Essays on Multivariate Modelling of Financial Markets Using Copula and Sentiment Networks

DISSERTATION

of the University of St. Gallen,
School of Management,
Economics, Law, Social Sciences
and International Affairs
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Anastasija Tetereva

from

Latvia

Approved on the application of

Prof. Francesco Audrino, PhD

and

Prof. Ostap Okhrin, PhD

Dissertation no. 4819

Difo-Druck GmbH, Bamberg, 2018

Essays on Multivariate Modelling of Financial Markets Using Copula and Sentiment Networks

DISSERTATION

of the University of St. Gallen,
School of Management,
Economics, Law, Social Sciences
and International Affairs
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Anastasija Tetereva

from

Latvia

Approved on the application of

Prof. Francesco Audrino, PhD

and

Prof. Ostap Okhrin, PhD
Prof. Dr. Michael Lechner

Dissertation no. 4819

Difo-Druck GmbH, Bamberg, 2018

The University of St. Gallen, School of Management, Economics, Law, Social Sciences and International Affairs hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St. Gallen, June 15, 2018

The President:

Prof. Dr. Thomas Bieger

Acknowledgements

Writing a doctoral thesis is always demanding and challenging. I would not have been able to complete this work without the motivation and support of many people whom I would like to acknowledge here.

I am deeply indebted to Prof. Dr. Francesco Audrino for taking me on board and affording me the opportunity to pursue a PhD under his supervision. I would like to express my deep gratitude for his patient guidance, enthusiastic encouragement and useful critiques of my research work. I am also grateful to Prof. Dr. Ostap Okhrin, who offered me much valuable advice in the early stages of this work. Moreover, I owe an important debt to Prof. Dr. Janis Valeinis for his warm encouragement.

I am thankful to my current and former fellow PhD students (in no particular order): Alexander, Wale, Constantin, Marcial, Simon, Pirmin, Tatiana, Katja, Thomas, Anselm, Yauhen, Dominik, Daniele. Many of them have become dear friends of mine. Sincere thanks also to my current and former colleagues at the faculty of mathematics and statistics, especially Prof. Dr. Juan-Pablo Ortega, Prof. Dr. Matthias Fengler, Prof. Dr. Lorenzo Camponovo and, Dr. Stefan Ott. I benefited a great deal from their interesting and helpful discussions regarding research and beyond. A special acknowledgement is necessary for Margit and Fadrina for their continued effort.

I am grateful to all the participants of the following workshops for their valuable suggestions and discussions: the Salzburg Workshop on Dependence Models & Copulas; the 10th Annual SoFiE Conference; the CFE-CMStatistics Conferences; European Meeting of Statisticians; Text, Herding and Sentiment workshop; and the 11th Annual SoFiE Conference.

My special thanks are extended to Thomson Reuters and the University of St. Gallen for providing the news analytics data.

Last, but definitely not least, I would like to acknowledge the support of my family and friends. Their love provided me with the necessary strength and motivation.

Anastasija Tetereva

St. Gallen, 2018

Summary

Multivariate dependence structures play an important role in finance. The modelling and accurate prediction of multivariate financial time series is an important component of asset pricing and portfolio management. This doctoral thesis comprises three essays that address the question of multivariate dependencies using high-frequency data and innovative sources of information such as news analytics. These essays make complementary contributions to the field of financial econometrics and can be read independently of each other.

The first essay focuses on the improvement of Value at Risk prediction based on high-frequency data. The novel concept of the realized hierarchical Archimedean copula is introduced. It is proposed estimating the structure and the parameters of the hierarchical Archimedean copula using the realized correlation matrix only. This approach allows one to estimate the multivariate distribution of daily returns based on intraday information. Moreover, the proposed estimator does not suffer from the curse of dimensionality. In this essay, the realized hierarchical Archimedean copula is applied to manage the risk of high-dimensional portfolios. The evidence of the superior forecasting power of our approach, compared to a set of existing models, is provided.

The second essay investigates the role of news sentiment data in improving forecasts in financial econometrics. The objective of this paper is to answer the question regarding whether the class of stock-price-relevant news is wider than firm-specific announcements. For this purpose, causal links between news sentiments and excess returns are studied by means of an adaptive lasso. It is concluded that unexpected returns in the whole economy can be explained by news originating from the financial and energy sectors. In other words, the news spillover effects are dominating the direct effects of sectoral news. Therefore, including exogenous financial or energy sentiment variables in econometric models can significantly improve forecasting properties.

The third and final essay extends the ideas presented in the second essay along several lines. First, it analyses the mutual relationship amongst financial news at the firm level. Second, it exploits the news data of higher granularity than weekly or daily. In this paper, the occurrences of firm-specific news announcements are considered to follow the Hawkes process. This approach provides a tool to identify systemically important financial companies in terms of news. Based on this information, the novel composite news intensity index is constructed. It is empirically demonstrated that the proposed index provides early warning signals of market instability.

Zusammenfassung

Multivariate Abhängigkeitsstrukturen spielen eine wichtige Rolle im Finanzwesen. Die Modellierung und genaue Vorhersage multivariater Finanzzeitreihen ist ein wichtiger Bestandteil des Asset Pricings und des Portfoliomanagements. Diese Doktorarbeit umfasst drei Aufsätze, die sich mit der Frage von multivariaten Abhängigkeiten befassen, indem Hochfrequenzdaten und innovative Informationsquellen, wie zum Beispiel die Nachrichtenanalytik, genutzt werden. Die drei Studien leisten ergänzende Beiträge zum Gebiet der Finanzökonometrie und können unabhängig voneinander gelesen werden.

Der erste Aufsatz konzentriert sich auf die Verbesserung der Vorhersage des Value at Risks auf Basis von Hochfrequenzdaten. Dafür wird das neuartige Konzept der realisierten hierarchischen archimedischen Kopula eingeführt. Es wird vorgeschlagen, die Struktur und die Parameter der hierarchischen archimedischen Kopula nur durch den Gebrauch der realisierten Korrelationsmatrix zu schätzen. Dieser Ansatz erlaubt es, die multivariate Verteilung täglicher Renditen basierend auf Intraday-Daten zu ermitteln. Außerdem leidet die vorgeschlagene Schätzfunktion nicht unter dem Fluch der Dimensionalität. In diesem Aufsatz wird die realisierte hierarchische archimedische Kopula angewendet, um das Risiko von hoch-dimensionalen Portfolios zu kontrollieren. Wir zeigen, dass unser Ansatz im Vergleich zu existierenden Modellen, eine höhere Vorhersagekraft besitzt.

Jüngste Studien im Finanzwesen behaupten, dass Börsenkurse oftmals mehr von der Stimmung der Investoren als von der Realität beinflusst werden. Der zweite Aufsatz untersucht die Rolle des auf Nachrichten basierenden Sentiments bei der Verbesserung der Vorhersagen in der Finanzökonometrie. Das Ziel dieses Artikels ist es, die Frage zu beantworten, ob die Klasse der börsenkursrelevanten Nachrichten umfangreicher ist als firmenspezifische Meldungen. Zu diesem Zweck werden kausale Verbindungen zwischen den Sentiments der Nachrichten und Überschussrenditen mittels der ökonometrischen Methode des Adaptive Lassos untersucht. Basierend auf Thomson Reuters Nachrichtendaten kommt die Studie zum Ergebnis, dass unerwartete Renditen in der ganzen Wirtschaft durch Nachrichten aus dem Finanz- und Energiesektor- erklärt werden können. Anders gesagt dominieren Auswirkungen branchenfremder Nachrichten direkte Auswirkungen der branchenspezifischen Nachrichten. Daher kann das Einbeziehen exogener Sentimentvariablen aus dem Finanz- und Energiesektor- in ökonometrische Modelle die Vorhersageeigenschaften ökonometrischer Modelle wesentlich verbessern.

Der dritte und letzte Aufsatz erweitert die im zweiten Aufsatz vorgestellten Ideen auf mehreren Ebenen. Zunächst analysiert er die wechselseitige Beziehung zwischen Finanznachrichten auf Firmenebene. Zweitens werden Nachrichtendaten von höherer Granularität als wöchentlich oder täglich ausgenutzt. In diesem Kapitel nehmen wir an, dass das Auftreten von firmenspezifischen Nachrichtenmeldungen einem Hawkes Prozess folgt. Genauer gesagt, wird die Granger Kausalität der Nachrichtenmeldungen mittels multivariater Hawkes Graphen modelliert. Dieser Ansatz stellt ein Instrument zur Verfügung, welches systemisch wichtige Finanzunternehmen bezüglich Nachrichten identifiziert. Basierend auf diesen Informationen wird ein neuer zusammengesetzter Nachrichtenintensitätsindex konstruiert. Es wird empirisch nachgewiesen, dass der vorgeschlagene Index frühe Warnsignale auf Marktinstabilitäten liefert.

Contents

List of Figures

List of Tables

Preface

"Essays on Multivariate Modelling of Financial Markets Using Copula and Sentiment Networks" is a set of three essays at the intersection of the multivariate modelling of complex time-varying dependencies and finance. The modelling and predicting of the dependencies of financial time series plays a prominent role in financial econometrics. This task is of primary importance when making investment decisions, performing optimal asset allocation and managing portfolio risk. During the last decades, the availability of innovative sources of data and affordable computing power influenced the ways in which multivariate modelling is addressed in finance. On the one hand, the frequency of observations is referred to as the measure of progress in financial econometrics; nowadays, high-frequency data and the realized measures of dependence are successfully incorporated into econometric models with the aim of predicting the future and gaining significant profits. On the other hand, text analytics tools have undergone substantial changes in recent years, and they have made the news sentiment data available for econometricians. The focus of this work is to provide the tools for the improvement of the forecasting power of econometric models by means of high-frequency and news sentiment data. This doctoral thesis comprises three essays (chapters) that share the feature of an application of multivariate statistical techniques to better understand the nature of financial markets.

Chapter 1 is based on the published paper by Okhrin and Tetereva (2017), in which the authors introduced the concept of the realized hierarchical Archimedean copula. Copula is a flexible tool that is applied to model complex multivariate dependencies that appear in all areas of human activities. In financial markets, copula models are helpful to manage the risk and price financial derivatives. The flexibility of the copulae is due to the fact that the multivariate distribution can be decomposed into marginal distributions and copula. Okhrin and Tetereva (2017) present research at the intersection of two increasingly popular areas of financial econometrics: copula and high-frequency data. Incorporating high-frequency data into copula models improves their time-variability and forecasting power in short-term risk management. This work contributes to the existing copula literature and suggests estimating the structure and the parameters of the hierarchical Archemedean copula nonparametrically using the realized correlation matrix only. This estimator can be applied to construct a

novel model of high-dimensional realized copula. The computational costs of the proposed estimator are low, and it does not suffer from the curse of dimensionality. The properties of the estimator, in comparison with the benchmarks, are discussed by means of Monte Carlo simulations. It is demonstrated that this estimator is preferable in many cases despite its parsimonious construction. The advantage of this estimator in risk modelling is the possibility to estimate the copula of the daily returns based on high-frequency data. It is important to mention here that, in general, the multivariate distribution of daily returns does not coincide with the distribution of intraday returns. The introduction of the realized hierarchical Archimedean copula allows one to estimate the copula daily and to improve the accuracy of the forecasts. The proposed realized hierarchical Archimedean copula outperforms many competing models, including Bauer and Vorkink (2011) and Salvatierra and Patton (2015) in terms of the computational speed and accuracy of the one-day-ahead Value at Risk prediction. The results of the paper demonstrate that high-frequency observations can be successfully incorporated into copula models, allowing one to capture time-varying high-dimensional dependencies with higher accuracy.

Chapter 2, which is co-authored by Francesco Audrino, investigates the impact of sentiment spillover effects on the excess returns in the US and European markets. In financial econometrics, it is usually assumed that all the relevant information about the current price of the asset is contained in the historical prices. However, empirical evidence during the times of financial instability contradicts this assumption. As stated in Shiller (1999), "Stock prices in the United States, when compared with measures of the true fundamental value or sensible investment value, are too high, too low, or about right". For this reason, researchers and practitioners have been exploring additional sources of information that would be able to explain the origins of economic shocks and increasing volatility. Rapidly increasing computational power and the development of text- and data-mining techniques made it possible to augment financial time series models by news data. Given that these models demonstrate superior forecasting power, research on news sentiment analysis has attracted much attention. The majority of research on sentiment analysis is devoted to the direct causality of the news; in other words, previous studies have primarily focused on how the news about a particular company influences that company's stock returns and volatility. The main objective of the second essay is to characterise the news spillover effects and assess whether these effects dominate the direct effects. To answer this question, adaptive lasso testing procedure for large data sets is applied to construct a dynamic news/returns network. New econometric characteristics are provided to infer the causality between news that originates from one sector and the stock returns of the companies in other sectors. By

analysing Thomson Reuters' news sentiment data, the dynamics of these characteristics are studied for a period of almost 10 years. This chapter finds empirical evidence of news that originates from just a few sectors being important for the returns in the whole economy. In addition, it is observed that the causality of the news increased just before periods of economic turbulence. The superior performance of ARMA-GARCH models augmented by sentiment data from the financial and energy sectors is observed in the prediction comparison. The results of this chapter are beneficial for economists and financial risk managers, and they have great potential for other applications.

In Chapter 3, the mutual contagion of financial news is explored. The results of Chapter 2 suggest that financial news drives the excess returns in the whole market. For this reason, more research should be conducted on the mutual dependencies of financial announcements. The last chapter of the thesis attempts to answer the question regarding whether the intensity of financial news can be considered to be a self- and mutually exciting process. To this end, the multivariate Hawkes graphs and the corresponding nonparametric estimation procedure by Kirchner (2016) is employed. The network representation of the branching structure of the multivariate Hawkes process offers a compact way in which to describe the contribution of each company to the information flow. Another advantage of this approach is the possibility of studying real-time news event data without aggregating them and therefore more precisely identifying the causal links between announcements. From the conducted analysis, it can be concluded that the news shocks from systemically important companies trigger the announcements related to other companies. Moreover, it is concluded that the mutual causality of news arrival times increases during times of economic turbulence. The systemic importance of US and UK financial institutions, UBS and Deutsche Bank is observed. It is empirically demonstrated that the contribution of a company to the information flow is not always proportional to the intensity of firm-specific news and the size of the company. Based on the results, a composite news intensity index is constructed for the US market. The estimated Hawkes graphs suggest uniquely defining the weights of the companies in the composite index. In contrast to the sentiment indices discussed in the literature, the proposed measure does not involve the construction of a sentiment score and therefore does not require bag-of-words techniques to be applied. Compared to the existing fear index, it provides a timelier signal of uncertainty in the market, and it Granger causes VIX, S&P 500 price and volume at a time lag of 6 months. Therefore, policy makers and practitioners in the field can successfully use the proposed news intensity index.

Chapter 1

The realized hierarchical Archimedean copula in risk modelling (published in *Econometrics*)

Ostap Okhrin [1], **Anastasija Tetereva** [2]

[1]Chair of Econometrics and Statistics esp. Transportation, Institute of Economics and Transport, Faculty of Transportation, Dresden University of Technology, Helmholtzstraße 10, 01069 Dresden, Germany, ostap.okhrin@tu-dresden.de

[2]Chair of Mathematics and Statistics, University of St Gallen, Bodanstrasse 6, 9000 St Gallen, Switzerland, anastasija.tetereva@unisg.ch

Abstract

This paper introduces the concept of the realized hierarchical Archimedean copula (rHAC). The proposed approach inherits the ability of the copula to capture the dependencies among financial time series, and combines it with additional information contained in high-frequency data. The considered model does not suffer from the curse of dimensionality, and is able to accurately predict high-dimensional distributions. This flexibility is obtained by using a hierarchical structure in the copula. The time variability of the model is provided by daily forecasts of the realized correlation matrix, which is used to estimate the structure and the parameters of the rHAC. Extensive simulation studies show the validity of the estimator based on this realized correlation matrix, and its performance, in comparison to the benchmark models. The application of the estimator to one-day-ahead Value at Risk (VaR) prediction using high-frequency data exhibits good forecasting properties for a multivariate portfolio.

1.1 Introduction

One of the main objectives of quantitative research is the modelling and approximation of multivariate distributions. A multivariate model should be flexible enough to capture the stylized facts of empirical finance. Moreover, increasing interest in short-term quantitative risk management requires the time-variability of such models. The current paper builds on two actively developing areas of financial econometrics: copulae and high-frequency data. On the one hand, copulae appear to be a helpful tool to analyse complex dependence structures, evaluate the risk, and are therefore widely used to price financial derivatives, see Embrechts et al. (2003), Rodriguez (2007), Hofert and Scherer (2011), Krämer et al. (2013). On the other hand, models based on high-frequency data yield superior predictions in comparison to approaches based on daily data. Among others, Andersen et al. (2002), Barndorff-Nielsen and Shephard (2004) and Zhang et al. (2005) made it possible to compute the daily realized covariances from high-frequency data. Many researchers have implemented the obtained realized measures to model financial time series. Most of those studies, however, employ models where the realized correlation matrix directly characterizes the multivariate distribution, see, for example, Bauer and Vorkink (2011), Chiriac and Voev (2011), Jin and Maheu (2012), or address GARCH type models, for example, Hansen et al. (2014), Bauwens et al. (2012), Noureldin et al. (2012) and Bollerslev et al. (2016). There are only a limited number of studies which discuss the implementation of high-frequency data in copula models. Breymann et al. (2003) and Dias et al. (2004) employ copulae to study the properties of intraday log-returns. Creal et al. (2013) consider an autoregressive updating equation and improve the predictive power in Salvatierra and Patton (2015) by including the lagged realized volatility in the equation.

To the best of our knowledge, the only model that parameterizes the whole Archimedean copula (AC) by the realized variance-covariance matrix is in Fengler and Okhrin (2016), who introduced the realized copula. The authors suggested capturing time-varying dependence by using high-frequency intraday data to estimate the parameter of an AC daily. It has been demonstrated empirically that the realized copula model outperforms the list of benchmark models in one-day-ahead out-of-sample VaR prediction. The realized copula model of Fengler and Okhrin (2016) has, however, several limitations. First, their realized copula is driven by one single parameter, which limits the flexibility of the model. Second, the estimation procedure is performed by applying a method of moments kind of estimator, which suffers from the curse of dimensionality.

We propose to extend the work of Fengler and Okhrin (2016) by introducing the realized hierarchical Archimedean copula (rHAC), which allows more flexibility and is applicable to managing high-dimensional portfolios. We adapt the estimation procedures described in Segers and Uyttendaele (2014) and Górecki et al. (2016a) to high-frequency data, which allows estimating the structure and the parameters of a copula based only on a realized covariance matrix. As a result, the estimate does not suffer from microstructure noise or jumps. Moreover, it can be applied to high-dimensional portfolios since the computationally expensive optimization procedure proposed in Fengler and Okhrin (2016) is reduced to a set of simple tasks. This result is of particular importance in many financial applications, especially in risk management.

This paper is structured as follows. Section 1.2 contains a literature review of the theory of the copula and introduces the concept of a realized copula. An estimator of the structure and the parameters of an rHAC is presented in Section 1.3. Simulation studies and a comparison with the benchmark models are provided in Section 1.4. Section 1.5 discusses the construction of the rHAC, and gives a short summary of competing models. Section 1.6 describes an application of the proposed models to one-day-ahead VaR prediction for a multidimensional portfolio. Finally, we summarize the main contribution of the paper.

1.2 The concept of the realized copula

The concept of the copula was introduced to the statistical literature by Sklar (1959) and further popularized in the world of finance by Embrechts et al. (1999) in the context of risk management. Sklar's theorem, see Sklar (1959), states that a d-dimensional distribution function $F(x_1, \ldots, x_d)$ with marginals $F_1, \ldots, F_d$ can be represented as

$$F(x_1, \ldots, x_d) = C_d\{F_1(x_1), \ldots, F_d(x_d)\}, \tag{1.1}$$

where $C_d(u_1, \ldots, u_d)$ is a d-dimensional copula. In addition, it states that the continuity of the marginal distributions $F_1, \ldots, F_d$ ensures the uniqueness of the copula.

Having a huge number of classes of bivariate copulae, see Nelsen (2007), there is still a lack of multivariate ones. The most popular classes of multivariate copulae currently are elliptical, factor, pair-copula constructions, and HAC. The first class is often used in practice due to its simplicity and intuitive interpretation. However, elliptical copulae are not able to capture the stylized facts observed in financial data. The factor approach overcomes this limitation and has attracted attention in the copula literature over the last decade,

see, for example, Andersen and Sidenius (2004), van der Voort (2007), Krupskii and Joe (2013), and Oh and Patton (2017). The limitation of the factor copula models is that the likelihood function is often not known in closed form, which complicates the estimation of the parameters. Pair-copula constructions are discussed in more detail by Joe (1996), Bedford and Cooke (2001), Czado (2010), and Kurowicka (2011), and are increasing in popularity. Another popular copula class is the Archimedean copulae (AC), which contains, among others, the Clayton, Gumbel and Frank copulae. The AC parametrized by the parameter θ is defined as $C_d(u_1,\ldots,u_d;\theta) = \psi_\theta\{\psi_\theta^{[-1]}(u_1)+\ldots+\psi_\theta^{[-1]}(u_d)\}$, $u_1,\ldots,u_d \in [0,1]$ with $(-1)^j\psi_\theta^{(j)}(t) \geq 0$ being non-decreasing and convex on $[0,\infty)$ for $t>0$, where $j \in$. $\psi_\theta(0)=1$, $\psi_\theta(\infty)=0$ and the pseudo inverse is defined as $\psi_\theta^{[-1]}(t)=\psi_\theta^{-1}(t)$ for $0 \leq t \leq \psi_\theta(0)$ and 0 otherwise. The generators and the densities of some AC are given in Appendix 1.A.

