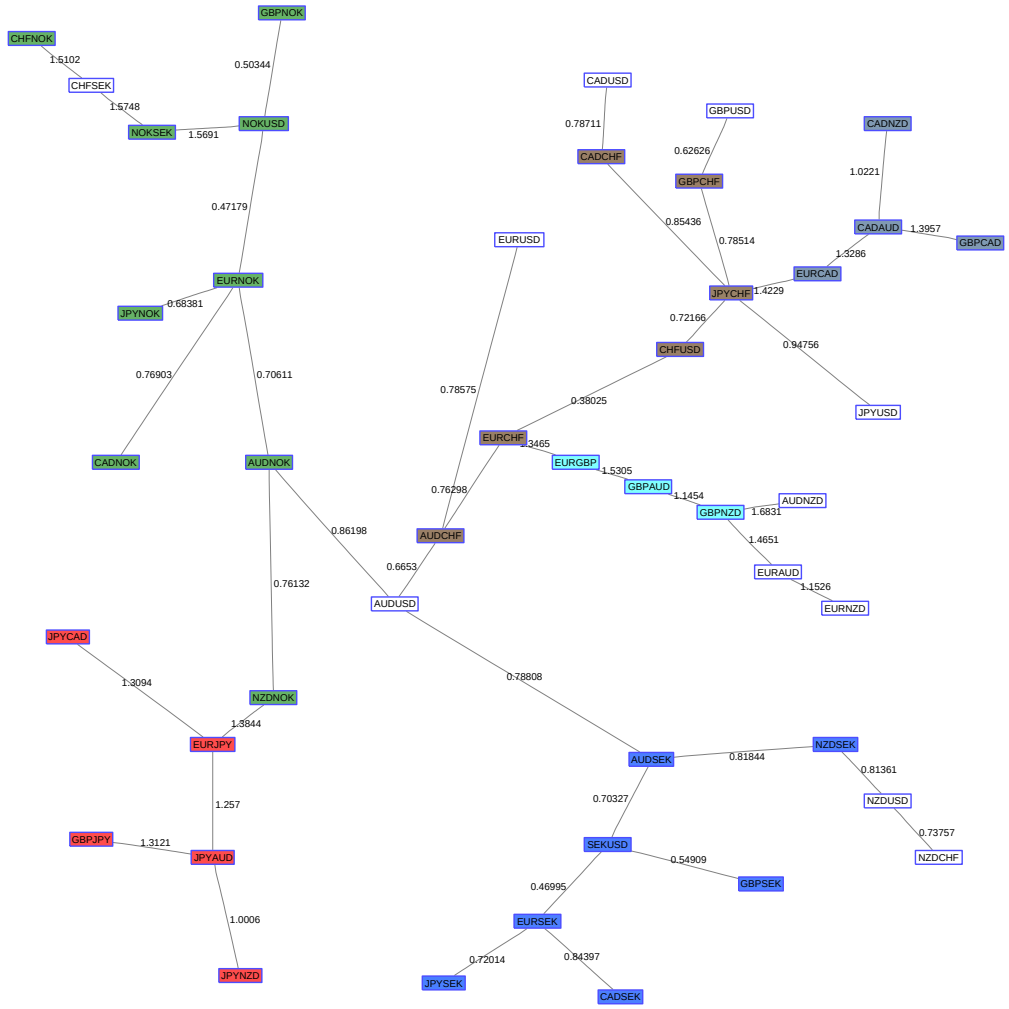

Fig. 4.22: MST showing the structure of dependencies between 45 currency pairs. The data is sampled at 2 samples per hour and covers a period of six months. Groups of currency pairs which contain a common currency are represented by the same colour.

So far in this chapter we have considered a range of examples which demonstrate some general properties of coupling in the FX markets. We now proceed to present a set of financial case studies which (in part) draw on the results presented so far to demonstrate the practical utility, accuracy and efficiency of the information coupling model (and its extensions) for extracting useful information from multivariate financial time series. For most of these case studies, we carry out a detailed comparative analysis of results obtained using the information coupling measure with other standard measures of symmetric interaction.

4.4.1 Case study 1: Studying interactions in FX markets during the 2008 financial crisis

There have been numerous academic studies on the causes and effects of the 2008 financial crisis [97, 259]. However, very few of these have focused on the nature of inter-dependencies in the global spot FX market during the crisis; here we present a set of examples which give us a unique insight into this area. Accurate estimation of dependencies at times of financial crises is of utmost importance, as these estimates are used by financial practitioners for a range of tasks, such as rebalancing portfolios, accurately pricing options, deciding on the level of risk-taking, etc. We first present an application of the information coupling model for detecting temporal changes in dependencies in bivariate FX data streams at times of financial crises. Figure 4.23 shows the daily closing mid-prices (P_t) for AUDUSD and USDJPY from January 2005 till April 2010 (the two plots are scaled for ease of comparison). The plots clearly show an abrupt change in the exchange rates in September-October 2008. This was caused at the height of the 2008 global financial crisis due to the unwinding of carry trades [293]. Figure 4.24(a) displays three plots showing the temporal variation of information coupling (η_t), linear correlation (ρ_t) and rank correlation ($\rho_{R,t}$) between AUDUSD and USDJPY log-returns. The plots are obtained using a six month long sliding-window. We notice the rise in uncertainty of the information coupling measure (Figure 4.24(b)) right before the crash, with uncertainty decreasing gradually thereafter; this information may be useful to systematically predict upheavals in the market, although we do not carry out this study in detail here. Information about the level of uncertainty can be used as a measure of confidence in the information coupling values and can be useful in various practical decision making scenarios, such as deciding on the capital to deploy for the purpose of trading or selecting stocks (or currencies) for inclusion in a portfolio. As daily sampled data is generally less non-Gaussian than data sampled at higher frequencies, therefore, the three plots in Figure 4.24(a) are somewhat similar during certain time periods. However, right after the September 2008 crash, the plots significantly deviate from each other. We believe that this is because the nature of the data, in particular its level of non-Gaussianity, has changed. As shown in Figure 4.25, the distance measure, $(\eta_t - |\rho_t|)^2$, between information coupling and linear correlation closely matches the non-Gaussianity of the data under consideration (the two plots are scaled for ease of comparison). The degree

of non-Gaussianity is calculated using the multivariate Jarque-Bera statistic (JB_{MV}) which we define for a N-dimensional multivariate data set as:

$$JB_{MV} = \sum_{j=1}^{N} \left[\frac{n_s}{6} \left(\gamma_j^2 + \frac{(\kappa_j - 3)^2}{4} \right) \right]^2 \tag{4.10}$$

where n_s is the number of data points (in this case the size of the sliding-window), γ is the skewness of the data under analysis and κ is its kurtosis. This shows that relying solely on correlation measures to model dependencies in multivariate financial time series, even when using data sampled at relatively lower frequencies, can potentially lead to inaccurate results. In contrast, the information coupling model takes into account properties of the data being analysed, resulting in an accurate approach to measure statistical dependencies.

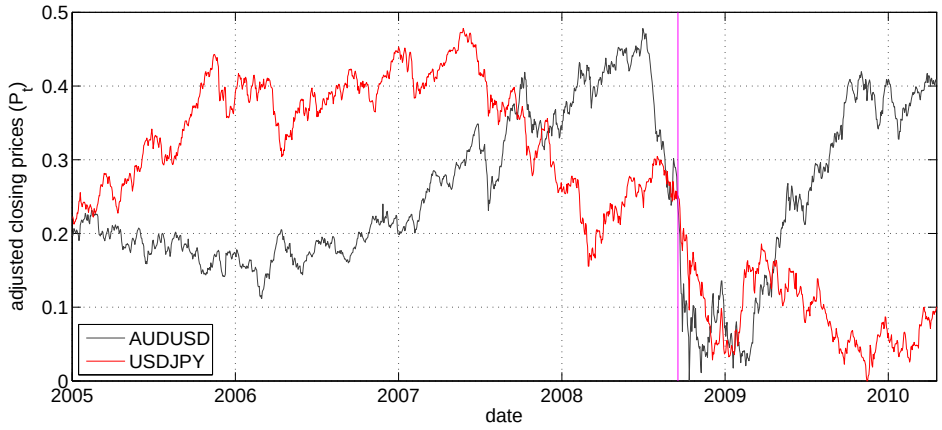

Fig. 4.23: Daily closing mid-prices (P_t) of AUDUSD and USDJPY. The two plots are scaled (such that they vary over the same range) for ease of comparison. The vertical line corresponds to the September 2008 financial meltdown.

We now show utility of the information coupling model for analysing multivariate statistical dependencies. Figure 4.26 shows the temporal variation of information coupling between four major liquid currency pairs (EURUSD, GBPUSD, USDCHF and USDJPY). The results are obtained using daily log-returns for a seven year period and a six month long sliding-window. Also plotted on the same figure is the FTSE-100 (Financial Times Stock Exchange 100) index for the corresponding time period (which has been scaled for ease of comparison). The plot clearly shows an abrupt upward shift in coupling between the four currency pairs right at the time of the September 2008 financial meltdown, with gradual decrease in coupling over the next year. We again notice an increase in uncertainty associated with the information cou-

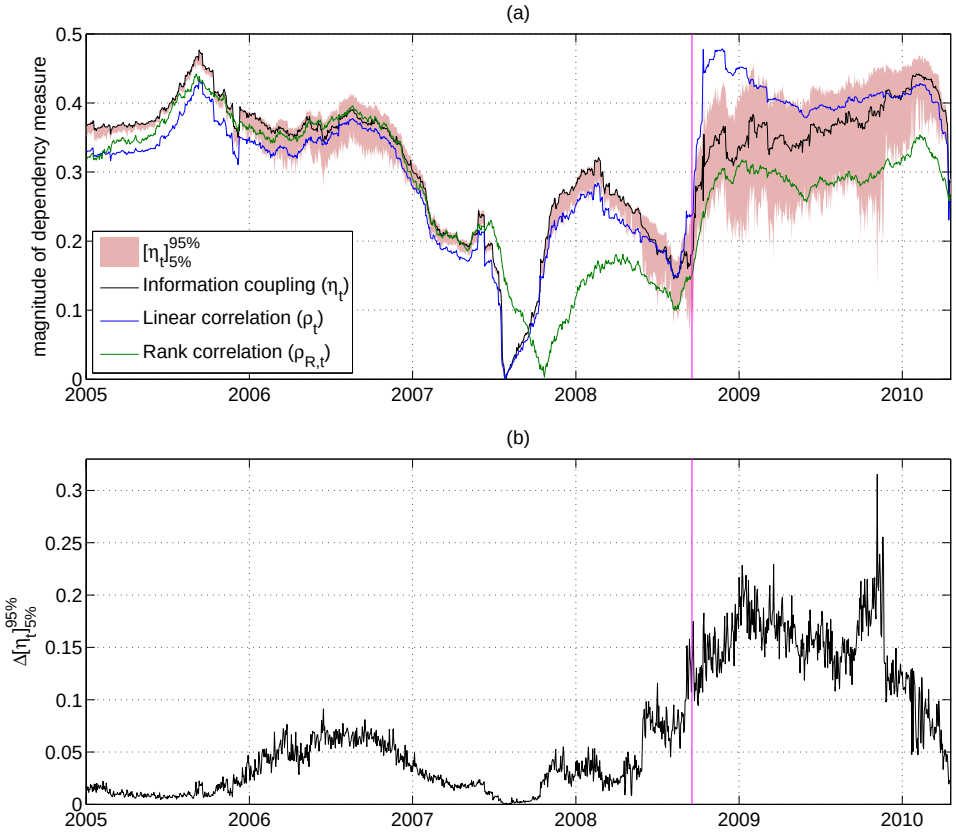

Fig. 4.24: (a). Three approaches used to measure temporal dependencies between AUDUSD and USDJPY log-returns; also plotted are the 95% confidence interval contours on the coupling measure. (b). Magnitude (range) of confidence intervals $\Delta[\eta]^{95\%}_{5\%}$ plotted as a function of time, showing the temporal variation of uncertainty associated with the information coupling measurements. The vertical lines correspond to the September 2008 financial meltdown.

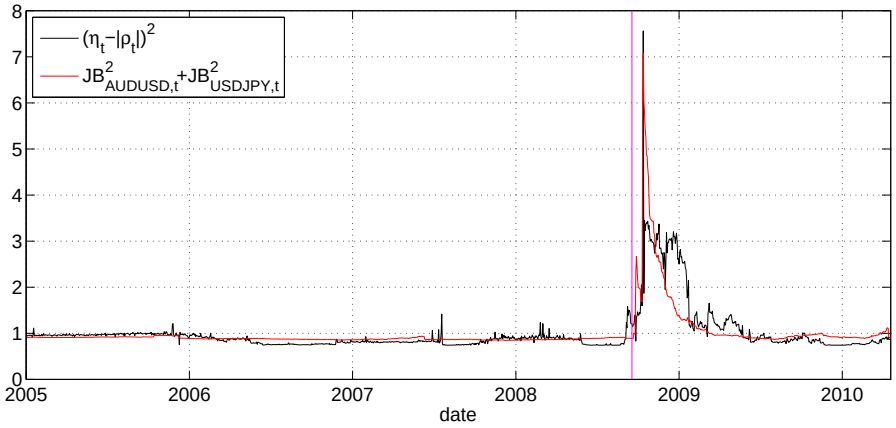

Fig. 4.25: Difference between information coupling and linear correlation plotted as a function of time. Also plotted is a measure of non-Gaussianity of the two time series as defined by (4.10). The two plots are scaled (such that they vary over the same range) for ease of comparison. The vertical line corresponds to the September 2008 financial meltdown.

pling measure before the 2008 crash. The increase in dependence of financial instruments in times of financial crises has been observed for other asset classes as well [311]. Our unique example, showing the dynamics of multivariate dependencies within the spot FX space, provides further insight into the nature of inter-dependencies in times of financial crises.

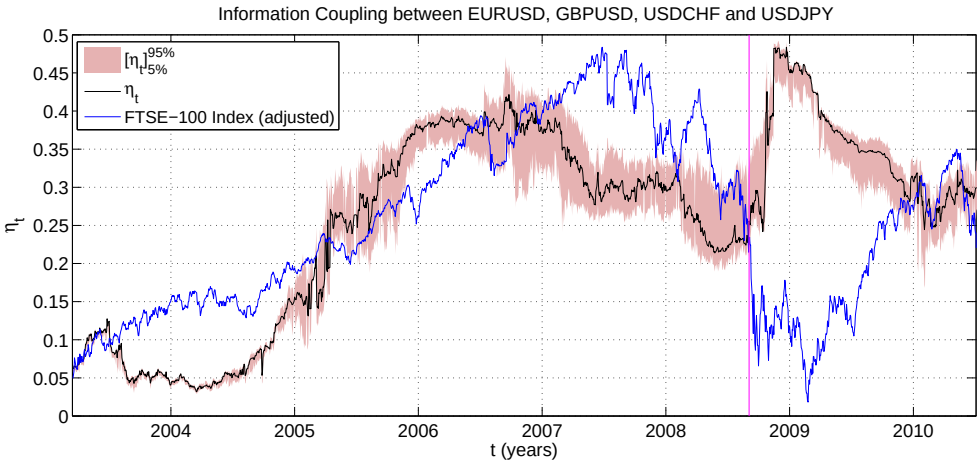

Fig. 4.26: Information coupling between EURUSD, GBPUSD, USDCHF and USDJPY log-returns plotted as a function of time. Also plotted is the FTSE-100 index, which is scaled (such that it varies over the same range as the information coupling plot) for ease of comparison. The vertical line corresponds to the September 2008 financial meltdown.

4.4.2 Case study 2: Information gain using a proxy exchange rate

In the FX market, different currency pairs have different levels of liquidity at different times of the day. One of the factors on which the speed of price discovery for any given currency pair depends is the relative liquidity level of that currency pair [110]. Price discovery refers to the process of how price movements react to the assimilation of relevant new information [75]; this information can range from order flow to macro-announcements, etc. In this case study we demonstrate how to make use of this characteristic of the FX market, together with the *predictive information coupling* approach, to develop a robust exchange rate forecasting model. All analysis which follows is based on strictly causal, out-of-sample methods, i.e. only information available up to time t is used to forecast relevant values at time $t+1$. As this model involves making tick level predictions using high-frequency sampled data, therefore we only compare those interaction measurement approaches, i.e. information coupling and linear correlation, which are computationally efficient enough to easily analyse data for this application.

Triangulated proxy rate as a predictor

We know that each FX currency pair consists of two currencies, the base (also called the denominator) currency and the quote (also called the numerator) currency. Defining the mid-price of the EURUSD currency pair at time t as $P_{EU}(t)$, the mid-price of EURCHF as $P_{EC}(t)$ and similarly the mid-price of USDCHF as $P_{UC}(t)$, the triangulated proxy USDCHF mid-price can be defined as:

$$\hat{P}_{UC}(t) = \frac{P_{EC}(t)}{P_{EU}(t)} \tag{4.11}$$

It may be possible to improve predictability of a relatively less liquid currency pair by using information contained in its triangulated proxy rate. This is because (by convention) FX currencies are often traded through one primary currency pair, which is therefore generally more liquid than other currency pairs containing that currency (the secondary currency pairs), and hence the exchange rates of the (more liquid) primary currency pairs often react more quickly to any new relevant market information than their corresponding (less liquid) secondary currency pairs [110]. As an example, USDCHF is relatively less liquid than both EURCHF and EURUSD, as CHF is traded mainly as EURCHF and USD as EURUSD (which is one of the most liquid currency pairs [313]); the primary market for all three currency pairs is EBS (Electronic Broking Services). Earlier (in Figure 4.47) we had noticed that USDCHF was the most closely *linked* currency pair in the dynamic FX network we had analysed, which again points to its importance as a stable pivot currency. Figure 4.20 also showed the pivotal role of USDCHF in a static FX coupling-MST. Therefore, it is quite likely that any change in the value of CHF or USD will generally be reflected first in EURCHF and EURUSD exchange rates respectively, before being assimilated in the USDCHF quoted price. This implies that a triangulated USDCHF proxy rate (as defined in (4.11)) will generally lead the quoted USDCHF spot rate and can therefore be used to potentially predict USDCHF spot movements. To demonstrate this effect, we first measure information coupling between the USDCHF exchange ($r_{UC}(t)$) and proxy ($\hat{r}_{UC}(t)$) log-returns at different time lags using 250 ms sampled data for five 8-hour trading sessions over five days. The resulting plots, presented in Figure 4.27, show a clear peak at a time lag of one tick (250 ms) for all five days, pointing to the presence of *predictive information coupling*.

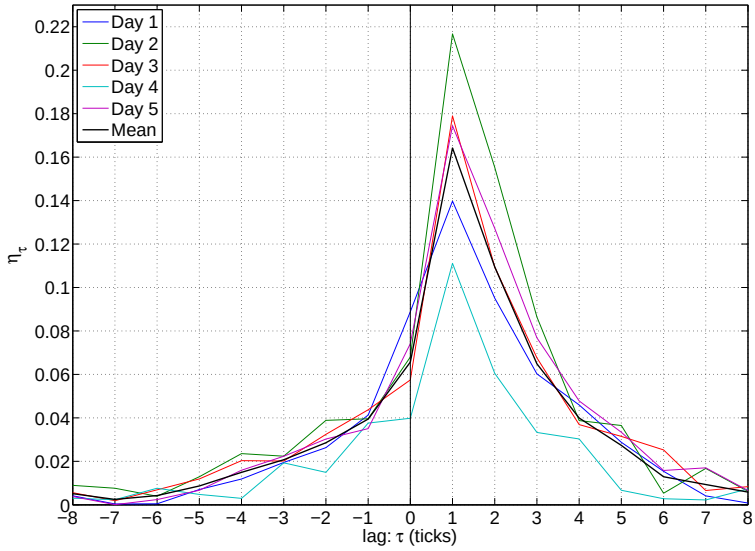

Fig. 4.27: Information coupling (η_τ) plotted as a function of time lag (τ) between $r_{UC}(t)$ and $\hat{r}_{UC}(t-\tau)$. The data is sampled at 4 samples per second, so each tick represents a time period of 250 ms. Results for 5 separate days are plotted together with the mean (black line).

We now develop a simple prediction model in which the proxy USDCHF spot rate at tick-time t is used to predict the value of USDCHF exchange rate at time $t+1$, i.e. the value at the next tick. Hence, the model buys USDCHF at time t if $\hat{P}_{UC}(t) > \hat{P}_{UC}(t-1)$ and sells USDCHF if $\hat{P}_{UC}(t) < \hat{P}_{UC}(t-1)$. The profit and loss (PnL), in pips, is then simply calculated as $P_{UC}(t+1) - P_{UC}(t)$ if going long (buying) USDCHF at time t or $P_{UC}(t) - P_{UC}(t+1)$ if going short (borrowing the asset and selling it). As FX currency pairs are often traded in lots of 1 million, therefore, a 1 pip move in the USDCHF mid-price represents a notional value of 100 CHF. The model was used to predict 250 ms sampled USDCHF spot rate over five 8-hour (115200 data points) trading sessions spread over five separate days. The resulting plots are presented in Figure 4.28(a). The plots show the robustness of this prediction model as the PnL gradient, i.e. the PnL per executed trade, is almost always positive for all five days; we hypothesise that the predictive accuracy of such a prediction model can be further improved by making use of causality analysis models (as we describe later). Prediction results obtained using a forecasting model may have a low error variance, but this can potentially be due to the predictable nature of the data, rather than the accuracy of the model. A simple prediction approach, based on a trend persistence model, can be used to act as a benchmark for accuracy of the prediction model presented above. FX spot returns at times exhibit some level of tick

level trend-following in short time horizons. This can be due to a range of factors, such as a large order being filled in smaller clips or the effect of a macro news release etc. This simple model assumes that the sign of one-step ahead USDCHF log-return, $r_{UC}(t+1)$, is the same as sign of the current log-return, $r_{UC}(t)$, i.e.:

$$\text{sgn}[r_{UC}(t+1)] = \text{sgn}[r_{UC}(t)] \tag{4.12}$$

Figure 4.28(b) shows the accuracy of this model when applied to the same data set as above. The mean of the five days (plotted in black) clearly shows the inability of this strategy to provide consistently positive returns. In contrast, the prediction model based on the proxy exchange rate can provide consistently positive returns over multiple days (Figure 4.28(a)). However, not all strategies resulting in a positive PnL are useful in practice. A more important measure is the PnL gradient. The PnL gradient can be optimised using statistical indicators, as discussed below.

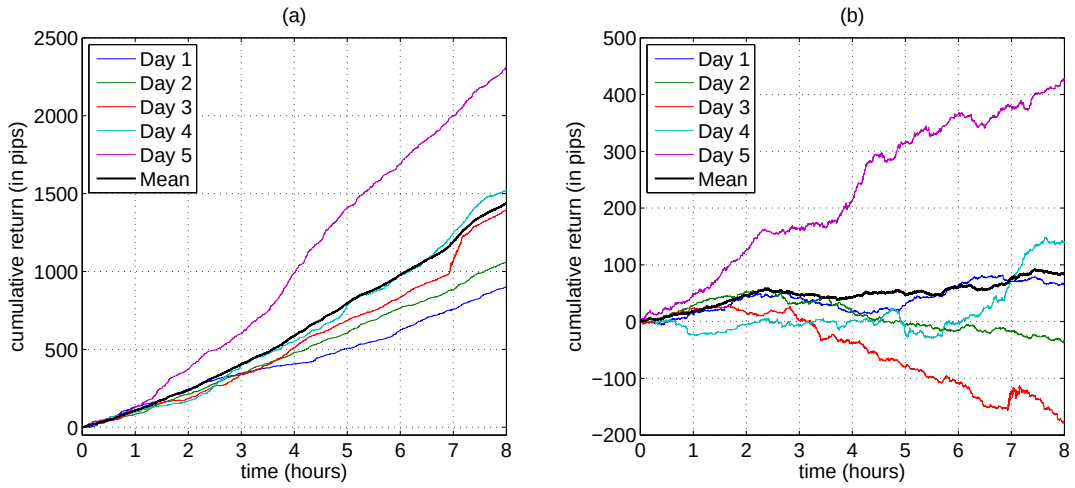

Fig. 4.28: (a). Plots showing the cumulative return (in pips) for predicting the USDCHF exchange rate (250 ms sampled) using the triangulated proxy rate. Results for five trading sessions over five different days are presented. The black line represents the mean of the five plots. The plots show the utility of incorporating information about the triangulated proxy rate in the simple predictive model. (b). Plots showing the accuracy of a simple trend persistence model (as given by (4.12)) when used to analyse the same data set without using the triangulated proxy rate. As the cumulative mean return (in black) shows, there is not much useful information to be obtained using this model. Note the different y-axes for the two figures.

Information coupling as an indicator

Trading FX pairs entails some transaction costs. These include the bid/ask spread, brokerage costs, slippage, etc. Therefore, it is generally preferred to only execute those trades which can overcome these trading costs and still result in a positive net PnL. The effect of bid/ask spread can be minimised by using a model that only places passive orders (bids and offers) instead of aggressing the market and paying the spread. Moreover, spot FX spreads have been tightening over the last few years [208], so even if the spread needs to be crossed it will result in relatively lower overall trading costs. However, brokerage and potentially slippage will still need to be paid for any trades executed. Therefore, for any live trading model, the higher the accuracy, the better it is. Estimating the mean PnL per tick, a measure of PnL gradient, which a model can generate is a standard and useful method to benchmark the level of accuracy of these prediction models. Statistical indicators are signals which can be used to indicate when a model should execute a trade such that the probability of correctly predicting the direction of the market, and hence the mean PnL per tick, is maximised. We now present use of the ICA-based information coupling measure as a statistical indicator, as described below.

An important pre-defined characteristic of any prediction model is its trading frequency. This represents the average number of trades the model is designed to execute in a given time interval. For the purpose of this example, the model we present is designed to make predictions on average every 5 seconds. This can be regarded as a typical trading frequency for a model making use of high-frequency sampled data, although different trading time periods can be selected based on the application and needs of the user. As the data is sampled at 4 samples per second, therefore on average the model will make one prediction for every 20 ticks, i.e. it will predict 5% of the ticks. Thus, the model will make 5760 predictions per 8-hour trading session each day. Now the problem arises as to how to select 5% of these ticks such that the mean PnL per prediction is maximised. For this purpose, we make use of information coupling as an indicator. Information coupling (η) is calculated between $r_{UC}(t)$ and $\hat{r}_{UC}(t-1)$ at each time step, t, i.e. we dynamically calculate $\eta\left[r_{UC}(t), L_\tau \hat{r}_{UC}(t)\right]$, where L_τ is a unit lag operator (i.e. $L_\tau r(t) = r(t-1)$, $L_\tau^2 r(t) = r(t-2)$, etc.). A sliding-window of size 40 data points (10 seconds) is used to produce temporal trailing coupling estimates for the last 10 seconds. This enables the model to capture the micro-trends in the data. We hypothesise that regions of high coupling

should result in a higher PnL gradient. Therefore, in order to select 5% of the ticks which will likely result in the highest mean PnL, we use the information coupling based indicator to only make predictions when $\eta \geq$ 95th percentile. Figure 4.29 shows normalised information coupling pdfs for five 8-hour trading sessions over five different days. Also indicated next to each plot is the 95th percentile threshold line. It is interesting to note that $\eta_{95\%}$ threshold values for all five days are very close together which can aid us in estimating $\eta_{95\%}$ values for other days. The mean PnL (in pips) per tick is obtained when the η_i threshold, i.e. i-th percentile of η, is reduced from 100% to 0. Figure 4.30 presents the average results obtained for the five trading days. As hypothesised, as the η_i threshold is reduced, the mean PnL per tick goes down. This shows the effectiveness of information coupling as an indicator to decide when to execute a trade which will result in a higher mean PnL. Using this information, it is also possible to select a η_i threshold value such that the expected return on average is greater than the estimated trading costs per trade.

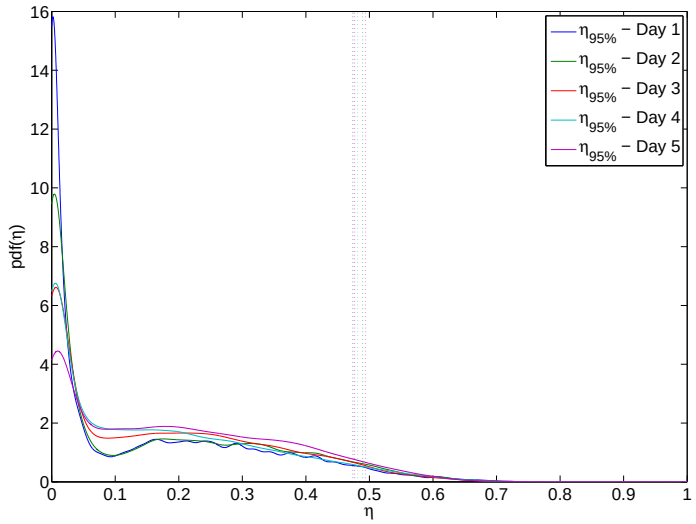

Fig. 4.29: Normalised pdf plots for the ICA-based information coupling measure (η) for five 8-hour trading sessions over five days. The plots reflect the distribution of η obtained when the information coupling model was used to dynamically compute coupling between the 250 ms sampled USDCHF exchange and triangulated proxy rates using a 40 data points wide sliding-window. Also plotted are the 95th percentile lines showing the $\eta_{95\%}$ threshold values for all five days.

Quantifying information gain

We are now in a position to quantify the accuracy of information gained using the coupling based indicator. We also compare results obtained with a linear correlation based indicator. For

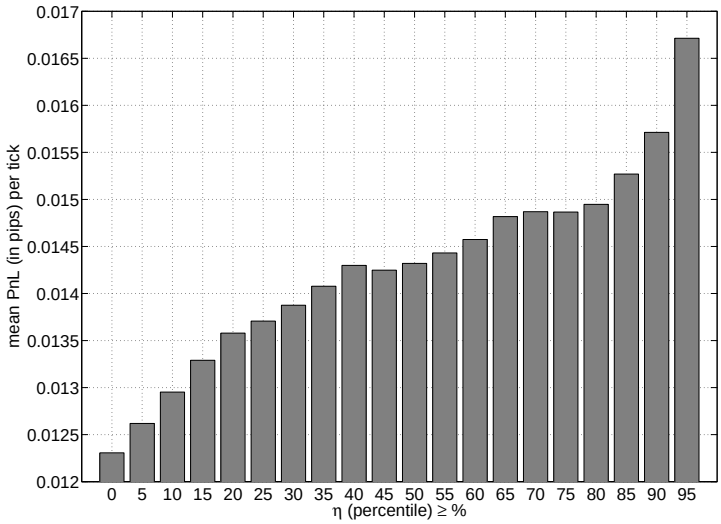

Fig. 4.30: Variation of mean USDCHF PnL (in pips) per tick as the η_i threshold, i.e. *i*-th percentile of η, is reduced from 100% to 0. The figure represents average of five 8-hour trading sessions over five days. Data is sampled every 250 ms.

each of the five days, the model is run over 8 hours (115200 data points) of data and predictions for the USDCHF exchange rate $P_{UC}(t+1)$ are made whenever $\eta(t) \geq$ 95th percentile. The cumulative return (in pips) for the 5760 predictions thus obtained for each of the five days is presented in Table 4.5. A similar model, but this time using linear correlation, $\rho(t)$, as an indicator is also used to get the cumulative returns and the results presented in the same table. Also presented in the table are the standard deviations of the returns[6]. As can be seen from the table, on average the information coupling based indicator outperforms the linear correlation based indicator by 6.7% in terms of PnL and by 10.2% in terms of the return-to-risk ratio. The coupling based indicator results in a higher cumulative PnL and return-to-risk ratio on four of the five days. Figure 4.31 shows the mean cumulative PnL of five days plotted against the number of ticks for both the indicators. Also plotted on the same graph is the cumulative PnL obtained using a model that uses no indicator but only makes predictions at fixed time intervals of 5 seconds. From the plots it can be ascertained that using the coupling based

[6]Highest possible returns are not always the preferred choice, there is a return-to-risk trade-off for all financial models. For financial returns, low standard deviation implies lower risks associated with the model. Prediction models with low standard deviations and high return-to-risk ratios are hence preferred [252]. A measure known as the Sharpe ratio is widely used in practise, which represents the ratio of the average return per trade to the standard deviation of the returns [320]. As choice of normalisation scale does not affect the relative values of the ratios, therefore for ease of comparison, return-to-risk ratios are calculated using average hourly returns for all five days and the results presented in Table 4.5.

indicator resulted in a cumulative PnL improvement of over 37% as compared to the fixed time interval prediction model (which acts as the benchmark).

	$\sum$PnL (in pips)		σ_{SD}(PnL)		Return-to-risk ratio	
Day	η	ρ	η	ρ	η	ρ
1	62.5	70.0	0.0900	0.1036	86.8	84.5
2	66.5	61.0	0.1026	0.0912	81.0	83.6
3	96.0	91.0	0.1566	0.1674	76.6	68.0
4	163.0	142.5	0.1982	0.1905	102.8	93.5
5	92.5	86.0	0.1641	0.1780	70.5	60.4
Mean	96.1	90.1	0.0652	0.0674	184.2	167.1

Table 4.5: Table showing accuracy of the two models, based on information coupling (η) and linear correlation (ρ) indicators, for analysis done using five days of 250 ms sampled data. Also included are values for the mean PnL and the standard deviation of the mean PnL. The return-to-risk ratio is calculated using hourly returns.

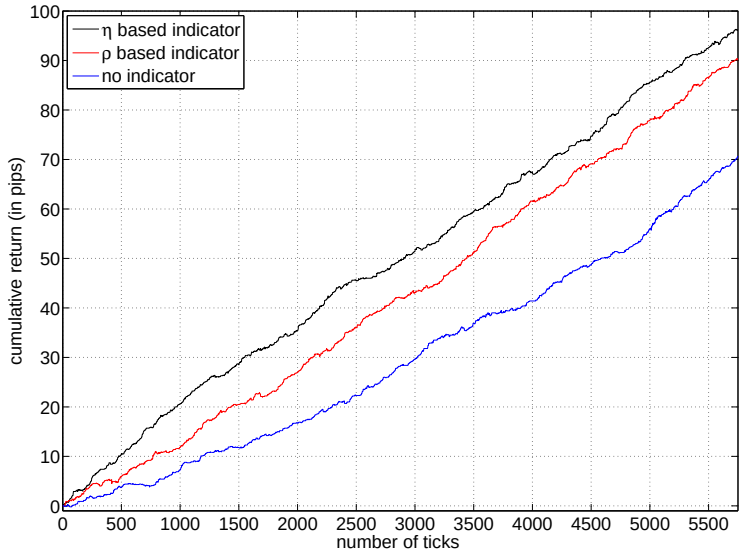

Fig. 4.31: Average cumulative PnL (in pips) for five 8-hour trading sessions over five days, obtained using the coupling and correlation based indicators. Also plotted is the cumulative PnL obtained without using any indicator, but instead making predictions at fixed time intervals of 5 seconds; this plot acts as a benchmark for accuracy of the two indicator-based models.

For each of the five days, we undertake another experiment in which we run the coupling and correlation indicator based models over a set of 1 hour (14400 data points) of "new" data and by using pre-fixed values for the $\eta_{95\%}$ and $\rho_{95\%}$ indicator thresholds obtained using the previous 8 hours of data for each day (as presented in Figure 4.29, $\eta_{95\%}$ threshold for all five days analysed is in a narrow range, i.e. between 0.4760 and 0.4964). The results obtained are presented in Table 4.6. It is interesting to note the relative similarity in the number of

predictions made using both indicators. As mentioned, the $\eta_{95\%}$ threshold for the model was set to a value which would on average result in one prediction every five seconds. Maintaining the same threshold in this example and even using new data resulted in the model making predictions on average every 4.66 seconds, which shows the reliability of using the model for new data. Once again, the total PnL, mean PnL and the return-to-risk ratios are higher for the η based indicator on four of the five days. On average, the coupling based indicator outperforms the correlation based one by 13.8% in terms of the total PnL, by 11.3% in terms of mean PnL (PnL gradient) and by 20.7% in terms of the return-to-risk ratio.

	$\sum$PnL (in pips)		$\sum$predictions		Mean PnL		σ_{SD}(PnL)		Return-to-Risk ratio	
Day	η	ρ	η	ρ	η	ρ	η	ρ	η	ρ
1	9.5	11.5	641	677	0.0148	0.0170	0.1016	0.1022	93	112
2	12.5	9.0	1028	1027	0.0121	0.0087	0.1015	0.0850	123	105
3	16.5	14.5	775	707	0.0212	0.0205	0.2198	0.2333	75	62
4	8.5	6.5	633	644	0.0134	0.0100	0.1468	0.1289	58	50
5	2.5	2.0	787	728	0.0032	0.0027	0.1273	0.1483	20	13
Mean	9.9	8.7	773	757	0.0128	0.0115	0.0559	0.0593	177.1	146.7

Table 4.6: Table showing accuracy of the η based indicator and the ρ based indicator when applied to 1 hour of new data using pre-fixed $\eta_{95\%}$ and $\rho_{95\%}$ thresholds. The return-to-risk ratio is calculated using hourly returns.

Discussion

Any statistical model dealing with high-frequency data (such as in this example) needs to be computationally efficient in order to reduce its impact on the latency of the trading system. Computationally, the information coupling indicator based prediction model takes on average 9 ms (on a 2.66 GHz processor) to make each prediction. Given that the data is sampled at 250 ms, this makes the model suitable for practical use, even when dealing with data sampled at high frequencies. In this example, transaction costs, such as bid/ask spread, brokerage, slippage, etc. are not taken into account. The effect of bid/ask spread can be minimised by placing passive orders rather than hitting bids and lifting offers. However, not all passively placed orders are filled within a reasonable time period, therefore any model used in practice will need to include effect of varying levels of fill ratios [9]. Brokerage is generally a constant cost and its effect can be included relatively easily once the actual values are known. However, the amount of any slippage depends on a number of factors and can only be modelled using detailed analysis [54], which is outside the scope of this thesis.

4.4.3 Case study 3: Dynamic portfolio optimisation

It is common practise to group together various financial assets in the form of a portfolio for investment purposes. There are various characteristics of a portfolio which have a direct impact on its performance. These include (amongst others) the number of financial assets making up the portfolio, the weight (or ratio) of each asset, the correlation of the assets with other assets which are not part of the portfolio and between themselves, and the time span for which each asset is kept in the portfolio. Active monitoring of correlation between different assets in a portfolio is one of the most important aspects of portfolio analysis, as portfolios with assets having low cross-correlation usually have a lower downside risk associated with them. Portfolio selection and optimisation is the primary task of many asset and risk managers in the equities sector. Any portfolio is designed to optimise the overall return-to-risk profile of all the assets it contains. Risk is generally considered to be the volatility of the expected returns of an asset and is usually estimated using the standard deviation of the returns. Due to the dynamically changing dependency structure between various instruments in the financial markets, the proportion of different stocks in a portfolio needs to be regularly rebalanced in order to take into account any changes in the coupling of the underlying instruments [33]. Modern portfolio theory (MPT) provides an elegant mathematical framework for analysing multivariate financial portfolios [240]. Over the years, numerous modifications and extensions to MPT have been proposed, many of which are still widely used in practise; however, most of them are based on the same basic principle of providing a relationship between the covariance matrix of asset returns and the optimum proportions (or weights) of different assets in the portfolio. Although the MPT is based on the assumption of normal distribution of returns, its utility can be potentially improved by making use of non-Gaussian measures of interaction. Therefore, in this section we propose the use of different measures of statistical dependence to estimate the returns covariance values and present empirical results obtained when using these different approaches to estimate optimum portfolio weights.

Global minimum variance portfolio

Any given portfolio is designed to reflect a certain type of return-to-risk profile. In the analysis presented in this case study, we consider the case of a global minimum variance (GMV)

portfolio which has the sole aim of minimising the variance of the expected returns [84, 350]. As the volatility of returns is a reflection of the risk of a portfolio, therefore GMV portfolio analysis can be considered as a value-at-risk (VaR) minimisation method [307]. Analysis of a GMV portfolio also allows us to focus on the purpose of comparing the effect of using different measures of dependence on the properties of a portfolio. As described in [247, 350], the N-dimensional vector of optimum portfolio weights at any given time, $\mathbf{w}_P(t) = [w_{P,1}(t), ..., w_{P,N}(t)]^\mathsf{T}$, for a GMV portfolio can be obtained by minimising the variance of the portfolio returns:

$$\min \sigma_P^2(t) = \min \left(\mathbf{w}_P^\top(t) \mathbf{\Sigma}(t) \mathbf{w}_P(t) \right) \tag{4.13}$$

subject to the following condition:

$$\sum_i w_{P,i}(t) = 1 \tag{4.14}$$

where i represents a single instrument in the portfolio and $\mathbf{\Sigma}(t)$ is an estimate of the temporal covariance matrix. It is common practise to put in place constraints on the range of values $w_{P,i}(t)$ can take [34]. In our analysis, we place a no-short sale constraint by restricting individual weights to the range $0 < w_{P,i}(t) < 1$, while still meeting the criterion set in (4.14). This allows us to study properties of the portfolio when a short sale restriction is placed in the market and to avoid cases where absolute values of the individual weights may become unrealistically large due to estimation errors [36]. For the case when $0 \leq w_{P,i}(t) \leq 1$, the covariance matrix has to be slightly modified as described in [194]. Optimum values of $\mathbf{w}_P(t)$ can be inferred from (4.13) using quadratic programming approaches [138]. For the analysis presented in this case study, we estimate the covariance matrices based on various measures of symmetric interaction. We achieve this by multiplying elements of the symmetric matrices (of different dependency measures) by standard deviations of their corresponding time series [349], and (if required) by computing the nearest (based on the 2-norm distance) positive semidefinite matrix [167]. For information coupling and mutual information, we obtain the sign of dependence using the rank correlation measure. This is because although rank correlation can give misleading information about the *strength* of dependence (due to potential loss of information from the data being analysed, as previously discussed), being a non-parametric

measure it does not assume any specific type of distribution of the data being analysed, hence the slope (positive if increasing, negative if decreasing) of the monotonic function being used to assess relationship between any two variables can give us a reliable indication of the *sign* of dependence between them.

Selecting stocks for an equities portfolio

In practise, a major issue in portfolio construction is selecting a sub-set of stocks, from a large set of thousands of listed stocks, which will result in the desired risk-return profile for any specific portfolio. For this purpose, a MST can be very useful [86], as it allows a user to easily visualise and analyse the hierarchical dependency structure among a large number of stocks. For our example, we demonstrate the use of an information coupling based equities MST to select a portfolio of 5 liquid stocks (from a basket of 25 stocks), as described below.

In the equity market, stocks within each sector are generally more closely correlated as compared to stocks in other sectors [20]. We can use the coupling-MST approach to study the structure of inter-sector and intra-sector relationships in the equity market. In the results presented here we use 25 stocks, each of which is a member of the S&P-500 (Standard and Poor's 500) index, representing 5 groups of 5 stocks each. The 5 stocks in each group are selected from the 10 largest stocks by market capitalisation in any given sector. The sectors, together with the symbols and names of stocks in each sector, are listed in Table 4.7.

