Multivariate density forecast evaluation and nonparametric Granger causality testing

Fang, H.

Creative Commons License (see https://creativecommons.org/use-remix/cc-licenses):
Other

Citation for published version (APA):
This dissertation studies the evaluation of density forecasts and conditional dependence structure for multivariate time series. Firstly, univariate and multivariate density forecasts are compared in the context of portfolio risk management. Secondly, simulation-based methodologies are applied to quantify the conditional dependence structure between two time series. Thirdly, this dissertation develops a robust statistical test for conditional independence, i.e. a non-parametric test for Granger non-causality by measuring the transfer entropy.

Hao Fang holds a B.Sc. in Economics and an M.Sc. in Finance from the Dongbei University of Finance and Economics in China. In 2014, he obtained his M.Phil. in Economics from Tinbergen Institute and then became a Ph.D. candidate in the Department of Quantitative Economics at the University of Amsterdam. His research focuses on various aspects of density estimation and evaluation, with applications in risk management and empirical finance.
Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing
ISBN 978 90 3610 487 6

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 711 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.
Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
Prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie, in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 27 maart 2018, te 10:00 uur
door

Hao Fang

geboren te Anhui, China
Promotiecommissie

**Promotor:** Prof. dr. C.G.H. Diks  
*Universiteit van Amsterdam*

**Copromotor:** Prof. dr. D.J.C. van Dijk  
*Erasmus Universiteit Rotterdam*

**Overige leden:**  
Prof. dr. V. Panchenko  
University of New South Wales  
Dr. C. Zhou  
Erasmus Universiteit Rotterdam  
Prof. dr. H.P. Boswijk  
De Nederlandsche Bank  
Dr. S.A. Broda  
Universiteit van Amsterdam  
Prof. dr. M.H. Vellekoop  
Universiteit van Amsterdam

Faculteit Economie en Bedrijfskunde
Acknowledgments

It has been a long and fun journey since I joined the Tinbergen MPhil program more than five years ago. Along this journey, I have encountered so many amazing people and received generous inspiration and encouragement from them. Without their support, this dissertation would never have been finished. I feel indebted towards them and wish to express my deep gratitude in these pages.

The most important person that I would like to thank is my supervisor prof. dr. Cees Diks. I firstly met Cees at one of his fascinating presentations, which I liked so much that I decided to write my PhD thesis with him. During the last four years, Cees has shown me not only his incredible skills in statistics and econometrics, but also tremendous patience in teaching me how to conduct academic research and correcting typos in my writings. I also greatly enjoyed our motivating discussions and the insightful feedback I receive from him, as well as the confidence he has put in me. Without his careful guidance, none of the chapters of my thesis would have come into existence.

My gratitude also goes to prof. dr. Dick van Dijk, who taught me Advanced Time Series Econometrics when I was a second year TI student. As my co-promoter, he provided me many valuable suggestions on how to polish my papers. Further, I would like to thank all committee members, prof. dr. Valentyn Panchenko, dr. Chen Zhou, prof. dr. Peter Boswijk, dr. Simon Broda and prof. dr. Michel Vellekoop for their time and advice while reviewing my dissertation.

While completing this dissertation, I was so lucky to work with several CeNDEFFers. I would like to thank Cars Hommes, Jan Tuinstra, Florian Wagener, Marco van der Leij, Dávid Kopányi, Anita Kopányi-Peuker, Isabelle Salle, Moutaz Altaghlibi, Tomasz Makarewicz, Marcin Wolski, Florian Sniekers, Juanxi Wang, Roy van der Weide, Roald Ramer, Maurice Koster, Hao Li and Johan de Jong. Special thanks go to Joep Lusstenhouwer, Gavin Goy and Myrna Hennequin, who have been sharing the office with me for two years. Without the flexible and cheerful environment in our office, writing an academic paper would have been much more boring.
Further, I would like to thank my colleagues from Tinbergen Institute for all the
good times we shared. I appreciate the company of my classmates: Yiming Zhang,
Wenjie Chen, Hang Liu, Yuan Gu, Yajie Sun, Wenqian Huang, Eglė Jakučionytė,
Nicole Ciurila, Iulian Ciobica, Sabina Albrecht, Oana Furtuna, Robin Döttling, Andro
Rilovič, Andries van Vlodorp, and Mehmet Kutluay; the Mphil students before and
after me at TI: Andrei Lalu, Shawny Xiao, Ieva Sakalauskaite, Stephanie Chen, Yue
Li, Zhaokun Zhang, Simin He, Zhiling Wang, Jindi Zheng, Swapnil Singh, Merrick Li,
Mengheng Li, Yuhao Zhu, Junze Sun, Lingwei Kong, Gabriele Ciminelli, and Rui Zhuo
among others. I also thank our secretariat for their excellent support in dealing with
administrative issues.

Whether inside academia or not, I met many friends thanks to whom my life and
studies in the Netherlands were so pleasant. First, there are Xiye Yang, Yuan Yue,
Zhenzhen Fan, Lin Zhao and Zhaochun Ren, who are not just doctors, but cool doctors.
Then I want to thank Yang Liu, Xuehan Zhang, Yue Liu for all the happiness that
we shared and all beers we consumed. Some distinct and equally grateful appreciation
goes to Shuo Xia and Bo Hu for all our meaningless discussions about life on our hiking
trips. May my wishes extend to Mieke and Yannan. Finally, the basketball group was
an important part of my PhD life. It was a great fun to play basketball and hang
out with Artūras Juodis, Simas Kučinskas, Vytis Kapočius, Will Wang, Pascal Golec,
Sándor Sóvágó, Kostats Mavromatis, Tadas Dob, and José Vives.

I would also like to thank Thomas Hufener, Robin van Boxsel, Matthew New-
bon, Jerry Tworek, Michal Garmulewicz, Erkki Silde, and Kostiantyn Zagranovskiyi for
opening the door to a whole new world of possibilities. My gratitude goes further to
China, where I have my best friends: in the chronological order, Hongzhe Shen, Jia
Sun, Zheming Zhang, Chang liu, Jingwen Wu, Zuowei Guo, Yong Yang, Jie Zhang,
Zongfei Yan, Zongwu Liu and Hong Qian. I thank them for the friendship that did not
fade with time.

I owe a big thank-you to my family. Without the support from my parents and
sisters, I wouldn’t have had the courage to embark on this journey in the first place.
I should not forget to thank my grandparents for showing me the wisdom about life.
Finally, the last words of the acknowledgment I have saved for my dear wife Yajing
Li, who has provided so much happiness and love during those years and made those
years the best in my life.

Amsterdam, December 2017
## Contents

**Acknowledgments**

1 **Introduction**

2 **Comparing Density Forecasts in a Risk Management Context**
   2.1 Introduction ........................................... 5
   2.2 Methodology ............................................. 8
      2.2.1 Density Comparison within Elliptical Families .... 8
      2.2.2 Testing Approach .................................... 9
      2.2.3 Weighted Logarithmic Scoring Rules ................. 11
      2.2.4 Skew Elliptical Distributions ....................... 14
   2.3 Monte Carlo Simulations ................................ 17
   2.4 Empirical Application .................................... 24
      2.4.1 Comparison of density forecasts across dimensions ... 24
      2.4.2 Dynamic selection of forecasted distributions ........ 30
      2.4.3 Value-at-Risk (VaR) Forecasts and Evaluation ........ 36
   2.5 Summary and Conclusions ................................ 44
   2.6 Appendix I ............................................. 46
      2.6.1 Density Function for Multivariate t-distribution .... 46
      2.6.2 Proof of the Properness of PWL ..................... 48
      2.6.3 Standardization of Skew Elliptical Distributions .... 49

3 **Transfer Entropy for Nonparametric Granger Causality Detection: An Evaluation of Different Resampling Methods**
   3.1 Introduction ............................................. 51
   3.2 Methodology ............................................. 53
      3.2.1 Transfer Entropy and its Estimator ................. 53
      3.2.2 Density Estimation and Bandwidth Selection .......... 55
      3.2.3 Resampling Methods ................................ 58
4 A Consistent Test for Granger Causality based on a Nonparametric information-based Statistic

4.1 Introduction ......................................................... 83

4.2 A Transfer Entropy-Based Test Statistic .......................... 85
  4.2.1 Nonparametric Granger non-Causality Tests ................. 85
  4.2.2 The DP Test .................................................. 86
  4.2.3 Information-theoretical Interpretation .......................... 91
  4.2.4 A Modified DP Test ........................................ 94
  4.2.5 Bandwidth Selection .......................................... 96

4.3 Monte Carlo Simulations ......................................... 97

4.4 Empirical Research .............................................. 103
  4.4.1 Stock Volume–Return Relation .............................. 103
  4.4.2 Application to Intraday Exchange Rates .................... 105

4.5 Summary and Conclusions ........................................ 112

4.6 Appendix II ...................................................... 113
  4.6.1 Proof of Positive-definiteness of the Transfer Entropy ...... 113
  4.6.2 Proof of Positive-definiteness of the First Order TE Statistic ... 114
  4.6.3 Proof of Non-Degeneracy of the Modified DP Test Statistic .... 114
  4.6.4 Proof of Lemma 2 ............................................ 119
  4.6.5 Proof of the Asymptotic Distribution of the Modified DP Statistic120
  4.6.6 Optimal Bandwidth for the Modified DP Test ................ 121

Bibliography .......................................................... 123

Summary .............................................................. 133

Samenvatting (Dutch Summary) ..................................... 135
Chapter 1

Introduction

Modern time series econometrics has been greatly reshaped in recent decades. The earlier emphasis on normally distributed processes and linear specifications for models had declined, and econometric modeling had moved its consideration from conditional expectations to conditional variance, and even further into other moments. However, as Granger (2003) puts it, “it is most natural to consider the whole conditional distribution [of a multivariate time series], particularly in many dimensions”. On the one hand, the interest in distribution estimation and forecasting is rapidly expanding in macroeconomics and finance, which requires statistical evaluation of the predictive ability of different density forecasts. On the other hand, the topic of measuring conditional dependence has been considered already for a long time in the econometrics field. The literature has witnessed a gradual shift of focus from linear, parametric methodologies to nonlinear, nonparametric techniques. My dissertation focuses on these two different, yet related, aspects of the estimation of distributions, i.e. the evaluation of density forecasts and the detection of conditional dependence.

The second chapter of this dissertation, based on Diks and Fang (2016), investigates predictive density comparison of a financial portfolio returns in the downside part of the support. The accuracy of the estimation of return distributions is crucial for all portfolio decisions, and especially for risk management, where the left-tail risk is very sensitive to the assumed family of distributions. A question that arises naturally is: what is the most appropriate return distribution to assume for evaluating the portfolio risk? Besides, portfolio risk management, typically involves estimating and evaluating a high-dimensional distribution, confronts with a dilemma: should the estimated density of the portfolio be based on multivariate modeling of the joint asset returns distribution, or the univariate modeled portfolio return distribution?

The second question may sound trivial, but it is highly important for portfolio risk
management. Fig. 1.1 presents an example. Suppose that the true density follows a standard bivariate Normal distribution, and that there are two competing density forecasts, \( \hat{f} \) and \( \hat{g} \). Contour plots of the true density and two density forecasts are shown in Fig. 1.1(a). The dotted circles correspond to the true bivariate normal distribution, while the solid and dashed lines are the contours of \( \hat{f} \) and \( \hat{g} \), respectively. Clearly, forecast \( \hat{g} \) is closer to the true distribution, and a reasonable test for the predictive ability of the bivariate densities should be in favor of \( \hat{g} \) over \( \hat{f} \).

However, if we construct a linear combination defined as \( Z = b_1 Y_1 + b_2 Y_2 \), the relative predictive ability of the forecasts may change. Consider the simple case where \( b_1 = b_2 = 1 \). Fig. 1.1(b) gives the univariate densities of the true distribution (dotted line), forecast \( \hat{f} \) (solid line) and \( \hat{g} \) (dashed line). In contrast with the previous conclusion, the density of forecast \( \hat{f} \) is clearly closer to the true distribution now. Therefore, even though density forecast \( \hat{g} \) is closer to the true distribution in the bivariate space, from the point of view of such a ‘portfolio return’ which consists of \( Y_1 \) and \( Y_2 \), it is better to base the portfolio risk assessment on the distribution of \( \hat{f} \).

By adopting the likelihood-based score test, Chapter 2 compares the multivariate and univariate approaches directly and sheds some light on the pitfalls of multivariate density evaluation in a portfolio context. We emphasize that it makes a difference whether you evaluate univariate or multivariate scores, and that the optimal forecast is not constant over time. The common practice to perform multivariate forecast comparisons can be problematic, since better multivariate forecasts do not necessarily correspond to better aggregate portfolio return forecasts. This is illustrated by
investigating analytically tractable (under projections) involving (skew) elliptical distributions as defined by Azzalini and Capitanio (2003); Azzalini and Dalla Valle (1996) and an application to daily returns of a number of US stock prices initially proposed by Giot and Laurent (2003). Additionally, an adaptive selection of forecasts is proposed and shown to improve the forecasting ability of portfolio risk measurement.

In addition to the evaluation of the forecast density accuracy, another challenging task of distribution modeling consists of estimating the conditional dependence structure of several variables of interest. Chapters 3 and 4 are devoted to the estimation of causal effect among variables in the sense of Granger. Granger (1969) mathematically defines a statistical concept of causality that is based on prediction; if a variable \( Y \) Granger causes another variable \( X \), it means that the past value of \( Y \) contains extra information that helps to predict the future value of \( X \) beyond the information contained in the history of \( X \) alone. In a general manner, the causal effects are defined as measures for the conditional impact from one variable on another. After classical linear, parametric modeling methods were criticized for overlooking nonlinear causal relations, plenty of simulation-based techniques and nonparametric approaches emerged.

To illustrate how linear tests can fail to spot nonlinear causal effect, Hiemstra and Jones (1994) considered the process given by

\[
X_t = \beta Y_{t-L} \cdot X_{t-M} + \varepsilon_t,
\]

where \( \{Y_t\} \) and \( \{\varepsilon_t\} \) are independent and identically distributed (i.i.d.) \( N(0,1) \) time series, \( \beta \) is the parameter of interest, and \( L \) and \( M \) represent the Lags of \( X \) and \( Y \), respectively. If one regresses \( \{X_t\} \) on the lagged values of \( \{Y_t\} \) naively and implement a traditional linear causality test, the result will be an incorrect rejection that there is no lagged impact from \( \{Y_t\} \) to \( \{X_t\} \). Linear approaches can have very low power against this kind of nonlinear causal relations, as the parametric estimation which the test is based on suffers from model mis-specification.

The concept of an information-theoretical measure — transfer entropy — a term coined by Schreiber (2000), which had appeared in the literature even earlier under different names, is a suitable basis for nonparametrically measuring directional information flows, i.e. the conditional dependence structure. However, this measure does not carry over directly to measures of conditional dependence. Various studies have shown the difficulties of applying entropy-based test in an econometric framework. Chapter 3, adapted from Diks and Fang (2017a), provides the practitioners some guidelines on how to apply some resampling techniques in implementing a nonparametric entropy-based test on causal effects. Namely, the time-shifted surrogates algorithms, the smoothed
bootstrap and the stationary bootstrap procedures are presented and compared. The spillover effect in the global equity market studied in Diebold and Yilmaz (2009) is revisited by using the proposed resampling-techniques to illustrate how to detect the pairwise Granger causality nonparametrically.

Chapter 4 takes one step further in the theoretical direction of nonparametrically testing for Granger causality, mainly based on Diks and Fang (2017b). This chapter develops a novel nonparametric test based on a first order Taylor expansion of the transfer entropy. To the best of our knowledge, the proposed test is the first transfer entropy-based test that maintains the asymptotic normality under the null hypothesis of non-Granger causality. The new test does not suffer from the lack of power problem of the frequently-used test developed by Diks and Panchenko (2006), which is not consistent against all fixed alternatives. Simulation results confirm the theoretical size and power properties of the new test. The proposed test is applied to two data sets to detect causal linkages between daily stock return and volume, and among major currencies on an intra-day frequency. In both applications, the proposed nonparametric test captures some strong causal effects that tend to be overlooked by parametric models.

The chapters of this dissertation can be read independently, providing an individual introduction, simulation study, empirical research and conclusion. To avoid redundancy, a combined bibliography is provided at the end of the dissertation.
Chapter 2

Comparing Density Forecasts in a Risk Management Context

2.1 Introduction

Portfolio managers are nowadays weaponed with various tools to estimate complex multivariate distributions of asset returns, which may largely be attributed to the ever-increasing computational power. For any given portfolio, a natural question that both researchers and practitioners face is: among a number of competing density forecasts of the portfolio returns, which one most accurately describes the distribution of losses (the left tail of the return distribution)? This is the first question this Chapter addresses.

Since Markowitz (1952) laid the foundation of modern portfolio theory, theoretical research in portfolio theory has mainly centered around the first two predictive moments of portfolio returns. Traditionally, multivariate normality has been taken for granted in portfolio theory research as such assumptions provide great convenience for risk measurement. After all, the dependence between returns on different assets is fully described by their variance-covariance structure under normality. However, over the past two decades it has become clear that financial multivariate time series are typically non-normally distributed, exhibiting fat tails, skewness and nonlinear co-movements or asymmetric dependence structure; for instance, Patton (2004) and Boubaker and Sghaier (2013) studied portfolio optimization with dependence measured by competing copula densities. Similarly, Giot and Laurent (2003) showed that a model with skew $t$ innovations outperforms those with symmetric distributions in VaR estimation; Huang et al. (2015) suggested that better portfolio performance is achieved with a time-varying copula, particularly the Clayton copula. Diks et al. (2010) and Diks et al. (2014) found that for daily exchange rate returns, the $t$ copula is favored over its counterparts, while
in the government bond market, a mixture of $t$ and Clayton copulas performs best.

In this chapter, we focus on assessing the accuracy of different distributional assumptions, with emphasis on the left-tail risk associated with a given portfolio. Tests for the relative accuracy of two or more competing density forecasts can be based on loss functions measuring the distance between predicted density and the (unknown) true density. Possible loss functions include the integrated squared difference discussed by Sarno and Valente (2004) and the mean squared error (MSE) considered by Corradi and Swanson (2005) and Giacomini and White (2006). Amisano and Giacomini (2007) proposed a so-called weighted logarithmic scoring rule, with a weight function allowing predictive accuracy comparison in a particular region of the distribution, which is a desirable property for testing in tails. A score-based approach is mathematically convenient as it leads to simple pseudo-likelihood ratio type $t$-tests.

An important restriction we wish to impose on weighted scoring rules is properness (Matheson and Winkler, 1976). Intuitively, a scoring rule is proper if it never assigns a higher expected score to any density forecast than to the true conditional density. Gneiting and Raftery (2007) suggested that apart from the logarithmic score, the continuous ranked probability score ($crps$) is proper. Gneiting and Ranjan (2011) pointed out that the weighted logarithmic scoring rule by Amisano and Giacomini (2007) is improper and introduced a proper weighted $crps$ rule to compare density forecasts. Diks et al. (2011) also noted the problem of weighted log likelihood tests, and proposed two proper related scoring rules: the conditional likelihood ($cl$) and the censored likelihood ($csl$) scoring rule. Like the log-likelihood scoring rule, these are based on the Kullback-Leibler Information Criterion (KLIC), a well-known and widely-used measure of divergence between two probability distributions; see, e.g., Vuong (1989) and Giacomini and White (2006). Motivated by desirable properties of weighted scoring rules, Pelenis (2014) also proposed two alternative scoring rules: the penalized weighted likelihood ($pwl$) scoring rule and the incremental weighted $crps$ rule. In this chapter, we restrict ourselves to the KLIC-based scoring rules, $cl$, $csl$ and $pwl$, which can be readily applied to multivariate density forecasts.

When the dimension (number of assets) increases, the tests for predictive accuracy of density forecasts based on alternative multivariate models will become increasingly demanding computationally. Hence it then becomes attractive to calculate scores based on univariate density forecasts for the portfolio return. This leads to the second question this chapter investigates: given a portfolio, should density forecast comparisons be based on comparing ($i$) multivariate density forecasts of the joint asset returns distribution, ($ii$) the corresponding univariate density forecasts obtained after projecting the multivariate forecasts onto the portfolio return, or ($iii$) direct univariate density
2.1. Introduction

forecasts of the portfolio return?

Common practice is to follow approach (i), that is, to construct a (usually model-based) multivariate density forecast first, and then use the implied univariate portfolio density forecast to assess portfolio risk. The problem of this routine is that the forecast that is better according to the multivariate evaluation needs not correspond to a better portfolio return forecast. Even worse, high-dimensional information could be misleading. Section 2.2.1 provides a simple example to illustrate this.

When comparing approach (i), the evaluation of multivariate density forecasts, with (ii), the evaluation of the corresponding univariate projections, we focus on density forecasts within (skew) elliptical families of distributions. These families of elliptical distributions are introduced here for mathematical convenience since they are closed under linear affine transformations. Typical elliptical families of distributions are the Normal and t-distributions. Besides these symmetric elliptical distributions, two more flexible families of distributions, the skew normal and skew t-distributions defined by Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (2003), respectively, are considered as well.

The last question we address in this chapter is whether due to the dynamics of the underlying distribution, it is possible to adaptively select the optimal distribution for out-of-sample VaR forecasting. For example, Opschoor et al. (2017) develop forecast combination schemes that assign weights to individual predictive densities based on their behavior. In a similar manner, the recently best-performing density can be selected to assess the downside risk of a given portfolio.

The contribution of this chapter is two-fold. First, by employing skew elliptical distributions that are closed under affine transformations and linear aggregation, we shed some light on the pitfalls of multivariate modeling in portfolio risk assessment. Second, the time-varying nature of the relative performance of families of forecast densities is demonstrated empirically and confirmed by the superior performance of dynamically selected densities in VaR estimation.

The remainder of this chapter is organized as follows. Section 2.2 describes the methodology, including the testing environment, the scoring rules, test statistics and the distributional assumptions. Section 2.3 provides Monte Carlo simulation experiments to study the empirical size and power of the tests. In Section 2.4 we firstly apply the tests to returns of a portfolio and on its underlying stocks. Based on the findings, we further develop a simple technique to dynamically select the distribution for out-of-sample VaR forecasts. Finally, Section 2.5 summarizes and concludes.
CHAPTER 2. DENSITY FORECASTS OF PORTFOLIO RETURNS

2.2 Methodology

2.2.1 Density Comparison within Elliptical Families

We are concerned with the evaluation of the predictive accuracy of two competing multivariate density forecasts versus evaluating their corresponding univariate projections. From a portfolio perspective, the forecasted densities of a portfolio consisting of some assets could be evaluated in the multivariate space of the returns of the individual assets, or in the univariate space of the portfolio return. The implied distribution of the portfolio return, if explicitly available, will be directly relevant to the assessment of the portfolio risk. There are some circumstances in which the higher-dimensional information is irrelevant, or even misleading. When we perform a statistical test on two competing density forecasts, \( \hat{f} \) and \( \hat{g} \), say, it may happen that the multivariate scoring rule favors distribution \( \hat{f} \) over \( \hat{g} \), while the univariate scoring rule suggests the opposite.

Recall that Fig. 1.1 illustrates such a paradox, where even though density forecast \( \hat{g} \) is closer to the true distribution in the bivariate space, for the portfolio risk evaluation it makes more sense to make decisions based on the distribution of \( \hat{f} \). This simple example addresses a dilemma faced by academia as well as industry; when the true distribution is unknown and with a linear combination of some marginal elements in hand, what is the basis for portfolio analysis, the high-dimensional information or the univariate distribution after aggregation?

In general, the comparison of density forecast evaluations across different dimensions is complicated by the fact that not many distributions are analytically tractable upon aggregation by taking linear combinations or projections. Starting from a multivariate density forecast, we also need to be able to analytically track the univariate density of the projection on the portfolio returns to study their properties. In this chapter, we therefore deliberately limit ourselves to distributions from the elliptical class (ED class) of distributions for both mathematical convenience and numerical accuracy. Elliptical distributions, which are closed under affine transformations and linear aggregation, provide a tool by which we can analytically trace closed-form expressions for the densities. Specifically, the property of the ED family which facilitates our analysis is the following.

**Property 1.** If the \( d \)-variate random variable \( Y \sim ED(\Delta, \Omega) \), where \( \Delta \) and \( \Omega \) are the first two moments, given any \( \ell \times d \) matrix \( A \) of rank \( \ell \leq d \), the random vector \( Z = AY \sim ED(A\Delta, A\Omega A') \).

Typical families of distributions in the elliptical class include the Normal, \( t \) and
2.2. Methodology

Cauchy families of distributions. We refer to Kelker (1970), Cambanis et al. (1981), Fang and Ng (1990), Arellano-Valle and Bolfarine (1995) for developments in multivariate ED distribution theory.

Here we apply the Normal and t-distributions as these two better appear to fit financial data than the Cauchy distribution. Prop. 1 is crucial to our study in the sense that a linear combination of elliptically distributed assets remains in the ED family. Hence compared to the corresponding multivariate density forecast, which is usually costly to estimate, the density forecast of the portfolio return is easier to handle. Some handy properties and analytical expressions for multivariate t-distributions are summarized in Appendix 2.6.1 for ease of reference.

Naturally, the next question is whether the class of ED models is rich enough to describe the random variables involving risk management and asset allocation accurately. To answer that, we wish to consider more families of elliptical distributions than just the symmetric Normal and t-distributions. Section 2.2.4 introduces skew normal and skew t-distributions developed in a number of papers by Azzalini and co-authors (1996, 2003, 2005, 2013). The merit of using these two skew families of distributions is that they allow for additional asymmetry while remaining closed under affine transformations.

2.2.2 Testing Approach

Consider a stochastic vector \( \{(Y_t', X_t)\}, t = 1, 2, ..., T \), where \( Y_t = (Y_{1,t}, ..., Y_{d,t})' \) represents the \( d \)-dimensional vector the density of which is of interest, and where \( X_t \) is a vector of exogenous or observable predictor variables. In the context of time series data it is natural to consider the conditional distribution \( F_{Y_{t+1}}(y|\mathcal{F}_t) \), where \( \mathcal{F}_t \) denotes the information available at time \( t \). For simplicity we restrict ourselves to one-step-ahead density forecasts; generalizations to multi-step-ahead forecasts are straightforward but would provide little extra insight.

Following Giacomini and White (2006), we compare forecast methods rather than forecast models. By forecast method, we mean the model on which the forecast is based, along with estimation methods, applied observation weights and all other choices one makes at the time of the prediction. The only restriction is that the density forecasts depend on a fixed number \( m \) of most recent observations. The advantage of this framework is that it allows for treating parameter estimation uncertainty as an integral part of the density forecasts. A fixed rolling window of length \( m \) is used for estimation. As shown by Giacomini and White (2006), this limited memory scheme affords considerable analytical convenience for the asymptotic theory of the test of
equal predictive accuracy of two competing density forecasts $\hat{f}_t$ and $\hat{g}_t$.

A prevalent approach to comparing the relative performance of density forecasts is based on scoring rules, which are loss functions whose arguments are the density forecast and the actual outcome of the variable; see Diebold and Lopez (1996). In the current context, the scoring rule for one-step-ahead forecasts is of the form $S(\hat{f}_t; y_{t+1})$, depending on the density forecast $\hat{f}_t$ of $Y_{t+1}$ given $\mathcal{F}_t$ and the actually observed value $y_{t+1}$, such that a ‘better’ prediction receives a higher score on average.

Given a scoring rule $S(\cdot; \cdot)$, two competing density forecasts $\hat{f}_t$ and $\hat{g}_t$ and the corresponding realizations of the $d$-dimensional variable $y_{t+1}$ for $t = m, ..., T - 1$, we may compare $\hat{f}_t$ and $\hat{g}_t$ based on their mean scores. The test we perform in this chapter, following Giacomini and White (2006), is an unconditional predictive ability test (although the density forecasts are made conditional on $m$ in-sample observations).

Under the moving window scheme, $n = T - m$ out-of-sample observations are available, and a formal test for whether the scores received by $\hat{f}_t$ and $\hat{g}_t$ are significantly different is performed. Defining the score differences as

$$d_{t+1} = S(\hat{f}_t; y_{t+1}) - S(\hat{g}_t; y_{t+1}),$$

the null hypothesis of equal predictive ability is given by

$$H_0 : \mathbb{E}(d_{t+1}) = 0, \quad \text{for } t = m, ..., T - 1,$$

which is tested against the alternative hypothesis $H_1 : \mathbb{E}(d_{t+1}) \neq 0$ (or $> 0$ or $< 0$ for a one-sided test). Let $\bar{d}_{m,n}$ denote the out-of-sample average score difference: $\bar{d}_{m,n} = n^{-1} \sum_{t=m}^{T-1} d_t$, a Diebold and Mariano (1995) type statistic is given by

$$(2.1) \quad t_{m,n} = \frac{\bar{d}_{m,n}}{\sqrt{\hat{\sigma}^2_{m,n}/n}},$$

where $\hat{\sigma}^2_{m,n}$ is a heteroskedasticity and autocorrelation-consistent (HAC) variance estimator for the long-run variance $\sigma^2_{m,n} = \text{Var}(\sqrt{n}\bar{d}_{m,n})$ given by $\hat{\sigma}^2_{m,n} = \hat{\gamma}_0^2 + 2 \sum_{k=1}^{K-1} a_k \hat{\gamma}_k$, in which $\hat{\gamma}_k$ denotes the sample covariance of sequence $\{d_{t+1}\}$ at lag $k$ and $a_k = 1 - k/K$ where $K = \lfloor n^{1/4} \rfloor$.

Under the assumptions of a fixed estimation window size $m$ and mild mixing conditions on the sequence of score differences $\{d_{t+1}\}$, the test statistic $t_{m,n}$ is asymptotically standard normally distributed by Theorem 4 in Giacomini and White (2006). A test of (asymptotic) significance level $\alpha$ therefore is obtained by rejecting the null hypothesis of equal performance when $|t_{m,n}| > z_{\alpha/2}$ for two-sided test, where $\alpha/2$ is the $1 - \alpha/2$
quantile of the standard normal distribution. Note that the sign of the test statistic $t_{m,n}$ indicates which of the two density forecasts \( \hat{f}_t \) and \( \hat{g}_t \) performs better on average, as scores will be defined such that a higher average score is preferred.

### 2.2.3 Weighted Logarithmic Scoring Rules

We focus on three different logarithmic scoring rules as they can be applied regardless of the dimension of \( Y \). A typical logarithmic scoring rule is of the form $S(\hat{f}_t; y_{t+1}) = \log \hat{f}_t(y_{t+1})$, and testing for equality of the average scores of \( \hat{f}_t \) and \( \hat{g}_t \) leads to the pseudo-likelihood ratio test originally developed by Vuong (1989). It has been shown in many studies that the pseudo-likelihood ratio test is closely related to the Kullback-Leibler Information Criterion (KLIC), an information theoretical goodness-of-fit measure of divergence between two probability distributions; see Vuong (1989) and Bao et al. (2004). KLIC-based scores therefore quantify the divergence between a candidate density forecast \( \hat{f}_t \) and the true density \( p_t \). For our purpose of comparing candidate density \( \hat{f}_t \) and \( \hat{g}_t \) on a particular region such as the left tail, a weighted scoring rule is used, with a weight function \( w(y_{t+1}) \) emphasizing the region of interest.

It is a natural requirement that the true density \( p_t \) is rewarded with the highest average score, otherwise the associated test might suggest that some incorrect density forecast is significantly better than the true predictive density. To avoid such situations, we focus on proper scoring rules, which implies that no density forecast \( \hat{f}_t \) receives a higher score, on average, than the true (unknown) density \( p_t \); that is,

$$\mathbb{E}_t(S(\hat{f}_t; y_{t+1})) \leq \mathbb{E}_t(S(p_t; y_{t+1})), \text{ for all } t.$$  

As mentioned above, Matheson and Winkler (1976) and Gneiting and Raftery (2007) suggested many suitable proper scoring rules, but these can not be readily extended to a multivariate context. We therefore restrict ourselves to KLIC-based scoring rules, which are practically convenient and closely related to pseudo-likelihood tests. Particularly, we consider the following three scoring rules, the first two of which were suggested by Diks et al. (2011), and the last by Pelenis (2014).

1. **The conditional likelihood scoring rule (cl)**

   \[
   S^{cl}(\hat{f}_t; y_{t+1}) = w_t(y_{t+1}) \log \left( \frac{\hat{f}_t(y_{t+1})}{\int w_t(s) \hat{f}_t(s) ds} \right); \tag{2.2}
   \]
2. The censored likelihood scoring rule \((csl)\)
\[
S^{csl}(\hat{f}_t; \mathbf{y}_{t+1}) = w_t(\mathbf{y}_{t+1}) \log(\hat{f}_t(\mathbf{y}_{t+1})) + (1 - w_t(\mathbf{y}_{t+1})) \log \left(1 - \int w_t(s) \hat{f}_t(s) ds\right);
\]

3. The penalized weighted likelihood scoring rule \((pwl)\)
\[
S^{pwl}(\hat{f}_t; \mathbf{y}_{t+1}) = w_t(\mathbf{y}_{t+1}) \log(\hat{f}_t(\mathbf{y}_{t+1})) - \int w_t(s) \hat{f}_t(s) ds + w_t(\mathbf{y}_{t+1})
\]

At this point, we make the following assumptions.

**Assumption 1.** The density forecasts \(\hat{f}_t\) and \(\hat{g}_t\) satisfy \(\text{KLIC}(\hat{f}_t) < \infty\) and \(\text{KLIC}(\hat{g}_t) < \infty\), where \(\text{KLIC}(h_t) = \int p_t(y) \log(p_t(y)/h_t(y)) dy\) is the Kullback-Leibler divergence between the density forecast \(h_t\) and the true conditional density \(p_t\).

**Assumption 2.** The weight function \(w_t(y)\) is such that (a) it is determined by the information available at time \(t\), and hence a function of \(\mathcal{F}_t\), (b) \(0 \leq w_t(y) \leq 1\), and (c) \(\int w_t(y) p_t(y) dy > 0\).

Under Assumptions 1 and 2, the \(cl\) scoring rule (Eq. (2.2)) and the \(csl\) scoring rule (Eq. (2.3)) are proper, as established by Lemma 1 in Diks et al. (2011). In a similar manner, it is shown in Appendix 2.6.2 that the \(pwl\) scoring rule (Eq. (2.4)) is proper in terms of Kullback-Leibler divergences between weighted density forecast and the true density.

To illustrate the usage of the above-mentioned weighted scoring rules, we revisit the example shown by Fig. 1.1, focusing on the left tail of the distributions by adopting the threshold weight function \(w(z) = I(z \leq r)\), where \(I(\cdot)\) denotes the indicator function taking the value 1 if its argument is true, and 0 otherwise, and where \(z = b_1 y_1 + b_2 y_2\). In order to make the rejection rate more comparable for different threshold values \(r\), we let the sample size \(n\) be determined by the threshold value as \(n = \lceil c/P(Z \leq r) \rceil\), for some \(c > 0\), so that the expected number of observations in the region of interest, is fixed at \(c\).

**Example** We generate 10,000 independent and identically distributed (i.i.d) observations from the bivariate vector \(\mathbf{Y} = (Y_1, Y_2) \sim \mathcal{N}(0, I)\) for \(c = 40\) and compare the
2.2. Methodology

Figure 2.1: One-sided rejection rates (at nominal size 5%) for the Diebold-Mariano type test statistic of equal predictive accuracy under the threshold weight function \( w(z) = I(z \leq r) \) for \( c = 40 \) expected observations in the left tail, based on 10,000 replications. The DGP is i.i.d. standard bivariate normal. Panel (a) shows high rejection rates against superior predictive ability of \( \hat{f} \) as a function of \( r \), while rejection in panel (b) suggests \( \hat{g} \) is more accurate.

Figure 2.1: One-sided rejection rates (at nominal size 5%) for the Diebold-Mariano type test statistic of equal predictive accuracy under the threshold weight function \( w(z) = I(z \leq r) \) for \( c = 40 \) expected observations in the left tail, based on 10,000 replications. The DGP is i.i.d. standard bivariate normal. Panel (a) shows high rejection rates against superior predictive ability of \( \hat{f} \) as a function of \( r \), while rejection in panel (b) suggests \( \hat{g} \) is more accurate.

performance of the two forecast candidates \( \hat{f} \) and \( \hat{g} \) defined by

\[
\hat{f} : \quad Y \sim N(\mu_1, \Sigma_1), \quad \text{with } \mu_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad \text{and } \Sigma_1 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix},
\]

\[
\hat{g} : \quad Y \sim N(\mu_2, \Sigma_2), \quad \text{with } \mu_2 = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix} \quad \text{and } \Sigma_2 = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}.
\]

Again we take \( b_1 = b_2 = 1 \) as the portfolio weights. Fig. 2.1 shows the observed one-sided rejection rates of the \( csl \) scoring rule for \( r \in [-2.5, 2.5] \), evaluated in both bivariate space and univariate space, denoted by the solid and dashed lines, respectively. The rejection rates for the \( cl \) and \( pwl \) scoring rule behave similarly, and are available from the authors upon request.