Due to the lack of flexibility of AC, caused by the fact that the whole copula is driven by just one parameter θ, generalizations such as nested copulae have been introduced. This paper employs a flexible multivariate copula family, the hierarchical Archimedean copulae (HAC), a special case of which may be defined recursively in the following way:

$$\psi_{\theta_{d-1}}\left\{\psi_{\theta_{d-1}}^{[-1]}(u_d)+\psi_{\theta_{d-1}}^{[-1]}\circ C_{d-1}\left(u_1,\ldots,u_{d-1};s_{d-2},(\theta_1,\ldots,\theta_{d-2})^\top\right)\right\}, \tag{1.2}$$

where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_{d-1})^\top$ is the parameter vector of the HAC and s is the structure of the HAC. As is evident from (1.2), the current study assumes that all generators of the HAC belong to the same parametric family and each of them depends on one single parameter. For simplicity, we compress the notation of (1.2) and denote the d-dimensional HAC with k generators which is parametrized by the structure s and the parameter vector $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_k)^\top$ as $C_d(u_1,\ldots,u_d;s,\boldsymbol{\theta})$. The structure s is the merging ordering $s=(\ldots(qr)s\ldots)$, where $q,r,s \in 1,\ldots,d$, $q\neq r\neq s$ is a reordering of the indices of the variables X_i, $i=1,\ldots,d$. The structure of a d-dimensional HAC s can be seen as a tree with $k \leq d-1$ non-leaf nodes that correspond to the generators and d leaves representing the variables $\mathcal{X}=(X_1,X_2,\ldots,X_d)^\top$. The leaves correspond to the lowest level of the tree. The root corresponding to the variable $C_d(u_1,\ldots,u_d;s,\boldsymbol{\theta})$ is assumed to be the highest level of the tree. The nodes, which are not the leaves are called internal nodes, each corresponds to the generator. A node which is directly connected to another node when moving away from the root is called the child node. A node which is directly connected to another node when moving from the leaves to the root is called the parent node. Descendants are the children nodes of the node, children of these children, etc. The set of ancestors includes the parent node of the node, parents of the parents, etc. The structure of the HAC is called binary if it corresponds to the binary

tree, i.e. if each internal node has exactly two children. Further on, we denote the nodes associated with the generators by $D_{\mathcal{X}_i}$, where $\mathcal{X}_i$ is the set of leaves (variables) that are descendant nodes of the node $D_{\mathcal{X}_i}$, $i = 1, \ldots, k$. Assuming this notation, the node $D_{\mathcal{X}_i}$ is an ancestor of the node $D_{\mathcal{X}_j}$ (the leave associated with the variable X_l) if $\mathcal{X}_j \subset \mathcal{X}_i$ ($X_l \subset \mathcal{X}_i$), $l = 1, \ldots, d$, $i, j = 1, \ldots, k$. Another concept that will be used later on is the concept of the lowest common ancestor (lca). The lca of the nodes $D_{\mathcal{X}_i}$ (the leave X_q) and $D_{\mathcal{X}_j}$ (the leave X_r) is the node $D_{\mathcal{X}_l}$ that is the lowest node satisfying $\mathcal{X}_i \subset \mathcal{X}_l$ ($X_q \subset \mathcal{X}_l$) and $\mathcal{X}_j \subset \mathcal{X}_l$ ($X_r \subset \mathcal{X}_l$), $q, r = 1, \ldots, d$, $i, j, l = 1, \ldots, k$.

To clarify the above-mentioned definitions and avoid introducing the comprehensive notation of the graph theory, we illustrate the above-named concepts by an example. Consider the 5-dimensional copula

$$\psi_{1.5}\Big\{\psi_{1.5}^{[-1]}\Big(\psi_2\big[\psi_2^{[-1]}\{\psi_4\big(\psi_4^{[-1]}(u_1)+\psi_4^{[-1]}(u_2)\big)\}+\psi_2^{[-1]}\{\psi_{2.5}\big(\psi_{2.5}^{[-1]}(u_3)+\psi_{2.5}^{[-1]}(u_4)\big)\}\big]\Big)+\psi_{1.5}^{[-1]}(u_5)\Big\}$$

that can be written as $C_5\left(u_1, u_2, u_3, u_4, u_5; s = ((12)(34)5), \boldsymbol{\theta} = (4, 2.5, 2, 1.5)^\top\right)$, where $u_i = F_i^{-1}(x_i, \nu_i)$ with ν_i being the parameters of the marginal distributions $F_i(\cdot)$, $i = 1, \ldots, 5$. The tree corresponding to this copula is presented in the Figure 1.1. This copula has the binary structure $s = ((12)(34)5)$. There are $k = 4$ non-leaf (internal) nodes. The leaves which correspond to the lowest level of the copula tree are given by the variables X_1, X_2, X_3, X_4 and X_5. The root $D_{\mathcal{X}_4}$ which represents the highest level of the copula tree corresponds to the variable $C_5(u_1, u_2, u_3, u_4, u_5; s, \boldsymbol{\theta})$, where $\mathcal{X}_4 = (X_1, X_2, X_3, X_4, X_5)^\top$. The root node is the parent node for the node corresponding to the variable X_5 and the node $D_{\mathcal{X}_3}$ associated with the variable generated by $C_4\left(u_1, u_2, u_3, u_4; s = (12)(34), \boldsymbol{\theta} = (4, 2.5, 2)^\top\right)$, where $\mathcal{X}_3 = (X_1, X_2, X_3, X_4)^\top$. The root node is the ancestor for all other nodes of the given copula tree. The lca of the nodes associated with the variables X_1 and X_2 is the node $D_{\mathcal{X}_1}$ that corresponds to the variable $C_2(u_1, u_2; s = (12), \boldsymbol{\theta} = 4)$, where $\mathcal{X}_1 = (X_1, X_2)^\top$. The lca of the nodes corresponding to the variables X_1 and X_5 is the root node $D_{\mathcal{X}_4}$ as it is the lowest node satisfying $X_1 \subset \mathcal{X}_l$ and $X_5 \subset \mathcal{X}_l$, $l = 1, \ldots, d$.

Although copula models are flexible enough to capture nonlinear dependencies, many empirical applications require the time variability of the parameters (and the structure) of the whole copula. For example, the empirical evidence makes it reasonable to assume that the dependence between asset log-returns gets stronger during periods of financial turbulence. A vast amount of literature is devoted to dynamic copula models, including the parsimonious rolling window approach and more sophisticated models, such as, for example, the local change point procedure of Härdle et al. (2013). Recent developments in time-varying copula models take advantage of the rapidly growing availability of high-frequency observations and include the realized measures (volatility and correlations) in

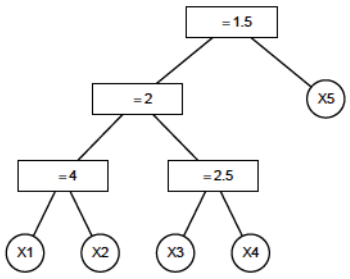


Figure 1.1 A 5-dimensional copula structure.

the copula models to improve their predictive power, see, for example, Salvatierra and Patton (2015). The improvement is obtained due to the fact that the actual realizations of the volatility of log-returns which are not directly observable can be estimated by the sum of finely-sampled squared realizations of log-return over a fixed time interval when the high-frequency observations are available. Such a nonparametric ex-post measurement of the log-return variation is called the realized volatility. In an analog manner, the realized covariances are defined by summing all the cross products of intraday log-returns. The formal definition of the realized measures is given in Appendix 1.B. Despite the constantly growing research on incorporating the realized measures into multivariate Gaussian models, discussed in Chiriac and Voev (2011) and Bauer and Vorkink (2011), and into GARCH type models, for example, Hansen et al. (2014) and Bollerslev et al. (2016), there is still a gap in the literature on how the parameters of non-Gaussian copula can be estimated daily based on high-frequency observations. It is important to note here that such standard copula estimation techniques as the Maximum Likelihood (ML) method or the inversion of Kendall's τ can not be directly applied to tick-by-tick observations. Estimating the copula by applying these approaches to high-frequency data would estimate the multivariate distribution of high-frequency log-returns, which in general does not coincide with the multivariate distribution of daily log-returns. Such a model would estimate the intraday dependence and produce the forecast of the multivariate distribution of log-returns in the next second and could not be used for one-day-ahead VaR forecasts. For further details on the standard estimation procedures, refer to Nelsen (2007), Trivedi et al. (2007), Jaworski et al. (2013), Cherubini et al. (2011), Joe (2014) and Durante and Sempi (2015). In contrast to the direct application of the ML approach to tick-by-tick data or high-frequency estimator of Kendall's τ, there is a considerable literature discussing how to estimate the

correlation matrix of daily log-returns via a realized correlation matrix or similar methods, see Barndorff-Nielsen et al. (2004), Barndorff-Nielsen and Shephard (2004), Zhang et al. (2005), Hayashi et al. (2005), and Pooter et al. (2008). The idea of using the information concentrated in the realized covariance matrix to estimate the parameters of a copula daily has been employed by Fengler and Okhrin (2016), who used a combination of the results from a lemma of Hoeffding (1940) and Sklar's theorem (1.1) to express the covariance σ_{ij} between two random variables X_i and X_j in terms of the marginal distributions $F_i(\cdot)$ and $F_j(\cdot)$ and the copula $C_2(\cdot,\cdot;\theta)$

$$\begin{aligned}\sigma_{ij}(\theta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{F_{i,j}(x,y;\theta,\nu_i,\nu_j) - F_i(x;\nu_i)F_j(y;\nu_j)\}\,\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[C_2\{F_i(x;\nu_i),F_j(y;\nu_j);\theta\} - F_i(x;\nu_i)F_j(y;\nu_j)\right]\mathrm{d}x\mathrm{d}y;\, i,j=1\ldots d,\end{aligned} \tag{1.3}$$

where θ is the parameter of the copula and ν_i, ν_j are the parameters of the marginal distributions $F_i(\cdot)$ and $F_j(\cdot)$. In the high-frequency framework, the covariance σ_{ij} in (1.3) is replaced by the element $r_{ij,t}$ of the realized covariance matrix R_t computed at day t. From now on, we denote the diagonal elements of matrix R_t by $r_{i,t}$ instead of $r_{ii,t}$, $i=1,\ldots,d$. As has been shown in Breymann et al. (2003) and discussed in more detail in Hautsch (2011), with an increasing sampling frequency, the marginal distributions of log-returns can be assumed to be Gaussian with zero mean and the standard deviation equal to $\sqrt{r_{i,t}}$, $t=1,\ldots,d$, this leads us to assume throughout this study that margins are $\mathrm{N}(0,r_{i,t})$. Thus, if the realized covariance matrix R_t can be computed, according to Fengler and Okhrin (2016), it can be assumed that for the Archimedean copula driven by one single parameter θ the integral in (1.3) depends on just the parameter of the copula which belongs to some parametric family $\mathcal{C}=\{C(\cdot;\theta),\theta\in\Theta\}$. Therefore, after replacing the covariances in (1.3) by their realized counterparts and standardizing the variables, the expression (1.3) can be rewritten for the realized correlations as

$$\begin{aligned}\rho_{ij,t} &= f(\theta_{ij,t}) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[C_2\{\Phi(x),\Phi(y);\theta_{ij,t}\} - \Phi(x)\Phi(y)\right]\mathrm{d}x\mathrm{d}y;\, i,j=1\ldots d, i\neq j,\end{aligned} \tag{1.4}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution and $\rho_{ij,t}=\frac{r_{ij,t}}{\sqrt{r_{i,t}\cdot r_{j,t}}}$ is the element of the realized correlation matrix $\mathcal{P}_t$ calculated at day t, $t=1,\ldots,T$. According to (1.4), the realized correlations depend solely on the copula parameter, under the assumption of some parametric family. Based on (1.4), the parameter of the copula can be estimated based on

just the realized correlation matrix:

$$\widehat{\theta}_t = \underset{\theta}{\operatorname{argmin}}\; g_t^{\top}(\theta)\, W\, g_t(\theta), \tag{1.5}$$

where $g_t(\theta)$ is a vector of length $\frac{d(d-1)}{2}$ where all the $g_{ij,t}(\theta) = \rho_{ij,t} - f(\theta)$ are stacked together and W is a $\left(\frac{d(d-1)}{2} \times \frac{d(d-1)}{2}\right)$-dimensional positive definite weighting matrix. When the copula parameter is estimated from (1.5) and the diagonal elements of the realized covariance matrix R_t are calculated, the multivariate distribution of $\mathcal{X} = (X_1, X_2, \ldots, X_d)^{\top}$ is fully specified. It is important to note that Fengler and Okhrin (2016) consider the restrictive setting of AC. Therefore, all bivariate copulae in (1.4) coincide and are driven by one single parameter θ.

In practice, one is usually interested in predicting a multivariate distribution, rather than just estimating it. This can be done in two ways. The parameter of the realized copula can be estimated daily and predicted using some time-series model. Alternatively, the realized correlation matrix can be predicted and the parameter of the copula can be estimated from $\widehat{\mathcal{P}}_{t+1|t}$, which is one-day-ahead prediction of the realized correlation matrix $\mathcal{P}_{t+1}$ obtained by applying the specific time series model in the spirit of Bauer and Vorkink (2011) or Chiriac and Voev (2011). The limitation of both approaches comes from the estimation procedure (1.5), which suffers from the curse of dimensionality and enables the estimation of the realized copula only in moderate dimensions. Moreover, as was mentioned earlier, the whole realized copula in Fengler and Okhrin (2016) is driven by just one parameter θ, which might be too restrictive for multivariate portfolios.

We propose to overcome these limitations by using the HAC instead of the simple AC. This extension is not straightforward, as in addition to the parameter vector $\boldsymbol{\theta}$ of $C_d(u_1, \ldots, u_d; s, \boldsymbol{\theta})$, the structure of the copula s needs to be estimated. The estimation of the parameter vector $\boldsymbol{\theta}$ of a d-dimensional copula $C_d(u_1, \ldots, u_d; s, \boldsymbol{\theta})$ should be addressed as well. The procedure of Fengler and Okhrin (2016) allows the estimation of the parameters at the bottom level of the copula. The estimation of the parameters of the higher levels is not trivial, as the realized correlation among the original variables and the variables determined by the copulae of the bottom levels can not be specified. This motivates the estimation of the structure and the parameters of the hierarchical copula based just on the realized correlation matrix. Recent studies in the copula literature address the question of how the structure (or the structure and the parameters) of a hierarchical copula can be estimated based on Kendall's τ correlation matrix, see, for example, Segers and Uyttendaele (2014) Górecki et al. (2016a), Uyttendaele et al. (2016), and Górecki et al. (2016b). We propose

to combine the methods discussed in Segers and Uyttendaele (2014) and Górecki et al. (2016a) and adapt them to the realized correlation matrix with the final goal of improving one-day-ahead VaR prediction for multivariate portfolios.

1.3 Estimating the realized hierarchical Archimedean copula

This section discusses how the structure and the parameters of an HAC can be estimated based on the realized correlation matrix $\mathcal{P}_t$ only. From now on, we refer to such a copula as an rHAC. In this section, the subindex t is dropped to simplify the notation. We suggest generalizing the clustering method proposed by Górecki et al. (2016a) by applying an adaptation of the algorithm introduced in Segers and Uyttendaele (2014) in order to estimate the structure of an HAC. Consequently, the parameters can be estimated by applying (1.4) to the specific average of the realized correlations. We restrict ourselves to the case when all the generators of the copula belong to the same Archimedean family and satisfy the nesting condition. A brief discussion of this will be provided later in this section.

1.3.1 Estimating the structure

In analog to the method mentioned in Górecki et al. (2016a) for Kendall's τ, we suggest defining the distance between two variables X_i and X_j as

$$h_{ij} = 1 - \rho_{ij}, \tag{1.6}$$

where ρ_{ij} is the realized correlation between X_i and X_j, $i, j = 1, \ldots, d$. Next, the dependence-based distance matrix is used as the input for an agglomerative cluster analysis. The obtained hierarchical clustering dendrogram corresponds to the estimated structure of the HAC. This approach is, however, valid only for HACs with binary (bivariate) structure. The introduction of an additional merging parameter that allows collapsing a binary structure into a general one is discussed in Uyttendaele et al. (2016). The optimal choice of such a parameter still needs to be addressed in the literature. To reduce the computational costs, we will adapt the method proposed in Segers and Uyttendaele (2014) to the distance (1.6) to recover the general structure of an rHAC.

Segers' and Uyttendaele's algorithm According to Segers and Uyttendaele (2014), the structure of a nested HAC s can be uniquely recovered from the structures of the set of $\binom{d}{3}$

triples (X_q, X_r, X_s) with distinct $q, r, s = 1, \ldots, d$ using the concept of the lowest common ancestor (lca). According to the definition given in Section 1.2, the lca of X_q and X_r is the node which is the lowest node that has both X_q and X_r as descendants, $q, r = 1, \ldots, d$. In the first step, the structures of the triples are estimated and the lcas of all pairs of variables in each triple are found. For a given tree, there are $d-2$ lcas that correspond to all possible pairs (X_q, X_r), $q, r = 1, \ldots, d$. In the second step, the pairs of variables which correspond to the same equivalence class are merged together step by step, resulting in the tree of the HAC. Two pairs of variables (X_q, X_r) and (X_p, X_s) are said to belong to the same equivalence class if they share the same lca in the tree s.

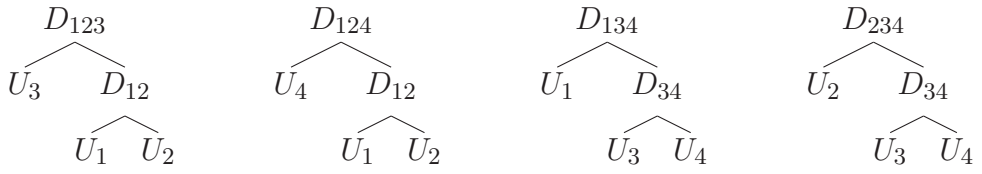


Figure 1.2 A set of trivariate structures corresponding to the copula with $s = ((12)(34))$.

As an example, we consider the 4-dimensional HAC with the predefined structures of the triples presented in Figure 1.2. Consider the first triple (U_1, U_2, U_3) with the structure $((12)3)$. The lca of (U_1, U_2) is the node $D_{U_1U_2}$. For simplicity of notation, we write D_{12} instead of $D_{U_1U_2}$. The parent node of U_1 and U_2 is given by D_{12}. The ancestor nodes of U_1 and U_2 are the nodes D_{12} and D_{123}. Therefore, the lca of (U_1, U_2) in the structure $((12)3)$ is the node D_{12} and the lca of (U_1, U_3) is the node D_{123}. The lcas of each pair are:

$$\begin{array}{c} \\ U_1 \\ U_2 \\ U_3 \\ U_4 \end{array} \begin{array}{c} \begin{array}{cccc} U_1 & U_2 & U_3 & U_4 \end{array} \\ \begin{pmatrix} & \{D_{12}, D_{12}\} & \{D_{123}, D_{134}\} & \{D_{124}, D_{134}\} \\ & & \{D_{123}, D_{234}\} & \{D_{124}, D_{234}\} \\ & & & \{D_{34}, D_{34}\} \\ & & & \end{pmatrix} \end{array}$$

In the given example, the pairs (U_1, U_2) and (U_3, U_4) do not share lcas with any other pair. Therefore, U_1 and U_2 belong to the same equivalence class and are merged together in the first step. The same is true for the pair (U_3, U_4). Consequently, it is observed that the pairs (U_1, U_3), (U_1, U_4), (U_2, U_3) and (U_2, U_4) belong to the same equivalence class and are merged together in the second step. The final structure of the copula is $s = ((12)(34))$. For further examples on how the structure of an HAC can be recovered by applying the concept of an lca, we refer to Segers and Uyttendaele (2014).

In this method, the structure of the individual triples should be found first. Each triple can have a binary or a trivial structure. The structure of the triple is called trivial if all three variables are merged together in one step, and binary otherwise. Formally speaking, for each triple of variables (X_q, X_r, X_s), $q,r,s = 1,\ldots,d$ we aim to test the null hypotheses H_0 : 'the structure is trivial (q,r,s)' against H_1 : 'the structure is binary $((q,r),s)$'. Segers and Uyttendaele (2014) suggest estimating the individual triples using a rank-based method. Let $K_{qr}(w) = \mathrm{P}\{C_2(X_q, X_r) \quad w\}$ be Kendall's distribution between X_q and X_r. Its empirical counterpart is then $\widehat{K}_{qr}(w) = \frac{1}{n}\sum_{m=1}^{n} \mathbf{I}(w_{m,qr} \quad w)$, where $0 < w < 1$, $w_{m,qr} = \frac{1}{n+1}\sum_{l=1}^{n} \mathbf{I}\left(x_{lq} < x_{mq}, x_{lr} < x_{mr}\right)$ and $\mathbf{I}(\cdot)$ is the identity function. The distance between the empirical Kendall distributions of pairs (X_s, X_q) and (X_s, X_r) is defined as

$$\delta_{sq,sr} = \int_0^1 |\widehat{K}_{sq}(x) - \widehat{K}_{sr}(x)| \mathrm{d}x = \frac{1}{n}\sum_{m=1}^{n} |w_{(m),sq} - w_{(m),sr}|, \tag{1.7}$$

where $w_{(1),ij}, \ldots, w_{(n),ij}$ are ordered pseudo-observations of $w_{1,sq} \ldots w_{n,sq}$. Segers and Uyttendaele (2014) point out that a trivial trivariate structure usually results in three distances which are approximately the same, but a binary structure results in one small distance and two larger approximately equal distances. In order to calculate the test statistic, Segers and Uyttendaele (2014) suggest drawing K samples from the nonparametrically estimated trivariate Archimedean copula using the work of Genest, Nešlehová and Ziegel (2011).

As the present paper addresses the framework when the copula family is assumed to be known, we modify the algorithm proposed in Segers and Uyttendaele (2014) and simulate from the copula coming from a predefined class. The test statistic is simulated under the assumption that the structure is trivial, therefore, the parameter of the copula can be found by inversion of the average empirical counterpart of Kendall's τ, i.e. $\widehat{\theta} = v^{-1}\left(\widehat{\tau}_{\text{avg}}\right)$, where $\widehat{\tau}_{\text{avg}} = \left(\widehat{\tau}_{qr} + \widehat{\tau}_{qs} + \widehat{\tau}_{rs}\right)/3$, $q,r,s = 1,\ldots,d$. The inverse $v^{-1}\left(\tau_{\text{avg}}\right)$ corresponds to the solution of the equation

$$\tau_{ij}(\theta) = v(\theta) = 4\int_0^1 \int_0^1 C_2(u_i, u_j; \theta) \mathrm{d}C_2(u_i, u_j; \theta) - 1;\ i,j = 1 \ldots d, \tag{1.8}$$

where $\tau_{ij} = 2\,\mathrm{P}\left\{(X_i - X_j)(Y_i - Y_j) > 0\right\} - 1$, with (X_i, Y_i) and (X_j, Y_j) are independent draws from (X, Y). For some copula functions, the integral in (1.8) is known in closed form as a function of θ, for example, for the Gumbel and Clayton copulae $\theta_{Gumbel}(\tau) = \frac{1}{1-\tau}$ and $\theta_{Clayton}(\tau) = \frac{2\tau}{1-\tau}$, respectively. To sum up, the modification of the algorithm of Segers and

Uyttendaele (2014) which allows identifying the structure of an HAC based on Kendall's distance is summarized in Algorithm 1.

Algorithm 1 Adaptation of the algorithm of Segers and Uyttendaele (2014).

Input : sample $(x_1, x_2, \ldots, x_d)^\top$ of size n, significance level $\alpha^\star$, parametric family of the HAC.

for $l = 1, \ldots, \binom{d}{3}$ **do**

Select a triple from $(x_q, x_r, x_s)^\top$, $q, r, s = 1, \ldots, d$, $q \neq r \neq s$, call it $(z_1, z_2, z_3)^\top$.