Energy	Financial	Technology	Healthcare	Transport
XOM	JPM	AAPL	JNJ	UPS
Exxon Mobil Corp.	JPMorgan Chase & Co.	Apple Inc.	Johnson & Johnson	United Parcel Service, Inc.
CVX	C	GOOG	PFE	UNP
Chevron Corp.	Citigroup Inc.	Google Inc.	Pfizer Inc.	Union Pacific Corp.
SLB	BAC	MSFT	MRK	FDX
Schlumberger Ltd.	Bank of America Corp.	Microsoft Corp.	Merck & Co., Inc.	FedEx Corp.
COP	GS	IBM	ABT	CSX
ConocoPhillips	Goldman Sachs Group, Inc.	IBM Corp.	Abbott Laboratories	CSX Corp.
OXY	WFC	ORCL	AMGN	NSC
Occidental Petroleum Corp.	Wells Fargo & Company	Oracle Corp.	Amgen, Inc.	Norfolk Southern Corp.

Table 4.7: Table representing a set of 25 stocks, each of which is a member of the S&P-500 index. The set represents 5 groups of 5 stocks each. The 5 stocks in each group are selected from the 10 largest stocks by market capitalisation in any given sector. These stocks are used for the analysis presented in this section.

For this set of stocks, Figure 4.32 presents a colour map showing pair-wise information coupling between each of the 25 stocks. To obtain the results, we make use of 5 years of daily

log-returns, covering the period 2005 to 2010. The dark lines are used to group together stocks belonging to different sectors of the economy. It is interesting to note the high inter-sector coupling of stocks belonging to the energy, financial and transport sectors, while stocks within the technology and healthcare sectors are least closely coupled. The high coupling of stock returns in the transport sector has been observed in previous studies as well, and is most likely due to sensitivity of these stocks to the global oil price [266]. The close coupling of energy sector stocks can also be explained by their dependence on the global demand and supply of oil. Similarly, financial sector stocks are (in-part) influenced by the interbank interest rates and governmental interventions (especially since the 2008 financial crisis), resulting in their high coupling values. In contrast, stock prices of technology and healthcare (pharmaceutical) companies are often driven by the sale and development of new and innovative products, which can explain the relatively low coupling values of stocks within each of these sectors. We now build an information coupling based equities MST, as presented in Figure 4.33. Nodes of the same colour represent stocks belonging to the same sector. As expected, the MST shows a number of groups, each of which generally contains stocks from the same sector.

As previously discussed, we make use of the coupling-MST approach (as our preferred choice for complex network analysis) due to its ability to extract the hierarchical coupling structure by exhibiting only the most relevant links, hence resulting in simplicity and visual clarity. However, it is also possible to combine the information coupling model with other network analysis approaches. Here we consider one such approach, by presenting an example of analysing complex coupling networks using a community detection model based on extremal optimisation techniques, details of which are presented in [111][7]. This example makes use of the same equities data as used in the previous example. The ICA-based information coupling measure is used as the dependency measure of choice for extracting the communities. Figure 4.34 presents the complex network obtained using the community detection algorithm. Once again, the network clearly shows the presence of five distinct clusters, with each cluster generally containing stocks from a specific sector of the economy. The only anomaly is the transport sector, which has two of its stocks present in the technology cluster.

[7]The extremal optimisation heuristic can be used to find solutions to hard optimisation problems by successively replacing "extremely" undesirable variables of a single sub-optimal solution with new random variables [45].

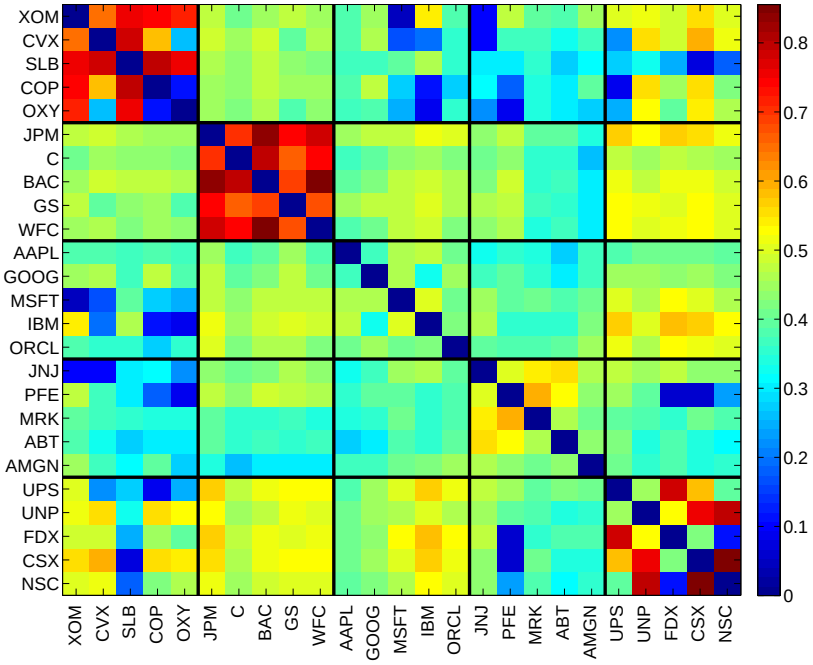

Fig. 4.32: Information coupling between 25 stocks, obtained using daily log-returns for the period 2005 to 2010. The dark lines are used to group together stocks belonging to different sectors of the economy. Coupling between the same stocks has been set to zero for ease of presentation. Names of the stocks represented by the symbols are presented in Table 4.7.

We now use the information coupling based equities MST (as presented in Figure 4.33) to select a portfolio of five stocks, each from a different sector of the economy, from a total of twenty-five stocks presented in the network. The complex coupling network shown in Figure 4.34 can also aid in this process. As already mentioned, in practise, the stocks selected for inclusion in a GMV portfolio should have low coupling between them, e.g. Figure 4.33 can be used to select five stocks, each from a different economy sector, based on two criteria; firstly, they should be linked to the least number of other nodes in the network, and secondly, the selected link's distance measure d should be the largest in the group in which the stock is placed, i.e. the stock should have low coupling to other stocks in that group. Making use of a MST for portfolio selection has the added advantage of identifying any outliers, e.g. although GOOG is a technology sector stock, the MST in Figure 4.33 shows it is statistically more closely coupled in the financial sector group and hence in this case should not be selected to represent the

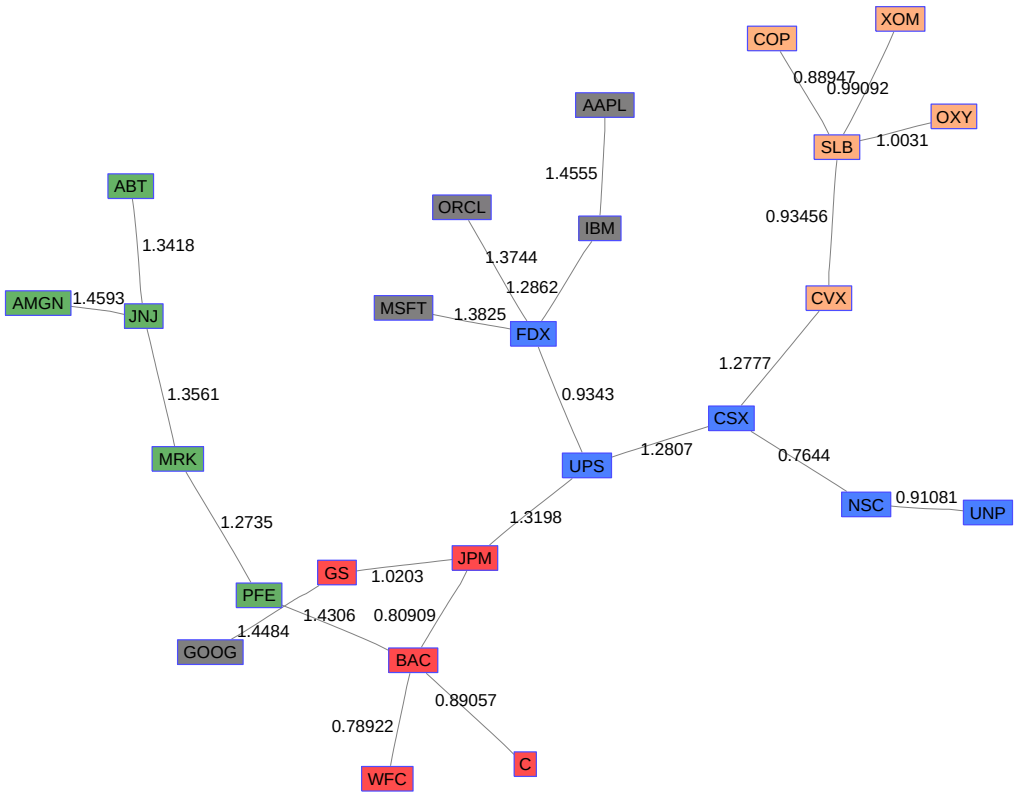

Fig. 4.33: MST showing various groups representing different sectors of the economy. Nodes of the same colour represent stocks belonging to the same sector. The plot was obtained using daily log-returns for the period 2005 to 2010. Each node represents one of the 25 stocks, all of which are constituent members of the S&P-500 index. The names of the stocks represented by the symbols are presented in Table 4.7.

technology sector group[8]. In practise, the structure of a financial MST keeps on changing with time (a good example being the dynamic coupling network which we present later in Figure 4.45), due to the temporal variation of coupling between different nodes. Hence, the stocks selected for inclusion in a portfolio need to be changed periodically as well. However, for the purposes of this example, we only analyse a portfolio of the same five stocks in order to simplify presentation of results. To show effectiveness of the model even under a regime of comparatively high coupling and to remove the possibility of any potential comparative bias in the results (as the MST is based on the information coupling measure), the five stocks we select from the five different groups meet two conditions; firstly, they are linked to at least one other group, and secondly, they have the most number of groups within a distance of two links.

[8] As an aside, the multivariate information coupling model can also be useful for the purpose of stock selection by selecting a sub-set of stocks which have the least multivariate coupling between them. This can be particularly useful when dealing with a very large number of stocks for which the MST approach may not be suitable.

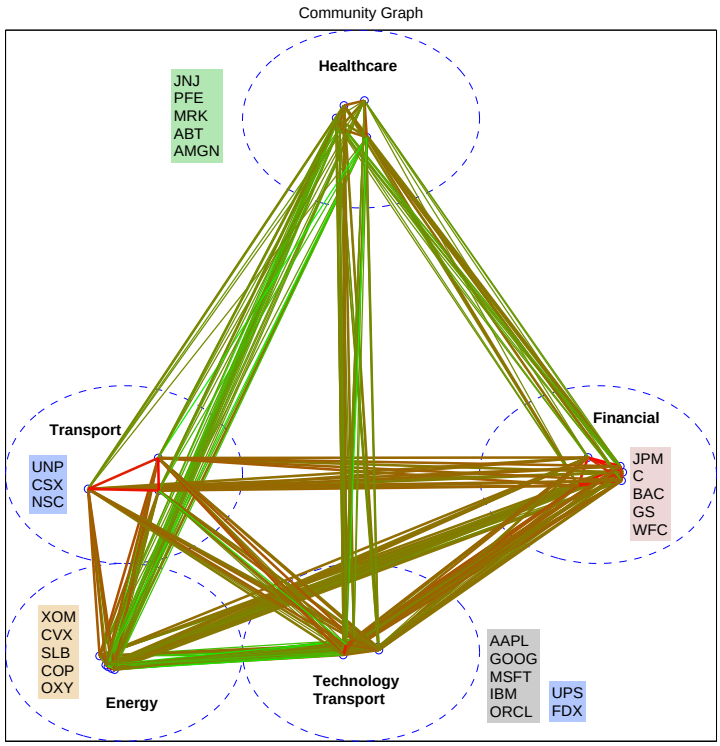

Fig. 4.34: Complex network showing the presence of five distinct communities in the data set being analysed. The network is obtained using the extremal optimisation approach, as presented in [111]. The colour of the lines represents strength of the connection between the nodes, with green representing a weak connection and red representing a strong connection. The stocks belonging to each community, and the sector to which these stocks belong, are listed next to the community. The names of the stocks represented by the symbols are presented in Table 4.7.

Based on these criteria, we select the following five stocks for analysis presented in this example: UPS (transport), CVX (energy), IBM (technology), JPM (financial) and PFE (healthcare). We analyse data over a 10 year period between 2002 and 2012, and use log-returns based on the daily closing prices for each of the five stocks. The data was obtained from NASDAQ and all prices are in USD [2]. The multivariate coupling between these five stocks (using the whole data set) is 0.2854 compared to 0.4505 for all the twenty-five stocks in the network; the low coupling value for the five stocks is as we would have expected, as they belong to different sectors of the economy.

Dynamically optimising the portfolio

Having selected the stocks for inclusion in our portfolio, we now present results obtained when dynamically optimising a GMV portfolio containing these stocks. For the results presented below, we dynamically estimate the optimum portfolio weights using covariance functions based on four different measures of dependence, i.e. information coupling (η), linear correlation (ρ), rank correlation (ρ_R) and normalised mutual information (I_N). We compare these results with a simple equal weighted portfolio (EWP) which assumes that at each time step, t, the individual portfolio weights of a N instrument portfolio are simply given by $\frac{1}{N}$, i.e. the expected return of an EWP is simply the average of the returns of all the assets in that portfolio at any given time. In each of the five cases, we compute the out-of-sample expected returns of the portfolio and the standard deviation of these returns. A low value of standard deviation implies lower volatility, which is the primary aim of a GMV portfolio. As already mentioned, in practise, to optimise the return-to-risk profile of a portfolio, the weightings of different assets in a portfolio need to be rebalanced frequently. The frequency of rebalancing depends on a number of factors, including the underlying volatility of the individual assets, the transaction costs associated with buying and selling the assets, as well as the dynamically changing dependency structure among the assets in the portfolio [215]. The portfolio weight rebalancing period ($t_{rebalance}$) as well as the sliding-window length (Δt) used to estimate the weights of a portfolio can have a significant impact on its performance. Therefore, we first compare the effect of varying $t_{rebalance}$ and Δt on the volatility of returns (when using different measures of interaction); the results obtained are presented in Figure 4.35. To obtain these results (over a 10 year period), we varied both $t_{rebalance}$ and Δt in increments of one day up to a maximum of three months. Figure 4.35(f) compares the plots obtained using different approaches, and excludes the EWP plot for ease of presentation.

As is evident, all four models that take into account the changing structure of the covariance matrix, i.e. plots (a) to (d), outperform the EWP model by a significant margin. It is worth noting that for small $\Delta t/t_{rebalance}$ ratios, the information coupling based model outperforms the rest. This is most likely due to the fact that as Δt decreases, the data analysed becomes increasingly non-Gaussian and can be regarded as locally linear, hence resulting in comparatively higher accuracy. It is also interesting to note that in this example, the variability

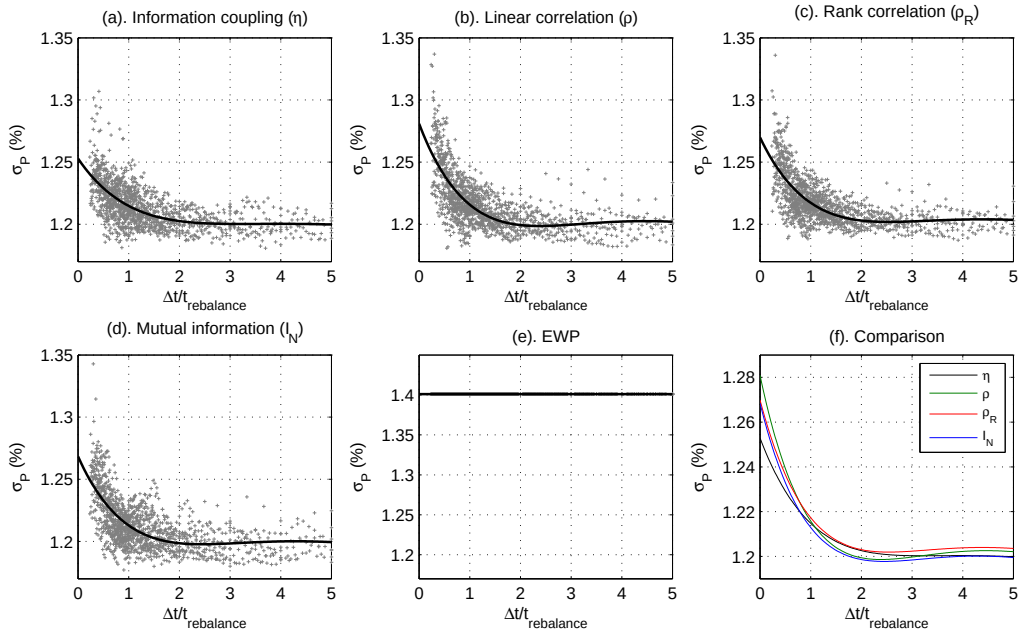

Fig. 4.35: Scatter plots showing effect of varying the ratio of sliding-window length (Δt) (in days) to the portfolio weight rebalancing period ($t_{rebalance}$) (in days) on the standard deviation of the portfolio returns, $\sigma_P(\%)$, for a portfolio of five stocks. The solid lines represent curves of best fit for the respective scatter plots and are useful to visualise the general trend in the data. The results are obtained using daily log-returns over a period of 10 year, and by varying both $t_{rebalance}$ and Δt in increments of one day up to a maximum of three months. Plot (f) compares the results presented in plots (a) to (d). Note that plots (e) and (f) have different scales to the four other plots for clarity.

of portfolio returns, σ_P, does not decrease significantly from the $\Delta t \approx 2t_{rebalance}$ point onwards. This information can be useful to select the maximum rebalancing period possible in order to minimise transaction costs associated with buying and selling stocks. The plots also show that as the $\Delta t/t_{rebalance}$ ratio keeps on increasing beyond ≈ 3.5, the information coupling and mutual information based models outperform the three other models. Overall, the mean values of σ_P over all values of $\Delta t/t_{rebalance}$ for the different models (in ascending order) are: 1.2144% (I_N), 1.2167% (η), 1.2189% (ρ), 1.2203% (ρ_R) and 1.4010% (EWP). These results show that using information coupling or mutual information measures to estimate the covariance matrix in the MPT framework can potentially result in improved portfolio weight estimates as compared to using a simple linear correlation based covariance matrix which is widely used in practise. As a general rule, if a portfolio is designed to be rebalanced less frequently and a large sliding-window is used to capture low frequency trends in the data, the I_N based measure may be better suited as it gives improved results for large data sets (by more accurately estimating the pdfs) and computational efficiency is usually not an issue in such cases. However,

if a comparatively small window is used, e.g. to capture recent market dynamics in a volatile market, or if not enough historic data is available, then the η based measure is potentially better suited as it can better capture the non-Gaussian dynamics of the data. Figure 4.36 shows the variation of volatility of returns obtained using different values of Δt and $t_{rebalance}$ for the information coupling based model. We note that for the set of equities analysed, it is possible to achieve a good GMV portfolio profile even when rebalancing the stocks in the portfolio at a relatively low frequency, provided ample historic data is used to estimate the weights of the portfolio. This can be advantageous in terms of transaction cost savings, especially if the portfolio is made up of a large number of stocks.

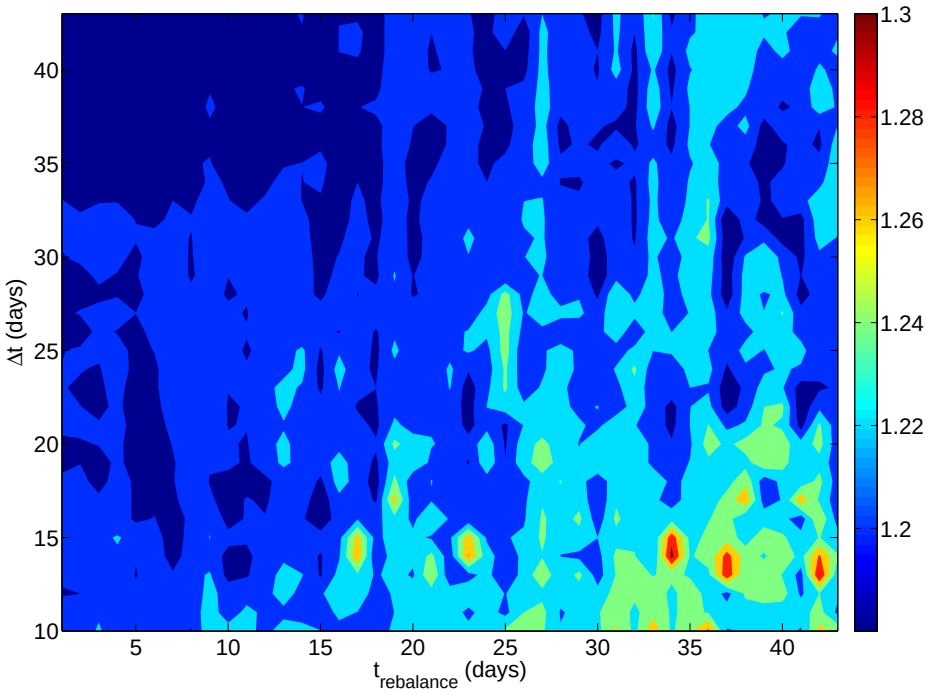

Fig. 4.36: Contour plot showing variability of the standard deviation of portfolio returns, $\sigma_P(\%)$, for different values of the portfolio rebalancing period ($t_{rebalance}$) and the sliding-window length (Δt), obtained using the information coupling (η) based portfolio optimisation model. The portfolio is made up of five stocks and daily sampled log-returns data is analysed over a period of 10 years. Values of $t_{rebalance}$ and Δt in the top left quadrant result in a near optimum GMV portfolio.

We now consider the case of optimising a GMV portfolio using relatively longer time scales. As an example, Figure 4.37 shows the standard deviation of returns for a set of different Δt and $t_{rebalance}$ values obtained using different models. The EWP values are omitted for clarity. When analysing the portfolio using Δt and $t_{rebalance}$ values of up to three months (66

days), we had shown that σ_P is generally minimised (and stays approximately constant) for $\Delta t \geq 2t_{rebalance}$, as shown in Figure 4.35. This effect can be seen in Figure 4.37(a). However, for larger values of Δt, the variability of returns generally increases with the rebalancing period, as shown by plots (b) to (d). Of the sixteen combinations of Δt and $t_{rebalance}$ considered in this particular example, the information coupling measure based portfolio outperforms the rest in eight cases, while it provides the second best results in a further four cases. It is interesting to see how using more data to estimate the weights of the portfolio, i.e. using a larger Δt, generally seems to decrease the accuracy of all the models, i.e. results in higher σ_P values for the same values of $t_{rebalance}$. This is due to the changing dynamics of the markets and further reinforces the need for careful selection of user-defined variables when optimising a portfolio of assets. This also points to the potential practical utility of using measures of dependence, such as information coupling, which can efficiently model non-Gaussian interactions in rapidly evolving market conditions.

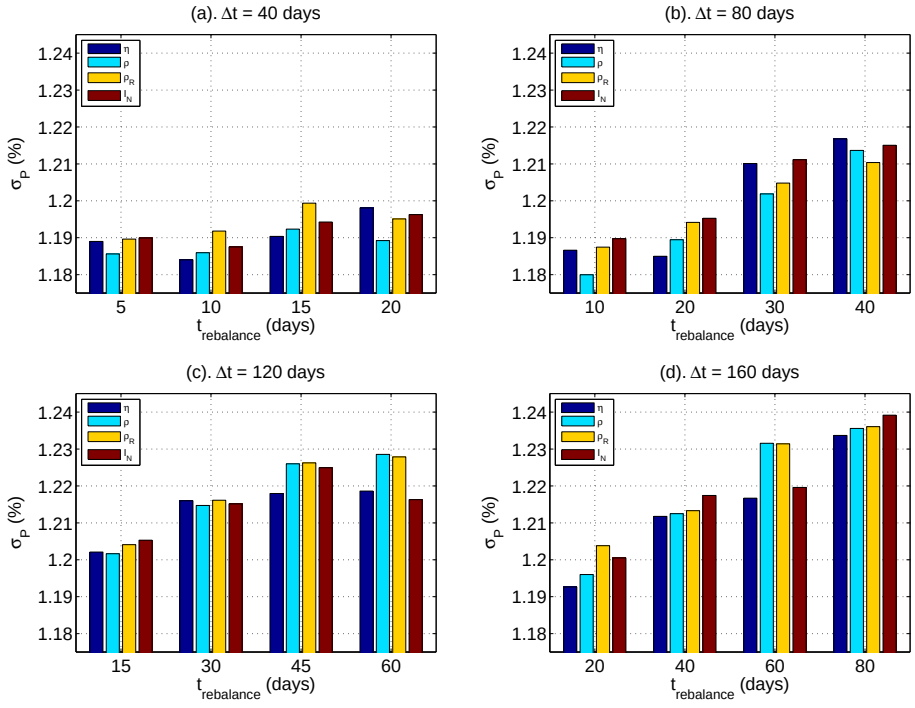

Fig. 4.37: Variation of the standard deviation of returns (σ_P) with the sliding-window length (Δt) and the rebalancing period ($t_{rebalance}$). All plots are obtained using 10 years of daily sampled log-returns for a five stock equities portfolio. See text for details.

4.4.4 Case study 4: Currency index tracking

In many financial sectors, tracking or replicating the temporal dynamics of a financial index, using a selected set of financial instruments, has wide-ranging practical applications. These range from tracking major stock market indices, e.g. S&P-500 [129], to tracking exchange traded funds [32]. There are many advantages to using index tracking methods instead of replicating the index in full, e.g. buying or selling all 500 stocks in the S&P-500 to fully replicate the index can lead to substantial transaction costs, a large part of which can be avoided by only investing in a selected set of the most liquid or highly capitalised stocks which form the index. Moreover, index tracking models can address any restrictions, e.g. limitations on foreign ownership, which may be in place in relation to trading any given financial instrument by not including it in the basket of instruments used for tracking purposes. From a statistical point of view, the problem of index tracking comes down to selecting, and then assigning appropriate weightings, to a basket of financial instruments such that the tracking error is minimised. In this case study we demonstrate utility of the ICA-based information coupling model for tracking a weighted USD index using a set of currency pairs, although the analysis and models presented here can be used to address index tracking problems in other asset classes as well. We compare our results with other standard measures of symmetric interaction.

Constructing a currency basket to track the USD index

In the FX market, currencies are quoted and traded in pairs. This implies that the actual strength of any single currency remains unclear if looking at the exchange rate of only a selected set of currency pairs containing that specific currency. For example, USDJPY may be strengthening at the same time as USDCAD is weakening, giving us no clear indication of USD's strength. To address this issue, it is possible to form a currency index which contains information about the exchange rates of most, if not all, currency pairs which contain that specific currency as one of its components. In this example, we make use of a normalised geometric mean based USD index which gives us the normalised mid-price, $P_I(t)$, of USD at time t, as a measure of the currency pair invariant strength of the USD. We make use of six major liquid currency pairs, i.e. AUDUSD, EURUSD, GBPUSD, NZDUSD, USDCAD and USDJPY, to form this index, which is given by [176]:

$$P_I(t) = \left[\prod_{k=1}^{n} \frac{P_k(t)}{P_k(t_0)}\right]^{\frac{1}{n}} \tag{4.15}$$

where $P_k(t)$ is mid-price (at time t) of the k-th USD containing currency pair, out of a total of n pairs (in this example $n = 6$). For all six currency pairs used in (4.15), USD is the base currency, i.e. we use the USDAUD mid-price, which is simply reciprocal of the AUDUSD quoted mid-price.

We can now construct a basket of m USD containing currency pairs which can then be used to track the six currency USD index. The normalised mid-price (at time t) of this currency basket, $P_B(t)$, is given by:

$$P_B(t) = \sum_{j=1}^{m} w_j(t) \frac{P_j(t)}{P_j(t_0)} \tag{4.16}$$

where $P_j(t)$ and $w_j(t)$ are the mid-price and the weight of the j-th USD containing currency pair at time t. The weights of the basket meet the following conditions:

$$\sum_{j=1}^{m} w_j(t) = 1, \quad 0 \leq w_j(t) \leq 1 \tag{4.17}$$

As before, we convert all currency pairs to have USD as the base currency.

Dynamically tracking the index

The tracking problem now reduces to finding the optimum weight vector, $\mathbf{w}(t) = [w_1(t)...w_m(t)]^{\mathsf{T}}$, at each time step, t, which maximises dependence between the USD index, $P_I(t)$, and the currency basket, $P_B(t)$. We compare performance of the different models using the tracking error, e_{TE}, which is defined as:

$$e_{TE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} [P_I(t) - P_B(t)]^2} \tag{4.18}$$

where T is length of the data set analysed (in this case, length of the sliding-window). We also estimate the correlation between the index and the basket (as a measure of accuracy). To simulate the index tracking model, we use 0.5 hour sampled FX spot data over a period of two years. We make use of six USD containing currency pairs to form the USD index (as mentioned earlier). To construct the tracking basket, we use three USD containing currency pairs. Using fewer instruments makes the tracking basket increasingly sensitive to weight

estimates, hence allowing us to compare the effectiveness and accuracy of the different models. We present results for two different three-pair baskets; the first basket is constructed using currency pairs whose mean gives the highest tracking error when used to track the USD index, while the mean of the second basket is least correlated with the index (as compared to all other possible combinations of currency pairs). Hence, using these two baskets allows us to test the accuracy of different models in "extreme" cases, where simply using a EWB is not suitable. In practise, the actual selection of instruments in the tracking basket will be based on a number of factors, some of which we mentioned earlier, i.e. liquidly, relative transaction costs, restrictions on trading, etc. Also, in practise, instruments may be included and excluded from the tracking basket depending on market dynamics, however, this is not considered in the results presented in this section.

We now use a 12 hour (24 data points) long sliding-window to model dynamics of the data over a period of two years, rebalancing the basket at each data point (0.5 hours). To causally track the USD index, we first use normalised log-returns of the three currency pairs in the basket to dynamically estimate the optimum weights at each time step and then use these weights to infer value of the index at the next time step. Trading spot FX incurs low transaction costs, so in practise rebalancing the basket on a regular basis may be feasible, depending on nature of the application and the number (and type) of currencies in the basket. Figure 4.38 shows variation of the USD index, $P_I(t)$, over the two year period together with the tracking basket, $P_B(t)$, based on four different dependency measures, i.e. information coupling (η), linear correlation (ρ), rank correlation (ρ_R) and normalised mutual information (I_N). The results are also compared with a simple equally weighted basket (EWB), which represents mean of the mid-prices of the three currency pairs in the basket.

From the plots presented in Figure 4.38, we can clearly see inability of the mutual information measure to accurately estimate weights of the instruments in the basket. This is because results obtained using direct computation of mutual information require large data sets for accurate estimation. The information coupling measure (a proxy for mutual information) gives relatively accurate results for the data analysed, as is evident from the values in Table 4.8. The table includes estimates for the tracking error, e_{TE}, as well as the correlation between the USD index and the tracking basket. Figure 4.39 compares the normalised pdf plots for the squared

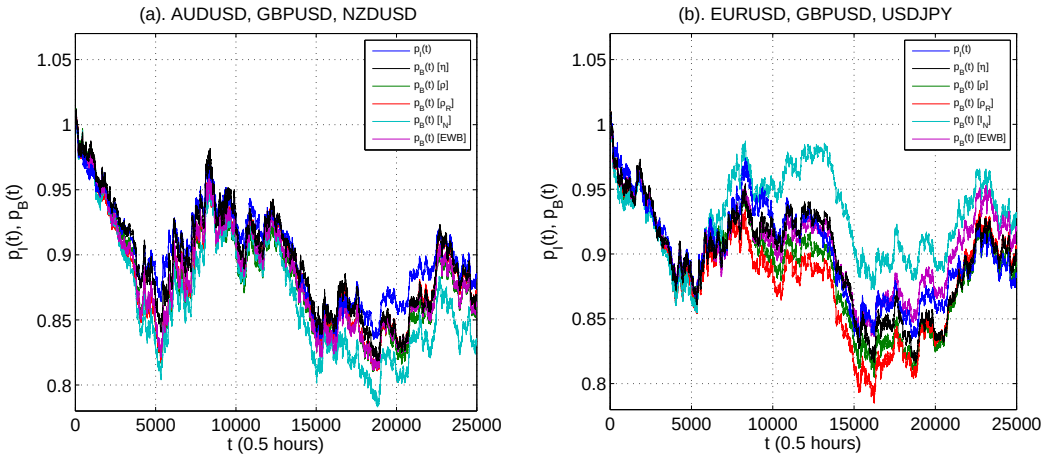

Fig. 4.38: The USD index, $P_I(t)$, plotted against time for a two years long (0.5 hour sampled) data set. Also plotted are the index values estimated using a tracking basket, $P_B(t)$, made up of three currency pairs, i.e. (a). AUDUSD, GBPUSD and NZDUSD, (b). EURUSD, GBPUSD and USDJPY. Results obtained using five different tracking models are presented, i.e. models based on using information coupling (η), linear correlation (ρ), rank correlation (ρ_R), normalised mutual information (I_N) and equally weighted basket (EWB).

tracking error (e_{TE}^2) resulting from using the five different models compared in this example. It is clear from these plots that the information coupling based model has the lowest mean and MLE for the tracking error and its pdf plot decays relatively quickly as compared to the other models. On the other hand, the model based on the direct computation of mutual information is clearly not suitable for this type of analysis for reasons discussed earlier.

	$e_{TE}(\times 10^{-3})$					$\rho(P_I(t),P_B(t))$				
j (USD-)	η	ρ	ρ_R	I_N	EWB	η	ρ	ρ_R	I_N	EWB
AUD,GBP,NZD	0.1190	0.2282	0.1840	0.3708	0.2392	0.9706	0.9598	0.9606	0.9638	0.9651
EUR,GBP,JPY	0.1241	0.1867	0.2753	0.3672	0.1549	0.9543	0.9525	0.8958	0.7256	0.9056

Table 4.8: Comparison of tracking error (e_{TE}) values obtained using index tracking models based on different measures of dependence, i.e. information coupling (η), linear correlation (ρ), rank correlation (ρ_R) and normalised mutual information (I_N). The results are also compared with a simple equally weighted basket (EWB). Also included are values for correlation between the USD index, $P_I(t)$, and the tracking basket, $P_B(t)$. All results are obtained using out-of-sample 0.5 hour sampled data over a period of two years.

4.4.5 Case study 5: Analysis and applications of scale-dependent FX coupling

Analysis of financial data at different frequencies has various applications (some of which we discussed earlier) and can enable us to extract interesting information about the scale-dependence of information coupling, which is the main focus of this case study. We first

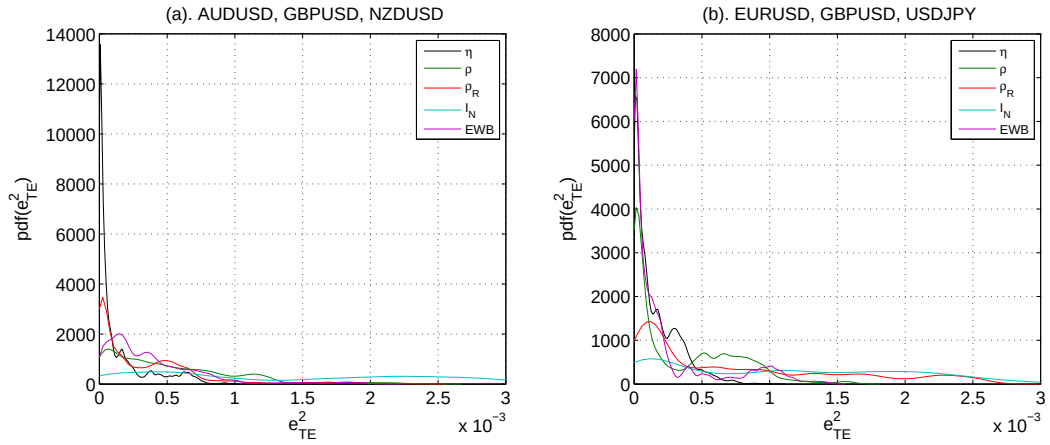

Fig. 4.39: Normalised pdf plots showing distribution of the squared tracking error, e^2_{TE}, obtained using models based on four different measures of dependence, i.e. information coupling (η), linear correlation (ρ), rank correlation (ρ_R) and normalised mutual information (I_N). Also plotted is the tracking error distribution for an equally weighted basket (EWB) which assumes that the weights stay constant. Plots in (a) are obtained using a tracking basket made up of AUDUSD, GBPUSD and NZDUSD while (b) represents results obtained using a basket made up of EURUSD, GBPUSD and USDJPY. The x-axis has been truncated for clarity.

present a simple example of application of the CWT based time-scale analysis approach to financial data. Figure 4.40 shows the temporal variation in values of CWT coefficients at six different scales for a section of 0.5 second sampled EURUSD log-returns data. For this high-frequency data set, a wavelet scale of 1 corresponds to a time period of 0.6154 seconds. We notice that the CWT coefficients follow an increasingly periodic path at scales greater than 2 minutes. It is also interesting to see the "zooming" property of the CWT, which is clear from this example; at lower scales the CWT captures the finer details of the signals, while at higher scales it captures the coarser signal structures, hence extracting scale-dependent structures in the data.

Earlier we presented the wavelet-ICA algorithm, which can be used to dynamically estimate the unmixing matrix, obtained using normalised CWT coefficients at different time-scales, and hence measure temporal variations in information coupling at different frequencies. We now use this algorithm to analyse the scale-based variations in information coupling in FX data streams. Figure 4.41 presents the coupling-scale plots (obtained using the wavelet-ICA algorithm) for a selected set of currency pairs. The plots show variation of the scale-based

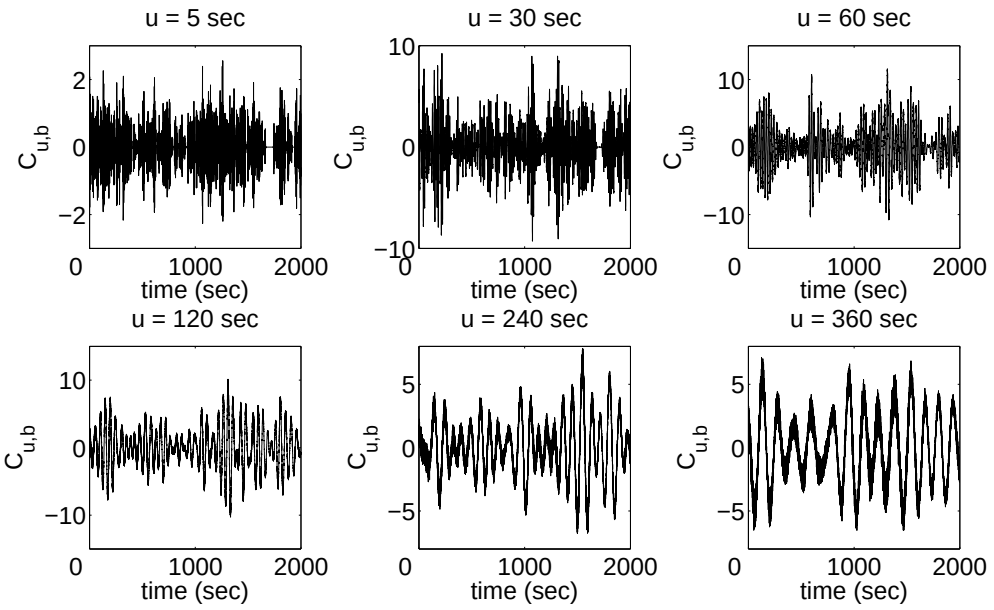

Fig. 4.40: CWT coefficients at different scales (in seconds) for EURUSD. Data is sampled at 2 samples per second.

magnitude of information coupling (η_u) for 0.5 second sampled FX log-returns[9]. The results are obtained using simulations carried out over 15 minute sections of the data and the contours represent the standard deviations of coupling over an 8-hour trading session. We notice that coupling generally increases with scale, although the rate of this increase in coupling is dependent on the currency pairs being analysed. For example, the coupling-scale gradient of plots (a) to (c) is noticeably higher than that for plots (d) to (f). We also note that at certain scales, information coupling drops to zero, implying that the data is symmetrically decoupled at these scales. This information can potentially be useful for certain risk management and portfolio selection applications [116]. To get a more general idea of the variation of information coupling with scale, we measure multivariate information coupling in all G10 currency pairs using eight 1-hour long data sets, with the results presented in Figure 4.42. We can clearly see the steady increase in average information coupling with scale and notice the relatively low standard deviation values associated with the measurements. The two examples presented above show that although the scale dependence of information coupling may vary significantly for any two currency pairs, multivariate coupling between major currency pairs

[9]Earlier (as presented in Figure 4.12) we showed that information coupling between non-Gaussian random variables is negligible at different time-scales, as we would expect. We use this information as a benchmark to ascertain the significance of frequency-dependent coupling values obtained using FX data in examples presented in this case study.

generally increases with scale. This is most likely due to the reason that at higher scales the general trend of major spot exchange rates is relatively more similar as compared to lower scales where currency pair specific factors (such as order flow) have a noticeable effect on exchange rate variations. This phenomena (i.e. presence of only weak dependencies at lower scales) has previously been observed in the equities market as well (where it is often known as the Epps effect) [49].

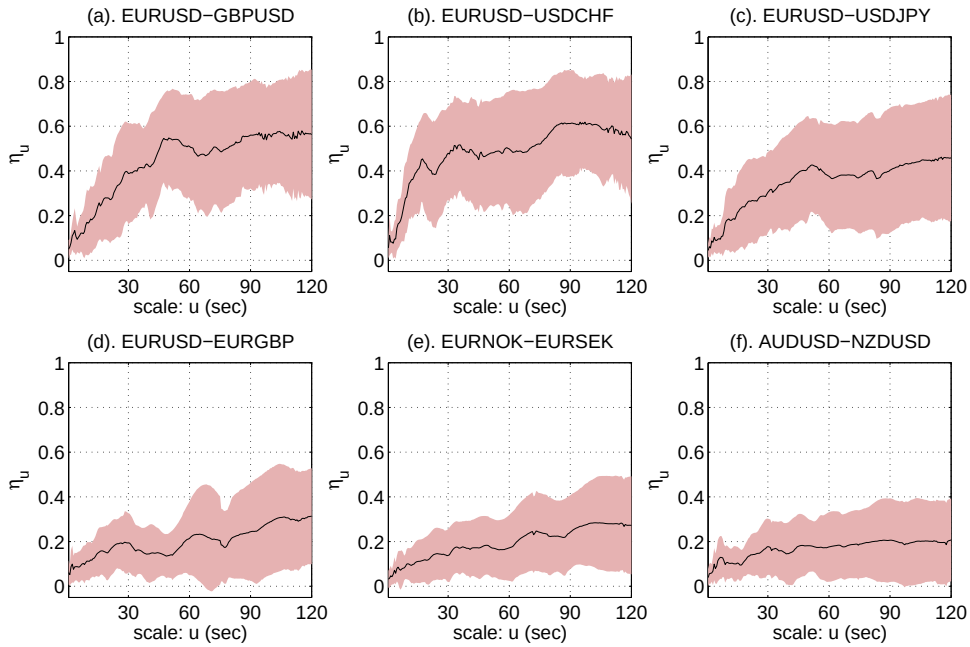

Fig. 4.41: Information coupling (η_u) between 0.5 second sampled FX spot returns at different scales (u) for various currency pairs. The solid lines represent average values for simulations carried out over 15 minute sections of the data and the contours represent standard deviation values of coupling over an 8-hour trading session (obtained using the 32 sections of data analysed). Plots (a) to (c) represent currency pairs with noticeably high coupling-scale gradient than the currency pairs analysed in plots (d) to (f).