Several interesting conclusions can be drawn from Fig. 2.1. First, the test based on the bivariate \( csl \) scoring rule has power against the null hypothesis of equal predictive accuracy and favors \( \hat{g} \), while the test based on the univariate \( csl \) scoring rule suggests the other way around. Especially for the left tail, when \( r \leq -2 \), the power of this one-sided test against the alternative that \( \hat{f} \) is better, is higher than 0.5. This result is quite robust, even for \( c = 20 \) observations in the region of interest (on average), the rejection rate for the univariate scoring rule against superior predictive ability of
\( \hat{f} \) tends to 0.4 in the left tail.

Second, when comparing Figs. 1.1 and 2.1, the apparently conflicting conclusions from the Diebold-Mariano type tests are not surprising after all. When constructing the linear combination \( Z = b_1 Y_1 + b_2 Y_2 \), essentially we project the bivariate vector \( Y \) from the plane \( \mathbb{R}^2 \) onto the real line \( \mathbb{R} \). During this projection, the relative distance between the density forecasts and the true distribution changes. Fig. 1.1 suggests that the univariate projection of \( \hat{f} \) is more accurate than \( \hat{g} \), even though \( \hat{f} \) is in fact the worse multivariate forecast.

Third, from this simple example it is not hard to see that it would be a mistake for a risk manager to believe that \( \hat{g} \) is more appropriate based on the bivariate scores. Since the projected distribution of \( \hat{g} \) is thinner-tailed than that of the DGP in the left tail (see Fig. 1.1(b)), the estimated VaR and ES will be insufficient to reflect the real risk level and large losses might be incurred.

The previous example illustrates the fact that a better forecast in \( \mathbb{R}^d \) space may not lead to a better risk assessment from a portfolio perspective. Note that if one forecast, say \( \hat{f} \), is perfect in the sense that \( \hat{f} \equiv p \), where \( p \) is the true conditional density implied by the DGP, we will not be in such a dilemma because the perfect forecast \( \hat{f} \) will dominate \( \hat{g} \) in both the univariate space and the bivariate space. However, the issue of evaluating competing densities with multivariate scoring rule concerns us because in reality \( p \) is never reached; \( \hat{f} \) and \( \hat{g} \) are always based on a misspecified model and therefore always incorrect to some extent. Even when one of the models would be correctly specified, parameter estimation uncertainty would prevent its associated density forecast from being perfect.

Hence, we observe that decisions based on the multivariate scoring rule need not coincide with that based on the corresponding univariate scoring rule. Arguably, from a portfolio risk assessment perspective, good univariate density forecasts are needed for portfolio returns, and these may not necessarily be the projected multivariate optimal forecasts.

### 2.2.4 Skew Elliptical Distributions

As Bauwens and Laurent (2005) noted, one well-established stylized fact of financial returns is that they often exhibit fat tails and skewness, and a more suitable distribution than the multivariate normal is of primary importance in modeling and inference. A general class of multidimensional distributions which allow for heavy tails and skewness will be useful to modeling multivariate random variables. Applications from Bauwens and Laurent (2005) and Giot and Laurent (2003) show that the skew \( t \)-distribution,
2.2. Methodology

when fully taking into account skewness and fat tails, provides a better fit for the value-at-risk of portfolios.

However, the density defined by Bauwens and Laurent (2005), which was derived using a nonlinear transformation of a symmetric density function, is inspiring, yet difficult to apply in the present context, where the aim is to compare the predictive distributions in different spaces. As their skew t-distribution does not have closed form expression under affine transformations, the distribution of a linear combination of marginal variables remains unknown. Jones and Faddy (2003) also considered a skew generalization of the t-distribution which is tractable for all moments, but which is difficult to generalize to the multivariate case.

Alternatively, skewness can be modeled by perturbing a symmetric distribution to generate asymmetric densities. This method was introduced by Azzalini and Dalla Valle (1996) to construct a multivariate skew-normal distribution. Later, Azzalini and Capitanio (2003) generalized the approach to distributions of elliptical families. A coincident result of multivariate skew-elliptical distributions is given by Branco and Dey (2001), although a different method is used. According to Azzalini and Capitanio (2003), there are at least two avenues to construct a skew-distributed random variable, first, the conditioning method, as used by Branco and Dey (2001), and second, the transformation method (e.g. Azzalini and Dalla Valle, 1996). We will not cover the mathematical details here, but instead refer to Azzalini (2005) and Azzalini (2013) for the intensive development of distribution theory in this direction.

Before introducing the density and some properties of skew t-distributions, we first introduce the simpler skew normal distributions. Following Azzalini and Capitanio (2003), given a $d \times d$ variance-covariance matrix $\Sigma$, define the square root of the main diagonal elements matrix $\omega = (\text{diag}(\Sigma))^{1/2}$ such that $\Sigma = \omega^{-1} \Sigma \omega^{-1}$ is the associated correlation matrix. A $d$-dimensional random vector $Y$ has a skew normal distribution, denoted by $Y \sim \text{SN}_d(\mu, \Sigma, \alpha)$ if it is continuous with density function of the type

$$f(y) = 2\phi(y, \mu, \Sigma)\Phi\left(\alpha^T \omega^{-1}(y - \mu)\right), \quad (2.5)$$

referring to $\mu$, $\Sigma$ and $\alpha$ as the location, scale and skewness parameters, respectively. Here $\phi(y, \mu, \Sigma)$ is the density function of a $d$-dimensional normal variate with mean $\mu$ and variance-covariance matrix $\Sigma$, and $\Phi\left(\alpha^T \omega^{-1}(y - \mu)\right)$ represents the standard univariate normal distribution function. The construction approach of our skew normal distribution is closely linked to the elliptical family of distributions; as a consequence, the distribution in (5) shares various properties with the elliptical family. Among these, particularly two are of major interest to us, namely, the properties related to moments
and affine transformations given below.

**Property 2.** (a) Moments: If $Y \sim SN_d(\mu, \Sigma, \alpha)$, $E(Y) = \mu + \omega \mu_S$ and $\text{Cov}(Y) = \Sigma - \omega \mu_S \mu_S^T \omega$, where $\mu_S = b\delta$ is the expectation of a standard skew normal variate (i.e. $S \sim SN_d(0, \bar{\Sigma}, \alpha)$) with a correlation matrix $\bar{\Sigma}$ such that $\Sigma = \omega \bar{\Sigma} \omega$, with $b$ and $\delta$ defined as

$$
    b = \sqrt{2/\pi}, \\
    \delta = (1 + \alpha^T \bar{\Sigma} \alpha)^{-1/2} \bar{\Sigma} \alpha.
$$

(b) Affine transformation: If $Y \sim SN_d(\mu, \Sigma, \alpha)$, given $Z = c + AY$, with $\ell \times d$ matrix $A$ of rank $\ell \leq d$ and $c \in \mathbb{R}^\ell$, we have $Z \sim SN_\ell(\mu_Z, \Sigma_Z, \alpha_Z)$ where

$$
    \mu_Z = c + A\mu, \\
    \Sigma_Z = A\Sigma A', \\
    \alpha_Z = (1 - \delta^T \omega_A \Sigma_Z^{-1} A \omega \delta)^{-1/2} \omega_Z \Sigma_Z^{-1} A \omega \delta.
$$

We adopt the skew normal and $t$-distribution defined by Azzalini and Capitanio (2003) based on two considerations, (i) the closed-form expression of the multidimensional density is available and straightforward for likelihood inference; (ii) the skew $t$-distribution is a special case of a skew elliptical density which possesses the properties of the distributions from the elliptical family, particularly the closure under affine transformations that allow us to project the multidimensional distribution to the real line. Next we provide the density of the skew $t$-distribution. We say that $Y$ follows a skew $t$-distribution, denoted by $Y \sim St_d(\mu, \Sigma, \alpha, \nu)$, if $Y$ is a $d$-variate skew $t$-distribution with location and scale parameters $\mu$ and $\Sigma$ respectively, defined as above; $\nu$ is the number of degrees of freedom and $\alpha \in \mathbb{R}^d$ is the shape or skewness parameter. When $\alpha = 0$, the $t_d(\mu, \Sigma, \nu)$ density is obtained again. The density function of $Y \sim St_d(\mu, \Sigma, \alpha, \nu)$ is

$$
    f(y) = 2t_d(\mu, \Sigma, \nu)T_1 \left( \alpha^T \omega^{-1}(y - \mu) \left( \frac{\nu + d}{Q_y + \nu} \right)^{1/2}; \nu + d \right),
$$

where $t_d(\mu, \Sigma, \nu)$ is the density function of a $d-$dimensional $t$-variable with $\nu$ degrees of freedom, $T_1(y; \nu + d)$ represents the standard univariate $t$-distribution function with $\nu + d$ degrees of freedom, and $\omega$, the diagonal matrix formed by the standard deviations within $\Sigma$, $Q_y$ are defined as

$$
    \omega = \text{diag}(\Sigma)^{1/2}, \\
    Q_y = (y - \mu)^T \Sigma^{-1} (y - \mu).
$$
Fig. 2.2 gives a graphical illustration of the skew $t$-distribution for univariate density in the upper two panels and contour plots in the lower panels. The above two panels show how $\alpha$ and $\nu$ may change the density function of the skew $t$-distribution. When $\alpha = 0$, we are back in the symmetric situation, i.e. the density for the standard $t$-distribution; and $St_1(y; \nu)$ will converge to a skew normal distribution as $\nu$ goes to infinity. Panel (c) shows that when the degrees of freedom parameter $\nu$ increases, the contours of the skew $t$-distribution converge to those of the skew normal distribution with the same skewness and correlation. In panel (d), we see how the correlation $\rho$ and the skewness vector $\alpha$ alter the appearance of the skew $t$ contours.

Analogous to Prop. 2 for the skew normal distribution, we have corresponding properties for the skew $t$-distribution, summarized in Prop. 3.

Property 3. (a) Moments: If $Y \sim St_d(\mu, \Sigma, \alpha, \nu)$, $E(Y) = \mu + \omega \mu_S$ for $\nu > 1$ and $\text{Cov}(Y) = \frac{\nu}{\nu-2} \Sigma - \omega \mu_S \mu_S^T \omega$ for $\nu > 2$, where $\mu_S = b_\nu \delta$ is the expectation of a standard skew $t$-variate (i.e. $S \sim St_d(0, \Sigma_S, \alpha, \nu)$ with correlation matrix $\Sigma_S$, hence $\Sigma = \omega \Sigma_S \omega$, with $\delta$ defined as in (6) and

$$b_\nu = \left( \frac{\nu}{\pi} \right)^{1/2} \frac{\Gamma\left(\frac{1}{2}(\nu - 1)\right)}{\Gamma\left(\frac{1}{2}\nu\right)}, \quad \text{for } \nu > 1.$$

(b) Affine transformation: If $Y \sim St_d(\mu, \Sigma, \alpha, \nu)$, given $Z = c + AY$, with $\ell \times d$ matrix $D$ of rank $\ell \leq d$ and $c \in \mathbb{R}^\ell$, we have $Z \sim St_\ell(\mu_Z, \Sigma_Z, \alpha_Z, \nu)$, with $\mu, \Sigma, \alpha$ as in Eq. (2.6).

Note that when $d = 1$ in the Eqs. (2.5) and (2.7), we obtain the special cases of the univariate skew normal distribution $SN(\mu, \Sigma, \alpha)$ and skew $t$-distribution $St(\mu, \Sigma, \alpha, \nu)$, respectively, with $\delta = \alpha/\sqrt{1 + \alpha^2}$ and $\omega = \sqrt{\Sigma}$.

### 2.3 Monte Carlo Simulations

In this section we will examine some finite-sample properties of our test of equal predictive ability. The proposed scoring rules are applied to the skew elliptical distributions. In the first two simulation experiments we illustrate the fact that the univariate skew $t$ assumption is more flexible in the sense that it incorporates skewness and fat tails simultaneously. The last simulation presents the difference between density forecast comparisons based on multivariate modeling and univariate modeling.

Since we are more interested in comparing density forecasts in the tails, we adopt a simple weight function $w(Y) = I(Y \leq r)$. To make the rejection rates obtained for
Figure 2.2: Graphical illustration for the skew t-distributions. The top two panels present the univariate skew t-density functions for $\alpha = 3$ in the left and $\alpha = 10$ in the right column, for different values of $\nu$. The bottom panels show contour plots for the standardized bivariate skew t-distribution. In the lower-left panel, given skewness vector $\alpha = (2, 6)$ and correlation $\rho = 0.5$, the contours of the skew t and skew normal distributions are shown; the lower-right panel illustrates changes in the contour plot when correlation and skewness are introduced.
2.3. Monte Carlo Simulations

different values of \( r \) more comparable, we determine the sample size \( n = \lceil c/P(Y \leq r) \rceil \) based on the threshold value in such a way that the expected number of observations in the region of interest is fixed at \( c \) across all selected values of \( r \). Besides, we standardize all forecast distributions to the same mean (zero) and variance (one) for a fair comparison. Appendix 2.6.3 presents the standardized skew Normal and \( t \)-distributions.

To be realistic, we limit ourselves to small values of \( c \). Fig. 2.3 shows the observed rejection frequency at the 5% significance level for \( c = 20 \), based on 10,000 replications. The data are drawn from the standardized univariate skew \( t(5) \)-distribution for the upper panels and the standardized univariate skew normal in the lower panels. In Fig. 2.3, the null hypothesis is equal predictive accuracy of univariate skew \( t(5) \) with 5 degrees of freedom and skewness parameter \( \alpha = -2 \) and univariate skew normal with the same \( \alpha \). Thus the left (right) column reports rejection rates against better density forecasts based on skew \( t \) density (skew normal) as a function of the threshold value \( r \).

In a similar manner, Fig. 2.4 presents rejection rates of equal predictive accuracy against better performance of the skew \( t(5) \)-distribution with \( \alpha = -2 \) (left column) and the \( t(5) \)-distribution (right column) based on 10,000 replications, when for the upper (lower) two panels data are generated from the skew \( t(5) \) (standardized \( t(5) \)) distribution, respectively.

Three conclusions can be drawn from Figs. 2.3 and 2.4. First, in both experiments the three scoring rules all give satisfactory power, and the power of \( pwl \) is close to that of \( csl \) all the time. They both outperform the \( cl \) rule when the threshold \( r \) takes very negative values, which is not strange as the \( cl \) rule does not take into account the probability of an outcome in the region of interest, as noted by Diks et al. (2011). As \( r \to \infty \), the three scoring rules behave similarly.

Second, it is quite straightforward to see that if the number of degrees of freedom \( \nu \) increases, the power curve in Fig. 2.3 will decrease as \( St(\nu) \) converges to SN in the limit; similarly, when \( \alpha \to 0 \), \( St(5) \) will become indistinguishable from \( t(5) \). This is observed in simulations (not included here due to space considerations).

Lastly, the power curves from the top-left and bottom-right panels in both simulations are decreasing in a non-monotonic way. The power is sloping downward because of the decreasing sample size; as the threshold \( r \) increases, the sample size \( n \) shrinks and as a consequence testing power drops. The non-monotonic power function can be explained by the fact that one density could gain a higher score in one region but a lower score in another. As \( r \) changes, the mean score difference between two competing densities could change sign and lead to a lower power. Diks et al. (2011) provide a numerical study on this when comparing \( t \) and normal densities.

Next, we illustrate the idea of density comparison under a linear projection of a den-
Figure 2.3: One-sided rejection rates (at nominal size 5%) for the Diebold-Mariano type test statistic of equal predictive accuracy under the threshold weight function \( w(y) = I(y \leq r) \) for \( c = 20 \) expected observations in the left tail, based on 10,000 replications. The DGP is i.i.d. skew \( t(5) \) (skew normal) with \( \alpha = -2 \) for the upper (lower) two panels. The left (right) column shows rejection rates against superior predictive ability of skew \( t(5) \) (skew normal) with \( \alpha = -2 \) as a function of the threshold parameter \( r \).
2.3. Monte Carlo Simulations

Figure 2.4: One-sided rejection rates (at nominal size 5%) for the Diebold-Mariano type test statistic of equal predictive accuracy under the threshold weight function \( w(z) = I(z \leq r) \) for \( c = 20 \) expected observations in the left tail, based on 10,000 replications. The DGP is i.i.d. skew \( t(5) \) with \( \alpha = -2 \) \((t(5))\) in upper (lower) two panels. The left (right) column shows rejection rates against superior predictive ability of skew \( t(5) \) with \( \alpha = -2 \) \((t(5))\) as a function of the threshold parameter \( r \).
sity from higher-dimensional space into one-dimensional space, in the framework of the skew elliptical distributions. Fig. 2.5 shows the rejection rates of tests on equal predictive accuracy of \(S_t^4(\mu, \Sigma, \alpha, \nu)\) and skew normal \(SN_t^4(\mu, \Sigma, \alpha)\) and their corresponding univariate projections with the equal-weight mapping vector \((1/4, 1/4, 1/4, 1/4)\) over the whole support, where \(\nu = 5\), \(\alpha = (-2, 1, -1, -1)\). We assume a strong correlation \(\rho_{ij} = 0.7\) with \(i = 1, \ldots, 4\) and \(j \neq i\). We standardize the distribution in such a way that the data are centered around the origin and each marginal distribution has unit variance. After the affine transformation (projection), the data are standardized again to have mean zero and variance one. The sample size \(n\) takes values 100, 200, 300, 500, 800, 1,000, 2,000, 3,000, 5,000 and 8,000.

The first and last panels in Fig. 2.5 clearly show that the null hypothesis of equal predictive ability is rejected. In both panels, it can be observed that the test based on the univariate density is less powerful than the test based on the four-dimensional density forecasts as the dashed power curve lies below the real line, especially for small sample sizes. However, the power of the test based on the univariate density increases rapidly with the sample size. As seen in the two off-diagonal panels, neither test has spurious power.

Fig. 2.5 suggests that the test based on the univariate scoring rule after projection to the portfolio linear combination is as reliable as the test based on the multivariate distribution, although the former suffers from lack of power for small samples, which is caused by information loss during the affine transformation. The relative lack of power, however, is not necessarily a bad thing. It only means that the two competing densities after projection are not distinguishable. The high rejection rates seen from the multivariate scoring rule might be just ‘spurious’ power in the sense that the strongly favored density forecast in the high-dimensional space may not be a better forecast for the portfolio density. Recall that we obtained contradicting test results for the example in Section 2.2.3 based on the univariate scoring rule and the bivariate scoring rule. Thus, from the perspective of a portfolio manager, it may not be optimal to perform portfolio risk measurement based on multivariate modeling.

A final comment concerns the number of degrees of freedom parameter \(\nu\). Although not presented in detail in this chapter, simulations show that when \(\nu = 20\) the power of the test deteriorates; since the two competing densities become very similar, the power based on the univariate densities drops, and spurious power can be observed in the off-diagonal panels.
2.3. Monte Carlo Simulations

Figure 2.5: One-sided rejection rates (at nominal size 5%) for the Diebold-Mariano type test statistic of equal predictive accuracy, based on 10,000 replications. The DGP is i.i.d. $S_t(5)$ ($SN_4$) for the upper (lower) two panels, with $\alpha = (-2, 1, -1, -1)$. The left (right) column shows rejection rates against superior predictive ability of $S_t(5)$ ($SN_4$) as a function of sample size. $wl$ denotes the test of weighted likelihood scoring rule directly on four-dimensional density, while $wl_u$ is the test based on the univariate density projected by the vector $(1/4, 1/4, 1/4, 1/4)$. 
2.4 Empirical Application

2.4.1 Comparison of density forecasts across dimensions

We present the comparison of out-of-sample density forecasts for a daily portfolio consisting of three asset returns using the parametric univariate and multivariate model of the GARCH family. The comparison is twofold: a comparison between using forecasted multivariate densities versus their univariate projections, and a comparison between the univariate projections and directly forecasted portfolio densities.

We consider daily data for three US stocks (Source: Yahoo Finance): the Alcoa Stock (AA), the McDonald’s stock (MCD) and the Merck stock (MRK) over the period from April 2, 1984 until May 14, 2015. For these stock prices, daily log-returns defined as \( y_t = \log(P_t/P_{t-1}) \) are examined, with the adjusted closing price \( P_t \) at day \( t \).\(^1\) This yields a total of 7,845 valid observations. A weighted portfolio with fixed weights allocated for 70% to AA, 20% to MCD and 10% to MRK is studied.\(^2\)

Our work is similar to that of Giot and Laurent (2003) in the following respects: first, both studies look at the same portfolio of daily asset returns with fixed weights. Secondly, they also provide a comparison of the performance of the assumed distributions that the innovation terms follow; there normal, standardized \( t \) and skew \( t \) are considered, although the expression of the skew \( t \)-distribution differs from ours. Thirdly, a ‘stability window’ of 50 days is applied there for parameter updating, which looks similar to our moving window scheme at first glance. However, a moving window is used here to satisfy a condition for the asymptotic distribution of the test statistic to be normal. Besides, their estimation sample is augmented as time moves forward, in contrast to the fixed estimation window of observations used here. More importantly, this chapter shifts focus from finding a better VaR measure to testing the more suitable distribution of the given portfolio return. Put differently, this chapter provides a direct comparison between multivariate densities with that based on the corresponding univariate projections.

Instead of back-testing the in-sample density, the out-of-sample forecasts are of

\(^1\)Strictly speaking the portfolio return is a linear combination of individual stock returns only for simple returns \((P_t - P_{t-1})/P_{t-1}\). However, for daily returns the differences between the two types of returns are small, and since the return distributions proposed in the literature commonly are defined for (joint) log-returns between \(-\infty \) to \(\infty \), it is more natural to work with log-returns in the applications.

\(^2\)We also performed the same procedure on two other portfolios, namely with the equal allocation weights \((1/3,1/3,1/3)\) and the reverse weights on AA and MRK, \((0.1,0.2,0.7)\), than that discussed in the main text. It turns out that the choice of portfolio weight may lead to slightly different evaluation results, which confirms our initial conjecture that the projection vector plays a role when we generate the univariate portfolio density from the combination of multivariate marginals.
2.4. Empirical Application

more practical interest. As Bao et al. (2007) found, the specification of the standardized innovations has a larger impact on density forecast accuracy than the volatility specification. Therefore, to illustrate the idea of comparing density forecasts under projections, we define four forecast methods which differ only in the prior assumption of the family of distributions from which the standardized innovations are drawn, namely Standard normal, $t$, skew normal and skew $t$. Parametric GARCH(1, 1) models, jointly with DCC(1, 1) updating of the correlations in the multivariate case, are used to estimate based on past returns and make one-step-ahead forecasts. The log-returns series are treated as having conditional mean zero, and conditional variance-covariance matrix, following Engle (2002), given by

$$Y_t | \mathcal{F}_{t-1} \sim \text{Dist}(0, H_t).$$

The covariance matrix $H_t$ can be decomposed as

$$H_t = D_t R_t D_t,$$

where $D_t$ is the $d \times d$ diagonal matrix of time-dependent standard deviations from the univariate GARCH(1,1) models with $\sqrt{h_{i,t}}$ the $i$-th diagonal element and $R_t$ is the estimated time-varying correlation matrix. The evolution of individual conditional variances $h_{i,t}$, $i = 1, \ldots, d$, and the correlations is determined by

$$h_{i,t} = \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{i,t-1},$$

$$Q_t = (1 - a - b)\bar{Q} + a \varepsilon_{t-1} \varepsilon_{t-1}' + b Q_{t-1},$$

$$R_t = Q_t^{-1} Q_t Q_t^{-1},$$

where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{d,t})'$ is the vector of the standardized residuals $\varepsilon_{i,t} = y_{i,t} / \sqrt{h_{i,t}}$ and where $\bar{Q} = \mathbb{E}(\varepsilon_t \varepsilon_t')$ represent the conditional correlation matrix. Lastly, $Q^*$ is a diagonal matrix with the square root of the $i$-th diagonal element of $Q$ on its $i$-th diagonal position. Once each individual GARCH(1, 1) model (2.9) has been estimated, $\varepsilon_t$ is stored to estimate the dynamics of the correlation (2.10)-(2.11).

We use a moving-window scheme to predict the one-day-ahead density. The estimation window is set to $m = 1,000$, and the remaining observations are used for out-of-sample evaluation. The Diebold-Mariano test defined in Eq. (2.1) is applied, with scoring rules Eqs. (2.2) to (2.4). As mentioned earlier, the left tail of the distribution of the portfolio return is of interest. Given the portfolio return series $\{r_t\}$ with the fixed individual weights, we use the truncated weight function $w_t(r_t) = I(r_t \leq \hat{r}_t^q)$. 

25
CHAPTER 2. DENSITY FORECASTS OF PORTFOLIO RETURNS

The time-varying threshold $\hat{r}_t^q$ is calculated as the empirical $q$-th quantile of all return observations in the most recent estimation window, where $q = 0.01, 0.05, 0.1, 0.15$ are considered.

Table 2.1 gives the average score differences and the test result for the censored likelihood scoring rule Eq. (2.3). Similar results from the conditional likelihood rule Eq. (2.2) and penalized weighted likelihood rule Eq. (2.4) are shown in Tables 2.2 and 2.3. The four sub-panels in Table 2.1 present results based on different choices for the threshold quantile $q$. The numbers in parentheses under the score differences are the values of the corresponding Diebold-Mariano test statistics, where those with an asterisk are significant at the 5% significance level. The test statistics are calculated based on the average score differences $\bar{d}_{m,n}$ with the HAC estimator adjusted for serial dependence as in Eq. (2.1). As mentioned before, we want to compare the evaluation of the density forecasts in both high-dimensional space and in the projected univariate space. The label $csl_{multi}$ denotes test results based on trivariate density evaluation, and $csl_{proj}$ uses the projected parameter values for density evaluation in one-dimensional space, where the asset allocation weight $(0.7, 0.2, 0.1)$ plays the role of the projection vector. The test results for $csl_{uni}$ in the last row is obtained using portfolio return series $\{r_t\}$ for estimation and prediction directly.

Table 2.1 indicates that, firstly, the scores of the $t$ and skew $t$ forecasts overwhelm those of the standard Normal and skew Normal in most comparisons, which is not surprising, as stock returns are often characterized by fat tails. Secondly, the order of priority among the other two pairs is hard to assess. The signs of the score differences suggest that the standard Normal and $t$ forecasts outperform the skew Normal and skew $t$ forecasts. However, significant test statistics are rarely seen.

Thirdly, test results from $csl_{proj}$ do not always lead to the same conclusion as $csl_{multi}$. For example, for $q = 0.05$, $csl_{multi}$ suggests that the $t$ forecast is significantly better than the skew $t$ forecast, while only insignificant score difference are witnessed by $csl_{proj}$. The divergence is even larger for the comparison between standard Normal and skew Normal for $q = 0.05$, where $csl_{multi}$ favors the skew Normal since it delivers negative score difference; on the contrary, $csl_{proj}$ prefers the standard Normal. The counterintuitive results from different dimensional spaces illustrates the idea of the example in Section 2.2.3 again; the better forecasts for the multivariate distribution may not yield better forecasts for the univariate projection. Therefore, from the portfolio risk assessment point of view, it seems better to consider with the univariate portfolio returns rather than the underlying multivariate returns.

Fourthly, when comparing the test results of $csl_{proj}$ and $csl_{uni}$, only two differences are observed. From the pair comparison $t$-$SN$ at $q = 1\%$ and $t$-$St$ at $q = 15\%$, both
### Table 2.1: Average score differences and test statistics based on the csl rule

<table>
<thead>
<tr>
<th>scoring rule</th>
<th>$q = 0.01$</th>
<th>$q = 0.05$</th>
<th>$q = 0.10$</th>
<th>$q = 0.15$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N - t$</td>
<td>$N - SN$</td>
<td>$N - St$</td>
<td>$t - SN$</td>
</tr>
<tr>
<td><strong>csl_multi</strong></td>
<td>-0.0104</td>
<td>0.0017</td>
<td>-0.0089</td>
<td>0.0122</td>
</tr>
<tr>
<td></td>
<td>(-2.7627*)</td>
<td>(0.7224)</td>
<td>(-2.6501*)</td>
<td>(3.3909*)</td>
</tr>
<tr>
<td><strong>csl_proj</strong></td>
<td>-0.0067</td>
<td>0.0001</td>
<td>-0.0061</td>
<td>0.0068</td>
</tr>
<tr>
<td></td>
<td>(-2.4209*)</td>
<td>(0.0736)</td>
<td>(-2.5974*)</td>
<td>(3.2254*)</td>
</tr>
<tr>
<td><strong>csl_uni</strong></td>
<td>-0.0057</td>
<td>0.0003</td>
<td>-0.0048</td>
<td>0.0060</td>
</tr>
<tr>
<td></td>
<td>(-2.1393*)</td>
<td>(0.2811)</td>
<td>(-2.0449*)</td>
<td>(1.8565)</td>
</tr>
</tbody>
</table>

Note: Average score difference $d$ and the corresponding test statistic (the number in parentheses) for the censored likelihood scoring rule (Eq. (2.3)) for different empirical quantiles $q$ in the indicator weight function $w_t(r_t) = I(r_t \leq \hat{r}_t^q)$, where for the four sub-panels, $q = 0.01, 0.05, 0.1, 0.15$ respectively. The superscript * denotes that the test statistic is significant at the 5% level of significance. There are six pair-comparisons since we have four distributional assumptions about the innovations. The test result in the first row of each sub-panel is based on multivariate regression model, and the corresponding projection onto one-dimensional space yields the outcome in the second row. The bottom row of each sub-panel gives test results based on the univariate regression model of portfolio return series.
Table 2.2: Average score differences and test statistics based on the $c_l$ rule

<table>
<thead>
<tr>
<th>scoring rule</th>
<th>$N - t$</th>
<th>$N - SN$</th>
<th>$N - St$</th>
<th>$t - SN$</th>
<th>$t - St$</th>
<th>$SN - St$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{multi}$</td>
<td>-0.0084</td>
<td>-0.0010</td>
<td>-0.0068</td>
<td>0.0074</td>
<td>0.0015</td>
<td>-0.0059</td>
</tr>
<tr>
<td></td>
<td>(-2.5060*)</td>
<td>(-0.5256)</td>
<td>(-2.3475*)</td>
<td>(2.0735*)</td>
<td>(1.5836)</td>
<td>(-1.8065)</td>
</tr>
<tr>
<td>$c_{proj}$</td>
<td>-0.0046</td>
<td>-0.0025</td>
<td>-0.0041</td>
<td>0.0021</td>
<td>0.0005</td>
<td>-0.0016</td>
</tr>
<tr>
<td></td>
<td>(-2.0268*)</td>
<td>(-2.9047*)</td>
<td>(-2.2145*)</td>
<td>(1.2272)</td>
<td>(0.9889)</td>
<td>(-1.2484)</td>
</tr>
<tr>
<td>$c_{uni}$</td>
<td>-0.0043</td>
<td>0.0006</td>
<td>-0.0032</td>
<td>0.0049</td>
<td>0.0010</td>
<td>-0.0039</td>
</tr>
<tr>
<td></td>
<td>(-1.8019)</td>
<td>(0.8646)</td>
<td>(-1.5998)</td>
<td>(1.6195)</td>
<td>(1.9506)</td>
<td>(-1.4604)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>scoring rule</th>
<th>$N - t$</th>
<th>$N - SN$</th>
<th>$N - St$</th>
<th>$t - SN$</th>
<th>$t - St$</th>
<th>$SN - St$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{multi}$</td>
<td>-0.0174</td>
<td>0.0225</td>
<td>0.0031</td>
<td>0.0011</td>
<td>-0.0055</td>
<td>(2.0735*)</td>
</tr>
<tr>
<td></td>
<td>(4.3397*)</td>
<td>(2.1550*)</td>
<td>(2.1449*)</td>
<td>(1.6494)</td>
<td>(1.9237)</td>
<td></td>
</tr>
<tr>
<td>$c_{proj}$</td>
<td>-0.0074</td>
<td>0.0044</td>
<td>0.0003</td>
<td>-0.0035</td>
<td>-0.0052</td>
<td>(2.2436*)</td>
</tr>
<tr>
<td></td>
<td>(-3.0417*)</td>
<td>(2.4364*)</td>
<td>(2.4364*)</td>
<td>(1.6494)</td>
<td>(1.9237)</td>
<td></td>
</tr>
<tr>
<td>$c_{uni}$</td>
<td>-0.0056</td>
<td>0.0066</td>
<td>0.0011</td>
<td>-0.0035</td>
<td>-0.0052</td>
<td>(2.0735*)</td>
</tr>
<tr>
<td></td>
<td>(-2.2393*)</td>
<td>(2.0364*)</td>
<td>(2.0364*)</td>
<td>(1.6494)</td>
<td>(1.9237)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>scoring rule</th>
<th>$N - t$</th>
<th>$N - SN$</th>
<th>$N - St$</th>
<th>$t - SN$</th>
<th>$t - St$</th>
<th>$SN - St$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_{multi}$</td>
<td>-0.0121</td>
<td>-0.0224</td>
<td>-0.1152</td>
<td>0.0991</td>
<td>0.0063</td>
<td>-0.0028</td>
</tr>
<tr>
<td></td>
<td>(-1.3915)</td>
<td>(-0.6834)</td>
<td>(-1.3184)</td>
<td>(1.8069)</td>
<td>(3.5978*)</td>
<td>(-1.6911)</td>
</tr>
<tr>
<td>$c_{proj}$</td>
<td>-0.0021</td>
<td>-0.0074</td>
<td>0.0065</td>
<td>0.0011</td>
<td>-0.0054</td>
<td>(3.0549*)</td>
</tr>
<tr>
<td></td>
<td>(-1.6109)</td>
<td>(-2.8563*)</td>
<td>(3.0549*)</td>
<td>(1.2834)</td>
<td>(2.8686*)</td>
<td></td>
</tr>
<tr>
<td>$c_{uni}$</td>
<td>-0.0062</td>
<td>0.0078</td>
<td>0.0013</td>
<td>-0.0065</td>
<td>-0.0054</td>
<td>(3.0549*)</td>
</tr>
<tr>
<td></td>
<td>(-2.3658*)</td>
<td>(2.3603*)</td>
<td>(2.3603*)</td>
<td>(1.6528)</td>
<td>(2.2425*)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Average score difference $\bar{d}$ and the corresponding test statistic (the number in parentheses) for the conditional likelihood scoring rule (Eq. (2.2)) for different empirical quantiles $q$ in the indicator weight function $w_t(r_t) = I(r_t \leq \hat{r}_q^t)$, where for the four sub-panels, $q = 0.01, 0.05, 0.1, 0.15$, respectively. The superscript * denotes that the test statistic is significant at the 5% level of significance. There are six pair-comparisons since we have four distributional assumptions about the innovations. The test result in the first row of each sub-panel is based on multivariate regression model, and the corresponding projection onto one-dimensional space yields the outcome in the second row. The bottom row of each sub-panel gives testing result based on the univariate regression model of portfolio return series.
2.4. Empirical Application

Table 2.3: Average score differences and test statistics based on the \(pwl\) rule

<table>
<thead>
<tr>
<th>scoring rule</th>
<th>(N - t)</th>
<th>(N - SN)</th>
<th>(N - St)</th>
<th>(t - SN)</th>
<th>(t - St)</th>
<th>(SN - St)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pwl_{multi})</td>
<td>-0.0104</td>
<td>0.0015</td>
<td>-0.0088</td>
<td>0.0119</td>
<td>0.0016</td>
<td>-0.0103</td>
</tr>
<tr>
<td></td>
<td>(-2.7584*)</td>
<td>(0.6313)</td>
<td>(-2.6425*)</td>
<td>(3.3268*)</td>
<td>(1.2167)</td>
<td>(-3.1664*)</td>
</tr>
<tr>
<td>(pwl_{proj})</td>
<td>-0.0066</td>
<td>-0.0001</td>
<td>-0.0061</td>
<td>0.0065</td>
<td>0.0005</td>
<td>-0.0060</td>
</tr>
<tr>
<td></td>
<td>(-2.4172*)</td>
<td>(-0.0606)</td>
<td>(-2.5927*)</td>
<td>(3.1492*)</td>
<td>(0.6457)</td>
<td>(-3.2981*)</td>
</tr>
<tr>
<td>(pwl_{uni})</td>
<td>-0.0057</td>
<td>0.0003</td>
<td>-0.0048</td>
<td>0.0060</td>
<td>0.0009</td>
<td>-0.0050</td>
</tr>
<tr>
<td></td>
<td>(-2.1356*)</td>
<td>(0.2768)</td>
<td>(-2.0393*)</td>
<td>(1.8519)</td>
<td>(1.4161)</td>
<td>(-1.7906)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>scoring rule</th>
<th>(N - t)</th>
<th>(N - SN)</th>
<th>(N - St)</th>
<th>(t - SN)</th>
<th>(t - St)</th>
<th>(SN - St)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pwl_{multi})</td>
<td>-0.0215</td>
<td>0.0060</td>
<td>-0.0173</td>
<td>0.0275</td>
<td>0.0042</td>
<td>-0.0233</td>
</tr>
<tr>
<td></td>
<td>(-4.3302*)</td>
<td>(1.7873)</td>
<td>(-3.7774*)</td>
<td>(5.3561*)</td>
<td>(2.5726*)</td>
<td>(-4.9076*)</td>
</tr>
<tr>
<td>(pwl_{proj})</td>
<td>-0.0088</td>
<td>0.0007</td>
<td>-0.0074</td>
<td>0.0095</td>
<td>0.0014</td>
<td>-0.0081</td>
</tr>
<tr>
<td></td>
<td>(-3.0078*)</td>
<td>(0.3123)</td>
<td>(-2.8031*)</td>
<td>(3.9747*)</td>
<td>(1.5495)</td>
<td>(-3.5782*)</td>
</tr>
<tr>
<td>(pwl_{uni})</td>
<td>-0.0074</td>
<td>0.0003</td>
<td>-0.0061</td>
<td>0.0077</td>
<td>0.0014</td>
<td>-0.0064</td>
</tr>
<tr>
<td></td>
<td>(-2.6340*)</td>
<td>(0.2664)</td>
<td>(-2.3442*)</td>
<td>(2.3619*)</td>
<td>(1.8647)</td>
<td>(-2.1863)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>scoring rule</th>
<th>(N - t)</th>
<th>(N - SN)</th>
<th>(N - St)</th>
<th>(t - SN)</th>
<th>(t - St)</th>
<th>(SN - St)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pwl_{multi})</td>
<td>-0.0281</td>
<td>0.0088</td>
<td>-0.0222</td>
<td>0.0369</td>
<td>0.0059</td>
<td>-0.0310</td>
</tr>
<tr>
<td></td>
<td>(-5.3676*)</td>
<td>(2.3429*)</td>
<td>(-4.4909*)</td>
<td>(6.8909*)</td>
<td>(3.3378*)</td>
<td>(-6.2352*)</td>
</tr>
<tr>
<td>(pwl_{proj})</td>
<td>-0.0101</td>
<td>0.0008</td>
<td>-0.0083</td>
<td>0.0108</td>
<td>0.0018</td>
<td>-0.0090</td>
</tr>
<tr>
<td></td>
<td>(-3.4151*)</td>
<td>(0.3408)</td>
<td>(-3.0963*)</td>
<td>(4.3593*)</td>
<td>(1.8907)</td>
<td>(-3.8081*)</td>
</tr>
<tr>
<td>(pwl_{uni})</td>
<td>-0.0084</td>
<td>0.0003</td>
<td>-0.0071</td>
<td>0.0087</td>
<td>0.0013</td>
<td>-0.0074</td>
</tr>
<tr>
<td></td>
<td>(-2.9532*)</td>
<td>(0.2873)</td>
<td>(-2.7100*)</td>
<td>(2.6515*)</td>
<td>(1.7423)</td>
<td>(-2.5323)</td>
</tr>
</tbody>
</table>

Note: Average score difference \(\bar{d}\) and the corresponding test statistic (the number in parentheses) for the penalized weighted likelihood scoring rule (Eq. (2.4)) for different empirical quantiles \(q\) in the indicator weight function \(w_t(r_t) = I(r_t \leq \hat{r}_q)\), where for the four sub-panels, \(q = 0.01, 0.05, 0.1, 0.15\) respectively. The superscript * denotes that the test statistic is significant at the 5% level of significance. There are six pair-comparisons since we have four distributional assumptions about the innovations. The test result in the first row of each sub-panel is based on multivariate regression model, and the corresponding projection onto one-dimensional space yields the outcome in the second row. The bottom row of each sub-panel gives testing result based on the univariate regression model of portfolio return series.
scoring rules favor the $t$ forecasts. However $csl_{proj}$ delivers a sharp conclusion at the 5\% level, while $csl_{uni}$ shows less significant results. Apart from that, the results from $csl_{proj}$ and $csl_{uni}$ are highly in agreement. This is numerical support for univariate rather than multivariate modeling, avoiding excessive computations for multivariate density forecast evaluation, without losing crucial information.