Compute the distances $\delta_{12,13}$, $\delta_{12,23}$ and $\delta_{13,23}$ according to (1.7), order them and call the result $\delta_{(1)}, \delta_{(2)}, \delta_{(3)}$.

Compute the test statistic

$$\delta = \frac{|\delta_{(1)} - \delta_{(2)}| + |\delta_{(1)} - \delta_{(3)}|}{2}. \tag{1.9}$$

Compute $\widehat{\tau}_{\text{avg}} = \frac{\widehat{\tau}_{12} + \widehat{\tau}_{13} + \widehat{\tau}_{23}}{3}$ and estimate $\widehat{\theta} = v^{-1}\left(\widehat{\tau}_{\text{avg}}\right)$ according to (1.8).

for $k = 1, \ldots, K$ **do**

Draw a sample of size n from $(U_1, U_2, U_3)^\top \sim C_3\left(u_1, u_2, u_3; (123), \widehat{\theta}\right)$ being a trivial copula.

Compute $\delta^{(k)}$ for the simulated sample k in analog to (1.9).

end for

Compute δ_{crit} by taking the $\alpha = \alpha^\star$ quantile of the empirical distribution of $\delta^{(k)}$, $k = 1, \ldots, K$.

if $\delta > \delta_{crit}$ **then** reject the H_0: the true trivariate structure is the trivial structure.

end if

end for

Recover the full structure of the d-dimensional HAC from the set of $\binom{d}{3}$ triples of variables using the concept of the lowest common ancestor (lca).

Return : the estimated structure of the HAC $\widehat{s}$.

The significance level of the individual tests $\alpha^\star$ should be selected considering the multiple testing procedure. For the significance level of the test to be $\overline{\alpha}$, the significance level of the individual tests should satisfy $\overline{\alpha} = 1 - (1 - \alpha^\star)^{\binom{d}{3}}$. However, this approach is not recommended for high-dimensional samples. Therefore, in the empirical part of the paper, we use the rule of thumb proposed in Uyttendaele et al. (2016) and choose the significance level of the individual tests to be smaller or equal than the overall significance level. It is worth noting that the method of Segers and Uyttendaele (2014) is much more

general as no prior specification of the copula generators is necessary and generators might differ from level to level of the hierarchy. In contrary, our method assumes that generators on all levels of the hierarchy belong to the same predefined family. However, the method proposed in Segers and Uyttendaele (2014) and its modification described in Algorithm 1 are not applicable to the case of high-frequency data because of the absence of a high-frequency estimator of Kendall's τ and Kendall's distribution. The computation of the empirical Kendall's distribution (1.7) involves realizations of $X_1, \ldots, X_d$. Therefore, the estimation of a multivariate distribution of daily observations would require data of a longer time horizon in comparison to the case when the copula is parametrized by solely the realized correlation matrix. The structure and the parameters would have to be fixed within some time window, resulting in the reduced time flexibility of the estimated multivariate distribution. Moreover, Algorithm 1 employs Kendall's distance as the test statistic, which leads to large computational costs in higher dimensions.

Clustering estimator of the structure We propose to proceed analogously to Segers and Uyttendaele (2014) and recover the full structure of an HAC from the set of triples of variables. The estimation of the structure of the individual triples is made using a test that, in contrast to Segers and Uyttendaele (2014), does not involve the observations themselves and is based solely on pairwise correlations.

Consider the triple (X_q, X_r, X_s) and assume that the estimated distance

$$\widehat{h}_{qr} = \min\left(\widehat{h}_{qr}, \widehat{h}_{qs}, \widehat{h}_{rs}\right),$$

where $\widehat{h}_{qr}$ is defined in (1.6). Therefore, the variables X_q and X_r are merged together into the variable (X_q, X_r) in the first step. The distance between the cluster (X_q, X_r) and X_s is calculated according to the complete linkage rule:

$$\widehat{h}_{qr,s} = \max\left\{\widehat{h}_{qs}, \widehat{h}_{rs}\right\}. \tag{1.10}$$

Preliminary simulation studies have shown that the choice of the clustering algorithm is of minor importance. We refer to Kaufman and Rousseeuw (2005) and Hastie et al. (2009) for more details on cluster analysis.

It can be observed that the difference between merging distances $\widehat{h}_{qr,s}$ and $\widehat{h}_{qr}$ is generally bigger if the trivariate copula has a binary structure. Therefore, the measure

$$\Delta\widehat{h} = \widehat{h}_{qr,s} - \widehat{h}_{qr} \tag{1.11}$$

can be chosen as the test statistic to distinguish between trivial and binary structure of a triple.

To sum up, the testing procedure is performed in the following way: for each triple, it is assumed that the structure is trivial, the average correlation is computed, and inverted to the parameter of the trivial copula $f^{-1}\left(\rho_{\text{avg}}\right)$ according to (1.4). The test statistic is obtained by simulating $k = 1, \ldots, K$ samples from the trivial copula and calculating K distances $\Delta\widehat{h}^{(k)}$ according to (1.11). The sample size of the simulated sample corresponds to the sample size of the original sample. Finally, the empirical difference of the merging distances is compared to the quantile of the simulated one. The proposed procedure is briefly summarized in Algorithm 2.

Algorithm 2 Structure determination using cluster analysis.

Input : the realized correlation matrix $\mathcal{P}$ of the dimension $d \times d$ calculated based on the sample $(x_1, x_2, \ldots, x_d)^\top$ of size n, significance level $\alpha^\star$, parametric family of the HAC.

for $l = 1, \ldots, \binom{d}{3}$ **do**

Select a triple from $(q, r, s)^\top$, $q, r, s = 1, \ldots, d$, $q \neq r \neq s$, call it $(1, 2, 3)^\top$.

Compute $\widehat{h}_{12}$, $\widehat{h}_{13}$, and $\widehat{h}_{23}$ according to (1.6).

Merge the two closest variables and calculate $\Delta\widehat{h}$ according to (1.11).

Compute $\rho_{\text{avg}} = \frac{\rho_{12}+\rho_{13}+\rho_{23}}{3}$ and estimate $\widehat{\theta} = f^{-1}\left(\rho_{\text{avg}}\right)$.

for $k = 1, \ldots, K$ **do**

Draw a sample of size n from $(U_1, U_2, U_3)^\top \sim C_3\left(u_1, u_2, u_3; (123)\widehat{\theta}\right)$ being a trivial copula.

Transform $(u_1, u_2, u_3)^\top$ to $\{F_1^{-1}(u_1), F_2^{-1}(u_2), F_3^{-1}(u_3)\}^\top$.

Compute $\Delta\widehat{h}^{(k)}$ for the simulated sample k according to (1.11).

end for

Compute h_{crit} by taking the $\alpha = \alpha^\star$ quantile of the empirical distribution of $\Delta\widehat{h}^{(k)}$, $k = 1, \ldots, K$.

if $\Delta\widehat{h} > h_{crit}$ **then** reject the H_0: the true trivariate structure is the trivial structure.

end if

end for

Recover the full structure of the d-dimensional rHAC from the set of $\binom{d}{3}$ triples of variables using the concept of the lowest common ancestor (lca).

Return : the estimated structure of the HAC $\widehat{s}$.

It is important to note that the estimation of the marginal distributions $F_i(\cdot)$ is a trivial task, as the distribution of the high-frequency log-returns can be assumed to be Gaussian $\mathrm{N}(0, r_i)$, $i = 1, \ldots, d$ based on the results described in Hautsch (2011).

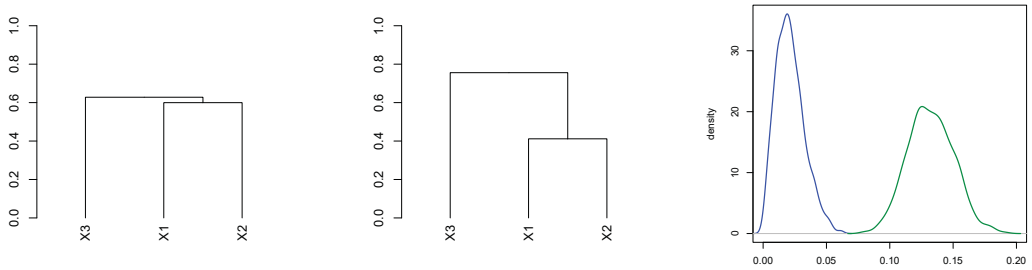


Figure 1.3 Dendrograms for the trivial Gumbel copula $C_3(u_1, u_2, u_3; s = (123); \theta = 1.4)$, the binary Gumbel copula $C_3\left(u_1, u_2, u_3; s = ((12)3); \boldsymbol{\theta} = (1.7, 1.2)^\top\right)$ (center) and kernel density estimate of $\widehat{h}_{12,3} - \widehat{h}_{12}$, where $\widehat{h}_{12,3} = \max\left\{\widehat{h}_{13}, \widehat{h}_{23}\right\}$, blue for the trivial structure and green for the binary structure.

Note: In order to illustrate the test statistic (1.11), samples from

$$C_3(u_1, u_2, u_3; s = (1, 2, 3); \theta = 1.4)$$

and

$$C_3\left(u_1, u_2, u_3; s = ((1, 2), 3); \boldsymbol{\theta} = (1.7, 1.2)^\top\right)$$

are drawn (the copulae are assumed to be Gumbel). The left plot in Figure 1.3 illustrates the dendrogram of the hierarchical cluster analysis based on the distance (1.6) and complete linkage merging rule for a random sample of size 100 from the trivial Gumbel copula. The central part of Figure 1.3 shows the dendrogram for the binary trivariate Gumbel copula. It can be observed that the difference between merging distances $\widehat{h}_{12,3} - \widehat{h}_{12}$ is much smaller for the trivial copula. We simulated $k = 1, \ldots, 100$ random samples from each of the above mentioned copulae, and each time calculated $\Delta\widehat{h}^{(k)}$ according to (1.11). The kernel density estimate of the $\Delta\widehat{h}$ based on 100 random samples is presented in the right part of Figure 1.3. For the given copulae, the density estimate of $\Delta\widehat{h}$ for the trivial copula is more concentrated. This example only illustrates the validity of the proposed test statistic. The distance between these two distributions is influenced by the values of the parameters, and more research should be done to find the asymptotic properties of the proposed test.

Benchmark models Many recent studies have addressed the question of the structure's estimation of an HAC, for example, Okhrin et al. (2013), Górecki et al. (2014), Okhrin et al. (2015), Uyttendaele et al. (2016) and Górecki et al. (2016b). Most of the studies illustrate the performance of the proposed methods by means of simulations. The consistency of the structure's estimator still has to be addressed in the literature. Some of these studies are much more general than Algorithm 2. However, they are not applicable in the current framework, where the observations can not be directly used, as discussed in the previous section. Moreover, in the overwhelming majority of cases, the methods perform in a similar way for big samples. To illustrate the validity of Algorithm 2, it will be compared, by means of simulations, to the recursive procedure proposed in Okhrin et al. (2013) and further improved by Górecki et al. (2014). It has been implemented in the R package HAC by Okhrin and Ristig (2014). The idea of the method is to construct a binary tree by recursively merging the variables with the largest values of the estimated parameter. Subsequently, the obtained tree is collapsed using a predefined merging parameter. As is the case with many others, this method can not be applied to high-frequency data. However, it will provide an opportunity to evaluate the loss of precision and gain in computational speed when the general structure is estimated based solely on the realized correlation matrix.

1.3.2 Estimating the parameters

As was mentioned in Section 1.2, the parameters of the copula can be estimated by the inversion of the realized correlation according to (1.4). However, this is usually done only for the correlation between two variables. Some generalizations for Kendall's τ have already been addressed in the literature. Nelsen (1996) discusses how the parameter of a three-dimensional binary copula can be found by inverting the average coefficient of agreement. Genest, Nešlehova and Ghorbal (2011) have described the average Kendall's τ based approach to the trivial copulae with an odd number of parameters. Górecki et al. (2016a) mention the estimation of the parameters of a binary HAC based on Kendall's τ correlation matrix and discuss a trivial extension to HAC with general structures in Górecki et al. (2016b).

We suggest following the idea of averaging the correlation coefficient ρ_{ij}, $i,j = 1,\dots,d$ over some given set of variables to estimate the parameters of the rHAC. The question whether the procedure based on the average realized correlation gives a valid estimate has not been addressed in the literature.

Suppose that k parameters of the HAC θ_i, $i = 1,\dots,k$ corresponding to k merging nodes need to be estimated. Let $\rho^\star(\mathcal{X}_i)$ be the average correlation of the pairs of variables

with the lca at node $D_{\mathcal{X}_i}$, $i = 1, \ldots, k$, where $\mathcal{X}_i$ is the set of descendant leaves (variables) of the node $D_{\mathcal{X}_i}$, $i = 1, \ldots, k$. Thus, the parameter θ_i of the HAC may be estimated by inverting the average correlation measure $\rho^\star(\mathcal{X}_i)$, $i = 1, \ldots, k$. For the HAC with the structure presented in Figure 1.1, the node associated with the parameter $\theta_3 = 2$ is the node D_{1234}. The children nodes of the node D_{1234} are the nodes D_{12} and D_{34}. The node D_{12} is associated with the parameter $\theta_1 = 4$ and the node D_{34} is associated with the parameter $\theta_2 = 2.5$. Moreover, the node D_{1234} is the ancestor for the nodes associated with the variables X_1, X_2, X_3 and X_4. The lca of the pair (X_1, X_2) is the node D_{12} and the lca of the pair (X_3, X_4) is the node D_{34}. Therefore, the pairs of variables with the lca at node D_{1234} are (X_1, X_3), (X_1, X_4), (X_2, X_3) and (X_2, X_4). Therefore, the average correlation corresponding to the parameter θ_3 is given by $\rho^\star(X_1, X_2, X_3, X_4) = \frac{1}{4}\{\rho_{13} + \rho_{23} + \rho_{14} + \rho_{24}\}$. The parameter θ_3 is estimated by inverting the mentioned above average correlation, i.e. $\widehat{\theta}_3 = f^{-1}\{\rho^\star(X_1, X_2, X_3, X_4)\}$. Analogically, $\rho^\star(X_1, X_2, X_3, X_4, X_5) = \frac{1}{4}\{\rho_{15} + \rho_{25} + \rho_{35} + \rho_{45}\}$, $\widehat{\theta}_4 = f^{-1}\{\rho^\star(X_1, X_2, X_3, X_4, X_5)\}$. A summary of the estimation procedure is given in Algorithm 3.

Algorithm 3 Average correlation estimator.

Input : the realized correlation matrix $\mathcal{P}$, the estimated structure $\widehat{s}$ from Algorithm 2, parametric family of the HAC.

Let $\theta_i, i = 1, \ldots, k$ be the set of the HAC parameters to be estimated.

Let $\mathcal{X}_i$, $i = 1, \ldots, k$ be the set of the descendants of the node $D_{\mathcal{X}_i}$; $\mathcal{X}$ is the set of all variables.

for $i = 1, \ldots k$ **do**

$$\rho^\star(\mathcal{X}_i) = \frac{1}{|(X_j, X_k) \in \mathcal{X} : \operatorname{lca}(X_j, X_k) = D_{\mathcal{X}_i}|} \sum_{(X_j, X_k) \in \mathcal{X} \,:\, \operatorname{lca}(X_j, X_k) = D_{\mathcal{X}_i}} \rho_{jk} \tag{1.12}$$

$$\widehat{\theta}_i(\mathcal{X}_i) = f^{-1}\{\rho^\star(\mathcal{X}_i)\} \tag{1.13}$$

end for

Truncate the parameters according to the nesting condition, i.e. $\widehat{\theta}_i \leq \widehat{\theta}_j$, if $\mathcal{X}_j \subset \mathcal{X}_i$, $i, j = 1, \ldots, k$.

Return : estimated parameter vector $\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \ldots, \widehat{\theta}_k\right)^\top$ of the HAC.

Simulation studies show that the proposed estimator is asymptotically unbiased and follows a Gaussian distribution. In the case when the realized correlation is replaced by Kendall's correlation and the parameter is estimated by applying (1.8) to the average

Kendall's τ. Let $\hat{\tau}^\star(\mathcal{U}_i)$ be the average empirical Kendall's τ of the pairs of variables with the lca at node $D_{\mathcal{U}_i}$ and is defined analogically to (1.12). Let L_i be a set of the pairs of variables with the lca at node $D_{\mathcal{U}_i}$, i.e. $L_i = (U_j, U_l) : \text{lca}(U_j, U_l) = D_{\mathcal{U}_i}, j < l, i \in 1, \ldots, k$, then the asymptotic variance of the average Kendall's τ associated with the node $D_{\mathcal{U}_i}$ and the parameter θ_i can be estimated as

$$\text{Var}\{\hat{\tau}^\star(\mathcal{U}_i)\} = \frac{1}{|L_i|^2} \sum\sum_{(U_j,U_l)\in L_i,(U_p,U_q)\in L_i} \text{cov}\{\hat{\tau}_{il}, \hat{\tau}_{pq}\}, \tag{1.14}$$

$n\,\text{cov}\{\hat{\tau}_{jl}, \hat{\tau}_{pq}\} \underset{n\to\infty}{\to} 16\,\text{cov}\{C_2(U_j, U_l; \hat{\theta}_i) + \bar{C}_2(U_j, U_l; \hat{\theta}_i), C_2(U_p, U_q; \hat{\theta}_i) + \bar{C}_2(U_p, U_q; \hat{\theta}_i)\}$, where $\bar{C}_2(U_j, U_l; \hat{\theta}_i) = U_j + U_l - 1 + C_2(1 - U_j, 1 - U_l; \hat{\theta}_i)$ is the survival copula and $|L|$ is the cardinality of the set L. Combined with the expression (1.8), this implies

$$\text{Var}\left(\hat{\theta}_i\right) = \left[v^{-1}\{\tau^\star(\mathcal{U}_i)\}'\right]^2 \text{Var}\{\hat{\tau}^\star(\mathcal{U}_i)\}.$$

The estimator of the variance is a straightforward application of the result developed in Genest, Nešlehova and Ghorbal (2011).

1.4 Simulation results

In this section, we show the validity of the clustering estimator (CE) presented in Algorithm 2 and Algorithm 3 and compare it to the adaptation of the method of Segers and Uyttendaele (2014) (SU) and the approach of Okhrin et al. (2013) (OOS) which was improved by Górecki et al. (2014) and was implemented in the R package HAC by Okhrin and Ristig (2014). We compare the introduced estimator only to a couple of currently available studies and leave the recent advances discussed in, for example, Górecki et al. (2014), Uyttendaele et al. (2016), Okhrin et al. (2015) and Górecki et al. (2016b) outside the scope of this study since the objective of the simulation studies is rather to answer the question whether the proposed algorithm is valid in the case of linear correlation, than to find the best possible estimator of an HAC. We are aware of the fact that the linear correlation based estimator might be not as efficient as an ML approach or a nonlinear correlation based estimator, as it contains information only about linear dependencies among the variables. However, in the framework of high-frequency data, this is so far the only possible way to proceed. Moreover, we aim to define a minimal recommended sample size.

In the current simulation study no high-frequency observations are presented. In order to compare different methods, the clustering estimator (CE) is applied to the Kendall's

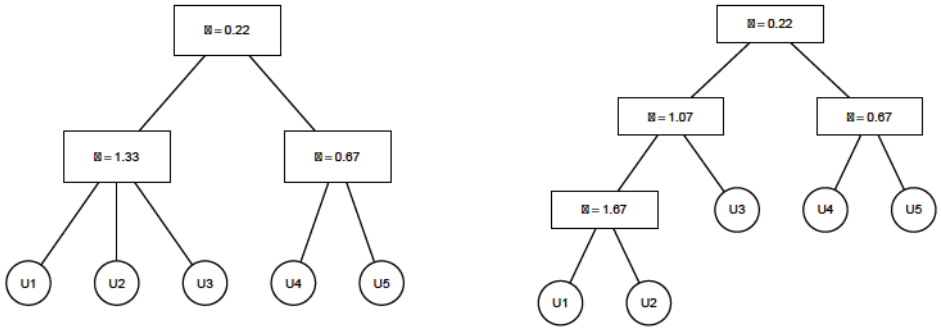


Figure 1.4 Structures of the 5-dimensional copulae used in the simulation studies.

correlation matrix and to the linear correlation matrix estimated in the usual manner over the whole sample path that corresponds to the correlation matrices of the daily log-returns. In the case of the SU estimator, the parameters are estimated by the sequential inversion of Kendall's τ. For the estimation of the structure according to Algorithm 1 and Algorithm 2, we set $K = 500$ and $\alpha^\star = 0.01$. A full ML is applied to the structures estimated by OOS. For illustrative purposes, the 5-dimensional copulae structures presented in Figure 1.4 are considered. For each structure, Clayton, Gumbel and Frank copulae are analysed with the parameters corresponding to $\boldsymbol{\tau} = (0.40, 0.25, 0.10)^\top$ and $\boldsymbol{\tau} = (0.45, 0.35, 0.25, 0.10)^\top$. The marginal distributions are assumed to be known. For each of the above mentioned estimators, we proceed as follows: a sample of size n is simulated from the copula, and the structure is estimated. If the estimated structure coincides with the true one, the parameters are estimated. The procedure is repeated m times until 200 structures are estimated correctly. Thus, the estimators of the structure are compared in terms of the proportion of correctly estimated structures $200/m$. For the comparison of the estimation of the parameters, we introduce the characteristic $E = \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|$, which is the Euclidean norm of the difference between the vector of true parameters and the estimated ones. Tables 1.1 and 1.2 present the mean $\overline{E}$, the variance $\mathrm{Var}(E)$ and the 25% $q_{0.25}(E)$, 50% $q_{0.5}(E)$ and 75% $q_{0.75}(E)$ quantiles of E for different structures.