Capturing discrete changes in scale-dependent coupling dynamics

Earlier we presented use of the HMICA model to detect regimes of high and low information coupling dynamics. We now demonstrate the utility of this approach using the wavelet-HMICA algorithm, which can be used to detect regions of coupling persistence at different time scales. For this representative example, we use 0.5 second sampled USDCHF and EURUSD spot log-returns data sets, covering a period of 8 hours. We detected state transitions (using a 2-state HMICA model) over the 8-hour period and also dynamically estimated in-

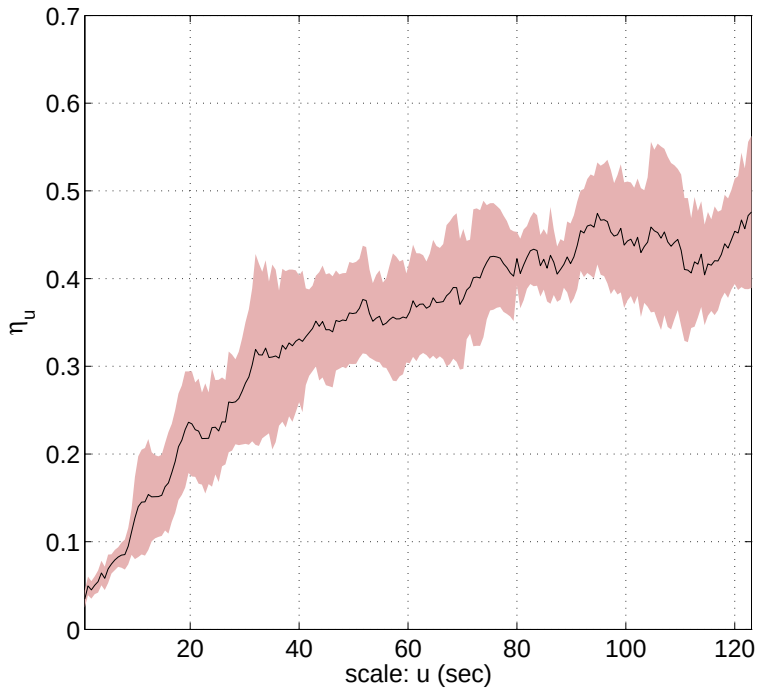

Fig. 4.42: Multivariate information coupling (η_u) at different time scales (u) between all G10 currency pairs sampled every 0.5 seconds. The solid line represents average of hourly values while the contour shows the standard deviation of η_u for eight hours of data.

formation coupling between the two currency pairs using a 20 second (40 data points) wide sliding-window. The results obtained are presented in Figure 4.43. Figure 4.43(a) shows the average value of information coupling (η_{state}) which corresponds to each HMICA state. We note that coupling in state 1 is generally greater than state 0 for all scales, exhibiting ability of the wavelet-HMICA algorithm to capture discrete state-based, frequency-dependent, information coupling dynamics. We also notice the gradual increase in information coupling with scale, a result which we had also observed in Figures 4.41 and 4.42. Figure 4.43(b) presents the corresponding time periods in each state, obtained using the state transition matrices, as per (3.50). We can notice the increasing state stability with scale, indicating that dynamics of information coupling become increasingly stable at lower frequencies. To get a more general indication of state stability across scale, we estimate the state transition probability matrices, $\mathbf{P}_{hmm}$, at different time scales, for the six 0.5 second sampled currency pairs we analysed earlier in Figure 4.41. We repeated the analysis 20 times over different parts of the data set, each 2500 samples in length, and estimated the average values of $\mathbf{P}_{hmm}$ for all six currency pairs. The resulting diagonal elements of the $\mathbf{P}_{hmm}$ matrix, corresponding to p_{00} and p_{11}, are plotted

in Figure 4.44. As before, state stability increases at higher time scales. We also notice that on average, values of p_{00} and p_{11} saturate at approximately a scale of 3 minutes for this data set, with no substantial increase in state stability thereafter. Later we notice a similar characteristic of the scale-based dynamics of information coupling when analysing dynamic coupling networks. We will discuss this further in the next case study.

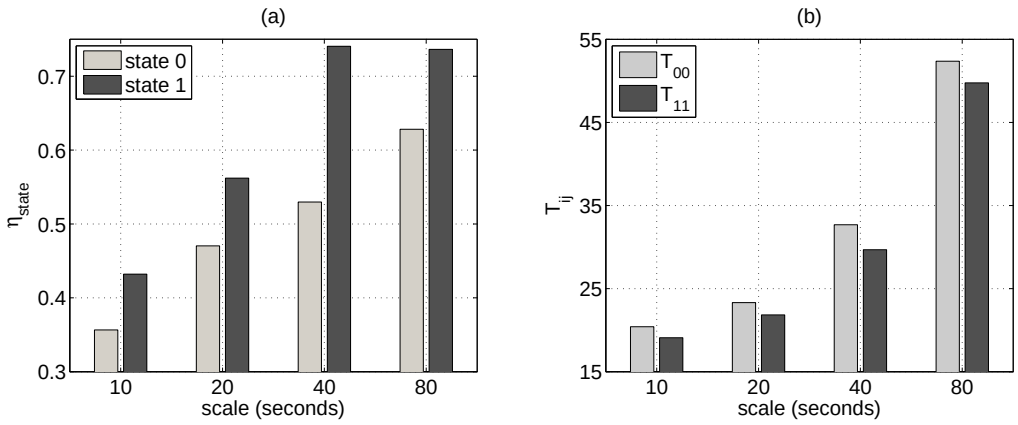

Fig. 4.43: (a). Average information coupling (η_{state}), between 0.5 second sampled USDCHF and EURUSD log-returns, corresponding to each HMICA state at different time-scales; results are obtained using a 20 second (40 data points) wide sliding-window, over a period of 8-hours. (b). Average time in each HMICA state (T_{ij}), which is a measure of state stability.

4.4.6 Case study 6: Analysis of dynamics of FX coupling networks

The coupling networks we have considered so far in this chapter had a static structure. However, (as previously discussed) the nature of interactions in financial markets dynamically changes with time. Therefore, for many practical applications, analysing the dynamics of financial networks can provide more useful information. To demonstrate this, in this case study we present a set of examples which provide us with a unique insight into the evolving structure of interactions in the FX market. As a first example, Figure 4.45 shows evolution of the structure of a FX coupling-MST over a period of 11 minutes. Each of the 12 graphs in the figure represents a MST at time intervals of 1 minute. In each MST, the 12 nodes represent individual currency pairs (sampled at 2 samples per second), and the weights of the links are a representation of information coupling between any two nodes. It is interesting to note that some links in the network are much more stable across time as compared to others. In this example, we let both the weights of the links, as well as the weights of the nodes, to dynamically evolve with time. Studying the evolution of the weights of the nodes provides us with a further

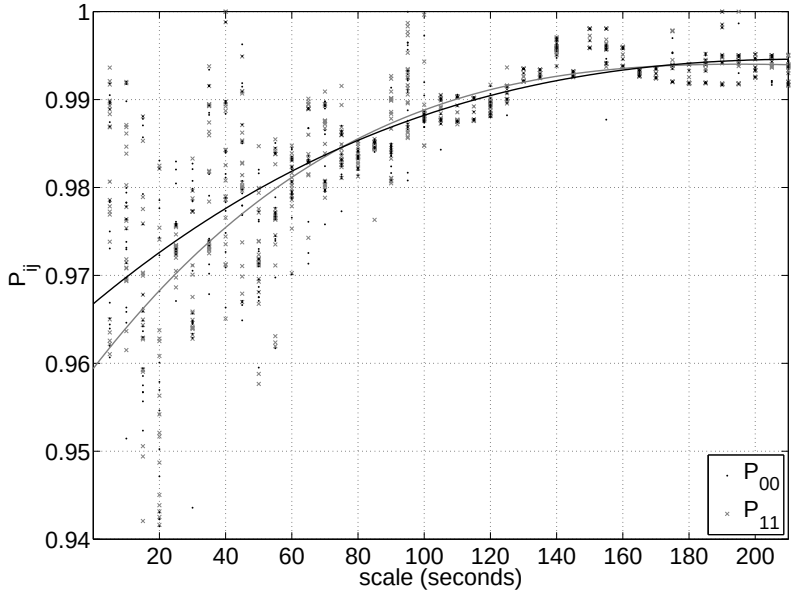

Fig. 4.44: Average scale-dependent values of the diagonals of the state transition probability matrix, $\mathbf{P}_{hmm}$, for the six 0.5 second sampled currency pairs analysed earlier in Figure 4.41. The scatter plots show values for p_{00} and p_{11} over 20 independent simulations over different parts of the data set, each 2500 samples in length. The solid lines represent lines of best fit for the scatter plots, giving us an indication of the general trend in the data. As discussed in the text, higher values of p_{ij} indicate increasing state stability.

aid to investigate the dynamical evolution of the MST; to calculate these weights, we make use of an approach based on the concept of "mean coupling". Indicated next to each node is a numeric value of the mean coupling, i.e. the average information coupling of that node with all the other nodes in the network at any specific point in time. The mean coupling gives us a measure of importance of a particular node (at any given time) in the network, and is defined as:

$$\hat{\eta}_{i,t} = \frac{1}{N} \sum_{j=1, j\neq i}^{N} \eta_{ij,t} \tag{4.19}$$

The time-varying parameter $\hat{\eta}_{i,t}$ is hence a measure of the extent of *linkage* of a given currency pair in a FX network at any given time.

Figure 4.46 shows the temporal variation of mean coupling for 4 of the 12 currency pairs. The top two plots show mean coupling of the USDCHF and EURUSD nodes respectively. We had previously observed that both these currency pairs are generally closely coupled with other

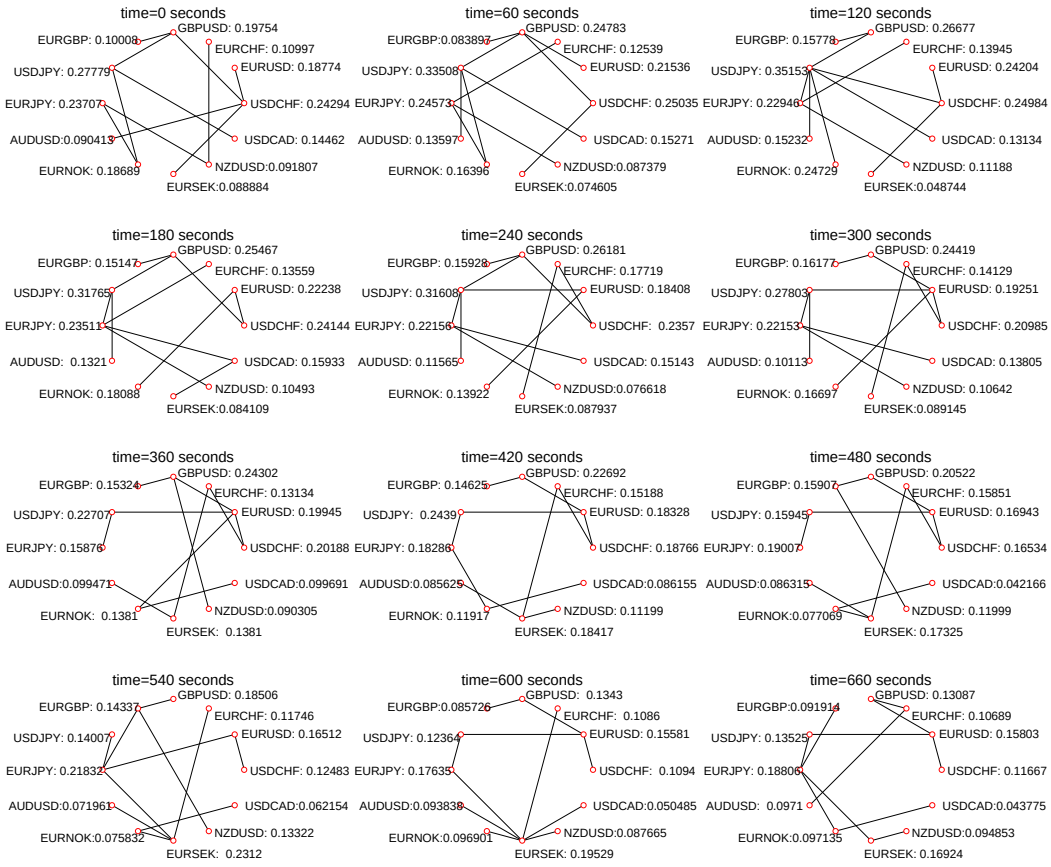

Fig. 4.45: Dynamically evolving coupling-MST showing a network of 12 FX currency pairs, each one sampled every 0.5 seconds. The 12 vertices (nodes) represent the currency pairs, while the 11 edges (links) represent the *distance* (d_{ij}) between the vertices, calculated using (3.44). Each MST represents the state of the hierarchical interaction structure between the 12 currency pairs at 60 second intervals. Each node is labelled with the name of the currency pair that it represents, as well as the mean coupling of each node with all other nodes, which is calculated using (4.19).

currency pairs[10]. The bottom two plots of Figure 4.46 show mean coupling of the EURNOK and NZDUSD nodes respectively. Both these currency pairs have a significantly lower mean coupling (which can be explained using the same reasoning as mentioned above for currency pairs with high mean coupling). It is also interesting to note that the mean coupling generally oscillates around an average value for all four currency pairs within a narrow range. This means that we can be fairly certain that although the coupling between any two given currency

[10]We can infer this from results shown earlier in Figures 4.15 and 4.20. Figure 4.15 shows that coupling of both these currency pairs is high with almost all other currency pairs. Figure 4.20 shows that both these currency pairs act as the major nodes in two separate groups.

pairs may vary significantly with time, the value of mean coupling of any given currency pair is relatively stable. We can also use the concept of mean coupling to get an idea of the general time-invariant importance of a currency pair in the FX market. The bar chart presented in Figure 4.47 shows the average of the mean coupling ($\hat{\eta}$) of 12 currency pairs over an 8-hour period. The chart ranks all the currency pairs according to their importance in the FX network, judged by their *linkage* with other currency pairs. From the figure it is clear that for the data set being analysed, USDCHF is the most *linked* currency pair in the FX network, followed closely by GBPUSD, EURUSD, USDJPY and AUDUSD respectively. A possible reason for the high degree of connectivity of USDCHF in the FX network could be due to its role as a classic arbitrage-leg currency pair [241]; a property which can be used for a range of practical applications, e.g. to develop an exchange rate forecasting model (as we presented earlier). It is also interesting to note that all five top ranked currency pairs contain the USD as one of the currencies. This shows that the USD is driving the dynamics of the FX network and is by far the most important currency during this 8-hour trading period, once again showing dominance of the USD in the global FX market.

Earlier we had discussed the concept of survival ratio as a useful approach for readily analysing the dynamics of high-dimensional financial networks. Using a set of examples, we now analyse the temporal variation of survival ratio coefficient, $\sigma_{SR}(t)$, of financial coupling networks across time as well as scale. Figure 4.48 shows four plots, each one representing the time-varying survival ratio of FX coupling-MSTs containing 45 nodes (currency pairs); details of the plots are included in the caption of the figure. Static versions of these two MSTs were presented in Figures 4.21 and 4.22. Figures 4.48(a) and 4.48(b) show that the survival ratio for the MST obtained using high-frequency (0.5 second sampled) data remains fairly high and stable across time. Also, by comparing Figures 4.48(a) and 4.48(c), or Figures 4.48(b) and 4.48(d), we can observe that the mean survival ratio is generally higher for data sampled at a higher frequency, implying that the coupling structure in FX networks is temporally more stable at lower frequencies. To analyse this scale-dependence of the survival ratio coefficient in more detail, we consider the case of scale-based variation of a coupling network's survival ratio. Figure 4.49 shows the variation of survival ratio with scale for a FX coupling-MST comprising of 12 currency pairs, each sampled at 2 samples per second. The wavelet-ICA

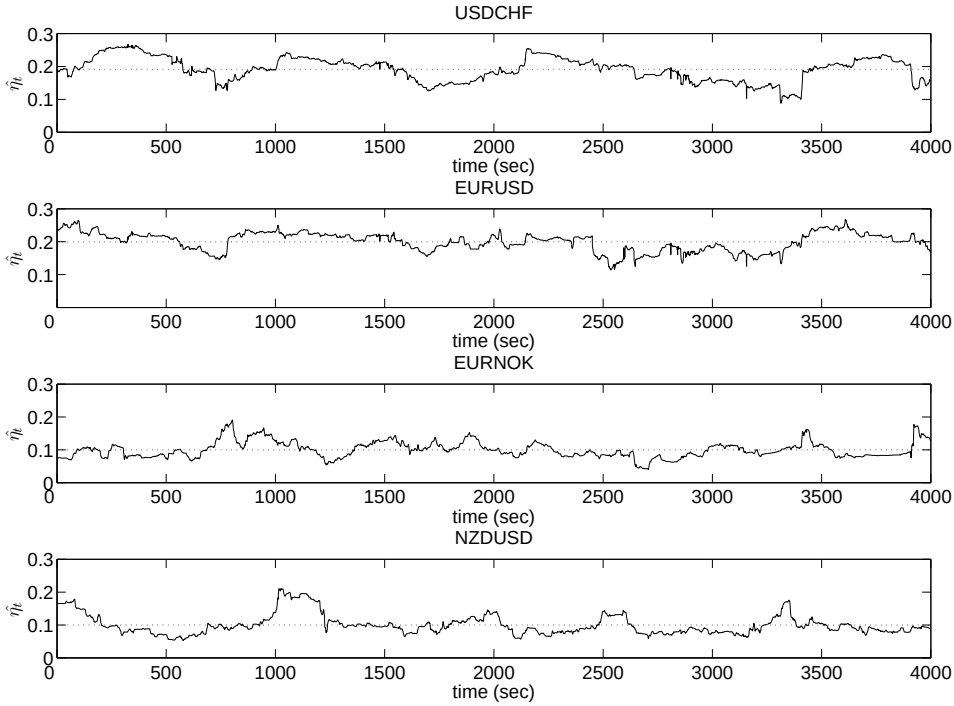

Fig. 4.46: Temporal variation of mean coupling, $\hat{\eta}_t$, of four currency pairs, each sampled at 2 samples per second. The top two plots relate to currency pairs (USDCHF, EURUSD) with high mean coupling, whereas the bottom two represent currency pairs (EURNOK, NZDUSD) with low mean coupling.

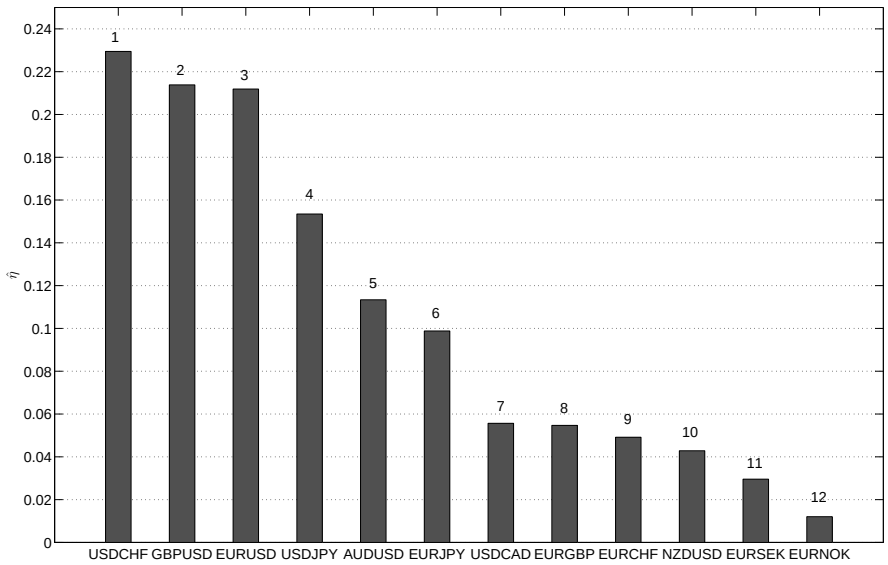

Fig. 4.47: Ranking of 12 currency pairs according to their time-invariant mean coupling, $\hat{\eta}$. See text for details.

algorithm was used to calculate information coupling at different time-scales. The plot shows that for the data being analysed, survival ratio generally increases with scale up to a scale of approximately 5 minutes. At higher scales, the survival ratio value remains relatively stable. Earlier (in Figure 4.44) we had noticed that the wavelet-HMICA model based state transition probability peaked at a scale of approximately 3 minutes. Both these results point to two general observations about the data set analysed. Firstly, the dynamics of information coupling become increasingly stable at higher time scales, i.e. lower frequencies, and secondly, at scales of larger than 3-5 minutes, the dynamics of information coupling stabilise.

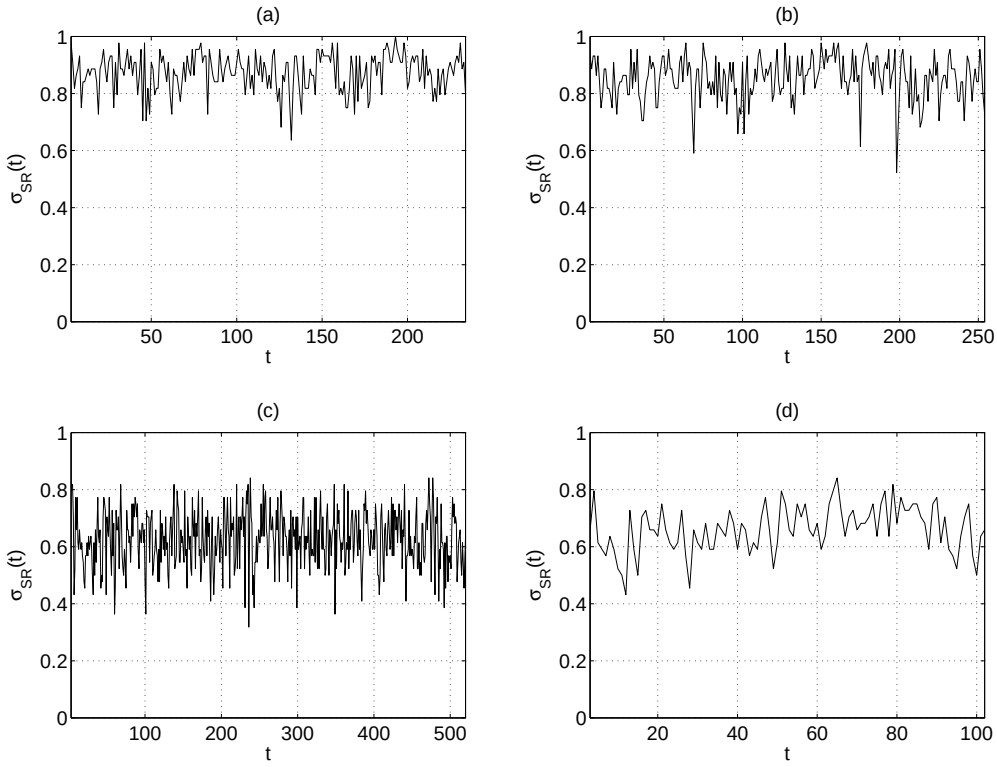

Fig. 4.48: Variation of survival ratio, $\sigma_{SR}(t)$, with time for a FX coupling-MST comprising of 45 currency pairs. The four plots were generated using data with the following properties. (a). sampling period: 2 samples per second, window size: 120 seconds (240 data points), step-size: 24 seconds (48 data points). (b). sampling period: 2 samples per second, window size: 10 minutes (1200 data points), step-size: 2 minutes (240 data points). (c). sampling period: 2 samples per hour, window size: 1 week (240 data points), step-size: 1 day (48 data points). (d). sampling period: 2 samples per hour, window size: 1 month (960 data points), step-size: 1 week (240 data points).

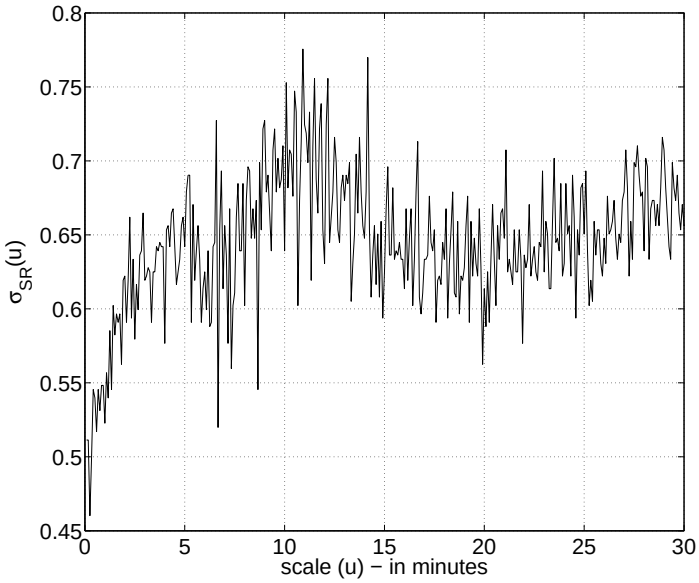

Fig. 4.49: Variation of survival ratio, $\sigma_{SR}(u)$, with scale, u, for a FX MST comprising of 12 currency pairs, each sampled at 2 samples per second. The window size and step-size used for the analysis was 5 minutes and 1 minute respectively.

4.5 Conclusions

In the previous chapter we presented the theoretical and analytical description of an ICA-based information coupling model (and its extensions) that can be used to efficiently and accurately measure symmetric interactions in multivariate non-Gaussian data streams. The model makes use of information about the source mixing process encoded in the ICA unmixing matrix to measure interactions in multivariate data sets, which, as far as we know, is the first such attempt. In this chapter we have provided an in-depth analysis of the application of our approaches using a range of synthetic and financial data examples. The comparative empirical results presented point to the accuracy and efficiency of the information coupling model when used to analyse financial data sets at a range of sampling frequencies. The results presented show a selection of the diverse range of practical applications in which our proposed approaches can be effectively used to extract interesting and useful information from financial data and can outperform standard interaction measurement approaches frequently used in practise. Most of our results point to inaccuracies in the direct computation of mutual information using finite data sets, limiting its usability in a dynamic framework. This further points to the practical utility of information coupling as a proxy measure for mutual

information. For an in-depth analysis of interactions in multivariate financial systems, we presented a range of extensions to the information coupling model. We exhibited the use of a 2-state HMICA model to capture discrete state-based dynamics of information coupling by identifying regimes of different coupling strength. To capture interactions across scale, we demonstrated use of the CWT, with a Morlet basis function, as a suitable time-scale analysis approach for financial data analysis. The wavelet-ICA model thus developed was used to analyse multivariate FX spot returns and the results obtained point to a general increase in information coupling across scale, indicating the presence of long-range dependencies in the FX market. To simultaneously measure both time- and scale-based interactions, we presented the wavelet-HMICA model which extracts state-based information coupling dynamics at different time-scales. Results obtained using this model point to increasing stability and persistence in coupling strength at higher scales. To accurately extract the hierarchical interaction structure in high-dimensional systems, we presented an approach which combines the information coupling model with techniques used for analysing MSTs. We demonstrated the utility of this coupling-MST approach to analyse multivariate financial data, both in static and dynamic environments. As our analysis shows, this approach has many useful applications, such as identifying dominant financial instruments in a network at any given time (e.g. by determining which currency pairs are "in play") or for extracting information from equities coupling networks with the aim of selecting stocks for inclusion in a portfolio. The financial case studies we presented in this chapter demonstrate the practical utility, of all the above mentioned extensions to the information coupling model, in dealing with a range of practical scenarios, such as dynamic portfolio optimisation, exchange rate prediction or index tracking.

The ICA-based information coupling model has multiple other benefits associated with its use. It provides estimates for the uncertainty associated with the information coupling metric; can be efficiently used to model dependencies in high-dimensional spaces without the need to resort to pair-wise dependency analysis; gives normalised and symmetric results; unlike many other approaches, such as direct computation of mutual information and use of copula functions, does not rely on user-defined parameters; by making use of the sliding-window based decorrelating manifold approach to ICA and by using a reciprocal cosh source model, results in increased accuracy and efficiency of the algorithm (in fact, the information coupling

model has computational complexity similar to that of linear correlation with the accuracy of mutual information); and unlike many information-theoretic and copula-based approaches, is not data intensive, i.e. it can be used even with relatively small data sets without having a significant impact on the results, an important requirement for analysing systems with rapidly changing dynamics such as financial returns. All these benefits make the information coupling model a very attractive choice for the online analysis of multivariate financial data. As with any other approach to interaction measurement, the ICA-based information coupling model also has some limitations. It is based on the assumption of linear mixing of latent source signals, which can be a potential limitation to its use for modelling non-linear interactions. However, as previously discussed, this is not an issue when analysing financial returns in an adaptive environment, as financial log-returns can be considered to be locally linear. It is possible to further extend this model by developing a non-linear ICA-based information coupling algorithm, as discussed in the future work section of this thesis (where we also discuss some other possible improvements and extensions to the model). In summary, if all constituent signals in a multivariate system are purely Gaussian with linear interactions, simply using a linear correlation approach will suffice. If the aim is simply to model non-linear interactions, a large enough data set is available, and computational efficiency is not a factor in consideration, then direct use of mutual information or copula-based methods may be suitable. However, as is the case in many practical financial applications, if the goal is to dynamically measure interactions between multivariate non-Gaussian data streams, in a computationally efficient framework, the ICA-based information coupling model is well suited and, as our analysis shows, can outperform other competing models. Having presented an in-depth theoretical, analytical and empirical analysis of measures of symmetric interaction in the last two chapters, we now move on to provide a detailed description and application of a set of asymmetric measures of interaction in the next two chapters.

Chapter 5
Asymmetric measures of interaction (causality)

We provided an introduction to the concept and theory of asymmetric interaction measurement (causality) in Chapter 2, which also included a brief review of some frequently used approaches for causal inference and provided a theoretical analysis of their relative merits and limitations. From our discussion in Chapter 2 it was evident that most standard causality detection approaches lack either the accuracy, robustness or computational efficiency required to dynamically analyse multivariate financial returns; this chapter aims to address this issue to some extent by presenting a set of asymmetric interaction measurement approaches. For most practical applications, a causality detection algorithm needs to be able to carry out online inference for it to be useful. This is because most real-world systems generate data which exhibits rapidly changing dynamics, often accompanied with variations in the underlying causal structures; this is especially true for financial returns [62, 65]. Rapidly changing properties of financial data may hence make previously discovered causal links to cease to exist or indeed become anti-causal. This is likely to happen even more frequently during periods of market turmoil and the accompanied high volatility, when many previously linked indices (signals) often get decoupled [153]; a process which has been observed during all major financial crises since the early twentieth century [226, 242]. This problem can be addressed by inferring the dynamics of causal interactions between multivariate time series using a sliding-window technique to compute a time-varying "causality matrix", with each element of the matrix representing the causal strength of any given link at any given time, and its variation with time giving us an idea of the temporal properties of causality. Hence, any causality detection approach used in practise (especially when used to analyse high-frequency sampled financial data) needs to be

well-suited to carry out online dynamic causal inference in multivariate data streams. In this chapter we present two such approaches, both of which are based on the principles of Granger causality; the first of these approaches is described below.

5.1 Granger independent component causality

The standard linear Granger causality model (which is one of the most commonly used causality detection approaches in practise) is based on a multivariate autoregressive (MAR) model, and infers causation between a set of signals if the linear predictability of one of them is improved by incorporating prior information about the others. The MAR model specifies the (linear) within-signal and cross-signal temporal dependencies in multivariate systems, and hence quantifies the informativeness of prior values of one (or more) observed variable on the current value of another [159]. However, a standard MAR model (with fixed parameters) is based on the assumption of data stationarity. Therefore, for our analysis we make use of MAR models which have time-varying parameters, in that we infer the model parameters within sliding-windows. Our choice of dynamically re-estimating parameters of the model is based on two primary reasons; firstly, within each window, financial log-returns can be considered to be locally stationary [211], and secondly, it allows us to capture the rapidly changing dynamics of financial markets. The dynamic-MAR model (for the bivariate case) is represented by the following set of equations; describing putative temporal causal relationships between time series $x(t)$ and $y(t)$:

$$x(t) = \sum_{j=1}^{p} \alpha_{11,j}(t)x(t-j) + \sum_{j=1}^{p} \alpha_{12,j}(t)y(t-j) + e_x(t) \tag{5.1}$$

$$y(t) = \sum_{j=1}^{p} \alpha_{21,j}(t)x(t-j) + \sum_{j=1}^{p} \alpha_{22,j}(t)y(t-j) + e_y(t) \tag{5.2}$$

where α's are real-valued regression parameters, p is the model order and $e(t)$ represent the regression error terms (residuals). As previously discussed, there are two separate (but related) tests that can be performed to infer the presence of a causal link; we can either make use of the weight parameters, for example, $Y \rightarrow X$ if values of α_{12} in (5.1) are significantly different from zero; or we can measure variance of the residuals, for example, $Y \rightarrow X$ if variance of $e_x(t)$ in (5.1) is significantly lower than the variance of residuals resulting from fitting an

autoregressive model (of the same order) to X only, i.e. if variance of one of the variables is reduced by including past terms from the other variable in the regression equation, then the second variable is said to Granger-cause the first one.

However, a Granger causality model making use of standard parameter estimation approaches, such as ordinary least squares, only makes use of second-order statistics and enforces the assumption of normally distributed regression residuals. Hence, it is unable to accurately infer causal relationships in non-Gaussian signals [28, 207]. This is a potential limitation, as many real data sets (in particular financial returns) have highly non-Gaussian distributions [98]. The Granger independent component (GIC) causality analysis approach presented in this chapter can overcome this limitation while maintaining the computational simplicity of a standard Granger causality approach. As many data sets can be considered as being generated by a linear, but non-Gaussian, causal process, the GIC approach makes use of ICA in combination with a dynamic-MAR model. It infers causation by taking into account the non-Gaussian nature of the data being analysed, by assuming the dynamic-MAR model residuals to be non-Gaussian and serially independent[1]. It is hence better suited to model asymmetric dependencies in multivariate signals with non-Gaussian distributions, such as financial returns.

5.1.1 Analytical framework

Equations (5.1) and (5.2) show a set of discrete-time dynamic-MAR model equations for a 2-dimensional system. Using (5.1), the regression error of $x(t)$ can be written as[2]:

$$e_x(t) = x(t) - \sum_{j=1}^{p} \alpha_{11,j}(t)x(t-j) - \sum_{j=1}^{p} \alpha_{12,j}(t)y(t-j) \tag{5.3}$$

Expanding the summation terms, $e_x(t)$ is rewritten as:

$$\begin{aligned} e_x(t) = x(t) &- \left[\alpha_{11,1}(t)x(t-1) + \alpha_{11,2}(t)x(t-2) + \ldots + \alpha_{11,p}(t)x(t-p)\right] \\ &- \left[\alpha_{12,1}(t)y(t-1) + \alpha_{12,2}(t)y(t-2) + \ldots + \alpha_{12,p}(t)y(t-p)\right] \end{aligned} \tag{5.4}$$

[1]The term "independent" here implies maximally statistically independent (using higher-order statistics), rather than only decorrelated (second-order independence). We achieve independence by using ICA.

[2]For brevity and clarity, we only present details for a bivariate GIC model in this chapter. However, it can be easily extended to high-dimensional spaces by following similar principles.

Using (5.4), the regression error terms at different time lags can be written in matrix form as:

$$\begin{bmatrix} e_x(t) \\ e_x(t-1) \\ e_x(t-2) \\ \dots \\ e_y(t-1) \\ e_y(t-2) \\ \dots \end{bmatrix} = \begin{bmatrix} 1 & -\alpha_{11,1}(t) & -\alpha_{11,2}(t) & \dots & -\alpha_{12,1}(t) & -\alpha_{12,2}(t) & \dots \\ 0 & 1 & -\alpha_{11,2}(t-1) & \dots & 0 & -\alpha_{12,2}(t-1) & \dots \\ 0 & 0 & 1 & \dots & 0 & 0 & \dots \\ & & & \dots & & & \\ 0 & 0 & -\alpha_{21,2}(t-1) & \dots & 1 & -\alpha_{22,2}(t-1) & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 & \dots \\ & & & \dots & & & \end{bmatrix} \begin{bmatrix} x(t) \\ x(t-1) \\ x(t-2) \\ \dots \\ y(t-1) \\ y(t-2) \\ \dots \end{bmatrix} \tag{5.5}$$

For a sliding-window of size T, we can represent the set of lagged regression error terms (at times $t-T:t$) by a $(pd+1)\times T$ dimensional matrix $\mathbf{E}_t$, the $(pd+1)$ dimensional square matrix representing the weights of the dynamic-MAR model at time t by $\mathbf{\Omega}_t$, and the data set representing the lagged values of the observed signals by a $(pd+1)\times T$ dimensional matrix $\mathbf{Z}_t$. Hence, (5.5) can be written as:

$$\mathbf{E}_t = \mathbf{\Omega}_t \mathbf{Z}_t \tag{5.6}$$

Similarly, for a dynamic univariate autoregressive (dynamic-AR) model, the residuals (in terms of lagged values of the observed variable) are given by:

$$e'_x(t) = x(t) - \sum_{j=1}^{p} \alpha'_{11,j}(t) x(t-j) \tag{5.7}$$

Again, these can be represented in matrix form as:

$$\begin{bmatrix} e'_x(t) \\ e'_x(t-1) \\ e'_x(t-2) \\ \dots \end{bmatrix} = \begin{bmatrix} 1 & -\alpha'_{11,1}(t) & -\alpha'_{11,2}(t) & \dots \\ 0 & 1 & -\alpha'_{11,2}(t-1) & \dots \\ 0 & 0 & 1 & \dots \\ & \dots & \dots & \end{bmatrix} \begin{bmatrix} x(t) \\ x(t-1) \\ x(t-2) \\ \dots \end{bmatrix} \tag{5.8}$$

Likewise, representing the set of lagged regression error terms of the dynamic-AR model (at times $t-T:t$) by a $(p+1)\times T$ dimensional matrix $\mathbf{E}'_t$, the $(p+1)$ dimensional square matrix representing the weights of the dynamic-AR model by $\mathbf{\Omega}'_t$, and the data set representing the lagged values of the observed signal by a $(p+1)\times T$ dimensional matrix $\mathbf{Z}'_t$, (5.8) can be written as:

$$\mathbf{E}'_t = \mathbf{\Omega}'_t \mathbf{Z}'_t \tag{5.9}$$

For multivariate financial returns, the regression residuals can be assumed to be non-Gaussian and serially independent[3] [12, 98, 336]. Under these conditions, (5.6) and (5.9) may be recast as an ICA model with $\mathbf{Z}_t$ (and $\mathbf{Z}'_t$) representing the sets of observed variables, $\mathbf{\Omega}_t$ (and $\mathbf{\Omega}'_t$) representing the unmixing matrices and $\mathbf{E}_t$ (and $\mathbf{E}'_t$) representing the sets of latent independent components. Hence, we can exploit the signal processing power of ICA to obtain estimates for the matrices $\mathbf{\Omega}_t$ and $\mathbf{\Omega}'_t$; this can be achieved by using the *icadec* algorithm (as previously presented).

5.1.2 Causal inference

The unmixing matrices $\mathbf{\Omega}$ and $\mathbf{\Omega}'$ obtained using ICA contain information about the causal structures (both within-signal and cross-signal) present in the signals being analysed. As before, noting that multiplication of the ICA unmixing matrix by a diagonal matrix does not effect the mutual information of the recovered sources, we row-normalise these unmixing matrices in order to address the ICA scale indeterminacy problem. Row-normalisation implies that the elements ω_{ij} of the unmixing matrices are constrained, such that the vectors, $\boldsymbol{\omega}_i = [\omega_{i1}, ..., \omega_{iN}]$, containing elements of the i-th row of the matrix are of unit length, i.e.:

$$\boldsymbol{\omega}_i \boldsymbol{\omega}_i^{\mathsf{T}} = 1 \tag{5.10}$$

We can solve the ICA row permutation problem by noting that the unmixing matrices in our model are structured such that their off-diagonal elements are smaller in magnitude than the diagonal elements (as $\mid \alpha \mid < 1$), i.e. the largest element (by magnitude) associated with the i-th row of the matrices lies in column i; in this way we are able to determine the ordering of rows in $\mathbf{\Omega}$ (and $\mathbf{\Omega}'$). We address the ICA sign ambiguity problem by making use of symmetric statistics for causal inference (as described later).

For a set of observed signals which are serially as well as mutually decoupled, i.e. all the constituent vectors in $\mathbf{Z}$ (or $\mathbf{Z}'$) are independent, the latent independent components will be the same as the observed signals; hence, the row-normalised unmixing matrix $\mathbf{\Omega}$ (or $\mathbf{\Omega}'$) will be a permutation of the identity matrix. Any temporal dependence (either within-signal or cross-signal) will be captured in the elements of the unmixing matrix and subsequently in the

[3] This assumption holds for many real-world systems which give rise to non-Gaussian signals and can generally be modelled using a MAR model, such as signals in biomedicine [145] and climatology [254].

matrix of residuals $\mathbf{E}$ (or $\mathbf{E}'$); hence, making it possible to identify causal links in the system, as described below.