Further investigation of the projected densities and univariate densities is carried out by direct pairwise comparison of the corresponding univariate density forecast accuracies. Table 2.4 reports the average score difference and the associated test statistics for the univariate model-based densities against their rivaling projected densities, based on scoring rules Eqs. (2.2) to (2.4). From there, we see very robust result, especially for the $csl$ and $pwl$ rules. The $t$ and skew $t$ forecasts strongly outperform those of the normal and skew normal, regardless of whether these forecasts are generated using univariate modeling or projected multivariate modeling. Besides, all diagonal statistics in each sub-panel are insignificant, which indicates again that the test results based on the projected densities are in line with the results from testing univariate densities.

### 2.4.2 Dynamic selection of forecasted distributions

So far, the results provided in Tables 2.1 to 2.4 made use of the full sample, yielding the conclusion that the skew $t$ and $t$ forecasts are best over the entire sample. However, assuming a fixed family of distributions for stock returns across the entire historical sample is not very realistic, especially for a sample over the last thirty years, which witnessed a number of severe crises. A more practically relevant question is, what is the underlying distribution of the portfolio returns within shorter sub-periods? The remaining part of this section will seek to provide insights into the time-varying nature of the conditional densities by using the test statistic dynamically within sub-samples.

We apply an iterative procedure on the univariate daily return series, namely, a portfolio consisting of 70\% investment in an aluminum producer AA, 20\% in service industry for MCD and the last 10\% in medicine manufacturer MRK. Each sub-sample consists of an estimation window of length $m = 1,000$, the same as before, and an out-of-sample evaluation window. The length of the evaluation window is also fixed at 1,000, which corresponds to about four years. At each iteration, the univariate GARCH(1,1) model is estimated using the portfolio return data available in the sample, and a one-step-ahead forecast is calculated and stored for later evaluation. As the estimation window rolls forward until all data except the last in the evaluation sample are consumed, we calculate the test statistic. For example, in Figs. 2.6 to 2.8 the first estimation window and the first evaluation window are denoted by a light blue
Table 2.4: Average score differences and test statistics of univariate and projected densities

<table>
<thead>
<tr>
<th></th>
<th>$q = 0.01$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>$q = 0.05$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>$q = 0.10$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>$q = 0.15$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N_{uni}$</td>
<td>$t_{uni}$</td>
<td>$SN_{uni}$</td>
<td>$St_{uni}$</td>
<td>$p_{ul}$</td>
<td>$N_{uni}$</td>
<td>$t_{uni}$</td>
<td>$SN_{uni}$</td>
<td>$St_{uni}$</td>
<td>$p_{ul}$</td>
<td>$N_{uni}$</td>
<td>$t_{uni}$</td>
<td>$SN_{uni}$</td>
<td>$St_{uni}$</td>
<td>$p_{ul}$</td>
<td>$N_{uni}$</td>
<td>$t_{uni}$</td>
<td>$SN_{uni}$</td>
<td>$St_{uni}$</td>
<td>$p_{ul}$</td>
<td>$N_{uni}$</td>
<td>$t_{uni}$</td>
</tr>
<tr>
<td>$N_{proj}$</td>
<td>0.0004</td>
<td>0.0047</td>
<td>-0.0002</td>
<td>0.0036</td>
<td></td>
<td>0.0011</td>
<td>0.0068</td>
<td>0.0009</td>
<td>0.0059</td>
<td></td>
<td>0.0011</td>
<td>0.0068</td>
<td>0.0009</td>
<td>0.0059</td>
<td></td>
<td></td>
<td>0.0004</td>
<td>0.0047</td>
<td>-0.0002</td>
<td>0.0036</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_{proj}$</td>
<td>-0.0042</td>
<td>0.0000</td>
<td>-0.0049</td>
<td>-0.0010</td>
<td></td>
<td>-0.0055</td>
<td>0.0002</td>
<td>-0.0058</td>
<td>-0.0007</td>
<td></td>
<td>-0.0055</td>
<td>0.0002</td>
<td>-0.0058</td>
<td>-0.0007</td>
<td></td>
<td></td>
<td>-0.0042</td>
<td>0.0000</td>
<td>-0.0049</td>
<td>-0.0010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SN_{proj}$</td>
<td>0.0021</td>
<td>0.0021</td>
<td>-0.0028</td>
<td>0.0011</td>
<td></td>
<td>0.0013</td>
<td>0.0070</td>
<td>0.0010</td>
<td>0.0061</td>
<td></td>
<td>0.0013</td>
<td>0.0070</td>
<td>0.0010</td>
<td>0.0061</td>
<td></td>
<td></td>
<td>0.0021</td>
<td>0.0021</td>
<td>-0.0028</td>
<td>0.0011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$St_{proj}$</td>
<td>-0.0037</td>
<td>0.0005</td>
<td>-0.0044</td>
<td>-0.0001</td>
<td></td>
<td>-0.0050</td>
<td>0.0007</td>
<td>-0.0053</td>
<td>-0.0002</td>
<td></td>
<td>-0.0050</td>
<td>0.0007</td>
<td>-0.0053</td>
<td>-0.0002</td>
<td></td>
<td></td>
<td>-0.0037</td>
<td>0.0005</td>
<td>-0.0044</td>
<td>-0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Note: Average score difference $d$ and the corresponding test statistic (the number in parentheses) of univariate densities against projected densities for scoring rules (Eq. (2.2))–(Eq. (2.4)) for different empirical quantiles $q$ in the indicator weight function $w_i(t) = 1_{t_i \leq t} I_p$, where for the four sub-panels, $q = 0.01, 0.05, 0.1, 0.15$, respectively. The superscript * denotes that the test statistic is significant at the 5% level of significance.
rectangle (labeled I) and a cyan rectangle (labeled II), respectively. The iteration procedure takes steps of 25 days, which means that after every sub-sample of length 2,000 the model parameters are estimated and the forecasts evaluated; this procedure repeats itself every 25 trading days. In total 234 sequential test statistics are calculated across time, and plotted.

Fig. 2.6 reports the dynamics of the test statistics for six pair-wise comparisons, based on scoring rules Eqs. (2.2) to (2.4) at empirical quantile level \( q = 0.05 \) and Fig. 2.7 compares the dynamics of the test statistics of the censored likelihood rule for the same competing densities of different quantile levels. The area between the two black dashed lines represents a (point-wise) 5% forecast interval. Whenever the upper boundary is reached by the test statistic, the former distribution from that pair comparison is preferred significantly, and vice versa. Even without significant results, since we have many sequential test statistics, the signs of these test statistics may be indicative of the relative predictive ability under the different distributional assumptions. For example, in the first sub-panel of Fig. 2.6, the standard Normal distribution is tested against the \( t \)-distribution; the statistics lie below zero most of the time, suggesting that \( t \) is overall more accurate, except for a short period in the 90s when the Normal distribution is slightly more suitable.

The dynamical graphs enable us to see which of the two competing innovation distributions is more appropriate during different periods. Figs. 2.6 and 2.7 show rather consistent results across different scoring rules and different quantiles \( q \). Some are in line with our findings in Table 2.1. For example, \( t \) innovations all along outperform standard Normal and skew Normal innovations, and the superiority of the standard Normal or skew Normal densities varies mostly within the significance bands. Interesting dynamics can be observed for the other three pairs, where more volatile patterns occur as a sign of regime switching from time to time. Another eye-catching phenomenon is the relatively stable regime of the statistics in the late 90s.

In the remaining part of this section we try to provide some insights into these phenomena. We take the S&P500 Index as a barometer of the stock market. The blue line in Fig. 2.8 corresponds to the S&P price index labeled on the left \( y \) axis while the right \( y \) axis denotes test statistic values. The red dash-dotted line is the \( csl \) statistic for equal predictive ability of \( t \) and skew \( t \) innovations, while the green dashed line corresponds to the test of skew Normal against skew \( t \), both replicated from Fig. 2.6. The first period of interest runs from November 1997 to October 2001, the test statistics stay quite steady for all pair-comparisons, denoted by the green rectangle (labeled III) in Fig. 2.8.

To understand this, recall that we use an estimation window of length 1,000 to
Figure 2.6: Dynamics of the two-sided Diebold-Mariano type test statistic (at nominal size 5%) of equal predictive accuracy, based on the cl, csl and pwl scoring rules. The empirical quantile for $q = 0.05$ in the indicator weight function $w_t(r_t) = I(r_t \leq \hat{r}_t^q)$ is applied for selecting the left tail region. The light blue (I) and cyan (II) rectangles represent the first estimation and evaluation window.
Figure 2.7: Dynamics of the two-sided Diebold-Mariano type test statistic (at nominal size 5%) of equal predictive accuracy, based on the csl scoring rule. The empirical quantiles for $q = 0.05, 0.10, 0.15$ in the indicator weight function $w_t(r_t) = I(r_t \leq \hat{r}_t^q)$ are applied for selecting the left tail region. The light blue (I) and cyan (II) rectangles represent the first estimation and evaluation window.
Figure 2.8: S&P500 Index and the dynamics of the two-sided Diebold-Mariano type test statistic (at nominal size 5%) of equal predictive accuracy for $t$ against skew $t$, and skew Normal against skew $t$, based on the $csl$ scoring rule. The empirical quantile for $q = 0.05$ in the indicator weight function $w_t(r_t) = I(r_t \leq \hat{r}_q t)$ is applied for selecting the left tail region. The light blue (I) and cyan (II) rectangles represent the first estimation and evaluation windows, respectively. The green (III) and pink (IV) regions indicate two periods discussed in the main text.
CHAPTER 2. DENSITY FORECASTS OF PORTFOLIO RETURNS

generate one-step-ahead forecasts and another 1,000 day data window to evaluate the two competing densities. Within this framework, the first statistic in the green region is based on a set of out-of-sample density forecasts roughly between December 1993 and October 1997. In other words, almost 8 years of out-of-sample forecasts are used to generate the dynamics inside the green area. Therefore, during the US stock market boom period of 1994–2001 before 9/11, all six pair comparisons of our test show stable, near constant, results, among which t innovations are suggested to perform better than their counterparts for the given portfolio.

Next, the pink rectangle (labeled IV) in Fig. 2.8 witnesses a striking decline of the csl statistics of the t against the skew t from 2.83 in October 2002 to -4.78 in August 2009, which starts right after the dot-com bubble burst and ends by the bear market rally during the recent financial crisis. The stock market experienced a wild fluctuation in the first decade of this century, reflected by frequent ups and downs of S&P500 Index. In this period, t innovations give way to the skew t assumption gradually, suggesting that the symmetric t-distribution is unable to model drastic negative shocks in a volatile market.

In this period we also observe an increase in the test statistics of the skew Normal against skew t before it drops to the negative significance boundary. A potential reason for the superiority of skew Normal innovations at the beginning could be the fact that the rolling estimation window does not incorporate enough negative return observations up to April 2000, when the market bubble did not burst yet. As the iterative procedure moves forward, more and more extreme shocks are included for estimation, giving rise to better model estimation and prediction using skew t innovations.

2.4.3 Value-at-Risk (VaR) Forecasts and Evaluation

This section illustrates how time-varying score averages can be used for dynamic one-step-ahead VaR forecasting. With the findings from the last subsection that the relative performances of four density assumptions are time-varying, apart from generating the VaR forecasts from the four density candidates, we may use the scoring rules Eqs. (2.2) to (2.4) to select the recently best performing distribution to base the out-of-sample VaR forecast on.

Intuitively, for a fixed estimation window up to and including time t, we estimate a parametric GARCH model. Apart from the one-day-ahead VaRs calculated based on each of the fixed distributions, we can also dynamically choose a distribution for VaR calculation at time t. With the observed returns, we can compute the scores based on Eqs. (2.2) to (2.4) and select the density with the highest average score over a short
2.4. Empirical Application

period for the one-day-ahead VaR calculation on day $t + m + 1$, where $m$ is the length of evaluation window. We consider two situations, one has the shortest evaluation window, one-day horizon; and a five-days evaluation horizon is applied in the other case.

Under the rolling window scheme, the parametric GARCH(1,1) and GARCH(1,1)-DCC(1,1) model are estimated and updated for every trading day for both the univariate modeling and multivariate modeling, respectively. In the next step, the one-step-ahead VaR forecasts are generated under different distributional assumptions. Apart from the four fixed densities, the dynamically selected density based on both the univariate and trivariate scoring rules for three different modeling methods, i.e. univariate modeling, multivariate modeling and the projection of the latter, are also used in forecasting the VaR. The estimation window is set equal to $m = 1,000$ and the evaluation windows under consideration are 1 day and 5 days, yielding 6,844 and 6,840 out-of-sample VaR forecasts, respectively.

Conceptually, the VaR is nothing but a time-varying conditional quantile, and our goal is to provide the evolution of the quantile forecast over time. In this sense, our technique and the CAViaR model proposed by Engle and Manganelli (2004) are alike, although we seek to select the highest scoring density for quantile forecasting indirectly while a direct quantile regression is performed by Engle and Manganelli (2004). One may also think of this adaptive selection technique as an analogue of the regime switching model proposed by Pelletier (2006) and Garcia and Tsafack (2011) for time-varying dependence structure. Chollete et al. (2009) constructed a multivariate regime switching copula model for VaR forecasting. This subsection is similar to their model not only in dynamically updating conditional dependence for outperforming fixed distributions in VaR forecasts, but also in the two-step estimation methodology for multivariate modeling. However, our technique is more parsimonious than regime switching models since no latent variable or transition probabilities are involved. Another advantage of our technique is that more than two ‘regimes’ can be easily accommodated with scoring rules, whereas regime switching models typically assume two regimes, e.g. a symmetric and asymmetric regime for normal and extreme situations, respectively. It is also noteworthy that selection of a single forecast can be refined to combining forecasts in an adaptive way, as has been done by Opschoor et al. (2017), where a technique of combining densities based on scoring rules is applied to obtain VaR forecasts.

To test the accuracy of the VaR forecasts from different distributions, namely four fixed families of densities $N, t, SN, St$ and the dynamically selected densities based on (Eq. (2.2))-(Eq. (2.4)), we first compute the coverage rates for 1% and 5% VaR forecasts, e.g. the percentages of the observations such that $y_{t+1} \leq \text{VaR}_{t+1}$, then consider
three different types of tests. The first two tests only keep track on the statistical accuracy of the VaR values by checking the violation frequency of the VaR forecasts with the correct unconditional coverage (CUC) of Kupiec (1995) and the correct conditional coverage (CCC) of Christoffersen (1998).

In addition, we compare two competing VaR forecasting models in the style of Diebold and Mariano (1995) by using a loss function which measures the distance between the observed returns and forecasted VaR values. The linear asymmetric linear loss function, defined by Giacomini and Komunjer (2005), is given by

\begin{equation}
L^q(e_{t+1}) = (q - I(e_{t+1} < 0))e_{t+1},
\end{equation}

where \( e_{t+1} = y_{t+1} - \text{VaR}^q_{t+1} \) and \( q \in \{0.01, 0.05\} \) is the quantile considered. Let \( d_{AB} = L^q_A - L^q_B \), where \( L^q_A \) and \( L^q_B \) are the loss functions for distributions \( A \) and \( B \), respectively. A negative value of \( d_{AB} \) indicates that the forecast based on \( A \) is superior to that based on \( B \). There are 21 models in total, namely, the four fixed distribution assumptions and three scoring-rule based selections from the univariate modeling, the projected modeling and the multivariate modeling. The Diebold-Mariano type statistic given in Eq. (2.1) will be calculated for forecast evaluation. Furthermore, the model confidence set (MCS) procedure is applied for decreasing the size of model set. The MCS procedure can yield a collection of models such that it will contain the best ones at a given confidence level. See Hansen et al. (2011) for more details. We implement the MCS procedure with the loss function Eq. (2.12) by stationary bootstrap, with a desired block length of 20 days, approximately one month.

A comparison is made in Table 2.5 for the univariate modeling of the portfolio return directly, the projected multivariate modeling with the allocation vector, and the multivariate modeling over the three individual stock returns. The results summarize the coverage rates, the \( p \)-values for the CUC test, the \( p \)-values for the CCC test and the \( t \)-statistic of the Diebold-Mariano test for the equal conditional predictive ability of two VaR models. In all three modeling frameworks, we choose the VaR based on the \( csl \) rule selected density as the benchmark, and compare the other six models against it. A negative statistic suggests that the VaR\( ^{csl}\) is preferred and vice versa. The upper two panels consider the length of evaluation window \( m = 1 \) and the bottom two panels apply \( m = 5 \).

Several conclusions can be drawn from Table 2.5. First, the empirical VaR violation rates using the dynamically selected density based rules are close to the nominal frequencies, independent of the modeling method. For the 5% quantiles, the the \( pwl \) and \( csl \) rule-based VaR forecasts outperform the forecasts generated by the fixed families.
### Table 2.5: Evaluation of One-day-ahead VaR Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate Modeling</td>
<td></td>
<td></td>
<td></td>
<td>Projected Modeling</td>
<td></td>
<td></td>
<td></td>
<td>Multivariate Modeling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.0111</td>
<td>0.3668</td>
<td>0.3525</td>
<td>-0.2753</td>
<td>0.0120</td>
<td>0.1101</td>
<td>0.0648</td>
<td>-0.6611</td>
<td>0.0120</td>
<td>0.1101</td>
<td>0.0648</td>
</tr>
<tr>
<td>t</td>
<td>0.0200</td>
<td>0.0909</td>
<td>0.3531</td>
<td>-0.1069</td>
<td>0.0076</td>
<td>0.2155</td>
<td>0.0006</td>
<td>-1.4613</td>
<td>0.0468</td>
<td>0.2175</td>
<td>0.0006</td>
</tr>
<tr>
<td>SN</td>
<td>0.0111</td>
<td>0.3668</td>
<td>0.3525</td>
<td>-1.0227</td>
<td>0.0113</td>
<td>0.9458</td>
<td>0.1938</td>
<td>0.0101</td>
<td>0.9458</td>
<td>0.1938</td>
<td>0.0101</td>
</tr>
<tr>
<td>St</td>
<td>0.0088</td>
<td>0.2949</td>
<td>0.2271</td>
<td>-0.0544</td>
<td>0.0096</td>
<td>0.7656</td>
<td>0.0001</td>
<td>0.0067</td>
<td>0.0370</td>
<td>0.0250</td>
<td>0.4752</td>
</tr>
<tr>
<td>cl</td>
<td>0.0200</td>
<td>0.1187</td>
<td>0.0865</td>
<td>-0.1587</td>
<td>0.0077</td>
<td>0.6058</td>
<td>0.3507</td>
<td>0.0067</td>
<td>0.6058</td>
<td>0.3507</td>
<td>0.0067</td>
</tr>
<tr>
<td>mwl</td>
<td>0.0112</td>
<td>0.0679</td>
<td>0.0669</td>
<td>1.0558</td>
<td>0.0132</td>
<td>0.0125</td>
<td>0.0121</td>
<td>0.0000</td>
<td>0.0132</td>
<td>0.0125</td>
<td>0.0121</td>
</tr>
<tr>
<td>csi</td>
<td>0.0112</td>
<td>0.0679</td>
<td>0.0669</td>
<td>-</td>
<td>0.0132</td>
<td>0.0125</td>
<td>0.0121</td>
<td>-</td>
<td>0.0132</td>
<td>0.0125</td>
<td>0.0121</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>m = 5, q = 0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate Modeling</td>
<td></td>
<td></td>
<td></td>
<td>Projected Modeling</td>
<td></td>
<td></td>
<td></td>
<td>Multivariate Modeling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.0443</td>
<td>0.0267</td>
<td>0.0015</td>
<td>-2.7276*</td>
<td>0.0432</td>
<td>0.0888</td>
<td>0.0001</td>
<td>-2.7276*</td>
<td>0.0432</td>
<td>0.0888</td>
<td>0.0001</td>
</tr>
<tr>
<td>t</td>
<td>0.0481</td>
<td>0.4613</td>
<td>0.0024</td>
<td>-0.3436</td>
<td>0.0469</td>
<td>0.2350</td>
<td>0.0007</td>
<td>-0.7552</td>
<td>0.0469</td>
<td>0.2350</td>
<td>0.0007</td>
</tr>
<tr>
<td>SN</td>
<td>0.0450</td>
<td>0.0538</td>
<td>0.0011</td>
<td>-1.8656*</td>
<td>0.0378</td>
<td>0.0000</td>
<td>0.0000</td>
<td>-1.5306*</td>
<td>0.0378</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>St</td>
<td>0.0492</td>
<td>0.7725</td>
<td>0.0015</td>
<td>-0.7333</td>
<td>0.0473</td>
<td>0.3086</td>
<td>0.0023</td>
<td>-1.8461*</td>
<td>0.0473</td>
<td>0.3086</td>
<td>0.0023</td>
</tr>
<tr>
<td>cl</td>
<td>0.0476</td>
<td>0.3653</td>
<td>0.0115</td>
<td>-0.3445</td>
<td>0.0465</td>
<td>0.1746</td>
<td>0.0036</td>
<td>-0.5367</td>
<td>0.0466</td>
<td>0.1933</td>
<td>0.0021</td>
</tr>
<tr>
<td>mwl</td>
<td>0.0491</td>
<td>0.8155</td>
<td>0.0409</td>
<td>-0.4504</td>
<td>0.0506</td>
<td>0.8334</td>
<td>0.1666</td>
<td>-0.7284</td>
<td>0.0506</td>
<td>0.8334</td>
<td>0.1666</td>
</tr>
<tr>
<td>csi</td>
<td>0.0491</td>
<td>0.8155</td>
<td>0.0409</td>
<td>-</td>
<td>0.0503</td>
<td>0.9205</td>
<td>0.1549</td>
<td>-</td>
<td>0.0506</td>
<td>0.8334</td>
<td>0.1666</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>m = 5, q = 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate Modeling</td>
<td></td>
<td></td>
<td></td>
<td>Projected Modeling</td>
<td></td>
<td></td>
<td></td>
<td>Multivariate Modeling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.0111</td>
<td>0.3642</td>
<td>0.3500</td>
<td>-0.9601</td>
<td>0.0120</td>
<td>0.1090</td>
<td>0.0642</td>
<td>-0.9721</td>
<td>0.0120</td>
<td>0.1090</td>
<td>0.0642</td>
</tr>
<tr>
<td>t</td>
<td>0.0080</td>
<td>0.0918</td>
<td>0.0658</td>
<td>-1.3005</td>
<td>0.0076</td>
<td>0.0374</td>
<td>0.0253</td>
<td>-0.2749</td>
<td>0.0076</td>
<td>0.0374</td>
<td>0.0253</td>
</tr>
<tr>
<td>SN</td>
<td>0.0111</td>
<td>0.3642</td>
<td>0.3500</td>
<td>-1.5471</td>
<td>0.0101</td>
<td>0.9420</td>
<td>0.1937</td>
<td>-0.2400</td>
<td>0.0101</td>
<td>0.9420</td>
<td>0.1937</td>
</tr>
<tr>
<td>St</td>
<td>0.0080</td>
<td>0.2971</td>
<td>0.0531</td>
<td>-0.7671</td>
<td>0.0086</td>
<td>0.7092</td>
<td>0.3921</td>
<td>-0.1894</td>
<td>0.0086</td>
<td>0.7092</td>
<td>0.3921</td>
</tr>
<tr>
<td>cl</td>
<td>0.0082</td>
<td>0.1198</td>
<td>0.0873</td>
<td>-1.2689</td>
<td>0.0076</td>
<td>0.0374</td>
<td>0.0253</td>
<td>-0.2889</td>
<td>0.0076</td>
<td>0.0374</td>
<td>0.0253</td>
</tr>
<tr>
<td>mwl</td>
<td>0.0121</td>
<td>0.0860</td>
<td>0.0847</td>
<td>-1.0006</td>
<td>0.0129</td>
<td>0.0226</td>
<td>0.0220</td>
<td>0.9732</td>
<td>0.0130</td>
<td>0.0167</td>
<td>0.0163</td>
</tr>
<tr>
<td>csi</td>
<td>0.0121</td>
<td>0.0860</td>
<td>0.0847</td>
<td>-</td>
<td>0.0129</td>
<td>0.0226</td>
<td>0.0220</td>
<td>-</td>
<td>0.0130</td>
<td>0.0167</td>
<td>0.0163</td>
</tr>
</tbody>
</table>

Note: Accuracy of one-day-ahead VaR estimates of the daily portfolio returns, obtained by assuming the densities to be Normal, t, skew Normal, skew t and three dynamically-selected densities based on the scoring rules cl, pwl and csi. 1% and 5% VaR results for univariate modeling, projected modeling and multivariate modeling are reported, respectively. The length of evaluation window, m, is 1 day (5 days) for upper (bottom) two panels. In each sub-panel, the table provides the corresponding coverage rate of VaR forecasts, the p-values of the CUC test and the CCC test and the t-statistic of the Diebold-Mariano test for equal conditional predictive ability against better cl rule based forecasts, using the loss function (Eq. (2.12)). Bold numbers highlight the p-values smaller than 1% and asterisks suggest that the t-statistic is significant at the 5% level.
of distributions, however this advantage is not seen from the \textit{cl} rule. For the small quantile 1\% the \textit{pwl} and \textit{csl} rule-based VaR forecasts deliver identical results, which is consistent with the finding before that the performance of the two scoring rules is very similar.

Second, the CUC and the CCC tests reject the null hypothesis of a correct coverage rate for the fixed density based VaR more than the dynamically selected density-based VaR, indicating again that it is more reasonable to assume that the underlying portfolio return distribution is time-varying rather than fixed. Especially for the 5\% quantile, both the \textit{pwl} and \textit{csl} rule-based VaR forecasts are not rejected under the CCC tests, no matter what modeling methods are used.

Among the four fixed distributional assumptions, the \textit{t} and skew \textit{t}-distributions perform as good as, if not better than, the other two distributions in the sense that the null hypothesis is rejected less often under the CUC tests, which is not surprising given the fact that \textit{t} and skew \textit{t} forecasts in general gain a higher score in the density forecast comparisons than the Normal and skew Normal forecasts. When comparing different modeling methods, the univariate modeling witnesses fewer rejections under the CUC tests than its multivariate counterparts.

Third, the last column for each modeling method reports the \textit{t}-value for the Diebold-Mariano test. It is suggested that the \textit{csl} rule based VaR forecast consistently outperforms the fixed distribution based VaR forecasts since the \textit{t}-statistics are negative more often. The Normal density and the skew Normal density based VaR estimates are significantly dominated by the benchmark for all three modeling methods for the 5\% quantile. The comparison among different dynamically selected densities is less conclusive. The \textit{pwl} rule based VaR forecasts tend to be superior to the \textit{csl} rule based ones sometimes.

Finally, we compare the VaR forecast results across the different modeling methods in Tables 2.6 and 2.7 by using the loss function Eq. (2.12) again. Table 2.6 reports the results of the Diebold-Mariano test for univariate modeling and projected modeling against the multivariate modeling with four fixed densities and three dynamic-selected densities. A negative test statistic suggests that the VaR forecasts from multivariate modeling are better. Similarly, Table 2.7 presents the results for tests on the equal VaR accuracy between the univariate modeling or projected modeling and projected modeling with different density assumptions. Furthermore, Table 2.8 summarizes the result of the MCS procedure.

From Tables 2.6 and 2.7, it can be seen that neither the multivariate modeling or the projected modeling improve the VaR accuracy. The diagonal statistics in each subpanel of the two tables suggest that the VaR based on the parsimonious GARCH(1,1)
<table>
<thead>
<tr>
<th>$m = 1$, $q = 0.01$</th>
<th>Univariate Modeling</th>
<th>Projected Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{uni}$</td>
<td>0.2070</td>
<td>0.0708</td>
</tr>
<tr>
<td>SN$_{uni}$</td>
<td>0.1901</td>
<td>0.2784</td>
</tr>
<tr>
<td>$St_{uni}$</td>
<td>0.1195</td>
<td>0.2134</td>
</tr>
<tr>
<td>$c_{uni}$</td>
<td>-0.0001</td>
<td>0.0087</td>
</tr>
<tr>
<td>$pwl_{uni}$</td>
<td>0.5228</td>
<td>0.0502</td>
</tr>
<tr>
<td>$csl_{uni}$</td>
<td>0.5314</td>
<td>0.0468</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m = 0.05$</th>
<th>Univariate Modeling</th>
<th>Projected Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{uni}$</td>
<td>-1.5868</td>
<td>-0.5601</td>
</tr>
<tr>
<td>SN$_{uni}$</td>
<td>1.4517</td>
<td>1.2094</td>
</tr>
<tr>
<td>$St_{uni}$</td>
<td>-0.2992</td>
<td>-0.1694</td>
</tr>
<tr>
<td>$c_{uni}$</td>
<td>-2.1173</td>
<td>-1.6346</td>
</tr>
<tr>
<td>$pwl_{uni}$</td>
<td>-0.6337</td>
<td>-0.8215</td>
</tr>
<tr>
<td>$csl_{uni}$</td>
<td>-2.4006</td>
<td>-2.2965</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m = 5$, $q = 0.01$</th>
<th>Univariate Modeling</th>
<th>Projected Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{uni}$</td>
<td>1.7816</td>
<td>2.2904</td>
</tr>
<tr>
<td>SN$_{uni}$</td>
<td>1.4517</td>
<td>1.4517</td>
</tr>
<tr>
<td>$St_{uni}$</td>
<td>-0.2992</td>
<td>-0.1694</td>
</tr>
<tr>
<td>$c_{uni}$</td>
<td>-2.1173</td>
<td>-1.6346</td>
</tr>
<tr>
<td>$pwl_{uni}$</td>
<td>-0.6337</td>
<td>-0.8215</td>
</tr>
<tr>
<td>$csl_{uni}$</td>
<td>-2.4006</td>
<td>-2.2965</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m = 5$, $q = 0.05$</th>
<th>Univariate Modeling</th>
<th>Projected Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{uni}$</td>
<td>1.7816</td>
<td>2.2904</td>
</tr>
<tr>
<td>SN$_{uni}$</td>
<td>1.4517</td>
<td>1.4517</td>
</tr>
<tr>
<td>$St_{uni}$</td>
<td>-0.2992</td>
<td>-0.1694</td>
</tr>
<tr>
<td>$c_{uni}$</td>
<td>-2.1173</td>
<td>-1.6346</td>
</tr>
<tr>
<td>$pwl_{uni}$</td>
<td>-0.6337</td>
<td>-0.8215</td>
</tr>
<tr>
<td>$csl_{uni}$</td>
<td>-2.4006</td>
<td>-2.2965</td>
</tr>
</tbody>
</table>

Note: The $t$-statistic of Diebold-Mariano type for the equal conditional predictive ability of pairs of competing VaR forecasts, using the loss function (Eq. (2.12)). The length of evaluation window, $m$, is 1 day (5 days) for upper (bottom) two panels. Both univariate modeling and projected modeling methods are evaluated against direct multivariate modeling for different quantiles $q$. Asterisks indicate that the $t$-statistic is significant at 5% level.
modelling for different quantiles. Asterisks suggest that the model is significant at 5% level.