Table 1.1 shows the simulation results for the 5-dimensional Clayton copula presented in Figure 1.4 with sample sizes $n = 30, 50, 70, 100, 200, 300, 500, 800, 1000$. The results make evident that the OOS method outperforms all the competitors for small samples for the Clayton copula with the structure $s = ((123)(45))$. However, there are some outliers, which

	n	$200/m$	$\overline{E}$	$\mathrm{Var}(E)$	$q_{0.25}(E)$	$q_{0.5}(E)$	$q_{0.75}(E)$
	30	0.262	0.738	0.175	0.465	0.686	0.930
	50	0.370	0.518	0.078	0.312	0.449	0.650
	70	0.449	0.435	0.040	0.290	0.401	0.543
	100	0.570	0.356	0.023	0.249	0.338	0.460
CE τ	200	0.797	0.236	0.013	0.158	0.221	0.279
	300	0.847	0.190	0.007	0.126	0.180	0.241
	500	0.873	0.137	0.004	0.091	0.124	0.177
	800	0.905	0.113	0.003	0.070	0.107	0.144
	1000	0.840	0.110	0.003	0.070	0.097	0.142
	30	0.268	1.716	3.813	0.525	0.816	1.519
	50	0.439	1.104	2.468	0.355	0.556	0.725
	70	0.472	0.853	1.828	0.309	0.466	0.650
	100	0.592	0.483	0.645	0.242	0.342	0.461
CE ρ	200	0.797	0.247	0.014	0.166	0.228	0.314
	300	0.866	0.198	0.008	0.128	0.181	0.255
	500	0.870	0.146	0.005	0.093	0.135	0.192
	800	0.917	0.115	0.004	0.067	0.110	0.155
	1000	0.873	0.115	0.003	0.070	0.106	0.153
	30	0.203	0.727	0.136	0.469	0.679	0.934
	50	0.276	0.532	0.069	0.336	0.513	0.663
	70	0.349	0.449	0.051	0.292	0.401	0.562
SU	100	0.441	0.360	0.024	0.259	0.336	0.464
	200	0.645	0.250	0.015	0.164	0.231	0.301
	300	0.722	0.188	0.008	0.123	0.171	0.239
	500	0.847	0.138	0.005	0.093	0.124	0.178
	800	0.905	0.113	0.003	0.070	0.107	0.144
	1000	0.840	0.110	0.003	0.070	0.097	0.142
	30	0.141	0.323	0.027	0.224	0.297	0.422
	50	0.216	0.298	0.021	0.188	0.267	0.376
	70	0.300	0.257	0.014	0.178	0.240	0.321
	100	0.402	0.225	0.011	0.154	0.212	0.270
OOS	200	0.647	0.154	0.006	0.093	0.151	0.194
	300	0.740	0.129	0.003	0.089	0.119	0.162
	500	0.915	0.103	0.002	0.069	0.099	0.134
	800	0.980	0.075	0.001	0.052	0.073	0.094
	1000	0.983	0.071	0.001	0.049	0.065	0.092

Table 1.1 Simulation results for the Clayton copula with the structure $((123)(45))$ and $\boldsymbol{\theta} = (1.33, 0.67, 0.22)^\top$.

	n	$200/m$	$\overline{E}$	$\mathrm{Var}(E)$	$q_{0.25}(E)$	$q_{0.5}(E)$	$q_{0.75}(E)$
	30	0.288	1.095	0.551	0.664	0.954	1.268
	50	0.374	0.766	0.145	0.480	0.727	0.966
	70	0.407	0.659	0.188	0.443	0.566	0.772
	100	0.601	0.506	0.050	0.357	0.485	0.638
CE τ	200	0.858	0.336	0.026	0.223	0.315	0.431
	300	0.939	0.288	0.017	0.192	0.271	0.361
	500	0.995	0.213	0.007	0.151	0.206	0.262
	800	1.000	0.167	0.005	0.113	0.153	0.209
	1000	1.000	0.158	0.004	0.114	0.148	0.202
	30	0.262	2.352	3.969	0.830	1.413	5.358
	50	0.421	1.420	2.553	0.543	0.838	1.260
	70	0.475	0.978	1.593	0.412	0.612	0.873
	100	0.621	0.687	0.755	0.364	0.516	0.713
CE ρ	200	0.885	0.352	0.028	0.242	0.318	0.424
	300	0.952	0.324	0.022	0.219	0.292	0.402
	500	1.000	0.228	0.013	0.154	0.216	0.276
	800	1.000	0.183	0.007	0.121	0.164	0.220
	1000	1.000	0.169	0.006	0.110	0.161	0.210
	30	0.252	1.072	0.329	0.657	0.959	1.329
	50	0.401	0.756	0.146	0.464	0.699	0.926
	70	0.448	0.657	0.097	0.447	0.598	0.809
SU	100	0.401	0.508	0.050	0.360	0.471	0.616
	200	0.615	0.353	0.026	0.234	0.339	0.447
	300	0.760	0.300	0.018	0.194	0.284	0.369
	500	0.939	0.207	0.006	0.147	0.206	0.253
	800	0.995	0.167	0.005	0.113	0.153	0.209
	1000	1.000	0.158	0.004	0.114	0.148	0.202
	30	0.388	0.539	0.096	0.333	0.447	0.657
	50	0.536	0.420	0.046	0.278	0.376	0.508
	70	0.666	0.359	0.024	0.244	0.328	0.451
	100	0.774	0.305	0.017	0.212	0.291	0.364
OOS	200	0.953	0.226	0.008	0.165	0.217	0.271
	300	0.985	0.198	0.007	0.135	0.183	0.246
	500	0.998	0.146	0.004	0.099	0.137	0.179
	800	1.000	0.112	0.002	0.081	0.108	0.141
	1000	1.000	0.106	0.002	0.070	0.101	0.133

Table 1.2 Simulation results for the Clayton copula with the structure $(((12)3)(45))$ and $\boldsymbol{\theta} = (1.67, 1.07, 0.67, 0.22)^\top$.

can be seen from the sample variance of E. This means that the full ML estimate had a large deviation from the true value of the parameter for a few samples. The interquartile range $q_{0.75}(E) - q_{0.25}(E)$ is still smaller for the ML in small samples. The same results for the variance are observed for the CE ρ, therefore, this estimator is not recommended for small samples. In contrast, Table 1.2 shows that for the structure $s = (((12)3)(45))$, OOS is not the best method for estimating the structure in small samples. This is due to the fact that the performance of this estimator depends on the choice of the merging parameter. The results for the other copulae are presented in Appendix 1.C and show that there is no leading method in terms of estimating the structure. The method to choose depends on the type of the copula and the values of the parameters. For a large enough sample, all the methods perform similarly. The general conclusion to be drawn for the estimation of the parameters is that the variance of the CE r estimator is the highest for small samples and that the full ML has the smallest variance, however, some exceptions are observed. It is worth noting that the simulation results are used just for comparison purposes, as the difference in the parameters influences the proportion of the correctly estimated structures more severely than does the type of the copula. Additionally, the dimension of the copula should always be taken into consideration in order to select the minimal sufficient sample size. The question of convergence of the estimator to the true structure still needs to be addressed in the literature.

In Figure 1.5, we take a closer look at the individual components of $\boldsymbol{\theta}$. We compare only CE based on Kendall's correlation and the full ML, as the CE ρ and SU behave very similarly in terms of the properties of $\widehat{\boldsymbol{\theta}}$. It is evident that both estimators are asymptotically unbiased, however, CE has a higher variance. In addition to the kernel density estimates of CE and ML, we add a kernel density estimate of the Gaussian sample (blue line) with the mean $\boldsymbol{\theta}$ and the variance estimated from (1.14) and observe that it coincides with the kernel density estimate of CE.

It is worth noting that the computational advantage is on the side of CE. Figure 1.6 shows the average computational time in seconds for all the above mentioned estimators over 100 trials. The difference in the computational time becomes crucial with growing dimensions, for example, in Segers and Uyttendaele (2014), the SU estimation of a 7-dimensional copula needs roughly 20 minutes versus 15 seconds for the proposed clustering estimator (CE).

The main conclusion of this section is that the linear correlation based clustering estimator is applicable in practice and can be applied to high-frequency data, where moderate samples are atypical.

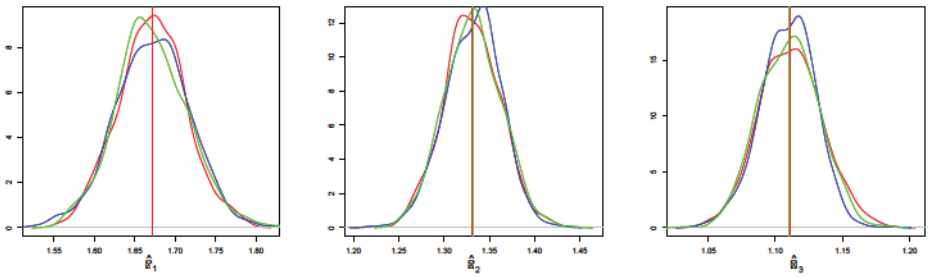


Figure 1.5 KDE of $\widehat{\theta}^{CE}$ (green), $\widehat{\theta}_{MLE}$ (red) and KDE of the Gaussian distribution $\mathrm{N}\{\widehat{\theta}^{CE}, \widehat{\mathrm{Var}}(\widehat{\theta}^{CE})\}$ sample (blue) for the Gumbel copula with the structure $((123)(45))$ and $\boldsymbol{\theta} = (1.67, 1.33, 1.11)^\top$.

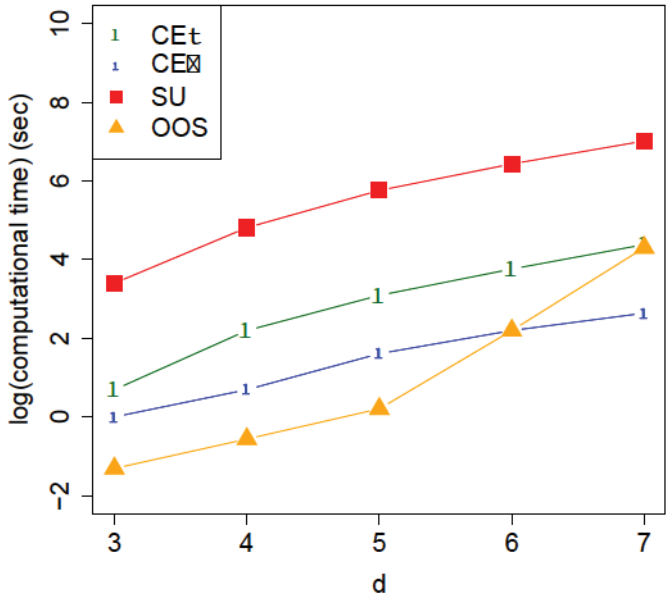


Figure 1.6 Average log computational time (in seconds) over 100 simulations for the estimation of the Clayton copula by CE and the benchmark models depending on the dimension.

1.5 Forecasting VaR using high-frequency data

1.5.1 Predicting rHAC

The model introduced in this section extends the work of Fengler and Okhrin (2016) to higher dimensions. The computationally expensive estimating procedure (1.5) is reduced

to a set of simple tasks of the form (1.13). Moreover, this procedure allows avoiding the question of the optimal choice of the weighting matrix W in (1.5).

As mentioned in Section 1.2, the combination of a lemma of Hoeffding (1940) and Sklar's theorem (1.1) allows to express the pairwise covariances in terms of the copula and the corresponding marginal distributions. Under the assumption that the marginal distributions $F_i(x_i, r_{i,t+1})$, $i = 1, \dots, d$, are Gaussian $\mathrm{N}(0, r_{i,t+1})$, the multivariate distribution of daily log-returns $\mathcal{X}_{t+1}|_{\mathcal{F}_t} \sim F(\cdot; R_{t+1})$ is parametrized solely by a $\mathcal{F}_t$-measurable covariance matrix R_{t+1}. This is due to the fact that the structure s_{t+1} and the parameters $\boldsymbol{\theta}_{t+1}$ of the HAC are estimated from realized correlation matrix $\mathcal{P}_{t+1}$ using Algorithm 2 and Algorithm 3 and the margins are fully specified by the realized volatilities $r_{i,t+1}$, $i = 1, \dots, d$, i.e.

$$F_{\mathcal{X}_{t+1}}(x, R_{t+1}) = C_d\left\{F_1(x_1, r_{1,t+1}), \dots, F_d(x_d, r_{d,t+1}); s_{t+1}; \boldsymbol{\theta}_{t+1}\right\}, \tag{1.15}$$

where $x = (x_1, x_2, \dots, x_d)^\top$. The prediction of the multivariate distribution of daily log-returns is based on the predicted realized covariance matrix $\widehat{R}_{t+1|t}$ obtained by the Heterogeneous Autoregressive (HAR) model introduced by Corsi (2009) and applied in the spirit of Bauer and Vorkink (2011). First, the individual elements of the realized covariance matrix are stacked together into a joint matrix. Then, the matrix logarithm $A_t = \operatorname{logm}(R_t)$ is calculated to guarantee that the matrix is positive definite. In the next step, the covariances are stacked into one vector $a_t = \operatorname{vech}(A_t)$ and modelled using the logarithmic version of the HAR model:

$$\log a_{t+1} = \beta_0 + \beta_D \log a_t^D + \beta_W \log a_t^W + \beta_M \log a_t^M + \varepsilon_{t+1}, \tag{1.16}$$

where $a_t^D = a_t$, $a_t^W = \frac{1}{5}\sum_{i=0}^{4} a_{t-i}$, $a_t^M = \frac{1}{22}\sum_{i=0}^{21} a_{t-i}$, and ε_{t+1} is an error term. When the coefficients in (1.16) are estimated using ordinary least squares, the prediction $\widehat{a}_{t+1}$ is obtained. The prediction $\widehat{R}_{t+1|t}$ is obtained by applying the reverse vech-operator to $\widehat{a}_{t+1}$ and taking the matrix exponential $\widehat{R}_{t+1|t} = \operatorname{expm}\left(\widehat{A}_{t+1|t}\right)$. The prediction of the realized correlation matrix $\widehat{\mathcal{P}}_{t+1|t}$ is obtained by dividing the elements of $\widehat{R}_{t+1|t}$ by the product of the square roots of the corresponding predicted realized volatilities, i.e. $\widehat{\rho}_{ij,t+1|t} = \frac{\widehat{r}_{ij,t+1|t}}{\sqrt{\widehat{r}_{i,t+1|t}\,\widehat{r}_{j,t+1|t}}}$. Since we consider only one-day-ahead prediction, we assume that the prediction bias caused by the nonlinear transformation is small and omit the bias adjustment, analogously to Chiriac and Voev (2011).

We stress once again that only the realized correlation matrix is used for the estimation procedure. The computational costs of such an estimator are low, and the rHAC model still shows excellent forecasting properties.

1.5.2 Competitor models

In order to show a competitive advantage of the rHAC, we apply it to one-day-ahead VaR prediction for a multidimensional portfolio and compare the performance of the rHAC to three classes of benchmark models:

- Rolling window copula models
- Dynamic copula models
  - Copula DCC model Engle (2002)
  - Dynamic copula model by Patton (2004)
  - GAS, GRAS by Creal et al. (2013) and Salvatierra and Patton (2015)
- Realized covariance model by Bauer and Vorkink (2011)

The first class employs copula models with parameters fixed over the given time interval. The second includes dynamic copula models which assume that the parameter of the copula follows some autoregressive process. The third class, which is both popular and successful, comprises the realized volatility models. A more detailed description of the benchmark models is given in Appendix 1.E.

1.6 Application

This section illustrates the rHAC model using high-frequency log-returns of stocks traded on the New York Stock Exchange. First, we give a description of the data used in the empirical part of this section. Thereafter, we apply the rHAC and the above mentioned competing models to one-day-ahead VaR prediction. The interpretation of the results is provided at the end of this section.

Value at Risk prediction The selected data set consists of the tick-by-tick prices of 6 assets obtained from TickData: AA (Alcoa Inc), AXP (American Express), BAX (Baxter International Inc.), C (Citigroup Inc.), INTC (Intel Corporation) and KO (Coca-Cola Co.).

The selection of the number of assets was motivated by the computational intensity of some of the competing models. A well-diversified portfolio was chosen. The selected companies represent the following industrial sectors: consumer products, technology, financial services, chemicals, health care, communications, and energy. The considered time period is from January 2005 to March 2010 which corresponds to $T = 1346$ trading days. This choice stems from the fact that the correlations among the log-returns increased during the financial crisis. We are interested in testing whether the rHAC model is able to capture the crashes appearing in 2008 and 2009. To answer this question, we compare the VaR level α to the exceedance ratio $\widehat{\alpha} = \frac{N}{T}$, where the VaR is defined as the quantile of the profit and loss (P& L) distribution $l_t = (V_{t+1} - V_t) = \sum_{j=1}^{d} a_{j,t} S_{j,t} \{\exp(x_{j,t+1}) - 1\}$, $j = 1, \ldots, d$. $V_t = \sum_{j=1}^{d} a_{j,t} S_{j,t}$ is the value of the portfolio at time t, $a_{j,t}$ are some weights, $S_{j,t}$ is the jth asset's closing price at day t, $x_{j,t+1}$ is the jth asset's log-return at day $t+1$, d is the number of assets in the portfolio, T is the sample size, and $N = \sum_{t=1}^{T} \mathbf{I}\{l_t < \widehat{\text{VaR}}_t(\alpha)\}$ is the number of exceedances of the realization of distribution l_t. From now on, portfolios with equal wealth allocation are considered, i.e. $a_{j,t} = V_t / (d \times S_{j,t})$, $j = 1, \ldots, d$.

Before applying the models, the dataset was cleaned according to Brownless and Gallo (2006), namely the quotes with normal trading conditions, positive price and volume with the timestamp within office trading hours of NYSE are used. Then, outliers have been removed according to a specific bid–ask spread rule.

After the dataset was cleaned, the log-returns were aggregated to the 1-minute frequency and the realized volatilities and correlations were obtained using the realized kernel estimator, which allows reducing the microstructure noise. More details on this estimator are given in Appendix 1.B.

The prediction of the realized volatilities and the realized correlations is made using the HAR model (1.16). The realized volatilities of the selected assets and their out-of-sample predictions are given in Appendix 1.D, Figure 1.A.1. The time series of the selected realized correlations together with the predicted values are given in Appendix 1.D, Figure 1.A.2. The results coincide with the conclusions of Audrino and Corsi (2010), who state that the prediction of the realized correlations is more difficult than the prediction of the realized volatility due to their large variance. When the realized correlations and the realized volatilities are estimated and the forecast is made, the out-of-sample prediction of the one-day-ahead VaR at the 0.5%, 1%, 5% and 10% levels can be made using the clustering estimator according to Algorithm 4.

Algorithm 4 Applying rHAC to the VaR.

Input: predicted realized covariance matrix $\widehat{R}_{t+1|t}$, predicted realized correlation matrix $\widehat{\mathcal{P}}_{t+1|t}$, log-returns $x_{j,t}$, $j = 1, \ldots, d$.

Predict the $\widehat{R}_{t+1|t}$ using HAR, compute $\widehat{\mathcal{P}}_{t+1|t}$.

Estimate the structure $\widehat{s}_{t+1|t}$ and the parameter vector $\widehat{\boldsymbol{\theta}}_{t+1|t}$ of the rHAC from $\widehat{\mathcal{P}}_{t+1|t}$ using Algorithm 2 and Algorithm 3 with $\alpha^\star = 0.01$.

for $i = k, \ldots, 1000$ **do**

Simulate a sample $u_{j,t+1|t}$ from $C_d\left(\cdot; \widehat{s}_{t+1|t}, \widehat{\boldsymbol{\theta}}_{t+1|t}\right)$, $j = 1, \ldots, d$.

Compute $x_{j,t+1|t} = \sqrt{\widehat{r}_{j,t+1|t}} \cdot \Phi^{-1}\left(u_{j,t+1|t}\right)$.

Calculate P&L $l_{t+1}^{(k)}$

end for

Calculate the $\widehat{\text{VaR}}_{t+1|\mathcal{F}_t}(\alpha)$ as

$$\widehat{\text{VaR}}_{t+1|\mathcal{F}_t}(\alpha) = \widehat{F}^{-1}_{l_{t+1}|\mathcal{F}_t}(\alpha)$$

Return: $\widehat{\text{VaR}}_{t+1|\mathcal{F}_t}(\alpha)$.

In the VaR modelling, it is required that the exceedances are independent and the percentage of the exceedances corresponds to the predefined VaR level. Three backtesting procedures have been used to test these properties. The first testing procedure is the unconditional coverage testing due to Kupiec (1995), which compares the exceedance ratio to the VaR level. The second procedure is the VaR duration test of Christoffersen and Pelletier (2004), which checks the independence of the exceedances. This backtesting tool is based on the number of days between the violations of the risk metric.

The dynamic quantile (DQ) test of Engle and Manganelli (2004) enables testing the two required properties simultaneously. In the most widespread version of the test, the demeaned exceedances are regressed on their first lag and the lagged values of the VaR:

$$\mathbf{I}\{l_t < \widehat{\text{VaR}}_t(\alpha)\} - \alpha = \gamma_0 + \gamma_1 \mathbf{I}\{l_{t-1} < \widehat{\text{VaR}}_{t-1}(\alpha)\} - \alpha + \gamma_2 \widehat{\text{VaR}}_{t-1}(\alpha) + \varepsilon_t. \tag{1.17}$$

The null hypothesis for independence and conditional coverage is given by $H_0 : \gamma_0 = 0$, $\gamma_1 = 0$ and $\gamma_2 = 0$.

To verify this method, the results are compared to the benchmark models described in Section 1.5.2. The backtesting results of the unconditional coverage and independence tests are presented in Table 1.3. The p-values indicate that the copula models give more accurate

prediction for the AA-AXP-BAX-C-INTC-KO portfolio, at the 0.5%, 1% and 5% levels, and do not match the 10% level quantile well. The unconditional coverage test supports both the rolling window and rHAC models. However, the independence test of Christoffersen and Pelletier (2004) speaks in favor of the rHAC model.

The time series of the P&L for the given portfolio and the corresponding VaR bounds are illustrated in Figure 1.7. The rHAC method has been found to be effective in handling the 1% and 0.5% quantiles, which is especially important in risk management. No models with a similar predictive power have been found. The hitting ratios of the dynamic copula and the realized covariance approaches are disappointing.

As was mentioned above, VaR prediction using the competing models gets computationally difficult in higher dimensions, which is not the case for the rHAC approach. The VaR regions of the rHAC model and the model of Bauer and Vorkink (2011) for a portfolio consisting of 17 assets (AA (Alcoa Inc.), AXP (American Express), BAX (Baxter International Inc.), BLK (BlackRock Inc.), C (Citigroup Inc.), DOW (Dow Chemical Company), GS(Goldman Sachs Group), HAS (Hasbro Inc.), HOG (Harley-Davidson Inc.), INTC (Intel Corporation), KO (Coca-Cola Co.), MET (Metlife Inc.), MSFT (Microsoft Corporation), NKE (Nike Inc.), PFE (Pfizer), VZ (Verizon Communications), XOM (Exxon Mobil Corporation)) are given in Figure 1.8. The p-values for three considered backtesting procedures can be found in Table 1.4. It is evident that the multidimensional realized copula model does not suffer from the curse of dimensionality, and performs satisfactorily in the sense of unconditional coverage for moderate α levels in higher dimensions. The null hypothesis of the unconditional coverage test for the Gaussian model of Bauer and Vorkink (2011) is rejected at all VaR levels.