Measuring causal strength

Following the principles of Granger causality, we can say that variable Y causes X, i.e. $Y \rightarrow X$, if the regression error of X is reduced by incorporating prior terms from Y. Therefore, to measure the strength of a causal link between variables Y and X, we need to measure the relative magnitude of the residuals obtained using the two models we are comparing. Given a T data points long data set (corresponding to the length of the sliding-window in a dynamic environment), we denote the vectors of residuals (corresponding to time t) for the two models by:

$$\mathbf{u} = \boldsymbol{\omega}_1 \mathbf{Z}, \quad \mathbf{u}' = \boldsymbol{\omega}_1' \mathbf{Z}' \tag{5.11}$$

where $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_1'$ are the first rows of the row-normalised, permutation-adjusted, unmixing matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}'$ respectively. To measure the causal strength, we compute the log-likelihood ratio by comparing the negative log-likelihood loss functions for the two models, given by:

$$g(\mathbf{u}) = -\sum_k \log p(u_k), \quad g(\mathbf{u}') = -\sum_k \log p(u_k') \tag{5.12}$$

where u_k and u_k' are the k-th elements of the vectors $\mathbf{u}$ and $\mathbf{u}'$ respectively and $g(\cdot)$ represents the loss function. Hence, the log-likelihood ratio $(g(\mathbf{u}') - g(\mathbf{u}))$ is given by:

$$\zeta = -\sum_k \log p(u_k') + \sum_k \log p(u_k) \tag{5.13}$$

A standard linear Granger model assumes the residuals to be normally distributed; hence, when using normalised data, $\zeta = \frac{1}{2}\sum_k u_k'^2 - \frac{1}{2}\sum_k u_k^2$, which gives us the commonly used sum of squares error function. For the GIC model, we implicitly assume the distribution of residuals to be heavy-tailed, therefore, we need to use alternate loss functions which can approximately model the non-Gaussian pdfs of the residuals. A good choice for such a loss function is based on the heavy-tailed reciprocal cosh distribution (as given by (3.14)) which we use as a prior source model in our ICA algorithm; this gives us the following measure of causal strength for

the $Y \rightarrow X$ causal link[4]:

$$\zeta(y,x) = \sum_k \log\cosh(u'_k) - \sum_k \log\cosh(u_k) \tag{5.14}$$

A generic representation of the loss function based on the reciprocal cosh distribution is presented in Figure 5.1. Also included in the figure are the loss functions associated with the normal and Laplace (which is simply the sum of absolute errors) distributions. We notice that the loss functions based on the reciprocal cosh and Laplace distributions are much more robust to outliers as compared to the loss function based on the normal distribution. We also note that the reciprocal cosh distribution based loss function locally ($\mid u \mid < 0.5$) mimics the normal distribution based loss function while globally ($\mid u \mid \geq 0.5$) mimics the Laplace distribution based loss function; hence, it is robust to outliers as well as locally quadratic.

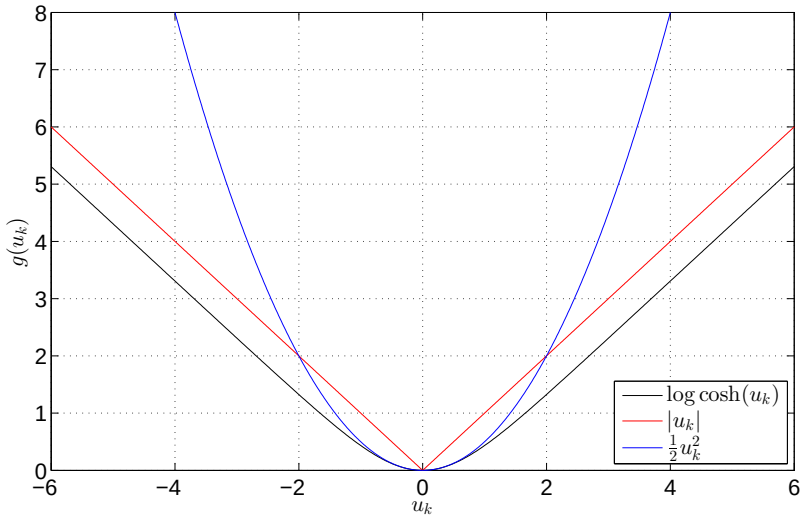

Fig. 5.1: Loss functions ($g(u_k)$) based on the reciprocal cosh distribution ($\log\cosh(u_k)$), the Laplace distribution ($\mid u_k \mid$) and the normal distribution ($\frac{1}{2}u_k^2$); u_k refers to the k-th element of the vector of residuals $\mathbf{u}$.

The correct (optimal) model order can be selected by minimising the Bayesian information criterion (BIC) as a function of the model order [343]. The BIC is applicable to most types of

[4]For certain applications, it may be useful to estimate the uncertainty associated with our measure of causal strength (ζ). This can be done using a similar approach as described previously (for estimating confidence intervals on the information coupling measure), i.e. by drawing samples of the skew-symmetric matrix ($\mathbf{J}'$) from a multivariate normal with a mean and covariance given by $\mathbf{J}$ and $\mathbf{H}^{-1}$ respectively (as presented in (3.43)), and transforming them into samples in $\mathbf{\Omega}$ using (3.13) and (3.12) respectively; these samples can then be used to obtain a set of values for ζ using the analysis presented earlier, hence, enabling us to obtain confidence bounds on the measure of causal strength.

likelihood functions, making it suitable for a wide variety of model order selection problems [67]. It is quite likely that for real data the BIC may decrease monotonically over the range of model orders considered, e.g. it may asymptotically reach a minimum. In such a scenario, we select the model order for which the reduction in BIC has reached 90% of the maximum possible reduction over the range of model orders considered; this approach ensures that we only use those model orders which result in significant information gain.

Statistical hypothesis testing

The statistic $\zeta(y,x)$ can be used to detect the presence of causality as well as to measure the causal strength. The null hypothesis (H_0) of no causality and the alternate hypothesis (H_1) of the presence of causation ($Y \rightarrow X$) are given by:

$$H_0 : \zeta(y,x) \leq 0, \quad H_1 : \zeta(y,x) > 0 \tag{5.15}$$

A value of $\zeta(y,x) > 0$ indicates that the $Y \rightarrow X$ causal link exists, and similarly if $\zeta(x,y) > 0$, then the $X \rightarrow Y$ link is likely to be causal; hence, we can easily infer the direction of information flow in the system. In practise, we need to test for the significance of these links by rejecting the null hypothesis only if the causal strength statistic (ζ) is greater than (or equal to) a critical value (ζ_c), i.e. a link is causal only if $\zeta \geq \zeta_c$. This critical value will always be positive, and the exact value can be set depending on the application and requirements of the user[5]. For a standard linear Granger model, we can make use of various standard hypothesis testing approaches based on comparison of the means or variances of the data (residuals or weights as appropriate), such as the F-test or the chi-square test [344]. However, these (and other similar tests) are based on the assumption that the samples are drawn from a known (usually the normal) distribution and are remarkably sensitive to any departure from this condition [52]. For the GIC model, we make use of one of two suitable hypothesis testing approaches (depending on the application), which are described below.

The first approach directly makes use of the causal strength statistic (ζ), and infers the presence of causation by finding its critical value (ζ_c) at a given significance level. As no fixed analytical distribution of ζ under the null hypothesis exists, therefore, for hypothesis

[5]For certain applications, it is possible to set the critical value using a training data set; we empirically demonstrate this approach later in the thesis.

testing we use a method which finds its empirical distribution under the null hypothesis using a bootstrapping approach [277, 339]. We now describe this four-step approach when testing for the presence of the $Y \rightarrow X$ causal link. In the first step, we generate an ensemble of surrogate data sets (e.g. 100) by random temporal realignment of the original time series $x(t)$ to obtain $\bar{x_q}(t)$, where $q = [1:100]$ indexes each realigned time series; the bivariate surrogate data sets $[Y, \bar{X_q}]$ thus generated will have similar statistical properties to the actual data being analysed but satisfy the null hypothesis of no causality. In the second step, we compute the statistic $\zeta(y, \bar{x_q})$, i.e. the strength of the $Y \rightarrow \bar{X_q}$ causal link for each of the surrogate data sets. In the third step, we calculate ζ_c, the γ-th quantile (we use γ=0.95) of the bootstrapping distribution of $\zeta(y, \bar{x_q})$; the value of γ can be set such that the false-positive detection rate (type I error) under the null hypothesis is below a pre-defined threshold value[6]. Finally, in the fourth step, we test for the presence of causation by rejecting the null hypothesis if $\zeta(y,x) \geq \zeta_c$. Although accurate, this approach can be computationally demanding due to the bootstrapping procedure.

This brings us to our second approach that makes use of the Mann-Whitney U-test (also called the Wilcoxon rank-sum test), which is a non-parametric test based on comparing the medians of two populations. As the U-test is based on rank statistics, therefore, it is suitable for use with non-Gaussian data [124]; hence, making it an attractive choice for the GIC model. For our analysis, we use the U-test to compare the distributions of the absolute values of the residuals, i.e. $\mid \mathbf{u} \mid$ and $\mid \mathbf{u}' \mid$, and reject the null hypothesis of no causality if the median of $\mid \mathbf{u} \mid$ is smaller than that of $\mid \mathbf{u}' \mid$ at the 5% significance level. Although efficient, the U-test is most appropriate when underlying distributions from which samples are drawn have the same (or very similar) second and higher-order moments; its performance can degrade if this condition is not met [125]. This is because being a rank-based approach, it ignores any other structure in the data, i.e. it does not fully utilise the information contained in the distributions. Therefore, it is not always suitable for use with the GIC model, for which this condition may not hold. We empirically compared the performance of the U-test with the surrogate data based approach using synthetic data with properties similar to those of financial returns. Our results indicate that the surrogate data based approach gives better results as compared to the U-test, however,

[6] A higher value of ζ_c will result in the model detecting fewer causal links with low uncertainty in the presence of causality and vice versa. All causality detection approaches face this trade-off between the number of causal links detected and the uncertainty associated with the results.

it can be computationally demanding (being based on a bootstrapping approach) and thus well-suited for online analysis of medium or low frequency data; when analysing high-frequency (e.g. sub-second sampled) data, the U-test may be the better option, due to its comparative efficiency.

5.1.3 Summary of the algorithm

We now recap the main steps to be taken when testing for causation using the GIC causal inference algorithm. The steps listed below are valid for a bivariate system, however, they can easily be extended to model causality in multivariate signals.

1. Let $x(t)$ and $y(t)$ be two time series between which we want to infer the presence or absence of the $Y \rightarrow X$ causal link. Arrange the time series data into two matrices, $\mathbf{Z}_t = [x(t),...,x(t-p),y(t-1),...,y(t-p)]^{\mathsf{T}}_{t=\{t-T:t\}}$ and $\mathbf{Z}'_t = [x(t),...,x(t-p)]^{\mathsf{T}}_{t=\{t-T:t\}}$, where T is the length of the sliding-window and p is the model order.

2. Normalise all constituent signals of $\mathbf{Z}_t$ and $\mathbf{Z}'_t$; this can easily be done by removing the mean and dividing by their standard deviation values.

3. Infer the unmixing matrices $\mathbf{\Omega}_t$ and $\mathbf{\Omega}'_t$ by using the *icadec* algorithm to decompose the matrices $\mathbf{Z}_t$ and $\mathbf{Z}'_t$ respectively.

4. Row-normalise the unmixing matrices using (5.10). Select $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}'_1$, the rows of these row-normalised matrices which have the largest element (by magnitude) associated with their first column.

5. Estimate the vectors of residuals, $\mathbf{u}$ and $\mathbf{u}'$, using (5.11).

6. Calculate a value for $\zeta(y,x)$ (the causal strength of the $Y \rightarrow X$ causal link) using (5.14) (the correct model order can be estimated using the BIC).

7. Estimate a value for ζ_c (the critical value of the causal strength statistic at a significance level of γ) using either the bootstrapping based surrogate data approach or the U-test (or both). Reject the null hypothesis of no causality (i.e. the $Y \rightarrow X$ link is causal at a significance level of γ) if $\zeta(y,x) \geq \zeta_c$.

Discussion

The standard Granger causality model uses least squares approximations to obtain estimates for the MAR model parameters, i.e. it only uses second-order statistics and is therefore only suitable for analysing Gaussian signals [28]. On the other hand, the GIC model (by making use of ICA) implicitly makes use of higher-order statistics to obtain an estimate for the model parameters and can therefore capture information in the tails of a distribution. The GIC approach is well-suited to efficiently analyse asymmetric dependencies in non-Gaussian data streams, making it especially relevant for use with financial returns, which have non-Gaussian (heavy-tailed) distributions [98, 258]. One of the main financial sector applications of causality detection approaches is financial time series forecasting. Most of these forecasting models deal with high-frequency data (often sampled at millisecond intervals), and therefore, to minimise the effect of latencies in the system, require the use of very efficient algorithms. The GIC algorithm achieves computational efficiency by making use of the sliding-window based decorrelating manifold approach to ICA, and reciprocal cosh source models[7]. This makes the GIC algorithm suitable for analysing high-frequency sampled financial data (as we empirically demonstrate later). There has been some previous work dealing with causality detection approaches using ICA [323, 324]. However, there are some key differences with the model we have presented. Other work has focused on inferring causal order (direction of causation) in a set of temporally aligned variables. The methods were based on inference in directed acyclic graphs and were inefficient in higher-dimensions. By making the assumption of serial independence in the regression residuals, we are able to base our approach on the principles of Granger causality, exploiting ICA to deal with non-Gaussianity. Moreover, our computationally efficient model is well-suited for online analysis of financial data streams and also provides us with a measure of significance of causal links.

5.1.4 Independent component autoregressive model

Autoregressive (AR) models can have two broad goals, i.e. they can either be used to infer the causal structure in multivariate systems (as in the GIC model) or to improve the predictability of a single variable. We now present an ICA-based approach for univariate prediction of

[7] As an example, the GIC algorithm takes on average 34 ms (averaged over 100 simulations, on a 2.66 GHz processor) for analysing 1000 data points long bivariate signals at $p = 3$.

non-Gaussian signals by estimating the parameters of an AR model. We call this approach (which is a variant of the GIC model) the ICA-AR sequential prediction model. Developing efficient approaches for accurate forecasting of financial time series is one of the primary goals of many market practitioners and academics alike. There is a considerable amount of literature available which demonstrates the application of various forecasting approaches across different asset classes. For example, [134] presents a detailed description of using neural networks for financial forecasting, in [63, 64] the authors present application of support vector machines for forecasting purposes, while [235] presents the use of genetic algorithms for financial data forecasting. However, most of these (and other similar) forecasting approaches have certain limitations with respect to their computational complexity and the need for optimising various parameters using training data; these limitations make most of these approaches unsuitable for real-time dynamic analysis of financial data, especially data sampled at high-frequencies. The studies mentioned above (as well as other similar studies) show that different financial time series have different levels of predictability. In stocks, there is evidence of short-horizon predictability but very little evidence of any statistically significant long-horizon predictability [16]. There is also evidence of presence of long-horizon predictable components in FX returns [239]; long-horizon models usually make use of secondary data, e.g. interest rates, consumer confidence or dividend yields in case of stocks. Predictability of volatility of financial instruments is also an area of relevance; studies have shown that exchange rate volatility predictability is statistically more significant for shorter time horizons [301]. An interesting overview of the predictability of high-frequency financial time series is presented in [98].

The autoregressive (AR) model is one of the most commonly used forecasting approaches in practise, mainly due to its simplicity and computational efficiency. Statistical models derived from the AR model, e.g. AR moving average (ARMA) models, AR integrated moving average (ARIMA) models, Generalised AR conditional heteroskedasticity (GARCH) models [46], etc. are frequently used by financial practitioners. A dynamic-AR model (which has time-varying parameters) is a simple univariate linear regression process which can be used to forecast the current value of a time series at time t, $x(t)$, using its past p values, i.e.[8]:

[8]Usually, AR models are used for obtaining one-step ahead prediction values in most practical applications. Using an AR model, it is also possible to do multi-step ahead predictions. However, in general the more steps ahead the predicted value is, the worse the results will be [191].

$$x(t) = \sum_{j=1}^{p} \alpha_j(t)x(t-j) + e(t) \tag{5.16}$$

where α_j are real-valued parameters (weights) of the model and $e(t)$ is the regression residual. There are various approaches to estimate the AR model parameters, e.g. the Yule-Walker approach, ordinary least squares (OLS), geometric lattice approach, forward-backward approach and Burg's lattice based method, etc. Further details of each of these methods are presented in [328]. The most commonly used of these methods is the OLS approach which estimates the values of the AR model parameters by minimising the variance of the regression error terms.

However, standard AR models suffer from many limitations. These include the assumption of Gaussian residuals and sensitivity to the presence of outliers [220]. These limitations can severely effect the accuracy of standard second-order statistics based AR models, especially when used to analyse financial returns, which have non-Gaussian pdfs with heavy-tails [98]. The ICA-AR sequential forecasting approach can overcome this limitation while maintaining the computational simplicity of a standard AR approach. It estimates the parameters of the AR process by taking into account the non-Gaussian nature of the data being analysed, by assuming the AR model residuals to be non-Gaussian and serially independent. Hence, it is better suited for carrying out predictive analysis in non-Gaussian dynamic environments. As shown earlier by (5.7), (5.8), (5.9) and (5.10), it is possible to make use of ICA to obtain a row-normalised vector $\boldsymbol{\omega}_1$ (corresponding to the first row of the unmixing matrix $\boldsymbol{\Omega}$) which contains information about the AR model parameters. In order to simultaneously address the ICA sign ambiguity problem and to compensate for the effect of row-normalisation on values of the AR model parameters, we divide the elements of $\boldsymbol{\omega}_1$ by $\boldsymbol{\omega}_1(1)$ (i.e. the first element of the vector $\boldsymbol{\omega}_1$), such that the results obtained are directly comparable to the analytical model presented in (5.8). Hence, we can obtain estimates for the AR model parameters at time-lag j as follows[9]:

[9] If required, we can also estimate the uncertainty associated with our forecasts by using the procedure described earlier, i.e. by drawing samples (e.g. 100) of a skew-symmetric matrix $\mathbf{J}'$ from a multivariate normal (as presented in (3.43)) with the mean and covariance given by $\mathbf{J}$ and $\mathbf{H}^{-1}$ respectively. These samples can be readily transformed to samples in $\boldsymbol{\Omega}_t$ using (3.13), (3.12) and (5.9) respectively. We can then proceed to obtain a set of values for α_j using the analysis presented earlier. Confidence bounds (e.g. the 95% bounds) may then be easily obtained from these samples for α_j, which give us an indication of the uncertainty associated with our forecasts.

$$\alpha_j = -\frac{\boldsymbol{\omega}_1(j+1)}{\boldsymbol{\omega}_1(1)} \tag{5.17}$$

The model order of the AR process can be estimated using the BIC.

Data used for benchmarking the performance of a prediction model can be either in-sample or out-of-sample. Test data is referred to as being in-sample if it is the same data that was used to determine the parameters of the model during the model training phase. On the other hand, an out-of-sample test involves using data which has not been used for parameter estimation. As the parameters are trained using the same data, therefore, in-sample prediction tests usually perform better than out-of-sample ones. In-sample tests are unreliable in the presence of even slight changes in the underlying structure of the data [192]. Therefore, it is important to test the accuracy of any prediction model using out-of-sample data in order to get more realistic results (as we do later in this thesis). At higher frequencies, the variation of financial data decreases within a window of any given size (in terms of data points). It is common for the value of a high-frequency financial mid-price to remain constant (i.e. for log-return values to remain at zero) for multiple time epochs; an effect that was illustrated earlier in Figure 4.2. Recently, some financial models have been making use of data sampled at ultra high-frequencies, e.g. milliseconds [9, 109]. Consider a statistical model that acquires financial data in real-time for a specific financial instrument at a high sampling frequency of say 0.25 seconds. It is very likely that values of the data will stay constant for multiple data points. This presents us with a new problem in terms of testing for the accuracy of a prediction model, as standard error measures, such as root mean square (rms) or normalised mean square (nms) will give significantly lower values when analysing high-frequency data in comparison to mid- or low- frequency sampled data, which can give us an erroneous indication of a model's performance. To address this issue, we present a modified rms error term as follows:

$$\hat{e}_{rms} = \sqrt{\frac{\sum_{t=1}^{T} [e(t)]^2}{T - \sum_{t=1}^{T} \delta\{x(t)x(t-1)\}}} \tag{5.18}$$

where $\delta\{x(t)x(t-1)\}$ is the Kronecker delta function given by:

$$\delta\{x(t)x(t-1)\} = \begin{bmatrix} 1 & \text{if} & x(t) = x(t-1) \\ 0 & \text{if} & x(t) \neq x(t-1) \end{bmatrix} \tag{5.19}$$

Hence, this measure calculates the rms error based on only those predicted values for which there is a change in the value of the observed time series, resulting in a more realistic (and easily comparable) model performance measure.

5.2 Variational Granger causality

The second causality detection approach we present in this chapter is based on the principles of a variational Bayesian MAR (VB-MAR) model [291] and a standard Granger causality model [147]. We call this the variational Granger (VG) approach to causality detection. Bayesian inference within a VB setting can result in significant computational benefits (as previously discussed). The VB-MAR approach can therefore be used to estimate parameters of a MAR model within a Bayesian framework, without incurring the large computational costs typically associated with standard Bayesian approaches[10]. Causal inference within a VB setting has multiple benefits, which will become clear as we go through this section, which is organised as follows. We first present an introduction to VB-MAR models. Next, we describe the methods we use to analyse Granger causality across both time- and frequency-domains using parameters estimated via the VB-MAR models. We conclude this section by discussing the advantages and benefits of using the VG approach. In the next chapter, we present empirical results, using both synthetic and financial data, which demonstrate the utility and accuracy of our approach for extracting interesting information from multivariate financial time series.

5.2.1 Variational Bayesian multivariate autoregressive models

Multivariate autoregressive (MAR) models (a specific type of general linear models) causally quantify the linear dependence of one time series on all the others in a multivariate system. Let $\mathbf{y}(t) = [y_1(t), y_2(t), ..., y_d(t)]$ be the t-th sample of a d-dimensional time series. Then,

[10] Bayesian inference can be computationally expensive [326]. Using VB techniques, it is possible to get a parametric approximation for the true posterior density of an intractable distribution by using an approximate distribution for which the required inferences are tractable [218]. The approximate posterior distribution can be obtained by finding a distribution, such that it minimises the Kullback-Leibler (KL) divergence between this distribution and the actual posterior distribution, as we describe in detail later.

using (5.1) and (5.2), we may present a d-dimensional MAR model as follows:

$$\mathbf{y}(t) = \sum_{m=1}^{p} \mathbf{y}(t-m)\mathbf{A}(m) + \mathbf{e}(t) \tag{5.20}$$

where $\mathbf{A}(m)$ is a d-dimensional square matrix of coefficients (weights) of the model and $\mathbf{e}(t) = [e_1(t), e_2(t), ..., e_d(t)]$ is the vector of residuals at time t. Now let $\mathbf{x}(t) = [\mathbf{y}(t-1), \mathbf{y}(t-2), ..., \mathbf{y}(t-p)]$ be the p previous multivariate time series samples. If $\mathbf{y}(t)$ and $\mathbf{x}(t)$ are the t-th rows of the matrices $\mathbf{Y}$ and $\mathbf{X}$ respectively, then, for a T data points long data set, the MAR model can be written in matrix form as:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{E} \tag{5.21}$$

where $\mathbf{W}$ is a multi-dimensional MAR model coefficient matrix of dimensions $(p \times d) \times d$ and $\mathbf{E}$ is a matrix of residuals (which we assume to be normally distributed). A wide range of multivariate linear regression models can be expressed in the form of (5.21); many of these models assume the likelihood of the data, $\mathbf{D} = \{\mathbf{X}, \mathbf{Y}\}$, to be given by [53, 291]:

$$p(\mathbf{D} \mid \mathbf{W}, \mathbf{\Lambda}) = (2\pi)^{\frac{-dT}{2}} |\mathbf{\Lambda}|^{\frac{T}{2}} \exp\left[-\frac{1}{2}\mathrm{Tr}(\mathbf{\Lambda}\mathbf{E}_{\mathbf{D}}(\mathbf{W}))\right] \tag{5.22}$$

where $\mathrm{Tr}(\cdot)$ is the trace of the matrix, $\mathbf{\Lambda}$ is the precision (inverse covariance) matrix of the regression error terms, $|\cdot|$ denotes the determinant of the matrix and $\mathbf{E}_{\mathbf{D}}(\mathbf{W})$ denotes the error covariance matrix, given by:

$$\mathbf{E}_{\mathbf{D}}(\mathbf{W}) = (\mathbf{Y} - \mathbf{X}\mathbf{W})^{\mathsf{T}}(\mathbf{Y} - \mathbf{X}\mathbf{W}) \tag{5.23}$$

Maximum likelihood estimation

Let us first describe the standard, maximum likelihood (ML) based, approach to MAR modelling (we use the ML parameters to initialise the Bayesian MAR algorithm, which we describe later). The ML solution to the MAR model coefficient matrix (for normally distributed residuals) is given by [291]:

$$\mathbf{W}_{ML} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y} \tag{5.24}$$

i.e. the coefficients are obtained by using pseudo-inverse of the matrix $\mathbf{X}$. The ML noise covariance can then be estimated as [291]:

$$\mathbf{S}_{ML} = \frac{1}{T-k}\mathbf{E}_{\mathbf{D}}(\mathbf{W}_{ML}) \tag{5.25}$$

where $k = pd^2$ is the total number of weight coefficients in the model. For simplifying relevant analysis presented later in this section, we introduce the notation $\mathbf{w}_{ML} = \text{vec}(\mathbf{W}_{ML})$, which implies that the columns of $\mathbf{W}_{ML}$ are stacked on top of each other to form a vector of length k. The ML parameter covariance matrix for $\mathbf{w}_{ML}$ is then given by [291]:

$$\mathbf{\Sigma}_{ML} = \mathbf{S}_{ML} \otimes \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \tag{5.26}$$

where $\otimes$ denotes the Kronecker product. The ML approach, however, suffers from the model overfitting problem [291, 303]; this can be addressed to some extent by using a standard model order selection criterion (such as the Bayesian or the Akaike information criteria) to estimate an optimum value of the model order [59]. For accurate estimation, ML approaches require large amounts of stationary data to reliably fit the $\mathcal{O}(k)$ parameters [291]; however, large data sets are not always available in practise and may be unsuitable for applications focused on capturing the dynamically evolving structure of interactions in a system. This problem can be addressed to some extent by introducing priors over the weights (in order to regularise coefficient magnitudes) and marginalising over the posterior distributions (hence taking into account the intrinsic uncertainty) of the variables to estimate the model parameters, i.e. by utilising the Bayesian framework for data analysis [291, 303], as described below.

Bayesian framework

In the Bayesian MAR model presented in this section, we assume the weights are drawn from a zero-mean Gaussian prior having an isotropic covariance with precision ς [291]:

$$p(\mathbf{w} \mid \varsigma) = \left(\frac{\varsigma}{2\pi}\right)^{\frac{k}{2}} \exp\left[-\varsigma E(\mathbf{w})\right] \tag{5.27}$$

where $E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w}$. The weight precision parameter, ς, is itself drawn from a Gamma prior:

$$p(\varsigma) = \text{Ga}(\varsigma; b_{\varsigma}, c_{\varsigma}) \tag{5.28}$$

where b_ς and c_ς are hyperparameters of the distribution. Similarly, the noise precision matrix has the prior:

$$p(\mathbf{\Lambda}) = |\mathbf{\Lambda}|^{-\frac{d+1}{2}} \tag{5.29}$$

which is the "uninformative prior" for multivariate linear regression, as discussed in [291]. These parameters can be represented by a single set, which we denote $\boldsymbol{\theta} = \{\mathbf{w}, \varsigma, \mathbf{\Lambda}\}$; writing the prior parameters in this form simplifies the subsequent analysis. As the weights depend on ς (but not on $\mathbf{\Lambda}$), the joint distribution over $\boldsymbol{\theta}$ factorises as [291]:

$$p(\boldsymbol{\theta}) = p(\mathbf{w} \mid \varsigma)p(\varsigma)p(\mathbf{\Lambda}) \tag{5.30}$$

The subsequent inference can now be described in terms of learning the model parameters $\boldsymbol{\theta}$, given the data $\mathbf{D}$, i.e. in terms of the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{D})$, which fully describes our knowledge about the model parameters. We can write the marginal likelihood (evidence) of data, given the set of model parameters, as:

$$p(\mathbf{D}) = \int p(\mathbf{D}, \boldsymbol{\theta}) d\boldsymbol{\theta} \tag{5.31}$$

In the above equation, the integral over the joint probability is intractable [326]. It is possible to estimate this marginal integral using various stochastic Markov chain Monte Carlo (MCMC) techniques, however, MCMC methods can be computationally demanding and have convergence problems [15]. We consider, therefore, the use of approximate Bayes methods; these make tractable approximations to the posterior distribution over all model parameters, providing efficient computation even for large, high-dimensional, data sets. Various approximate Bayes approaches exist, such as the Laplace approximation or expectation propagation, however, (due to reasons discussed previously) these approaches can give misleading results in practise. Therefore, we make use of the variational Bayes framework for Bayesian approximations [169], as described below.

Variational Bayesian inference

Here we present a brief overview of using the variational Bayes (VB) approach for MAR modelling. The main steps involved in estimating the parameters of the VB-MAR model are

presented in Appendix B; a more in-depth description of the model is contained in [291]. Let $q(\boldsymbol{\theta} \mid \mathbf{D})$ be a tractable posterior proposal density, then, using (5.31), we can write the log evidence as:

$$\log p(\mathbf{D}) = \log \int q(\boldsymbol{\theta} \mid \mathbf{D}) \frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \mathbf{D})} d\boldsymbol{\theta} \tag{5.32}$$

Making use of Jensen's inequality [214], we can infer that:

$$\log \int q(\boldsymbol{\theta} \mid \mathbf{D}) \frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \mathbf{D})} d\boldsymbol{\theta} \geq \int q(\boldsymbol{\theta} \mid \mathbf{D}) \log \frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \mathbf{D})} d\boldsymbol{\theta} \tag{5.33}$$

Noting that the right-hand-side term of this inequality is equivalent to the *negative variational free energy*, i.e. [130]:

$$F(p) = \int q(\boldsymbol{\theta} \mid \mathbf{D}) \log \frac{p(\mathbf{D}, \boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \mathbf{D})} d\boldsymbol{\theta} \tag{5.34}$$

we can infer (by comparing (5.32), (5.33) and (5.34)) that:

$$\log p(\mathbf{D}) \geq F(p) \tag{5.35}$$

$F(p)$ therefore acts as a strict lower bound to the log evidence with equality only if $q(\boldsymbol{\theta} \mid \mathbf{D}) = p(\boldsymbol{\theta} \mid \mathbf{D})$. Using Bayes' rule, $p(\mathbf{D}, \boldsymbol{\theta}) = p(\mathbf{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$, the negative free energy term (as given by (5.34)) can then be expanded as:

$$F(p) = \int q(\boldsymbol{\theta} \mid \mathbf{D}) \log p(\mathbf{D} \mid \boldsymbol{\theta}) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta} \mid \mathbf{D}) \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta} \mid \mathbf{D})} d\boldsymbol{\theta} \tag{5.36}$$

Noting that the first term on the right-hand-side of this equation is equivalent to the average log-likelihood of the data, i.e.:

$$L_{av} = \int q(\boldsymbol{\theta} \mid \mathbf{D}) \log p(\mathbf{D} \mid \boldsymbol{\theta}) d\boldsymbol{\theta} \tag{5.37}$$

and that that last term is equal to the negative KL divergence (which acts as a penalty term by penalizing more complex models) between the two distributions, we can rewrite (5.36) as:

$$F(p) = L_{av} - KL[q \,||\, p] \tag{5.38}$$

This is the fundamental equation of the VB framework, with the primary aim of VB learning being maximisation of $F(p)$ by successive iterations over $\boldsymbol{\theta}$. To make the subsequent analysis easier, we impose the following factorisation on the approximating posterior proposal density [291]:

$$q(\boldsymbol{\theta} \mid \mathbf{D}) = q(\mathbf{w} \mid \mathbf{D})q(\varsigma \mid \mathbf{D})q(\boldsymbol{\Lambda} \mid \mathbf{D}) \tag{5.39}$$

The parameters of the VB-MAR model can now be iteratively updated by maximising the negative free energy with respect to each of the parameters[11]. We present details of the parameter update procedure in Appendix B; here we only present a set of update equations which are relevant immediately. For detecting Granger causality (as described later) we require estimates for the posterior model weights ($\hat{\mathbf{w}}$) and the covariance matrix $\left(\hat{\boldsymbol{\Sigma}}\right)$. Defining $\hat{\varsigma}$ and $\hat{\boldsymbol{\Lambda}}$ as the weight and noise precisions respectively, we define $\boldsymbol{\Lambda}_D = \hat{\boldsymbol{\Lambda}} \otimes (\mathbf{X}^{\mathsf{T}}\mathbf{X})$, and hence write the posterior covariance matrix as:

$$\hat{\boldsymbol{\Sigma}} = (\boldsymbol{\Lambda}_D + \hat{\varsigma}\mathbf{I})^{-1} \tag{5.40}$$

and the weight posteriors as:

$$\hat{\mathbf{w}} = \hat{\boldsymbol{\Sigma}}\boldsymbol{\Lambda}_D\mathbf{w}_{ML} \tag{5.41}$$

We determine the model order using the negative free energy as a proxy measure for the true data evidence, as presented in Appendix B.

In the VB-MAR model described so far, an isotropic Gaussian prior over the weights is used (as given by (5.27)), where each coefficient has the same prior variance. However, it is possible to assign different prior variances to different groups of coefficients (or indeed to each coefficient). Using such structured priors (which are also referred to as *automatic relevance determination* priors) allows us to constrain the effective number of degrees of freedom in the model by constraining subsets of coefficients to be of similar values, hence, enabling us

[11] $F(p)$ can be used to evaluate convergence of the algorithm at each iteration. This can be done by iteratively updating the model parameters until the proportional increase in $F(p)$ at each subsequent iteration is less than a pre-defined threshold (we use 0.01%). As $F(p)$ converges relatively quickly if a larger data set is analysed, therefore putting in place this criterion enables the algorithm to rapidly reach a consistent solution, irrespective of the size of the data set analysed, hence offering significant computational benefits.

to design MAR models with sparse connectivities [291]. Use of such a sparsity inducing shrinkage mechanism also enables us to highlight coefficients which are contributing most in the linear regression process and acts as a further tool to prevent model overfitting by naturally biasing the analysis towards simpler models [303]. We present further details about structured priors, including a description of some of the possible types of priors that can be defined, in Appendix B.

5.2.2 Causal inference

We now proceed to present details of the approaches we take to infer the presence or absence of Granger causality (across both time- and frequency-domains) in multivariate systems using the parameters estimated via the VB-MAR model [22, 159]. As MAR models are based on the assumption of data stationarity, therefore (as a pre-processing step) we transform (via (4.4)) the time series representing (non-stationary) mid-prices of a set of instruments into a set of log-returns time series, which are regarded as a stationary process [211].

Time-domain inference

For time-domain causal inference in multivariate systems, we first define a d-dimensional sparse (square) connectivity matrix, $\mathbf{C}_{con}$, with $\mathbf{C}_{con}(i,j) = 1$ if testing for the $i \rightarrow j$ (direct) causal link and with all other elements set to zero. We then calculate the Kronecker product between $\mathbf{C}_{con}$ and each successive column ($\mathbf{i}_m$) of a p-dimensional identity matrix, where $m = 1, 2, \ldots, p$. This gives us p separate matrices ($\mathbf{C}_{i\rightarrow j,m}$), each of dimensions $(p \times d) \times d$:

$$\mathbf{C}_{i\rightarrow j,m} = \mathbf{i}_m \otimes \mathbf{C}_{con} \tag{5.42}$$

The vector $\hat{\mathbf{w}}$ (of length $k = pd^2$) representing the posterior weight coefficients of the MAR model can be inferred using the analysis presented previously. Defining $\mathbf{C}_{i\rightarrow j} = \left[\text{vec}(\mathbf{C}_{i\rightarrow j,m})_{m=1}^{m=p}\right]$ as a $k \times p$ dimensional matrix, the vector $\hat{\mathbf{w}}_{i\rightarrow j}$ (of length p), which picks off all the relevant weights for the $i \rightarrow j$ connection, is given by:

$$\hat{\mathbf{w}}_{i\rightarrow j} = \mathbf{C}_{i\rightarrow j}^{\mathsf{T}} \hat{\mathbf{w}} \tag{5.43}$$

Similarly, the corresponding posterior weight covariance matrix $\hat{\mathbf{\Sigma}}_{i\to j}$ (of dimensions $p \times p$) for this specific connection can be obtained as:

$$\hat{\mathbf{\Sigma}}_{i\to j} = \mathbf{C}_{i\to j}^{\mathsf{T}} \hat{\mathbf{\Sigma}} \mathbf{C}_{i\to j} \tag{5.44}$$

where $\hat{\mathbf{\Sigma}}$ is the $k \times k$ dimensional posterior weight covariance matrix.

For a connection $i \to j$ to be causal, the magnitude of its associated weight parameters $\hat{\mathbf{w}}_{i\to j}$ must be significantly greater than zero. However, large weights on their own may not accurately reflect the strength of a causal connection, as the weights may have a high uncertainty (as given by the weight covariance matrix $\hat{\mathbf{\Sigma}}_{i\to j}$) associated with their estimation. Similarly, weights with smaller magnitudes, but lower uncertainty, may in-fact indicate the presence of a causal link. Hence, we fold in the uncertainty associated with the estimated weights when inferring causation. We achieve this by making use of the squared Mahalanobis distance to measure the *distance* of the vector $\hat{\mathbf{w}}_{i\to j}$ from a vector of zeros (of the same length as $\hat{\mathbf{w}}_{i\to j}$, which is representative of the magnitude of the weights under the null hypothesis of no causality), while taking into account information contained in the covariance matrix $\hat{\mathbf{\Sigma}}_{i\to j}$ [159, 234]:

$$\chi^2_{i\to j} = \hat{\mathbf{w}}_{i\to j}^{\mathsf{T}} \hat{\mathbf{\Sigma}}_{i\to j}^{-1} \hat{\mathbf{w}}_{i\to j} \tag{5.45}$$

For p-dimensional multivariate normally distributed data, the squared Mahalanobis distance is asymptotically distributed as a chi-squared distribution with p degrees of freedom [100]. We can hence estimate the probability, $\mathrm{Pr}_{i\to j}$, that any single observation from a chi-square distribution with p degrees of freedom falls in the interval $\left[0, \chi^2_{i\to j}\right]$. This probability can be computed using the chi-square cumulative distribution function $D\left(\chi^2_{i\to j} \mid p\right)$ [57, 217]:

$$\mathrm{Pr}_{i\to j} = D\left(\chi^2_{i\to j} \mid p\right) = \frac{1}{2^{\frac{p}{2}}\Gamma\left(\frac{p}{2}\right)} \int_0^{\chi^2_{i\to j}} u^{\frac{p-2}{2}} \exp\left(-\frac{u}{2}\right) du \tag{5.46}$$

where $\Gamma(\cdot)$ is the gamma function given by:

$$\Gamma\left(\frac{p}{2}\right) = \int_0^{\infty} \exp(-u) u^{\frac{p-2}{2}} du \tag{5.47}$$

Element (i, j) of the significance value matrix, $\mathbf{P}_{sig}$, is then given by:

$$\mathbf{P}_{sig}(i,j) = 1 - \Pr_{i \to j} \tag{5.48}$$

We refer to this as the *Granger probability matrix* which gives us a probability of any specific link ($i \to j$) being causal. Element (i,j) of the *Granger causality matrix*, giving a binary indication of the presence or absence of causality at a confidence level of γ, is then given by[12]:

$$\mathbf{G}_c(i,j) = \begin{bmatrix} 1 & \text{if} & \mathbf{P}_{sig}(i,j) \geq \gamma \\ 0 & \text{if} & \mathbf{P}_{sig}(i,j) < \gamma \end{bmatrix} \tag{5.49}$$

For bivariate time series, we set the value of γ at 0.95 (95%). However, when analysing d-dimensional multivariate time series, we use a "corrected" cutoff value for γ by setting it at $\frac{0.95}{d}$; this is a standard approach for multiple hypothesis testing and is known as the Bonferroni correction [172].

It is also possible to detect non-linear causal links by projecting the observed signals onto the universal Hilbert space for the purpose of non-linear causal inference. The vector of responses to the input $\mathbf{x}$ are given by [227]:

$$\varphi(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \Phi(\mathbf{x}) \\ 1 \end{bmatrix} \tag{5.50}$$

where $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), ..., \phi_L(\mathbf{x})]$ represent the responses of $\mathbf{x}$ under L non-linear Gaussian kernel functions, each of which is given by [58, 227]:

$$\phi_l(\mathbf{x}) = \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu}_l)^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}_l)\right] \tag{5.51}$$

where $\boldsymbol{\mu}_l$ are the location parameters and the scale parameter σ can be used to adjust the complexity of the Gaussian kernel [227, 238]. It is also possible to use other types of kernels, e.g. the thin plate spline [355]. Similarly, we can obtain basis responses of other variables to obtain a projection of the data set in the Hilbert space, e.g. in a bivariate system the data set is given by $\mathbf{D} = [\varphi(\mathbf{x}), \varphi(\mathbf{y})]$. Linear VG causality can now be performed in this feature space

[12] The value of the confidence level for inferring the presence of causation can be set depending on the application and requirements of the user. A higher value will result in the model detecting fewer causal links with low uncertainty in the presence of causality and vice versa. All causality detection approaches face this trade-off between the number of causal links detected and the uncertainty associated with the results. We return to this issue later in the thesis.

to obtain estimates for the significance of causal links ($\mathbf{P}_{sig}$) and hence infer the presence or absence of non-linear causal structures in the data set.

Frequency-domain inference

Causal inference in the time-domain (as discussed so far) provides us with the mean spectral Granger causality over all frequencies (upto the Nyquist frequency) [29]. It does not convey any information about the scale-dependence of asymmetric interactions, which is one of its potential limitations in analysing FX returns (that can exhibit scale-based interaction dynamics [47, 317]). It is possible to make use of MAR models for spectral causal inference, as we discuss below. The standard MAR model based approach to study frequency-domain interactions in dynamic systems is the use of coherence (cross-spectrum) functions [146]. However, ordinary coherence analysis only describes instances when pairs of structures are in synchronous activity (by assessing temporal information between two signals) and is not suitable for identifying the direction of information-flow [22, 77]. The generalised partial directed coherence (gPDC) analysis approach (which has its origins in the field of biomedical signal processing) addresses this limitation by identifying the frequency-domain strength, as well as direction, of causal links in multivariate signals [22, 21]. Hence, by providing scale-dependent structural information for MAR models, the gPDC analysis approach can be viewed as a direct frequency-domain alternative to Granger causality (a time-domain approach) [22].