### Chapter 2: Density Forecasts of Portfolio Returns

#### Table 2.7: Evaluation of One-day-ahead VaR Results against the Projected Model

<table>
<thead>
<tr>
<th>Method</th>
<th>0.05 = 0</th>
<th>0.05 = 1</th>
<th>(q = 0.05)</th>
<th>(q = 0.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projected Modeling</td>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{proj})</td>
</tr>
<tr>
<td>Univariate Modeling</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{uni})</td>
</tr>
<tr>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{proj})</td>
</tr>
<tr>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{proj})</td>
</tr>
<tr>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{proj})</td>
<td>(\tilde{\alpha}_{uni})</td>
<td>(\tilde{\alpha}_{proj})</td>
</tr>
</tbody>
</table>

Note: The table provides the evaluation of one-day-ahead VaR results against the projected model for different quantiles. The asterisks indicate significance at the 5% level.
### Table 2.8: MCS for the One-day-ahead VaR Forecasts

<table>
<thead>
<tr>
<th></th>
<th>Univariate Modeling</th>
<th>Projected Modeling</th>
<th>Multivariate Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m = 1, q = 0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.6760</td>
<td>0.2570</td>
<td>0.2570</td>
</tr>
<tr>
<td>$t$</td>
<td>0.8450</td>
<td>0.7490</td>
<td>0.8820</td>
</tr>
<tr>
<td>SN</td>
<td>0.3120</td>
<td>0.9140</td>
<td>0.9140</td>
</tr>
<tr>
<td>St</td>
<td>1.0000</td>
<td>0.8770</td>
<td>0.8770</td>
</tr>
<tr>
<td>cl</td>
<td>0.5320</td>
<td>0.7740</td>
<td>0.8210</td>
</tr>
<tr>
<td>pwl</td>
<td>0.9620</td>
<td>0.3220</td>
<td>0.7490</td>
</tr>
<tr>
<td>csl</td>
<td>0.3470</td>
<td>0.3220</td>
<td>0.7380</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 1, q = 0.05$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.0290*</td>
<td>0.0310*</td>
<td>0.0290*</td>
</tr>
<tr>
<td>$t$</td>
<td>0.2270</td>
<td>0.2800</td>
<td>0.2460</td>
</tr>
<tr>
<td>SN</td>
<td>0.0030*</td>
<td>0.0310*</td>
<td>0.0310*</td>
</tr>
<tr>
<td>St</td>
<td>0.2800</td>
<td>0.2660</td>
<td>0.2660</td>
</tr>
<tr>
<td>cl</td>
<td>0.1140</td>
<td>0.6880</td>
<td>1.0000</td>
</tr>
<tr>
<td>pwl</td>
<td>0.1480</td>
<td>0.1790</td>
<td>0.8490</td>
</tr>
<tr>
<td>csl</td>
<td>0.1970</td>
<td>0.9810</td>
<td>0.0920</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 5, q = 0.01$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.6340</td>
<td>0.3140</td>
<td>0.3140</td>
</tr>
<tr>
<td>$t$</td>
<td>0.3210</td>
<td>0.3630</td>
<td>0.5040</td>
</tr>
<tr>
<td>SN</td>
<td>0.3630</td>
<td>0.4170</td>
<td>0.3800</td>
</tr>
<tr>
<td>St</td>
<td>0.4200</td>
<td>0.4170</td>
<td>0.4170</td>
</tr>
<tr>
<td>cl</td>
<td>0.2740</td>
<td>0.4590</td>
<td>0.4940</td>
</tr>
<tr>
<td>pwl</td>
<td>0.1610</td>
<td>0.2280</td>
<td>0.3770</td>
</tr>
<tr>
<td>csl</td>
<td>1.0000</td>
<td>0.2570</td>
<td>0.2910</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 5, q = 0.05$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.0400*</td>
<td>0.0430*</td>
<td>0.0430*</td>
</tr>
<tr>
<td>$t$</td>
<td>0.8310</td>
<td>0.9010</td>
<td>0.8780</td>
</tr>
<tr>
<td>SN</td>
<td>0.0050*</td>
<td>0.0070*</td>
<td>0.0070*</td>
</tr>
<tr>
<td>St</td>
<td>0.1010</td>
<td>0.4270</td>
<td>0.4270</td>
</tr>
<tr>
<td>cl</td>
<td>0.7210</td>
<td>0.7810</td>
<td>0.6980</td>
</tr>
<tr>
<td>pwl</td>
<td>0.8460</td>
<td>0.3450</td>
<td>0.6830</td>
</tr>
<tr>
<td>csl</td>
<td>0.8380</td>
<td>1.0000</td>
<td>0.5830</td>
</tr>
</tbody>
</table>

Note: The $p$-values of MCS for the losses of one-day-ahead VaR forecasts generated by different models. The length of evaluation window, $m$, is 1 day (5 days) for upper (bottom) two panels. Both univariate modeling and projected modeling methods are evaluated against direct multivariate modeling for different quantiles $q$. The asterisks suggest that the test statistic is significant at 5% level of significance.
model is at least as good as the forecasts from the involved DCC(1,1)-GARCH(1,1) model when comparing identical distribution families with different dimensions. Besides, the comparisons of the dynamic-selection of multivariate densities based forecasts against all fixed univariate distribution based forecasts end up with negative $t$-statistics most of the time, especially the Normal and skew Normal density based VaR forecasts are rejected more often for testing equal predictive accuracy against the multivariate scoring rule-based VaR forecasts. Next, Table 2.8 presents the $p$-values for the MCS procedure when screening all possible combinations. The results suggest that the Normal and skew Normal density based modeling should be excluded from the model confidence set of the 5% VaR forecast. For different combinations of forecasting quantile $q$ and evaluation window $m$, the dynamically-selection based models are never excluded.

To summarize, the Normal and the skew Normal consistently perform worse than the $t$ or the Skew $t$ assumption in forecasting VaR given the portfolio, and the multivariate modeling method does not deliver better forecasts than the univariate modeling method. Further, by taking advantage of the time-varying distribution for the given portfolio, we may come up with a simple selection method for the underlying distribution to be assumed for VaR estimation based on scoring rules Eqs. (2.2) to (2.4), evaluated for both univariate densities and multivariate densities. Consistent results are observed in Tables 2.6 and 2.7, which suggest that the dynamically-selected distribution-based VaR forecasts outperform the fixed distributions.

2.5 Summary and Conclusions

To evaluate the risk of a given portfolio of assets, one can evaluate and select competing density forecasts for the relevant future random variables. The traditional way is to focus on forecasting the multivariate distribution of the composing assets first, and then assess the downside risk of the portfolio. However, as we showed in a simple example, information in high-dimensional space can be irrelevant or even misleading for portfolio risk measurement.

In this chapter we emphasize that it makes a difference whether competing densities are evaluated by univariate or multivariate scores. We also demonstrated empirically that, adaptive selection of forecasts improves the forecasting ability in terms of portfolio risk measurement. Based on the Kullback-Leibler information criterion, three proper scoring rules including conditional likelihood, censored likelihood and penalized weighted likelihood were used for evaluating the predictive ability of (out-of-sample) densities. We focused on distributions from skew elliptical families since these distri-
butions are closed under affine transformations. The merits are that we may explicitly trace the distribution function of the aggregate portfolio return. Monte Carlo simulations indicate that the test in higher-dimensional space could lead to a similar or opposite conclusion as the univariate test, depending on the shape and location of the projected densities.

Our empirical application to daily return series of a fixed portfolio consisting of three S&P500 stocks indicates that the $t$-distribution outperforms the Normal, Skew $t$ and Skew Normal distributions in the left tail of the support over a thirty years time horizon. The parametric GARCH(1,1) model, jointly with DCC updating for correlation in the multivariate setting, is used to successively generate one-step-ahead density forecasts. Given the fact that the dependence structure may vary across time with respect to sign and/or magnitude, we iteratively perform our test on portfolio return forecasts in a moving sub-sample to generate the dynamics of test statistics. The changes in the test statistics are visualized, from which it can be observed that the $t$-distribution overwhelms during the booming period, while the skew $t$-distribution is more appropriate in a declining market. The dynamical results show that the optimal family of distributions changes more than once in the sample period.

To further investigate the time-varying nature of the conditional distributions, a natural extension to the univariate and multivariate VaR estimation is proposed; a scoring rule-based technique to choose the distribution for VaR forecasts is applied, allowing dynamically selected forecast distributions for VaR forecasts. The empirical VaR forecasts based on this approach show improvements, confirming the time-varying nature of the underlying distribution.

The method discussed in this chapter can also be used to compare multi-step ahead density forecasts, or to compare non-parametrically estimated densities. It may be possible to improve the forecasts by considering more general distributions, but this is left for future research. Besides, given the potential dynamics of the underlying distribution, more advanced techniques could be applied in the context of density forecasting and VaR estimation. For example, possible leverage effects are ignored in our empirical study, which are known to be important in practice and worth taking into account in future work. Lastly, dynamic portfolio optimization driven by time-varying density forecasts form another topic for further research.
2.6 Appendix I

2.6.1 Density Function for Multivariate $t$-distribution

There are many different candidates for the multivariate generalization of Student’s $t$-distribution; Kotz and Nadarajah (2004) provide an extensive survey. The multivariate $t$-distribution is defined as an extension of the classical univariate $t$-distribution. For example, the probability density function of a $d$-dimensional $t$-distributed vector $Y = \{Y_1, ..., Y_d\}$ is defined by Kotz and Nadarajah (2004) as

$$ f(y) = \frac{\Gamma(\frac{\nu+d}{2})|R|^{-\frac{d}{2}}}{\Gamma(d)(\frac{\nu}{2})\nu^\frac{d}{2}} \left(1 + \frac{1}{\nu}(y - \mu)^T R^{-1}(y - \mu)\right)^{-\frac{\nu+d}{2}}, \tag{2.13} $$

with mean vector $\mu$, correlation matrix $R$ and degrees of freedom $\nu$, provided $R$ is positive definite. When $\mu = 0$, we render the standard $d$–dimensional form

$$ f(y) = \frac{\Gamma(\nu+\frac{1}{2})|R|^{-\frac{1}{2}}}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{1}{\nu}y^T R^{-1}y\right)^{-\frac{\nu+1}{2}}. $$

For $d = 1$, together with $R = 1$, we have the univariate probability density function

$$ f(y) = \frac{\Gamma(\nu+\frac{1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}. $$

The generalization of the multivariate $t$-distribution in Eq. (2.13) indeed uses the elliptical methodology, by which we mean the distribution theory for elliptical distributions (ED) and for the distributions obtained from elliptical distributions by introducing a location vector and a scale matrix, i.e. $\mu$ and $R$ used above. Shaw and Lee (2008) pointed out a difficulty with the standard ED representation of the multivariate $t$-distribution, which is that the joint distribution in the form of Eq. (Eq. (2.13)), does not factorize into the product of the $d$ marginal one-dimensional densities. However, we stick with the traditional ED representation for the multivariate $t$-distribution based on the desirable properties of ED family distributions. Indeed, since Chamberlain (1983) and Owen and Rabinovitch (1983), elliptical distributions are widely applied to characterize the distribution of portfolio returns. For example, Landsman and Valdez (2003) decomposed the portfolio risk into individual components under multivariate elliptical distribution setting. See Gupta et al. (2013) for a detailed discussion with the application of elliptical models in portfolio theory.

In this chapter, we consider a generalized form of Eq. (2.13) for the multivariate
t-distribution, with location vector $\mu$ and scale matrix $\Sigma$ and degrees of freedom $\nu$. The density is defined as

$$f(y) = \frac{\Gamma\left(\frac{\nu+d}{2}\right) |\Sigma|^{-\frac{1}{2}}}{\Gamma\left(\frac{d}{2}\right) \Gamma\left(\frac{\nu}{2}\right) \nu^\frac{d}{2}} \left(1 + \frac{1}{\nu} (y - \mu)^T \Sigma^{-1} (y - \mu)\right)^{-\frac{\nu+d}{2}}.$$  \hspace{1cm} (2.14)

It is worth mentioning that in Eq. (2.14), $\Sigma$ is not correlation matrix anymore, but a positive definite matrix which is proportional to the covariance matrix. Here we show two useful properties particularly for multivariate t-distribution defined in Eq. (2.14):

1. **Moments** If $Y \sim t_d(\mu, \Sigma, \nu)$, then $E(y) = \mu$ and $Cov(y) = \nu \Sigma / (\nu - 2)$, provided with $\nu > 2$.

2. **Affine Transformation** Given any $(\ell \times d)$ matrix $A$ of rank $\ell \leq d$, the random vector $Z = AY \sim t_d(A\mu, A\Sigma A', \nu)$.

Notice that the second property is a special case of Property 1 summarized in Section 2.1.

Next, we intend to show the density of ‘standardized’ t-distribution, i.e. the contour is centered around the origin and the covariance is an identity matrix, or for the univariate scenario, the variance equals one. Following Theorems 2.1.5 and 3.5.6 in Casella and Berger (2002), suppose $Y \sim t_d(\mu, \Sigma, \nu)$ with the density function as Eq. (2.14), we may construct $Z = \sqrt{\frac{\nu-2}{\nu}} \Sigma^{-1/2} (Y - \mu)$, which follows $t_d(0, \frac{\nu-2}{\nu} I, \nu)$, then the density function of $Z$ and relationship between $Y$ and $Z$ are

$$f_Z(z) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{d}{2}\right) \Gamma\left(\frac{\nu}{2}\right) \nu^\frac{d}{2}} \left(1 + \frac{1}{\nu} z^T z\right)^{-\frac{\nu+d}{2}},$$

$$f_Y(y) = \sqrt{\frac{\nu-2}{\nu}} |\Sigma|^{-\frac{1}{2}} f_Z\left(\sqrt{\frac{\nu-2}{\nu}} \Sigma^{-1/2} (y - \mu)\right).$$  \hspace{1cm} (2.15)

The density of aggregated marginal variables from the multivariate distribution is also of interest. The linear combination of two arguments from a bivariate t-distributed random vector $Y = (Y_1, Y_2)'$ is derived for simplicity. Assuming the weighting vector $A = (d_1, d_2)$, we could partition $\mu$ and $\Sigma$ into $\mu = (\mu_{Y_1}, \mu_{Y_2})'$, $\Sigma = (\sigma_{Y_1}^2, \sigma_{Y_1,Y_2}; \sigma_{Y_1,Y_2}, \sigma_{Y_2}^2)$ and derive the portfolio distribution of $Z = AY$, with location parameter $\mu_Z$ and scale parameter $\sigma_Z^2$ given by:

$$\mu_Z = d_1 \mu_{Y_1} + d_2 \mu_{Y_2}$$

$$\sigma_Z^2 = d_1^2 \sigma_{Y_1}^2 + d_2^2 \sigma_{Y_2}^2 + 2d_1 d_2 \sigma_{Y_1,Y_2}.$$
Note that $\sigma_Z^2$ is not the variance of $Z$; the variance of $Z$ is defined by $\text{Var}(Z) = \nu \sigma_Z^2 / (\nu - 2)$ as Property 1 states. The density function of $Z \sim t(\mu_Z, \sigma_Z^2, \nu)$ is defined as

$$f(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi \sigma_Z^2} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(z - \mu_Z)^2}{\nu \sigma_Z^2}\right)^{-\frac{\nu+1}{2}}.$$  

Finally, we provide the ‘standardized’ univariate $t$-distribution for a fair comparison with standard normal distribution $N(0, 1)$. Given $Z \sim t(0, \frac{\nu-2}{\nu}, \nu)$, which is a $t$-distributed random variable with mean 0 and variance 1, we have $Z = \frac{Y - \mu_Y}{\sqrt{\frac{\nu-2}{\nu} \sigma_Y^2}}$ and

$$f_Z(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{(\nu-2)\pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{z^2}{\nu - 2}\right)^{-\frac{\nu+1}{2}},$$

$$f_Y(y) = \sqrt{\frac{\nu-2}{\nu \sigma_Y^2}} f_Z\left(\frac{Y - \mu_Y}{\sqrt{\frac{\nu-2}{\nu} \sigma_Y^2}}\right).$$

### 2.6.2 Proof of the Properness of PWL

Define $P_t = \int w_t(s)p_t(s)ds$, and similarly $\hat{F}_t = \int w_t(s)\hat{f}_t(s)ds$. Assumption 2(b), scaling $w_t(y)$ between 0 and 1, implies that $P_t$ and $\hat{F}_t$ can be interpreted as the corresponding probability distribution functions for the true conditional density $p_t(s)$ and forecasted density $\hat{f}_t(s)$, respectively.

It is to be proven that $E_t(S(p_t; y_{t+1}) - S(\hat{f}_t; y_{t+1})) \geq 0$. Denote the conditional score difference for density forecast $p_t$ and $\hat{f}_t$ as $d_{t+1}^{\text{pwl}}(p_t, \hat{f}_t) = S(p_t; y_{t+1}) - S(\hat{f}_t; y_{t+1})$,
then
\[
E_t \left( d_{1:t+1}^\text{rel}(p_t, \hat{f}_t) \right) = \int p_t(y) \left( w_t(y) \log(p_t(y)) - P_t - w_t(y) \log(\hat{f}_t(y)) + \hat{F}_t \right) \, dy \\
= \int w_t(y)p_t(y) \log \left( \frac{p_t(y)}{\hat{f}_t(y)} \right) \, dy + \hat{F}_t - P_t \\
= \int \left( w_t(y)p_t(y) \log \left( \frac{p_t(y)}{P_t} \right) + w_t(y)p_t(y) \log \left( \frac{P_t}{\hat{f}_t} \right) \right) \, dy + \hat{F}_t - P_t \\
= P_t \int \frac{w_t(y)p_t(y)}{P_t} \log \left( \frac{w_t(y)p_t(y)}{P_t} \frac{P_t}{w_t(y)\hat{f}_t(y)/\hat{F}_t} \right) \, dy + P_t \log (\frac{P_t}{\hat{F}_t}) + \hat{F}_t - P_t \\
= P_tK \left( \frac{w_t(y)p_t(y)}{P_t}, \frac{w_t(y)\hat{f}_t(y)}{\hat{F}_t} \right) + P_t \log (P_t) - P_t \log (\hat{F}_t) + \hat{F}_t - P_t,
\]

(2.18)

where \( K(.,.) \) is the Kullback-Leibler divergence, which measures the difference between two probability distributions. According to Assumptions 1 and 2(c), \( w_t(y)p_t(y)/P_t \) and \( w_t(y)\hat{f}_t(y)/\hat{F}_t \) can be treated as pdfs, and Assumption 2(a) makes sure that the weighting function \( w_t(y) \) is a given function of \( y \) conditional on \( \mathcal{F}_t \). Thus the first term in the last line of (A.6), as a distance measurement of its arguments, is non-negative and finite. The remaining terms reach the minimum value zero if and only if \( \hat{F}_t = P_t \), provided with the positive semidefinite Hessian matrix except for the trivial situation where \( P_t = 0 \).

2.6.3 Standardization of Skew Elliptical Distributions

In this subsection, we give the density expressions for the ‘standardized’ skew Normal and ‘standardized’ t-distributions. We consider this problem because the way we form skew Normal and t-distributions does not assign location and scale parameter \( \mu \) and \( \Sigma \) as the expectation and variance, respectively. When we assume the value of \( \mu \) and \( \Sigma \), the expectation and covariance are pinned down according to Properties 2 and 3. To make the two competing densities more comparable, we adjust the parameters such that the two competing densities are centered around the origin and of unit variance, i.e. the ‘standardized’ covariance matrix coincides with the correlation matrix.

First consider the skew Normal distribution. Given \( Y \sim \text{SN}_n(\mu, \Sigma, \alpha) \), and the pdf of the form in Eq. (2.5), the expectation and covariance of \( Y \) are defined by Property 2 as \( E(Y) = \mu + \omega \mu_S \) and \( \text{Cov}(Y) = \Sigma - \omega \mu_S \mu_S^T \omega \), with \( \mu_S = \sqrt{2/\pi}(1 + \alpha^T \Sigma \alpha)^{-1/2} \Sigma \alpha \). Note that \( \Sigma = \omega \Sigma \omega \), where \( \omega \) is defined as \( \omega = (\text{diag}(\Sigma))^{1/2} \). Thus to standardize \( Y \) we may
express the covariance of $Y$ as

\begin{equation}
\text{Cov}(Y) = \omega \Sigma \omega - \omega \mu S \mu S^T \omega \equiv \Sigma.
\end{equation}

Eq. (2.19) gives the solution to $\omega$. For $n$-dimensional variate, $\omega$ is a diagonal matrix of the form

$$
\omega = \begin{pmatrix}
\omega_1 \\
\vdots \\
\omega_n 
\end{pmatrix},
$$

where $\omega_i$ is the standard deviation for the $i$th component. Using the Hadamard product instead of the inner product, we may rewrite Eq. (2.19) as

$$
\text{Cov}(Y) = \Sigma = \omega (\Sigma - \mu S \mu S^T) \omega = \omega^2 \circ (\Sigma - \mu S \mu S^T).
$$

Given the standardized $\omega$, $\mu = -\omega \mu S$ and $\Sigma = \omega \Sigma \omega$ will give the standardized skew Normal distribution.

For $Y \sim St_n(\mu, \Sigma, \alpha, \nu)$, the only difference is now $\mu S = (\nu)^{1/2} \Gamma\left(\frac{\nu-1}{2}\right) (1 + \alpha^T \Sigma \alpha)^{-1/2} \Sigma \alpha$, and Eq. (2.19) still holds. Similar as before, we could solve for the standardizing parameters $\mu$ and $\Sigma$.

When we project a multivariate distribution onto the real line, we must do the standardization again for the univariate distribution. Following the previous notation, if $Y \sim SN_n(\mu, \Sigma, \alpha, \nu)$, $Z = AY \sim SN(\mu Z, \Sigma Z, \alpha Z, \nu)$ with $(1 \times n)$ matrix $A$ and the scalar parameter $\mu Z$, $\Sigma Z$ and $\alpha Z$ are defined same as Eq. (2.7) (see Property 2(b)). Recall $Y$ is already standardized centered around zero, hence $E(Z) = E(AY) = 0$ and we only need to consider re-scale variance to be unit, which requires

\begin{equation}
\text{Var}(Z) = \Sigma' Z - \omega' Z \mu Z^2 \omega' Z \equiv 1,
\end{equation}

where $\mu_s = \sqrt{2} \alpha Z / \sqrt{\pi (1 + \alpha Z^2)}$ and $\omega'_Z = \sqrt{\Sigma' Z}$. From Eq. (Eq. (2.20)), it is easy to show that $\Sigma'_Z = 1/(1 - \mu Z^2)$. The derivation for standardizing the skew $t$-distribution is similar, apart from the definition of $\mu_s = \sqrt{\pi \Gamma(\nu-1)/\Gamma(\nu/2)}$ and the requirement for the variance becomes

$$
\text{Var}(Z) = \frac{\nu}{\pi} \Sigma'_Z - \omega'_Z \mu Z^2 \omega'_Z \equiv 1.
$$

Solving $\text{Var}(Z) = 1$ gives the scalar parameter $\Sigma'_Z$. 
Chapter 3

Transfer Entropy for Nonparametric Granger Causality Detection: An Evaluation of Different Resampling Methods

3.1 Introduction

Entropy, introduced by Shannon (1948, 1951), is an information theoretical concept with several appealing properties, and therefore wide applications in information theory, thermodynamics and time series analysis. Based on this classical measure, transfer entropy (TE) has become a popular information theoretical measure for quantifying the flow of information. This concept, which was coined by Schreiber (2000), was applied to distinguish a possible asymmetric information exchange between the variables of a bivariate system. When based on appropriate non-parametric density estimates, the TE is a flexible non-parametric measure for conditional dependence, coupling structure, or Granger causality in a general sense.

The notion of Granger causality was developed by the pioneering work of Granger (1969) to capture causal interactions in a linear system. In a more general model-free world, the Granger causal effect can be interpreted as the impact of incorporating the history of another variable on the conditional distribution of a future variable in addition to its own history. Recently, various nonparametric measures have been developed to capture such difference between conditional distributions in a more complex, and typically nonlinear system. There is a growing list of such methods, based on, among others, correlation integrals (Hiemstra and Jones, 1994), kernel density estimation (Diks and Panchenko, 2006), Hellinger distances (Su and White, 2008), copula functions (Bouezmarni et al., 2012) and empirical likelihood (Su and White, 2014).
In contrast with the above-mentioned methods, TE-based causality tests do not attempt to capture the difference between two conditional distributions explicitly. Instead, with the information theoretical interpretation, the TE offers a natural way to measure directional information transfer and Granger causality. We refer to Hlaváčková-Schindler et al. (2007) and Amblard and Michel (2012) for detailed reviews of the relation between Granger causality and directed information theory. However, the direct application of entropy and its variants, though attractive, turns out to be difficult, if not impossible altogether, due to the lack of asymptotic distribution theory for the test statistics. For example, Granger and Lin (1994) normalize the entropy to detect serial dependence with critical values obtained from simulations. Hong and White (2005) provide the asymptotic distribution for the Granger-Lin statistic with a specific kernel function. Barnett and Bossomaier (2012) derive a $\chi^2$ distribution for the TE at the cost of the model-free property.

On the other hand, to obviate the asymptotic problem, several resampling methods on TE have been developed for providing empirical distributions of the test statistics. Two popular techniques are bootstrapping and surrogate data. Bootstrapping is a random resampling technique proposed by Efron (1979) to estimate the properties of an estimator by measuring those properties from approximating distributions. The “surrogate” approach developed by Theiler et al. (1992) is another randomization method initially employing Fourier transforms to provide a benchmark in detecting nonlinearity in a time series setting. It is worth mentioning that the two methods are different with respect to the statistical properties of the resampled data. For the surrogate method the null hypothesis is maintained, while the bootstrap method does not seek to impose the null hypothesis on the bootstrapped samples. We refer to Schreiber and Schmitz (2000) and Horowitz (2001) for detailed applications of the two methods.

However, not all resampling methods are suitable for entropy-based dependence measures. As Hong and White (2005) put it, a standard bootstrap fails to deliver a consistent entropy-based statistic because it does not preserve the statistical properties of a degenerate $U$-statistic. Similarly, with respect to traditional surrogates based on phase randomization of the Fourier transform, Hinich et al. (2005) criticize the particularly restrictive assumption of linear Gaussian process, and Faes et al. (2008) point out that it cannot preserve the whole statistical structure of the original time series.

As far as we are aware, there are several applications of both methods in entropy-based tests, for example, Su and White (2008) propose a smoothed local bootstrap for entropy-based test for serial dependence, Papana et al. (2014) apply stationary bootstrap in partial TE estimation, Quiroga et al. (2002) use time-shifted surrogates to test the significance of the asymmetry of directional measures of coupling, and Marschinski and Kantz (2002) introduce the effective TE, which relies on random shuffling surrogate in estimation. Kugiumtzis (2008) and Papana et al. (2014) provide some comparisons between bootstrap and surrogate methods for entropy-based tests.
3.2. Methodology

In this chapter, we adopt TE as a test statistic for measuring conditional independence (Granger non-causality). Being aware of the fact that the analytical null distribution may not always be accurate or available in an analytically closed form, we resort to resampling techniques for constructing the empirical null distribution. The techniques under consideration include smoothed local bootstrap, stationary bootstrap and time-shifted surrogates, all of which are shown in literature to be applicable to entropy-based test statistics. Using different dependence structures, the size and power performance of all methods are examined in simulations.

The remainder of this chapter is organized as follows. Section 3.2 first provides the TE-based testing framework and a short introduction to kernel density estimation; then bandwidth selection rules are discussed. After presenting the resampling methods, including the smoothed local bootstrap and time-shifted surrogates for different dependence settings. Section 3.3 examines the empirical performance of different resampling methods, presenting the size and power of the tests. Section 3.4 considers a financial application with the TE-based nonparametric test and Section 3.5 summarizes.

3.2 Methodology

3.2.1 Transfer Entropy and its Estimator

Information theory is a branch of applied mathematical theory of probability and statistics. The central problem of classical information theory is to measure transmission of information over a noisy channel. Entropy, also referred to as Shannon entropy, is one key measure in the field of information theory brought by Shannon (1948, 1951). Entropy measures the uncertainty and randomness associated with a random variable. Supposing that $S$ is a random vector with density $f_S(s)$, its Shannon entropy is defined as

$$H(S) = - \int f_S(s) \log (f_S(s)) \, ds.$$  

There is a long history of applying information theoretical measures in time series analysis. For example, Robinson (1991) applies the Kullback and Leibler (1951) information criterion to construct an one-sided test for serial independence. Since then, nonparametric tests using entropy measures for dependence between two time series are becoming prevalent. Granger and Lin (1994) normalize the entropy measure to identify the lags in a nonlinear bivariate time series model. Granger et al. (2004) study dependence with a transformed metric entropy, which turns out to be a proper measure of distance. Hong and White (2005) provide a new entropy-based test for serial dependence, and the test statistic follows a standard normal distribution asymptotically.
Although those heuristic approaches work for entropy-based measures of dependence, these methodologies do not carry over directly to measures of conditional dependence, i.e., Granger causality. The term TE was coined by Schreiber (2000), although it appeared in the literature earlier under different names, is a suitable measure to serve this purpose. The TE quantifies the amount of information contained in one series at \( k \) steps ahead from the state of another series, given the current and past state of itself. Suppose we have two series \( \{X_t\} \) and \( \{Y_t\} \), for brevity put \( X = \{X_t\}, Y = \{Y_t\} \) and \( Z = \{Y_{t+k}\} \), further we define a three-variate vector \( W_t = (X_t, Y_t, Z_t) \), where \( Z_t = Y_{t+k} \); and \( W = (X, Y, Z) \) is used when there is no danger of confusion. Within this bivariate setting, \( W \) is a three dimensional continuous vector. In this chapter, we limit ourselves to \( k = 1 \) for simplicity, but the method can be generalized into multiple steps easily. The quantity \( \text{TE}_{X \rightarrow Y} \) is a nonlinear measure for the amount of information explained in \( Z \) (future \( Y \)) by \( X \), accounting for the information on \( Z \) already contained in \( Y \). Although TE defined in Schreiber (2000) applies to discrete variables, it is easily generalized to continuous variables. Conditional on \( Y \), \( \text{TE}_{X \rightarrow Y} \) is defined as

\[
\text{TE}_{X \rightarrow Y} = E_W \left( \log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X,Y}(X|Y)f_{Z|Y}(Z|Y)} \right)
\]

\[
= \int \int \int f_{X,Y,Z}(X,Y,Z) \log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X,Y}(X|Y)f_{Z|Y}(Z|Y)} \, dx \, dy \, dz
\]

\[
= E_W \left( \log \frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \log \frac{f_{X,Y}(X,Y)}{f_Y(Y)} - \log \frac{f_{Y,Z}(Y,Z)}{f_Y(Y)} \right)
\]

\[
= E_W \left( \log f_{X,Y,Z}(X,Y,Z) + \log f_Y(Y) - \log f_{X,Y}(X,Y) - \log f_{Y,Z}(Y,Z) \right).
\]

Using conditional mutual information \( I(Z,X|Y = y) \), the TE can be equivalently formulated in terms of four Shannon entropy terms as

\[
\text{TE}_{X \rightarrow Y} = I(Z,X|Y)
\]

\[
= H(Z|Y) - H(Z|X,Y)
\]

\[
= H(Z,Y) - H(Y) - H(Z,X,Y) + H(X,Y).
\]

In order to construct a test for Granger causality based on the TE, one first needs to show quantitatively that the TE is a proper basis for detecting whether the null hypothesis is satisfied. The following theorem, as a direct application of the Kullback–Leibler criterion, lays the quantitative foundation for testing based on the TE.

**Theorem 1.** \( \text{TE}_{X \rightarrow Y} \geq 0 \) with equality if and only if \( f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y) \).

**Proof.** The proof of Theorem 1 is given in Diks and Fang (2017a). \( \square \)

It is not difficult to verify that the condition for \( \text{TE}_{X \rightarrow Y} = 0 \) coincides with the null hy-
3.2. Methodology

Hypothesis of Granger non-causality defined in Eq. (3.4), also referred to as conditional independence or no coupling. Mathematically speaking, the null hypothesis of Granger non-causality, $H_0 : \{X_t\} \text{ is not a Granger cause of } \{Y_t\}$, can be phrased as

$$H_0 : \frac{f_{X,Y,Z}(x,y,z)}{f_Y(y)} = \frac{f_{Y,Z}(y,z)}{f_Y(y)} \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

for $(x,y,z)$ in the support of $W$. A nonparametric test for Granger non-causality seeks to find statistical evidence of violation of Eq. (3.4). There are many nonparametric measures available for this purpose, some of which are mentioned above. Eq. (3.4) provides the basis for a model-free test without imposing any parametric assumptions about the data generating process or underlying distributions for $\{X_t\}$ and $\{Y_t\}$. We only assume two things here. First, $\{X_t, Y_t\}$ is a strictly stationary bivariate process. Second, the process has finite memory, i.e., variable lags $l_X, l_Y \ll \infty$. The second (finite Markov order) assumption is needed in this nonparametric setting to make conditioning on past information feasible by conditioning on a finite number of past observations. Moreover, strict stationarity and the mixing properties implied by the finite Markov order assumption ensure that the transfer entropy can be estimated consistently through kernel density estimation of the underlying densities.

As far as we are aware, the direct use of TE to test Granger non-causality in nonparametric setting is difficult, if not impossible at all, due to the lack of asymptotic theory for the test statistic. As Granger and Lin (1994) put it, very few asymptotic distribution results for entropy-based estimators are available. Although over the years several break-throughs have been made with application of entropy to testing serial independence, the limiting distribution of TE statistic is still unknown. One may wish to use simulation techniques to overcome the lack of asymptotic distributions. However, as noted by Su and White (2008), there are estimation biases of the TE statistics for non-parametric dependence measures under the smoothed bootstrap procedure. Even for the parametric test statistic used by Barnett and Bossomaier (2012), the authors noticed that the TE-based estimator is generally biased.

3.2.2 Density Estimation and Bandwidth Selection

The non-negativity property in Theorem 4 makes $\text{TE}_{X \rightarrow Y}$ a desirable measure for constructing an one-sided test of conditional independence; any positive divergence from zero is a sign of conditional dependence of $Y$ on $X$. To estimate TE there are several different approaches, such as histogram-based estimators by Moddemeijer (1989), correlation sums by Diks and Manzan (2002) and nearest neighbor estimators by Kraskov et al. (2004). However, the optimal rule for the number of neighbor points is unclear, and as Kraskov et al. (2004) comment, a small value of neighbor points may lead to large statistical errors. A more natural method, kernel density estimators, the properties of which have been well studied, is applied here. With the plug-in kernel estimates of densities, we may replace the expectation in Eq. (3.2) by a sample
average to get an estimate for $\text{TE}_{X \rightarrow Y}$.

A local density estimator of a $d_W$-variate random vector $W$ at $W_i$ is given by

$$
\hat{f}_W(W_i) = ((n - 1)h)^{-d_W} \sum_{j \neq i} \mathbb{K} \left( \frac{W_i - W_j}{h} \right),
$$

where $\mathbb{K}$ is a kernel function and $h$ is the bandwidth. We take $\mathbb{K}(\cdot)$ to be a product kernel function defined as $\mathbb{K}(\cdot) = \prod_{s=1}^{d_W} \kappa(w_s)$, where $w_s$ is the $s^{th}$ element in $W$. Using a standard univariate Gaussian kernel, $\kappa(w) = (2\pi)^{-1/2}e^{-1/2(w^2)}$, $\mathbb{K}(\cdot)$ is the standard multivariate Gaussian kernel as described by Wand and Jones (1994) and Silverman (1986). Using Eq. (3.5) as the plug-in density estimator, and replacing the expectation by the sample mean, we obtain the estimator for the TE given by

$$
\hat{I}(Z, X | Y) = \frac{1}{n} \sum_{i=1}^{n} \left( \log \hat{f}(x_i, y_i, z_i) + \log \hat{f}(y_i) - \log \hat{f}(x_i, y_i) - \log \hat{f}(y_i, z_i) \right).
$$

If we estimate the Shannon entropy in Eq. (3.1) based on a sample of size $n$ from the $d_W$-dimensional random vector $W$, by the sample average of the plug-in density estimates, we obtain

$$
\hat{H}(W) = \frac{1}{n} \sum_{i=1}^{n} \left( \log(\hat{f}_W(W_i)) \right),
$$

then Eq. (3.6a) can be equivalently expressed in terms of four entropy estimators, that is,

$$
\hat{I}(Z, X | Y) = -\hat{H}(x_i, y_i, z_i) - \hat{H}(y_i) + \hat{H}(x_i, y_i) + \hat{H}(y_i, z_i).
$$

To construct a statistical test, we develop the asymptotic properties of $\hat{I}(Z, X | Y)$ defined in Eqs. (3.6a) and (3.6c) through two steps. In the first step, given the density estimates the consistency of entropy estimates is achieved and then the linear combination of four entropy estimates would converge in probability to the true value. The following two theorems ensure the consistency of $\hat{I}(Z, X | Y)$.