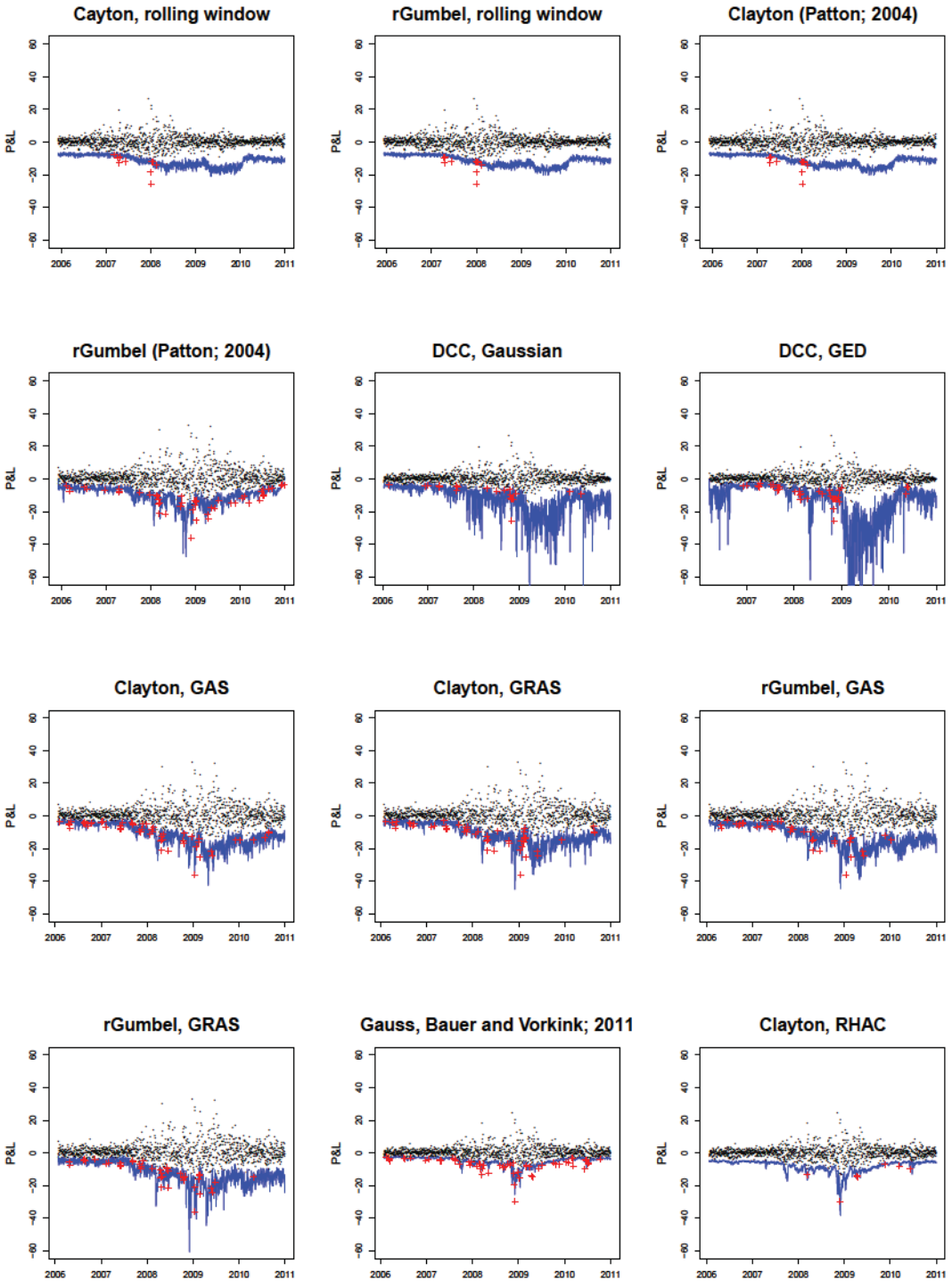


Figure 1.7 Exceedances for the VaR(0.01) of the AA-AXP-BAX-C-INTC-KO portfolio. P & L (black dots), the lower VaR(0.01) (blue solid line), exceedances (red crosses).

Model	Level	$\widehat{\alpha}$	K	C	DQ
Rolling window, Clayton, GED	$\alpha = 0.005$	0.0030	0.2712	0.0317	0.6756
	$\alpha = 0.01$	0.0076	0.3510	0.0000	0.6619
	$\alpha = 0.05$	0.0514	0.8163	1.0000	0.1290
	$\alpha = 0.10$	0.1043	0.0000	0.0000	0.0070
Rolling window, rGumbel, GED	$\alpha = 0.005$	0.0045	0.8076	0.0018	0.4378
	$\alpha = 0.01$	0.0083	0.5257	0.0000	0.0053
	$\alpha = 0.05$	0.0506	0.9148	0.0000	0.0186
	$\alpha = 0.10$	0.0990	0.0000	0.0000	0.0016
DCC, t-copula	$\alpha = 0.005$	0.0232	0.0001	0.0000	0.0139
	$\alpha = 0.01$	0.0304	0.0000	0.0000	0.0041
	$\alpha = 0.05$	0.0728	0.0000	0.0000	0.0001
	$\alpha = 0.10$	0.1112	0.0000	0.0000	0.0000
DCC, t-copula, GED	$\alpha = 0.005$	0.0054	0.0671	0.0000	0.9796
	$\alpha = 0.01$	0.0162	0.0403	0.0000	0.0000
	$\alpha = 0.05$	0.0470	0.0000	0.0000	0.3045
	$\alpha = 0.10$	0.0924	0.0000	0.0000	0.2985
Patton, Clayton	$\alpha = 0.005$	0.0509	0.0000	0.0377	0.6360
	$\alpha = 0.01$	0.0616	0.0000	0.0601	0.7315
	$\alpha = 0.05$	0.1036	0.0000	0.2041	0.7414
	$\alpha = 0.10$	0.1366	0.0001	0.3031	0.4549
Patton, rGumbel	$\alpha = 0.005$	0.0332	0.0000	0.0786	0.8460
	$\alpha = 0.01$	0.0370	0.0000	0.0425	0.6558
	$\alpha = 0.05$	0.0612	0.0709	0.0653	0.5654
	$\alpha = 0.10$	0.0937	0.4372	0.1178	0.3615
GAS, Clayton, GED	$\alpha = 0.005$	0.0303	0.0000	0.0549	0.0726
	$\alpha = 0.01$	0.0427	0.0000	0.0079	0.1935
	$\alpha = 0.05$	0.0822	0.0000	0.0493	0.0078
	$\alpha = 0.10$	0.1404	0.0000	0.4827	0.0046
GRAS, Clayton, GED	$\alpha = 0.005$	0.0303	0.0000	0.0002	0.0001
	$\alpha = 0.01$	0.0388	0.0000	0.0000	0.0001
	$\alpha = 0.05$	0.0869	0.0000	0.0234	0.0014
	$\alpha = 0.10$	0.1381	0.0000	0.5202	0.0164
GAS, rGumbel, GED	$\alpha = 0.005$	0.0217	0.0000	0.0208	0.0838
	$\alpha = 0.01$	0.0295	0.0000	0.0052	0.0150
	$\alpha = 0.05$	0.0760	0.0001	0.0035	0.0007
	$\alpha = 0.10$	0.1296	0.0007	1.0000	0.0027
GRAS, rGumbel, GED	$\alpha = 0.005$	0.0202	0.0000	0.0052	0.0884
	$\alpha = 0.01$	0.0326	0.0000	0.0001	0.0639
	$\alpha = 0.05$	0.0706	0.0014	0.0242	0.0228
	$\alpha = 0.10$	0.1327	0.0002	1.0000	0.0345
RCov, Bauer and Vorkink	$\alpha = 0.005$	0.0350	0.0000	0.2920	0.0009
	$\alpha = 0.01$	0.0474	0.0000	0.1937	0.0008
	$\alpha = 0.05$	0.1213	0.0000	0.8088	0.0038
	$\alpha = 0.10$	0.1773	0.0000	0.0017	0.0017
rHAC, Clayton	$\alpha = 0.005$	0.0047	0.8589	0.5042	0.0000
	$\alpha = 0.01$	0.0085	0.5873	0.5064	0.0028
	$\alpha = 0.05$	0.0551	0.4098	0.1521	0.0000
	$\alpha = 0.10$	0.1140	0.0995	0.1482	0.0000

Table 1.3 VaR performance for the AA-AXP-BAX-C-INTC-KO. The hitting ratio $\widehat{\alpha}$ and the p-values of the Kupiec test (K), Christoffersen (C), and the DQ test.

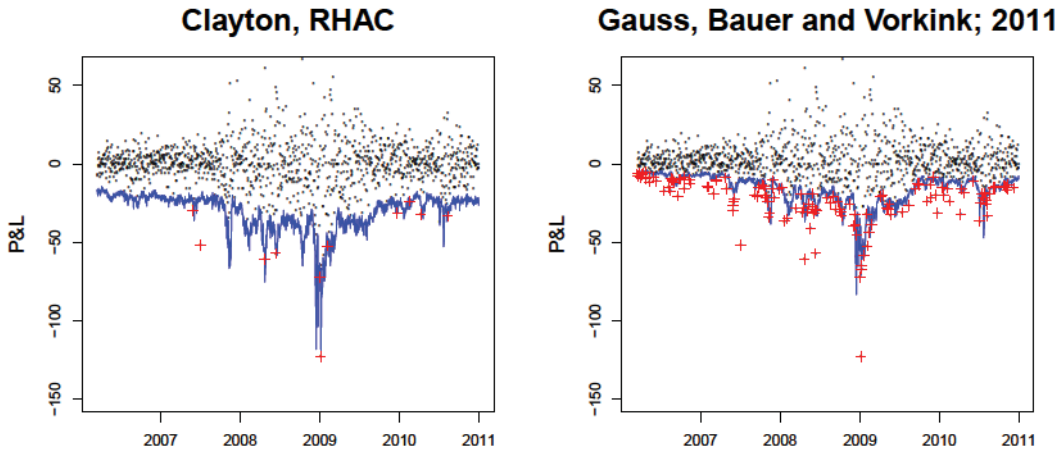


Figure 1.8 Exceedances for the VaR(0.01) of the AA-AXP-BAX-BLK-C-DOW-GS-HAS-HOG-INTC-KO-MET-MSFT-NKE-PFE-VZ-XOM portfolio. P & L (black dots), the lower VaR(0.01) (blue solid line), exceedances (red crosses).

Model	Level	$\widehat{\alpha}$	K	C	DQ
rHAC, Clayton	$\alpha = 0.005$	0.0040	0.6008	0.1006	0.0006
	$\alpha = 0.01$	0.0088	0.6593	0.5534	0.0000
	$\alpha = 0.05$	0.0192	0.0000	0.3231	0.0000
	$\alpha = 0.10$	0.0791	0.0107	0.6305	0.0000
RCov, Bauer and Voev	$\alpha = 0.005$	0.0799	0.0000	0.7393	0.9752
	$\alpha = 0.01$	0.1102	0.0000	0.7745	0.0710
	$\alpha = 0.05$	0.1294	0.0000	0.3221	0.0582
	$\alpha = 0.10$	0.1925	0.0000	0.0002	0.1289

Table 1.4 VaR performance for the AA-AXP-BAX-BLK-C-DOW-GS-HAS-HOG-INTC-KO-MET-MSFT-NKE-PFE-VZ-XOM. The hitting ratio $\widehat{\alpha}$ and the p-values of the Kupiec test (K), Christoffersen (C), and the DQ test.

Conclusions

The concept of the realized hierarchical Archimedean copula has been introduced. This model allows combining the flexibility of copula models with the additional information contained in high-frequency data. It has been suggested to combine the estimation procedures described in Segers and Uyttendaele (2014) and Górecki et al. (2016a) and adapt them to high-frequency data. This estimator is of particular importance in short-term financial

risk management, as the structure and the parameters of the copula are estimated daily based solely on the realized correlation matrix.

Based on the simulation results, it has been concluded that the linear correlation matrix based estimator performs well for large enough samples; it is unbiased but less efficient that the full maximum likelihood estimator. However, it is less computationally intensive than benchmark models and does not suffer from the curse of dimensionality.

In the empirical part of the study, the proposed estimator has been applied to predict the VaR based on high-frequency data for two portfolios, one of 6 and the other of 17 assets. The results have been compared to the benchmark approaches including dynamic copulas and realized covariance models. Based on three tests (Kupiec, Christoffersen, DQ), it has been concluded that the VaR regions obtained by the high-dimensional realized copula models outperform the benchmark models in higher dimensions, especially for lower VaR levels.

Appendices

Appendix 1.A The generators and the densities of some ACs

Copula	Generator	Distribution	Parameter
Gumbel	$(-\log t)^{\theta}$	$\exp\left[-\left\{\sum_{i=1}^{d}(-\log u_i)^{\theta}\right\}^{\frac{1}{\theta}}\right]$	$\theta \in [1,\infty)$
Clayton	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$	$\max\left[\left\{\left(\sum_{i=1}^{d}u_i^{-\theta}\right)-d+1\right\}^{-\frac{1}{\theta}},0\right]$	$\theta \in (0,\infty)$
Frank	$-\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right)$	$\frac{1}{\theta}\log\left\{1+\frac{\prod_{i=1}^{d}(\exp(-\theta u_i)-1)}{(\exp(-\theta)-1)^{d-1}}\right\}$	$\theta \in (0,\infty)$

Table 1.A.1 Archimedean copulae: Gumbel, Clayton and Frank.

Appendix 1.B Realized covariance and realized kernel estimator

Assume that the d-dimensional log-price process follows a Brownian semimartingale

$$X_t = X_{t-1} + \int_{t-1}^{t} \sigma_u \, dW_u$$

where $[t-1;t]$ is a period corresponding to one trading day, σ_t is a càdlàg volatility matrix process and W_t is a d-dimensional vector of independent Brownian motions. It is important to note that the price process is superimposed by the market microstructure noise U_{τ_i}, i.e. one observes

$$P_{\tau_i} = X_{\tau_i} + U_{\tau_i},$$

where $t-1 = \tau_0 < \tau_1 < \ldots < \tau_N = t$, $\mathrm{E}(U_{\tau_i}) = 0$, $\sum_h |h\Omega_h| < \infty$ and $\Omega_h = \mathrm{cov}(U_{\tau_i}, U_{\tau_{i-h}})$ for $h > 0$. The realized covariance over the time interval $[t-1;t]$ is defined as the sample

analog of the quadratic variation of X given by

$$[X]_{t,t-1} = \int_{t-1}^{t} \Sigma_u du$$

with $\Sigma = \sigma\sigma^\top$ and is denoted by R_t in Section 1.2.

One of the estimators which reduces the effect of microstructure noise is the realized kernel estimator proposed by Barndorff-Nielsen et al. (2008). As the realized covariances are obtained by summing all the cross products of log-returns that have a non zero overlapping of their respective time span, the data should be synchronized first. The procedure which is called refresh time sampling and described in Hautsch (2011) is applied to synchronize the data. The first refresh time is defined as $\tau_1^* = \max\{\tau_{1,1}, \ldots, \tau_{d,1}\}$ and $\tau_{i+1}^* = \min\{\tau_{j,k_j} | \tau_{j,k_j} > \tau_i^*,\ \forall k_j = 1, \ldots, N_j;\ j \in 1 \ldots d\}$, where N_j is the number of price observations for asset j. As a result, a new high-frequency vector of returns $p_i = P_{\tau_i^*} - P_{\tau_{i-1}^*}$ is produced, where $i = 1, \ldots, n$, and n is the number of the synchronized observations.

The multivariate realized kernel estimator is given by

$$K(P) = \sum_{h=-H}^{H} k\left(\frac{|h|}{H+1}\right)\Gamma_h,$$

where Γ_h is the autocovariance matrix defined as

$$\Gamma_h = \begin{cases} \sum_{j=|h|+1}^{n} p_j p_{j-h}^\top, & h \geq 0 \\ \sum_{j=|h|+1}^{n} p_{j-h} p_j^\top, & h < 0 \end{cases},$$

and $k(y)$ is the Parzen kernel

$$k(y) = \begin{cases} 1 - 6y^2 + 6y^3 & 0 \leq y \leq 1/2 \\ 2(1-y)^3 & 1/2 \leq y \leq 1 \\ 0 & y > 1 \end{cases}.$$

The multivariate bandwidth parameter H is selected according to Barndorff-Nielsen et al. (2008).

Appendix 1.C Simulation results

	n	$200/m$	$\overline{E}$	$\mathrm{Var}(E)$	$q_{0.25}(E)$	$q_{0.5}(E)$	$q_{0.75}(E)$
CE τ	30	0.290	0.391	0.037	0.254	0.358	0.492
	50	0.377	0.247	0.014	0.169	0.222	0.313
	70	0.493	0.215	0.013	0.131	0.203	0.268
	100	0.552	0.183	0.008	0.116	0.159	0.235
	200	0.707	0.117	0.003	0.073	0.110	0.153
	300	0.784	0.099	0.002	0.069	0.088	0.124
	500	0.844	0.072	0.001	0.047	0.064	0.095
	800	0.897	0.063	0.001	0.042	0.059	0.081
	1000	0.881	0.053	0.001	0.031	0.046	0.068
CE ρ	30	0.251	0.372	0.041	0.245	0.319	0.475
	50	0.401	0.234	0.013	0.152	0.219	0.297
	70	0.404	0.219	0.010	0.139	0.210	0.272
	100	0.463	0.178	0.007	0.111	0.169	0.240
	200	0.571	0.123	0.004	0.082	0.110	0.161
	300	0.633	0.101	0.002	0.070	0.096	0.124
	500	0.651	0.071	0.001	0.047	0.067	0.090
	800	0.714	0.062	0.001	0.043	0.061	0.077
	1000	0.707	0.054	0.001	0.033	0.048	0.069
SU	30	0.247	0.368	0.034	0.233	0.348	0.467
	50	0.292	0.259	0.018	0.172	0.241	0.316
	70	0.412	0.221	0.014	0.138	0.206	0.275
	100	0.410	0.175	0.007	0.117	0.158	0.219
	200	0.604	0.127	0.003	0.088	0.118	0.159
	300	0.680	0.098	0.002	0.068	0.087	0.122
	500	0.820	0.074	0.001	0.047	0.068	0.096
	800	0.877	0.061	0.001	0.041	0.057	0.078
OOS	30	0.160	0.218	0.015	0.142	0.192	0.256
	50	0.284	0.179	0.006	0.129	0.175	0.216
	70	0.428	0.143	0.005	0.093	0.135	0.175
	100	0.526	0.125	0.004	0.077	0.116	0.159
	200	0.743	0.090	0.002	0.059	0.085	0.112
	300	0.855	0.075	0.001	0.050	0.070	0.093
	500	0.960	0.059	0.001	0.038	0.056	0.076
	800	0.997	0.045	0.000	0.028	0.044	0.059
	1000	1.000	0.042	0.000	0.027	0.039	0.054

Table 1.A.2 Simulation results for the Gumbel copula with the structure $((123)(45))$ and $\boldsymbol{\theta} = (1.67, 1.33, 1.11)^\top$.

	n	$200/m$	$\overline{E}$	$\mathrm{Var}(E)$	$q_{0.25}(E)$	$q_{0.5}(E)$	$q_{0.75}(E)$
CE τ	30	0.325	1.843	0.681	1.211	1.799	2.260
	50	0.394	1.343	0.431	0.879	1.260	1.722
	70	0.503	1.176	0.231	0.852	1.077	1.440
	100	0.513	0.893	0.127	0.641	0.840	1.107
	200	0.714	0.636	0.080	0.453	0.597	0.749
	300	0.772	0.500	0.052	0.327	0.458	0.664
	500	0.866	0.405	0.031	0.264	0.388	0.524
	800	0.893	0.305	0.019	0.217	0.289	0.393
	1000	0.909	0.267	0.013	0.187	0.254	0.331
CE ρ	30	0.264	2.564	2.372	1.420	2.081	3.888
	50	0.403	1.800	1.661	0.943	1.384	2.176
	70	0.430	1.306	0.824	0.777	1.104	1.473
	100	0.423	0.996	0.437	0.652	0.913	1.177
	200	0.557	0.628	0.082	0.414	0.579	0.813
	300	0.637	0.532	0.049	0.361	0.524	0.663
	500	0.685	0.415	0.034	0.284	0.392	0.513
	800	0.667	0.324	0.021	0.220	0.310	0.416
	1000	0.709	0.295	0.016	0.207	0.286	0.379
SU	30	0.222	1.812	0.737	1.150	1.678	2.279
	50	0.272	1.399	0.422	0.934	1.355	1.758
	70	0.401	1.140	0.256	0.787	1.062	1.433
	100	0.425	0.886	0.147	0.593	0.830	1.111
	200	0.601	0.661	0.079	0.479	0.633	0.790
	300	0.662	0.502	0.050	0.336	0.474	0.653
	500	0.813	0.399	0.033	0.256	0.375	0.522
	800	0.905	0.304	0.019	0.214	0.294	0.385
	1000	0.917	0.268	0.013	0.185	0.255	0.331
OOS	30	0.186	1.249	0.343	0.853	1.152	1.472
	50	0.296	1.028	0.234	0.752	0.942	1.273
	70	0.442	0.891	0.149	0.595	0.846	1.135
	100	0.524	0.707	0.096	0.486	0.651	0.885
	200	0.828	0.548	0.054	0.363	0.529	0.699
	300	0.905	0.453	0.042	0.303	0.448	0.574
	500	0.985	0.362	0.025	0.259	0.353	0.473
	800	1.000	0.289	0.017	0.198	0.266	0.371
	1000	1.000	0.255	0.013	0.168	0.249	0.328

Table 1.A.3 Simulation results for the Frank copula with the structure $((123)(45))$ and $\boldsymbol{\theta} = (4.16, 2.37, 0.91)^\top$.