We can re-frame the MAR model (as given by (5.20)) in the spectral-domain as:

$$\mathbf{y}(f) = \mathbf{H}(f)\mathbf{e}(f) \tag{5.52}$$

where f denotes the frequency and the transfer function $\mathbf{H}(f)$ is given by:

$$\mathbf{H}(f) = \bar{\mathbf{A}}^{-1}(f) = (\mathbf{I} - \mathbf{A}(f))^{-1} \tag{5.53}$$

hence, the transfer function is only dependent on elements of the complex-valued matrix $\bar{\mathbf{A}}(f)$, which is obtained by transforming the matrix of MAR model weights into the frequency-domain via the discrete Fourier transform [161], as follows:

$$\bar{\mathbf{A}}(f) = \mathbf{I} - \sum_{m=1}^{p} \mathbf{A}(m) \exp(-i2\pi f m) \tag{5.54}$$

Using (5.54), we can write the squared gPDC function (describing the flow of information from time series i to j at frequency f) as [22]:

$$|\pi_{i\to j}(f)|^2 = \frac{\frac{1}{\sigma_j^2}|\bar{\mathbf{A}}_{ij}(f)|^2}{\sum_{k=1}^{d}\frac{1}{\sigma_k^2}|\bar{\mathbf{A}}_{ik}(f)|^2} \tag{5.55}$$

where σ^2's are the variances of the residuals, obtained using the noise covariance matrix $\hat{\mathbf{\Lambda}}^{-1}$ (as given by (B.11)); hence, we note that the gPDC statistic also folds in the uncertainty associated with the estimated parameters. As the squared gPDC statistic makes direct use of the matrix $\mathbf{A}(f)$, therefore, it offers computational advantages (over using the transfer function $\mathbf{H}(f)$) by avoiding matrix inversion and hence possible numerical imprecisions that may result due to the matrix being ill-conditioned at certain frequencies. The statistic is bounded in the range $0 \leq |\pi_{i\to j}(f)|^2 \leq 1$ and is normalised, i.e. $\sum_{j=1}^{d}|\pi_{i\to j}(f)|^2 = 1$ for $1 \leq i \leq d$ [22]. It provides us with a measure of the power-transfer between signals by estimating fraction of the power density of the i-th time series that has an influence on the j-th time series [19]. We note that the quantity $|\bar{\pi}_i(f)|^2 = 1 - |\pi_{i\to i}(f)|^2$ hence gives us the fraction of the power density of time series i which is providing "explanatory" information to all the other time series under analysis; hence, a larger value of $|\bar{\pi}_i(f)|^2$ (or correspondingly, a smaller value of $|\pi_{i\to i}(f)|^2$) provides us information regarding the extent to which time series i is *driving* other variables in the system at any given frequency, hence, we can use the statistic $|\bar{\pi}_i(f)|^2$ to measure the explanatory power-contribution of time series i in a multivariate system.

The statistic $|\pi_{i\to j}(f)|^2$ can be used to detect the presence of causality as well as to measure the causal strength. The null hypothesis (H_0) of no causality at any given frequency (f) is given by:

$$H_0 : |\pi_{i\to j}(f)|^2 = 0 \tag{5.56}$$

We note that if $|\pi_{i\to j}(f)|^2 = 0$ at all frequencies then time series i does not Granger-cause time series j, however, non-zero values of $|\pi_{i\to j}(f)|^2$ do not necessarily imply the presence of causation; in practise we need to measure the statistical significance of the statistic. This can be done using a similar approach to that described in [312]; the significance value of the statistic $|\pi_{i\to j}(f)|^2$ obtained using this approach depends on the frequency as well as the normalising

factor (denominator of (5.55)), hence, it is possible that larger values of the statistic may be insignificant while smaller values with a larger normalising factor are significant. When H_0 can be rejected, we can also compute the confidence intervals over $|\pi_{i\to j}(f)|^2$ under the normal approximation [333].

Discussion

We have presented suitable approaches (based on the VB-MAR model) for investigating the existence of causal structures (in both time- and frequency-domains) present in multivariate financial time series. We now discuss the benefits these causal inference approaches offer. Standard (ML based) MAR models suffer from the model overfitting problem [303]; inferring the MAR model parameters within a Bayesian setting allows us to address this problem by introducing priors over the weights and marginalising over the posterior distributions of the variables to estimate the model parameters. As Bayesian approaches can be computationally expensive, therefore, we use the VB approach to MAR modelling [291]. This results in better estimates for the MAR model parameters [290], and hence accurate causal inference. The VB-MAR model provides us with a unified framework under which we can accurately perform parameter estimation and model order selection (by maximising the negative free energy); the model order estimate thus obtained is often more accurate when compared to other approaches [290, 291]. The estimated MAR model weights can be highly dependent on the value of the model order, therefore selecting the correct model order is very important [83]. The VB-MAR model also provides us with estimates for the uncertainties associated with the estimated parameters [289, 291]. This allows us to fold in the uncertainty associated with estimation of the model weights by giving more "significance" to weights with lower uncertainty when inferring causation (and vice-versa). The VB-MAR model also enables us to make use of "structured priors" for causal inference, by identifying the dominant types of interaction structures present in multivariate data sets. Lastly, the VB-MAR based causality detection model can be used to easily incorporate non-linear basis functions and hence measure causality in non-linear multivariate systems.

Chapter 6
Analysis of asymmetric measures of interaction

We now proceed to carry out an in-depth empirical analysis of the asymmetric measures of interaction presented in the last chapter. This chapter provides a set of applications, using both synthetic and financial data, which allow us to test the accuracy and robustness of the causality analysis approaches. We compare our results with a linear Granger causality model, which is the standard causality detection approach frequently used in practise. Using a set of financial case studies, we demonstrate the practical utility of our approaches in extracting interesting and useful information from multivariate financial data streams, and end the chapter by providing concluding remarks focused on their merits and limitations. The FX data used in this chapter is same as that described earlier in Section 4.2.1. Unless otherwise indicated, the synthetic data used is also sampled from the same Pearson type IV distributions (as previously described in Section 4.2.2) and a causal structure is induced as required.

6.1 Analysis of synthetic data

We first test the relative accuracy of the Granger independent component (GIC) algorithm in comparison with a standard Granger causality model using a synthetic data example. Later in this section we demonstrate the use of the variational Granger (VG) algorithm for causal inference.

GIC algorithm

Given two independent, randomly distributed, non-Gaussian Pearson type IV distributions, $p_{IV,x}(t)$ and $p_{IV,y}(t)$, we generate a causal system as follows:

$$\begin{aligned} x(t) &= p_{IV,x}(t) \\ y(t) &= \alpha_{11} x(t-3) + (1-\alpha_{11}) p_{IV,y}(t) \end{aligned} \tag{6.1}$$

where the parameter α_{11} can be used to adjust the strength of the $X \to Y$ causal link and is a measure of the signal-to-noise ratio in the system. We first set $\alpha_{11} = 0$ and use the bootstrapping approach to estimate the critical values of the causal strength parameters for the two models (both with $p = 3$) we are comparing such that the false-positive detection rate (type I error) under the null hypothesis of no causality (for 1000 independent simulations) is less than 0.01 (1%), i.e. the models detect presence of the $X \to Y$ causal link in a non-causal system (as given by (6.1)) in no more than 1% of the cases; for this analysis (and in rest of this example), we make use of 1000 data points long samples, which makes it possible to generate data using a Pearson type IV distribution with relatively accurate average kurtosis and skewness values[1]. This gives us values of $\zeta_c = 4.51$ for the GIC model and F-statistic critical value of 3.87 for the linear Granger model.

We now gradually increase the value of α_{11} from 0 to 1 in steps of 0.001 and at each value of α_{11} run 100 independent simulations of the causality detection models; this allows us to test the sensitivity of the models to variations in α_{11}. The proportion of simulations correctly detecting the presence of causality gradually increases as α_{11} is increased, as shown in Figure 6.1(a). We see that the GIC model outperforms the linear Granger model in terms of the proportion of correctly detected causal links at most values of α_{11}. Both models detect all links correctly once $\alpha_{11} > 0.17$. The comparative accuracy of the GIC model is shown in Figure 6.1(b), which shows the difference in the proportion of correctly detected causal links when comparing the two causality analysis models. On average, the GIC approach detected 11.9% more causal links compared to the linear Granger approach in this example. Figures 6.1(c) and 6.1(d) show the variation of the two measures of causal strength with α_{11}, together with their associated standard deviation confidence levels. We note the significantly higher values of ζ for the GIC model as compared to ζ_{LG} values (representing the strength of causal links obtained using the standard linear Granger model); this demonstrates ability of the GIC

[1]The mean kurtosis values of the signals analysed were 12.13 and 13.16 and the signals had skewness values of -0.1454 and -0.1663 respectively. These higher-order moment values closely match properties of the financial returns presented in Figure 4.4, once again showing the utility of sampling from Pearson type IV distributions for generating synthetic data.

model to gain significantly more information about the causal structure present in the system. It is also interesting to note how the ζ value starts to accurately pick up the presence of a causal structure in the data set even for very small values of α_{11}; in contrast, the value of ζ_{LG} stays almost constant for $\alpha_{11} \leq 0.08$, indicating its inability to detect causality for low values of the causal strength parameter, i.e. in weakly causal systems.

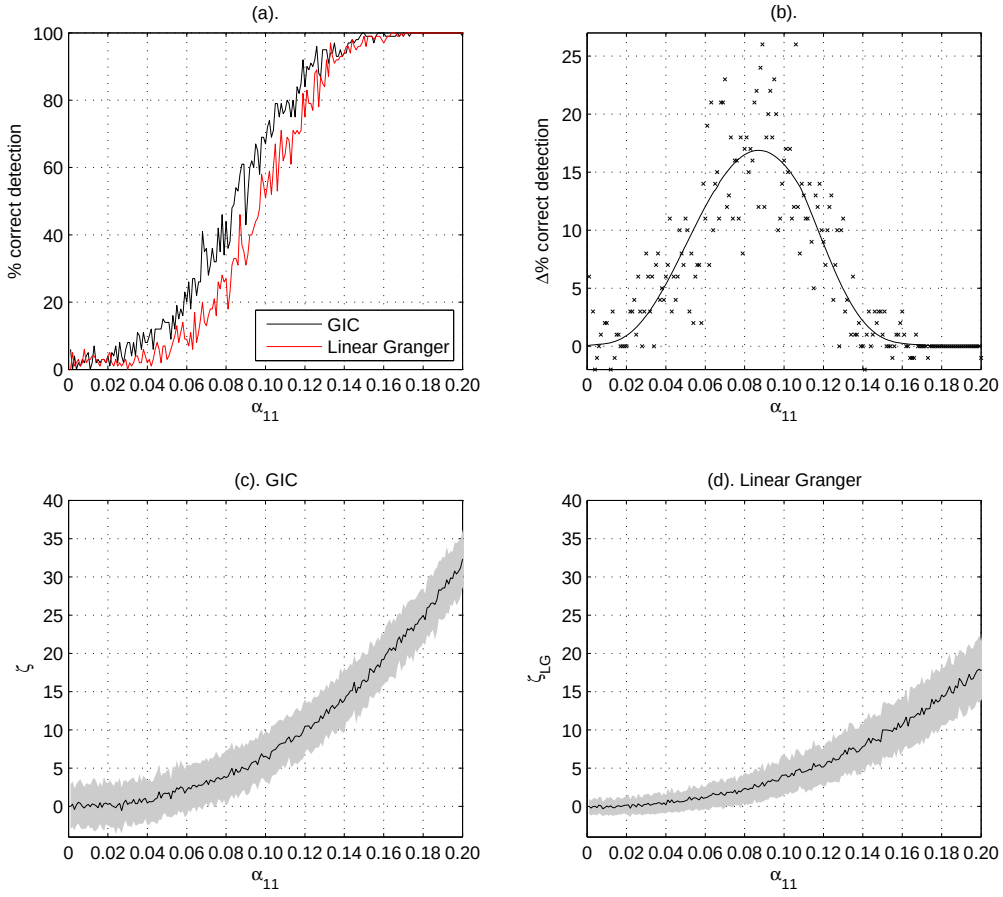

Fig. 6.1: (a,b). Comparison of GIC and linear Granger (LG) models when used to detect the proportion of causal links in the non-Gaussian bivariate system (given by (6.1)). Plot (a) shows performance of the two models as the causal strength parameter, α_{11}, is gradually increased. Plot (b) shows differences in the percentage of correctly detected links when using the GIC model as compared to the linear Granger model; the solid line represents curve of best fit for the scatter plot and is useful to visualise the general trend in the data. All results are obtained as an average of 100 independent simulations at each value of α_{11}. (c,d). Variation of measures of causal strength with α_{11} for GIC (ζ) and linear Granger (ζ_{LG}). The solid lines represent average values over 100 independent simulations at each value of α_{11}, while the shadings represent the standard deviation confidence intervals. We note that the GIC model is able to pick up the presence of a causal structure in the data even for very small values of α_{11}, and is able to gain significantly more information (as compared to a standard linear Granger model) about the presence of a causal structure in the data.

By setting $\alpha_{11} = 0.1$, we repeated the analysis (presented above) using two other standard causal inference approaches. One of these is transfer entropy, an information theoretic measure, and the other is based on correlation integrals. Both these approaches do not make any assumption about the underlying distributions and hence should be able to capture non-Gaussian asymmetric interactions. The reason they are not widely used in practise is based on their computational complexity and their requirement for large stationary data sets for accurate results (as previously discussed in detail). Here we test for the accuracy as well as the computational efficiency of these two methods. The transfer entropy approach was used to estimate the flow of information between the two variables. The average value obtained using 1000 data points long samples, over 100 independent simulations, of normalised transfer entropy was $T_{E,X\to Y} = 0.2441$. In the opposite direction, the average transfer entropy value obtained was $T_{E,Y\to X} = 0.0471$. These values correctly indicate that the flow of information in the system is from X to Y, i.e. X causes Y. The approach we used to estimate entropies using correlation integrals is based on a method employing Gaussian kernels [357]. Using this algorithm, we can obtain an estimate for the actual information transfer between variables X and Y at non-zero time-lags, i.e. an estimate for $I_{C,X\to Y}$. The average value of information transfer obtained was $I_{C,X\to Y} = 5.6813$; as $I_{C,X\to Y} > 0$, we can infer that the model is able to detect a causal link between the two variables. The average computation time required to run each simulation using the transfer entropy and correlation integral based approaches was 4.2 seconds and 12.7 seconds respectively. On the other hand, the GIC model takes an average of 0.03 seconds per simulation. The results indicate that although the transfer entropy and correlation integral based approaches can model non-Gaussian data sets, they are computationally very demanding (and require large stationary data sets, which are often not available in dynamic environments), which limits their practical use for most applications. In contrast, the computational efficiency of the GIC model indicates its ability to dynamically analyse financial data, even if it is sampled at high frequencies.

VG algorithm

We now present a set of synthetic data examples which demonstrate the accuracy and utility of the VG algorithm. Consider a 4-dimensional causal system (of order $p = 5$), given by:

$$\begin{aligned} y_1(t) &= 0.5y_1(t-1) + e_1(t) \\ y_2(t) &= 0.2y_2(t-1) - 0.6y_1(t-2) + e_2(t) \\ y_3(t) &= 0.7y_3(t-1) + 0.3y_1(t-3) - 0.8y_2(t-5) + e_3(t) \\ y_4(t) &= 0.6y_4(t-1) + 0.4y_3(t-4) + e_4(t) \end{aligned} \tag{6.2}$$

where $e(t)$'s represent unit-variance white noise terms. The structure of this causal system is illustrated by the directed graph presented in Figure 6.2. We note that some causal links are direct while others are indirect; as an example, the link between variables Y_1 and Y_4 ($Y_1 \to Y_3 \to Y_4$) is indirect due to presence of the variable Y_3 (which acts as the intervening variable in this case).

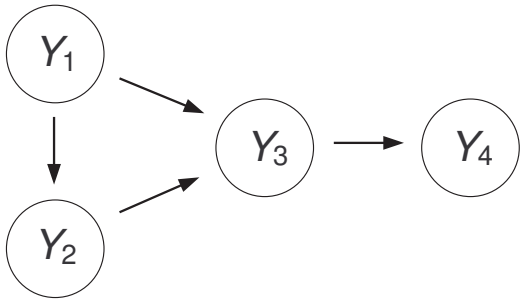

Fig. 6.2: Directed graph representing the causal structure in a four-variable system, as given by (6.2). We note the presence of direct (e.g. $Y_1 \to Y_3$) as well as indirect (e.g. $Y_1 \to Y_3 \to Y_4$) causal links.

We generate 1000 data points from this system and analyse the presence of causal links between the four variables using a MAR(5) (i.e. a MAR model with $p = 5$) model. We repeat this process 100 times and obtain the mean value of the causal strength significance for each link; the resulting Granger probability matrix is present in Figure 6.3. We note the model is correctly able to pick up the causal structure (as illustrated by Figure 6.2). We also note that the model correctly identifies only the direct links, while ignoring the indirect ones (e.g. the probabilities of $Y_1 \to Y_4$ and $Y_2 \to Y_4$ links being causal are insignificant); a bivariate pair-wise analysis of the same data failed to distinguish between direct and indirect links by resulting in significance values of 0.98 ± 0.09 and 0.85 ± 0.11 for the $Y_1 \to Y_4$ and $Y_2 \to Y_4$ links respectively. We also measure the corresponding generalised partial directed coherence (gPDC) for the data analysed above. The squared gPDC plots (for one simulation) are presented in Figure 6.4, showing the frequency-domain strength and direction of information-flow in the 4-dimensional causal system. Once again, we note that the plots correctly indicate only the direct causal links present

in the system and also provide us information regarding the frequency-dependence of asymmetric interactions. The four plots on the diagonal represent the proportion of power-density of each of the time series that does not contribute "explanatory power" to the other time series in the system.

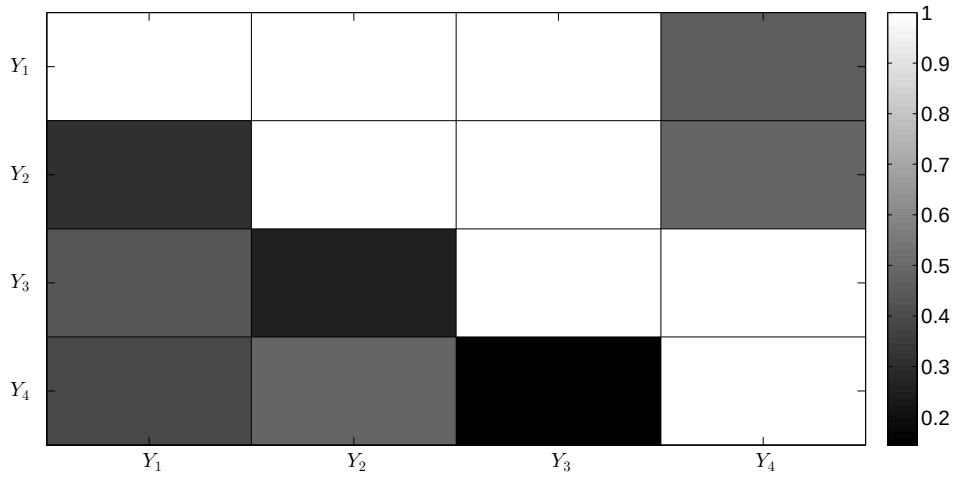

Fig. 6.3: Granger probability matrix showing the significance of causal links, obtained by analysing data generated via a 4-dimensional causal system (as illustrated in Figure 6.2 and analytically represented by (6.2)). Results are obtained using a MAR(5) model, as an average of 100 simulations.

For the data analysed in the above examples, the evidence (as indicated by the negative free energy) for models based on different types of structured priors (which we define in Appendix B) is: *Global* (2.9 ± 1.0), *Lag* (-3.7 ± 0.7), *Interaction* (6.5 ± 0.8) and *Lag-interaction* (-5.7 ± 1.3); we obtain these results by subtracting the mean evidence value for each simulation (in order to make the results meaningful). We note that although the difference between these values is not significantly large, the *Interaction* prior gives the highest evidence (due to generally larger diagonal elements in the weight matrix at a lag of one and relatively smaller off-diagonal elements). To provide a more in-depth analysis of the effect of using different type of priors, we set up the following data-generation process for the four variables, obtained by removing all cross-variable connections from the causal system described in (6.2):

$$\begin{aligned} y_1(t) &= 0.5y_1(t-1) + e_1(t) \\ y_2(t) &= 0.2y_2(t-1) + e_2(t) \\ y_3(t) &= 0.7y_3(t-1) + e_3(t) \\ y_4(t) &= 0.6y_4(t-1) + e_4(t) \end{aligned} \tag{6.3}$$

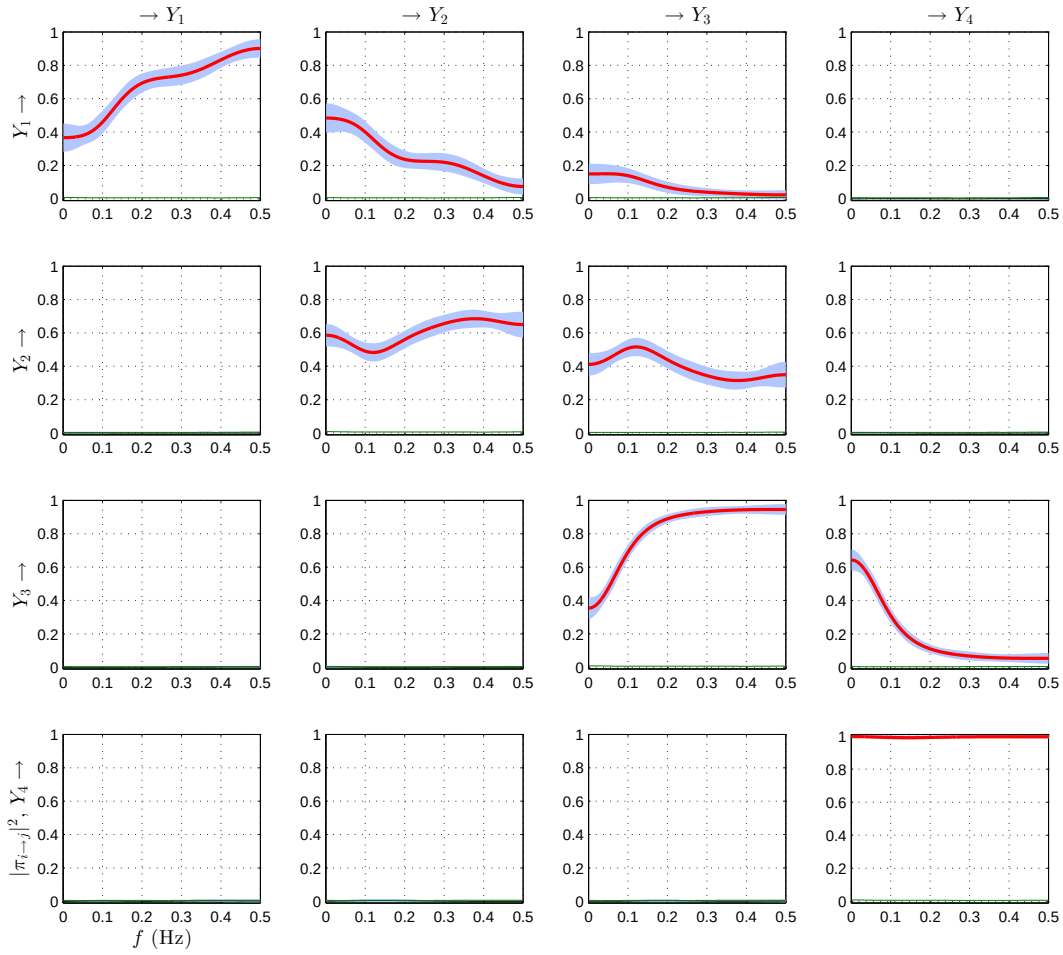

Fig. 6.4: Squared generalised partial directed coherence (gPDC) plots showing the frequency-dependent direction and strength of information-flow in the 4-dimensional causal system illustrated in Figure 6.2 and analytically represented by (6.2). The red lines represent the $|\pi_{i\to j}(f)|^2$ values (for a plot present in row i and column j) which are significantly greater than zero (with threshold significance values represented by the green lines, which, in all plots, are very close to zero). Results are obtained using a MAR(5) model and the shadings represent the 95% confidence intervals. The threshold significance values and the confidence intervals are computed using a similar approach to that described in [312].

We now generate 1000 data points from this system and (using a MAR(1) model) obtain estimates for the weights of the MAR model as well as significance of any causal links present (due to the absence of any cross-variable causal structure in this example, the off-diagonal elements of the weight matrix should ideally be zero). We repeat this process 100 times (using the *Global* and *Interaction* priors) and present the results obtained in Figure 6.5 (as we are using a MAR(1) model in this example, therefore, the *Lag* and *Lag-interaction* priors will

result in the same model evidences as the *Global* and *Interaction* priors respectively). Average values of the weights obtained when using the *Global* and *Interaction* priors are illustrated by the Hinton diagrams presented in Figures 6.5(a) and 6.5(c) respectively. We note that using the *Interaction* prior "shrinks" the off-diagonal elements of the weight matrix to close to zero (which is their correct value), hence resulting in accurate estimates for the weights, while using the *Global* prior results in several spurious off-diagonal elements in the weight matrix. The effect of using these priors (for parameter estimation) on causal inference is illustrated in Figures 6.5(b) and 6.5(d), which present the corresponding Granger probability matrices. We notice that using the *Interaction* prior in this example correctly results in much sharper contrast (in comparison to the *Global* prior) between the within-signal and cross-signal causal significance values. Using the *Interaction* prior also results in a higher evidence (7.2 ± 0.9) as compared to the *Global* prior (-7.2 ± 0.9), which once again correctly points to the presence of a diagonal structure in the weight matrices; as before, we obtain these values for the model evidences by subtracting the mean evidence value for each simulation (in order to make the results meaningful).

Non-linear causality detection

Earlier we had presented kernel methods for non-linear Granger causality analysis using the VG approach. To test the accuracy of such models, a non-linear set of time series is required. For this purpose, we make use of unidirectionally coupled Henon maps for testing the model, which can be represented by the following set of equations [230, 314]:

$$\begin{aligned} x(t) &= a - x^2(t-1) + b_x x(t-2) \\ y(t) &= a - \{ex(t-1) + (1-e)y(t-1)\}y(t-1) + b_y y(t-2) \end{aligned} \tag{6.4}$$

where $e \in [0,1]$ determines the causal strength between variables X and Y. The values of a, b_x and b_y are fixed at $a = 1.4$, $b_x = b_y = 0.3$, keeping in line with [230, 314]. It is evident from the equations that (for non-zero values of e) $X \rightarrow Y$. We vary the causal strength parameter (e) between 0.02 and 0.10 (in steps of 0.02) and run 100 independent simulations (using 1000 data points long samples) at each value of e in order to test accuracy of the model. For each simulation, in the first step, 15 basis responses of the data were obtained using the method de-

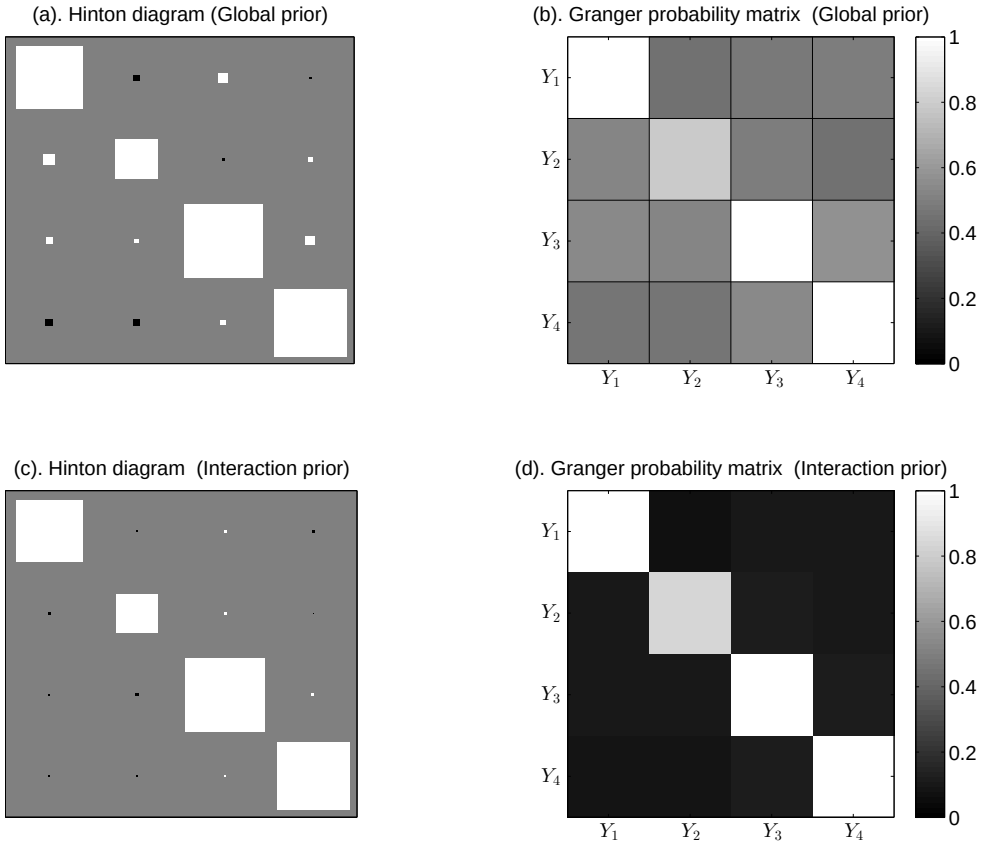

Fig. 6.5: (a,c): Hinton diagrams showing the magnitude (represented by the area of the squares) and the sign (white and black squares indicate positive and negative values respectively) of the weights of the MAR process, obtained using MAR(1) models with (a). *Global* and (b). *Interaction* priors. The results are obtained as an average over 100 simulations, using data generated via the process presented by (6.3). (b,d): Corresponding Granger probability matrices.

scribed in the previous chapter. Gaussian kernels, with randomly selected location parameters $\boldsymbol{\mu}_j$ and with scaling parameter $\sigma = 0.1$, were used for obtaining the basis responses. In the next step, Granger causality was inferred in the kernel space using the VG causality detection model, using a maximum model order of eight. The average significance values (P_{sig}), together with their associated standard deviations, for the $X \rightarrow Y$ causal links, for different values of e are plotted in Figure 6.6. Figures 6.6(a) and 6.6(b) show the results obtained with and without using the non-linear basis responses respectively. In both cases, there was no indication of causality in the opposite direction, as expected. As is evident from the plots, using non-linear basis responses allows the VG model to accurately detect the presence of causality even for very low values of the causal strength parameter (e). It is also interesting to note the low un-

certainty in the values of P_{sig} when using the kernel method. This example shows the ability of the VG model to detect causality in non-linear systems by making use of a kernel based approach. One disadvantage of such approaches is their dependence on user-defined parameters, such as the number of kernels and the scaling parameter (σ); special care needs to be taken when using these approaches for practical applications, especially in dynamic environments where these parameters may require frequent updating.

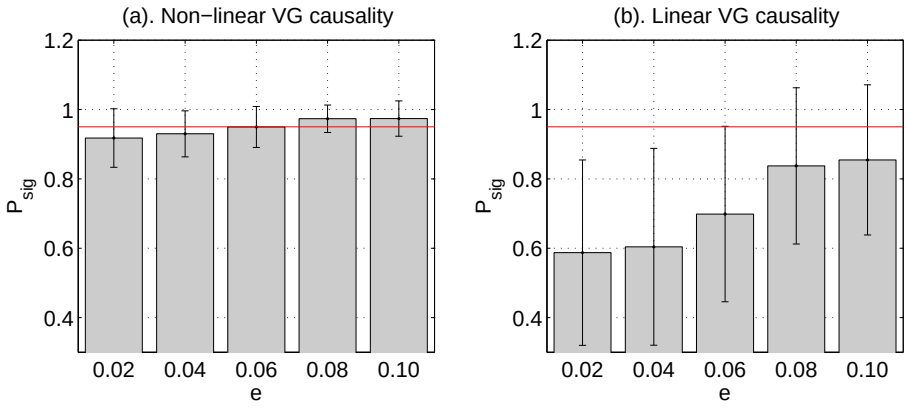

Fig. 6.6: Plots showing variation of the average significance value (P_{sig}) of the $X \rightarrow Y$ causal link as the strength of causality (e) in the unidirectionally coupled Henon maps, as given by (6.4), is increased. Plot (a) shows results for non-linear VG causality obtained using 15 individual basis responses. Gaussian kernels, with randomly selected location parameters $\boldsymbol{\mu}_j$ and with scaling parameter $\sigma = 0.1$, were used for obtaining the basis responses. Plot (b) shows the linear VG causality results obtained without using non-linear basis. All results are obtained as an average of 100 independent simulations and the error bars show the standard deviation values. The horizontal red line shows the $P_{sig} = 0.95$ threshold value for detecting the presence of causality.

ICA-AR sequential prediction algorithm

We now present an example to test the accuracy of the ICA-AR sequential prediction model. In this example, we make use of the following AR data generation process, where the variables are sampled from a non-Gaussian Pearson type IV distribution:

$$x(t) = \sum_{i=1}^{3} \alpha_i x(t-i) + p_{IV,x}(t) \tag{6.5}$$

For this example, we set the three AR model weight parameters α_i to be 0.1. We compare three prediction models, i.e. the ICA-AR, ordinary least squares AR (OLS) and VB-AR[2], using 5000 independent simulations, with each simulation using 1000 data points to estimate the model parameters (the average kurtosis of the data analysed was 12.2 with a skewness of -0.12). Figure 6.7(a) shows the average serial coupling of the data at different time lags. We now calculate the root mean square (rms) error values, e_{rms}, for the 5000 simulations using the three models. Figure 6.7(b) shows the normalised distribution of e_{rms} for 100 sets of 50 predictions, obtained using the three models. The average rms errors for the 5000 simulations were 6.81×10^{-3} for the ICA-AR model, 7.01×10^{-3} for the VB-AR model and 7.23×10^{-3} for the OLS model. As these results show, the ICA-AR sequential prediction model performs better than the two alternative AR models considered in this example. This is most likely because it takes into account the non-Gaussianity of the data being analysed and prevents model overfitting.

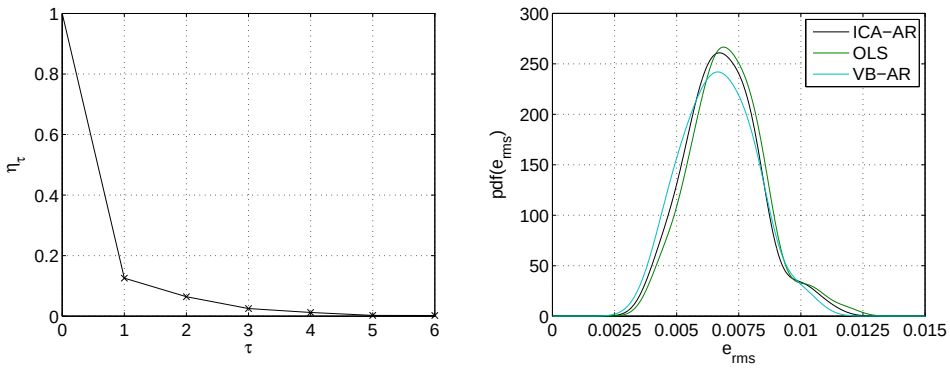

Fig. 6.7: Plot (a) shows the serial coupling at different time lags for the non-Gaussian time series, $x(t)$. Plot (b) shows the normalised pdf plots obtained using the three prediction models when used to forecast 5000 independent instances of $x(t)$.

[2]Earlier we presented the VB framework for MAR modelling. A similar method for AR processes is presented in [306], which describes a VB learning algorithm for generalised autoregressive (GAR) models. The VB-AR model is well suited for estimating the parameters of AR processes in which the noise is non-Gaussian, or non-stationary, or both. This non-Gaussian model can be used for modelling stationary stochastic processes in which a mixture of Gaussians is used to model the noise term. The model also provides a model order selection criterion using the variational free energy and average log-likelihood procedure similar to that described for the multivariate case. We use the VB approach to AR modelling to carry out a comparative analysis of the results of the ICA-AR approach.

6.2 Analysis of financial data

Let us now proceed to empirically demonstrate the utility of our asymmetric interaction measurement approaches when analysing financial data. The results presented in this section include a set of practical financial case studies which showcase some of the possible applications of causal inference approaches in the financial markets. All results presented here are obtained within a causal framework and for most examples we compare results obtained with standard approaches frequently used in practise.

6.2.1 Case study 7: Causal strength as a predictive indicator

Earlier (in Section 4.4.2) we had presented a case study which made use of a proxy rate to forecast 250 ms sampled USDCHF exchange rate, and demonstrated use of statistical indicators to maximise the mean PnL per prediction, i.e. the PnL gradient. In that case study we compared performance of the ICA-based information coupling indicator ($\eta_{95\%}$) with a linear correlation based indicator ($\rho_{95\%}$). These indicators acted as signals for their respective forecasting models to make predictions when the relevant symmetric interaction measure was greater than or equal to the 95th percentile threshold. In this case study we carry out a similar comparative analysis, but this time using indicators based on the strength of the causal links between two variables; these indicators should allow us to maximise the PnL gradient by only making predictions when the strength of the causal link is above a predefined threshold. These asymmetric (causality-based) indicators take into account the actual improvement in predictive accuracy rather than being based on lagged symmetric interaction measures (as in the previous case study). All analysis presented in this section is based on strictly causal, out-of sample methods, i.e. only information available up to time t is used to forecast relevant values at time $t+1$.

Description of the indicators

We now describe these asymmetric indicators in more details, starting with the VG causality model based indicator. The VG causality analysis approach can be used to sequentially obtain estimates for the significance value matrix, $\mathbf{P}_{sig,t}(i,j)$, at each time step t, using a fixed-length sliding-window. Elements of this matrix quantify the strength of causal link between time

series i and j, such that $i \rightarrow j$. In the following analysis, we use the notation $\mathbf{P}_{sig}(1,2)$ to represent the causal link significance from time series 1 (proxy USDCHF) to 2 (USDCHF exchange rate). As in the previous case study (Section 4.4.2), we obtain all results using a 10 second (40 data points) wide sliding-window. Our analysis indicates that $p = 1$ is by far the most frequently occurring optimal model order (for all three models compared); therefore, we fix the model order at one in this example in order to avoid any bias caused in the comparative results (the model order is often fixed in practice when analysing high-frequency sampled data in real-time, in order to reduce computational load). Figure 6.8(a) shows the normalised pdf plots of $\mathbf{P}_{sig}(1,2)$ for five 8-hour trading sessions over five days. Each of the five plots has two distinct peaks at $\mathbf{P}_{sig}(1,2) \approx 0$ and $\mathbf{P}_{sig}(1,2) \approx 1$. This shows that the temporal strength of the causal links switches between a low and a high state. Our discussion and analysis so far has pointed to the presence of unidirectional causality. However, to test for any significant presence of bidirectional causality, we also estimate the significance of causal links from exchange ($r_{UC}(t)$) to proxy ($\hat{r}_{UC}(t)$) log-returns, i.e. $\mathbf{P}_{sig}(2,1)$. The pdf plots thus obtained for the five days are presented in Figure 6.8(b). These plots only have one clear peak at $\mathbf{P}_{sig}(2,1) \approx 0$, which is as expected. Similarly, when using the GIC approach we make use of the $\zeta(1,2)$ statistic to measure the asymmetric flow of information between time series 1 and 2 ($1 \rightarrow 2$) where, as before, time series 1 represents the proxy USDCHF log-returns and time series 2 is the USDCHF exchange log-returns.

Comparative analysis

We now quantify the information gained using three separate causality detection approaches, i.e. GIC, VG and standard (OLS-based) linear Granger, by analysing the causal structure between the proxy rate and actual USDCHF exchange rate. We analyse relationship between the causal strength measure and the mean PnL for all three models; the results obtained are presented in Figure 6.9. The plots show variation of the normalised mean PnL (per tick) as the respective causal strength significance threshold (to detect causality) is reduced from the maximum (100th percentile) value to zero in buckets of 5 percentile points using a sliding-window of length 10 seconds (40 data points). The mean PnL values are normalised to start from zero in order to make comparison easier. From the plots it can be seen that the mean PnL obtained using the GIC based measure gradually decreases as the threshold value, $\zeta(1,2)$, is

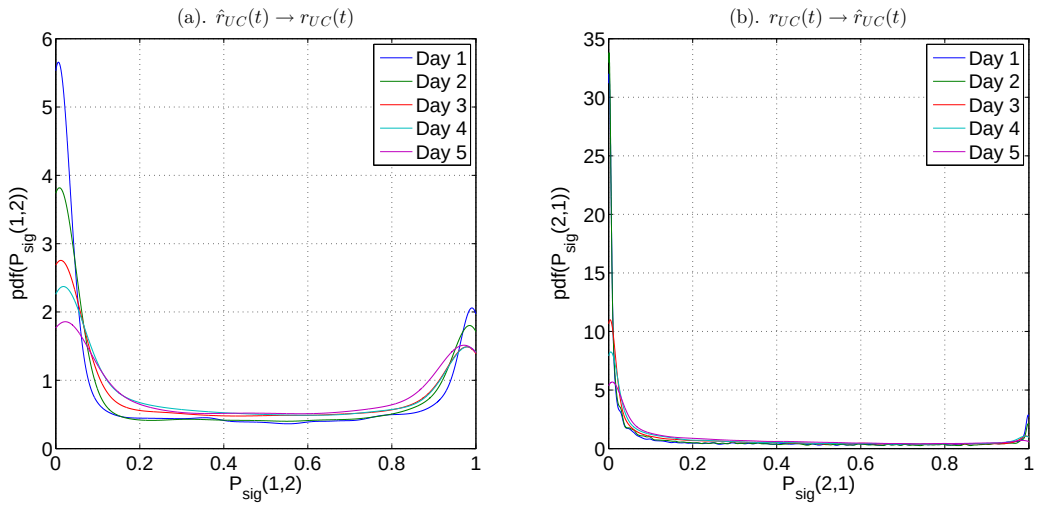

Fig. 6.8: Plots showing normalised pdfs of $\mathbf{P}_{sig}(1,2)$ and $\mathbf{P}_{sig}(2,1)$ for five days, where: (a). $\mathbf{P}_{sig}(1,2)$ represents the significance of the $\hat{r}_{UC} \rightarrow r_{UC}$ causal link, (b). $\mathbf{P}_{sig}(2,1)$ represents the significance of the $r_{UC} \rightarrow \hat{r}_{UC}$ causal link. The two distinct peaks for all plots in (a) show that the temporal causal strength between the proxy ($\hat{r}_{UC}$) and exchange (r_{UC}) USDCHF log-returns varies between a high and a low state when $\hat{r}_{UC} \rightarrow r_{UC}$, while plots in (b) only have one clear peak at $\mathbf{P}_{sig}(2,1) \approx 0$, which shows absence of any significant bidirectional causality.

reduced from $\zeta(1,2) \geq$ 95th percentile to $\zeta(1,2) \geq$ 0th percentile. On the other hand, both the VG and linear Granger based measures do not decrease monotonically, which signifies their inability to accurately capture strength of the causal links for the non-Gaussian data streams being analysed.