**Theorem 2.** Given the kernel density estimate $\hat{f}_W(W_i)$ for $f_W(W_i)$, where $W$ is a $d_W$-dimensional random vector with length $n$, let $\hat{H}(W)$ be the plug-in estimate for the Shannon entropy as defined in Eq. (3.6b). Then $\hat{H}(W) \xrightarrow{P} H(W)$.

**Proof.** The proof of Theorem 2 is given in Granger and Lin (1994) using results from Joe (1989). \(\square\)

The basic idea of the proof is to take the Taylor series expansion of $\log(\hat{f}_W(W_i))$ around the true value $\log(f_W(W_i))$ and use the fact that $\hat{f}_W(W_i)$, given an appropriate bandwidth
3.2. Methodology

sequence, converges to \( f_W(W_i) \) pointwise to obtain consistency. In the next step, the consistency of \( \hat{I}(Z, X|Y) \) is provided by the continuous mapping theorem.

**Theorem 3.** Given \( \hat{H}(W) \xrightarrow{P} H(W) \), with \( \hat{I}(Z, X|Y) \) defined as in Eq. (3.6c),
\[
\hat{I}(Z, X|Y) \xrightarrow{P} I(Z, X|Y).
\]

**Proof.** The proof is straightforward if one applies the Continuous Mapping Theorem. See Theorem 2.3 in Van der Vaart (2000).

Before we move to the next section, it is worth having a careful look at the issue of bandwidth selection. The bandwidth \( h \) for kernel estimation determines how smooth the density estimation is; a smaller bandwidth reveals more structure of the data, whereas a larger bandwidth delivers a smoother density estimate. Bandwidth selection is essentially a trade-off between the bias and variance in density estimation. A very small value of \( h \) could eliminate estimation bias, with a large variance. On the other hand, a large bandwidth reduces estimation variance at the expense of incorporating more bias. See Chapter 3 in Wand and Jones (1994) and Jones et al. (1996) for details.

However, the bandwidth selection for the TE statistic is more involved. To the best of our knowledge, there is a blank in the field of optimal bandwidth selection in kernel-based TE estimator. As He et al. (2012) show, when estimating the entropy estimator, two types of errors would be generated, one is from entropy estimation and the other from density estimation, and the optimal bandwidth for density estimation may not coincide with the optimal one for entropy estimation. Thus, rather than the rule-of-thumb bandwidth in Silverman (1986), which aims at optimal density estimation, the bandwidth in our study should provide an accurate estimator for \( I(Z, X|Y) \) in the minimal mean squared error (MSE) sense, say. In Diks and Fang (2017a) we develop such a bandwidth rule for a TE-based estimator. In that paper, rather than directly developing asymptotic properties for TE, we study the properties of the first order Taylor expansion of TE under the null hypothesis. The suggested bandwidth is shown to be MSE optimal and allows us to obtain asymptotic normality for the test statistic. In principal, the convergence rate of the TE estimator should be the same as the leading term of its Taylor approximation. We therefore propose to use the same rate also here, giving

\[
(3.7) \quad h = Cn^{-2/7},
\]

where \( C \) is an unknown parameter. This bandwidth would deliver a consistent test since the variance of local estimate of \( \hat{I}(Z_i, X_i|Y_i) \) will dominate the MSE. In Diks and Fang (2017a) we suggest to use \( C = 4.8 \) based on simulations, while Diks and Panchenko (2006) suggest to set \( C \approx 8 \) for autoregressive conditional heteroskedasticity (ARCH) processes. Our simulations here also may possibly prefer a larger value of \( C \) because the squared bias is of higher order and hence less concern for the TE-based statistic. A larger bandwidth could better control the estimation variance and deliver a more powerful test. As a robustness check, we adopt \( C = 8 \).
as well as $C = 4.8$. To match the Gaussian kernel, we standardize the data before estimate Eqs. (3.6a) to (3.6c) such that the transformed time series have mean zero and unit variance; very similar results are obtained by matching the mean absolute deviation instead of the variance of the standard Gaussian kernel for TE estimation.

3.2.3 Resampling Methods

To develop simulation-based tests for the null hypothesis, given in Eq. (3.4), of no Granger causality from $X$ to $Y$, or equivalently, for conditional independence, we consider three resampling techniques, i.e., (1) time shifted surrogates developed by Quiroga et al. (2002), (2) the smoothed bootstrap of Su and White (2008) and (3) the stationary bootstrap introduced by Politis and Romano (1994). The first technique is widely applied in coupling measures, as for example by Kugiumtzis (2013) and Papana et al. (2013), while the latter two have already been used for detecting conditional independence for decades. It worth mentioning that the surrogates and bootstrap methods treat the null quite differently. Surrogate data are supposed to preserve the dependence structure imposed by $H_0$ while bootstrap data are not restricted to $H_0$. It is possible to bootstrap the dataset without imposing the conditional independence structure of $\{X, Y, Z\}$ implied by the null hypothesis; see, for instance, Efron and Tibshirani (1994) for more details. To avoid resampling errors and to make different methods more comparable, we limit ourselves to methods that impose the null hypothesis on the resampled data. The following three different resampling methods are implemented with different sampling details.

1. Time-Shifted Surrogates

- (TS.a) The first resampling method only deals with the driving variable $X$. Suppose we have observations $\{x_1, ..., x_n\}$, the time-shifted surrogates are generated by cyclically time-shifting the components of the time series. Specifically, an integer $d$ is randomly generated within the interval $([0.05n], [0.95n])$, and then the first $d$ values of $\{x_1, ..., x_n\}$ would be moved to the end of the series, to deliver the surrogate sample $X^* = \{x_{d+1}, ..., x_n, x_1, ..., x_d\}$. Compared with the traditional surrogates based on phase randomization of the Fourier transform, the time-shifted surrogates can preserve the whole statistical structure in $X$. The couplings between $X$ and $Y$ are destroyed, although the null hypothesis of $X$ not causing $Y$ is imposed.

- (TS.b) The second scheme resamples both the driving variable $X$ and the response variable $Y$ separately. Similar to (TS.a), $Y^* = \{y_{c+1}, ..., y_n, y_1, ..., y_c\}$ is created given another random integer $c$ from the range $([0.05n], [0.95n])$. In contrast with the standard time-shifted surrogates described in (TS.a), in this setting we add more noise to the coupling between $X$ and $Y$. 

58
2. Smoothed Local Bootstrap

The smoothed bootstrap selects samples from a smoothed distribution instead of drawing observations from the empirical distribution directly. See Shao and Tu (2012) for a discussion of the smoothed bootstrap procedure. Based on rather mild assumptions, Neumann and Paparoditis (2000) show that there is no need to reproduce the whole dependence structure of the stochastic process to get an asymptotically correct nonparametric dependence estimator. Hence a smoothed bootstrap from the estimated conditional density is able to deliver a consistent statistic. Specifically, we consider two versions of the smoothed bootstrap that are different in dependence structure to some extent.

- (SMB.a) In the first setting, \( Y^* \) is resampled without replacement from the smoothed local bootstrap. Given the sample \( Y = \{y_1, \ldots, y_n\} \), the bootstrap sample is generated by adding a smoothing noise term \( \varepsilon_i^Y \) such that \( \tilde{y}_i^* = y_i^* + h_b \varepsilon_i^Y \), where \( h_b > 0 \) is the bandwidth used in bootstrap procedure, \( \varepsilon_i^Y \) represents a sequence of i.i.d. \( N(0, 1) \) random variables. Without random replacement from the original time series, this procedure does not disturb the original dynamics of \( Y = \{y_1, \ldots, y_n\} \) at all. After \( Y^* \) is resampled, both \( X^* \) and \( Z^* \) are drawn from the smoothed conditional densities \( f(x|Y^*) \) and \( f(z|Y^*) \) as described in Paparoditis and Politis (2000).

- (SMB.b) Secondly, we implement the smoothed local bootstrap as in Su and White (2008). The only difference between this setting and (SMB.a) is that the bootstrap sample \( Y^* \) is drawn with replacement from the smoothed kernel density.

3. Stationary Bootstrap

Politis and Romano (1994) propose the stationary bootstrap to maintain serial dependence within the bootstrap time series. This method replicates the time dependence of original data by resampling blocks of the data with randomly varying block length. The lengths of the bootstrap blocks follows a geometric distribution. Given a fixed probability \( p \), the length \( L_i \) of block \( i \) is decided as \( P(L_i = k) = (1 - p)^{k-1}p \) for \( k = 1, 2, \ldots \), and the starting points of block \( i \) are randomly and uniformly drawn from the original \( n \) observations. To restore the dependence structure exactly under the null, we combine the stationary bootstrap with the smoothed local bootstrap for our simulations.

- (STB) In short, firstly \( y_1^* \) is picked randomly from the original \( n \) observations of \( Y = \{y_1, \ldots, y_n\} \), denoted as \( y_1^* = y_s \) where \( s \in [1, n] \). With probability \( p \), \( y_2^* \) is picked at random from the data set; and with probability \( 1 - p \), \( y_2^* = y_{s+1} \), so that \( y_2^* \) would be the next observation to \( y_s \) in original series \( Y = \{y_1, \ldots, y_n\} \). Proceeding in this way, \( \{y_1^*, \ldots, y_n^*\} \) can be generated. If \( y_n^* = y_s \) and \( s = n \), the “circular boundary condition” would kick in, so that \( y_{n+1}^* = y_1 \). After \( Y^* = \{y_1^*, \ldots, y_n^*\} \) is
generated, both $X^*$ and $Z^*$ are randomly drawn from the smoothed conditional densities $f(x|Y^*)$ and $f(z|Y^*)$ as in (SMB.b).

The resampling procedure works as the follows: once the TE statistic $\hat{I}$ for the original data $W = \{(X_i, Y_i, Z_i), i = 1, ..., n\}$ is estimated according to Eqs. (3.6a) to (3.6c), we start to generate the resampled data set, which is denoted by $W^*_j$ with $j = 1, ..., B$, where $B$ is the number of simulations. Using the simulated sample, for each $j$ we compute the TE statistic $\hat{I}^*_j$, in exactly the same way as $\hat{I}$ was computed. The $p$-value for the one-sided test is calculated as

$$
\hat{p} = \frac{1}{B+1} \left( 1 + \sum_{j=1}^{B} 1(\hat{I}^*_j \geq \hat{I}) \right),
$$

where the constant 1 is added to avoid $p$-values equal to zero.

A final remark concerns the difference between this chapter and Papana et al. (2014); there both time-shifted surrogate and the stationary bootstrap are implemented for an entropy-based causality test. However, this chapter provides additional insights into several aspects. Firstly, the smoothed bootstrap, being shown in the literature to work for nonparametric kernel estimators under general dependence structure, is applied here. Secondly, they treat the bootstrap and surrogate sample in a similar way, but as we noted above, the bootstrap method is not designed to impose the null hypothesis but designed to keep the dependence structure present in the original data. The stationary bootstrap procedure in Papana et al. (2014) might be incompatible with the null hypothesis of conditional independence since it destroys the dependence completely. Because they restore independence between $X$ and $Y$ rather than conditional independence between $X|Y$ and $Z|Y$ during resampling, the distribution of the estimated statistics from the resampled data may not necessary correspond to that of the statistic under the null of only conditional independence. Thirdly, we provide rigorous size and power result in our simulations, which is missing in their paper.

### 3.3 Simulation Study

In this section, we investigate the performance of the five resampling methods in detecting conditional dependence for several data generating processes\footnote{In Diks and Fang (2017b), we consider 10 popular processes that are widely used in empirical researches for macroeconomics, finance and physics. In this dissertation, we provide only five of them for simplicity.}. In Equations (3.9)–(3.12), we use a single parameter $a$ to control the strength of the conditional dependence. The size assessment is obtained based on testing Granger non-causality from $\{X_t\}$ to $\{Y_t\}$, and for the power we use the same process but we test for Granger non-causality from $\{Y_t\}$ to $\{X_t\}$. We set $a = 0.4$ to represent moderate dependence in the size performance investigation and
3.3. Simulation Study

$a = 0.1$ to evaluate the power of the tests. Further, Eq. (3.13) is included to investigate the power performance in the presence of two-way causal linkages, where the two control parameters are $b = -0.2$ and $c = 0.1$.

In each experiment, we run 500 simulations for sample sizes $n = \{200, 500, 1,000, 2,000\}$. The surrogate and the bootstrap sample size is set to $B = 999$. For fair comparisons between (TS.a) and (TS.b), as well as between (SMB.a) and (SMB.b), we fix the seeds of the random number generator in the resampling functions to eliminate the potential effect of randomness. Besides, we use the empirical standard deviation of $\{Y_t\}$ as the bootstrapping bandwidth and $C = \{4.8, 8\}$ in the bandwidth equation Eq. (3.7) for the kernel density estimation.

The processes under consideration include a nonlinear vector autoregressive (VAR) process in Eq. (3.9), a bilinear process in Eq. (3.10), a bivariate AR(1)-EGARCH process in Eq. (3.11) where ‘EGARCH’ represents ‘exponential GARCH’, a vector error correction (VECM) process in Eq. (3.12), and a two-way VAR process in Eq. (3.13).

It is worth mentioning that the data generating processes in Eqs. (3.9), (3.10) and (3.13)) are stationary and of finite memory as we assumed earlier. However, it is also important to be aware of the behavior of the proposed non-parametric test for robustness consideration when the two assumptions are not satisfied. The finite memory assumption is violated in Equations (3.11) since the GARCH process, being equivalent to an infinite ARCH process, strictly speaking is of infinite Markov order; and the stationarity assumption does not hold in Eq. (3.12) where $X_t$ and $Y_t$ are cointegrated of order one. Since for the VECM process the two time series $\{X_t\}$ and $\{Y_t\}$ are not stationary, we can not directly apply our nonparametric test. In this case, we perform the Engle and Granger (1987) approach first to eliminate the influence of the co-integration, and then perform the nonparametric test on the collected stationary residuals from the linear regression of $\Delta X_t$ and $\Delta Y_t$ on a constant and the co-integration term. The procedure is similar to that in Bekiros and Diks (2008).

1. Nonlinear VAR. This process is considered in Baek and Brock (1992) to show the failure of linear Granger causality test.

\begin{align}
X_t &= aX_{t-1}Y_{t-1} + \varepsilon_{x,t}, & \varepsilon_{x,t} &\sim N(0,1) \\
Y_t &= 0.6Y_{t-1} + \varepsilon_{y,t}, & \varepsilon_{y,t} &\sim N(0,1).
\end{align}

(3.9)

2. Bilinear process considered in Davidson (2002).

\begin{align}
X_t &= 0.3X_{t-1} + aY_{t-1}\varepsilon_{y,t-1} + \varepsilon_{x,t}, & \varepsilon_{x,t} &\sim N(0,1) \\
Y_t &= 0.4Y_{t-1} + \varepsilon_{y,t}, & \varepsilon_{y,t} &\sim N(0,1).
\end{align}

(3.10)
3. Bivariate AR(1)-EGARCH process.

\[
X_t = 0.5X_{t-1} + \varepsilon_{x,t}, \\
\varepsilon_{x,t} = \sqrt{h_{x,t}} v_{x,t}, \quad v_{x,t} \sim t(5), \\
\log(h_{x,t}) = -0.5 + a|\varepsilon_{y,t-1}^2| + 0.2 \frac{\varepsilon_{y,t-1}^2}{\sqrt{h_{x,t-1}}} + 0.9 \log(h_{x,t-1}); \\
Y_t = 0.6Y_{t-1} + \varepsilon_{y,t}, \\
\varepsilon_{y,t} = \sqrt{h_{y,t}} v_{y,t}, \quad v_{y,t} \sim t(5), \\
\log(h_{y,t}) = -0.6 + 0.05|\varepsilon_{y,t-1}^2| + 0.02 \frac{\varepsilon_{x,t-1}^2}{\sqrt{h_{x,t-1}}} + 0.8 \log(h_{x,t-1}).
\] (3.11)

4. VECM process. Note that in this situation both \{X_t\} and \{Y_t\} are not stationary.

\[
X_t = 1.2 + 0.6Y_{t-1} + \varepsilon_{x,t}, \\
\varepsilon_{x,t} = \sqrt{1 - a^2} v_{x,t} + a \varepsilon_{y,t-1}, \quad v_{x,t} \sim N(0,1) \\
Y_t = Y_{t-1} + \varepsilon_{y,t}, \quad \varepsilon_{y,t} \sim N(0,1).
\] (3.12)

5. Two-way VAR process.

\[
X_t = 0.7X_{t-1} + bY_{t-1} + \varepsilon_{x,t}, \quad \varepsilon_{x,t} \sim N(0,1) \\
Y_t = cX_{t-1} + 0.5Y_{t-1} + \varepsilon_{y,t}, \quad \varepsilon_{y,t} \sim N(0,1).
\] (3.13)

The empirical rejection rates are summarized in Tables 3.1–3.5. The top panels in each table summarize the empirical rejection rates obtained for the 5% and 10% nominal significance levels for processes (3.9)–(3.13) under the null hypothesis, and the bottom panels report the corresponding empirical power under the alternatives. Generally speaking, the size and power are quite satisfactory for almost all combinations of the constant C, sample size n and nominal significance level. The performance differences for the various resampling schemes are not substantial.

With respect to the size performance, most of the time we see that the realized rejection rates stay in line with the nominal size. Besides, the bootstrap methods outperform the time-shifted surrogate methods in that their empirical size is slightly closer to the nominal size. Lastly, the size of the tests is not very sensitive to the choice of the constant C apart from the cases for the model given in Eq. (3.11), where the data generating process has infinite memory.

From the point of view of power, (TS.a) and (SMB.a) seem to outperform their counterparts, yet, the differences are subtle. Along the dimension of the sample size, clearly we see that the empirical power increases in the sample size in most cases. Furthermore, the results are very robust with respect to choices for the constant C in the kernel density estimation bandwidth. For the VAR and nonlinear processes given by Eqs. (3.9) and (3.10) a smaller C seems to give more powerful tests while a larger C is more beneficial for detecting conditional dependence structure in the (G)ARCH processes Eq. (3.11).
Finally, Table 3.5 presents the empirical power for the two-way VAR process, where the two variables \{X\} and \{Y\} are inter-tangled with each other. Due to the setting \( b = -0.2 \) and \( c = 0.1 \) in Eq. (3.13), it is obvious that \{Y\} is a stronger Granger cause for \{X\} than the other way around. As a consequence, the reported rejection rates in Table 3.5 are overall higher when testing \( Y \to X \) than \( X \to Y \).

To visualize the simulation results, Figures 3.1–3.5 report the empirical size and power against the nominal size. Since the performance of the five difference resampling methods is quite similar, we only show the results for (SMB.a) for simplicity. In each figure, the left (right) panels show the realized size (power), and we choose \( C = 4.8 \) (\( C = 8 \)) for the top (bottom) two panels. We can see from the figures that the empirical performance of the TE test are overall satisfactory, apart for the (G)ARCH process where a small \( C \) may lead to conservative testing size for large sample sizes (see Fig. 3.3(a)). The under-rejection problem is caused by the inappropriate choice \( C = 4.8 \), which makes the bandwidth for kernel estimation too small. The influence of an inappropriately small bandwidth can also be seen in Fig. 3.3(b), where the test has limited power for the alternative.
Table 3.1: Observed size and power of the TE-based test for the nonlinear VAR process in Eq. (3.9). For the simulation study, we consider $a = 0$ and $C = 8$. The control parameter $a = 0.4$ for size evaluation and $a = 0.1$ for establishing powers. For all simulation setups, we use the same seed across simulations to ensure comparability. Significant sizes lie between 0.05 and 0.10 for different resampling methods. The values represent observed rejection rates over 500 realizations for nominal size 0.05. Sample sizes go from 200 to 2,000. The control parameter $a = 0.4$ for size evaluation and $a = 0.1$ for establishing powers. For this simulation study, we consider $C = 4.8$ and $C = 8$.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Size $\alpha = 0.05$</th>
<th>Size $\alpha = 0.10$</th>
<th>Power $\alpha = 0.05$</th>
<th>Power $\alpha = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.0400 0.0300 0.0200</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
</tr>
<tr>
<td>500</td>
<td>0.0400 0.0300 0.0200</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
</tr>
<tr>
<td>1,000</td>
<td>0.0400 0.0300 0.0200</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
</tr>
<tr>
<td>2,000</td>
<td>0.0400 0.0300 0.0200</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
<td>0.1000 0.0500 0.0300</td>
</tr>
</tbody>
</table>

Note: Empirical size and power of the TE-based test at 5% and 10% significance levels for process Eq. (3.9) for different resampling methods. The values represent observed rejection rates over 500 realizations for nominal size 0.05. Sample sizes go from 200 to 2,000. The control parameter $a = 0.4$ for size evaluation and $a = 0.1$ for establishing powers. For this simulation study, we consider $C = 4.8$ and $C = 8$.
Table 3.2: Observed size and power of the TE-based test for the bilinear process in Eq. (3.10).

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 0.05$</td>
<td>$\alpha = 0.10$</td>
</tr>
<tr>
<td></td>
<td>TS.a</td>
<td>TS.b</td>
</tr>
<tr>
<td>$n$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.0460</td>
<td>0.0420</td>
</tr>
<tr>
<td>500</td>
<td>0.0540</td>
<td>0.0480</td>
</tr>
<tr>
<td>1,000</td>
<td>0.0440</td>
<td>0.0440</td>
</tr>
<tr>
<td>2,000</td>
<td>0.0460</td>
<td>0.0480</td>
</tr>
</tbody>
</table>

Note: Empirical size and power of the TE-based test at 5% and 10% significance levels for process Eq. (3.10) for different resampling methods. The values represent observed rejection rates over 500 realizations for nominal size 0.05. Sample sizes go from 200 to 2,000. The control parameter $a = 0.4$ for size evaluation and $a = 0.1$ for establishing powers. For this simulation study, we consider $C = 4.8$ and $C = 8$. 
<table>
<thead>
<tr>
<th>n</th>
<th>TS.a</th>
<th>TS.b</th>
<th>SMB.a</th>
<th>SMB.b</th>
<th>STB</th>
<th>TS.a</th>
<th>TS.b</th>
<th>SMB.a</th>
<th>SMB.b</th>
<th>STB</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
<td>0.0880</td>
</tr>
<tr>
<td>500</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
<td>0.0760</td>
</tr>
<tr>
<td>1,000</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
<td>0.0380</td>
</tr>
<tr>
<td>2,000</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
<td>0.0360</td>
</tr>
</tbody>
</table>

Size

Table 3.3: Observed size and power of the TE-based test for the AR(1)-EGARCH process in Eq. (3.11).

For this simulation study, we consider $a = 0.4$ and $C = 8$. The control parameter $a = 0.4$ for size evaluation and $a = 0.1$ for establishing powers. For this simulation study, we consider $C = 4.8$ and $C = 8$. Sample sizes go from 200 to 2,000. The empirical size and power of the TE-based test at 5% and 10% significance levels for different resampling methods. The values represent observed rejection rates over 500 realizations for nominal size 0.05 significance levels for process in Eq. (3.11).
### Table 3.4: Observed size and power of the TE-based test for the VECM process in Eq. (3.12).

<table>
<thead>
<tr>
<th>n</th>
<th>(C = 4.8)</th>
<th>(C = 8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\alpha = 0.05)</td>
<td>(\alpha = 0.10)</td>
</tr>
<tr>
<td></td>
<td>TS.a</td>
<td>TS.b</td>
</tr>
<tr>
<td>200</td>
<td>0.0060</td>
<td>0.0060</td>
</tr>
<tr>
<td>500</td>
<td>0.0220</td>
<td>0.0220</td>
</tr>
<tr>
<td>1,000</td>
<td>0.0400</td>
<td>0.0400</td>
</tr>
<tr>
<td>2,000</td>
<td>0.0440</td>
<td>0.0420</td>
</tr>
</tbody>
</table>

Note: Empirical size and power of the TE-based test at 5% and 10% significance levels for process Eq. (3.12) for different resampling methods. The values represent observed rejection rates over 500 realizations for nominal size 0.05. Sample sizes go from 200 to 2,000. The control parameter \(a = 0.4\) for size evaluation and \(a = 0.1\) for establishing powers. For this simulation study, we consider \(C = 4.8\) and \(C = 8\).
### Table 3.5: Observed Power of the TE-based Test for the Two-way VAR process in Eq. (3.13)

<table>
<thead>
<tr>
<th>n</th>
<th>TS(a)</th>
<th>SMB(a)</th>
<th>STB</th>
<th>TS(b)</th>
<th>SMB(b)</th>
<th>STB</th>
<th>C</th>
<th>(\alpha = 0.05)</th>
<th>(\alpha = 0.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.960</td>
<td>0.920</td>
<td>0.980</td>
<td>0.930</td>
<td>0.900</td>
<td>0.960</td>
<td>4.8</td>
<td>0.960</td>
<td>0.920</td>
</tr>
<tr>
<td>500</td>
<td>0.980</td>
<td>0.940</td>
<td>0.980</td>
<td>0.930</td>
<td>0.900</td>
<td>0.960</td>
<td>4.8</td>
<td>0.980</td>
<td>0.940</td>
</tr>
<tr>
<td>1,000</td>
<td>0.990</td>
<td>0.950</td>
<td>0.990</td>
<td>0.950</td>
<td>0.920</td>
<td>0.980</td>
<td>4.8</td>
<td>0.990</td>
<td>0.950</td>
</tr>
<tr>
<td>2,000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>4.8</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Note:** Empirical power of the TE-based test at 5% and 10% significance levels for process Eq. (3.13) for different resampling methods. The values represent observed rejection rates over 500 realizations for nominal size 0.05. Sample sizes go from 200 to 2,000. The control parameter \(b = -0.2\) and \(c = 0.0\) for establishing powers. For this simulation study, we consider \(C = 4.8\) and \(C = 8.68\).
3.3. Simulation Study

Figure 3.1: Size-size and size-power plots of Granger non-causality tests, based on 500 replications and smoothed local bootstrap (a). The DGP is the bivariate non-linear VAR process in Eq. (3.9), with $Y$ affecting $X$. The left (right) column shows observed rejection rates under the null (alternative) hypothesis. The sample size varies from $n = 200$ to $n = 2,000$. 
Figure 3.2: Size-size and size-power plots of Granger non-causality tests, based on 500 replications and smoothed local bootstrap (a). The DGP is the bilinear process in Eq. (3.10), with $Y$ affecting $X$. The left (right) column shows observed rejection rates under the null (alternative) hypothesis. The sample size varies from $n = 200$ to $n = 2,000$. 
Figure 3.3: Size-size and size-power plots of Granger non-causality tests, based on 500 replications and smoothed local bootstrap (a). The DGP is the bivariate AR1-EGARCH process in Eq. (3.11), with Y affecting X. The left (right) column shows observed rejection rates under the null (alternative) hypothesis. The sample size varies from $n = 200$ to $n = 2,000$. 

\begin{figure}[h]
\centering
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{fig3.3a}
\caption{Size-Size Plot, C=4.8}
\end{subfigure}
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{fig3.3b}
\caption{Size-Power Plot, C=4.8}
\end{subfigure}
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{fig3.3c}
\caption{Size-Size Plot, C=8}
\end{subfigure}
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{fig3.3d}
\caption{Size-Power Plot, C=8}
\end{subfigure}
\end{figure}
Figure 3.4: Size-size and size-power plots of Granger non-causality tests, based on 500 replications and smoothed local bootstrap (a). The DGP is the VECM process in Eq. (3.12), with $Y$ affecting $X$. The left (right) column shows observed rejection rates under the null (alternative) hypothesis. The sample size varies from $n = 200$ to $n = 2,000$. 

CHAPTER 3. SIMULATION-BASED TEST FOR GRANGER CAUSALITY
Figure 3.5: Size-power plots of Granger non-causality tests, based on 500 replications and smoothed local bootstrap (a). The DGP is the two-way VAR process in Eq. (3.13), with $X$ affecting $Y$ and $Y$ affecting $X$. The left (right) column shows observed rejection rates for testing $X \rightarrow Y$ (respectively, $Y \rightarrow X$). The sample size varies from $n = 200$ to $n = 2,000$. 

---

3.3. Simulation Study
3.4 Application

In this section, we apply the TE-based nonparametric test on detecting financial market interdependence, in terms of both return and volatility. Diebold and Yilmaz (2009) performed a variance decomposition of the covariance matrix of the error terms from a reduced-form VAR model to investigate the spillover effect in the global equity market. More recently, Gamba-Santamaria et al. (2017) extended the framework and considered the time-varying feature in global volatility spillovers. Their research, although providing simple and intuitive methods for measuring directional linkages between global stock markets, may suffer from the limitation of the linear parametric modeling, as discussed above. We revisit the topic of spillovers in the global equity market by the nonparametric method.

For our analysis, we use daily nominal stock market indexes from January 1992 to March 2017, obtained from Datastream, for six developed countries including the US (DJIA), Japan (Nikkei 225), Hong Kong (Hangseng), the UK (FTSE 100), Germany (DAX 30) and France (CAC 40). The target series are weekly return and volatility for each index. The weekly returns are calculated in terms of differenced log prices multiplied by 100, from Friday to Friday. Where the price for Friday is not available due to a public holiday, we use the Thursday price instead.

The weekly volatility series are generated following Diebold and Yilmaz (2009) by making use of the weekly high, low, opening and closing prices, obtained from the underlying daily high, low, opening and closing data. The volatility $\sigma_t^2$ for week $t$ is estimated as

$$\hat{\sigma}_t^2 = 0.511 (H_t - L_t)^2 - 0.019 [(C_t - O_t)(H_t + L_t - 2O_t) - 2(H_t - O_t)(L_t - O_t)] - 0.383(C_t - O_t)^2,$$

where $H_t$ is the Monday-Friday high, $L_t$ is the Monday–Friday low, $O_t$ is the Monday-Friday open and $C_t$ is the Monday-Friday close (in natural logarithms multiplied by 100). Further, after deleting the volatility estimates for the New Year week in 2002, 2008 and 2013 due to the lack of observations for Nikkei 225 index, we have 1313 observations in total for weekly returns volatilities. The descriptive statistics, Ljung Box (LB) test statistics and Augmented Dickey Fuller (ADF) test statistics for both series are summarized in Table 3.6. From the ADF test results, it is clear that all time series are stationary for further analysis (we also performed the Johansen cointegration test pair-wisely on the price levels and no cointegration was found for the six market indexes).

We firstly provide a full-sample analysis of global stock market return and volatility spillovers over the period from January 1992 to March 2017, summarized in Tables 3.7 and 3.8. The two tables report the pairwise test statistics for conditional independence between index $X$ and index $Y$, given the constant $C$ in the bandwidth for kernel estimation is 4.8 or 8 and 999 resampling time series. In other words, we test for the absence of the one-week-ahead directional linkage from index $X$ to $Y$ by using the five resampling methods described in
Table 3.6: Descriptive statistics for global stock market return and volatility.

<table>
<thead>
<tr>
<th></th>
<th>DJIA</th>
<th>Nikkei</th>
<th>Hangseng</th>
<th>FTSE</th>
<th>DAX</th>
<th>CAC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>0.1433</td>
<td>-0.0126</td>
<td>0.1294</td>
<td>0.0823</td>
<td>0.1535</td>
<td>0.0790</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>0.2911</td>
<td>0.1472</td>
<td>0.2627</td>
<td>0.2121</td>
<td>0.4029</td>
<td>0.1984</td>
</tr>
<tr>
<td><strong>Std. Dev.</strong></td>
<td>2.2321</td>
<td>3.0521</td>
<td>3.3819</td>
<td>2.3367</td>
<td>3.0972</td>
<td>2.9376</td>
</tr>
<tr>
<td><strong>Skewness</strong></td>
<td>-0.8851</td>
<td>-0.6978</td>
<td>-0.3773</td>
<td>-0.8643</td>
<td>-0.6398</td>
<td>-0.6803</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>10.8430</td>
<td>8.9250</td>
<td>5.9522</td>
<td>13.2777</td>
<td>7.9186</td>
<td>8.0780</td>
</tr>
<tr>
<td><strong>LB Test</strong></td>
<td>49.9368**</td>
<td>15.4577</td>
<td>28.7922</td>
<td>61.0916**</td>
<td>28.0474</td>
<td>43.5004**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DJIA</th>
<th>Nikkei</th>
<th>Hangseng</th>
<th>FTSE</th>
<th>DAX</th>
<th>CAC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>4.5983</td>
<td>7.7167</td>
<td>9.6164</td>
<td>5.5306</td>
<td>8.5698</td>
<td>8.2629</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>2.2155</td>
<td>4.6208</td>
<td>4.5827</td>
<td>2.7161</td>
<td>4.0596</td>
<td>4.7122</td>
</tr>
<tr>
<td><strong>Maximum</strong></td>
<td>208.2227</td>
<td>265.9300</td>
<td>379.4385</td>
<td>149.1572</td>
<td>175.0968</td>
<td>179.8414</td>
</tr>
<tr>
<td><strong>Minimum</strong></td>
<td>0.0636</td>
<td>0.1882</td>
<td>0.1154</td>
<td>0.1154</td>
<td>0.1263</td>
<td>0.2904</td>
</tr>
<tr>
<td><strong>Skewness</strong></td>
<td>10.9980</td>
<td>9.6361</td>
<td>10.2868</td>
<td>6.7179</td>
<td>5.3602</td>
<td>6.0357</td>
</tr>
<tr>
<td><strong>Kurtosis</strong></td>
<td>180.0844</td>
<td>140.7606</td>
<td>152.4263</td>
<td>67.0128</td>
<td>42.0810</td>
<td>58.3263</td>
</tr>
<tr>
<td><strong>LB Test</strong></td>
<td>1924.0870**</td>
<td>933.3972**</td>
<td>1198.8722**</td>
<td>1970.7366**</td>
<td>2770.5973**</td>
<td>1982.4141**</td>
</tr>
</tbody>
</table>

Note: Descriptive statistics for six globally leading indexes. The sample size is 1313 for both Returns and Volatilities. The nominal returns are measured by weekly Friday-to-Friday log price difference multiplied by 100 and the Monday-to-Friday volatilities are calculated following Diebold and Yilmaz (2009). For the LB test and the ADF test statistics, the asterisks indicate the significance of the corresponding p-value at the 1% level of significance.
Section 3.2. For example, the first line in the top panel in Table 3.7 reports the one-week-ahead influence of DJIA returns upon other indexes by using the first time-shifted surrogates method (TS.a). Given $C = 8$, DJIA return is shown to be a strong Granger cause for Nikkei, FTSE and CAC at the 1% level, and for DAX at the 5% level.

Based on Tables 3.7 and 3.8, we may draw several conclusions. Firstly, the US index and German index are the most important return transmitters and Hong Kong is the largest source for volatility spillover, judged by the numbers of significant linkages. Note that this finding is similar as the result in Diebold and Yilmaz (2009), where the total return (volatility) spillovers from US (Hong Kong) to others are found to be much higher than from any other country. Fig. 3.6 provides a graphical illustration of the global spillover network based on the result of (SMB.a) from Tables 3.7 and 3.8. Apart from the main transmitters, we can clearly see that Nikkei and CAC are the main receivers in the global return spillover network, while DAX is the main receiver of global volatility transmission.

Secondly, the result obtained is very robust, no matter which re-sampling method is applied. Although the differences between the five resampling methods are small, (TS.a) is seen to be slightly more powerful than (TS.b) in Table 3.7. The three different bootstrap methods are very consistent almost all the time, similarly to what we observed in Section 3.3.

However, the summary results in Tables 3.7 and 3.8 are static in the sense that they do not take into account possible time-variation. The statistics are measurements for averaged-out directional linkages over the whole period from 1992 to 2017. The conditional dependence structure of the time series, at any point in time, can be very different. Hence, the full-sample
Table 3.7: Detection of Conditional Dependence in Global Stock Returns.