	n	$200/m$	$\overline{E}$	$\mathrm{Var}(E)$	$q_{0.25}(E)$	$q_{0.5}(E)$	$q_{0.75}(E)$
CE τ	30	0.335	0.521	0.087	0.320	0.466	0.614
	50	0.398	0.363	0.023	0.264	0.336	0.445
	70	0.493	0.313	0.026	0.192	0.287	0.392
	100	0.557	0.270	0.015	0.188	0.256	0.325
	200	0.772	0.172	0.005	0.117	0.160	0.217
	300	0.885	0.144	0.004	0.095	0.133	0.182
	500	0.990	0.105	0.003	0.070	0.097	0.130
	800	1.000	0.086	0.001	0.062	0.078	0.105
	1000	1.000	0.079	0.001	0.052	0.074	0.099
CE ρ	30	0.345	0.481	0.091	0.300	0.408	0.587
	50	0.427	0.358	0.030	0.238	0.321	0.441
	70	0.475	0.315	0.029	0.189	0.277	0.403
	100	0.581	0.276	0.016	0.187	0.259	0.324
	200	0.781	0.168	0.005	0.117	0.152	0.218
	300	0.881	0.143	0.005	0.097	0.127	0.173
	500	0.990	0.106	0.002	0.069	0.100	0.128
	800	1.000	0.087	0.002	0.060	0.077	0.103
	1000	1.000	0.080	0.001	0.052	0.077	0.103
SU	30	0.255	0.536	0.086	0.322	0.458	0.622
	50	0.290	0.367	0.029	0.258	0.341	0.452
	70	0.402	0.326	0.027	0.220	0.306	0.377
	100	0.418	0.274	0.016	0.185	0.251	0.329
	200	0.631	0.173	0.005	0.123	0.162	0.218
	300	0.697	0.140	0.004	0.094	0.128	0.173
	500	0.885	0.105	0.003	0.068	0.096	0.131
	800	0.990	0.086	0.001	0.062	0.078	0.105
	1000	0.990	0.078	0.001	0.051	0.074	0.098
OOS	30	0.291	0.333	0.038	0.188	0.280	0.427
	50	0.356	0.235	0.017	0.146	0.202	0.304
	70	0.512	0.219	0.014	0.138	0.189	0.264
	100	0.552	0.170	0.007	0.108	0.153	0.210
	200	0.772	0.128	0.003	0.087	0.122	0.156
	300	0.863	0.106	0.002	0.072	0.101	0.134
	500	0.953	0.086	0.001	0.059	0.080	0.106
	800	0.983	0.068	0.001	0.048	0.067	0.084
	1000	0.993	0.062	0.001	0.042	0.061	0.079

Table 1.A.4 Simulation results for the Gumbel copula with the structure $(((12)3)(45))$ and $\boldsymbol{\theta} = (1.82, 1.54, 1.33, 1.11)^\top$.

	n	$200/m$	$\overline{E}$	$\mathrm{Var}(E)$	$q_{0.25}(E)$	$q_{0.5}(E)$	$q_{0.75}(E)$
CE τ	30	0.324	2.466	1.389	1.648	2.253	2.975
	50	0.400	1.842	0.529	1.293	1.754	2.335
	70	0.459	1.542	0.398	1.106	1.404	1.950
	100	0.536	1.174	0.222	0.877	1.117	1.407
	200	0.749	0.861	0.101	0.651	0.829	1.048
	300	0.881	0.649	0.059	0.467	0.652	0.803
	500	1.000	0.529	0.043	0.379	0.505	0.637
	800	1.000	0.421	0.024	0.308	0.404	0.502
	1000	1.000	0.354	0.023	0.244	0.335	0.442
CE ρ	30	0.344	3.009	2.284	1.870	2.690	3.931
	50	0.403	2.214	1.594	1.424	1.867	2.523
	70	0.451	1.723	0.872	1.086	1.528	2.116
	100	0.625	1.321	0.457	0.944	1.240	1.561
	200	0.800	0.944	0.122	0.697	0.887	1.169
	300	0.909	0.694	0.076	0.499	0.673	0.837
	500	1.000	0.559	0.053	0.405	0.537	0.675
	800	1.000	0.434	0.028	0.319	0.416	0.522
	1000	1.000	0.384	0.023	0.267	0.361	0.477
SU	30	0.226	2.539	1.191	1.792	2.368	2.896
	50	0.273	1.863	0.617	1.292	1.767	2.330
	70	0.400	1.635	0.489	1.153	1.494	2.031
	100	0.401	1.253	0.215	0.946	1.214	1.499
	200	0.606	0.877	0.107	0.648	0.846	1.055
	300	0.719	0.665	0.065	0.466	0.668	0.814
	500	0.909	0.514	0.042	0.362	0.497	0.619
	800	0.995	0.420	0.024	0.308	0.404	0.502
	1000	0.995	0.353	0.023	0.244	0.335	0.439
OOS	30	0.261	1.829	1.017	1.188	1.582	2.216
	50	0.374	1.357	0.440	0.838	1.226	1.666
	70	0.468	1.203	0.326	0.775	1.136	1.464
	100	0.580	0.932	0.163	0.631	0.884	1.163
	200	0.802	0.720	0.085	0.530	0.702	0.889
	300	0.890	0.596	0.053	0.442	0.570	0.723
	500	0.953	0.462	0.032	0.340	0.434	0.573
	800	0.983	0.389	0.021	0.284	0.372	0.471
	1000	0.990	0.337	0.020	0.226	0.324	0.431

Table 1.A.5 Simulation results for the Frank copula with the structure $(((12)3)(45))$ and $\boldsymbol{\theta} = (4.89, 3.51, 2.37, 0.91)^\top$.

Appendix 1.D Realized volatilities and correlations

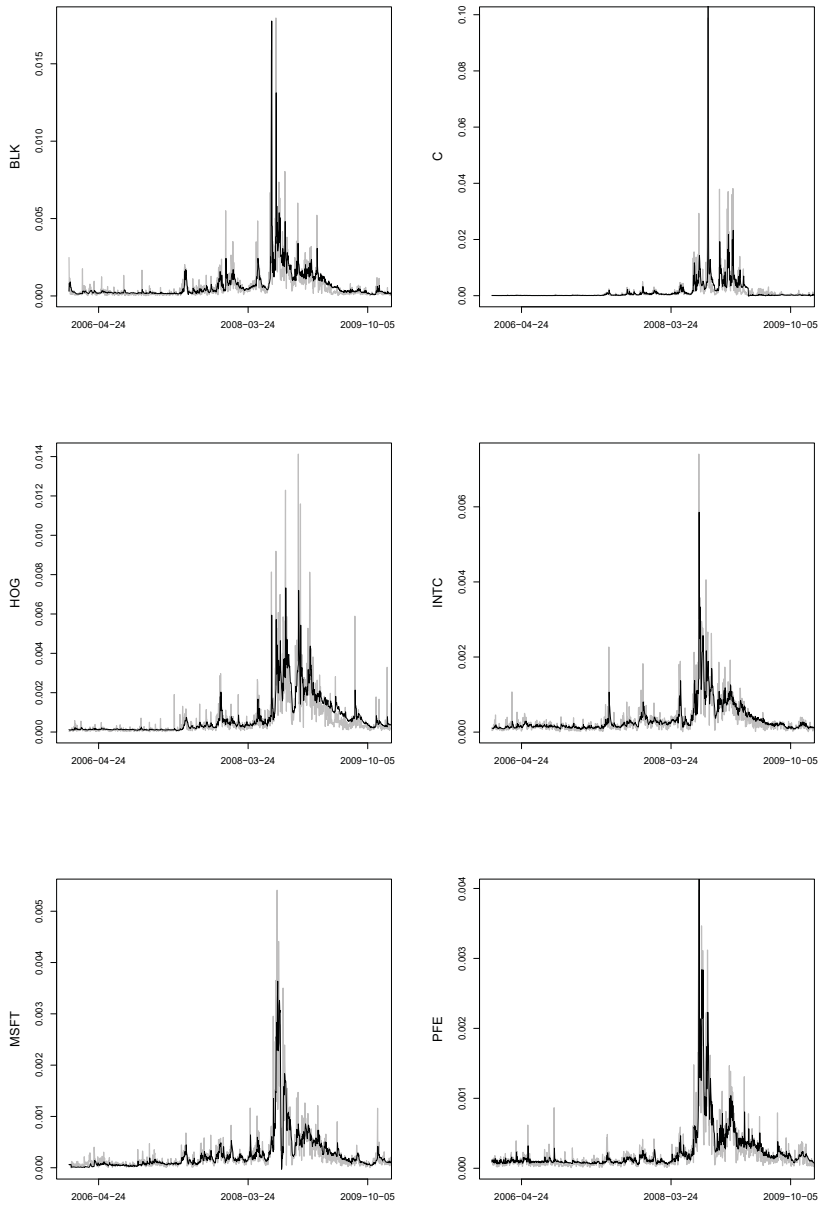


Figure 1.A.1 Time series of the selected daily realized volatilities (lines) and their one-day-ahead out-of-sample predictions (bold black).

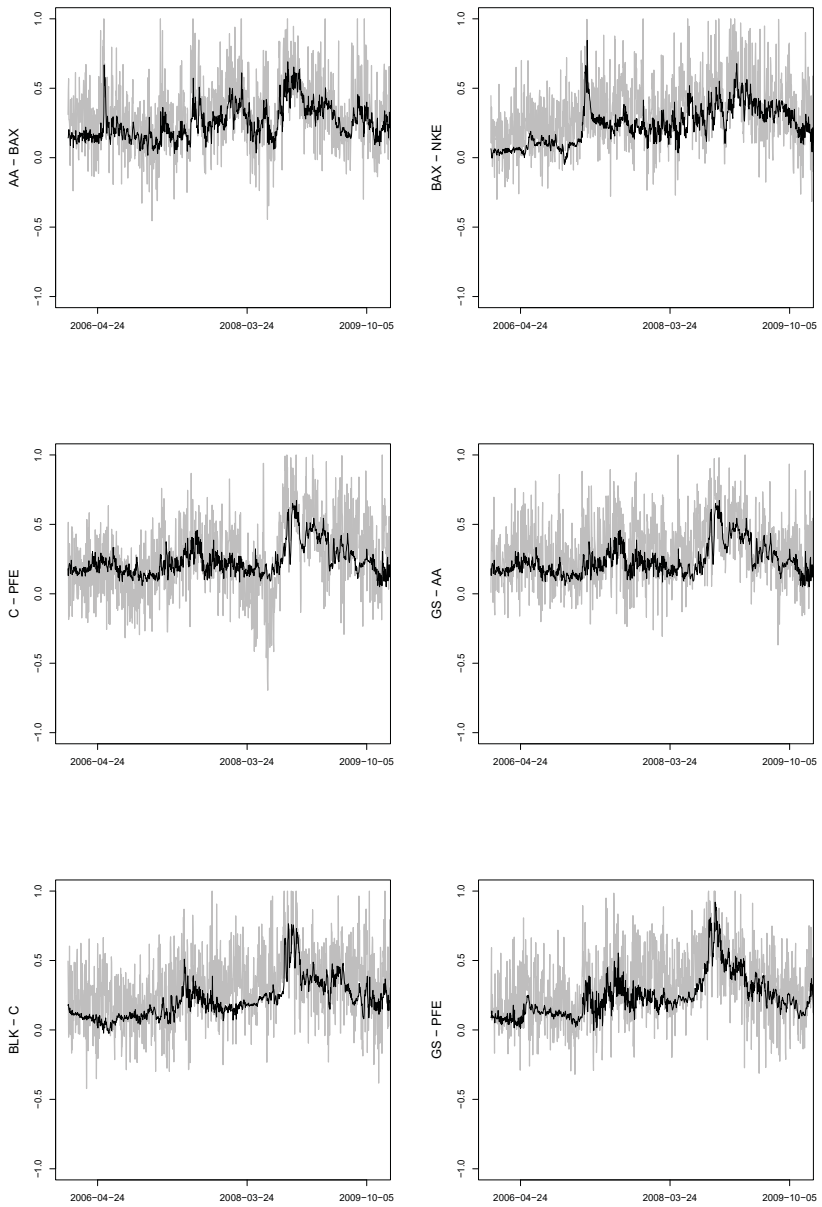


Figure 1.A.2 Time series of the selected daily realized correlations (grey) and their one-day-ahead out-of-sample predictions (bold black).

Appendix 1.E Benchmark models

Rolling window copula model The rolling window copula setting models the joint distribution of the standardized innovations $\varepsilon_t = \frac{x_{i,t}}{\sqrt{r_{i,t}}}$, $i = 1, \ldots, d$, $t = 1, \ldots, T$ via a copula with a parameter that is constant over some time period, where $x_{i,t}$ is the log-return and

$r_{i,t}$ is the realized volatility of the ith asset at day t. In this study, the Clayton copula with a rolling window of $w = 200$ days is applied. For the generalization of this approach, we refer to the locally adaptive change point algorithm of Härdle et al. (2013). This model is more flexible due to the time-varying rolling window. However, this model falls outside of the scope of this paper, due to its computational complexity.

Dynamic copula models

Copula DCC model Another essential class of VaR models incorporates the DCC models of Engle (2002). The mean process of the log-returns is assumed to be $\mu_t = 0$ and the correlation R_t of the standardized residuals $\varepsilon_t = \frac{x_{i,t}}{\sqrt{r_{i,t}}}$, $i = 1, \dots, d$, $t = 1, \dots, T$ is assumed to follow a dynamic process. These correlations are used as the input for the Student's t copula, i.e.

$$\left(\varepsilon_{1,t}, \dots, \varepsilon_{d,t}\right)^\top \sim C_d\{F_{1,t}\left(\varepsilon_{1,t}\right), \dots, F_{d,t}\left(\varepsilon_{d,t}\right); \nu, R_t\}.$$

The number of degrees of freedom ν is kept constant, while R_t is the conditional correlation matrix of the DCC model. In this study, we use a GJR-GARCH$(1,1)$ model for the univariate time series and DCC $(1,1)$ for the correlation of the log-returns. The normal and GED distributions are used to capture the margins $F_{1,t}\left(\varepsilon_{1,t}\right), \dots, F_{d,t}\left(\varepsilon_{d,t}\right)$.

The Patton (2004) model While in the previous setting the mean process is assumed to be $\mu_t = 0$, Patton (2004) suggests that the parameter of the copula should depend on a conditional mean process μ_t. This can be formalized as follows:

$$\left(\varepsilon_{1,t}, \dots, \varepsilon_{d,t}\right)^\top \sim C_d\{F_{1,t}\left(\varepsilon_{1,t}\right), \dots, F_{d,t}\left(\varepsilon_{d,t}\right); \theta_t\}, \theta_t = \Lambda\left(\sum_{i=0}^{d} \gamma_i \mu_{i,t}\right).$$

$\varepsilon_t = \frac{x_{i,t}}{\sqrt{r_{i,t}}}$, $i = 1, \dots, d$, $t = 1, \dots, T$ are the standardized residuals, γ_i, $i = 1, \dots, d$ are unknown parameters, and the function $\Lambda(\cdot)$ ensures the validity of the copula parameter, $\Lambda(x) = \exp(x)$ for the Clayton copula and $\Lambda(x) = \exp(x) + 1$ for the Gumbel copula. The marginal time series are modelled as AR(1)-GARCH$(1,1)$ processes with GED innovations.

GAS and GRAS models Even more complex models have been proposed by Creal et al. (2013) and Salvatierra and Patton (2015). In the GAS model of Creal et al. (2013), the copula parameter follows the autoregressive process

$$\Lambda\left(\theta_t\right) = \omega + \beta\Lambda\left(\theta_{t-1}\right) + \alpha s_{t-1},$$

where $s_{t-1} = S_{t-1}\delta_{t-1}$, $\delta_{t-1} = \frac{\partial \log c(u_{t-1},\theta_{t-1})}{\partial \theta_{t-1}}$ is the score function of the copula of the transformed standardized residuals $u_{i,t} = F_{i,t}(\varepsilon_{i,t})$ and S_{t-1} is a scaling matrix. The univariate time series are assumed to be GARCH$(1,1)$ with GED margins.

The updating equation of the GRAS model of Salvatierra and Patton (2015) additionally includes the realized measure $\text{RM}_t = \frac{2}{d(d-1)} \sum_{i>j}^{d} r_{ij,t}$

$$\Lambda(\theta_t) = \omega + \beta\Lambda(\theta_{t-1}) + \alpha s_{t-1} + \gamma \text{RM}_{t-1},$$

where $r_{ij,t}$ is the realized correlation.

Realized covariance models The third popular class of the models are the realized covariance models. According to the methodology proposed by Bauer and Vorkink (2011), the time series of the realized covariance matrices R_t are transformed using the matrix logarithm $A_t = \log(R_t)$. Thus, the positive-definiteness of the matrix A_t is guaranteed. In the next step, the upper-triangular elements of the matrix A_t are stacked together in a vector $a_t = \text{vech}(A_t)$, which is modelled using the HAR model. Thereafter, the vector $\widehat{a}_{t+1}$ is transformed back into the matrix $\widehat{A}_{t+1}$. The final prediction is obtained by taking the matrix exponential, i.e. $\widehat{R}_{t+1} = \text{expm}\left(\widehat{A}_{t+1}\right)$. The predicted realized covariance matrix is used as the input for a multivariate Gaussian distribution.

Another realized volatility model which uses the Cholesky decomposition instead of the logarithmic transformation is addressed in Chiriac and Voev (2011). As it performs similarly to that of Bauer and Vorkink (2011), we do not use it in the empirical part of the study.

Chapter 2

Sentiment spillover effects for US and European companies

Francesco Audrino [1], **Anastasija Tetereva** [2]

[1]Chair of Mathematics and Statistics, University of St Gallen, Bodanstrasse 6, 9000 St Gallen, Switzerland, francesco.audrino@unisg.ch

[2]Chair of Mathematics and Statistics, University of St Gallen, Bodanstrasse 6, 9000 St Gallen, Switzerland, anastasija.tetereva@unisg.ch

Abstract

The fast-growing literature on the news and social media analysis provide empirical evidence that the financial markets are often driven by sentiments rather than facts. However, the direct effects of sentiments on the returns are of main interest. In this paper, we propose to study the cross-industry influence of the news for a set of US and European stocks. The graphical Granger causality of the news sentiments-excess return networks is estimated by applying the adaptive lasso procedure. We introduce two characteristics to measure the influence of the news coming from each sector and analyze their dynamics for a period of 10 years ranging from 2005 to 2014. The results obtained provide insight into the news spillover effects among the industries and the importance of sentiments related to certain sectors during periods of financial instability.

2.1 Introduction

The influence of the news and social media in politics and economics has grown consistently over the last decades. The availability of real-time online sources and recent developments in machine learning algorithms have made the news relevant for the area of quantitative finance. The news influences the opinions and expectations of investors, which find expression in sentiments. A growing number of agencies are developing news indices, which can potentially help improve trading strategies, as recent research shows that the behavior of market participants may be more highly influenced by the news than by reality.

Some authors have explored the sensitivity of stock returns to stock-related news; see, for example, Fang and Peress (2009), Peress (2014), Akyildirim et al. (2015), Narayan and Bannigidadmath (2015), Ding et al. (2015) and Luss and d'Aspremont (2015). Akyildirim et al. (2015) describe the role of firm-specific public announcements on liquidity, price and the volatility of individual stocks. Allen et al. (2015) have analyzed how the performance of the GARCH, GJR and EGARCH can be improved by including sentiment data in the model. Additional, general conclusions have been obtained by Cahan et al. (2009) who have empirically shown that the information coming from the news can be seen as an additional factor in the Fama French factor models. The most recent research by Borovkova et al. (2016) makes an effort to construct a systematic risk indicator based on the news related to the biggest financial companies. They show that the proposed risk measure outperforms the conditional capital shortfall measure of systemic risk (SRISK) by Brownlees and Engle (2015) and the CBOE volatility index (VIX) by Brenner and Galai (1989) in signalling the periods of financial stress. Most of the above-mentioned studies concentrate, however, on the direct effects of the asset-related news on the price and do not consider spillover effects.

In this paper, we study the news spillover effects based on the news data provided by Thomson Reuters. The daily sentiment indices for 10 US sectors, 10 non-US sectors and 5 countries are considered together with the prices of more than 100 stocks. The analyzed time interval ranges from the January 1, 2005 to December 31, 2014 and covers the global financial crisis, the US debt-ceiling crisis, and the European sovereign debt crisis.

We conduct a rolling window analysis of the cross-industry influence of news sentiments. The influence of news on the stock excess returns is defined by means of the graphical Granger causality, which is estimated by constructing a sparse network. In order to reduce the number of false positive edges of the network, the adaptive lasso methodology is applied for the estimation of the networks with a related testing procedure introduced in Audrino and Camponovo (2015). Two characteristics describing the relevance and the strength

of the news coming from an individual sector or a specific country are suggested, and their dynamic behavior is studied. The news sentiments coming from US and non-US industrial sectors show similar behavior and are highly correlated; for this reason, the spillover effects are analyzed separately for the US and the European companies, and the results are compared.

In this study, we provide strong empirical evidence that the class of stock-relevant sentiments is wider than just the stock-specific announcements and the news coming from the related sector. We show that the returns of the whole industry are driven by the news coming from several sectors. Moreover, we investigate how the importance of the news changes over time, getting stronger just before periods of financial turbulence, which can be seen as an early warning signal for investors.

The paper is organized as follows: The first section features an introduction to the Thomson Reuters MarketPsych news data, and Section 2 discusses how the signal can be extracted from the noisy news. The methodology related to the estimation of the graphical Granger causality by means of sparse regression models is provided in Section 3. Full results are presented in Section 4. Section 5 contains an empirical illustration of how the predictive power of time series models can be improved by augmenting the conditional mean equation by specific sentiment indices. Finally, we summarize the main contribution of the paper.

2.2 TRMI construction

The sentiment data used in the current paper are provided by Thomson Reuters MarketPsych and include the MarketPsych indices (TRMI) ranging from 2005 to 2014. In this section, we provide more details on the construction of the MarketPsych index and its characteristics.

In analogy to a variety of approaches in the sentiment literature, the TRMI is constructed using the lexical analysis or the so-called "bag of words" technique: the frequency of a series of predefined words in the text is counted and the quantitative index is extracted by means of sentiment dictionaries such as code libraries by Tim Loughran and Bill McDonald and Harvard General Inquirer. For more details on text analysis as applied to economic news we refer to, for example, Loughran et al. (2009), Loughran and McDonald (2011), Loughran and McDonald (2013) and Loughran and McDonald (2014). In contrast to the most popular approaches, the TRMI is sensitive to grammatical structures and accounts for correlations among the words.

TRMI is a multidimensional index which considers the sentiments beyond positive and negative. Each TRMI index is constructed based on a set of over 4 000 variables (for example,

Ambiguity, *EarningsUpFuture*, *AccountingBad*, *AccountingGood*) which are generated by the linguistic machine learning algorithms. In the next step, a numerical value which considers the tense, proximity and many other multipliers is assigned to each variable. Finally, the variables are associated with the assets. To illustrate this procedure, consider the following example from Peterson (2016):

Analysts expect Mattel to report much higher earnings next quarter.

The linguistic analysis of the sentence will be performed in the following steps:

1. The entity "*Mattel*" will be associated with the ticker *MAT*.
2. Word *"earnings"* is associated with the variable *Earnings*.
3. Word *"expect"* assigns future tense to the phrase.
4. Word *"higher"* is an Up-Word.
5. Word *"higher"* is multiplied by 2 due to the presence of the word "much."
6. Word *"higher"* is associated with the word "earnings" due to proximity.

The above-described procedure leads to the score of 2 for the variable *EarningsUpFuture* for the ticker *MAT*.