For any dynamic causality detection model, there is a threshold value of the causal strength which needs to be crossed for the model to infer the presence or absence of causality, i.e. there is a trade-off between the strength of causality and the number of causal links detected. Comparing different causality models in a dynamic environment therefore requires us to set causality detection thresholds to a value such that the number of causal links detected is the same for all models. As previously discussed (in Section 4.4.2), any practical trading model needs to be able to causally make predictions at predefined average time intervals. As before, in the analysis that follows, we set the causal strength threshold values to a level which results in the three models each making a prediction on average once every 5 seconds, i.e. every 20 data points; this is a reasonable estimate for trading frequency of a trading strategy making use of data sampled at high-frequency. To do this, the threshold value for the three measures is set at $\geq$ 95th percentile. Figure 6.10 shows cumulative return (in pips) obtained for five trading sessions over five days using the three different causality measures. Also included are the

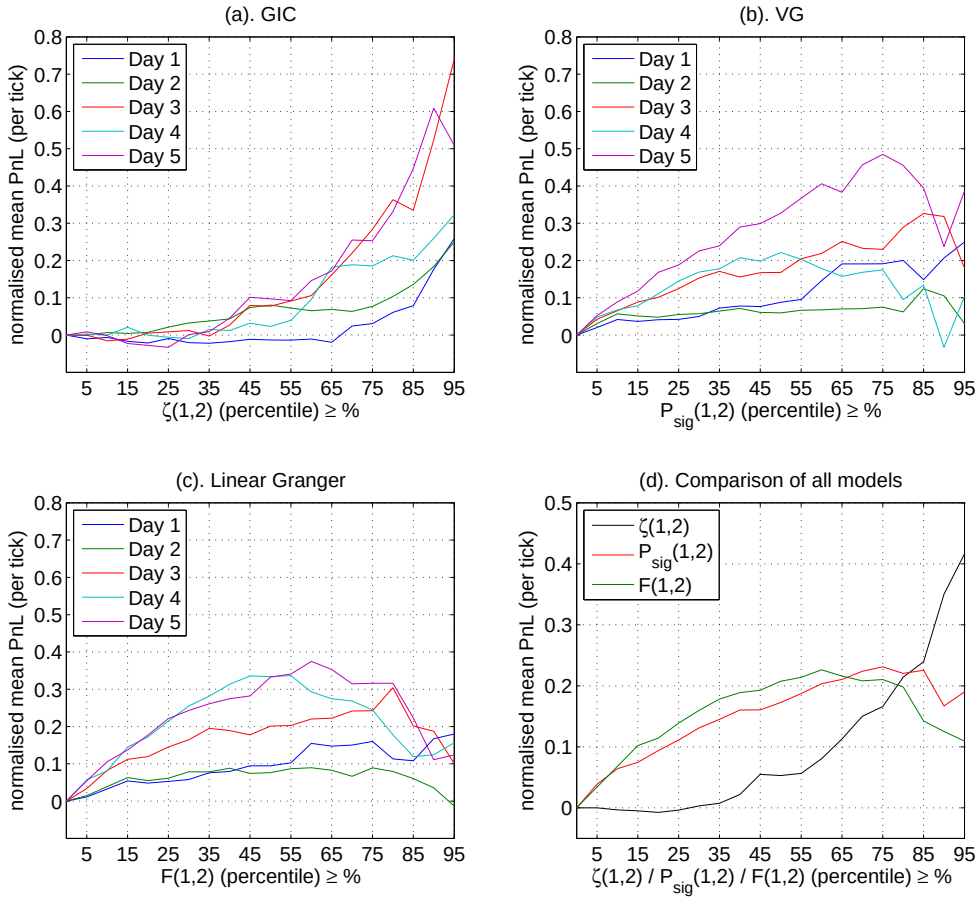

Fig. 6.9: Normalised mean PnL (per tick) plotted against the percentile values of the strength of causality measures obtained using three different causality analysis models, i.e. GIC, VG and the OLS-based linear Granger for five 8-hour trading sessions over five days. Plots in (d) represent the mean values for all five days for the three models compared in plots (a) to (c). Only the GIC plot in (d) decreases monotonically with the causal strength significance threshold value, showing its relative accuracy in modelling the causal structure in these non-Gaussian data streams. Time series 1 and 2 represent the proxy ($\hat{r}_{UC}$) and exchange (r_{UC}) USDCHF log-returns respectively.

plots (Figure 6.10(d)) for the case when no indicator is used but instead predictions are made at equally spaced time intervals of 5 seconds over the entire length of the data. These plots act as a benchmark for the relative accuracy of the three causality analysis models considered in this case study.

Figure 6.11 shows the mean daily cumulative return (in pips) for the three models compared, obtained using data for the five trading sessions. The PnL gradient is maximum for the GIC model, followed by the VG and linear Granger models respectively. This shows the relative utility of the GIC model to accurately identify regions of high causal strength in the

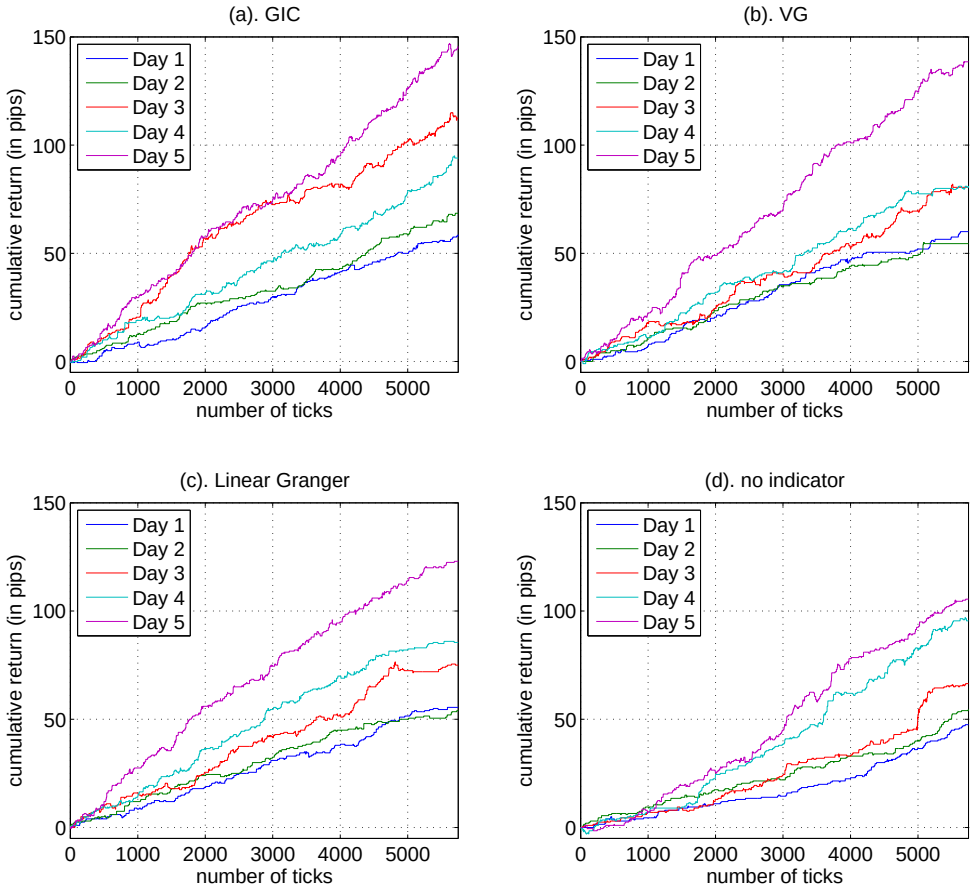

Fig. 6.10: Cumulative return (in pips) for five 8-hour trading sessions over five days obtained using three different causality model based indicators. The four graphs represent models based on the following indicators: (a). GIC, $\zeta(1,2) \geq$ 95th percentile; (b). VG, $\mathbf{P}_{sig}(1,2) \geq$ 95th percentile; (c). Linear Granger, $F(1,2) \geq$ 95th percentile; (d) without using any indicator, but instead making predictions at equally spaced 5 second intervals. Time series 1 and 2 represent the proxy ($\hat{r}_{UC}$) and exchange (r_{UC}) USDCHF log-returns respectively.

system by taking into account the non-Gaussian nature of the underlying data. Earlier (in Section 4.4.2) we described the standard deviation of returns, σ_{SD}(PnL), as a widely used measure of the risk associated with any given trading model. Table 6.1 shows daily values for the cumulative PnL, standard deviation of the PnL and the hourly return-to-risk ratio for the three causality models. As these values indicate, the GIC model outperforms the VG and linear Granger models by 15.6% and 21.9% respectively in terms of PnL. The GIC model results in a higher cumulative PnL on four out of the five days with respect to the VG model and on all five days compared to the linear Granger model. The data shows that the variability of PnL is relatively higher for the GIC model, however, it still results in a higher return-to-risk ratio on

average. Computationally, the GIC model takes on average 19 ms to estimate the strength of causal links and make a prediction at each time step. Given that the data is sampled at 250 ms, this makes the model suitable for practical use, even when dealing with data sampled at high frequencies.

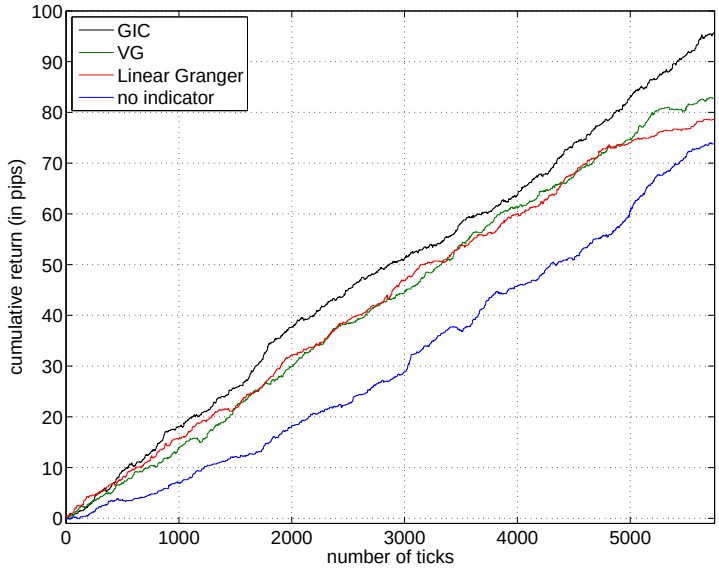

Fig. 6.11: Average daily cumulative return (in pips) for five 8-hour trading sessions over five days obtained using three different causality model based indicators. Each indicator results in the respective model making a prediction when the relevant causal strength measure is $\geq$ 95th percentile. Also included is a plot (no indicator) obtained using a simple model that makes predictions at equally spaced 5 second intervals. Time series 1 and 2 represent the proxy ($\hat{r}_{UC}$) and exchange (r_{UC}) USDCHF log-returns respectively.

	$\sum$PnL (in pips)			σ_{SD}(PnL)			Return-to-risk ratio		
Day	$\zeta(1,2)$	$\mathbf{P}_{sig}(1,2)$	$F(1,2)$	$\zeta(1,2)$	$\mathbf{P}_{sig}(1,2)$	$F(1,2)$	$\zeta(1,2)$	$\mathbf{P}_{sig}(1,2)$	$F(1,2)$
1	58.5	60.0	55.3	0.1073	0.0999	0.1002	68.1	75.1	70.0
2	68.4	54.3	54.0	0.1051	0.0976	0.0996	81.4	69.5	67.8
3	112.5	80.6	75.1	0.1910	0.1616	0.1490	73.6	62.3	63.0
4	94.0	80.9	85.4	0.1634	0.1403	0.1424	71.9	72.1	74.9
5	145.6	138.5	123.0	0.2125	0.1936	0.1767	85.6	89.4	87.1
Mean	95.8	82.9	78.6	0.0719	0.0641	0.0618	167	162	159

Table 6.1: Table showing accuracy of the three causality analysis model based indicators for analysis carried over five 8-hour trading sessions across five days. Included in the table are values for the PnL (in pips), standard deviation of the PnL (σ_{SD}(PnL)) and the return-to-risk ratio (which is calculated using hourly returns). Time series 1 and 2 represent the proxy ($\hat{r}_{UC}$) and exchange (r_{UC}) USDCHF log-returns respectively.

6.2.2 Case study 8: Forecasting financial market volatility

Let us now consider an example which demonstrates utility of the ICA-AR sequential prediction model for forecasting financial market volatility. Financial returns generally have negligible (real) autocorrelation [14]. However, financial volatility, i.e. the squared log-returns, typically exhibits slowly decaying positive serial correlation at multiple time lags. This is a well-known property of financial data sets at a wide range of sampling frequencies and is often used in practise to forecast volatility of any specific financial instrument [295]. Volatility forecasting has numerous applications in the financial markets [13]. It is of central importance in many risk management models, e.g. to accurately estimate the value-at-risk (VaR) of a portfolio at any given time requires volatility forecasts [140]. Fund managers often rely on future volatility estimates to decide on the asset allocation proportions in any given portfolio (we described this in detail in Section 4.4.3). Volatility forecasts are also the primary variable taken into account by almost all options pricing models, which is another important area where these models are used [210]. Most of these applications do not require the use of high-frequency sub-second sampled data, therefore in our analysis in this section we make use of relatively lower frequency half-hour sampled FX data. In practise, AR conditional heteroskedasticity (ARCH) and generalised ARCH (GARCH) models are generally used to model volatility dynamics [55]. These (and other similar) models are primarily based on standard AR models. Hence, it is possible to improve the volatility forecasts using these models if an accurate AR model is used as the basis. Therefore, for the purpose of this example, we only focus on comparing the merits of different AR models for volatility forecasting. As described in [353], GARCH models making use of non-Gaussian AR residuals can be more accurate than those using normally distributed residuals. As the results presented later in this case study show, the ICA-AR model (which takes into account the higher-order statistics of the data being analysed) is potentially well-suited for forecasting volatility of financial instruments.

We make use of the ICA-AR model to predict the one step ahead out-of-sample volatility, i.e. the squared log-returns, $r^2(t)$, for seven liquid currency pairs (sampled every 0.5 hours). The currency pairs analysed in this example are: AUDUSD, EURCHF, EURUSD, GBPUSD, NZDUSD, USDCAD and USDJPY. Figure 6.12 shows the average information coupling between the volatility and its time delayed copies at different lags (obtained using data covering

a 2 year period) for all seven currency pairs; as expected, all currency pairs have noticeable serial coupling at time lags ranging from 1 data point (0.5 hours) to 6 data points (3 hours). In this case study we compare results obtained using the ICA-AR model with three other forecasting approaches, i.e. the OLS based AR, VB-AR (with four mixture components) and a simple unit-lag trend-following (TF) model (as given by (4.12)). The four approaches were used for one-step ahead out-of-sample forecasts over the 104 weeks (25,000 data points) period. A 12 hour (24 data points) wide sliding-window was used to capture the changing dynamics of the markets, and hence to dynamically update the AR model parameters at each time step. The BIC was used to decide on the optimal lag of three, which was fixed for all the currency pairs and models in order to stay consistent (when analysing relatively small samples, as is the case with our dynamic model, BIC outperforms other AR model order estimation methods [232]). To quantify the accuracy of the different models, the rms error, e_{rms}, between the realised volatility of the observed data, i.e. the absolute values of the log-returns, $|r_{obs}(t)|$, and the predicted data, $|r_{pred}(t)|$, is calculated, in order to stay consistent with previous similar empirical studies [93, 213]. Realised volatility, i.e. the standard deviation of the log-returns, is simply the square root of the volatility of the observed data. Similarly, the information coupling, $\eta(|r_{obs}(t)|,|r_{pred}(t)|)$, between the observed and predicted realised volatilities is also measured, with higher values representing more accurate forecasts. The results for all currency pairs are presented in Table 6.2. Figure 6.13 shows the normalised pdf plots of the rms error values, e_{rms}, obtained using 104 weeks of data. Results for each of the seven currency pairs obtained using the four models are plotted for comparative analysis.

As the results indicate, on average the ICA-AR sequential prediction approach outperforms the three other approaches for the data analysed in this case study. The ICA-AR model's average rms error is 27% lower than that for the OLS model, which is the standard model used in practise. The results presented clearly indicate (relative) inability of the OLS model to accurately forecast volatility for all currency pairs. As previously stated, in this example we used four mixture components for the VB-AR model. As the accuracy of the VB-AR model is dependent on the number of components used in the analysis, therefore, in practical applications results obtained using different number of components should be considered in order to select the most suitable model. However, this can potentially be a limitation of the

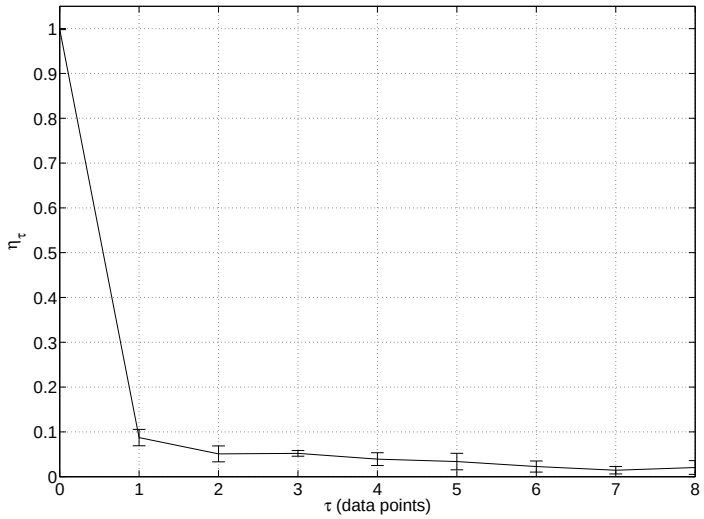

Fig. 6.12: Serial information coupling η_τ (at different lags τ) for the volatility data $r^2(t)$ which is sampled at 2 samples per hour. The plot shows the mean coupling for the seven liquid currencies, together with the error bars representing the standard deviation of coupling at each time lag.

	$e_{rms}(\times 10^{-3})$				$\eta(\lvert r_{obs}(t)\rvert, \lvert r_{pred}(t)\rvert)$			
	ICA-AR	OLS	VB-AR	TF	ICA-AR	OLS	VB-AR	TF
AUDUSD	0.9912	1.2289	0.9701	1.0076	0.2060	0.1412	0.1801	0.1905
EURCHF	0.3264	0.3791	0.3271	0.3297	0.2110	0.1758	0.1930	0.2155
EURUSD	0.8280	1.2554	0.9268	0.8397	0.2199	0.1460	0.1732	0.2175
GBPUSD	0.7369	0.9204	0.7351	0.7478	0.2158	0.1701	0.1929	0.2135
NZDUSD	1.0920	1.5571	1.1139	1.1047	0.2282	0.1341	0.1769	0.2142
USDCAD	0.8129	1.1444	0.8546	0.8263	0.2425	0.1823	0.2011	0.2401
USDJPY	0.8237	1.2816	0.8833	0.8322	0.2106	0.1497	0.1833	0.2105
Mean	0.8016	1.1096	0.8301	0.8126	0.2191	0.1570	0.1858	0.2145

Table 6.2: Comparison of four different forecasting models when used to carry out one-step ahead out-of-sample forecasts for seven 0.5 hour sampled liquid currency pairs representing data over a period of two years. The table includes values for the rms error, e_{rms}, as well as the information coupling between the observed and predicted time series, $\eta(|r_{obs}(t)|, |r_{pred}(t)|)$; in all cases, the 95% confidence bounds on $\eta(|r_{obs}(t)|, |r_{pred}(t)|)$ were very close to the actual values. TF represents a simple unit-lag trend-following model, as given by (4.12).

VB-AR model as well, as the optimal number of components may change dynamically with time. It is interesting to note how for some of the currency pairs the rms error and information coupling accuracy measures appear to give conflicting information regarding the accuracy of the models, e.g. for AUDUSD, although the VB-AR model appears to be the best choice when considering the rms errors, the ICA-AR model has the highest information coupling between the observed and predicted values. This is because the rms errors are a measure of the absolute difference between the observed and predicted values while the information coupling measure only captures the directional accuracy of the models without taking into account by how much

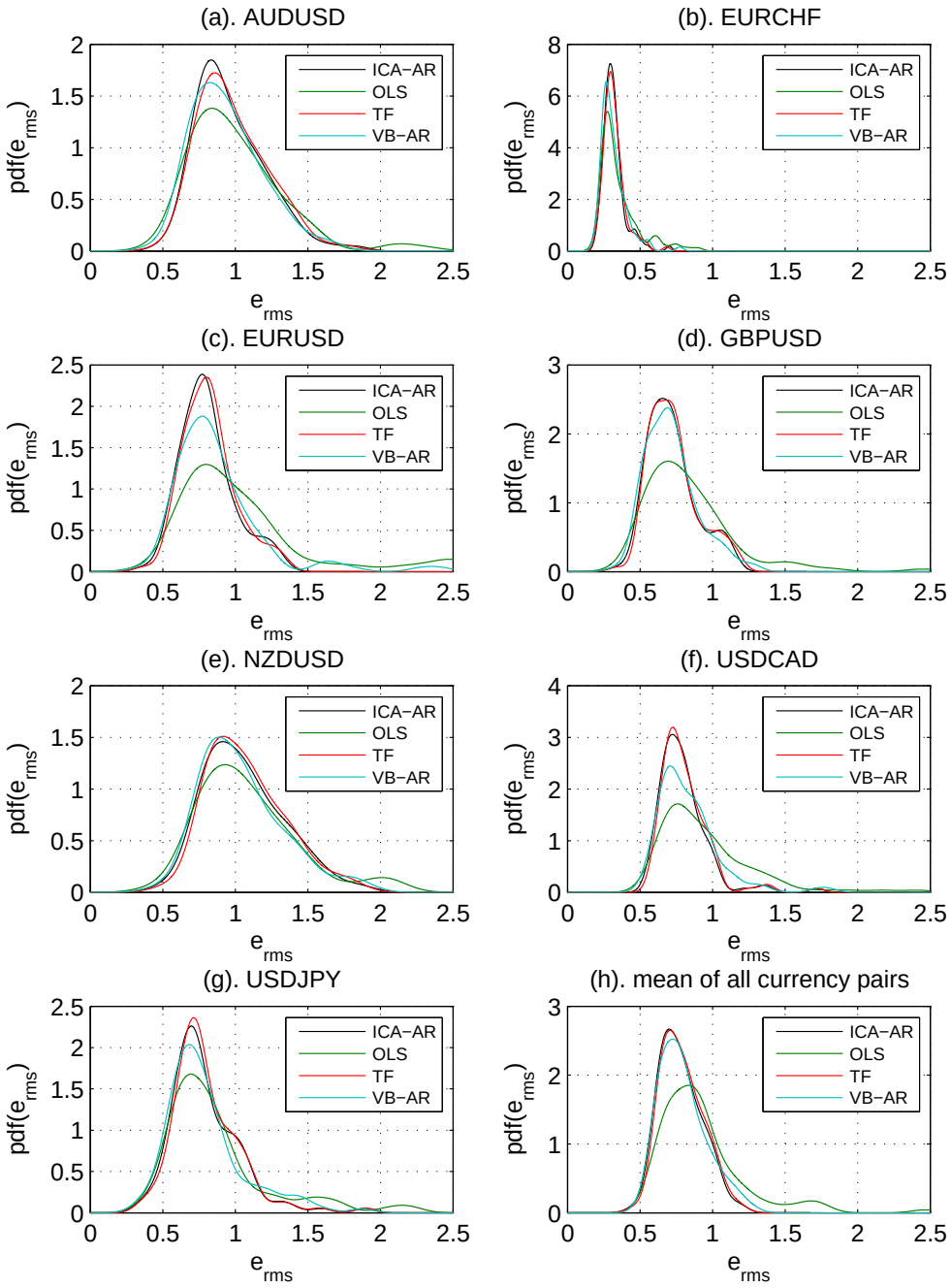

Fig. 6.13: Normalised pdf plots showing distributions of the rms errors, $e_{rms}(\times 10^{-3})$, between the observed, $|r_{obs}(t)|$, and predicted, $|r_{pred}(t)|$, realised volatilities. The plots are obtained using 0.5 hour sampled data for seven liquid currency pairs over a 104 week period. Plot (h) shows the mean of e_{rms} for all seven currency pairs.

the predicted value differs from the observed value in magnitude. Therefore, this indicates that for AUDUSD the VB-AR model more accurately captures the exact values of the observed data, but the ICA-AR model outperforms in terms of the proportion of accurate directional predictions. On the whole, the ICA-AR model outperforms the other models for five of the seven currency pairs in terms of the rms error and for six of the seven currency pairs in terms of the information coupling accuracy measure for the data analysed. Although both the ICA-AR and VB-AR approaches assume the residuals to be non-Gaussian, the ICA-AR model has computational advantages over the VB-AR model. On average, the ICA-AR model took 17.6 ms for each prediction, compared to 70.1 ms for the VB-AR model (which is fine for the data analysed in this example, but can potentially be a limiting factor in its usability if used to analyse high-frequency sampled data). The computational speed and accuracy of the VB-AR model is highly dependent on the number of mixture components used in the model with the computational cost increasing almost exponentially with the number of components used; this is another factor which may need to be considered when using it in practise.

6.2.3 Case study 9: Extracting information from directed FX networks

So far we have studied asymmetric interactions in bivariate FX data streams, in this case study we investigate causal relationships in multivariate FX time series with the aid of directed FX networks. We first run a 12-dimensional MAR(4) model over 8 hours (one trading session) of 0.5 second sampled FX data representing log-returns of 12 currency pairs. For our analysis, we use ten minute (1200 data points) sections of the data in order to make sure the data set (within each section) is large enough to be as Gaussian as possible (as larger log-returns data sets are generally more Gaussian than smaller ones [98][3]), while at the same time small enough to ensure stationarity and to capture short-run dynamics of the market. For each of the 48 ten minute sections of data we analyse over the 8-hour trading session, we compute the 12-dimensional Granger probability matrix ($\mathbf{P}_{sig}$) in order to investigate the cross-currency causal structures present in FX spot data. We first present a set of results and later discuss the implications of our findings. The Granger probability matrix representing the mean significance values of the causal links over 48 simulations is represented in Figure 6.14 (with diagonal elements set to

[3]Financial data is often non-Gaussian [318]. Standard (ML based) MAR models (as well as the VB-MAR model) are based on the assumption of normality of the data, which is one of their limitations. Therefore, we use relatively large (hence, less non-Gaussian) data sets for analysis presented in this case study.

zero for clarity). The figure provides some interesting information, e.g. the rows representing EURUSD and USDJPY seem to have the highest significance values, which points to their role as the *driving* currency pairs during this trading session. We also note that the columns representing USDCHF and EURJPY seem to have the highest mean values, indicating that these currency pairs are generally *driven* by other pairs.

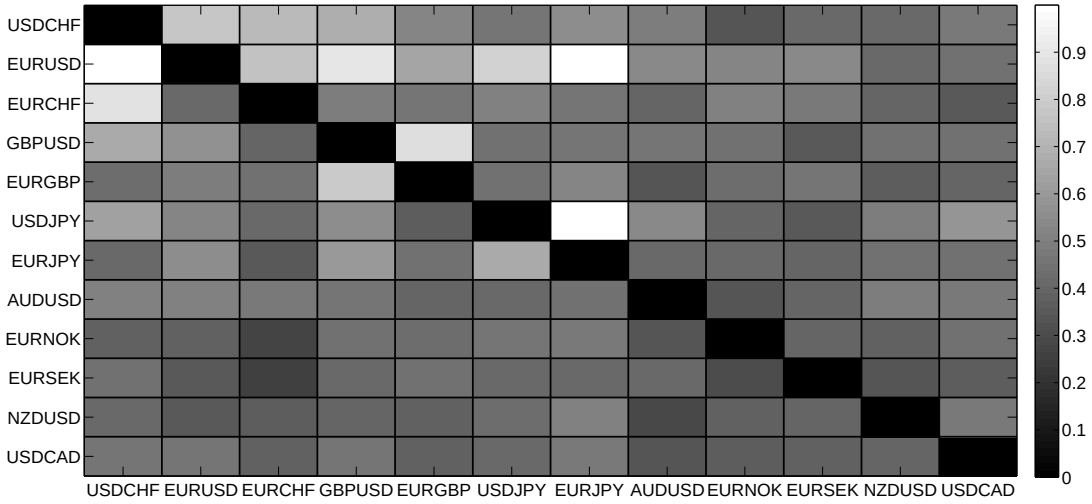

Fig. 6.14: Granger probability matrix showing significance of any given link ($i \rightarrow j$) being causal, where i and j refer to the rows and columns respectively. The matrix shows the significance of the links ($\mathbf{P}_{sig}(i,j)$) for 12 currency pairs obtained as an average of 10 minute sections of 0.5 second sampled data over an 8-hour period. The within-series significance values are set to zero for clarity (as we are interested in analysing cross-currency causation).

To get a more detailed insight into the structure of these asymmetric interactions, we make use of directed causal networks. Ideally, we should include all available information about the causal structure present in a multivariate system in such networks, however, a fully connected directed network with d nodes will have $d(d-1)$ links, which will make information retrieval a complex undertaking. Therefore, we make use of alternate approaches for investigating the properties of such networks. We first extract the fully connected unidirectional network, as presented in Figure 6.15. It shows the direction and strength of all the dominant (i.e. the more significant connection between each pair of nodes) causal links present in the network. The network provides us with some interesting information about the overall structure of asymmetric interactions in the FX market (during this trading session). The role of EURUSD, USDJPY, EURCHF and GBPUSD as the main *driving* currency pairs is clearly visible, similarly, we note

that EURJPY, EURGBP and USDCHF are being *driven* by other currency pairs. The position of USDCHF as the most actively *driven* currency pair once again points to its role as a classic arbitrage-leg pair. We also observe that some currency pairs, e.g. EURSEK, EURNOK and NZDUSD, are relatively *dormant*, i.e. they are neither noticeably *driving*, nor being *driven*, by other currency pairs. This is a reflection of their illiquidity relative to the other G10 currency pairs, due to which their exchange rates generally exhibit relatively stable dynamics, hence resulting in their *dormant* position.

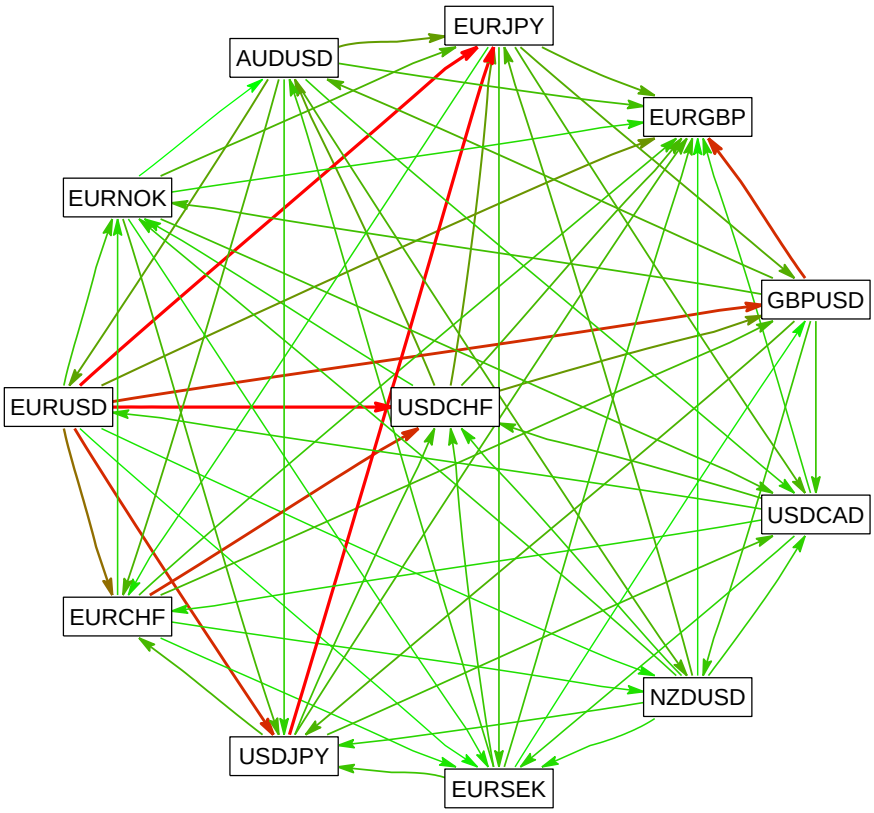

Fig. 6.15: Fully connected unidirectional network, showing all the dominant (i.e. the more significant connection between each pair of nodes) links between 12 currency pairs. The colour of the links smoothly transitions between red (for strong connections) and green (for weaker connections).

However, by definition, a fully connected unidirectional network will contain nodes for all the currency pairs, even if these nodes represent *dormant* pairs. They also do not show any bidirectional information-flow. To address these issues, we make use of a partially-connected network, which we call the Granger causality network (GCN), in which we include all the causal links which have a significance value above a pre-defined threshold. Putting in place

this threshold enables us to prune-away less significant edges (from a fully-connected directed network) by placing a restriction on the inclusion of less relevant links (and thus nodes) in the network, and hence presents the direction and magnitude of all the causal links in a network which we have a relatively high probability of occurrence. The exact value of this threshold can be selected based on requirements of the user; a higher value will result in the GCN displaying fewer (but more significant) connections, while a lower value will result in a larger number of connections (some of which may be insignificant). For our analysis, we set the threshold value at 0.75, such that only connections with $\mathbf{P}_{sig}(i,j) \geq 0.75$ will be included in the network. The resulting GCN is presented in Figure 6.16. We notice that the GCN provides much more useful information about unidirectional as well as bidirectional causal links present in the network, while still maintaining representational simplicity. We can clearly see the *driving* position of EURUSD, which once again points to its dominance during this trading session. An interesting pair of causal links we can identify from this network is {EURUSD,USDJPY}$\rightarrow$EURJPY. Both these links have very high significance values; information which can be used to develop a forecasting model to predict EURJPY spot rates. The low standard deviations of these (and some other) significance values indicates that these links remain relatively stable across time, hence, making it possible to develop robust prediction models. There has been some work done previously on using directed networks in the spot FX market, e.g. in [243] the authors use lagged linear correlation values to determine the direction of information-flow. However, as far as we know, ours is the first attempt to construct directed causal networks using Granger causality in order to quantify the magnitude, and determine the direction, of information-flow between spot FX currency pairs.

Let us now discuss our main findings from the results presented so far in this case study. Some of the prominent causal links we have identified are {EURUSD,EURCHF}$\rightarrow$USDCHF and {EURUSD,USDJPY}$\rightarrow$EURJPY. We notice that for all these links, the *driven* currency pair is made up of two currencies whose more liquid, primary, pairs are in the *driving* position[4]. This indicates that the speed of the price discovery process (i.e. how price movements react to

[4]In the FX market, each currency is generally traded through one or more primary currency pairs. Often, this is a currency pair containing the USD as one of the currencies, e.g. JPY is usually traded as USDJPY, EUR as EURUSD, AUD as AUDUSD, etc. Within the G10 space, major exceptions to this convention are the Swiss franc, which is generally traded as EURCHF, and the Scandinavian currencies, which are also traded through their EUR crosses (non-USD currency pair), i.e. EURSEK and EURNOK.

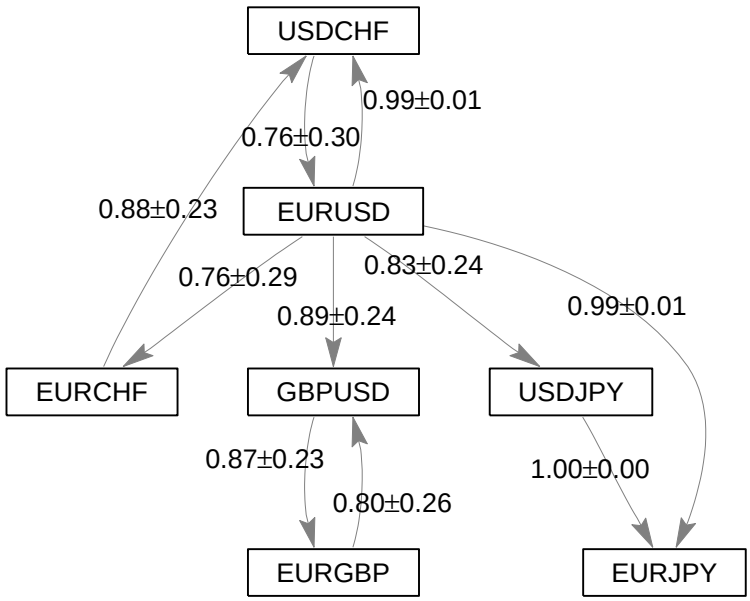

Fig. 6.16: Granger causality network (GCN) showing all the causal links which have a significance value of $\mathbf{P}_{sig}(i,j) \geq 0.75$, i.e. the figure is obtained by pruning-away less significant edges from a fully-connected network. The means and standard deviations of the significance values are obtained as an average of 10 minute sections of 0.5 second sampled data over an 8-hour period.

the assimilation of relevant new market information) in the spot market in-part depends on the relative liquidity level of that currency pair. Earlier (in Sections 4.4.2 and 6.2.1) we had presented a set of case studies that made use of this property of the FX markets to develop robust exchange rate forecasting models. Another interesting trio of relatively strong causal links we have identified are EURUSD→GBPUSD, EURGBP→GBPUSD and GBPUSD→EURGBP. These last two links indicate that during this trading session GBPUSD and EURGBP (the two major GBP pairs) exhibited relatively strong bidirectional causality. This can possibly be explained by GBP being "in play" (i.e. very active and liquid) during this trading session, because of which USD and EUR had relatively less impact on the price dynamics of GBPUSD and EURGBP respectively as compared to GBP; this can also explain the absence of a strong EURUSD→EURGBP link which we would have expected (based on our previous discussion), given that EURUSD is regarded as a relatively more liquid and active currency pair.

6.2.4 Case study 10: Frequency-domain causal inference in the FX market

So far we have investigated the time-domain properties of asymmetric interactions in the FX market; this has provided us information about the mean spectral Granger causality across the full range of frequencies. To study the scale-dependence of information-flow between currency pairs, we make use of the generalised partial directed coherence (gPDC) causal inference framework (as previously discussed). We now analyse spectral properties of causation between five currency pairs (we have not included all 12 pairs for ease of presentation), i.e. USDCHF, EURUSD, EURCHF, USDJPY and EURJPY. Our results in the time-domain indicate the presence of significantly strong causal links between some of these currency pairs, we now investigate whether the strength of these interactions exhibit any frequency-dependence. For this purpose, we sequentially run a 5-dimensional MAR(4) model (inferring the optimum value of p in the range [1:4] at each time-step) over the entire 8-hour period of the trading session (using a 10 minute wide sliding-window), and at each time-step infer the squared gPDC statistic for all causal links at a range of frequencies (upto 1 Hz, the Nyquist frequency). The resulting spectrograms obtained are presented in Figure 6.17, with the off-diagonal plots representing values for the $|\pi_{ij}(f)|^2$ statistic. We note that some plots represent significantly high gPDC values, prominent examples being EURUSD$\rightarrow${USDCHF,EURJPY}, EURCHF$\rightarrow$USDCHF and USDJPY$\rightarrow$EURJPY. We also notice that the strength of causation for these connections is most significant at lower frequencies (0.1-0.4 Hz) and gradually decreases at higher frequencies. This is most likely due to the fact that at higher frequencies, effects of market microstructure act as "noise" to diminish the significance of any causal links; this noise can originate due to a range of factors, such as market-makers continuously (and asynchronously) updating their bids and offers for individual currency pairs in order to stay in front of the stack, or indeed as a result of pre-processing the resulting asynchronous data before inclusion in a (synchronous) database [98]. It is also interesting to note that the strength of these causal connections, at lower-frequencies, stays relatively high for long time periods (measured in tens of minutes), indicating the presence of regions of temporal persistence in causality between particular currency pairs at lower frequencies; this information can lead to the development of robust algorithms for the purpose of trading or hedging.

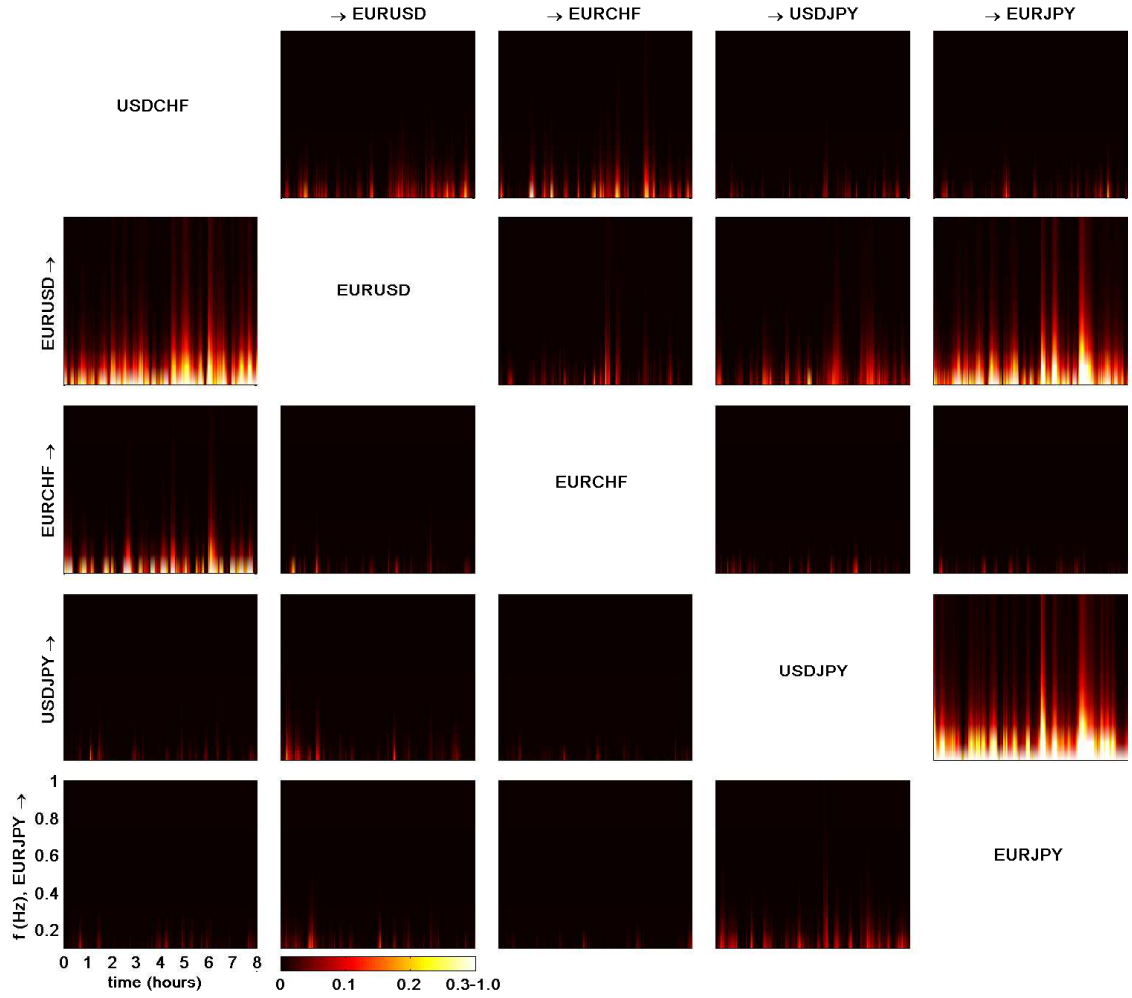

Fig. 6.17: Spectrograms (based on the squared gPDC statistic) for five currency pairs, obtained using a 5-dimensional MAR(4) model (i.e. inferring the optimum value of p in the range [1:4] at each time-step). The spot data is sampled at 2 Hz.