<table>
<thead>
<tr>
<th></th>
<th>DJIA</th>
<th>Nikkei</th>
<th>Hangseng</th>
<th>FTSE</th>
<th>DAX</th>
<th>CAC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C = 4.8$</td>
<td>$C = 8$</td>
<td>$C = 4.8$</td>
<td>$C = 8$</td>
<td>$C = 4.8$</td>
<td>$C = 8$</td>
</tr>
<tr>
<td><strong>TS.a</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJIA</td>
<td>0.367</td>
<td>0.002 **</td>
<td>0.972</td>
<td>0.273</td>
<td>0.609</td>
<td>0.009 **</td>
</tr>
<tr>
<td>Nikkei</td>
<td>0.231</td>
<td>0.081</td>
<td>-</td>
<td>0.997</td>
<td>0.951</td>
<td>0.883</td>
</tr>
<tr>
<td>Hangseng</td>
<td>0.898</td>
<td>0.197</td>
<td>0.963</td>
<td>0.407</td>
<td>-</td>
<td>0.701</td>
</tr>
<tr>
<td>FTSE</td>
<td>0.483</td>
<td>0.004 **</td>
<td>0.004 **</td>
<td>0.001 **</td>
<td>0.419</td>
<td>0.001 **</td>
</tr>
<tr>
<td>DAX</td>
<td>0.977</td>
<td>0.149</td>
<td>0.027 *</td>
<td>0.001 **</td>
<td>0.009 **</td>
<td>0.001 **</td>
</tr>
<tr>
<td>CAC</td>
<td>0.995</td>
<td>0.713</td>
<td>0.001 **</td>
<td>0.001 **</td>
<td>0.294</td>
<td>0.001 **</td>
</tr>
<tr>
<td><strong>TS.b</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJIA</td>
<td>0.402</td>
<td>0.004 **</td>
<td>0.967</td>
<td>0.341</td>
<td>0.595</td>
<td>0.020 *</td>
</tr>
<tr>
<td>Nikkei</td>
<td>0.219</td>
<td>0.110</td>
<td>-</td>
<td>0.999</td>
<td>0.956</td>
<td>0.853</td>
</tr>
<tr>
<td>Hangseng</td>
<td>0.899</td>
<td>0.269</td>
<td>0.975</td>
<td>0.485</td>
<td>-</td>
<td>0.696</td>
</tr>
<tr>
<td>FTSE</td>
<td>0.477</td>
<td>0.016 *</td>
<td>0.006 **</td>
<td>0.003 **</td>
<td>0.404</td>
<td>0.018 *</td>
</tr>
<tr>
<td>DAX</td>
<td>0.976</td>
<td>0.221</td>
<td>0.027 *</td>
<td>0.002 **</td>
<td>0.009 **</td>
<td>0.005 **</td>
</tr>
<tr>
<td>CAC</td>
<td>0.993</td>
<td>0.729</td>
<td>0.002 **</td>
<td>0.002 **</td>
<td>0.346</td>
<td>0.016 *</td>
</tr>
<tr>
<td><strong>SMB.a</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJIA</td>
<td>0.564</td>
<td>0.001 **</td>
<td>0.999</td>
<td>0.349</td>
<td>0.796</td>
<td>0.003 **</td>
</tr>
<tr>
<td>Nikkei</td>
<td>0.321</td>
<td>0.147</td>
<td>-</td>
<td>0.988</td>
<td>0.946</td>
<td>0.957</td>
</tr>
<tr>
<td>Hangseng</td>
<td>0.946</td>
<td>0.273</td>
<td>0.967</td>
<td>0.483</td>
<td>-</td>
<td>0.860</td>
</tr>
<tr>
<td>FTSE</td>
<td>0.579</td>
<td>0.005 **</td>
<td>0.021 *</td>
<td>0.001 **</td>
<td>0.701</td>
<td>0.003 **</td>
</tr>
<tr>
<td>DAX</td>
<td>0.988</td>
<td>0.240</td>
<td>0.044 *</td>
<td>0.001 **</td>
<td>0.012 *</td>
<td>0.001 **</td>
</tr>
<tr>
<td>CAC</td>
<td>0.993</td>
<td>0.762</td>
<td>0.006 **</td>
<td>0.001 **</td>
<td>0.594</td>
<td>0.001 **</td>
</tr>
<tr>
<td><strong>SMB.b</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJIA</td>
<td>0.583</td>
<td>0.002 **</td>
<td>0.994</td>
<td>0.334</td>
<td>0.797</td>
<td>0.001 **</td>
</tr>
<tr>
<td>Nikkei</td>
<td>0.351</td>
<td>0.155</td>
<td>-</td>
<td>0.992</td>
<td>0.940</td>
<td>0.952</td>
</tr>
<tr>
<td>Hangseng</td>
<td>0.945</td>
<td>0.276</td>
<td>0.970</td>
<td>0.506</td>
<td>-</td>
<td>0.866</td>
</tr>
<tr>
<td>FTSE</td>
<td>0.637</td>
<td>0.006 **</td>
<td>0.008 **</td>
<td>0.001 **</td>
<td>0.714</td>
<td>0.003 **</td>
</tr>
<tr>
<td>DAX</td>
<td>0.980</td>
<td>0.256</td>
<td>0.034 *</td>
<td>0.001 **</td>
<td>0.011 *</td>
<td>0.001 **</td>
</tr>
<tr>
<td>CAC</td>
<td>0.993</td>
<td>0.825</td>
<td>0.003 **</td>
<td>0.001 **</td>
<td>0.591</td>
<td>0.002 **</td>
</tr>
<tr>
<td><strong>STB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJIA</td>
<td>0.645</td>
<td>0.001 **</td>
<td>0.996</td>
<td>0.334</td>
<td>0.787</td>
<td>0.002 **</td>
</tr>
<tr>
<td>Nikkei</td>
<td>0.363</td>
<td>0.114</td>
<td>-</td>
<td>0.989</td>
<td>0.944</td>
<td>0.946</td>
</tr>
<tr>
<td>Hangseng</td>
<td>0.964</td>
<td>0.272</td>
<td>0.984</td>
<td>0.491</td>
<td>-</td>
<td>0.859</td>
</tr>
<tr>
<td>FTSE</td>
<td>0.652</td>
<td>0.005 **</td>
<td>0.016 *</td>
<td>0.001 **</td>
<td>0.688</td>
<td>0.003 **</td>
</tr>
<tr>
<td>DAX</td>
<td>0.982</td>
<td>0.234</td>
<td>0.048 *</td>
<td>0.001 **</td>
<td>0.017 *</td>
<td>0.001 **</td>
</tr>
<tr>
<td>CAC</td>
<td>0.996</td>
<td>0.814</td>
<td>0.007 **</td>
<td>0.001 **</td>
<td>0.578</td>
<td>0.001 **</td>
</tr>
</tbody>
</table>

Note: Statistics for pairwise T-based test on returns of global stock indexes for one-week ahead conditional non-independence. The results are shown both for the five different resampling methods in Section 3.2.3. The constant $C$ takes value 4.8 and 8 for robustness-check. The asterisks indicate the significance of the corresponding $p$-value, at the 5% (*) and 1% (**) levels.
### Table 3.8: Detection of Conditional Dependence in Global Stock Volatilities

<table>
<thead>
<tr>
<th></th>
<th>CAC</th>
<th>DAX</th>
<th>FTSE</th>
<th>TSF.a</th>
<th>RMB</th>
<th>HKSE</th>
<th>DJIA</th>
<th>CCA</th>
<th>TS.a</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>CIPE</td>
<td>0.00</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Note: Statistics for pairwise TE-based test on volatilities of global stock-indexes for one-week ahead conditional non-independence. The results are shown both for the five different resampling methods in Section 3.2.3. The constant levels. The columns indicate the significance of the corresponding $p$-value. (*) Significant at the 5%. (** Significant at the 1% level.
3.5 Conclusions

This chapter provides guidelines for the practical application of TE in detecting conditional dependence, i.e., Granger causality in a more general sense, between two time series. Although there already is a tremendous literature that tried to apply the TE in this context, the asymptotics of the statistic and the performance of the resampling-based measures are still not understood well. We have considered tests based on five different resampling methods,
CHAPTER 3. SIMULATION-BASED TEST FOR GRANGER CAUSALITY

Figure 3.7: Time-varying $p$-values for the TE-based Granger causality test in Return series. The causal linkages from DJIA to other markets, as well as the linkages from other markets to DJIA are tested.
3.5. Conclusions

all of which were shown in the literature to be suitable for entropy-related tests, and investigated the size and power of the associated tests numerically. Two time-shifted surrogates and three smoothed bootstrap methods are tested on simulated data from several processes. The simulation results in this controlled environment suggest that all five measures achieve reasonable rejection rates under the null as well as the alternative hypotheses. Our results are very robust with respect to the density estimation method, including the procedure used for standardizing the location and scale of the data and the choice of the bandwidth parameter, as long as the convergence rate of the kernel estimator of TE is consistent with its first order Taylor expansion.

In the empirical application, we have shown how the proposed resampling techniques can be used on real world data for detecting conditional dependence in the data set. We use global equity data to carry out the detection in pairwise causalities in the return and volatility series among the world leading stock indexes. Our work can be viewed as a nonparametric extension of the spillover measures considered by Diebold and Yilmaz (2009). In accordance with them, we found evidence that the DJIA and the DAX are the most important return transmitters and Hong Kong is the largest source for volatility spillover. Furthermore, the rolling window-based test for Granger causality in pairwise return series demonstrated that the causal linkages in the global equity market are time-varying rather than static. The overall dependence is more tight during the most recent financial crisis, and the fluctuations of the p-values are shown to be event dependent.

As for future work, there are several directions for potential extensions. On the theoretical side, it would be practically meaningful to consider causal linkage detection beyond the single period lag and to deal with the infinite order issue in a nonparametric setting. Further nonparametric techniques need to be developed to play a similar role as the information criterion does for order selection of an estimation model in the parametric world. On the empirical side, it will be interesting to further exploit entropy-based statistics in testing conditional dependence when there exists a so-called common factor, i.e., looking at multivariate systems with more than two variables. One potential candidate for this type of test in the partial TE has been coined by Vakorin et al. (2009), but its statistical properties still need to be thoroughly studied yet.
Chapter 4

A Consistent Test for Granger Causality based on a Nonparametric Information-based Statistic

4.1 Introduction

Characterizing causal interactions between time series was a challenging issue until Granger (1969) in his pioneering work brought forward the concept later known as Granger causality. Since then, testing causal effects has attracted massive attention not only in Economics and Econometrics research, but also in the domains of neuroscience (Bressler and Seth, 2011; Ding et al., 2006), biology (Guo et al., 2010) and physics (Barrett et al., 2010), among others.

The vector autoregressive (VAR) modeling-based test has become a popular methodology over the last decades, with repeated debates on its validity. As we see it, there are at least two critical problems with parametric causality tests. First, being based on a classical linear VAR model, traditional Granger causality tests may overlook a significant nonlinear relationship between two processes. As Granger (1989) puts it, nonlinear models represent the proper way to model the real world which is ‘almost certainly nonlinear’. Secondly, parametric causality testing approaches bear the risk of model mis-specification. The conclusion drawn from a wrong regression model could be either misleading or lacking power. For example, Baek and Brock (1992) construct an example where nonlinear causal relations cannot be detected by a traditional linear causality test.

A series of studies tried to relax parametric model assumptions and provide many non-parametric versions of Granger causality tests. Among all nonparametric approaches, the Hiemstra and Jones (1994) test is one of the most far-reaching tests. By modifying the Baek and Brock (1992) nonparametric methodology, Hiemstra and Jones firstly developed a nonparametric test to detect nonlinear causality in weakly dependent stochastic data. How-
ever, the Hiemstra-Jones test is suffering from an over-rejection problem, leading Diks and
Panchenko (2006) to propose a new test (hereafter referred to as the DP test). Other alter-
native tests include additive models used by Bell et al. (1996), the Hellinger distance measure
used by Su and White (2008), and the empirical likelihood ratio-based test used by Su and
White (2014).

The scope of this chapter is to provide a novel test for Granger causality, based on the
information theoretical notion transfer entropy (hereafter TE), which was coined by Schreiber
(2000). Transfer entropy was initially used to measure asymmetric information exchange in
a bivariate system. By using appropriate conditional densities, transfer entropy is able to
distinguish information that is actually transferred from shared information due to common
history. This property makes it attractive for detecting conditional dependence, i.e. Granger
causality. We refer to Hlaváčková-Schindler et al. (2007), Amblard and Michel (2012) for
detailed reviews of the relation between Granger causality and directed information theory.

Although attractive, the application of concepts from information theory to time series
analysis has proved difficult, due to the lack of asymptotic theory. For example, Granger and
Lin (1994) utilize entropy to detect serial dependence using critical values obtained by sim-
ulation. Hong and White (2005) prove asymptotic normality for an entropy-based statistic,
but the asymptotics only hold for a specific kernel function. Barnett and Bossomaier (2012)
establish an asymptotic $\chi^2$ distribution for transfer entropy estimator under finite Markov
chain assumption. Establishing asymptotic distribution theory for fully nonparametric trans-
fer entropy measure is challenging, if not impossible.

In this research, we propose a test statistic based on a first order Taylor expansion of
the transfer entropy, which follows a Normal distribution asymptotically. Instead of deriving
the limiting distribution of the transfer entropy itself, we bypass the problem by testing an
implication of the null hypothesis. Further we show that this new test statistic is closely
related to the DP test, but with the appealing property of non-negative definiteness.

This chapter is organized as follows. Section 4.2 provides a short introduction to nonpara-
metric tests for Granger non-causality and the DP test, followed by an example to show that
the DP test has no nontrivial power against some alternatives because the test is not based
on a positive definite quantity. Subsequently, information theory, and the testing framework
based on series expansion on transfer entropy statistic is introduced. The close linkage of
this novel test statistic with the DP test is shown, and asymptotical normality is proved by
using a third order $U$-statistic representation. Section 4.2 also discusses the optimal band-
width selection rule for specific cases. Section 4.3 deals with Monte Carlo simulations; three
different data generating processes are considered, enabling a direct comparison of size and
power between the modified DP test and the DP test. Section 4.4 considers two financial
applications. In the first, we apply our test on the stock volume and return data to make a
direct comparison with the DP test; in the second application high frequency exchange rates
of main currencies are tested. Finally, Section 4.5 summarizes.
4.2 A Transfer Entropy-Based Test Statistic for Granger non-Causality

4.2.1 Nonparametric Granger non-Causality Tests

This subsection provides some basic concepts and definitions for Granger causality, and the idea of nonparametrically testing for conditional independence. We restrict ourselves to the bivariate setting as it is the most-accepted implementation, although generalization to multivariate densities is possible.

Intuitively, in a bivariate system \( \{X_t, Y_t\} \), given two strictly stationary processes \( \{X_t\} \) and \( \{Y_t\} \), it is said that \( \{X_t\} \) Granger causes \( \{Y_t\} \) if current and past values of \( \{X_t\} \) contain some additional information beyond current and past values of \( \{Y_t\} \) about future values of \( \{Y_t\} \). Linear Granger causality tests based on a parametric VAR model can be seen as special cases for testing the difference between two conditional distributions, i.e. the difference in conditional means.

In a more general setting, the null hypothesis of Granger causality can be rephrased in terms of conditional dependence between two series: \( \{X_t\} \) is a Granger cause of \( \{Y_t\} \) if the distribution of \( \{Y_t\} \) conditional on its own history is not the same as that conditional on the histories of both \( \{X_t\} \) and \( \{Y_t\} \). If we denote the information set of \( \{X_t\} \) and \( \{Y_t\} \) until time \( t-1 \) by \( \mathcal{F}_{Y,t-1} \) and \( \mathcal{F}_{X,t-1} \), respectively, and use ‘\( \sim \)’ to denote equivalence in distribution, we may give a formal and general definition for Granger causality. For a strictly stationary bivariate process \( \{X_t, Y_t\}, t \in \mathbb{Z} \), \( \{X_t\} \) is a Granger cause of \( \{Y_t\} \) if, given lags \( l_X, l_Y \) and \( k \),

\[
(Y_t, \ldots, Y_{t+k}) \mid (\mathcal{F}_{Y,t-l_Y}, \mathcal{F}_{X,t-l_X}) \sim (Y_t, \ldots, Y_{t+k}) \mid \mathcal{F}_{Y,t-l_Y}.
\]

In the absence of Granger causality, i.e.

\[
(Y_t, \ldots, Y_{t+k}) \mid (\mathcal{F}_{Y,t-l_Y}, \mathcal{F}_{X,t-l_X}) \sim (Y_t, \ldots, Y_{t+k}) \mid \mathcal{F}_{Y,t-l_Y},
\]

\( \{X_t\} \) has no influence on the distribution of future \( \{Y_t\} \). This is also referred to as Granger non-causality and often expressed as conditional independence between \( \{X_t\} \) and \( \{Y_t\} \) as

\[
(4.1) \quad (Y_t, \ldots, Y_{t+k}) \perp (X_{t-1}, \ldots, X_{t-l_X}) \mid \mathcal{F}_{Y,t-l_Y}.
\]

Granger non-causality, as expressed in Eq. (4.1), lays the first stone for a nonparametric test without imposing any parametric assumptions about the data generating process or underlying distributions for \( \{X_t\} \) and \( \{Y_t\} \). The orthogonality here concerns not only the conditional mean, but also higher conditional moments. We only assume two things here. First, \( \{(X_t, Y_t)\} \) is a strictly stationary bivariate process. Second, the process has finite memory, i.e. \( l_X, l_Y \ll \infty \). The second assumption is needed in a nonparametric setting to
allow conditioning on the past, as a finite Markov property.

The null hypothesis of Granger non-causality test is that \( H_0 : \{X_t\} \text{ is not a Granger cause of } \{Y_t\} \). For simplicity, we limit ourselves to single lags in the past such that \( l_X = l_Y = 1 \) and \( k = 0 \), which is the case of most practical interest. Further we define a three-variate vector \( W_t \) as \( W_t = (X_t, Y_t, Z_t) \), where \( Z_t = Y_{t+1} \) and \( W = (X, Y, Z) \) is used when there is no danger of confusion. Within the bivariate setting, \( W \) is a three-dimensional continuous vector. By using density functions \( f(.) \) and given \( l_X = l_Y = 1 \) and \( k = 0 \), Eq. (4.1) can be phrased as

\[
(4.2) \quad H_0 : \frac{f_{X,Y,Z}(x,y,z)}{f_Y(y)} = \frac{f_{Y,Z}(y,z)}{f_Y(y)} \frac{f_{X,Y}(x,y)}{f_Y(y)},
\]

for \((x,y,z)\) in the support of \( W \), or equivalently as

\[
(4.3) \quad H_0 : \frac{f_{X,Y,Z}(x,y,z)}{f_Y(y)} - \frac{f_{Y,Z}(y,z)}{f_Y(y)} \frac{f_{X,Y}(x,y)}{f_Y(y)} = 0,
\]

for \((x,y,z)\) in the support of \( W \). A nonparametric test for Granger non-causality seeks to find statistical evidence of violation of Eq. (4.2) or Eq. (4.3). There are many nonparametric measures available for this purpose, some of which are mentioned above. However, as far as we know, the DP test, to be described below, is the only fully nonparametric test that is known to have correct asymptotic size under the null hypothesis of no Granger causality.

### 4.2.2 The DP Test

Hiemstra and Jones (1994) first propose to test the condition given by Eq. (4.2) by calculating correlation integrals for each density and measuring the discrepancy between two sides of the equation. However their test has been shown to suffer from severe size distortion due to a simple fact that measuring each density separately may not deliver the same quantity implied by Eq. (4.2). To overcome this problem, Diks and Panchenko (2006) suggest to use a conditional dependence measure by incorporating a local weighting function \( g(x,y,z) \) and formulate Eq. (4.3) as

\[
(4.4) \quad E \left[ \left( \frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \frac{f_{Y,Z}(Y,Z)}{f_Y(Y)} \frac{f_{X,Y}(X,Y)}{f_Y(Y)} \right) g(X,Y,Z) \right] = 0.
\]

Under the null hypothesis of no Granger causality, the term within the large round brackets vanishes, and the expectation goes to zero. Eq. (4.4) can be treated as infinite number of moment restriction, as noted by Diks and Wolski (2015). Although testing for Eq. (4.4) instead of Eq. (4.2) or Eq. (4.3) may lead to a loss of power against some specific alternatives, there is also an advantage to do so. For example, the weighting function \( g(x,y,z) \) is assumed to be \( g(x,y,z) = f^2_Y(y) \) in the DP test as it delivers a U-statistic representation of the corresponding
estimator, which enables the analytical derivation of the asymptotic distribution of the test statistic. In principle, other choices for \( g(x, y, z) \) will also do as long as the test has satisfactory power against alternatives of interest. By plugging in \( g(x, y, z) = f_Y^2(y) \), the DP test tests on the implication of \( H_0 \),

\[
H'_0 : \quad q \equiv E [f_{X,Y,Z}(X,Y,Z)f_Y(Y) - f_{X,Y}(X,Y)f_{Y,Z}(Y,Z)] = 0. 
\]

Given a local density estimator of a \( d_W \)-variate random vector \( W \) at \( W_i \) as

\[
\hat{f}_W(W_i) = ((n-1)h)^{-d_W} \sum_{j \neq i} \kappa \left( \frac{W_i - W_j}{h} \right),
\]

where \( \kappa \) is a kernel function and \( h \) is the bandwidth, the DP test develops a third order \( U \)-statistic estimator for the functional \( q \), given by

\[
T_n(h) = \frac{(n-1)}{n(n-2)} \sum_i \left( \hat{f}_{X,Y,Z}(X_i,Y_i,Z_i)\hat{f}_Y(Y_i) - \hat{f}_{X,Y}(X_i,Y_i)\hat{f}_{Y,Z}(Y_i,Z_i) \right),
\]

where the normalization factor \( (n-1)/(n(n-2)) \) is inherited from the \( U \)-statistic representation of \( T_n(h) \). It is worth mentioning that a second order square kernel \( \kappa \) is adopted by Diks and Panchenko (2006). However there are two main drawbacks of using a square kernel. First, a square kernel will yield a discontinuous density estimate \( \hat{f}(\cdot) \), which is not attractive from a practical perspective. Second, it weighs all neighbor points \( W_j \) equally, overlooking their relative distance to the estimation point \( W_i \). Therefore, a smooth kernel function — the Gaussian kernel — is applied here, namely, \( \kappa(\cdot) \), the product kernel function defined as \( \kappa(W) = \prod_{s=1}^{d_W} \kappa(w^s) \), where \( w^s \) is \( s \)-th element in \( W \). Using a standard univariate Gaussian kernel, \( \kappa(w^s) = (2\pi)^{-1/2} e^{-\frac{1}{2}(w^s)^2} \), \( \kappa(\cdot) \) is the standard multivariate Gaussian kernel as described in Wand and Jones (1994) and Silverman (1986).

For \( l_X = l_Y = 1 \), Diks and Panchenko (2006) prove the asymptotic normality of \( T_n(h) \). Namely, if the bandwidth \( h \) depends on the sample size in such way that \( h = Cn^{-\beta} \) for some \( C > 0 \) and \( \beta \in \left( \frac{1}{3}, \frac{1}{2} \right) \), the test statistic in Eq. (4.7) satisfies

\[
\sqrt{n} \frac{T_n(h) - q}{S_n} \overset{d}{\to} N(0,1),
\]

where \( \overset{d}{\to} \) denotes convergence in probability and \( S_n^2 \) is a consistent estimator of the asymptotic variance of \( T_n(h) \). Diks and Panchenko (2006) suggest to implement an one-sided version of the test, rejecting \( H'_0 \) against the alternative \( H_a \): \( q > 0 \) if \( T_n(h) \) is too large. In other words, given the asymptotic critical value \( z_{1-\alpha} \), the null hypothesis \( H'_0 \) is rejected at significance level \( \alpha \) if \( \sqrt{n}T_n(h)/S_n > z_{1-\alpha} \).

The drawback of the DP test arises from the fact that \( H'_0 \) in Eq. (4.5) need not be equivalent to \( H_0 \) in Eqs. (4.2) and (4.3). If \( H'_0 \) is equivalent to \( H_0 \), then \( q \) forms a suitable
CHAPTER 4. A NONPARAMETRIC TEST FOR GRANGER CAUSALITY

basis for the DP test for Granger non-causality. However, this equivalence depends on the positive definiteness stated in the following property, which $q$ does not satisfy.

Property 4. $q$ is non-negative in such a way that $q \geq 0$ with equality if and only if $X_t$ and $Z_t$ are conditionally independent given $Y_t$.

From previous reasoning, it is obvious that Eq. (4.5) is implied by Eq. (4.3), and Prop. 4 states that a strictly positive $q$ is achieved if and only if $H_0$ is violated. In other words, the null hypothesis of Granger non-causality requires that $X_t$ and $Z_t$ are independent conditionally on $Y_t$, which is just a sufficient, but not necessary, condition for $q = 0$. With Prop. 4, $H'_0$ coincides with $H_0$ and a consistent estimator of $q$, i.e. $T_n(h)$ as suggested by DP, will have unit asymptotic power. If this property is not satisfied, a test on $q = 0$ could deviate from the test on $H_0$. Although Diks and Wolski (2015) identified specific classes of processes for which the property holds, we could easily construct a counterexample to show that the DP test has no power even in cases where $X_t$ strongly Granger causes $Z_t$.

Inspired by the example in Skaug and Tjostheim (1993), where a closely related test for unconditional independence is performed, we consider a conditional counterpart to illustrate that $q$ is not positive definite. The one-sided DP test will be seen to suffer from a lack of power for this example process. As we show below, in the extreme case when $q = 0$, this drawback cannot be overcome even with a two sided DP test.

Consider the process $\{(X_t, Y_t, Z_t)\}$ where $Z_t \equiv Y_{t+1}$ as before. We assume that the i.i.d continuous variable $X_t \in [-1, 1]$, with probability $1 - d$ of being positive, where $0 < d < 1$. Further, there is no dependence between $X_t$ and $Y_t$, and $Z_t$ does not depend on $Y_t$ but on $X_t$ in such a way that the joint density of $(X_t, Z_t)$ conditional on $Y_t$ is given by

\[
\begin{align*}
  f(x_t, z_t|y_t) = f(x_t, z_t) = \begin{cases} 
    1 - 2d, & \text{if } 0 \leq x_t \leq 1, \quad 0 \leq z_t \leq 1, \\
    d, & \text{if } 0 \leq x_t \leq 1, \quad -1 \leq z_t < 0, \\
    d, & \text{if } -1 \leq x_t < 0, \quad 0 \leq z_t \leq 1, \\
    0, & \text{if } -1 \leq x_t < 0, \quad -1 \leq z_t < 0.
  \end{cases}
\end{align*}
\]

Given Eq. (4.9), the probability of $Z_t$ being larger than zero conditional on $\{X_t, Y_t\}$ can be calculated as $P(0 < z_t \leq 1|x_t, z_t) = 1$ if $-1 \leq x_t < 0$, and $P(0 \leq z_t \leq 1|x_t, z_t) = (1 - 2d)/(1 - d)$ if $0 \leq x_t \leq 1$. In this setting, as long as $d \neq 0$, $X_t$ is a Granger cause of $Y_t$ since it has an impact on the distribution of future values of $Y_t$. Then it is not difficult to analytically calculate $q$ defined in Eq. (4.5). Particularly, one finds that $q = d^2 ((1 - d)^3 + d^3) (4d - 1).$ For $0 < d < \frac{1}{4}$, $q$ has a negative value. In this situation, the one-sided DP test, which rejects for large $q$, is not a consistent test for Granger non-causality. One may argue that this is not a problem if we use a two-sided test at the price of losing some power. However, the inconsistency of the DP test - which tests on $H'_0$ rather than $H_0$ - is illustrated by considering $q = 0$, which is the case when $d = \frac{1}{4}$. In that case $X_t$ is clearly a Granger cause of $Y_t$, however
4.2. A Transfer Entropy-Based Test Statistic

Figure 4.1: Power for the one-sided DP test for the artificial process \((X_t, Y_t, Z_t)\) given \(q = 0\) at nominal size, from the bottom to the top, 5%, 10% and 15%, respectively, based on 10,000 independent simulations.

The DP test fails to detect this specific conditional independence.

Fig. 4.1 reports the power of the one-sided DP test as a function of the sample size for different significance levels, based on 10,000 independent simulations. Three nominal sizes are illustrated here: 5%, 10% and 15%, and the sample size ranges from 100 to 20,000. It is striking from Fig. 4.1 that the DP does not have power at all against the alternative for this process. The same conclusion can be drawn based on Fig. 4.2, where the power-size plots are given. For almost all sub-panels with different sample sizes, the power of the DP test is around the diagonal line for this particular example when \(q = 0\), which indicates that DP test has only trivial power to detect Granger causality from \(X_t\) to \(Y_t\).

The lack of power of the one-sided DP test in this example is hardly to be alleviated by its two-sided counterpart, as a result of the absence of equivalence between \(q = 0\) and conditional independence. The difference between \(H_0\) and its implication \(H'_0\) gives rise to the lack of power of the DP test as the positive definiteness in Prop. 4 is violated. In the next subsection, a new test statistic, based on the information theoretical concept transfer entropy, is introduced and the test statistic is shown to be positive definite, which in return
Figure 4.2: Size-power plots for the one-sided DP test for the artificial process \{(X_t, Y_t, Z_t)\} with \(q = 0\), based on 10,000 independent simulations. Each subplot draws the actual power against the nominal size for different sample sizes, ranging from 100 to 10,000. The solid curve represents the actual power and the red dash line indicates the diagonal, indicating the nominal size of a test.
4.2. A Transfer Entropy-Based Test Statistic

overcomes the inherited drawback of the DP test. In fact, this new test statistic shares many similarities with the DP test statistic, but allows for an information theoretical interpretation for its non-negativity.

4.2.3 Information-theoretical Interpretation

In a very different context from testing conditional independence, the problem of information feedback and impact also has drawn many attentions since 1950. Information theory, as a branch of applied mathematical theory of probability and statistics, studies the transmission of information over a noisy channel. Entropy, also referred to as Shannon entropy, is a key measure in the field of information theory brought forward by Shannon (1948, 1951). Entropy measures the uncertainty and randomness associated with a random variable. Suppose that $S$ is a random vector with density $f_S(s)$, then the Shannon entropy is defined as

$$H(S) = -\int f_S(s) \log\{f_S(s)\} ds.$$ 

There is a long history of applying information measures in econometrics. For example, Robinson (1991) uses the Kullback-Leibler information criterion (Kullback and Leibler (1951), KLIC) to construct a one-sided test for serial independence. Since then, nonparametric tests using entropy-based measures for independence between two time series are becoming prevalent. Granger and Lin (1994) use entropy measure to identify the lags in a nonlinear bivariate model. Granger et al. (2004) study dependence with a transformed metric entropy, which has the additional advantage of allowing multiple comparisons of distances and turns out to be a proper measure of distance. Hong and White (2005) provide a new entropy-based test for serial dependence, and show that the test statistic follows a standard normal distribution asymptotically.

Although inspiring, those results cannot be applied directly to measure conditional dependence, i.e. Granger causality. On the other hand, transfer entropy (TE) named by Schreiber (2000), which appeared in literatures earlier under different names, is a suitable measure to serve this purpose. The TE quantifies the amount of information explained in one series $k$ steps ahead from the state of another series, given the current state of itself. We briefly introduce the TE and KLIC before we further discuss its relation with the modified DP test.

Suppose that we have a bivariate process $\{(X_t, Y_t)\}$, and for brevity we put $X = \{X_t\}$, $Y = \{Y_t\}$ and $Z = \{Y_{t+k}\}$. Again we limit lag period $k = 1$ for simplicity, and define the three-dimensional vector $W = (X, Y, Z)$ as before. The transfer entropy $TE_{X \rightarrow Y}$ is a nonlinear and nonparametric measure for the amount of information contained in $X$ about $Z$, in addition to the information about $Z$ that already contained in $Y$. Although the TE defined by Schreiber (2000) applies to discrete variables, it is easily generalized to continuous
variables. Conditional on \( Y \), \( \text{TE}_{X \rightarrow Y} \) is defined as

\[
\text{TE}_{X \rightarrow Y} = E_W \left( \log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} \right)
\]

\[= \int \int \int f_{X,Y,Z}(X,Y,Z) \log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} \, dx \, dy \, dz
\]

\[= E_W \left( \log \frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \log \frac{f_{X,Y}(X,Y)}{f_Y(Y)} - \log \frac{f_{Y,Z}(Y,Z)}{f_Y(Y)} \right)
\]

\[= E_W \left( \log f_{X,Y,Z}(X,Y,Z) + \log f_Y(Y) - \log f_{X,Y}(X,Y) - \log f_{Y,Z}(Y,Z) \right).
\]

Using the conditional mutual information \( I(Z,X|Y = y) \), the TE can be equivalently formulated in terms of four Shannon entropies as

\[
\text{TE}_{X \rightarrow Y} = I(Z,X|Y) = H(Z|Y) - H(Z|X,Y)
\]

\[= H(Z,Y) - H(Y) - H(Z,X,Y) + H(X,Y).
\]

In order to construct a test for Granger causality based on the TE, one first needs to show quantitatively that the TE is a proper basis for testing the null hypothesis. Namely, it has to be proved that there exists an analogue property for TE such as Prop. 4 that the \( q \) in Eq. (4.5) of the DP test fails to satisfy. The following theorem, as a direct application of KLIC, lays the quantitative foundation for testing for TE.

**Theorem 4.** \( \text{TE}_{X \rightarrow Y} \geq 0 \) with equality if and only if \( f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y) \).

**Proof.** The proof to Thm. 4 follows from generalizing Theorem 3.1 in Chapter 2 of Kullback (1968), where the divergence between two different densities has been considered. An alternative proof is given in Appendix 4.6.1 by using Jensen’s inequality and concavity of the log function.

It is not difficult to verify that the condition for \( \text{TE}_{X \rightarrow Y} = 0 \) coincides with Eqs. (4.2) and (4.3) for Granger non-causality under the null hypothesis. This positive definiteness makes \( \text{TE}_{X \rightarrow Y} \) a desirable measure for constructing a one-sided test of Granger causality; any divergence from zero is a sign of conditional dependence of \( Y \) on \( X \). To estimate TE, one may follow the recipe of Kraskov et al. (2004) by measuring \( k \)-nearest neighbor distances. A more natural method, as we applied in this chapter, is to use the plug-in kernel estimates given in Eq. (4.6), and replace the expectation by the sample average.

However, the direct use of the TE to test Granger non-causality is not easy due to the lack of asymptotic theory for the test statistic. As shown by Granger and Lin (1994), the asymptotic distribution of entropy based estimators usually depends on strict assumptions on the dataset. Although over the years several break-throughs have been made with the
4.2. A Transfer Entropy-Based Test Statistic

application of entropy to testing serial independence, e.g. Robinson (1991) obtains an asymptotic $N(0,1)$ distribution for an entropy measure by a sample-splitting technique and Hong and White (2005) derive an asymptotic normal distribution under bounded support data and quartic kernel assumptions, the limiting distribution of TE statistic is still in myth.

One may argue in favor of using simulation techniques to overcome the problem of lacking asymptotic distributions. However, as suggested by Su and White (2008), there exist estimation biases of TE statistics for non-parametric dependence measures under the smoothed bootstrap procedure. Even with parametric test statistics, Barnett and Bossomaier (2012) notice that the TE based estimator is generally biased. Surrogate data are also applied widely by, among others, Wibral et al. (2013) and Papana et al. (2016) to detect information transfer. Direct usage of the TE for non-parametric Granger non-causality test is considered difficult, if not impossible.

In this chapter we show that a first order Taylor expansion of the TE provides a way out to construct the asymptotic distribution of this meaningful information measure. In the next section, we will show that the first order Taylor expansion of the TE can form the basis of a modified DP test for conditional dependence. This not only helps to circumvent the problem of asymptotic distribution for entropy based statistic, but also endows the modified DP test with positive definiteness, which is missing in DP test. We simply kill two birds with one stone.

In the remaining part of this section we will introduce the first order Taylor expansion of the TE, and the positive definiteness of the measure will be given afterwards. Starting with Eq. (4.10), we perform the first order Taylor expansion locally at $TE_{X \rightarrow Y} = 0$, which is

$$TE_{X \rightarrow Y} = E_W \left[ \log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} \right]$$

(4.11)

$$= E_W \left[ \log \left( 1 + \left( \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} - 1 \right) \right) \right]$$

$$= E_W \left[ \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} - 1 \right] + \text{h.o.t.},$$

where ‘h.o.t’ stands for ‘higher order terms’. By ignoring higher order terms, we define the first order expansion $t = E_W \left[ \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} - 1 \right]$ as a measure for conditional dependence. The following theorem states that $t$ inherits the positive definiteness of the TE.