In the next step, the variables (*PsychVar*) are combined into the TRMI sentiment indices which are computed as a ratio of the sum of all relevant variables to the absolute values of all TRMI-contributing variables called *Buzz*. In addition, each variable is classified as additive or subtractive. Thus, the index is computed in the following way:

$$TRMI = \frac{\sum_{c \in C(A),\, v \in V} \left[I(v,t) \times PsychVar_v(c) \right]}{Buzz\,(A)}, \tag{2.1}$$

where $I(v,t) = 1$, if the variable is additive and $I(v,t) = -1$, if the variable is subtractive, $Buzz(A) = \sum_{c \in C(A), v \in V} |PsychVar_v(c)|$, V is the set of all variables *PsychVar*, A denotes a specific asset. $C(A)$ is the set of all constituents of A, i.e. set of all entities which are relevant for the specific asset; for example, Mattel is a constituent of the Nasdaq 100 index. It is worth noting that a single variable can contribute to multiple TRMI; for example, the above-mentioned variable *earningsUpFuture* is a constituent of such TRMI sentiment indices as *Optimism* and *fundamentalStrength*, and Mattel is a constituent of the Consumer Goods sector and the Nasdaq 100. Additionally, such characteristics as relevance and novelty are

taken into account when the variables are summed up in the index. The weights of all constituents are constantly recalculated, giving bigger coefficients to the more influential companies.

In the current paper, we make use of the TRMI sentiment index for 10 industries and 5 countries which are listed in Appendix 2.A. This analysis considers a daily frequency, as we are interested in the analysis of the global financial crisis which goes back to 2008 when tick-by-tick data were not available for all the asset classes we are interested in.

2.3 Extracting signal from the news sentiment

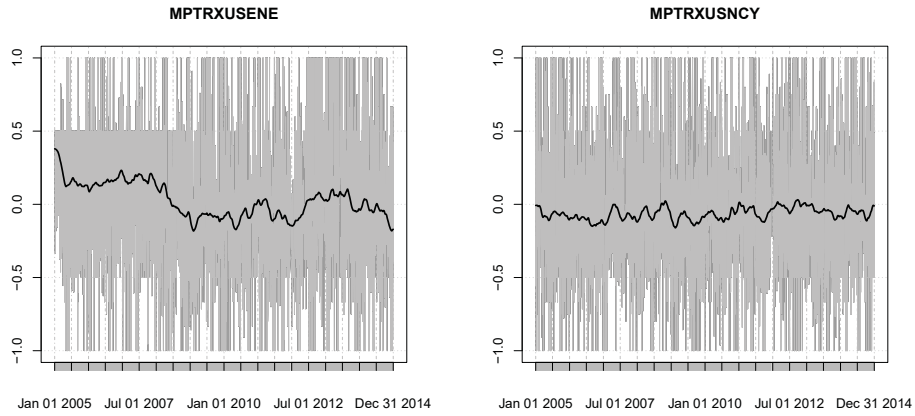


Figure 2.1 News sentiment for the US Energy and Non-Cyclical Consumer Goods and Services and the Kalman smoothed news sentiment.

As can be seen from Figure 2.1, the news data are very noisy and cannot be directly used for modelling. Two main approaches for extracting the information from the noise which appeared in the recent sentiment literature could be potentially used.

The first approach is the moving average convergence/divergence oscillator (MACD). It was first developed by Appel (2003) and has been applied to sentiment data by, for example, Peterson (2016), Kirange et al. (2016) and Lugmayr and Gossen (2013). Becker (2016b) shows that 10-30 MACD of the TRMI sentiment about Starbucks has an influence on its price. Becker (2016a) points out the connection between Volkswagen share prices and the 30-200 MACD of the TRMI Media sentiment. However, different studies consider different time windows for the long and the short components of the MACD with a lack of a clear statistical and economic reasoning for the (data-driven) choice of the time window.

In the current study, we follow instead a more rigorous approach which employs the Kalman filter first introduced by Kalman (1960) and discussed by Borovkova and Mahakena (2015) and Borovkova et al. (2016) for applications to sentiment data. The real unobserved sentiment μ_t is extracted from the noisy news y_t by applying the Local Level model by Durbin and Koopman (2001). The starting point of the model is the following system of equations:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \varepsilon_t \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right), \\ \mu_{t+1} &= \mu_t + \eta_t, \eta_t \sim \mathrm{N}\left(0, \sigma_\eta^2\right). \end{aligned} \tag{2.2}$$

In the sentiment literature, it is assumed that economic agents are at least as intelligent as animals and aggregate the news as a function of time decay. This fact is motivated by the findings in the behavioral literature; see Nickerson (2011) and Reilly et al. (2012), for example. The sentiment can be therefore considered to follow some autoregressive process.

In (2.2), the first equation corresponds to the noisy observed news y_t, whereas the second equation is the signal equation which corresponds to the unobserved sentiment μ_{t+1}, $t = 1, \ldots, T$. The state moments $\tilde{\mu}_t = \mathrm{E}\left(\mu_t | \mathcal{F}_{t-1}\right)$ and $P_t = \mathrm{Var}\left(\mu_t | \mathcal{F}_{t-1}\right)$ are computed recursively by solving the following equations:

$$\begin{aligned} \varepsilon_t &= y_t - \tilde{\mu}_t, F_t = P_t + \sigma_\varepsilon^2, \\ K_t &= \frac{P_t}{F_t}, \\ \tilde{\mu}_{t+1} &= \tilde{\mu}_t + K_t \varepsilon_t, P_{t+1} = P_t\left(1 - K_t\right) + \sigma_\eta^2, \\ \tilde{\mu}_1 &= \mu_1, P_1 = e^7. \end{aligned} \tag{2.3}$$

Thus, the unobserved state μ_t is updated each time a new noisy observation y_t arrives. The state at time t is calculated by exponentially weighting the previous states.

In the next step, the states are estimated by applying the Kalman smoother and solving the following backward recursion equations:

$$\begin{aligned} \widehat{\mu}_t &= \tilde{\mu}_t + P_t r_{t-1} \\ r_{t-1} &= \frac{\varepsilon_t}{F_t} + L_t r_t; N_{t-1} = F_t^{-1} + L_t^2 N_t, \\ L_t &= 1 - K_t, \\ V_t &= P_t - P_t^2 N_{t-1}, \end{aligned} \tag{2.4}$$

where $t = 1, \ldots, T$, $r_T = 0$ and $\mathrm{N}\left(\widehat{\mu}_t | \mathcal{D}\right)$ is the conditional density of μ_t with $\widehat{\mu}_t = \mathrm{E}\left(\mu_t | \mathcal{F}_T\right)$ and $V_t = \mathrm{Var}\left(\mu_t | \mathcal{F}_T\right)$. For the technical details of the estimation procedure we refer to Durbin and Koopman (2001). In the current study, the Kalman smoother is applied to the daily averages of the news sentiment. The original data and the Kalman smoothed versions for US and Energy and Non-Cyclical Consumer Goods and Services sectors are shown in the Figure 2.1. It is worth noting, that for most sectors news sentiments of the US and non-US sectors behave in a very similar way. The smoothed US and non-US financial news sentiments are given in Figure 2.2. It is evident that the correlation between the two series is very high and it would be difficult to disentangle the influence of the US and non-US news on the different industries under investigation in a linear regression framework. We decided, therefore, to analyze the US and the European markets separately.

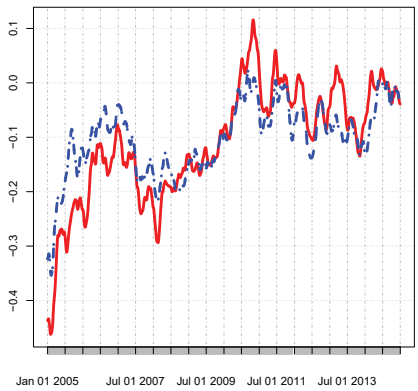


Figure 2.2 Kalman smoothed news sentiment for US (red solid) and non-US (blue dotted) financial sector.

2.4 Penalized estimation of the sentiment networks

In the current study, we are interested in analyzing if and in which cases the individual time series of the companies' prices can be potentially influenced by the news on sectors and countries. Cross-industry news spillover effects have not been directly addressed in the literature thus far. Most of the recent studies investigate the direct influence of the sentiment about an asset on the asset itself and do not consider that, for example, news coming from the energy sector can influence financial companies. Borovkova and Mahakena (2015) and Borovkova (2015) have shown that extreme positive and extreme negative sentiment days influence the future price momentum of natural gas and that there

is a complex relationship between the arrival of the news and the price jumps. Similar results have been developed for the energy markets in Borovkova and Lammiman (2010). Erawan (2015) uses the sentiment data for different sectors as an input for classification and regression trees and finds empirical evidence that the trading strategy can be improved by including the sector-specific news data into the model. However, no spillover effects are studied. Borovkova and Mahakena (2015) show that abnormal stock returns calculated from Fama-French and Carhart factor models might be explained by the specific sentiment data. The recent study by Borovkova et al. (2016) makes an effort to construct the risk measure based on the financial news sentiment data only and tests for the ability of this indicator to forecast the stress in the market.

As mentioned above, the aim of this work is to conduct a systematic empirical investigation of the spillover effect of the sentiment coming from different industries and countries on individual stock prices. For this purpose, we employ sparse graphical models in order to obtain insight into the joint causal relationship between the individual time series.

First, we introduce the notation specific to the network literature. In the first step, a graph $\mathcal{G} = \langle V, E \rangle$ is considered, where V is the set of p nodes corresponding to each variable $X^1, X^2, \ldots, X^p$ and $E \subset V \times V$ is the set of the edges corresponding to the pairwise association of all variables. Causal relations are usually represented by directed graphs: for example, if the variable X^i is assumed to be influenced by a set of variables X^j, $j = 1, \ldots, p$, $j \neq i$, the associations $j \to i$ are studied. Thus, the dependence of X^i and a set of the nodes is stated as:

$$X^i = f_i\left(X^1, X^2, \ldots, X^p\right) + \varepsilon_i,\ i = 1, \ldots, p, j \neq i. \tag{2.5}$$

In (2.5), ε_i is an error term. The function f_i is usually assumed to be linear. In this way, the estimation of the network can be reduced to the estimation of the individual regressions

$$X^i = \sum_{j \neq i} \beta^{ij} X^j + \varepsilon_i, \tag{2.6}$$

where the associations $j \to i$ are expressed in terms of the coefficients β^{ij}.

The individual regressions can be estimated by OLS (ordinary least squares) or more sophisticated methods. For example, Peng et al. (2009) suggest estimating nonzero partial correlations by joint sparse regression models.

Where the causal relationship among the individual time series is important, it is useful to consider the concept of Granger causality, first introduced by Granger (1980) and now widely discussed in the literature. Per definition, the time series process $\{X_t^j\}$, $t = 1, \ldots, T$

Granger causes the time series process $\{X_t^i\}$, $t = 1, \ldots, T$ if the regression of X_t^i on the past values of X_t^i and X_t^j gives a better fit than the regression of the past values of X_t^i.

One way to define Granger causality for a network where X_t^i is the response time series which is caused by a high-dimensional time series $X_t^{[-i]} = X_t^1, X_t^2, \ldots, X_t^p, t = 1, \ldots, T, j \neq i$ is mentioned in Lozano et al. (2009) and Arnold et al. (2007) in the context of bioinformatics. It has been proposed to test for Granger causality within the network by applying some kind of variable selection procedure. In particular X_t^j is Granger causing X_t^i if the lags of X_t^j are selected by some sparse estimation procedure for any time lag $l \in L$, $j = 1, \ldots, p$, $t = 1, \ldots, T$. Formally speaking, the following regression model is estimated:

$$X_t^i = \sum_{j=1}^{p} \sum_{l \in L} \beta_l^{i,j} X_{t-l}^j + \varepsilon_t^i \; i = 1, \ldots, p. \tag{2.7}$$

In practical applications, the networks often appear to be high-dimensional and the number of nodes can exceed the sample size. This fact leads to inaccuracy and overfitting of the estimates. The remaining part of this section discusses how these drawbacks can be overcome by means of sparse estimation methods.

A natural way to improve the performance of the estimator in a high-dimensional regression model is to introduce a regularization penalty. Many new regularized methods have been developed in the literature over the last decades, including the least absolute shrinkage and selection operator (lasso) estimator by Tibshirani (1996), SCAD by Fan and Li (2001), elastic net by Zou and Hastie (2005), fused lasso by Tibshirani et al. (2005) and extensions of the lasso models, such as adaptive lasso by Zou (2006) and group lasso by Yuan and Lin (2006). The application of sparse estimation methods for the estimation of high-dimensional networks in computer biology is widely discussed in the literature; see, for example, Gustafsson et al. (2005) , Shimamura et al. (2007), Li and Li (2008), Friedman et al. (2008) and Jacob et al. (2009). The present work employs the lasso approach originally formulated by Tibshirani (1996), i.e.

$$\widehat{\beta}^i(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\sum_{t=\max(L)+1}^{T} \left[X_t^i - \sum_{j=1}^{p} \sum_{l \in L} \beta_l^{i,j} X_{t-l}^j \right]^2}{T - \max(L)} + \lambda \|\beta^i\|_1 \right), \tag{2.8}$$

where $\|\beta^i\|_1 = \sum_{j=1}^{p} \sum_{l \in L} |\beta_l^{i,j}|$ and $\lambda > 0$ is a penalty parameter. As a result of l_1 penalization, the lasso solution is sparse, i.e. some coefficients are set exactly to zero. However, lasso has been shown to lack the consistency for selecting the relevant variables and to produce small false positive non-zero coefficients.

The two-stage adaptive lasso procedure introduced by Zou (2006) corrects the behavior of the lasso and reduces the number of false positives by re-weighting the penalty function, i.e.:

$$\widehat{\beta}^{i}_{adaptive}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\sum_{t=\max(L)+1}^{T} \left[X_t^i - \sum_{j=1}^{p} \sum_{l \in L} \beta_l^{i,j} X_{t-l}^j \right]^2}{T - \max(L)} + \lambda \sum_{j=1}^{p} \sum_{l \in L} \frac{|\beta_l^{i,j}|}{|\widehat{\beta}^{i,j}_{initial,\, l}|} \right), \tag{2.9}$$

where $\widehat{\beta}^{i,j}_{initial,\, l}$, $j = 1, \ldots, p$, $l \in L$ is the initial estimator from the first step of the procedure. Thus, if $\widehat{\beta}^{i,j}_{initial,\, l}$ is large, the adaptive lasso employs a small penalty for the $\beta_l^{i,j}$, $j = 1, \ldots, p$ and improves the estimation of the effective variables. The simple OLS, the ridge regression or any other consistent estimator can be used to obtain the initial estimates. The theoretical properties and the details of the estimation algorithms can be found in Bühlmann and Van De Geer (2011).

In the current work, we include several lags of each regressor in the regression model in order to check for Granger causality. In particular, we are interested in whether the news for a given sector and not the specific lags are Granger causing the price of the asset. We implement the adaptive lasso procedure with the OLS coefficients as the initial estimators. In addition, we employ the testing procedure based on the finite sample properties of the adaptive lasso developed by Audrino and Camponovo (2015) to reduce the number of false positive selected variables.

In the current framework, the causality of the news data and its lags on the prices of the assets in different sectors needs to be estimated. For this reason, we define two sets of variables. Let $\{X_t^i\}$, $t = 1, \ldots, T$, $i = 1, \ldots, p_1$ be the set of the prices or excess returns and let $\{X_t^j\}$, $t = 1, \ldots, T$, $j = p_1 + 1, \ldots, p$ be the set of the news sentiment and $\{X_{t-l}^i\}$ be the corresponding lags. Thus, the network consists of p nodes. As we are interested in Granger causality of the news rather than returns of other assets, the edges connecting the price variable X_t^i to the lags of other prices X_{t-l}^j, $j = 1, \ldots, p_1$, $j \neq i$ are set to zero. This restrictive assumption is introduced because the sentiment data of the industries contain the price information of the biggest companies. If both price and sentiment lags are included in the regression, the estimated coefficients might be misleading. The combination of the Granger causality concept with the adaptive lasso variable selection procedure results in the following algorithm:

1. Define the input as the p_1-dimensional vectors of returns $\{X_t^i\}_{t=1,\ldots,T}$, $i = 1, \ldots, p_1$ and p_2-dimensional vector of sentiment $\{X_t^j\}_{t=1,\ldots,T}$, $j = p_1 + 1, \ldots, p$.

2. Set the adjacency matrix corresponding to the network $G = \langle V, E \rangle$ equal to the zero matrix.

3. For $i = 1, \ldots, p_1$:

   - apply the adaptive lasso variable selection with its related testing procedure for false positives to the model

$$X_t^i = \sum_{l \in L} \beta_l^{i,i} X_{t-l}^i + \sum_{j=p_1+1}^{p} \sum_{l \in L} \beta_l^{i,j} X_{t-l}^j + \varepsilon_t^i, \tag{2.10}$$

     where L is the set of the predefined lags.

   - If for X_t^i at least one lag of X_t^j is selected as significant, place the edge $X^j \to X^i$ into $E, l \in L$.

2.5 Results

In this section, the news spillover effects of 10 industrial sectors are analyzed using the methodology of the graphical Granger causality. The data set contains the stock excess returns and TRMI sentiment indices on 10 sectors listed in Appendix 2.A and 5 countries, namely, US, China, Germany, Italy, and Greece. The country information plays the role of a control variable. 78 US companies and 78 European companies are investigated, which corresponds to approximately 8 companies per sector. The full list of the companies can be found in Appendix 2.B. In order to extract the signal from the noisy sentiment data, the Kalman smoothing approach described in Section 2.3 is applied. All the data are standardized after applying the Kalman smoother. As mentioned above, the sentiments for the US and non-US industries show similar trends over time. Therefore, the analysis is performed separately for the US and European market. The rates of returns in excess are calculated by applying the CAPM model by Sharpe (1964) using as a benchmark the S&P 500. The preliminary analysis showed that the results for the prices and excess returns using different benchmarks coincide, which is consistent with the discussion in Erawan (2015). Therefore, the same benchmark is used for the US and European companies. The daily data ranging from the January 1, 2005 to December 31, 2014 are used. A daily rolling window approach is employed in order to analyze how the connectedness between the news and the excess returns changes over time. The size of the rolling window is set to be equal to 200 trading days, which is the shortest possible size addressed in the literature; see Audrino and Knaus

(2016) for more details. The small size of the rolling window is motivated by the possible time breakpoints.

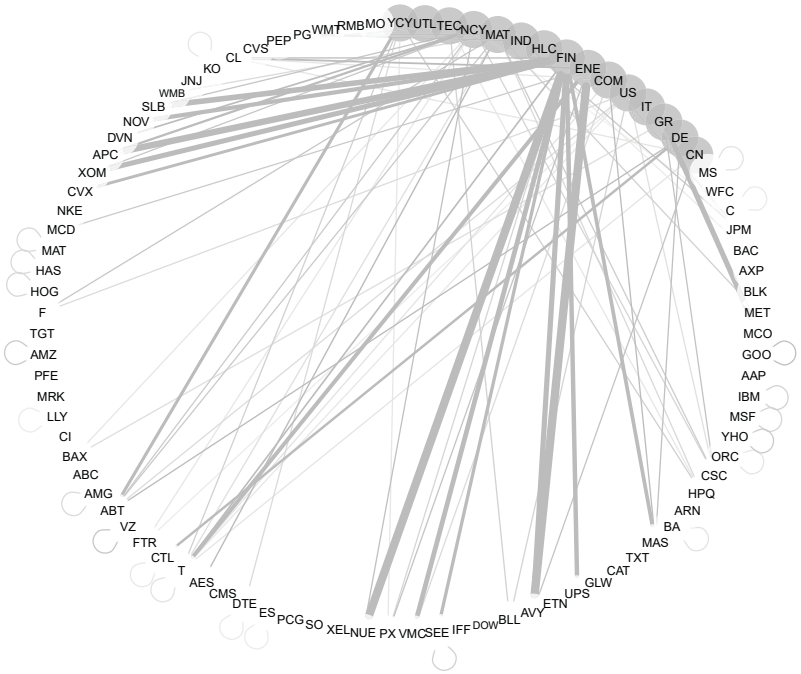


Figure 2.3 The graphical Granger network for the set of US assets on Nov. 3, 2010.

Employing the notation of Section 2.4, we say that the news from one industry is Granger causing the excess returns in the other industry if the lags of the sentiment on one sector are selected by the adaptive lasso procedure (2.10) for the excess returns of the assets from the other industry, i.e. $X^j \longrightarrow X^i$, $i = 1, \ldots, p_1$, $j = p_1 + 1, \ldots, p$. In the current studies, we distinguish between direct effects and spillover effects, i.e. if the sentiment of a particular sector is Granger causing the excess returns of the assets from the same sector, we call them direct effects, whereas in case the assets belong to different sectors we call them spillover effects.

This study focuses on the information spillover effect due to sentiment. For this reason, the network at time t, $t = 1, \ldots, T$ is constructed by regressing the individual excess returns X_t^i on their own lags X_{t-l}^i, $l \in L$ and the set of the news sentiment lags on the sectors and the countries X_{t-l}^j, $j = p_1 + 1, \ldots, p$, $l \in L$. As explained above, lags of excess returns of other stocks are not taken into consideration in the individual regressions. The estimation is performed using the R package `glmnet` by Friedman et al. (2009). In this study, we suggest

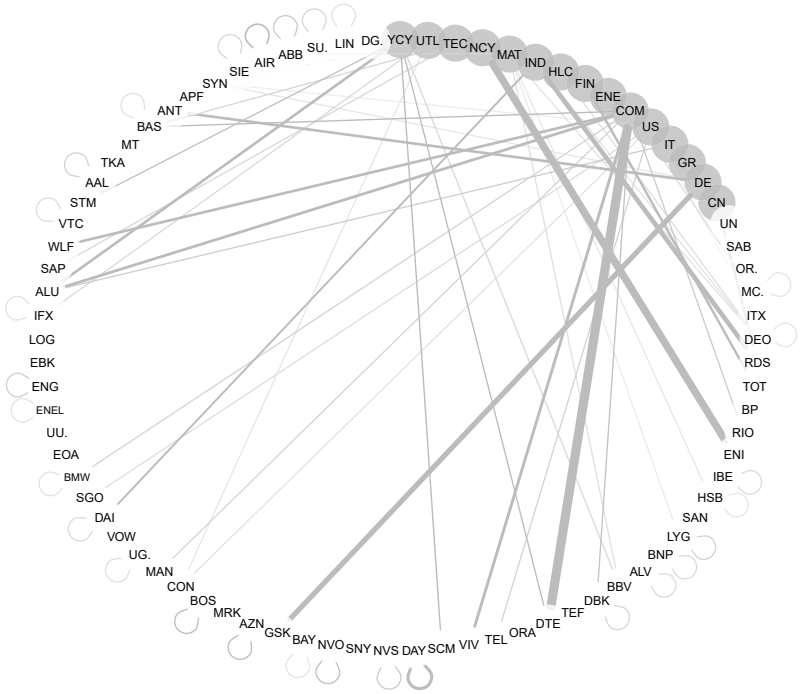


Figure 2.4 The graphical Granger network for the set of European assets on Nov. 3, 2010.

including 1 day, 1 week and 1 month lags in the individual regressions, i.e. $l \in \{1, 5, 22\}$. Sentiment data of the same day (0 lag) are not included in the analysis because they might contain the information about the prices and thus overestimate the causality of the news. The choice of 1 month as mentioned above is motivated by the recent studies in the sentiment literature which reveal that the shocks in the sentiment can drive the price of the underlying during the period of the next month; see Borovkova et al. (2016), for instance.