Although at higher frequencies (0.6-1.0 Hz) the strength of causation seems to be relatively lower, at certain times during the trading session we can clearly see the presence of high gPDC values for short time durations (represented as spikes on the plots in Figure 6.17). To further investigate what might be the cause behind these spikes, we study the mid-price dynamics of the exchange rates, and present the comparative results in Figure 6.18. The plots represent $|\overline{\pi}_i(f)|^2$ values for EURUSD and USDJPY (the two major *driving* currency pairs during this trading session) at a set of discrete higher frequencies, together with their respective mid-price time series; as previously discussed, the quantity $|\overline{\pi}_i(f)|^2 = 1 - |\pi_{i\to i}(f)|^2$ gives us the fraction of the power density of time series i which is providing "explanatory" information to all the other time series under analysis. We note that the peaks in the $|\pi_{ij}(f)|^2$ plots at higher

frequencies (the source of the spikes in the spectrograms) correspond to sudden big moves in the spot market. This implies that the amount of power transferred (the strength of the causal links) by these *driving* currency pairs to other pairs increases substantially at times of market turmoil (which could possibly be occurring due to news releases). Hence, we can infer that currency pairs that exhibit asymmetric interactions are generally (during periods of normal market activity) causally linked only at lower frequencies, however, the links become causal at all frequencies at times of market rallies or collapses. This is possibly because at times of high market volatility, significantly more market participants actively trade in order to take on risk or to cover their existing positions, hence the quote (as well as trading) frequency increases substantially, resulting in reducing the effect of market microstructure noise.

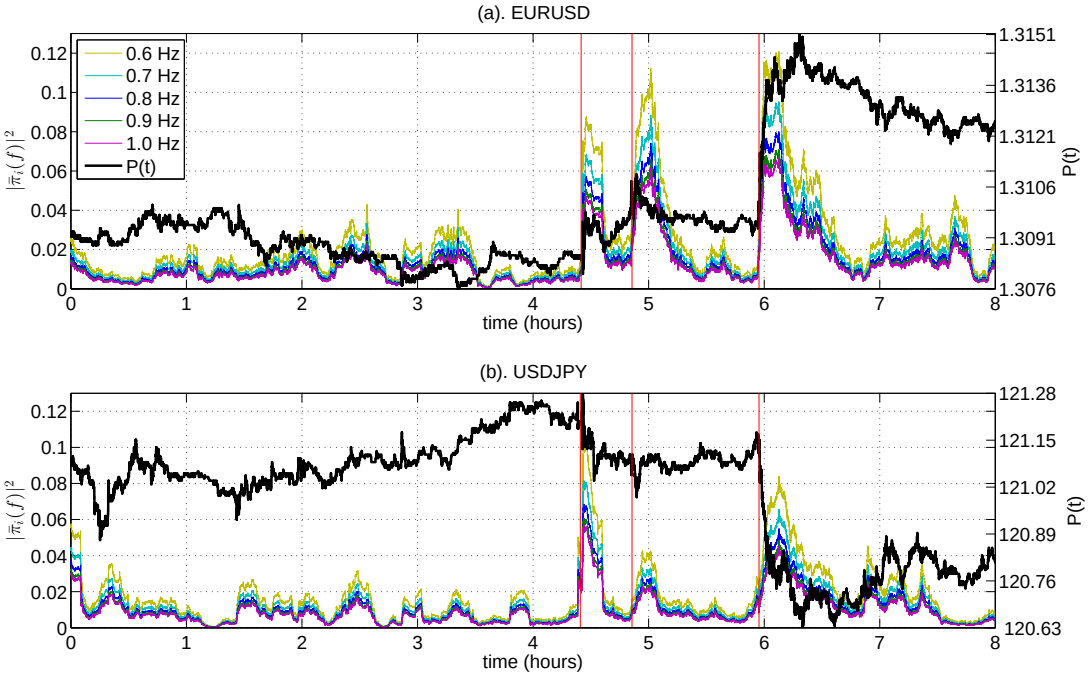

Fig. 6.18: Temporal variation of the $|\bar{\pi}_i(f)|^2$ statistic at a set of discrete higher frequencies for EURUSD and USDJPY; the quantity $|\bar{\pi}_i(f)|^2 = 1 - |\pi_{i \to i}(f)|^2$ gives us the fraction of the power density of time series i which is providing "explanatory" information to all the other time series under analysis (see text for details). Also shown are plots representing the mid-price dynamics, $P(t)$, of these two currency pairs (the bold black lines), with the right y-axes representing the spot rate values. The red vertical lines represent times of sudden moves in the spot rates.

6.3 Conclusions

Asymmetric approaches to interaction measurement have numerous practical applications in the financial sector, such as inferring causation in a set of instruments or for forecasting purposes. Unlike symmetric interaction measurement approaches, most asymmetric approaches are primarily based on the principle of improvement in the predictability of a variable based on the past information contained in one or more other variables. This makes them well-suited for developing forecasting models in a multivariate environment. The FX market provides us with an ideal setting for utilising causality detection models to extract useful information about asymmetric interactions. This is because different currency pairs capture and reflect new market information in their spot prices at varying rates, hence inducing causal effects. In this thesis we have presented a set of efficient causality detection approaches based primarily on the MAR model and ICA. These approaches are suitable for use with multivariate financial log-returns, as indicated by our theoretical arguments and empirical results. Using synthetic and financial data examples, we compared results obtained using our approaches with a standard (ML based) linear Granger causality model, which is the most frequently used causality detection approach in practise. Using a set of practical case studies, we demonstrated utility of our proposed causality detection and forecasting approaches for the purpose of extracting interesting and practically useful information from multivariate financial time series.

To accurately measure asymmetric interactions in non-Gaussian data streams in a computationally efficient framework, we presented the Granger independent component (GIC) causal inference approach. Unlike standard approaches, the GIC model is able to infer causation even in the presence of very weak asymmetric interactions in non-Gaussian signals and, due to its low computational cost, can be easily deployed in an online dynamic environment for analysing high-frequency sampled data. This makes it particularly useful for analysing financial returns, which often exhibit asymmetric interactions that dynamically change in strength and direction. We also presented the ICA-AR model (a variant of the GIC model) which can be used for sequential prediction of univariate non-Gaussian time series. By making use of the sliding-window based decorrelating manifold approach to ICA and by using a reciprocal cosh source model, the GIC and ICA-AR models are not only computationally efficient but also provide improved results for the ICA unmixing process. The GIC model also provides us with

a causal strength statistic (which measures the strength of any given causal link) and enables us to set its critical value, hence allowing us to manage the level of uncertainty we require for any given application (as we will have more confidence in the accuracy of causal links detected using a higher critical threshold value). The second causality detection approach we presented in the previous chapter is based on the VB-MAR model and the principles of Granger causality; we called this the variational Granger (VG) approach to causality detection. We demonstrated (theoretically and empirically) the utility of this approach for accurately and efficiently measuring asymmetric interactions (across both time- and frequency-domains) in multivariate systems by accurately estimating the model parameters and the model order. We also showcased use of the VG model to analyse non-linear Granger causality in multivariate data sets. However, the VG approach assumes Gaussian MAR model residuals, and is thus not well suited for dynamically analysing high-frequency sampled financial data (as windowed high-frequency financial data can be highly non-Gaussian), which is one of its limitations.

Chapter 7
Summary and future directions

7.1 Summary

We are now in a position to re-evaluate the work presented in this thesis in light of our two major objectives, as presented towards the start of the thesis. Our first main goal was the development of efficient approaches for symmetric and asymmetric interaction measurement between multivariate non-Gaussian data streams, with the aim of analysing financial returns. Our second goal was to demonstrate the utility of these approaches by extracting interesting and practically useful information from multivariate financial time series, focusing on the FX spot market. We started the thesis by critically examining the existing literature and identifying the merits and limitations of various standard interaction measurement approaches. We pointed to the current lack of availability of suitable non-Gaussian models of interaction for real-time data analysis by noting that most standard measures have certain limitations, either based on their underlying assumptions, computational complexity or the amount of data they require for accurate computation.

To address these issues, we presented a set of symmetric and asymmetric approaches to efficiently measure interactions in multivariate financial data streams. For the symmetric case, we presented an ICA-based approach which can be used to measure information coupling, as a proxy measure for mutual information. This approach makes use of ICA as a tool to capture information in the tails of the underlying distributions, and is suitable for efficiently and accurately measuring symmetric interactions between multiple non-Gaussian signals. As far as we know, this is the first attempt to measure multivariate interactions using information encoded in the ICA unmixing matrix. Our proposed information coupling model has multiple other benefits associated with its practical use. It provides a framework for estimating con-

fidence bounds on the information coupling metric, can be efficiently used to directly model dependencies in high-dimensional spaces and gives normalised, symmetric results. The model makes use of a sliding-window ICA approach to estimate the ICA unmixing matrix, which results in increased accuracy and efficiency of the algorithm; this results in a model which has computational complexity similar to that of linear correlation with the accuracy of mutual information. It has the added advantage of not depending on any user-defined parameters, in contrast to some other competing interaction measurement approaches, and is not data intensive, i.e. it can be used even with relatively small data sets without significantly effecting its performance, an important requirement for analysing data with rapidly changing dynamics such as financial returns. We also proposed a number of extensions to the information coupling model in order to accurately capture time- and scale-based dynamics of interactions as well as to analyse static and dynamic complex coupling networks. By noting that financial returns exhibit rapidly changing dynamics, often characterised by regions of quasi-stability punctuated by abrupt changes, we proposed that changes in their underlying information coupling may be captured using a Markov process model with switching states, and hence presented use of the HMICA model to capture variations in information coupling dynamics. We showed that changes in information coupling dynamics in financial returns can generally be modelled using a 2-state HMICA model, with each state representing regions of low and high coupling, hence, making it possible for us to detect regimes of the data showing persistence in information coupling values. Knowing that many real-world signals, including financial returns, exhibit time-scale behaviour, we presented the wavelet-ICA model, an extension of the information coupling model, that can be used to analyse dependencies at different time-scales by making use of the continuous wavelet transform, with a Morlet basis function. We noticed that financial returns become increasingly coupled together across scale, pointing to the presence of long-range dependencies. To more efficiently analyse the information coupling structure in high-dimensional systems, we combined techniques used for building minimum spanning trees with the ICA-based information coupling model to obtain complex coupling networks, which can be used to extract the hierarchical interaction structure from multivariate financial data. We extended our analysis to study the dynamics of complex coupling networks and demonstrated the utility of information gained using these networks in portfolio selection

applications and for identifying currency pairs which are "in play" at any given time. We also noticed that the structure of dynamic coupling networks, as analysed using their survival ratios, becomes increasingly stable across scale, pointing to rapid variations in the strength of coupling in financial systems at higher frequencies. To analyse the utility, efficiency and accuracy of our proposed models, we carried out an in-depth comparative study with some other standard measures of symmetric interaction, using a range of synthetic and financial data examples. We also presented a set of practical financial case studies, which utilised various measures of symmetric interaction, and discussed the merits of using different approaches in each case. The empirical results we obtained backed our theoretical arguments concerning the suitability of the ICA-based information coupling model for accurately and efficiently modelling dependencies in multivariate financial systems in an online dynamic environment. Some other possible improvements and extensions to the model are proposed in the future work section.

We next turned to developing efficient approaches for analysing asymmetric interactions, i.e. causality, in multivariate non-Gaussian data streams with the aim of analysing financial returns. Asymmetric interaction measurement approaches, as opposed to symmetric ones, are based on the principle of measurement of information flow at non-zero time lags, and hence are well-suited for forecasting purposes in multivariate environments. This makes them a very attractive choice for predicting changes in FX spot rates, as FX returns are interlinked due to the effect of a range of macro- and micro-economic factors. There is often a time-delay introduced in the price formation process for relatively less liquid FX currency pairs, as new information is captured and reflected in their prices at a slower rate. This phenomena also induces causal effects in multivariate FX returns; information about which can be retrieved using causality detection approaches. We presented two causal inference approaches in this thesis. One of these, the variational Granger (VG) model, is suitable for analysing large, high-dimensional, data sets (across both time- and frequency-domains) and provides accurate results as compared to a standard (ML based) Granger causality model by preventing model overfitting. It also provides us with a unified framework under which we can accurately perform parameter estimation and model order selection for causal inference. Moreover, it enables us to estimate uncertainties associated with the estimated parameters, hence, allowing us to fold in these uncertainties by

giving more "significance" to weights with lower uncertainty when inferring causation (and vice-versa). However, the VG approach is based on the assumption of Gaussian residuals, which is one of its limitations. This brings us to our second causality detection approach, which addresses this issue. The Granger independent component (GIC) causal inference approach, which is based on a combination of ICA, MAR and Granger causality models, is suitable for measuring asymmetric interactions in multivariate non-Gaussian data streams in a computationally efficient manner. By assuming the MAR model residuals to be non-Gaussian and serially independent, the GIC model allows us to make use of ICA to take into account the higher-order moments of the data while inferring causation, hence resulting in improved accuracy when dynamically analysing multivariate financial returns. We also presented the ICA-AR model (a variant of the GIC model) that can be used for sequential prediction of univariate non-Gaussian time series. We carried out a detailed comparative study of the efficiency, utility and accuracy of our proposed asymmetric approaches with a standard Granger causality model by testing the ability of different approaches to accurately find a causal structure in both synthetic and financial data streams. By making use of a set of practical financial case studies, we demonstrated the utility of our proposed asymmetric interaction measurement and forecasting approaches for extracting interesting and useful information from financial data. We propose some possible improvements to the causality analysis approaches in the future work section which follows.

7.2 Future directions

The interaction measurement approaches we have presented in this thesis are accurate, robust and computationally efficient. However, there is room for further refining and extending them, as we now discuss. Most of the approaches presented in this thesis are valid under the assumption that the data under analysis is stationary. This assumption can be made because in the log-returns space financial data can be considered to be locally stationary. Moreover, the sliding-window ICA algorithm which we use for analysis is also good at handling non-stationary data. However, the assumption of local stationarity might not hold if mid-prices are used directly instead of the log-returns (as may be required for certain applications) or if a very large log-returns data set needs to be analysed. To deal with such a scenario, non-

stationary ICA models can be used, as discussed in [121, 123]; these models are essentially based on the assumption that the latent source signals are stationary while the mixing process is non-stationary. Hence, all models making use of ICA in this thesis (such as information coupling, GIC and ICA-AR) can be extended to deal with non-stationary signals (and hence used to directly analyse mid-prices). Likewise, the VG causality detection approach is valid for stationary data sets. However, we can adapt it to deal with non-stationary data by making use of non-stationary VB-MAR techniques, as presented in [66]. Another study providing a good starting point towards developing non-stationary statistical models for financial time series analysis is presented in [148]. The information coupling model we have presented is based on the assumption of linear mixing of latent source signals. This assumption is valid in dynamic environments as financial log-returns can be considered to be locally linear. However, if larger financial data sets need to be analysed, this assumption may not always hold. To address this issue, it is possible to make use of non-linear ICA models to dynamically estimate the unmixing matrix, some of which are described in [10, 203, 304]. However, most of these (and other similar) approaches are computationally complex, therefore the benefits of using these models should be carefully weighed against their limitations. Once the unmixing matrix is obtained using non-linear ICA, the metric presented in this thesis can be used to calculate information coupling. Similarly, performance of the GIC and ICA-AR models can be potentially improved by making use of non-linear ICA (when analysing larger data sets). Another possible way forward is to make use of local linear ICA for obtaining a more accurate estimation of information coupling in non-linear systems. We can achieve this in three main steps to obtain an estimate for non-linear information coupling. The first step involves using some clustering method for dividing the data set into separate clusters [216]. We can then calculate information coupling within each cluster, using the linear ICA-based information coupling model, and hence estimate the global information coupling using the estimates for each separate cluster. In theory, we can use any of the commonly used clustering techniques, e.g. the k-means clustering algorithm; however, a major problem associated with using such methods is choosing the number of clusters beforehand. Fortunately, we can address this problem by using a VB inference approach to estimate the most likely number of clusters in the data space [82, 305]. We can then develop a model which dynamically switches between the linear ICA

coupling measure and the non-linear one based on the temporal properties of the data being analysed.

The independent components extracted by ICA are not always mutually independent. Independent subspace analysis (ISA), an extension of the standard ICA model, assumes the components are divided into subspaces and components in different subspaces are assumed to be independent, whereas components in the same subspace have dependencies [185, 188]. Therefore, certain dependencies between the independent components can be modelled using ISA. Because of the local dependencies between different currency pairs, ISA can potentially have very useful applications in modelling dependencies in multivariate financial time series. This can lead to an ISA-coupling model which measures coupling within each subspace, while treating individual subspaces as independent entities. It will also be interesting to make use of topographic ICA (a generalisation of ISA) for measuring interactions in financial returns. Topographic ICA is a generative model that combines topographic mapping with ICA [186]. In topographic mappings, the distance in the representation grid is related to the distance of the represented components. Likewise, in topographic ICA, the distance between represented components is defined by the mutual information implied by the higher-order correlations, which gives the natural distance measure in the context of ICA [186]. The topography in topographic ICA can be useful for visualising relationships between the independent components. Earlier we had described the use of structured priors for causal inference and presented a set of synthetic data examples demonstrating some of their uses. It will be interesting to further investigate the utility of using such priors for analysing real financial data, for example, to study the dominant types of interactions present in multivariate financial time series or to highlight individual financial instruments that are contributing most to the predictive power of the model. Another area of active research, with some very diverse and important applications in the financial markets, is online changepoint detection [219, 222]. Changepoints can occur in various features of the multivariate time series, e.g. in amplitude, mixing process, volatility and possibly in some latent variables. The changepoint model developed has to be causal in order to be useful in practical applications. An interesting starting point for such a model can be the Bayesian online changepoint detection algorithm [5]. However, as this algorithm (and most other changepoint detection algorithms) only detects changepoints in a univariate time series,

therefore, a multivariate online changepoint detection algorithm can be developed which can be used to detect changepoints in multivariate financial time series. In Chapters 4 and 6, we presented some financial applications of interaction measurement approaches for prediction purposes. It is possible to combine information about a system's dependency structure with some standard forecasting models to obtain improved performance. There are various methods that can be used for this purpose, e.g. the Kalman filter (KF), extended KF or non-linear methods such as artificial neural networks (ANN). A very brief overview of some useful forecasting methods is given below. A KF measures the states of a dynamic system with a series of noisy measurements which can be used for prediction purposes [279]. Due to their computational efficiency and accuracy, KFs are potentially very useful for the online forecasting of financial time series [155, 279]. It is also possible to develop a predictive ICA model which forecasts the independent components. These components can then be mixed together to obtain an estimate for the predicted value of the observed signals. As this model uses the underlying sources to make future predictions, therefore it can potentially outperform standard prediction methods [180]. This can result in a wavelet-ICA-KF model, that forecasts the independent components at different scales using KFs. Another possibility is to make use of ANNs, a well-known non-linear method for time series prediction [168]. Knowledge about a system's information coupling or causality can potentially be incorporated as prior information in ANNs in order to obtain improved prediction results for multivariate time series [228].

The scale and complexity of modern financial markets presents great challenges and opportunities for the successful implementation of various statistical signal processing models. We provided some possible applications of such models in the financial markets in Chapters 4 and 6. There are numerous other financial applications where these models can play a central role, some of which we now discuss. Improved statistical arbitrage models employing various signal processing techniques can be developed. These models make use of statistical mispricing in an asset's value to generate positive returns [341]; the triangulated proxy exchange rate example we presented in this thesis is one such model. ICA can also be used to extract the underlying structure of multivariate financial time series, a process which can aid in the development of an efficient statistical arbitrage model [356]. Wavelets can be used to identify arbitrage opportunities at different time-scales [359]. Managing risk associated with

the returns of a financial asset is of great importance [175]. Earlier we had presented a set of practical case studies which demonstrated the utility of both symmetric and asymmetric interaction measurement approaches for value-at-risk (VaR) estimation. Specific models dealing with the active management of VaR can be developed and used as part of a real-time trading (RTT) model or a portfolio to monitor and manage risk [104]. As ICA focuses on higher-order statistics, therefore it can potentially have very useful applications in the management of financial risk [79]. Managing risk associated with financial derivatives is also an area of interest in financial markets [164], therefore, risk management models can be developed for financial derivatives as well. RTT models make use of various signal processing techniques to make online and informed trading decisions [135]. Models developed for finding information coupling and causality in multivariate financial returns, forecasting financial time series, and risk management of financial assets, can be combined to form RTT models. One possible example is a pairs trading model. It is often the case that a set of financial instruments are coupled due to some fundamental sector based reasons, this fact was also exhibited by the equities complex coupling networks we presented in Figures 4.32 and 4.33. Pairs trading models make use of two closely coupled instruments and place trades as soon as these instruments become decoupled for short time periods. It is also possible to develop pairs trading models operating at different time-scales using scale-dependent information coupling models. Earlier we had presented an application of the information coupling model to analyse a GMV portfolio. There are various other possible applications of interaction measurement approaches for portfolio optimisation, such as development of maximum return portfolios or portfolios with other pre-defined return-to-risk profiles. Information about coupling at different time-scales can potentially be useful for estimating the time period for which assets need be kept in a portfolio.

Appendix A
Inference in hidden Markov ICA models

A Markov model is a statistical process in which future probabilities are determined by only its most recent values. Using the product rule, the joint probability of a variable x can be written as [41]:

$$p(x_1,...,x_T) = \prod_{t=1}^{T} p(x_t \mid x_1,...,x_{t-1}) \tag{A.1}$$

A first-order Markov model assumes that all the conditional probabilities in the product part of (A.1) are dependent on only the most recent observation and independent of all others. Thus, a first order Markov model can be represented by:

$$p(x_1,...,x_T) = p(x_1)\prod_{t=2}^{T} p(x_t \mid x_{t-1}) \tag{A.2}$$

A hidden Markov model (HMM) is a statistical model consisting of a set of observations which are produced by an unobservable set of latent states, the transitions between which are a Markov process. Mathematically, the model can be represented by [41]:

$$p(X \mid Z, \boldsymbol{\theta}) = p(z_1 \mid \boldsymbol{\pi}) \left\{ \prod_{t=2}^{T} p(z_t \mid z_{t-1}, \mathbf{P}_{hmm}) \right\} \prod_{t'=1}^{T} p(x_{t'} \mid z_{t'}, \mathbf{B}_{hmm}) \tag{A.3}$$

where $X = [x_1,...,x_T]$ is the observation set, $Z = [z_1,...,z_T]$ is the set of latent variables and $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{P}_{hmm}, \mathbf{B}_{hmm}\}$ represents the set of parameters governing the model, with $\mathbf{B}_{hmm}$ denoting the set of parameters of the observation model, $\mathbf{P}_{hmm}$ denoting the state transition matrix with entries p_{ij}, and $\boldsymbol{\pi}$ denoting the initial state probability matrix. The HMM can be trained using an expectation-maximisation (EM) algorithm, as described in [101].

We may combine ICA and HMM to form the hidden Markov ICA (HMICA) model [288], which may be seen as a HMM with an ICA observation model (i.e. $\mathbf{B}_{hmm}$ contains parameters of the ICA observation model). For a HMICA parameterised by some vector $\hat{\boldsymbol{\theta}}$, the EM algorithm requires us to maximise an auxiliary function $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, where:

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = Q(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) + Q(\mathbf{P}_{hmm}, \hat{\mathbf{P}}_{hmm}) + Q(\mathbf{B}_{hmm}, \hat{\mathbf{B}}_{hmm}) \tag{A.4}$$

It is possible to obtain parameter update equations for each part of the model by separately maximising the three terms on the right-hand-side of (A.4) (details are presented in [288, 298]). The update equa-

tions obtained via maximising $Q(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}})$ and $Q(\mathbf{P}_{hmm}, \hat{\mathbf{P}}_{hmm})$ are the same as those for a HMM model. Analysis presented in [288] shows that for an observation sequence $\mathbf{x}_t$ the term for the observation model parameters is given by:

$$Q(\mathbf{B}_{hmm}, \hat{\mathbf{B}}_{hmm}) = \sum_k \sum_t \gamma_k[t] \log p(\mathbf{x}_t \mid z_t) \tag{A.5}$$

where $\gamma_k[t]$ is the probability of being in state k. The log-likelihood of the ICA observation model (as given earlier by (3.22), assuming noiseless mixing), with unmixing matrix $\mathbf{W}$ and M sources can be written as [288]:

$$\log p(\mathbf{x}_t) = \log |\det(\mathbf{W})| + \sum_{i=1}^{M} \log p(a_i[t]) \tag{A.6}$$

where $a_i[t]$ is the i-th source component (generated as a result of a generalised AR process with non-Gaussian noise). Substituting the ICA log-likelihood ((A.6)) into the HMM auxiliary function ((A.5)) gives:

$$Q_k = \log |\det(\mathbf{W}_k)| + \frac{\sum_t \gamma_k[t] \sum_i \log p(a_i[t])}{\sum_t \gamma_k[t]} \tag{A.7}$$

where $\mathbf{W}_k$ is the unmixing matrix corresponding to state k. The auxiliary function, summed over all states k, is hence:

$$Q = \sum_k Q_k \tag{A.8}$$

The HMICA model finds the unmixing matrix $\mathbf{W}_k$ for state k, by minimising the cost function given by (A.8) over all underlying parameters using a set of iterated update equations, as described in detail in [288].

Appendix B
Inference in variational Bayesian MAR models

B.1 Updating parameters of the VB-MAR model

Here we present the main steps for inferring the parameters of a MAR model within the VB framework. A comprehensive description of these update equations is presented in [291]. We first present the outline of the method used to update the weights of the model. Let:

$$I(\mathbf{w}) = \int\int q(\mathbf{\Lambda} \mid \mathbf{D}) q(\varsigma \mid \mathbf{D}) \log[p(\mathbf{D} \mid \mathbf{w}, \mathbf{\Lambda}) p(\mathbf{w} \mid \varsigma)] d\varsigma d\mathbf{\Lambda} \tag{B.1}$$

Then, substituting (5.30) and (5.39) into (5.34) (and dropping those terms which are not a function of $\mathbf{w}$), the negative free energy is given by [291]:

$$F(p) = -KL[q(\mathbf{w} \mid \mathbf{D}), \exp\{I(\mathbf{w})\}] \tag{B.2}$$

This term is maximised when:

$$q(\mathbf{w} \mid \mathbf{D}) \propto \exp\{I(\mathbf{w})\} \tag{B.3}$$

Now, substituting (5.22) and (5.27) into (B.1) gives:

$$I(\mathbf{w}) = -\int q(\mathbf{\Lambda} \mid \mathbf{D}) \mathrm{Tr}(\mathbf{\Lambda} \mathbf{E}_{\mathbf{D}}(\mathbf{w})) d\mathbf{\Lambda} - \int q(\varsigma \mid \mathbf{D}) \varsigma E(\mathbf{w}) d\varsigma \tag{B.4}$$

Defining $\hat{\mathbf{\Lambda}}$ and $\hat{\varsigma}$ as the noise and weight precisions from the approximating densities, (B.4) can be simplified to:

$$I(\mathbf{w}) = -\frac{1}{2} \mathrm{Tr}(\hat{\mathbf{\Lambda}} \mathbf{E}_{\mathbf{D}}(\mathbf{w})) - \hat{\varsigma} E(\mathbf{w}) \tag{B.5}$$

The weight posterior is therefore a normal density $q(\mathbf{w} \mid \mathbf{D}) = \mathcal{N}(\mathbf{w}; \hat{\mathbf{w}}, \hat{\mathbf{\Sigma}})$, where:

$$\mathbf{\Lambda}_D = \hat{\mathbf{\Lambda}} \otimes (\mathbf{X}^{\mathsf{T}} \mathbf{X}), \quad \hat{\mathbf{\Sigma}} = (\mathbf{\Lambda}_D + \hat{\varsigma} \mathbf{I})^{-1}, \quad \hat{\mathbf{w}} = \hat{\mathbf{\Sigma}} \mathbf{\Lambda}_D \mathbf{w}_{ML} \tag{B.6}$$

in which $\otimes$ represents the Kronecker product of the matrices and $\mathbf{w}_{ML}$ is the ML estimate of the weight parameter. To update the weight precisions of the model, the following approach is taken. Let:

$$I(\varsigma) = \int q(\mathbf{w} \mid \mathbf{D}) \log[p(\mathbf{w} \mid \varsigma) p(\varsigma)] d\mathbf{w} \tag{B.7}$$

Then, following a similar approach as taken to update the weights of the model (but this time dropping all those terms which are not a function of ς), it can be shown that the negative free energy is maximised when [291]:

$$q(\varsigma \mid \mathbf{D}) \propto \exp\{I(\varsigma)\} \tag{B.8}$$

It can also be shown that by substituting the weight and weight precision priors, the weight precision posterior is a Gamma density $q(\varsigma \mid \mathbf{D}) = \text{Ga}(\varsigma; b'_\varsigma, c'_\varsigma)$, where:

$$\frac{1}{b'_\varsigma} = E(\hat{\mathbf{w}}) + \frac{1}{2}\text{Tr}(\hat{\mathbf{\Sigma}}) + \frac{1}{b_\varsigma}, \quad c'_\varsigma = \frac{k}{2} + c_\varsigma, \quad \hat{\varsigma} = b'_\varsigma c'_\varsigma \tag{B.9}$$

Finally, we present the method used to update the noise precisions. Let:

$$I(\mathbf{\Lambda}) = \int q(\mathbf{w} \mid \mathbf{D}) \log[p(\mathbf{D} \mid \mathbf{w}, \mathbf{\Lambda}) p(\mathbf{\Lambda})] d\mathbf{w} \tag{B.10}$$

Then, following a similar procedure to that described earlier for updating the model weights and weight precisions (but this time dropping all those terms which are not a function of $\mathbf{\Lambda}$), the negative free energy is maximised when [291]:

$$q(\mathbf{\Lambda} \mid \mathbf{D}) \propto \exp\{I(\mathbf{\Lambda})\} \tag{B.11}$$

By substituting the weight and weight precision priors, the noise precision posterior is a Wishart density $q(\mathbf{\Lambda}) = \text{Wi}(\mathbf{\Lambda}; a, \mathbf{B}_\Lambda)$, where:

$$\mathbf{B}_\Lambda = \mathbf{E}_\mathbf{D}(\hat{\mathbf{w}}) + \sum_t (\mathbf{I}_d \otimes \mathbf{x}(t)) \hat{\mathbf{\Sigma}} (\mathbf{I}_d \otimes \mathbf{x}(t))^\mathsf{T}, \quad a = T, \quad \hat{\mathbf{\Lambda}} = a\mathbf{B}_\Lambda^{-1} \tag{B.12}$$

B.2 VB model order selection

The negative free energy, given by (5.34), is an approximation to the likelihood (evidence) of the data [291], and can therefore be used for model order selection. Using (5.22) and (5.39), the average log-likelihood term, as given by (5.37), can be written as:

$$L_{av} = -\frac{dT}{2}\log 2\pi + \frac{T}{2}\int q(\mathbf{\Lambda} \mid \mathbf{D}) \log|\mathbf{\Lambda}| d\mathbf{\Lambda} - \frac{1}{2}\int\int q(\mathbf{\Lambda} \mid \mathbf{D}) q(\mathbf{w} \mid \mathbf{D}) \text{Tr}(\hat{\mathbf{\Lambda}} \mathbf{E}_\mathbf{D}(\mathbf{w})) d\mathbf{\Lambda} d\mathbf{w} \tag{B.13}$$

Noting that the entropy of a Wishart distribution is given by [255]:

$$L(a, \mathbf{B}_\Lambda) = \int \text{Wi}(\mathbf{\Lambda}; a, \mathbf{B}_\Lambda) \log|\mathbf{\Lambda}| d\mathbf{\Lambda} \tag{B.14}$$

the average log-likelihood can be expressed as [291]:

$$L_{av} = -\frac{dT}{2}\log 2\pi e + \frac{T}{2}L(a,\mathbf{B}_\Lambda) \tag{B.15}$$

Substituting (B.15) into (5.38) [291]:

$$F(p) = -\frac{T}{2}\log|\mathbf{B}_\Lambda| - KL[q(\mathbf{w}\mid\mathbf{D}),p(\mathbf{w})] - KL[q(\varsigma\mid\mathbf{D}),p(\varsigma)] + \log\Gamma_d\left(\frac{T}{2}\right) \tag{B.16}$$

where Γ_d is the generalised gamma function. The last term of (B.16) is constant for any given value of T and d, and therefore has no effect when the negative free energy method is used for model order selection [291]. The optimum model order is one which maximises the value of the negative free energy, as given by (B.16). We note that as the number of samples increases, i.e. $T\to\infty$, $F(p)$ becomes equivalent to the BIC [78].

B.3 Structured priors

Using (5.27) as the basis, structured priors take the following form [291]:

$$p(\mathbf{w}\mid\varsigma_g) = \prod_{g=1}^{G_n}\left(\frac{\varsigma_g}{2\pi}\right)^{\frac{k_g}{2}}\exp\left[-\varsigma_g E_g(\mathbf{w})\right] \tag{B.17}$$

where $g=1,2,...,G_n$ indexes the G_n different groups of weight parameters, k_g is the number of weights in the g-th group and $E_g(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{I}_g\mathbf{w}$ (where the diagonal indicator matrix $\mathbf{I}_g$ is used to pick off coefficients in the g-th group). When using structured priors, the posterior weight covariance update equation can be rewritten as:

$$\hat{\mathbf{\Sigma}} = \left(\mathbf{\Lambda}_D + \sum_{g=1}^{G_n}\hat{\varsigma}_g\mathbf{I}_g\right)^{-1} \tag{B.18}$$

All other weight update equations remain the same. Likewise, the weight precision update equations can be rewritten as:

$$\frac{1}{b'_\varsigma(g)} = E_g(\hat{\mathbf{w}}) + \frac{1}{2}\mathrm{Tr}(\mathbf{I}_g\hat{\mathbf{\Sigma}}\mathbf{I}_g) + \frac{1}{b_\varsigma}, \quad c'_\varsigma(g) = \frac{k_g}{2} + c_\varsigma, \quad \hat{\varsigma}(g) = b'_\varsigma(g)c'_\varsigma(g) \tag{B.19}$$

The update equations for the noise precision matrix remain the same, as they are independent of the MAR coefficients. We can make use of negative free energy to estimate the evidence for different models which make use of different types of structured priors, with models resulting in a higher evidence signifying the suitability of a given prior for analysing the data at hand [291]. We now describe a set of structured priors which may be useful for analysing multivariate financial time series (although it is possible to define various other types of priors based on the requirements of the user).

- *Global*: The most obvious choice of prior, with one group with equal weights.
- *Lag*: As the coefficients of a MAR model are associated with different time-lags, therefore we can define a *Lag* prior, whereby the coefficients are split into different groups depending on the

time-lags with which they are associated. This results in p groups, each picking out weights associated with different time-lags.

- *Interaction*: It is also possible to make use of *Interaction* priors, in which the MAR model coefficients are grouped into two separate groups based on whether they relate to within-series predictions or between-series predictions (interactions) [291]. These priors are suitable for data sets which result in weight matrices in which the weights naturally lie in such a way that the magnitude of the diagonal and off-diagonal elements are of relatively similar values. The resulting prior has two groups, one for within-series weights and one for between-series weights.

- *Lag-interaction*: These priors can be used to group the within-series and between-series prediction coefficients at each time-lag into separate groups. This will result in two separate groups of priors at each time-lag, e.g. for a MAR model of order p, there will be $2p$ separate *Lag-interaction* prior groups. As an example, for a bivariate system, weights for the within-series and the between-series priors (at $p = 1$) can be picked off using the matrices:

$$\hat{\mathbf{I}}_{ws,1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{\mathbf{I}}_{bs,1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{B.20}$$

respectively. Similarly, at $p = 2$, the weights can be picked off using the matrices:

$$\hat{\mathbf{I}}_{ws,2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{\mathbf{I}}_{bs,2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{B.21}$$

and so on. These matrices can then be converted into their equivalent diagonal form (for use in analysis presented previously) using the transformations $\mathbf{I}_g = \text{diag}\left[\text{vec}\left(\hat{\mathbf{I}}_{ws,p}\right)\right]$, etc.

Bibliography

[1] Triennial central bank survey of foreign exchange and derivatives market activity in April 2010. *Bank for International Settlements, Monetary and Economic Department, Switzerland*, 2010.

[2] The NASDAQ OMX Group, Inc. 2012.

[3] Wavelet toolbox documentation. *The MathWorks, Inc.*, 2012.

[4] F. Abramovich, T. Sapatinas, and B.W. Silverman. Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):725–749, 1998.

[5] R.P. Adams and D. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[6] P.S. Addison. *The illustrated wavelet transform handbook: Introductory theory and applications in science, engineering, medicine and finance*. Taylor & Francis, 2010.

[7] W. Addison and S. Roberts. Blind source separation with non-stationary mixing using wavelets. *ICA Research Network Workshop, The University of Liverpool*, 2006.

[8] Y. Aiba, N. Hatano, H. Takayasu, K. Marumo, and T. Shimizu. Triangular arbitrage as an interaction among foreign exchange rates. *Physica A: Statistical Mechanics and its Applications*, 310(3-4):467–479, 2002.

[9] I. Aldridge. *High-frequency trading: A practical guide to algorithmic strategies and trading systems*, volume 459. Wiley, 2009.

[10] L.B. Almeida. Linear and nonlinear ICA based on mutual information. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 117–122. IEEE, 2000.

[11] N. Ancona, D. Marinazzo, and S. Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70(5):56221, 2004.

[12] T.G. Andersen. *Handbook of financial time series*. Springer, 2009.

[13] T.G. Andersen, T. Bollerslev, P.F. Christoffersen, and F.X. Diebold. Volatility and correlation forecasting. *Handbook of economic forecasting*, 1:777–878, 2006.

[14] T.G. Andersen, T. Bollerslev, F.X. Diebold, and C. Vega. Micro effects of macro announcements: Real-time price discovery in foreign exchange. *American Economic Review*, 2003.

[15] C. Andrieu, N.D. Freitas, A. Doucet, and M.I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1):5–43, 2003.

[16] A. Ang and G. Bekaert. Stock return predictability: Is it there? *Review of Financial Studies*, 2006.

[17] R. Artusi, P. Verderio, and E. Marubini. Bravais-Pearson and Spearman correlation coefficients: Meaning, test of hypothesis and confidence interval. *The International journal of biological markers*, 17(2):148–151, 2002.

[18] I. Asimakopoulos, D. Ayling, and W.M. Mahmood. Non-linear Granger causality in the currency futures returns. *Economics Letters*, 68(1):25–30, 2000.

[19] L. Astolfi, F. Cincotti, D. Mattia, M.G. Marciani, L.A. Baccala, F. Fallani, S. Salinari, M. Ursino, M. Zavaglia, and F. Babiloni. Assessing cortical functional connectivity by partial directed coherence: Simulations and application to real data. *Biomedical Engineering, IEEE Transactions on*, 53(9):1802–1812, 2006.

[20] S.P. Baca, B.L. Garbe, and R.A. Weiss. The rise of sector effects in major equity markets. *Financial Analysts Journal*, 56(5):34–40, 2000.

[21] L.A. Baccalá et al. Generalized partial directed coherence. In *Digital Signal Processing, 2007 15th International Conference on*, pages 163–166. IEEE, 2007.

[22] L.A. Baccalá and K. Sameshima. Partial directed coherence: A new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474, 2001.

[23] A.D. Back and A.S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International journal of neural systems*, 8(04):473–484, 1997.

[24] E. Baek and W. Brock. A general test for nonlinear Granger causality: Bivariate model. *Technical Report, Iowa State University and University of Wisconsin, Madison.*, 1992.

[25] Z. Bai, W. Wong, and B. Zhang. Multivariate linear and nonlinear causality tests. *Mathematics and Computers in Simulation*, 81(1):5–17, 2010.

[26] R.T. Baillie and T. Bollerslev. Intra-day and inter-market volatility in foreign exchange rates. *The Review of Economic Studies*, 58(3):565–585, 1991.

[27] E. Balaban, A. Bayar, and J. Ouenniche. High-frequency distribution of foreign exchange changes. *Congrès ASAC 2004, Quebec, Canada*, 2004.

[28] L. Barnett, A.B. Barrett, and A.K. Seth. Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters*, 103(23):238701, 2009.

[29] A.B. Barrett and L. Barnett. Granger causality is designed to measure effect, not mechanism. *Frontiers in neuroinformatics*, 7, 2013.

[30] T. Bayes. Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, London*, 1764.

[31] M.J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

[32] J.E. Beasley, N. Meade, and T.J. Chang. An evolutionary heuristic for the index tracking problem. *European Journal of Operational Research*, 148(3):621–643, 2003.

[33] T.S. Beder and C.M. Marshall. *Financial engineering: The evolution of a profession*. Wiley, 2011.

[34] P. Behr, A. Guttler, and F. Miebs. Is minimum-variance investing really worth the while? An analysis with robust performance inference. *Technical Report, Department of Finance, Goethe University, Frankfurt*, 2008.

[35] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[36] C. Bengtsson and J. Holst. On portfolio selection: Improved covariance matrix estimation for Swedish asset returns. In *31st Meeting, Euro Working Group on Financial Modeling*. Citeseer, 2002.

[37] D.A. Bessler and J.L. Kling. A note on tests of Granger causality. *Applied Economics*, 16(3):335–342, 1984.

[38] R. Bhar and S. Hamori. *Hidden Markov models: Applications to financial economics*. Kluwer Academic Publishers, 2004.

[39] A. Bifet and R. Gavalda. Kalman filters and adaptive windows for learning in data streams. In *Discovery Science*, pages 29–40. Springer, 2006.

[40] A. Bifet and R. Gavalda. Learning from time-changing data with adaptive windowing. In *SIAM International Conference on Data Mining*, pages 443–448. Citeseer, 2007.

[41] C.M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[42] V. Bjorn. Multiresolution methods for financial time series prediction. *Computational Intelligence for Financial Engineering, 1995., Proceedings of the IEEE/IAFE 1995*, 1995.

[43] K.J. Blinowska, R. Kuś, and M. Kamiński. Granger causality and information flow in multivariate processes. *Physical Review E*, 70(5):50902, 2004.

[44] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

[45] S. Boettcher and A.G. Percus. Optimization with extremal dynamics. *Physical Review Letters*, 86(23):5211–5214, 2001.

[46] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

[47] T. Bollerslev. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *The Review of Economics and Statistics*, pages 498–505, 1990.