**Theorem 5.** $t \geq 0$ with equality if and only if $f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)$.

**Proof.** See Appendix 4.6.2

Thm. 5 indicates that the measure $t$ has the desirable property of positive definiteness, which is absent for the measure $q$ of the DP test. However, direct estimation of Eq. (4.11) does not lead to a practically useful test statistic without the asymptotic distribution. In the next subsection, by incorporating a different weight function $v(x,y,z)$ into Eq. (4.5), a modified
CHAPTER 4. A NONPARAMETRIC TEST FOR GRANGER CAUSALITY

DP measure actually reproduces the first order Taylor expansion of transfer entropy measure, which provides a theoretical interpretation for the directional information interaction, i.e. Granger causality between \( \{X_t\} \) and \( \{Y_t\} \).

4.2.4 A Modified DP Test

In comparing Eqs. (4.3) and (4.4), it can be seen that the discrepancy between \( H'_0 \) and \( H_0 \) arises from incorporation of the weight function \( g(x, y, z) = f^2_X(y) \) to the null hypothesis. In principle, a different positive function \( g(x, y, z) \) will also work, such as those discussed by Diks and Panchenko (2006). As long as the weight function allows for a U-statistic representation of the corresponding estimator, the asymptotic normality is maintained. Particularly, we propose to modify the DP test by dividing all terms in the expectation of Eq. (4.5) by a function \( v(X, Y, Z) \) given by

\[
\frac{f_{X,Y,Z}(X, Y, Z)f_Y(Y) - f_{X,Y}(X, Y)f_{Y,Z}(Y, Z)}{v(X, Y, Z)}
\]

Similar to the discussion in Section 4.2.2, we test the implication of \( H_0 \) rather than the null, but will choose \( v(x, y, z) \) so that the \( H''_0 \) and \( H_0 \) coincide. This new implication of \( H_0 \) has the form

\[
H''_0 : \quad E \left[ \left( f_{X,Y,Z}(X, Y, Z)f_Y(Y) - f_{X,Y}(X, Y)f_{Y,Z}(Y, Z) \right) \frac{1}{v(X, Y, Z)} \right] = 0.
\]

One can also think of Eq. (4.12) as the result of plugging in a different weighting function in Eq. (4.4). By defining \( g(x, y, z) = f^2_X(y)/(f_{X,Y}(x, y))(f_{Y,Z}(y, z)) \) instead of \( g(x, y, z) = f^2_Y(y) \) suggested by DP, Eq. (4.12) simplifies to

\[
H'_0 : \quad t \equiv E \left[ \frac{f_{X,Y,Z}(X, Y, Z)f_Y(Y)}{f_{X,Y}(X, Y)f_{Y,Z}(Y, Z)} - 1 \right] = 0,
\]

which is equivalent to the first order Taylor expansion in Eq. (4.11) and hence to \( H_0 \) by Thm. 5. To estimate this \( t \), we propose to use the following statistic with density estimator defined in Eq. (4.6):

\[
T'_n(h) = \frac{(n - 1)}{n(n - 2)} \sum_i \left[ \left( \hat{f}_{X,Y,Z}(X_i, Y_i, Z_i)\hat{f}_Y(Y_i) - \hat{f}_{X,Y}(X_i, Y_i)\hat{f}_{Y,Z}(Y_i, Z_i) \right) \right] \frac{1}{\hat{v}(X_i, Y_i, Z_i)}.
\]

The reason for estimating \( t \) in this form is that, with the sample statistic \( T'_n(h) \), we can obtain an order three U-statistic representation of \( t \), similar to that one for the DP test statistic, from where the asymptotic normality can be derived. Before introducing the formal theorem, Thm. 6 shows that the modified DP statistic \( T'_n(h) \) generically is non-degenerate thus may be used for performing a statistical test based on the usual first-order asymptotics. In other words, although the statistic \( T'_n(h) \) may be degenerate in some special cases, these are shown to be non-generic.
4.2. A Transfer Entropy-Based Test Statistic

**Theorem 6.** The limiting distribution of $T'_n(h)$ generically is non-degenerate.

**Proof.** See Appendix 4.6.3. \qed

The asymptotic normality of $T'_n(h)$ is given by Thm. 7, which relies on the following two lemmas concerning the consistency of density estimators.

**Lemma 1.** Let $\{W_i = (X_i, Y_i, Z_i)\}, i \in \mathbb{N}$ be a sequence of k-dimensional random variables with Lebesgue density $f$. For the estimation of $f$, we use the kernel estimator $\hat{f}$ with kernel function $K(w)$ given in Eq. (4.6). If $f(w)$ is continuous at $w \in \mathbb{R}^k$ and $K(w)$ is of bounded variation, then

$$\sup_{w \in \mathbb{R}^k} |\hat{f}(w) - f(w)| \rightarrow 0 \quad \text{a.s.},$$

provided that any of the following two conditions holds:

(A1) $W_i$ is an independent sequence and either

$$\sum_{i=1}^{\infty} e^{-\gamma i h^{2k}} < \infty, \quad \text{for all } \gamma \in \mathbb{R}_+,$$

or

$$\left(\frac{\log \log i}{i}\right)^{1/2k} = o(h).$$

(A2) $W_i$ is $\phi$-mixing, $A_1(\phi) < \infty$ (for definition of $A_1(\phi)$ see Sen et al. (1974) (2.1)) and

$$\sum_{i=1}^{\infty} \left(\frac{\gamma}{h k i^{1/2}}\right)^{2(l+1)} < \infty, \quad \text{for all } \gamma \in \mathbb{R}_+.$$

**Proof.** Lemma 1 is Theorem 1 in Rüschendorf (1977) and the proof is given there. \qed

Lemma 1 provides the uniform consistency with probability one for a class of kernel estimators of multivariate density functions. This is a generalization of the consistency result of the univariate density estimation of Nadaraya (1965) and Schuster (1969) to the multivariate case. Note that we impose a stronger consistency than the pointwise weak consistency. The pointwise convergence, which does not assure global convergence after aggregation over all points, may not be sufficient to serve our purpose here. We refer to Wegman (1972) and Wied and Weißbach (2012) for a detailed discussion between different types of convergence. Based on Lemma 1, we do not only have local convergence as $\hat{f}(.) \overset{a.s.}{\longrightarrow} f(.)$ when $n \rightarrow \infty$, but also $\hat{v}(x, y, z) \overset{a.s.}{\longrightarrow} v(x, y, z)$ globally as a result.

**Lemma 2.** Suppose that $\hat{v}(x_i, y_i, z_i)$ is a consistent estimator of $v(x_i, y_i, z_i)$ a.s., as stated in Lemma 1, $T'_n$ has the same limiting distribution as the one with $\hat{v}(.)$ replaced by $v(.)$ if the asymptotic distribution exists. Formally, if we denote the statistic containing a true $v(.)$ by $\hat{T}'_n$, then we have

$$(T'_n - \hat{T}'_n) = o_p \left( \frac{1}{\sqrt{n}} \right).$$
CHAPTER 4. A NONPARAMETRIC TEST FOR GRANGER CAUSALITY

Proof. See Appendix 4.6.4.

Theorem 7. Given the Gaussian kernel, let the bandwidth tend to zero with \( n \to \infty \) as
\[
h = Cn^{-\beta}, C > 0, \beta \in \left(\frac{1}{4}, \frac{1}{3}\right),
\]
the statistic \( T'_n(h) \) satisfies
\[
\sqrt{n} \frac{T'_n(h) - t}{S_n} \xrightarrow{d} N(0, 1),
\]
where \( S_n^2 \) is a consistent estimator of the asymptotic variance
\[
\sigma^2 = 9\text{var}\left[\lim_{h \to \infty}(\tilde{K}_1(W_i, h)/v(W_i, h))\right].
\]

Proof. See Appendix 4.6.5.

We wish to make a final comment on Eq. (4.14) regarding the treatment of the marginals. Although our testing framework does not depend crucially on the restrictive assumption of a uniform distribution for the time series as in Pompe (1993) and Hong and White (2005), we recommend to use the probability integral transformation (PIT), which usually improves the performance of statistical dependence test, as Diks and Panchenko (2006) suggest. The reason is that, contrary to directly calculating the test statistics on the original data, the bounded support after transforming the marginals to a uniform distribution avoids non-existing moments during the bias and variance evaluation, which helps to stabilize the test statistic. There are more ways to transform the marginal variables into a bounded support than the PIT, for example, by using logistic function as Hong and White (2005). For practical convenience, we would only apply the PIT. The procedure is to transform the original series \( \{X_t\} (\{Y_t\}) \) to \( \{U^X_t\} (\{U^Y_t\}) \) such that \( \{U^X_t\} (\{U^Y_t\}) \) is the empirical CDF of \( \{X_t\} (\{Y_t\}) \) and the empirical distribution of \( \{U^X_t\} (\{U^Y_t\}) \) is uniform. Assuming \( X_t \) and \( Y_t \) are continuous, the PIT transformation is almost everywhere invertible and the dependence structure between \( X_t \) and \( Y_t \) remains intact in the pair of \( U^X_t \) and \( U^Y_t \). An extra merit of our test is that \( t \) is constructed to be invariant under the PIT transformation due to the fraction form in Eq. (4.13); the DP test, on the other hand, is not invariant since marginal transformation has a direct impact on the value of \( q \) in Eq. (4.5).

Since the transfer entropy based measure \( t \) is non-negative, tests based on the statistic \( T'_n(h) \) can be implemented as one-sided tests, rejecting the null hypothesis if
\[
\sqrt{n} T'_n(h)/S_n > z_{1-\alpha},
\]
where \( z_{1-\alpha} \) is the \((1-\alpha)\)th quantile of standard Normal distribution for a given significance level \( \alpha \).

4.2.5 Bandwidth Selection

In nonparametric settings, there typically is no uniformly most powerful test against all alternatives. Hence it is unlikely that a uniformly optimal bandwidth exists. As long as the bandwidth tends to zero with \( h = Cn^{-\beta}, C > 0, \beta \in \left(\frac{1}{4}, \frac{1}{3}\right) \), our test has asymptotical power. Yet, we may define the optimal bandwidth in the sense of minimal asymptotical mean
squared error (MSE). When balancing the first and forth leading terms in Eq. (4.33) to minimize squared bias and variance, for a second order kernel, it is easy to show that the optimal bandwidth for the DP test is given by

\[ h_{DP} = C n^{-2/7}, \quad \text{where} \quad C = \left( \frac{18 \times 3 q_2}{4 (E[s(W)]^2)} \right)^{1/7}, \]

with \( q_2 \) and \( E[s(W)] \) the series expansion for second moment of kernel function and expectation of bias, respectively. Since the convergence rate of the MSE, derived in Appendix 4.6.5, is not affected by the way we construct the new test statistic, the derivation of Eq. (4.15) remains intact and we simply calibrate the optimal bandwidth for our new test as

\[ h^* \approx 0.6 h_{DP}, \]

where the scale factor 0.6 involved as a result of bias and variance adjustment for replacing the Square kernel by Gaussian kernel. Intuitively, the \( q_2 \) and \( E[s(W)] \) terms are different from those for the DP test; more details can be found in Appendix 4.6.6.

The optimal value for \( C \) is process-dependent and difficult to track analytically. For example, Diks and Panchenko (2006) demonstrated that for a (G)ARCH process, the optimal bandwidth is approximately given by \( h_{DP} = C n^{-2/7} \) where \( C \approx 8 \). Applying Eq. (4.16), we proceed with \( h^* = 4.8 n^{-2/7} \) for (G)ARCH processes. To gain some insights into the bandwidth, we illustrate the test size and power with a 2-variate ARCH process, given by

\[ X_t \sim N(0, 1 + a Y_{t-1}^2), \]
\[ Y_t \sim N(0, 1 + a Y_{t-1}^2). \]

We let \( 0 < a < 0.4 \) and run 5,000 Monte Carlo simulations for time series length varying from 200 to 5,000. The size is assessed based on testing Granger non-causality from \( \{X_t\} \) to \( \{Y_t\} \), and for the power we use the same process but testing on Granger non-causality from \( \{Y_t\} \) to \( \{X_t\} \). The results are presented in Table 4.1, from which it can be seen that the MDP test is conservative in the sense that empirical size is lower than the nominal 0.05 in all cases, while the power converges to one when \( a \) increases, or the sample size increases.

### 4.3 Monte Carlo Simulations

This section investigates the performance of the modified DP test. Before proceeding with new data generating processes, we first revisit the example illustrated in Eq. (4.9) for which the DP test fails to detect the impact of \( X \) on \( Y \). The modified DP test is performed with 10,000 replications again, with the same bandwidth. The counterpart of the power-size plots for the DP test in Fig. 4.2 is delivered in Fig. 4.3. In contrast with the lack of power of the DP test, for time series length \( n = 500 \) and larger, the modified DP test already has a very
CHAPTER 4. A NONPARAMETRIC TEST FOR GRANGER CAUSALITY

Table 4.1: Observed Size and Power of the $T_n'(h)$ test for bivariate ARCH process Eq. (4.17)

<table>
<thead>
<tr>
<th>n</th>
<th>200</th>
<th>500</th>
<th>1,000</th>
<th>2,000</th>
<th>5,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0563</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6670</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5471</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4211</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a = 0.1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.0020</td>
<td>0.0016</td>
<td>0.0036</td>
<td>0.0016</td>
<td>0.0048</td>
</tr>
<tr>
<td>Power</td>
<td>0.0032</td>
<td>0.0128</td>
<td>0.0288</td>
<td>0.0792</td>
<td>0.4000</td>
</tr>
<tr>
<td>$a = 0.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.0020</td>
<td>0.0008</td>
<td>0.0032</td>
<td>0.0016</td>
<td>0.0048</td>
</tr>
<tr>
<td>Power</td>
<td>0.0208</td>
<td>0.0932</td>
<td>0.2824</td>
<td>0.7292</td>
<td>0.9992</td>
</tr>
<tr>
<td>$a = 0.3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.0020</td>
<td>0.0012</td>
<td>0.0032</td>
<td>0.0032</td>
<td>0.0044</td>
</tr>
<tr>
<td>Power</td>
<td>0.0816</td>
<td>0.3668</td>
<td>0.7916</td>
<td>0.9972</td>
<td>1.0000</td>
</tr>
<tr>
<td>$a = 0.4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.0020</td>
<td>0.0016</td>
<td>0.0032</td>
<td>0.0028</td>
<td>0.0084</td>
</tr>
<tr>
<td>Power</td>
<td>0.1928</td>
<td>0.6968</td>
<td>0.9848</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Note: Empirical size and power of the modified DP test for the process given in Eq. (4.17) for different sample sizes and parameter $a$. The values represent observed rejection rates over 5,000 realizations for nominal size 0.05.

High power in this artificial experiment, as expected.

Next, we use numerical simulations to study the behavior of the modified DP test, while direct comparisons between the modified DP test $T_n'$ and the DP test $T_n$ are also given. Three processes are being considered. In the first experiment, we consider a simple bivariate VAR process, given by

$$X_t = aY_{t-1} + \varepsilon_{x,t}, \varepsilon_{x,t} \sim N(0, 1),$$

$$Y_t = aY_{t-1} + \varepsilon_{y,t}, \varepsilon_{y,t} \sim N(0, 1).$$

(4.18)

The second process is designed as a nonlinear VAR process in Eq. (4.19). Again the size and power are investigated by testing for Granger non-causality in two different directions.

$$X_t = 0.6X_{t-1} + aX_{t-1}Y_{t-1} + \varepsilon_{x,t}, \varepsilon_{x,t} \sim N(0, 1),$$

$$Y_t = 0.6Y_{t-1} + \varepsilon_{y,t}, \varepsilon_{y,t} \sim N(0, 1).$$

(4.19)

The last process is same as the example we used for illustrating the performance of the bandwidth selection rule, which is a bivariate ARCH process also given in Eq. (4.17),

$$X_t \sim N(0, 1 + aY_{t-1}^2),$$

$$Y_t \sim N(0, 1 + aY_{t-1}^2).$$

(4.20)

The results which are shown in Figs. 4.4 to 4.6 are obtained with 5,000 simulations for each process. We present the DP test and the modified DP test with both empirical size-size plot and size-power for three processes in Eqs. (4.18) to (4.20) for sample size $n = 500$ and $n = 5,000$, respectively. The control parameter $a$ is considered to take the values 0.1 and 0.4. As before, the empirical size is obtained by testing Granger non-causality for $\{X_t\}$ to $\{Y_t\}$, and the empirical power is the observed rejection rate of testing for Granger non-causality from $\{Y_t\}$ to $\{X_t\}$.

It can be seen from Figs. 4.4 to 4.6 that the modified DP test is slightly more conservative.
Figure 4.3: Size-power plots for the one-sided modified DP test for the artificial process \( \{(X_t, Y_t, Z_t)\} \) with \( q = 0 \), based on 10,000 independent replications. Each subplot draws the actual power against the nominal size for different sample sizes, ranging from 100 to 10,000. The solid curve represents the actual power and the red dash line indicates the diagonal, indicating the ideal size of a test.
Figure 4.4: Size-size and size-power plots of Granger non-causality tests, based on 5,000 simulations. The DGP is bivariate linear VAR as in Eq. (4.18), with $Y$ affecting $X$. The left (right) column shows observed rejection rates under the null (alternative), blue color stands for DP test while red line indicates modified DP test. Real line and dash line present results with sample size $n = 500$ and $n = 5,000$, respectively.
Figure 4.5: Size-size and size-power plots of Granger non-causality tests, based on 5,000 simulations. The DGP is bivariate Non-linear VAR as in Eq. (4.19), with $Y$ affecting $X$. The left (right) column shows observed rejection rates under the null(alternative), blue color stands for DP test while red line indicates modified DP test. Real line and dash line present results with sample size $n = 500$ and $n = 5,000$, respectively.
Figure 4.6: Size-size and size-power plots of Granger non-causality tests, based on 5,000 simulations. The DGP is bivariate ARCH as in Eq. (4.20), with $Y$ affecting $X$. The left (right) column shows observed rejection rates under the null (alternative), blue color stands for DP test while red line indicates modified DP test. Real line and dash line present results with sample size $n = 500$ and $n = 5,000$, respectively.
than the DP test under the null hypotheses. However the size distortion reduces when the sample size increases. The modified DP test is more powerful than the DP test in the linear and nonlinear VAR settings given in Eqs. (4.18) and (4.19). Overall, we see that the larger the sample size and the stronger the causal effect are, the better the asymptotic performance of the modified DP test is.

4.4 Empirical Research

4.4.1 Stock Volume–Return Relation

In this section, we first revisit the stock return-volume relation following Hiemstra and Jones (1994) and Diks and Panchenko (2006). This topic has a long research history. Early empirical work mainly focused on the positive correlation between volume and stock price change, see Karpoff (1987). The later literature exposed directional relations, for example, Gallant et al. (1992) found that large price movements are followed by high volume; Gervais et al. (2001) observed a high-volume return premium, namely, periods of extremely high (low) volume tend to be followed by positive (negative) excess returns. More recently, Podobnik et al. (2009) investigated the power law cross-correlations between price changes and volume changes of the S&P 500 Index over a long period.

We use daily volume and returns data for the three most-followed indices in US stock markets, the Standard and Poor’s 500 (S&P), the NASDAQ Composite (NASDAQ) and the Dow Jones Industrial Average (DJIA), between January 1985 and October 2016. Yahoo Finance is the source of the daily volume and adjusted daily closing prices. The time series are converted by taking log returns multiplied by $100$. In order to adjust for the day-of-the-week and month-of-the-year seasonal effects in both mean and variance of stock returns and volumes, we perform a two-stage adjustment process, similar to the procedure applied in Hiemstra and Jones (1994)$^1$. We apply our test not only on the raw data, but also on VAR filtered residuals and EGARCH$(1,1,1)$ filtered residuals.$^2$ The idea of filtration is to remove linear dependence and the effect of heteroskedasticity to isolate the nonlinear and higher moment relationships among series, respectively.

Tables 4.2 to 4.4 report the resulting $t$ statistics for both the DP test and our modified DP test in both directions. The linear Granger F-values based on the optimal VAR models are also given. Two bandwidth values are used: 1.5 and 0.6, where the latter one roughly

---

$^1$We replace Akaike’s information criterion used by Hiemstra and Jones (1994) with the Schwarz et al. (1978) information criterion to be more stringent on picking up variables. Having no intention to provoke a debate over the two criteria, we simply prefer a more parsimonious liner model to avoid potential overfitting.

$^2$We have tried different error distributions like Normal, Students’$t$, GED and Hansen (1994)’s skewed $t$. The difference caused by different distribution assumptions are ignorable. Thus we only report for the result based on the Students’$t$ distribution for simplicity.
corresponds to the optimal bandwidth \((h = 0.6138)\) and the larger bandwidth, also used in Diks and Panchenko (2006), is added as a robustness check.

**Table 4.2:** Test Statistics for the S&P500 returns and volume data

<table>
<thead>
<tr>
<th></th>
<th>Volume → Return</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Linear</td>
<td></td>
<td>MDP</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>h = 1.5</td>
<td>h = 0.6</td>
<td>h = 1.5</td>
</tr>
<tr>
<td>Raw data</td>
<td></td>
<td>0.8503</td>
<td>2.8769**</td>
<td>2.9952**</td>
<td>3.8526**</td>
<td>2.7929**</td>
</tr>
<tr>
<td></td>
<td>VAR residuals</td>
<td></td>
<td></td>
<td>3.6880**</td>
<td>3.5683**</td>
<td>4.2696**</td>
</tr>
<tr>
<td></td>
<td>EGARCH residuals</td>
<td></td>
<td></td>
<td>1.4403</td>
<td>1.2347</td>
<td>1.2672</td>
</tr>
<tr>
<td></td>
<td>Return → Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw data</td>
<td></td>
<td>18.8302**</td>
<td>4.9309**</td>
<td>4.5138**</td>
<td>4.9253**</td>
<td>3.8359**</td>
</tr>
<tr>
<td></td>
<td>VAR residuals</td>
<td></td>
<td></td>
<td>5.2732**</td>
<td>5.1692**</td>
<td>5.2239**</td>
</tr>
<tr>
<td></td>
<td>EGARCH residuals</td>
<td></td>
<td></td>
<td>3.0067**</td>
<td>3.1214**</td>
<td>3.1176**</td>
</tr>
</tbody>
</table>

Note: Test statistics for the Granger causality between S&P500 returns and volume data. Results for bandwidth values 1.5 and 0.6 are reported. The asterisks indicate significance at the 5% (*) and 1% (**) levels.

**Table 4.3:** Test Statistics for the NASDAQ returns and volume data

<table>
<thead>
<tr>
<th></th>
<th>Volume → Return</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Linear</td>
<td></td>
<td>MDP</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>h = 1.5</td>
<td>h = 0.6</td>
<td>h = 1.5</td>
</tr>
<tr>
<td>Raw data</td>
<td></td>
<td>0.0979</td>
<td>3.5894**</td>
<td>3.3751**</td>
<td>4.1311**</td>
<td>3.3532**</td>
</tr>
<tr>
<td></td>
<td>VAR residuals</td>
<td></td>
<td></td>
<td>4.3932**</td>
<td>4.2931**</td>
<td>5.3026**</td>
</tr>
<tr>
<td></td>
<td>EGARCH residuals</td>
<td></td>
<td></td>
<td>0.8282</td>
<td>0.5604</td>
<td>1.0430</td>
</tr>
<tr>
<td></td>
<td>Return → Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw data</td>
<td></td>
<td>11.1736**</td>
<td>5.1100**</td>
<td>5.0201**</td>
<td>4.9980**</td>
<td>4.5935**</td>
</tr>
<tr>
<td></td>
<td>VAR residuals</td>
<td></td>
<td></td>
<td>5.6293**</td>
<td>6.4855**</td>
<td>5.5750**</td>
</tr>
<tr>
<td></td>
<td>EGARCH residuals</td>
<td></td>
<td></td>
<td>3.5959**</td>
<td>4.0693**</td>
<td>4.0745**</td>
</tr>
</tbody>
</table>

Note: Test statistics for the Granger causality between NASDAQ returns and volume data. Results for bandwidth values 1.5 and 0.6 are reported. The asterisks indicate significance at the 5% (*) and 1% (**) levels.

Generally speaking, the results indicate that the effect in the return-volume direction is stronger than the other way around. For the test results on the raw data, the \(F\)-tests based on the linear VAR model and both nonparametric tests suggest evidence of return affecting volume for all three indexes. For the other direction, causality from Volume to Return, the linear Granger test finds no evidence of causal impact while the nonparametric tests claim strong causal effect except for the DJIA where only the modified DP test finds a causal link from volume to return. As argued before, the results of linear test are suspicious since it only examines linear causal effect in the mean, information exchange from higher moments are completely ignored.
### 4.4 Empirical Research

#### Table 4.4: Test Statistics for the DJIA returns and volume data

<table>
<thead>
<tr>
<th></th>
<th>Volume $\rightarrow$ Return</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear DP $h = 1.5$</td>
<td>MDP $h = 1.5$</td>
<td>MDP $h = 0.6$</td>
</tr>
<tr>
<td>Raw data</td>
<td>0.9761</td>
<td>1.2557</td>
<td>1.3834*</td>
</tr>
<tr>
<td></td>
<td>1.9450*</td>
<td>2.2697*</td>
<td>2.7207**</td>
</tr>
<tr>
<td>VAR residuals</td>
<td>-</td>
<td>1.8711*</td>
<td>2.0951*</td>
</tr>
<tr>
<td></td>
<td>2.0998*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGARCH residuals</td>
<td>-</td>
<td>1.4543</td>
<td>1.4317</td>
</tr>
<tr>
<td></td>
<td>1.0566</td>
<td></td>
<td>1.4801</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Return $\rightarrow$ Volume</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear DP $h = 1.5$</td>
<td>MDP $h = 1.5$</td>
<td>MDP $h = 0.6$</td>
</tr>
<tr>
<td>Raw data</td>
<td>17.0779**</td>
<td>2.3076*</td>
<td>1.7236*</td>
</tr>
<tr>
<td></td>
<td>2.4972**</td>
<td>3.3222**</td>
<td></td>
</tr>
<tr>
<td>VAR residuals</td>
<td>-</td>
<td>2.3086*</td>
<td>2.0454*</td>
</tr>
<tr>
<td></td>
<td>2.6734**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGARCH residuals</td>
<td>-</td>
<td>0.8033</td>
<td>1.0589</td>
</tr>
<tr>
<td></td>
<td>1.8989*</td>
<td></td>
<td>3.0707**</td>
</tr>
</tbody>
</table>

Note: Test statistics for the Granger causality between DJIA returns and volume data. Results for bandwidth values 1.5 and 0.6 are reported. The asterisks indicate significance at the 5% (*) and 1% (**) levels.

A direct comparison between the DP test and the modified DP test shows that the new test is more powerful overall. For the unfiltered data, both tests find a strong causal effect in two directions for S&P and NASDAQ, but for the DJIA, the $t$-statistics of DP test are weaker than those of the modified DP test. The bi-directional causality between return and volume remains unchanged after linear VAR filtration, although the DP test again shows weaker evidence. The result also suggests that the causality is strictly nonlinear. The linear test ($F$-test) either overlooks the nonlinear linkages, or is unable to spot its nonlinear nature.

Further, in the direction from Volume to Return, these nonlinear causalities tend to vanish after EGARCH filtering. Thus the bi-directional linkage is reduced to an one-directional relation from return to volume. The modified DP statistics, however, are in general larger than the DP $t$-values, and indicate more causal relations. In contrast with the DP test, our test suggests that the observed nonlinear causality cannot be completely attributed to the second moment effect. Heteroskedasticity modeling may reduce this nonlinear feature to some extent, but its impact is not as strong as the DP test would suggest. As argued above, the DP test may suffer from the problem of lacking power as it is based on a quantity that is not positive definite.

#### 4.4.2 Application to Intraday Exchange Rates

In the second application, we apply the modified DP test to intraday exchange rates. We consider five major currency: JPY, AUD, GBP, EUR and CHF, all against the USD. The data, obtained from Dukascopy Historical Data Feed, contain 5-minute bid and ask quotes for the third quarter of 2016; from July 1 to September 30, with a total of 92 trading days and 26496 high frequency observations. We use 5-minute data, corresponding to the sampling frequency of 288 times per day, which is high enough to avoid measurement errors (see Andersen and Bollerslev (1998)) but also low enough such that the micro-structure is not the major concern.
Although the foreign exchange market is one of the most active financial markets in the world, where trading takes place 24 hours a day, intraday trading is not always active. Thus we delete the thin trading period, from Friday 21:00 GMT until Sunday 20:55 GMT, also to keep the intraday periodicity intact. We calculate the exchange rate returns as in Diebold et al. (1999). First the average log bid and log ask prices are calculated, then the difference between the log prices at consecutive time stamps are obtained. Next, we remove the conditional mean dynamics by fitting a MA(1) model and using the residuals as our return series following Bollerslev and Domowitz (1993). Finally intraday seasonal effects are filtered out using estimated time-of-day dummies following Diebold et al. (1999), i.e.

\[ r_{i,n,t} = d_{i,t} z_{i,n,t}, \]

where \( r_{i,n,t} \) denotes the continuously intraday log returns after MA(1) filtration. The subscript \( i = 1, \ldots, 5 \) indicates five different currency and \( n,t \) stands for time \( t \) on day \( n \). The first component of return series \( d_{i,t} \) refers to a deterministic intraday seasonal component while \( z_{i,n,t} \) is the nonseasonal return portion, which is assumed to be independent of \( d_{i,t} \). To distinguish \( d_{i,t} \) from \( z_{i,n,t} \), we fit the time-of-day dummies to \( 2 \log |r_{i,n,t}| \) and use the estimated \( \hat{d}_{i,t} \) to standardize the return \( r_{i,n,t} \) with the restriction \( \sum_{t=1}^{T} d_{i,t} = 1 \). Figs. 4.7 to 4.9 report the first 200 autocorrelations of returns, absolute returns and squared returns, when checking on the raw series, MA(1) residuals and EGARCH residuals, respectively.

We perform pairwise nonparametric Granger causality tests on the MA(1) filtered and seasonally adjusted data, as well as on the standardized residuals after EGARCH(1,1,1) filtering. We use Hansen (1994)’s skewed t distribution to model the innovation terms. We choose a bandwidth of 0.2768, according to Eq. (4.16).

The test results are shown in Table 4.5 for both MA(1) demeaned and de-seasoned data, as well as EGARCH filtered data. Although not reported here, there is statistical evidence for strong bi-directional causality among all currency pairs on raw return data at 5-minute lag. These bi-directional causalities do not disappear after removing MA(1) component and seasonal component. However, the observed information spillover is significantly weaker after the EGARCH filtering. When testing on the EGARCH standardized residuals, only a few pairs still show signs of a strong causal relation. Especially, the directional relation of EUR → CHF is the only one detected by both the DP test and the modified DP test at the 1% level of significance. A graphical representation is provided in Fig. 4.10, where one can clearly see that most causal links are gone after EGARCH filtering. The modified DP test exposes five uni-directional linkages among the EGARCH filtered returns at the 5% level. The EUR and GBP are the most important driving currencies. While the DP test also admits the importance of JPY and particularly AUD, which shows bi-directional causality between JPY and GBP.

To sum up, we find evidence of strong causal links among exchange returns on intraday high-frequency level. Each currency has predictive power on other currencies, implying the
Figure 4.7: Autocorrelations of Returns, Absolute Returns and Square Returns, up to 200 lags.
Figure 4.8: Autocorrelation of Returns, Absolute Returns and Square Returns after MA(1) Component removed, up to 200 lags.
Figure 4.9: Autocorrelation of Returns, Absolute Returns and Square Returns after MA(1) and GARCH Component removed, up to 200 lags.


### Table 4.5: Test Statistics for the Pairwise Granger Causality on Raw Exchange Returns

<table>
<thead>
<tr>
<th>Pair</th>
<th>MA residuals</th>
<th>EGARCH residuals</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DP residuals</td>
<td>MDP residuals</td>
<td>DP residuals</td>
<td>MDP residuals</td>
</tr>
<tr>
<td>JPY AUD</td>
<td>4.0180**</td>
<td>3.0086**</td>
<td>1.9843*</td>
<td>1.2611</td>
</tr>
<tr>
<td>JPY EUR</td>
<td>4.4724**</td>
<td>3.7586**</td>
<td>0.9336</td>
<td>0.5734</td>
</tr>
<tr>
<td>JPY GBP</td>
<td>4.4096**</td>
<td>3.9775**</td>
<td>0.4305</td>
<td>0.3480</td>
</tr>
<tr>
<td>JPY CHF</td>
<td>4.3236**</td>
<td>3.9867**</td>
<td>1.6542*</td>
<td>1.5474</td>
</tr>
<tr>
<td>AUD JPY</td>
<td>4.4872**</td>
<td>3.6505**</td>
<td>2.0162*</td>
<td>1.4173</td>
</tr>
<tr>
<td>AUD EUR</td>
<td>4.5398**</td>
<td>3.5291**</td>
<td>2.7414**</td>
<td>1.9737*</td>
</tr>
<tr>
<td>AUD GBP</td>
<td>3.9936**</td>
<td>2.8616**</td>
<td>1.6532*</td>
<td>0.6208</td>
</tr>
<tr>
<td>AUD CHF</td>
<td>3.2458**</td>
<td>3.2727**</td>
<td>1.5546</td>
<td>1.5006</td>
</tr>
<tr>
<td>EUR JPY</td>
<td>4.0257**</td>
<td>3.2913**</td>
<td>1.1139</td>
<td>0.3133</td>
</tr>
<tr>
<td>EUR AUD</td>
<td>3.7456**</td>
<td>3.1796**</td>
<td>1.5543</td>
<td>1.0551</td>
</tr>
<tr>
<td>EUR GBP</td>
<td>5.3053**</td>
<td>4.3236**</td>
<td>3.0613**</td>
<td>2.2103*</td>
</tr>
<tr>
<td>EUR CHF</td>
<td>5.5101**</td>
<td>4.6634**</td>
<td>3.8299**</td>
<td>3.5006**</td>
</tr>
<tr>
<td>GBP JPY</td>
<td>4.2506**</td>
<td>3.4310**</td>
<td>0.3216</td>
<td>0.0284</td>
</tr>
<tr>
<td>GBP AUD</td>
<td>4.7248**</td>
<td>4.0036**</td>
<td>2.3418**</td>
<td>1.9653*</td>
</tr>
<tr>
<td>GBP EUR</td>
<td>4.7092**</td>
<td>3.9164**</td>
<td>1.3648</td>
<td>0.5487</td>
</tr>
<tr>
<td>GBP CHF</td>
<td>2.7094**</td>
<td>2.2224*</td>
<td>2.0109*</td>
<td>1.6580*</td>
</tr>
<tr>
<td>CHF JPY</td>
<td>4.0033**</td>
<td>3.4545**</td>
<td>0.9972</td>
<td>0.3293</td>
</tr>
<tr>
<td>CHF AUD</td>
<td>3.3506**</td>
<td>2.5622**</td>
<td>0.8823</td>
<td>0.0981</td>
</tr>
<tr>
<td>CHF EUR</td>
<td>3.8227**</td>
<td>2.6958**</td>
<td>1.6378</td>
<td>0.2864</td>
</tr>
<tr>
<td>CHF GBP</td>
<td>3.6522**</td>
<td>3.0242**</td>
<td>1.8381*</td>
<td>1.5102</td>
</tr>
</tbody>
</table>

Note: Statistics for pairwise Granger non-causality tests on high-frequency returns of five major currencies. The data are first cleaned by MA(1) component and seasonal component, and then standardized by EGARCH variance. Results are shown both for the DP test and the modified DP test with bandwidth $h = 0.2877$. The asterisks indicate significance at the 5% (*) and 1% (**) levels.
Figure 4.10: Graphical representation of pairwise causalities on MA and seasonal component filtered residuals, as well as EGARCH filtered residuals. All "→" in the graph indicate a significant directional causality at 5% level.
high co-movement in the international exchange market. Although those directional linkages are not affected by demeaning procedure, we may reduce most of them by taking the volatility dynamic into account. When filtering out heteroskedasticity by EGARCH estimation, there only exist a few pairs containing spillover effect.

4.5 Summary and Conclusions

Borrowing the concept of transfer entropy from Information Theory, this chapter develops a novel non-parametric test statistic for Granger non-causality. The asymptotic normality is derived for the test statistic by taking advantage of an order-three $U$-statistic representation initially applied in the DP test. The modified DP statistic, however, overwhelms the DP statistic in at least the two respects: firstly, positive definiteness is a natural property for our statistic, which paves the way for properly testing for differences between conditional densities; secondly, the weight function in our test is justified by the theoretical information representation, while the weight function in the DP test was selected in an ad hoc manner. Simulation studies show that our modified DP test has reasonable size and power for a wide range of data generating processes. In the first application, a direct comparison with the DP test confirms that the DP test may suffer from the lack of power problem, while the second application to high frequency exchange return data helps us better understand whether the spillover channel in exchange rate markets arises from mean, variance or higher moments. A possible extension to future work is incorporating more lags in a multivariate setting, and conditioning on additional, possibly confounding variables, in which cases new asymptotic theory is needed.
4.6 Appendix II

4.6.1 Proof of Positive-definiteness of the Transfer Entropy

According to the definition of TE in Eq. (4.14), the expectation over the logarithm of the density ratio is evaluated:

$$ TE_{X \rightarrow Y} = E_W \left( \log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} \right). $$

Define the reciprocal of the density ratio in the logarithm as a random variable $R$ in such a way:

$$ R = \frac{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}{f_{Z,X|Y}(Z,X|Y)}. $$

we could rewrite TE as

$$ TE_{X \rightarrow Y} = E \left( -\log(R) \right). $$

Since $\log(R)$ is a concave function of $R$, following Jensen’s inequality we have

$$ E(\log(R)) \leq \log(E(R)). \tag{4.22} $$

Next, as random variable $R$ is nonnegative since it is defined as a fraction of densities. For any realization of $R = r > 0$, $\log(r) \leq r - 1$. This is because as a concave function, $\log(r)$ is bounded from above by the tangent line at point $(1,0)$, which is given by $r - 1$. It follows that

$$ \log(E(R)) \leq E(R) - 1. \tag{4.23} $$

On combining Eqs. (4.22) and (4.23), we have $E(\log(R)) \leq E(R) - 1 = 0$, where the last equality holds simply as a result of integral of the pdf over its full support delivering 1. A similar argument can be found in Diks (2009). Thus, we have proved that $TE_{X \rightarrow Y} \equiv -E(\log(R)) \geq 0$. It is obvious that the equality holds if and only if $R = 1$, which is equivalent to $f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)$. This completes the proof of Thm. 7.
4.6.2 Proof of Positive-definiteness of the First Order TE Statistic

Starting from Eq. (4.11) and the definition of $t$,

$$
(4.24) \quad t = E_W \left( \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} - 1 \right)
$$

$$
= \int \int \left( \frac{f_{Z,X|Y}^2(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} - f_{Z,X|Y}(z,x|y) \right) \, dz | y \, dz | y 
$$

$$
= \int \int f_{X|Y}(x|y)f_{Z|Y}(z|y) \left( \frac{f_{Z,X|Y}^2(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} - \frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} \right) \, dz | y \, dz | y 
$$

$$
= \int \int f_{X|Y}(x|y)f_{Z|Y}(z|y) \left( \frac{f_{Z,X|Y}^2(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} - 2 \frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} + 1 \right) \, dz | y \, dz | y 
$$

$$
= \int \int f_{X|Y}(x|y)f_{Z|Y}(z|y) \left( \frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} - 1 \right)^2 \, dz | y \, dz | y \geq 0,
$$

with equality if and only if $\frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} = 1$ for any $(z,x|y)$ in the support of $(X,Y,Z)$. In the fourth step we completed the square by using the fact that the integral over the whole support of pdf is 1. Finally, $t \geq 0$ follows naturally from the integrand being non-negative.

4.6.3 Proof of Non-Degeneracy of the Modified DP Test Statistic

To show that the asymptotic normality in Thm. 7 is a non-degenerate distribution, it is sufficient to prove that with the plug-in weighting function $v(\cdot)$, the modified DP test U-statistic kernel is non-degenerate.

The symmetrized U-statistic representation of the modified DP test statistic defined in Eq. (4.14) is given by

$$
(4.25) \quad K(w_1, w_2, w_3) = [(\kappa_{X|Y}(w_1 - w_2)\kappa_{Y}(w_1 - w_3) - \kappa_{X|Y}(w_1 - w_2)\kappa_{Y}(w_1 - w_3)) / v(w_1)] / 6 
$$

$$
+ \text{permutations of } w_1, w_2, w_3,
$$

where $\kappa$ is a (bandwidth $h$ dependent) density estimation kernel function, and $w_i = (x_i, y_i, z_i)'$, $i \in \{1,2,3\}$.

Let $r_1$ be the Hájek projection

$$
r_1(w_1; h) = E(K(w_1, W_2, W_3))
$$
of the U-statistic kernel.

Define

\[ r_1(w_1) = \lim_{h \to 0} r_1(w_1; h), \]

the U-statistic is degenerate (in that the variance is of higher order than in the derivation of the modified DP test statistic) if \( r_1(w_1) \) is constant as a function of \( w_1 = (x_1, y_1, z_1)' \).

Combining the above equations and Eq. (4.25), we obtain

\[
(4.26) \\
r_1(w_1; h) = E \left[ (\kappa_{XYZ}(w_1 - W_2)\kappa_Y(w_1 - W_3) - \kappa_{XY}(w_1 - W_2)\kappa_{YZ}(w_1 - W_3)) / v(w_1) \right] / 6 \\
+ E \left[ (\kappa_{XYZ}(w_2 - w_1)\kappa_Y(W_2 - W_3) - \kappa_{XY}(w_2 - w_1)\kappa_{YZ}(W_2 - W_3)) / v(w_1) \right] / 6 \\
+ E \left[ (\kappa_{XYZ}(W_3 - W_2)\kappa_Y(w_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_{YZ}(W_3 - w_1)) / v(w_1) \right] / 6 \\
+ E \left[ (\kappa_{XYZ}(w_1 - W_3)\kappa_Y(w_1 - W_2) - \kappa_{XY}(w_1 - W_3)\kappa_{YZ}(w_1 - W_2)) / v(w_1) \right] / 6 \\
+ E \left[ (\kappa_{XYZ}(W_3 - W_1)\kappa_Y(W_3 - W_2) - \kappa_{XY}(W_3 - W_1)\kappa_{YZ}(W_3 - W_2)) / v(w_1) \right] / 6 \\
+ E \left[ (\kappa_{XYZ}(W_3 - W_2)\kappa_Y(W_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_{YZ}(W_3 - w_1)) / v(w_1) \right] / 6 \\
\equiv E_1(w_1; h) / 3 + E_2(w_1; h) / 3 + E_3(w_1; h) / 3,
\]

where in the last step we used the fact that the terms with \( W_2 \) and \( W_3 \) swapped are identical.

We next consider

\[
r_1(w_1) = \lim_{h \to 0} r_1(w_1; h) = \lim_{h \to 0} E_1(w_1; h) / 3 + \lim_{h \to 0} E_2(w_1; h) / 3 + \lim_{h \to 0} E_3(w_1; h) / 3 \\
\equiv E_1(w_1) / 3 + E_2(w_1) / 3 + E_3(w_1) / 3.
\]

For \( E_1(w_1) \) we find

\[
(4.27) \\
E_1(w_1) = \lim_{h \to 0} E \left[ (\kappa_{XYZ}(w_1 - W_2)\kappa_Y(w_1 - W_3) - \kappa_{XY}(w_1 - W_2)\kappa_{YZ}(w_1 - W_3)) / v(w_1) \right] \\
= \lim_{h \to 0} \int \int f_W(w_2)f_W(w_3) [ (\kappa_{XYZ}(w_1 - W_2)\kappa_Y(w_1 - W_3) \\
- \kappa_{XY}(w_1 - W_2)\kappa_{YZ}(w_1 - W_3)) / v(w_1) ] \, dw_2dw_3 \\
= \frac{1}{v(w_1)} \int \int f_{XY}(w_2)f_{YZ}(w_3) (\delta_{XYZ}(w_1 - w_2)\delta_Y(w_1 - w_3) \\
- \delta_{XY}(w_1 - w_2)\delta_{YZ}(w_1 - w_3)) \, dw_2dw_3 \\
= \frac{1}{v(w_1)} (f_{XYZ}(w_1)f_Y(w_1) - f_{XY}(w_1)f_{YZ}(w_1)),
\]
where in the third step, $\delta(\cdot)$ is the Dirac delta function, also referred to as the unit impulse symbol. Using convolution, we have the last equality. Under $H_0$ for all $w_1$ in the support of $W$, Eq. (4.27) is zero by construction. However, the other terms, $E_2(w_1)$ and $E_3(w_1)$, need not be constant even under $H_0$. For instance,
(4.28)
\[ E_2(w_1, h) = E[(\kappa_{XYZ}(W_2 - w_1)\kappa_Y(W_2 - W_3) - \kappa_{XY}(W_2 - w_1)\kappa_{YZ}(W_2 - W_3))/v(w_1)] \]
\[ E_2(w_1) = \lim_{h \to 0} E[(\kappa_{XYZ}(W_2 - w_1)\kappa_Y(W_2 - W_3) - \kappa_{XY}(W_2 - w_1)\kappa_{YZ}(W_2))/v(w_1)] \]
\[ = \lim_{h \to 0} E[(\kappa_{XYZ}(W_2 - w_1)f_Y(W_2) - \kappa_{XY}(W_2 - w_1)f_{YZ}(W_2))/v(w_1)] \]
\[ = \lim_{h \to 0} \int f_{XYZ}(w_2)\kappa_{XYZ}(w_2 - w_1)f_Y(w_2)/v(w_1)dw_2 \]
\[ - \lim_{h \to 0} \int f_{XYZ}(w_2)\kappa_{XY}(w_2 - w_1)f_{YZ}(w_2)/v(w_1)dw_2 \]
\[ = \int f_{XYZ}(w_2)\delta_{XYZ}(w_2 - w_1)f_Y(w_2)/v(w_1)dw_2 \]
\[ - \lim_{h \to 0} \int f_{XYZ}(w_2)\kappa_{XY}(w_2 - w_1)f_{YZ}(w_2)/v(w_1)dw_2 \]
\[ = f_{XYZ}(w_1)f_Y(w_1)/v(w_1) - \lim_{h \to 0} \int f_{XYZ}(w_2)\kappa_{XY}(w_2 - w_1)f_{YZ}(w_2)/v(w_1)dw_2 \]
\[ = f_{XYZ}(w_1)f_Y(w_1)/v(w_1) \]
\[ - \lim_{h \to 0} \int f_{XYZ}(x_2, y_2, z_2)\kappa_{XY}(x_2 - x_1, y_2 - y_1)f_{YZ}(y_2, z_2)/v(w_1)dx_2dy_2dz_2 \]
\[ = f_{XYZ}(w_1)f_Y(w_1)/v(w_1) \]
\[ - \frac{1}{v(w_1)} \int f_{XYZ}(x_2, y_2, z_2)\delta_{XY}(x_2 - x_1, y_2 - y_1)f_{YZ}(y_2, z_2)dx_2dy_2dz_2 \]
\[ = \frac{1}{v(w_1)}(f_{XYZ}(w_1)f_Y(w_1) - \int f_{XYZ}(x_1, y_1, z_2)f_{YZ}(y_1, z_2)dz_2). \]

Since the last term in the bracket does not depend on \( z_1 \), while the first typically depends on \( x_1, y_1 \) and \( z_1 \) under \( H_0 \), \( E_2(w_1) \) typically isn’t constant. A similar argument also can be used to show that \( E_3(w_1) \) is not constant (and neither is \( E_2(w_1) + E_3(w_1) \), because \( E_3(w_1) \) is a function of \( (y_1, z_1) \) only, while \( E_2(w_1) \) also depends on \( x_1 \) typically).

For completeness, \( E_3(w_1, h) \) and \( E_3(w_1) \) are given below:
(4.29)  
\[ E_3(w_1, h) = E \left[ (\kappa_{XYZ}(W_3 - W_2)\kappa_Y(W_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_Y(W_3 - w_1)) / v(w_1) \right] \]

\[ E_3(w_1) = \lim_{h \to 0} E \left[ (\kappa_{XYZ}(W_3 - W_2)\kappa_Y(W_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_Y(W_3 - w_1)) / v(w_1) \right] \]

\[ = \lim_{h \to 0} \int f_{XYZ}(w_3)\kappa_Y(w_3 - w_1)f_{XYZ}(w_3)/v(w_1) \, dw_3 \]

\[- \lim_{h \to 0} \int f_{XY}(w_3)\kappa_Y(w_3 - w_1)f_{XYZ}(w_3)/v(w_1) \, dw_3 \]

\[ = \lim_{h \to 0} \int f_{XYZ}(x_3, y_3, z_3)\kappa_Y(y_3 - y_1)f_{XYZ}(x_3, y_3, z_3)/v(w_1) \, dx_3dy_3dz_3 \]

\[- \lim_{h \to 0} \int f_{XY}(x_3, y_3)\kappa_Y(y_3 - y_1, z_3 - z_1)f_{XYZ}(x_3, y_3, z_3)/v(w_1) \, dx_3dy_3dz_3 \]

\[ = \frac{1}{v(w_1)} \left( \int f_{XY}(x_3, y_1)f_{XYZ}(x_3, y_1, z_1) \, dx_3dz_3 - \int f_{XY}(x_3, y_1)f_{XYZ}(x_3, y_1, z_1) \, dx_3dz_3 \right). \]

Because \( E_2(w_1) \) and \( E_3(w_1) \) are not constant, \( r_1(w_1) \) cannot be constant, hence the U-statistic defined in Eq. (4.14) is non-degenerate.

To illustrate that \( E_2(w_1) \) and \( E_3(w_1) \) are typically not constant, consider the example where \( W \sim N(0, I_3) \). In this case \( H_0 \) holds, so we have \( E_1 = 0 \), while

\[ v(w_1) = f_{XY}(x_1, y_1)f_{YZ}(y_1, z_1) = \left( \frac{1}{\sqrt{2\pi}} \right)^4 e^{-(x_1^2 + 2y_1^2 + z_1^2)/2}, \]

upon plugging this into Eqs. (4.28) and (4.29) we obtain

\[ E_2(w_1) = \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^4 e^{-(x_1^2 + 2y_1^2 + z_1^2)/2} - \left( \frac{1}{\sqrt{2\pi}} \right)^5 \int e^{-(x_1^2 + 2y_1^2 + 2z_1^2)/2} \, dz_2 \right] / v(w_1) \]

\[ = \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^4 e^{-(x_1^2 + 2y_1^2 + z_1^2)/2} - \left( \frac{1}{\sqrt{2\pi}} \right)^5 \int e^{-(x_1^2 + 2y_1^2)/2} \, dx_3 \right] / v(w_1) \]

\[ = \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^4 e^{-(x_1^2 + 2y_1^2 + z_1^2)/2} - \left( \frac{1}{\sqrt{2\pi}} \right)^4 \frac{1}{\sqrt{2}} e^{-(x_1^2 + 2y_1^2)/2} \right] / v(w_1) \]

\[ = \left( \frac{1}{\sqrt{2\pi}} \right)^4 \left( e^{-z_1^2/2} - \frac{1}{\sqrt{2}} \right) e^{-(x_1^2 + 2y_1^2)/2} / v(w_1) \]

\[ = \left( e^{-z_1^2/2} - \frac{1}{\sqrt{2}} \right) e^{x_1^2/2} \]

\[ = 1 - \frac{1}{\sqrt{2}} e^{x_1^2/2} \]
and

$$E_3(w_1) = \left[ \int f_{X,Y,Z}^2(x, y_1, z) \, dx \, dz - \int f_{X,Y}(x, y_1) f_{X,Y,Z}(x, y_1, z_1) \, dx \right] / v(w_1)$$

$$= \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^6 \int \left( e^{-(x^2+y_1^2+z^2)/2} \right)^2 \, dx \, dz - \left( \frac{1}{\sqrt{2\pi}} \right)^5 \int e^{-(2x^2+2y_1^2+z_1^2)/2} \, dx \right] / v(w_1)$$

$$= \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^6 e^{-y_1^2} \int e^{-(x^2+z^2)} \, dx \, dz - \left( \frac{1}{\sqrt{2\pi}} \right)^5 \int e^{-(2y_1^2+z_1^2)/2} \, dx \right] / v(w_1)$$

$$= \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^6 e^{-y_1^2} \pi - \left( \frac{1}{\sqrt{2\pi}} \right)^5 e^{-(2y_1^2+z_1^2)/2} \frac{1}{\sqrt{2\pi}} \right] / v(w_1)$$

$$= \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^5 \frac{1}{\pi^2} \left( \frac{1}{\sqrt{2}} e^{-y_1^2} - e^{-y_1^2} e^{-z_1^2/2} \right) \right] / v(w_1)$$

$$= \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^5 \frac{1}{\pi^2} e^{-y_1^2} \left( \frac{1}{\sqrt{2}} - e^{-z_1^2/2} \right) \right] / v(w_1)$$

$$= \frac{1}{\sqrt{2}} \left( \frac{1}{\sqrt{2}} - e^{-z_1^2/2} \right) e^{(x_1^2+z_1^2)/2},$$

so clearly both $E_2(w_1)$ and $E_3(w_1)$ are not constant.

### 4.6.4 Proof of Lemma 2

Suppose we denote the asymptotic distribution of $\tilde{T}'_n$ in such a way that

$$\sqrt{n} \tilde{T}'_n(h) - t \frac{d}{\sigma} \rightarrow N(0,1),$$

with appropriate variance $\sigma$. To prove Lemma 2, it is sufficient to show that

$$\sqrt{n} T'_n(h) - t \frac{d}{\sigma} \rightarrow N(0,1),$$

because the two limiting distributions imply that $\sqrt{n}(T'_n - \tilde{T}'_n) = o_P(1)$.

To construct the equivalence of the limiting distributions, we need to prove that if $\hat{v}(.)$ is consistent, the limiting distribution in Eq. (4.30) remains for $T'_n$. Lemma 1 guarantees that the estimate of a density is consistent. For a consistent density estimate, it can always be expressed as the corresponding true density plus an extra error term which vanishes in limit, for example $\hat{f}_{X,Y}(x, y_i) = f_{X,Y}(x, y_i) + \varepsilon_{X,Y}(x, y_i)$. Following this approach and ignoring the subscripts for simplicity, we can show that the estimated $\hat{v}(.)$ differs from $v(.)$ only in the
CHAPTER 4. A NONPARAMETRIC TEST FOR GRANGER CAUSALITY

higher order:

\begin{equation}
\frac{1}{v(x_i, y_i, z_i)} = \frac{1}{f(x_i, y_i) f(y_i, z_i)}
\end{equation}

4.6.5 Proof of the Asymptotic Distribution of the Modified DP Statistic

Applying Lemma 2, it remains to be proven that

\[ \sqrt{n} \frac{T^I_n(h) - t}{S_n} \overset{d}{\to} N(0, 1). \]

According to the definition of \( T^I_n(h) \) in Eq. (4.14) is a re-scaled DP statistic with the scaling factor \( 1/v(\cdot) \). In a similar manner of Theorem 1 in Diks and Panchenko (2006), we can obtain the asymptotic behavior of \( T^I_n(h) \) by making use of the optimal mean squared error (MSE) bandwidth developed by Powell and Stoker (1996) for this point estimator. The test statistic \( T^I_n(h) \) can be expressed by an order three \( U \)-statistic \( \tilde{K}(W_i, W_j, W_k) \) by symmetrization with respect to the indices \( i, j, k \). Further, defining two kernel functions as \( \tilde{K}_1(w_i) = E[\tilde{K}(W_i, W_j, W_k)] \) and \( \tilde{K}_2(w_i, w_j) = E[\tilde{K}(w_i, W_j, W_k)] \), we assume the three mild conditions adapting from Powell and Stoker (1996) for controlling the rate of convergence of
the point-wise bias as well as the serial expansions of the kernel functions, being

\[
\tilde{K}_1(w_i, h) - \lim_{h \to 0} \tilde{K}_1(w_i, h) = s(w_i)h^\alpha + s^*(w_i, h), \quad \alpha > 0,
\]

\[
(4.32) \quad E[(\tilde{K}_2(W_i, W_j))^2] = q_2h^{-\gamma} + q_2^*(h), \quad \gamma > 0,
\]

\[
E[(\tilde{K}(W_i, W_j, W_k))^2] = q_3h^{-\delta} + q_3^*(h), \quad \delta > 0,
\]

where all remainder terms are of higher orders, i.e. \(E|s^*(W_i, h)|^2 = oP(h^{2\alpha})\), \(q_2^*(h) = oP(h^{-\gamma})\) and \(q_3^*(h) = oP(h^{-\delta})\) and the convergence rate is controlled by the parameters \(\alpha, \gamma\) and \(\delta\). The conditions in Eq. (4.32) are satisfied if \(\alpha\) is set as the order of kernel function \(K(.)\), which is 2 for the Gaussian kernel, and \(\gamma, \delta\) depend on the dimensions of the variables under consideration via \(\gamma = d_X + d_Y + d_Z\) and \(\delta = d_X + 2d_Y + d_Z\). Define \(C_0 = 2\text{cov}[\lim_{h \to 0} \tilde{K}_1(W_i, h), s(W_i)]\), we can show that the mean squared error of DP statistic as a function of sample size dependent bandwidth is given by

\[
(4.33) \quad \text{MSE}[T_n(h)] = (E[s(W_i)])^2h^{2\alpha} + \frac{9}{n}C_0h^\alpha + \frac{9}{n}\text{Var}\left[\lim_{h \to 0} \tilde{K}_1(W_i, h)\right] + \frac{18}{n^2}q_2h^{-\gamma} + \frac{6}{n^3}q_3h^{-\delta} + \text{h.o.t.}
\]

The scaling factor \((1/v(\cdot))\) in \(\tilde{T}_n(h)\) would enter the MSE in Eq. (4.33) by mainly changing the variance term. For other bandwidth-dependent terms, \((1/v(\cdot))\) just re-scale their coefficients without affecting the convergence rates. Thus, we may still allow for all the \(h\)-dependent terms to be \(oP(n^{-1})\) to make the 9Var \(\left[\lim_{h \to 0} \tilde{K}_1(W_i, h)\right]\) dominant as in Diks and Panchenko (2006). Therefore, adopting a sample size-dependent bandwidth \(h = Cn^\beta\), with \(C, \beta > 0\), one finds

\[
(4.34) \quad \sqrt{n}\frac{\tilde{T}_n(h) - t}{S_n} \xrightarrow{d} N(0, 1) \quad \text{if} \quad \frac{1}{2\alpha} < \beta < \frac{1}{d_X + d_Y + d_Z},
\]

where \(S_n^2\) is a consistent estimator of the asymptotic variance 9Var \(\left[\lim_{h \to 0} \tilde{K}_1(W_i, h)\right]\). In our bivariate case, \(\alpha = 2\) and \(d_X + d_Y + d_Z = 3\), and we would have \(\beta \in (1/4, 1/3)\).

### 4.6.6 Optimal Bandwidth for the Modified DP Test

The optimal bandwidth should balance the squared bias and variance of the test statistic, given in Eq. (4.33). Particularly, the first and fourth terms are leading, and all reminders are of higher order. The optimal bandwidth should have a similar form as the one for DP test in Eq. (4.15).

In fact, the difference between two bandwidths is up to a scalar as a result of replacing the Square kernel by Gaussian one. Assuming the product kernel in Eq. (4.6), the bias and variance of the density estimator are described following Wand and Jones (1994) and Hansen.
CHAPTER 4. A NONPARAMETRIC TEST FOR GRANGER CAUSALITY

(2009),

\[
\text{Bias}(\hat{f}(x)) = \frac{\mu_\nu(\kappa)}{\nu!} \sum_{j=1}^{k} \frac{\partial^\nu}{\partial x_j^\nu} f(x) h_j^\nu + \alpha_P(h_1^\nu + ... + h_k^\nu),
\]

(4.35)

\[
\text{Var}(\hat{f}(x)) = \frac{f(x) R(\kappa)^k}{(n-1)h_1 h_2...h_k} + \alpha_P((n-1)h_1 h_2...h_k),
\]

where \(\mu_\nu(\kappa) = \int_{-\infty}^{\infty} t^\nu \kappa(t) dt\) is the \(\nu\)th moment of a kernel function, with \(\nu\) the corresponding order of the kernel. For Gaussian kernel \(\kappa(.)\), \(\nu = 2\). The function \(R(\kappa) = \int_{-\infty}^{\infty} \kappa(t)^2 dt\) is the so called roughness function of the kernel. For a \(k\)-dimensional vector, the multivariate density estimation is carried out with a bandwidth vector \(H = (h_1, ..., h_k)'\). It is not difficult to see that \(E[s(W)]\) and \(q_2\) defined in Eq. (4.32) depend on the kernel function used through the functions \(\mu_\nu(\kappa)\) and \(R(\kappa)\).

Using the superscripts ‘G’ and ‘SQ’ to denote the Gaussian and square kernels respectively, Hansen (2009) shows

\[
\mu_{SQ}^\nu(\kappa) = 1/3, \quad R_{SQ}^\nu(\kappa) = 1/2, \\
\mu_G^\nu(\kappa) = 1, \quad R_G^\nu(\kappa) = 1/2\sqrt{\pi}.
\]

(4.36)

In our research, when we substitute the square kernel by the Gaussian kernel, the squared bias-related \(E[s(W)]\) and the variance-related \(q_2\) will change correspondingly. Directly applying Eq. (4.36), we have \(q_G(\kappa)_2 = 3q_{SQ}(\kappa)_2\), \(R_G(\kappa) = R_{SQ}(\kappa)/\sqrt{\pi}\). After plugging this into Eq. (4.15) and performing some calculations, one finds

\[
h^* \approx 0.6 h_{DP}.
\]

(4.37)
Bibliography


124


128


Summary

This dissertation, titled “Multivariate Density Forecast Evaluation and Nonparametric Granger Causality Testing”, is devoted to the study of the evaluation of density forecasts and conditional dependence structure for multivariate time series. Specifically, this dissertation tries to answer the following questions: (1) For a given portfolio of financial assets, what is the most appropriate distribution to assume, supposing there are several candidate distributions, and what is the most relevant modeling method, univariate or multivariate, for the purpose of risk management? (2) How to use simulation-based methodologies to quantify the transfer entropy, and measure conditional dependence between two time-series, if the asymptotic theory of a formal statistical test for conditional independence, i.e. Granger non-causality is missing? (3) Is it possible to develop a robust statistical test for nonparametric Granger non-causality by measuring the transfer entropy; and if so, how does it perform compared to existing tests?

In the first chapter, the motivation of this dissertation is briefly summarized. In particular, time series econometrics has been greatly reshaped in recent decades; the literature has gradually shifted its focus from linear modeling and point forecast to nonparametric modeling and density forecast in many dimensions.

Chapter 2 investigates the first question by comparing multivariate and univariate approaches to assessing the accuracy of competing density forecasts of a portfolio return in the downside part of the support. The common practice to perform multivariate forecast comparisons can be problematic in the context of assessing portfolio risk, since better multivariate forecasts do not necessarily correspond to better aggregate portfolio return forecasts. This is illustrated by examples involving (skew) elliptical distributions and an application to daily returns of a number of US stock prices. Additionally, time-varying test statistics and Value-at-Risk forecasts provide empirical evidence for regime changes over the last decades.

Chapter 3 provides numerical comparisons for simulation procedure-based tests to gain some insights into the statistical behavior of an information theoretic measure — transfer entropy — in the context of detecting conditional independence, i.e. Granger non-causality in a more general sense. Namely, time-shifted surrogates algorithms, the smoothed bootstrap and the stationary bootstrap procedures are presented and compared. By using those proposed resampling techniques, a financial application is provided to illustrate how to detect pairwise
Granger causalities nonparametrically.

Chapter 4 introduces a novel nonparametric test based on a first order Taylor expansion of the transfer entropy. The new test statistic is shown to have an information-theoretical interpretation in terms of Granger non-causality. The proposed test avoids the lack of power of the frequently-used test proposed by Diks and Panchenko (2006), which is not consistent against specific fixed alternatives. Attributed to the U-statistic representation, asymptotic normality of the proposed test statistic is achieved when all densities are estimated with the appropriate sample-size dependent bandwidth. Simulation results confirm the theoretical size and power properties of this test. Finally, two applications to financial data at daily and intra-day frequency conclude this chapter.

Concerning the future work, there are several directions for potential extensions. The method discussed Chapter 2 can be generalized to compare multi-step ahead density forecasts, or to compare nonparametrically estimated densities. Chapter 3 leaves an open question of exploiting entropy-based statistics in testing conditional dependence when there exists a so-called common factor, i.e., looking at the multivariate system containing more than two variables. One potential candidate for this type of test is the partial transfer entropy, coined by Vakorin, Krakovska, and McIntosh (2009). On the theoretical side, it would be practically meaningful to consider causal linkage detection incorporating more lags in a multivariate setting, which requires new asymptotic theory.
Samenvatting (Dutch Summary)

Dit proefschrift, getiteld “Multivariate Dichtheidsvoorspellingsevaluatie en Niet-parametrische Granger Causaliteit” is gewijd aan de studie van de evaluatie van voorspellers en de voorwaardelijke afhankelijkheidsstructuur van multivariate tijdreeksen. In het bijzonder poogt deze dissertatie de volgende vragen te beantwoorden: (1) Gegeven een portefeuille met financiële producten, welke verdeling kan het best worden aangenomen, in gevallen dat er verschillende kandidaat-verdelingen zijn, en wat is de meest relevante modelleermethode, univariaat of multivariaat, voor toepassingen op het gebied van risicomanagement? (2) Kunnen simulatie-gebaseerde methoden worden gebruikt om de overdracht-entropie te quantificeren en de voorwaardelijke afhankelijkheid tussen twee tijdreeksen te meten, gegeven dat de asymptomotische theorie voor een formele toets voor voorwaardelijke afhankelijkheid, d.w.z. Granger causaliteit, vooralsnog ontbreekt? (3) Is het mogelijk om een robuuste statistische toets te ontwikkelen voor afwezigheid van niet-parametrische Granger-causaliteit via het meten van de overdracht-entropie; en zo ja, hoe presteert deze dan in vergelijking met bestaande toetsen?

In het eerste hoofdstuk wordt de motivatie van dit proefschrift gegeven. In het bijzonder is de tijdreekseconometrie behoorlijk veranderd in de recente decennia; de literatuur heeft geleidelijk de aandacht verschoven van lineair modelleren en punctvoorspellingen naar niet-parametrisch modelleren en dichtheidsvoorspellingen in meerdere dimensies.

Hoofdstuk 2 onderzoekt de eerste vraag door multivariate en univariate benaderingen te vergelijken om de nauwkeurigheid van concurrerende dichtheidsvoorspellers te evalueren in de linkerstaart van de verdeling. De gebruikelijke aanpak om multivariate methoden te gebruiken om voorspellingen te vergelijken kan problematisch zijn in de context van het vaststellen van portefeuille-risico, omdat betere multivariate voorspellingen niet noodzakelijkerwijs corresponderen met betere voorspellingen van het geïaggregeerde rendement van de portefeuille. Dit wordt geïllustreerd met voorbeelden van (scheve) elliptische verdelingen en een toepassing op dagelijkse rendementen van aandelen uit de VS. Daarnaast verschaffen tijdsafhankelijke toetsingsgrootheden en Value-at-Risk voorspellingen empirisch bewijs voor regime-veranderingen gedurende de laatste decennia.

In Hoofdstuk 3 worden numerieke vergelijkingen gemaakt van toetsen op basis van simulatie-procedures om inzicht te verkrijgen in het statistische gedrag van een informatietheoretische maat — de overdracht-entropie — in de context van het detecteren van voor-
waardelijke afhankelijkheid, d.w.z. Granger causaliteit in een bredere zin. Namelijk, het tijdverschoven surrogaat-algoritme, de ‘smoothed’ bootstrap en de stationaire bootstrap procedures worden gepresenteerd en vergeleken. Via het gebruik van deze her-bemonsteringstechnieken, wordt een financiële toepassing gegeven die illustreert hoe men op niet-parametrische wijze bivariate Granger causaliteit kan detecteren.


Wat betreft toekomstig werk zijn er verscheidene richtingen voor potentiële uitbreidingen. De methode van Hoofdstuk 2 is uit te breiden naar het vergelijken van voorspellingen over meerdere perioden, of het vergelijken van niet-parametrisch geschatte dichtheden. Hoofdstuk 3 laat een open vraag liggen m.b.t. het exploiteren van entropie-gebaseerde statistische grootheden voor het toetsen van voorwaardelijke afhankelijkheid in het geval dat er een zogenaamde gemeenschappelijke factor is, d.w.z. het bekijken van een multivariaat systeem met meer dan twee variabelen. Een potentiële kandidaat voor dit type toets is de partiële overdracht-entropie, geïntroduceerd door Vakorin, Krakovska en McIntosh (2009). Vanuit praktisch oogpunt zou het wenselijk zijn om causale links te kunnen detecteren met meerdere vertragingen in het multivariate geval, maar dit vereist nieuwe asymptotische theorie.
The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

661. J. CIZEŁ, Essays in Credit Risk, Banking, and Financial Regulation
662. I. MIKOLAJUN, Empirical Essays in International Economics
663. J. BAKENS, Economic Impacts of Immigrants and Ethnic Diversity on Cities
664. I. BARRA, Bayesian Analysis of Latent Variable Models in Finance
665. S. OZTÜRK, Price Discovery and Liquidity in the High Frequency World
666. J. JI, Three Essays in Empirical Finance
667. H. SCHMITTDIEL, Paid to Quit, Cheat, and Confess
668. A. DIMITROPOULOS, Low Emission Vehicles: Consumer Demand and Fiscal Policy
669. G. H. VAN HEUVELEN, Export Prices, Trade Dynamics and Economic Development
670. A. RUSECKAITE, New Flexible Models and Design Construction Algorithms for Mixtures and Binary Dependent Variables
671. Y. LIU, Time-varying Correlation and Common Structures in Volatility
672. S. HE, Cooperation, Coordination and Competition: Theory and Experiment
673. C. G.F. VAN DER KWAAK, The Macroeconomics of Banking
675. F. J.T. SNIEKERS, On the Functioning of Markets with Frictions
676. F. GOMEZ MARTINEZ, Essays in Experimental Industrial Organization: How Information and Communication affect Market Outcomes
677. J. A. ATTEY, Causes and Macroeconomic Consequences of Time Variations in Wage Indexation
678. T. BOOT, Macroeconomic Forecasting under Regime Switching, Structural Breaks and High-dimensional Data
681. Y. GAO, Stability and Adaptivity: Preferences over Time and under Risk
682. M. J. ZAMOJSKI, Panta Rhei, Measurement and Discovery of Change in Financial Markets
683. P. R. DENDERSKI, Essays on Information and Heterogeneity in Macroeconomics
684. U. TURMUNKH, Ambiguity in Social Dilemmas
686. M. LAMMERS, Financial Incentives and Job Choice
687. Z. ZHANG, Topics in Forecasting Macroeconomic Time Series
688. X. XIAO, Options and Higher Order Risk Premiums
689. D. C. SMERDON, ‘Everybody’s doing it’: Essays on Trust, Norms and Integration
690. S. SINGH, Three Essays on the Insurance of Income Risk and Monetary Policy
691. E. SILDE, The Econometrics of Financial Comovement
692. G. DE OLIVEIRA, Coercion and Integration
693. S. CHAN, Wake Me up before you CoCo: Implications of Contingent Convertible Capital for Financial Regulation
694. P. GAL, Essays on the role of frictions for firms, sectors and the macroeconomy
695. Z. FAN, Essays on International Portfolio Choice and Asset Pricing under Financial Contagion
696. H. ZHANG, Dealing with Health and Health Care System Challenges in China: Assessing Health Determinants and Health Care Reforms
697. M. VAN LENT, Essays on Intrinsic Motivation of Students and Workers
698. R. W. POLDERMANS, Accuracy of Method of Moments Based Inference
699. J. E. LUSTENHOUWER, Monetary and Fiscal Policy under Bounded Rationality and Heterogeneous Expectations
700. W. HUANG, Trading and Clearing in Modern Times
701. N. DE GROOT, Evaluating Labor Market Policy in the Netherlands
702. R. E.F. VAN MAURIK, The Economics of Pension Reforms
703. I. AYDOGAN, Decisions from Experience and from Description: Beliefs and Probability Weighting
704. T. B. CHILD, Political Economy of Development, Conflict, and Business Networks
705. O. HERLEM, Three Stories on Influence
706. J. D. ZHENG, Social Identity and Social Preferences: An Empirical Exploration
707. B. A. LOERAKKER, On the Role of Bonding, Emotional Leadership, and Partner Choice in Games of Cooperation and Conflict
708. L. ZIEGLER, Social Networks, Marital Sorting and Job Matching. Three Essays in Labor Economics
709. M. O. HOYER, Social Preferences and Emotions in Repeated Interactions
710. N. GHEBRIHIWET, Multinational Firms, Technology Transfer, and FDI Policy