[48] G. Bonanno, G. Caldarelli, F. Lillo, S. Micciche, N. Vandewalle, and R.N. Mantegna. Networks of equities in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):363–371, 2004.

[49] G. Bonanno, F. Lillo, and R.N. Mantegna. High-frequency cross-correlation in a set of stocks. *Quantitative Finance*, 1(1):96–104, 2001.

[50] K. Boudt, J. Cornelissen, and C. Croux. The Gaussian rank correlation estimator: Robustness properties. *Statistics and Computing*, 22(2):471–483, 2012.

[51] C.G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

[52] G.E.P. Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

[53] G.E.P. Box and G.C. Tiao. Bayesian inference in statistical analysis. 1992.

[54] A. Brabazon and M. Neill. Evolving technical trading rules for spot foreign-exchange markets using grammatical evolution. *Computational Management Science*, 1(3):311–327, 2004.

[55] T.J. Brailsford and R.W. Faff. An evaluation of volatility forecasting techniques. *Journal of Banking & Finance*, 20(3):419–438, 1996.

[56] W. Breymann, A. Dias, and P. Embrechts. Dependence structures for multivariate high-frequency data in finance. *Quantitative finance*, 3(1):1–14, 2003.

[57] M.W. Browne and R. Cudeck. Alternative ways of assessing model fit. *Testing structural equation models*, 154:136–162, 1993.

[58] M.D. Buhmann. Radial basis functions. *Acta numerica*, 9:1–38, 2001.

[59] K.P. Burnham and D.R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[60] C. Calderón and L. Liu. The direction of causality between financial development and economic growth. *Journal of Development Economics*, 72(1):321–334, 2003.

[61] J.R. Calderon-Rossel and M. Ben-Horim. The behavior of the foreign exchange rates: Empirical evidence and economic implications. *J. Intern. Business Studies*, 13:99–111, 1982.

[62] J.Y. Campbell and L. Hentschel. No news is good news: An asymmetric model of changing volatility in stock returns. *Journal of Financial Economics*, 31(3):281–318, 1992.

[63] L. Cao and F.E.H. Tay. Financial forecasting using support vector machines. *Neural Computing & Applications*, 10(2):184–192, 2001.

[64] L. Cao and F.E.H. Tay. Support vector machine with adaptive parameters in financial time series forecasting. *Neural Networks, IEEE Transactions on*, 14(6):1506–1518, 2003.

[65] L. Cappiello, R.F. Engle, and K. Sheppard. Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial Econometrics*, 2006.

[66] M.J. Cassidy and W. Penny. Bayesian nonstationary autoregressive models for biomedical signal analysis. *IEEE transactions on biomedical engineering*, 49(10):1142–1152, 2002.

[67] J.E. Cavanaugh and A.A. Neath. Generalizing the derivation of the Schwarz information criterion. *Communications in Statistics-Theory and Methods*, 28(1):49–66, 1999.

[68] A. Cerny. Introduction to fast Fourier transform in finance. *Cass Business School Research Paper*, 2006.

[69] A. Chaboud, B. Chiquoine, E. Hjalmarsson, and C. Vega. Rise of the machines: Algorithmic trading in the foreign exchange market. *International Finance Discussion Papers*, 2009.

[70] F.K.P. Chan, A.W.C. Fu, and C. Yu. Haar wavelets for efficient similarity search of time-series with and without time warping. *Knowledge and Data Engineering, IEEE Transactions on*, 15(3):686–705, 2003.

[71] H. Chao, H. Li-li, and H. Ting-ting. Financial time series forecasting based on wavelet kernel support vector machine. In *Natural Computation (ICNC), 2012 Eighth International Conference on*, pages 79–83. IEEE, 2012.

[72] A. Charpentier, J.D. Fermanian, and O. Scaillet. The estimation of copulas: Theory and practice. *Copulas: From theory to application in finance. Risk Publications*, 2007.

[73] M. Chávez, J. Martinerie, and M.L.V. Quyen. Statistical assessment of nonlinear causality: Application to epileptic EEG signals. *Journal of Neuroscience Methods*, 124(2):113–128, 2003.

[74] B. Chazelle. A faster deterministic algorithm for minimum spanning trees. In *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*, pages 22–31. IEEE, 1997.

[75] Y.L. Chen and Y.F. Gau. News announcements and price discovery in foreign exchange spot and futures markets. *Journal of Banking & Finance*, 34(7):1628–1636, 2010.

[76] R. Cheng. Using Pearson type IV and other cinderella distributions in simulation. In *Simulation Conference (WSC), Proceedings of the 2011 Winter*, pages 457–468. IEEE, 2011.

[77] J. Chiang, Z.J. Wang, and M.J. McKeown. Sparse multivariate autoregressive (MAR)-based partial directed coherence (PDC) for electroencephalogram (EEG) analysis. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 457–460. IEEE, 2009.

[78] D.M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, 29(2):181–212, 1997.

[79] E. Chin, A.S. Weigend, and H. Zimmermann. Computing portfolio risk using Gaussian mixtures and independent component analysis. *Computational Intelligence for Financial Engineering, 1999.(CIFEr) Proceedings of the IEEE/IAFE 1999 Conference on*, pages 74–117, 1999.

[80] S.Z. Chiou-Wei, C.F. Chen, and Z. Zhu. Economic growth and energy consumption revisited: Evidence from linear and nonlinear Granger causality. *Energy Economics*, 30(6):3063–3076, 2008.

[81] E.K.P. Chong and S.H. Zak. *An introduction to optimization*. Wiley-Interscience, 2008.

[82] R. Choudrey and S. Roberts. Variational mixture of Bayesian independent component analyzers. *Neural Computation*, 15(1):213–252, 2003.

[83] O. Ciftcioglu, J.E. Hoogenboom, and H.V. Dam. A consistent estimator for the model order of an autoregressive process. *Signal Processing, IEEE Transactions on*, 42(6):1471–1477, 1994.

[84] R.G. Clarke, H. DeSilva, and S. Thorley. Minimum-variance portfolios in the US equity market. *The Journal of Portfolio Management*, 33(1):10–24, 2006.

[85] S. Clémençon and S. Slim. Statistical analysis of financial time series under the assumption of local stationarity. *Quantitative Finance*, 4(2):208–220, 2004.

[86] R. Coelho, C.G. Gilmore, B. Lucey, P. Richmond, and S. Hutzler. The evolution of interdependence in world equity markets–Evidence from minimum spanning trees. *Physica A: Statistical Mechanics and its Applications*, 376:455–466, 2007.

[87] D.J. Colwell and J.R. Gillett. Spearman versus Kendall. *The Mathematical Gazette*, 66(438):307–309, 1982.

[88] T. Conlon, H.J. Ruskin, and M. Crane. Multiscaled cross-correlation dynamics in financial time-series. *Advances in Complex Systems*, 12(04n05):439–454, 2009.

[89] R. Cont. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[90] R. Cont, M. Potters, and J.P. Bouchaud. Scaling in stock market data: Stable laws and beyond. *Proceedings of the Les Houches workshop, Les Houches, France*, 1997.

[91] G.F. Cooper. An overview of the representation and discovery of causal relationships using Bayesian networks. *Computation, causation, and discovery*, pages 3–62, 1999.

[92] A. Corana. Adaptive box-assisted algorithm for correlation-dimension estimation. *Physical Review E*, 62(6):7872–7881, 2000.

[93] F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

[94] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-interscience, 2006.

[95] A.R. Cowan. Nonparametric event study tests. *Review of Quantitative Finance and Accounting*, 2(4):343–358, 1992.

[96] C. Croarkin, P. Tobias, and C. Zey. *Engineering statistics handbook*. National Institute of Standards and Technology (US), 2001.

[97] J. Crotty. Structural causes of the global financial crisis: A critical assessment of the new financial architecture. *Cambridge Journal of Economics*, 33(4):563–580, 2009.

[98] M.M. Dacorogna and R. Gencay. *An introduction to high-frequency finance*. Academic Press, 2001.

[99] K.B. Datta. *Matrix and linear algebra*. PHI Learning Pvt. Ltd., 2004.

[100] R. Davidson and J.G. MacKinnon. *Econometric theory and methods*. Oxford University Press, 2004.

[101] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[102] S.J. Devlin, R. Gnanadesikan, and J.R. Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):531, 1975.

[103] A. Dias and P. Embrechts. Dynamic copula models for multivariate high-frequency data in finance. *Manuscript, ETH Zurich*, 2004.

[104] F.X. Diebold, J. Hahn, and A.S. Tay. Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange. *Review of Economics and Statistics*, 81(4):661–673, 1999.

[105] C. Diks and J. DeGoede. A general nonparametric bootstrap test for Granger causality. *Global Analysis of Dynamical Systems*, pages 391–403, 2010.

[106] C. Diks and V. Panchenko. A new statistic and practical guidelines for nonparametric Granger causality testing. *Journal of Economic Dynamics and Control*, 30(9-10):1647–1669, 2006.

[107] M. Ding, S.L. Bressler, W. Yang, and H. Liang. Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: Data preprocessing, model validation, and variability assessment. *Biological cybernetics*, 83(1):35–45, 2000.

[108] A. Dionisio, R. Menezes, and D.A. Mendes. Mutual information: A measure of dependency for nonlinear time series. *Physica A: Statistical Mechanics and its Applications*, 344(1-2):326–329, 2004.

[109] B.S. Donefer. Algos gone wild: Risk in the world of automated trading strategies. *The Journal of Trading*, 5(2):31–34, 2010.

[110] C. D'Souza. Where does price discovery occur in FX markets? *Available at SSRN 966446*, 2007.

[111] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72(2):27104, 2005.

[112] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Wiley, 2001.

[113] C. Dunis and M. Williams. Modelling and trading the EUR/USD exchange rate: Do neural network models perform better? *Derivatives Use, Trading and Regulation*, 8(3):211–239, 2002.

[114] C. Dunis and B. Zhou. *Nonlinear modelling of high frequency financial time series*. Wiley, 1998.

[115] P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. *Handbook of heavy tailed distributions in finance*, 8(329-384):1, 2003.

[116] P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: Properties and pitfalls. *Risk management: value at risk and beyond*, pages 176–223, 2002.

[117] F. Emmert-Streib and M. Dehmer. *Information theory and statistical learning*. Springer, 2008.

[118] C. Engel and K.D. West. Exchange rates and fundamentals. *Journal of Political Economy*, 113(3):485–517, 2005.

[119] R.F. Engle and C.W. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, pages 251–276, 1987.

[120] D. Evans. A computationally efficient estimator for mutual information. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 464(2093):1203–1215, 2008.

[121] R. Everson. Non-stationary ICA. In *IEE Seminar Digests*, volume 4, 2004.

[122] R. Everson and S. Roberts. Independent component analysis: A flexible nonlinearity and decorrelating manifold approach. *Neural computation*, 11(8):1957–1983, 1999.

[123] R. Everson and S. Roberts. Non-stationary independent component analysis. In *IEE Conference Publication*, volume 1, pages 503–508. Institution of Electrical Engineers, 1999.

[124] M.P. Fay and M.A. Proschan. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, 4:1, 2010.

[125] N. Feltovich. Critical values for the robust rank-order test. *Communications in Statistics: Simulation and Computation*, 34(3):525–547, 2005.

[126] J.D. Fermanian and O. Scaillet. Some statistical pitfalls in copula modeling for financial applications. In *Capital formation, governance and banking*, pages 59–74. Nova Publishers, 2005.

[127] M. Frenkel, C. Pierdzioch, and G. Stadtmann. The effects of Japanese foreign exchange market interventions on the yen/US dollar exchange rate volatility. *International Review of Economics and Finance*, 14(1):27–39, 2005.

[128] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2):131–163, 1997.

[129] A. Frino and D. Gallagher. Tracking S&P 500 index funds. *Journal of Portfolio Management*, 28(1), 2001.

[130] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny. Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234, 2007.

[131] K.J. Friston, J.T. Ashburner, S.J. Kiebel, T.E. Nichols, and W. Penny. *Statistical parametric mapping: The analysis of functional brain images*. Academic Press, 2011.

[132] P. Fryzlewicz, T. Sapatinas, and S.S. Rao. A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, 93(3):687, 2006.

[133] H.N. Gabow, Z. Galil, T. Spencer, and R.E. Tarjan. Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica*, 6(2):109–122, 1986.

[134] E. Gately. *Neural networks for financial forecasting*. John Wiley & Sons, Inc., 1995.

[135] R. Gençay, G. Ballocchi, M. Dacorogna, R. Olsen, and O. Pictet. Real-time trading models and the statistical properties of foreign exchange rates. *International Economic Review*, 43(2):463–491, 2002.

[136] R. Gencay, F. Selçuk, and B. Whitcher. *An introduction to wavelets and other filtering methods in finance and economics*. Academic Press, 2002.

[137] Z. Ghahramani, M.J. Beal, et al. Variational inference for Bayesian mixtures of factor analysers. *Advances in neural information processing systems*, 12:449–455, 2000.

[138] P.E. Gill, W. Murray, and M.H. Wright. *Practical optimization*. Academic Press, 1981.

[139] C.G. Gilmore, B.M. Lucey, and M.W. Boscia. Comovements in government bond markets: A minimum spanning tree analysis. *Physica A: Statistical Mechanics and its Applications*, 2010.

[140] P. Giot and S. Laurent. Modelling daily value-at-risk using realized volatility and ARCH type models. *Journal of Empirical Finance*, 11(3):379–398, 2004.

[141] G.J. Glasser and R.F. Winter. Critical values of the coefficient of rank correlation for testing the hypothesis of independence. *Biometrika*, 48(3-4):444, 1961.

[142] S. Godsill, A. Doucet, and M. West. Maximum a posteriori sequence estimation using Monte Carlo particle filters. *Annals of the Institute of Statistical Mathematics*, 53(1):82–96, 2001.

[143] L.S. Goldberg. Is the international role of the dollar changing? *Current Issues in Economics and Finance*, 16(1), 2010.

[144] L.S. Goldberg and C. Tille. Vehicle currency use in international trade. *Journal of International Economics*, 76(2):177–192, 2008.

[145] G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J.L. Cantero. Measuring directional coupling between EEG sources. *NeuroImage*, 43(3):497–508, 2008.

[146] C.W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[147] C.W. Granger. Causal inference. *The New Palgrave: Econometrics*, 1987.

[148] C.W. Granger and C. Starica. Non-stationarities in stock returns. *The Review of Economics & Statistics*, 87(3):523–538, 2005.

[149] P. Grassberger and I. Procaccia. Dimensions and entropies of strange attractors from a fluctuating dynamics approach. *Physica D: Nonlinear Phenomena*, 13(1-2):34–54, 1984.

[150] S. Greenland and B. Brumback. An overview of relations among causal modelling methods. *International journal of epidemiology*, 31(5):1030, 2002.

[151] D.M. Guillaume, M.M. Dacorogna, R.R. Davé, U.A. Müller, R.B. Olsen, and O.V. Pictet. From the bird's eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets. *Finance and stochastics*, 1(2):95–129, 1997.

[152] M.C. Guisan. Causality and cointegration between consumption and GDP in 25 OECD countries: Limitations of the cointegration approach. *Applied Econometrics and International Development*, 1-1, 2001.

[153] L. Gulko. Decoupling. *The Journal of Portfolio Management*, 28(3):59–66, 2002.

[154] C.S. Hakkio and M. Rush. Cointegration: How short is the long run? *Journal of International Money and Finance*, 10(4):571–581, 1991.

[155] S.G. Hall. Modelling structural change using the Kalman Filter. *Economics of Planning*, 26(1):1–13, 1993.

[156] A.K. Han. Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, 35(2):303–316, 1987.

[157] M.H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[158] F. Harmantzis and L. Miao. Evolution of developed equity and foreign exchange markets: The minimum spanning tree approach. *Social Science Research Network, Available at SSRN 1397650*, 2009.

[159] L. Harrison, W. Penny, and K. Friston. Multivariate autoregressive modeling of fMRI time series. *NeuroImage*, 19(4):1477–1491, 2003.

[160] M. Haugh. The Monte Carlo framework, examples from finance and generating correlated random variables. *Monte Carlo Simulation: IEOR E4703, Columbia University*, 2004.

[161] M. Havlicek, J. Jan, M. Brazdil, and V.D. Calhoun. Dynamic Granger causality based on Kalman filter for evaluation of functional network connectivity in fMRI data. *NeuroImage*, 53(1):65–77, 2010.

[162] D. Heckerman. A tutorial on learning with Bayesian networks. *Innovations in Bayesian Networks*, pages 33–82, 2008.

[163] D.F. Hendry and K. Juselius. Explaining cointegration analysis: Part II. *The Energy Journal*, pages 75–120, 2001.

[164] L. Hentschel and C.W. Smith. Risk and regulation in derivatives markets. *Journal of Applied Corporate Finance*, 7(3):8–22, 1994.

[165] W. Hesse, E. Möller, M. Arnold, and B. Schack. The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, 124(1):27–44, 2003.

[166] C. Hiemstra and J.D. Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 2012.

[167] N.J. Higham. Matrix nearness problems and applications. *Applications of matrix theory*, 1989.

[168] T. Hill, L. Marquez, M. Connor, and W. Remus. Artificial neural network models for forecasting and decision making. *International Journal of Forecasting*, 10(1):5–15, 1994.

[169] G.E. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on computational learning theory*, pages 5–13. ACM New York, 1993.

[170] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

[171] J. Hlinka, M. Paluš, M. Vejmelka, D. Mantini, and M. Corbetta. Functional connectivity in resting-state fMRI: Is linear correlation sufficient? *NeuroImage*, 2010.

[172] Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[173] B.P.T. Hoekstra, C.G.H. Diks, M.A. Allessie, J. DeGoede, V. Barbaro, P. Bartolini, G. Calcagnini, and G. Boriani. Non-linear time series analysis: Methods and applications to atrial fibrillation. *Ann. Ist. Super. Sanità*, 37:325–333, 2003.

[174] P.W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[175] K.A. Horcher. *Essentials of financial risk management*. Wiley, 2005.

[176] N.V. Hovanov, J.W. Kolari, and M.V. Sokolov. Synthetic money. *International Review of Economics & Finance*, 16(2):161–168, 2007.

[177] D. Howitt and D. Cramer. *First steps in research and statistics: A practical workbook for psychology students*. Routledge, 2000.

[178] D.A. Hsieh. Modeling heteroskedasticity in daily foreign exchange rates. *Journal of Business and Economic Statistics*, 7(3):307–317, 1989.

[179] D.A. Hsieh. Testing for nonlinear dependence in daily foreign exchange rates. *The Journal of Business*, 62(3):339–368, 1989.

[180] L. Huang, F. Li, and L. Xin. ICA/RBF-based prediction of varying trend in real exchange rate. In *Services Computing, 2006. APSCC'06. IEEE Asia-Pacific Conference on*, pages 572–580, 2006.

[181] N.E. Huang, M.L. Wu, W. Qu, S.R. Long, and S.S.P. Shen. Applications of Hilbert-Huang transform to non-stationary financial time series analysis. *Applied Stochastic Models in Business and Industry*, 19(3):245–268, 2003.

[182] J.S. Hunter. The exponentially weighted moving average. *Journal of Quality Technology*, 18(4):203–210, 1986.

[183] D. Husmeier, R. Dybowski, and S. Roberts. *Probabilistic modeling in bioinformatics and medical informatics*. Springer, 2005.

[184] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2(4):94–128, 1999.

[185] A. Hyvärinen and P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[186] A. Hyvärinen, P.O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[187] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley, 2001.

[188] A. Hyvärinen and U. Koster. FastISA: A fast fixed-point algorithm for independent subspace analysis. In *Proceedings of European Symposium on Artificial Neural Networks (ESANN)*, 2006.

[189] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[190] R.L. Iman and W.J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation*, 11(3):311–334, 1982.

[191] C.K. Ing. Multistep prediction in autoregressive processes. *Econometric Theory*, 19(02):254–279, 2003.

[192] A. Inoue and L. Kilian. In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews*, 23(4):371–402, 2005.

[193] K. Ishiguro, N. Otsu, M. Lungarella, and Y. Kuniyoshi. Comparison of nonlinear Granger causality extensions for low-dimensional systems. *Physical Review E*, 77(3):36217, 2008.

[194] R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1684, 2003.

[195] J.P. Jarvis and D.E. Whited. Computational experience with minimum spanning tree algorithms. *Operations Research Letters*, 2(1):36–41, 1983.

[196] E.T. Jaynes. *Probability theory: The logic of science*. Cambridge University Press, 2003.

[197] R.L. Jenison and R.A. Reale. The shape of neural dependence. *Neural computation*, 16(4):665–672, 2004.

[198] H. Joe. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164, 1989.

[199] N.F. Johnson, M. McDonald, O. Suleman, S. Williams, and S. Howison. What shakes the FX tree? Understanding currency dominance, dependence, and dynamics. In *SPIE Third International Symposium on Fluctuations and Noise*, pages 86–99. 2005.

[200] N.L. Johnson, S. Kotz, and N. Balakrishnan. Continuous univariate distributions. *Wiley*, 1995.

[201] E. Jondeau, S.H. Poon, and M. Rockinger. *Financial modeling under non-Gaussian distributions*. Springer Verlag, 2007.

[202] B.H. Juang and L.R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[203] C. Jutten and J. Karhunen. Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. *International Journal of Neural Systems*, 14(5):267–292, 2004.

[204] R.E. Kass, L. Tierney, and J.B. Kadane. Laplace's method in Bayesian analysis. In *Statistical multiple integration: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference*, volume 115, page 89. American Mathematical Society, 1991.

[205] D.A. Kenny. *Correlation and causation*. Wiley, 1979.

[206] H.S. Kim, R. Eykholt, and J.D. Salas. Nonlinear dynamics, delay times, and embedding windows. *Physica D: Nonlinear Phenomena*, 127(1-2):48–60, 1999.

[207] S. Kim and E.N. Brown. A general statistical framework for assessing Granger causality. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2222–2225. IEEE, 2010.

[208] M. King and D. Rime. The $4 trillion question: What explains FX growth since the 2007 survey? *BIS Quarterly Review, December*, 2010.

[209] J. Knight and S. Satchell. *Return distributions in finance*. Butterworth-Heinemann, 2001.

[210] J. Knight and S. Satchell. *Forecasting volatility in the financial markets*. Butterworth-Heinemann, 2007.

[211] G. Koutmos, C. Negakis, and P. Theodossiou. Stochastic behaviour of the Athens stock exchange. *Applied Financial Economics*, 3(2):119–126, 1993.

[212] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):66138, 2004.

[213] K.F. Kroner, K.P. Kneafsey, and S. Claessens. Forecasting volatility in commodity markets. *Journal of Forecasting*, 14(2):77–95, 1995.

[214] M. Kuczma. *An introduction to the theory of functional equations and inequalities: Cauchy's equation and Jensen's inequality*. Birkhauser, 2008.

[215] D. Kuhn and D.G. Luenberger. Analysis of the rebalancing frequency in log-optimal portfolio selection. *Quantitative Finance*, 10(2):221–234, 2010.

[216] T. Lan and D. Erdogmus. Local linear ICA for mutual information estimation in feature selection. In *2005 IEEE Workshop on Machine Learning for Signal Processing*, pages 3–8, 2005.

[217] H.O. Lancaster. *Chi-squared distribution*. John Wiley & Sons New York, 1969.

[218] H. Lappalainen and J. Miskin. Ensemble learning. *Advances in Independent Component Analysis*, pages 75–92, 2000.

[219] M. Lavielle and G. Teyssière. Detection of multiple changepoints in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

[220] J. Ledolter. The effect of additive outliers on the forecasts from ARIMA models. *International Journal of Forecasting*, 5(2):231–240, 1989.

[221] T.W. Lee, M.S. Lewicki, M. Girolami, T.J. Sejnowski, et al. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6(4):87–90, 1999.

[222] M.J. Lenardon and A. Amirdjanova. Interaction between stock indices via changepoint analysis. *Applied Stochastic Models in Business and Industry*, 22(5):573–586, 2006.

[223] R. Levine, N. Loayza, and T. Beck. Financial intermediation and growth: Causality and causes. *Journal of Monetary Economics*, 46(1):31–78, 2000.

[224] W. Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60(5):823–837, 1990.

[225] S. Lieberson. Limitations in the application of non-parametric coefficients of correlation. *American Sociological Review*, 29(5):744–746, 1964.

[226] M. Loretan and W.B. English. Evaluating "correlation breakdowns" during periods of market volatility. *International Finance Discussion Papers*, 2000.

[227] D.R. Lowne, S. Roberts, and R. Garnett. Sequential non-stationary dynamic classification with sparse feedback. *Pattern Recognition*, 2009.

[228] B. Lu, K. Hirasawa, and J. Murata. A new learning method using prior information of neural networks. *Artificial Life and Robotics*, 4(2):78–83, 2000.

[229] C.J. Lu, T.S. Lee, and C.C. Chiu. Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, 47(2):115–125, 2009.

[230] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu. Methods for quantifying the causal structure of bivariate time series. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 17(3):903, 2007.

[231] M. Lungarella, A. Pitti, and Y. Kuniyoshi. Information transfer at multiple scales. *Physical Review E*, 76(5):56117, 2007.

[232] H. Lütkepohl. Comparison of criteria for estimating the order of a vector autoregressive process. *Journal of Time Series Analysis*, 6(1):35–52, 1985.

[233] R.K. Lyons and M.J. Moore. An information approach to international currencies. *Journal of International Economics*, 2009.

[234] R.D. Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[235] S. Mahfoud and G. Mani. Financial forecasting using genetic algorithms. *Applied Artificial Intelligence*, 10(6):543–566, 1996.

[236] Y. Malevergne and D. Sornette. Testing the Gaussian copula hypothesis for financial assets dependences. *Quantitative Finance*, 3(4):231–250, 2003.

[237] R.N. Mantegna and H.E. Stanley. *Introduction to econophysics: Correlations and complexity in finance*. Cambridge University Press, 1999.

[238] D. Marinazzo, M. Pellicoro, and S. Stramaglia. Kernel method for nonlinear Granger causality. *Physical Review Letters*, 100(14):144103, 2008.

[239] N.C. Mark. Exchange rates and fundamentals: Evidence on long-horizon predictability. *The American Economic Review*, 85(1):201–218, 1995.

[240] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[241] B. Marshall, S. Treepongkaruna, and M. Young. Exploitable arbitrage opportunities exist in the foreign exchange market. In *American Finance Association Annual Meeting, New Orleans*, 2008.

[242] J. McAndrews. Segmentation in the US dollar money markets during the financial crisis. In *International Conference on Financial System and Monetary Policy Implementation*, 2009.

[243] M. McDonald, O. Suleman, S. Williams, S. Howison, and N.F. Johnson. Detecting a currency's dominance or dependence using foreign exchange network trees. *Physical Review E*, 72(4):46106, 2005.

[244] S. McLaughlin, A. Stogioglou, and J. Fackrell. Introducing higher order statistics (HOS) for the detection of nonlinearities. *UK Nonlinear News*, 15, 1995.

[245] R.A. Meese and A.K. Rose. An empirical assessment of non-linearities in models of exchange rate determination. *The Review of Economic Studies*, pages 603–619, 1991.

[246] N.B. Mehr. *Portfolio allocation using wavelet transform*. ProQuest, 2008.

[247] R.C. Merton. An analytic derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis*, 7(4):1851–1872, 1972.

[248] S. Miccichč, G. Bonanno, and F. Lillo. Degree stability of a minimum spanning tree of price return and volatility. *Physica A: Statistical Mechanics and its Applications*, 324(1-2):66–73, 2003.

[249] T. Mikosch and C. Starica. Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics*, 86(1):378–390, 2004.

[250] R.E. Miller. *Optimization: Foundations and applications*. Wiley-Interscience, 2000.

[251] T. Mizuno, H. Takayasu, and M. Takayasu. Correlation networks among currencies. *Physica A: Statistical Mechanics and its Applications*, 364:336–342, 2006.

[252] F. Modigliani and L. Modigliani. Risk-adjusted performance. *The Journal of Portfolio Management*, 23(2):45–54, 1997.

[253] A. Morimoto, S. Ozawa, and R. Ashino. An efficient identification method of the structural parameters of MDOF structures using the wavelet transform and neural networks. In *Proceedings of the Second World Conference on Structural Control*, pages 2133–2140, 1999.

[254] M. Mudelsee. *Climate time series analysis: Classical statistical and bootstrap methods*, volume 42. Springer, 2010.

[255] R.J. Muirhead. Aspects of multivariate statistical theory. *John Wiley & Sons, Inc.*, 1982.

[256] U.A. Muller, M.M. Dacorogna, R.D. Dave, O.V. Pictet, R.B. Olsen, and J.R. Ward. Fractals and intrinsic time: A challenge to econometricians. *Olsen and Associates Publisher, Geneva*, 1993.

[257] U.A. Müller, M.M. Dacorogna, R.B. Olsen, O.V. Pictet, M. Schwarz, and C. Morgenegg. Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis. *Journal of Banking & Finance*, 14(6):1189–1208, 1990.

[258] U.A. Muller, M.M. Dacorogna, and O.V. Pictet. Heavy tails in high-frequency financial data. *A practical guide to heavy tails: Statistical techniques and applications*, pages 55–78, 1998.

[259] J.A. Murphy. An analysis of the financial crisis of 2008: Causes and solutions. *Social Science Research Network, Available at SSRN 1295344*, 2008.

[260] K.P. Murphy. *Dynamic Bayesian networks: Representation, inference and learning*. PhD thesis, University of California, 2002.

[261] F. Murtagh, J.L. Starck, and O. Renaud. On neuro-wavelet modeling. *Decision Support Systems*, 37(4):475–484, 2004.

[262] I.J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100, 2003.

[263] Y. Nagahara. The PDF and CF of Pearson type IV distributions and the ML estimation of the parameters. *Statistics & probability letters*, 43(3):251–264, 1999.

[264] K. Nagarajan, B. Holland, C. Slatton, and A.D. George. Scalable and portable architecture for probability density function estimation on FPGAs. In *Proceedings of the 2008 16th International Symposium on Field-Programmable Custom Computing Machines*. IEEE Computer Society, 2008.

[265] M. Nakken. Wavelet analysis of rainfall–runoff variability isolating climatic from anthropogenic patterns. *Environmental Modelling & Software*, 14(4):283–295, 1999.

[266] M. Nandha and R. Brooks. Oil prices and transport sector returns: An international analysis. *Review of Quantitative Finance and Accounting*, 33(4):393–409, 2009.

[267] G.P. Nason and R.V. Sachs. Wavelets in time series analysis. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 357(1760):2511–2526, 1999.

[268] M.J. Naylor, L.C. Rose, and B.J. Moyle. Topology of foreign exchange markets using hierarchical structure methods. *Physica A: Statistical Mechanics and its Applications*, 382(1):199–208, 2007.

[269] R.B. Nelsen. *An introduction to copulas*. Springer Verlag, 2006.

[270] M. Novey and T. Adali. Complex ICA by negentropy maximization. *Neural Networks, IEEE Transactions on*, 19(4):596–609, 2008.

[271] E. Oja, K. Kiviluoto, and S. Malaroiu. Independent component analysis for financial time series. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, pages 111–116. IEEE, 2000.

[272] M. Okutomi and T. Kanade. A locally adaptive window for signal matching. *International Journal of Computer Vision*, 7(2):143–162, 1992.

[273] I. Osorio, M.A.F. Harrison, Y.C. Lai, and M.G. Frei. Observations on the application of the correlation dimension and correlation integral to the prediction of seizures. *Journal of Clinical Neurophysiology*, 18(3):269, 2001.

[274] P. Oswiecimka, J. Kwapien, S. Drozdz, and R. Rak. Investigating multifractality of stock market fluctuations using wavelet and detrending fluctuation methods. *Acta Physica Polonica B*, 36(8):2447, 2005.

[275] A. Ozun and A. Cifter. Multi-scale causality between energy consumption and GNP in emerging markets: Evidence from Turkey. *Investment Management and Financial Innovations*, 4(2):61–70, 2007.

[276] J. Palmer, K. Kreutz-Delgado, and S. Makeig. Super-Gaussian mixture source model for ICA. *Independent Component Analysis and Blind Signal Separation*, pages 854–861, 2006.

[277] M. Paluš and M. Vejmelka. Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections. *Physical Review E*, 75(5):056211, 2007.

[278] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, pages 1065–1076, 1962.

[279] G.K. Pasricha. Kalman filter and its economic applications. *Munich Personal RePEc Archive*, MPRA Paper No. 22734, 2006.

[280] J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2000.

[281] J. Pearl. Causal inference in statistics: An overview. *Statistical Surveys*, 2009.

[282] M.C. Peel, G.E. Amirthanathan, G.G.S. Pegram, T.A. McMahon, and F.H.S. Chiew. Issues with the application of empirical mode decomposition analysis. *International Congress on Modelling and Simulation*, pages 1681–1687, 2005.

[283] A. Péguin-Feissolle, B. Strikholm, and T. Teräsvirta. Testing the Granger noncausality hypothesis in stationary nonlinear models of unknown functional form. *Communications in Statistics-Simulation and Computation*, 42(5):1063–1087, 2013.

[284] B. Peiers. Informed traders, intervention, and price leadership: A deeper view of the microstructure of the foreign exchange market. *Journal of Finance*, 52(4):1589–1614, 1997.

[285] A. Peiro. Skewness in financial returns. *Journal of Banking & Finance*, 23(6):847–862, 1999.

[286] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[287] Z.K. Peng, P.W. Tse, and F.L. Chu. A comparison study of improved Hilbert–Huang transform and wavelet transform: Application to fault diagnosis for rolling bearing. *Mechanical Systems and Signal Processing*, 19(5):974–988, 2005.

[288] W. Penny, R. Everson, and S. Roberts. Hidden Markov independent components analysis. *Advances in Independent Component Analysis. Springer*, pages 3–22, 2000.

[289] W. Penny, S. Kiebel, and K. Friston. Variational Bayesian inference for fMRI time series. *NeuroImage*, 19(3):727–741, 2003.

[290] W. Penny, S. Kiebel, and K. Friston. Variational Bayes. In *Statistical parametric mapping: The analysis of functional brain images*, pages 303–312. Elsevier, 2007.

[291] W. Penny and S. Roberts. Bayesian multivariate autoregressive models with structured priors. In *Vision, Image and Signal Processing, IEE Proceedings-*, volume 149, pages 33–41. 2002.

[292] M.E. Pflieger and R.E. Greenblatt. Using conditional mutual information to approximate causality for multivariate physiological time series. *International Journal of Bioelectromagnetism*, 7:285–288, 2005.

[293] M. Pojarliev and R.M. Levich. Detecting crowded trades in currency funds. *Financial Analysts Journal*, 67(1), 2011.

[294] A. Pole. *Statistical arbitrage: Algorithmic trading insights and techniques*. Wiley, 2007.

[295] S.H. Poon and C.W. Granger. Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539, 2003.

[296] A.S. Posen. Why the euro will not rival the dollar. *International Finance*, 11(1):75–100, 2008.

[297] D. Prichard and J. Theiler. Generalized redundancies for time series analysis. *Physica D: Nonlinear Phenomena*, 84(3-4):476–493, 1995.

[298] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[299] S.T. Rachev. *Bayesian methods in finance*. Wiley, 2008.

[300] M. Rajković. Extracting meaningful information from financial data. *Physica A: Statistical Mechanics and its Applications*, 287(3):383–395, 2000.

[301] B. Raunig. The predictability of exchange rate volatility. *Economics Letters*, 98(2):220–228, 2008.

[302] I.A. Rezek and S. Roberts. Causal analysis with information flow. *Manuscript, University of London*, 1998.

[303] S. Roberts, R. Cain, and M.S. Dawkins. Prediction of welfare outcomes for broiler chickens using Bayesian regression on continuous optical flow data. *Journal of The Royal Society Interface*, 9(77):3436–3443, 2012.

[304] S. Roberts and R. Everson. *Independent component analysis: Principles and practice*. Cambridge University Press, 2001.

[305] S. Roberts, D. Husmeier, I.A. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

[306] S. Roberts and W. Penny. Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing*, 50(9):2245–2257, 2002.

[307] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

[308] A. Rossi and G.M. Gallo. Volatility estimation via hidden Markov models. *Journal of Empirical Finance*, 13(2):203–230, 2006.

[309] T. Ryden. Estimating the order of hidden Markov models. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):345–354, 1995.

[310] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. Akaike information criterion statistics. *Journal of the American Statistical Association*, 1986.

[311] L. Sandoval and I.D.P. Franca. Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 391(1):187–208, 2012.

[312] B. Schelter, M. Winterhalder, M. Eichler, M. Peifer, B. Hellwig, B. Guschlbauer, C.H. Lücking, R. Dahlhaus, and J. Timmer. Testing for directed influences among neural signals using partial directed coherence. *Journal of neuroscience methods*, 152(1):210–219, 2006.

[313] A. Schmidt. Describing impact of trading in the global FX market. *ICAP, Available at SSRN 1978977*, 2012.

[314] A. Schmitz. Measuring statistical dependence and coupling of subsystems. *Physical Review E*, 62(5):7508–7511, 2000.

[315] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

[316] S. Sello. Time series forecasting: A nonlinear dynamics approach. *Los Alamos National Laboratories, Archive Physics 9906035*, 1999.

[317] N. Shah and S. Roberts. Hidden Markov independent component analysis as a measure of coupling in multivariate financial time series. In *2008 ICA Research Network International Workshop*, 2008.

[318] N. Shah and S. Roberts. Dynamically measuring statistical dependencies in multivariate financial time series using independent component analysis. *ISRN Signal Processing Journal*, 2013.

[319] C.E. Shannon and W. Weaver. A mathematical theory of communications. *Bell System Technical Journal*, 27(2):632–656, 1948.

[320] W.F. Sharpe. The Sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58, 1994.

[321] M.J. Shensa. The discrete wavelet transform: Wedding the a trous and Mallat algorithms. *Signal Processing, IEEE Transactions on*, 40(10):2464–2482, 1992.

[322] A. Shephard. *Structural models of the labour market and the impact and design of tax policies*. PhD thesis, Department of Economics, University College London, 2010.

[323] S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

[324] S. Shimizu, A. Hyvärinen, Y. Kawahara, and T. Washio. A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 506–513. AUAI Press, 2009.

[325] J. Shlens. A tutorial on principal component analysis. *Technical Report, Systems Neurobiology Laboratory, University of California at San Diego*, 2005.

[326] A.F.M. Smith. Bayesian computational methods. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337:369–386, 1991.

[327] S. Stavroyiannis, I. Makris, V. Nikolaidis, and L. Zarangas. Econometric modeling and value-at-risk using the Pearson type-IV distribution. *International Review of Financial Analysis*, 22:10–17, 2012.

[328] P. Stoica and R.L. Moses. *Introduction to spectral analysis*. Prentice Hall, New Jersey, 1997.

[329] H.R. Stoll. Electronic trading in stock markets. *The Journal of Economic Perspectives*, 20(1):153–174, 2006.

[330] J.V. Stone. *Independent component analysis: A tutorial introduction*. MIT Press, 2004.

[331] Z.R. Struzik. Wavelet methods in (financial) time-series processing. *Physica A: Statistical Mechanics and its Applications*, 296(1-2):307–319, 2001.

[332] A. Stuart. *Kendall's advanced theory of statistics*, volume 2. Charles Griffin, 1987.

[333] D.Y. Takahashi, L.A. Baccal, and K. Sameshima. Connectivity inference between neural structures via partial directed coherence. *Journal of Applied Statistics*, 34(10):1259–1273, 2007.

[334] Y.Y. Tang, V. Wickerhauser, P.C. Yuen, and C. Li. *Wavelet analysis and its applications*, volume 2. Springer, 2001.

[335] J.W. Taylor. Volatility forecasting with smooth transition exponential smoothing. *International Journal of Forecasting*, 20(2):273–286, 2004.

[336] S.J. Taylor. *Modelling financial time series*. World Scientific Pub Co Inc, 2007.

[337] L. Te-Won. *Independent component analysis: Theory and applications*. Kluwer Academic Publishers, 1998.

[338] T. Thadewald and H. Buning. Jarque–Bera test and its competitors for testing normality: A power comparison. *Journal of Applied Statistics*, 34(1):87–105, 2007.

[339] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J.D. Farmer. Testing for nonlinearity in time series: The method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1):77–94, 1992.

[340] F.J. Theis, E.W. Lang, and C.G. Puntonet. A geometric algorithm for overcomplete linear ICA. *Neurocomputing*, 56:381–398, 2004.

[341] N.S. Thomaidis, N. Kondakis, and G.D. Dounias. An intelligent statistical arbitrage trading system. pages 596–599, 2006.

[342] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[343] D. Tjøstheim. Granger-causality in multiple time series. *Journal of Econometrics*, 17(2):157–176, 1981.

[344] H.Y. Toda and P.C.B. Phillips. Vector autoregression and causality: A theoretical overview and simulation study. *Econometric reviews*, 13(2):259–285, 1994.

[345] K. Torkkola. On feature extraction by mutual information maximization. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1. IEEE, 2002.

[346] A. Venelli. Efficient entropy estimation for mutual information analysis using B-splines. *Information Security Theory and Practices*, pages 17–30, 2010.

[347] P.F. Verdes. Assessing causality from multivariate time series. *Physical Review E*, 72(2):26222, 2005.

[348] P. Verhoeven and M. McAleer. Fat tails and asymmetry in financial volatility models. *Mathematics and Computers in Simulation*, 64(3):351–361, 2004.

[349] S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91(2):557–575, 2000.

[350] D.I. Vortelinos. Portfolio analysis of intraday covariance matrix in the Greek equity market. *Research in International Business and Finance*, 2012.

[351] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[352] E.W. Weisstein. Minimum spanning tree. *MathWorld–A Wolfram Web Resource*, 2012.

[353] A. Wilhelmsson. Garch forecasting performance under different distribution assumptions. *Journal of Forecasting*, 25(8):561–578, 2006.

[354] R. Willink. A closed-form expression for the Pearson type IV distribution function. *Australian & New Zealand Journal of Statistics*, 50(2):199–205, 2008.

[355] S.N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003.

[356] F. Yip and L. Xu. An application of independent component analysis in the arbitrage pricing theory. *Proceedings of the International Joint Conference on Neural Networks*, 5:279–284, 2000.

[357] D. Yu, M. Small, R.G. Harrison, and C. Diks. Efficient implementation of the Gaussian kernel algorithm in estimating invariants and noise level from noisy time series data. *Physical Review E*, 61(4):3750–3756, 2000.

[358] J. Yu, V.A. Smith, P.P. Wang, A.J. Hartemink, and E.D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 2004.

[359] C. Zapart. Statistical arbitrage trading with wavelets and artificial neural networks. In *Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering*, pages 429–435. 2003.

[360] C. Zapart. Long-short equity pairs trading with optimum wavelet correlation measures. In *Financial Engineering and Applications*. ACTA Press, 2004.

[361] F. Zhang. High-frequency trading, stock volatility, and price discovery. *Available at SSRN 1691679*, 2010.