The examples of the networks obtained by the adaptive lasso procedure for the US and European stocks on November 3, 2010 are presented in the Figures 2.3 and 2.4, respectively. The industries and countries are grey-colored in the graph, the abbreviations are given in Appendix 2.A, and the companies are white-colored. The whole network is constructed by regressing excess returns of each company on the past values of excess returns and lags of the sentiments of the sectors and countries. Therefore, the connections coming from the sentiment (grey nodes) to the returns (white nodes) represent the Granger causality of the sentiment indices to the excess returns. If the node is self-connected, for example Amazon (AMZ) in Figure 2.3, the lags of the excess returns of the company itself were selected by the adaptive lasso procedure. The date always corresponds to the last day of the

rolling window. The US network (2.3) is more connected than the European network (2.4). However, in general European networks contain more edges; that is, more coefficients are estimated as non-zero by the lasso procedure, on average 14.6 % nonzero coefficients for Europe versus 13.9 % for USA. The countries that have more outgoing edges in Figure 2.3 (Germany, Greece and Italy) are considered to have more influence on the excess returns during this time period. The most relevant industries during this time period are Financials and Industrials. Similar results are observed for the European companies in the network in Figure 2.4.

In order to analyze the spillover effects over time, one needs to introduce some measures of connectedness among the excess returns and the news sentiments of the sectors and countries. Dynamic analysis of these measures can provide insight into the importance of particular sectors and countries during the global financial crisis starting in 2008 and the subsequent European sovereign debt crisis. In this study, we are interested in two characteristics of the sentiments: the relevance and the strength. We propose to measure the relevance of each sentiment by the number of outgoing edges. This measure reflects the share of regressions which have selected the considered sentiment index as an active variable. Therefore, the relevance is not informative for expressing the strength of the causality. For this reason, we define the second characteristic (strength) as the mean absolute value of the outgoing edges.

Formally speaking, the overall relevance of the sentiment variable X^j can be defined from the estimated coefficients over p_1 regressions of the form (2.10):

$$R\left(X^j \to \{X^1, \ldots, X^{p_1}\}\right) = \frac{\sum_{i=1}^{p_1} \mathbb{1}\{|\widehat{\beta}_1^{i,j}| + |\widehat{\beta}_5^{i,j}| + |\widehat{\beta}_{22}^{i,j}|\}}{p_1}, \tag{2.11}$$

$j = p_1 + 1, \ldots, p$,

$$\mathbb{1}\{x\} = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases}$$

is an indicator function.

The overall strength of the l-th lag of the sentiment variable X^j can be defined as the average absolute value of the coefficients of the outgoing edges:

$$S_l\left(X^j \to \{X^1, \ldots, X^{p_1}\}\right) = \frac{\sum_{i=1}^{p_1} |\widehat{\beta}_l^{i,j}|}{\sum_{i=1}^{p_1} \mathbb{1}\{|\widehat{\beta}_l^{i,j}|\}}, \; l \in \{1, 5, 22\}, \tag{2.12}$$

if $R\left(X^j \to \{X^1, \ldots, X^{p_1}\}\right)$ is different than zero and is zero if the variable is irrelevant. Thus, this characteristic represents the average absolute value of the coefficients $\widehat{\beta}_l^{i,j}$ in the regressions of the excess returns of company i for the sentiment j. If only the sum of the absolute values were considered, it would be impossible to distinguish between the nodes with many links with small coefficients and the nodes with a small number of links and bigger coefficients. Speaking in terms of graphical representation, the relevance characterizes the average number of the outgoing links and the strength characterizes the average width of the link. If the sentiment (node) has small relevance and high strength, it is selected as significant in a small number of regressions but estimated with relatively high coefficients. In definition (2.12), the strength of the regulator is defined separately for each lag.

Similarly, the same measures can be defined for particular groups of companies. In this case, the average is taken over the companies contained in the specific group. For example, if the average is taken over the companies corresponding to the financial sector with the Industrials sentiment as the variable under consideration, the relevance and the strength will show the spillover effect of the Industrials-related news to the financial sector.

2.5.1 US results

	FIN	TEC	IND	MAT	UTL	COM	HLC	NCY	ENE	YCY
asset	0.72	0.77	0.72	0.69	0.61	0.71	0.64	0.68	0.66	0.71
MPTRXUSFIN	0.16	0.17	0.17	0.19	0.17	0.17	0.16	0.18	0.22	0.15
MPTRXUSTEC	0.11	0.13	0.15	0.12	0.17	0.20	0.16	0.14	0.23	0.14
MPTRXUSIND	0.12	0.16	0.17	0.13	0.14	0.19	0.13	0.14	0.15	0.14
MPTRXUSMAT	0.10	0.07	0.15	0.13	0.13	0.11	0.12	0.12	0.18	0.11
MPTRXUSUTL	0.09	0.10	0.14	0.15	0.10	0.17	0.12	0.10	0.21	0.09
MPTRXUSCOM	0.13	0.12	0.11	0.10	0.11	0.18	0.11	0.11	0.18	0.10
MPTRXUSHLC	0.15	0.14	0.21	0.14	0.12	0.16	0.12	0.14	0.13	0.13
MPTRXUSNCY	0.13	0.12	0.20	0.17	0.14	0.17	0.14	0.14	0.20	0.13
MPTRXUSENE	0.14	0.13	0.16	0.14	0.11	0.16	0.17	0.15	0.18	0.19
MPTRXUSYCY	0.12	0.10	0.12	0.11	0.12	0.14	0.11	0.12	0.16	0.11
US	0.14	0.16	0.16	0.13	0.11	0.17	0.15	0.17	0.17	0.11
IT	0.12	0.14	0.12	0.12	0.12	0.13	0.12	0.15	0.13	0.11
GR	0.17	0.15	0.16	0.16	0.15	0.17	0.11	0.16	0.14	0.14
DE	0.14	0.12	0.15	0.12	0.14	0.15	0.13	0.12	0.21	0.13
CN	0.09	0.10	0.14	0.12	0.09	0.10	0.08	0.13	0.11	0.12

Table 2.1 Mean relevance of the news sentiments of the US sectors and selected countries for the US companies by sector ranging from Jan. 1, 2005 to Dec. 31, 2014.

To begin with, we discuss the average news spillover effects within the US market. In Table 2.1 we present the cross-industry mean relevance of the news over the full time period under consideration. For example, the number 0.12 in the first column and the tenth row

means that YCY news is Granger causing the excess returns of 12 companies from 100 in the financial sector on average. It is worth noting that the stocks are mostly self-connected, i.e. the own lags of the excess returns of the stocks are chosen to be relevant for the future excess returns by the adaptive lasso procedure (see the numbers in the first row). From Table 2.1, one can observe that the relevance of the country-related news spreads almost uniformly among the sectors. There is weak evidence of a bigger influence of US news on NCY and Energy. A similar effect is observed for the industries; however, a slightly higher relevance of the financial news for all sectors is evident. It is worth mentioning that the results of Table 2.1 should be interpreted with caution, as the average is taken over a long period of time. The time interval of almost 10 years might contain several structural breakpoints in the relevance of the news.

The dynamics of the relevance over time becomes clearer from Figure 2.5 where the rolling window relevance is presented for the selected countries and industries. It can be concluded that, in general, the relevance is not constant and has a fluctuating behavior. For some sectors, several breakouts can be observed. First, the relevance of the own lags of the stock drops during the period of the global financial crisis. This finding coincides with the results shown by Audrino and Knaus (2016) in the context of the HAR volatility model. Second, the relevance of Technology and Industrials rises right after the crisis.

Similarly, the relevance of the country-specific news shows a fluctuating behavior. The exception is the growing relevance of the US-related news around 2008 and the high relevance of the Germany-related news around 2008-2010 and 2013. These results support the common belief of the global leading role played by the US information flow during the financial crisis and by the Germany-related news (as the leading European country) during the financial and subsequent European sovereign debt crises.

In contrast to the analysis of relevance, a closer look at the strength of the sector-related news gives better insight into the cross-industry news spillover effects. The mean cross-industry strength of the 1, 5 and 22 lags of the news sentiments is presented in Appendix 2.C.

From Figure 2.6 it is clearly observed that the news of Financials and Energy have the strongest causal influence on the excess returns in all sectors. Such a high strength of the financial news could be explained by the fact that the global financial crisis and the European sovereign debt crisis are part of our sample period. Interestingly, the lags related to the stock itself are estimated with smaller coefficients than the lags of the sentiments. This finding supports the result that the past values of the news contain additional information

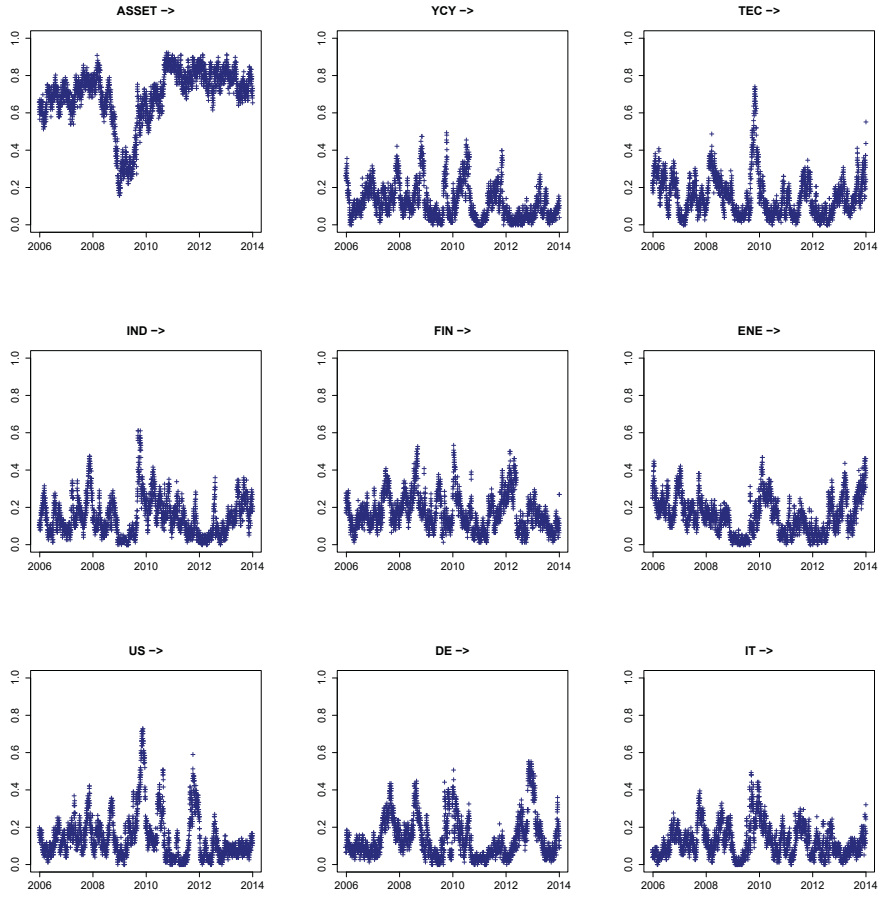


Figure 2.5 The overall relevance of the news sentiments of the selected sectors and countries on the US stocks ranging from Jan. 1, 2005 to Dec. 31, 2014.

about the future stock returns; moreover, the importance of the news rises during financial turbulence and macroeconomic recessions.

To get a deeper understanding, we present the rolling window results averaged over all stocks in Figure 2.6 for one-day lag variables: results for other lags are similar. The strength of the financial news starts rising in 2007, showing that the growing influence of the news can be considered an early warning of future instability. Higher average strength of YCY news is observed during the first half of the sample and decreases thereafter. This might be explained by the fact that the aggregated consumption of US households started to grow rapidly at the end of 2005, went through recession in 2008, and continued to increase in 2009. Consumption has shown stable upward trends since 2009; therefore, the news on the YCY sector is less important in the second half of the period. Moreover, it is observed that the strength of Technology and Industrials grows right after the recession period, as

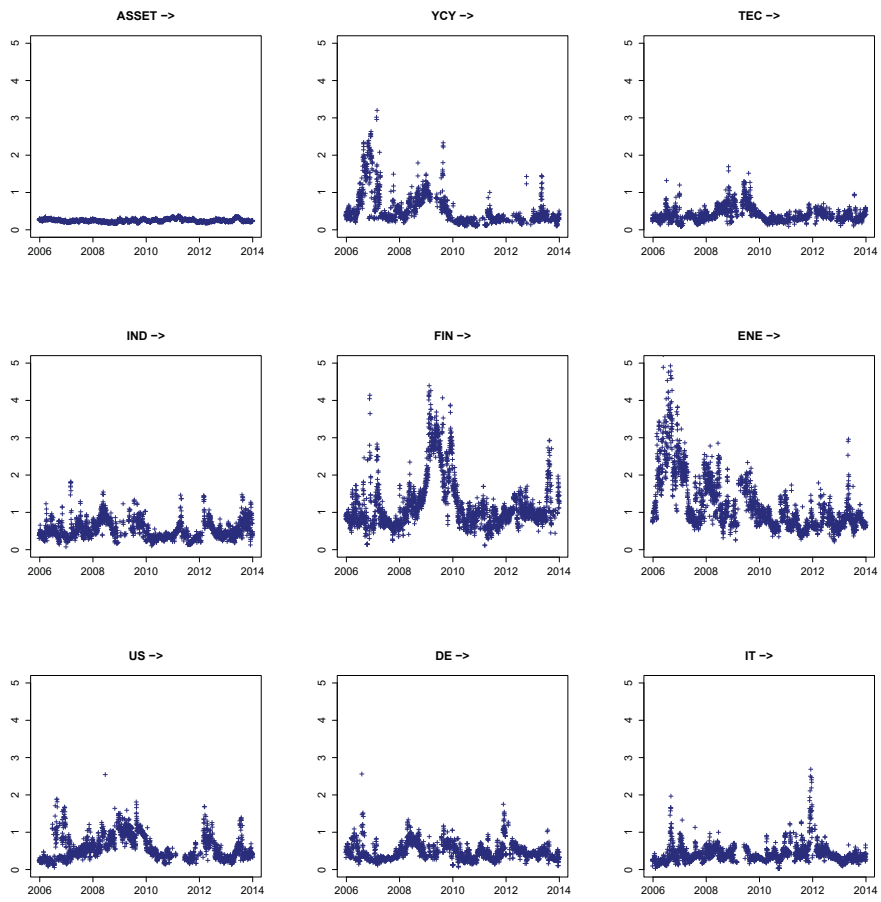


Figure 2.6 The overall strength of the 1 day lags of the news sentiments on the selected US sectors and countries for the US companies ranging from Jan. 1, 2005 to Dec. 31, 2014.

innovations in these sectors contribute notably to the recovery process and gain additional attention of the media. The smooth decline in the strength of the energy sector-related news on the stock returns might be caused by the introduction of US energy independence politics. The reduction in imports might be transferred to greater confidence in the market and, as a result, less weight for the energy-related news. Figure 2.6 depicts the strength of selected countries, and it appears that the US news shows relatively high overall strength and relativity. On the other hand, short peaks with low persistence in the strength of the news related to Italy and Germany are observed. The first peak in the strength of Italy corresponds to the currency crisis of 2006 in Europe: there were some concerns that several countries might leave the Eurozone, which could have potentially lead to the total collapse of the Euro. This event would influence the currency market worldwide, in particular in the US. The second peak could potentially be caused by the European debt crisis, which

	FIN	TEC	IND	MAT	UTL	COM	HLC	NCY	ENE	YCY
asset	0.70	0.74	0.70	0.71	0.59	0.68	0.69	0.78	0.70	0.76
MPTRXFIN	0.13	0.17	0.17	0.16	0.12	0.11	0.19	0.17	0.14	0.14
MPTRXTEC	0.13	0.16	0.08	0.15	0.09	0.14	0.18	0.09	0.18	0.10
MPTRXIND	0.14	0.12	0.12	0.15	0.13	0.10	0.17	0.12	0.18	0.11
MPTRXMAT	0.15	0.15	0.12	0.15	0.12	0.14	0.18	0.12	0.21	0.11
MPTRXUTL	0.17	0.12	0.17	0.17	0.10	0.11	0.24	0.18	0.15	0.18
MPTRXCOM	0.20	0.17	0.13	0.14	0.12	0.13	0.19	0.13	0.19	0.12
MPTRXHLC	0.14	0.13	0.15	0.11	0.10	0.09	0.16	0.13	0.15	0.14
MPTRXNCY	0.14	0.14	0.14	0.11	0.14	0.14	0.21	0.12	0.17	0.13
MPTRXENE	0.18	0.15	0.13	0.16	0.12	0.15	0.18	0.16	0.16	0.13
MPTRXYCY	0.18	0.16	0.16	0.17	0.14	0.17	0.23	0.15	0.22	0.22
US	0.16	0.12	0.11	0.16	0.12	0.15	0.16	0.11	0.15	0.14
IT	0.15	0.15	0.17	0.19	0.10	0.12	0.16	0.16	0.15	0.13
GR	0.16	0.16	0.16	0.19	0.12	0.15	0.21	0.17	0.18	0.12
DE	0.12	0.15	0.14	0.15	0.11	0.14	0.18	0.11	0.19	0.13
CN	0.12	0.11	0.12	0.12	0.10	0.09	0.17	0.10	0.13	0.12

Table 2.2 Mean relevance of the news sentiments of the European sectors and selected countries for the European companies by sector ranging from Jan. 1, 2005 to Dec. 31, 2014.

severely affected such countries as Italy and Greece. 2010 in Italy was characterized by a rapid decrease in GDP and increasing unemployment. The fluctuations in Germany's strength can be explained by the global financial crisis of 2008, the European debt crisis of 2010 and the decision of Germany in 2012 to provide more support to its European partners for more centralized control over the Eurozone. The results for other sectors and countries are presented in Appendix 2.D. A detailed analysis of the strength by sector can be obtained from the authors upon request.

The findings presented above must be interpreted with caution. It is important to bear in mind that the omitted market factor could be responsible for the obtained results. Herein, we discuss the extent to which the spillover remains unchanged after controlling for a market-wide sentiment. The analysis described above is replicated, and the corresponding lags of the VIX index are incorporated into the model. The VIX – CBOE volatility index – measures the expected volatility implied by the S&P 500 index option and therefore reflects uncertainty in the market. This index is often called a 'fear index' due to its ability to mimic investors' sentiments. The mean values of the relevance and strength of all sectors after controlling for VIX influence are presented in Appendix 2.E. These results confirm the previous finding – the biggest spillover effects originate in the financial and energy sectors. The main difference of the VIX augmented model is that the adaptive lasso procedure selects the lags of the assets' returns less frequently, while VIX lags are relevant for 90% of the assets on average. However, due to the small values of the estimated coefficients, the overall strength-related results remain unchanged. It is evident that an additional VIX covariate adds more variance to the pattern of the strength. This difference can be explained by the

more volatile nature of the VIX index in comparison to excess returns and a smoothed sentiment index.

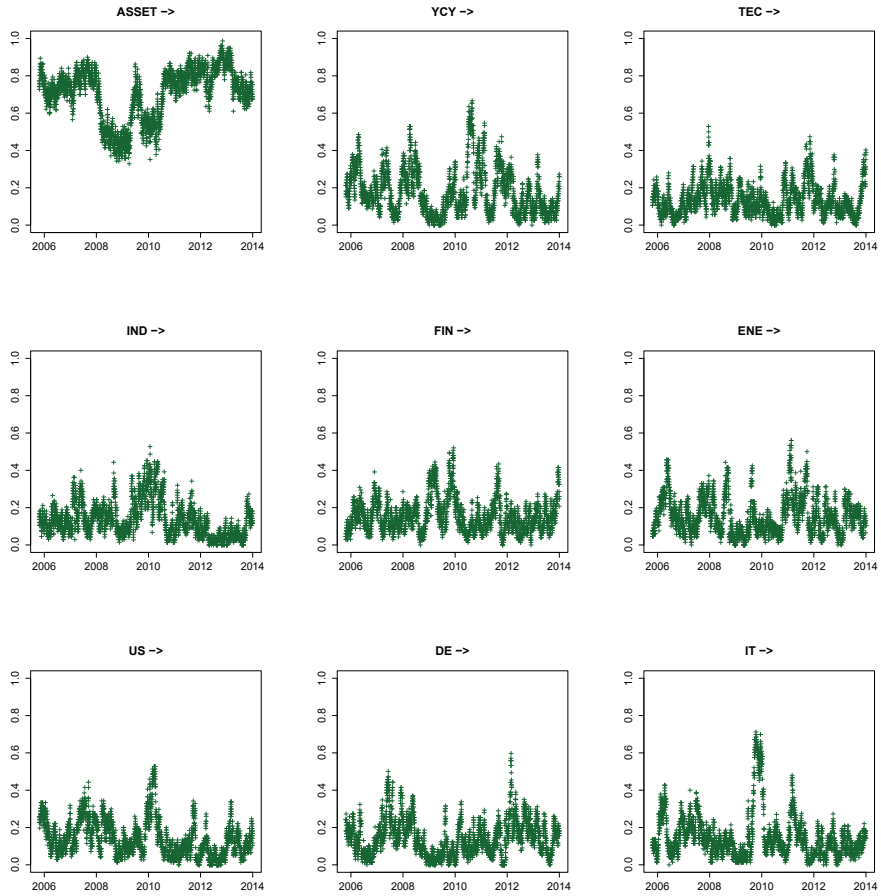


Figure 2.7 The overall relevance of the news sentiments of the selected sectors and countries on the European stocks ranging from Jan. 1, 2005 to Dec. 31, 2014.

2.5.2 EU results

Comparing the news spillover effects in the European market to the US market, we observe the similarity in terms of the relevance of the news. However, the strength of the news sentiments is less strong on average; see the results summarized in Table 2.2. The fluctuating behavior of the average relevance of some sectors for all stocks is shown in Figure 2.7. The analysis of the strength of the sectors in terms of the average absolute lasso estimated coefficients coincides with the results for the US market. The news related to Financials and Energy seems to be important for all other sectors; see Appendix 2.C. Similarly to the US

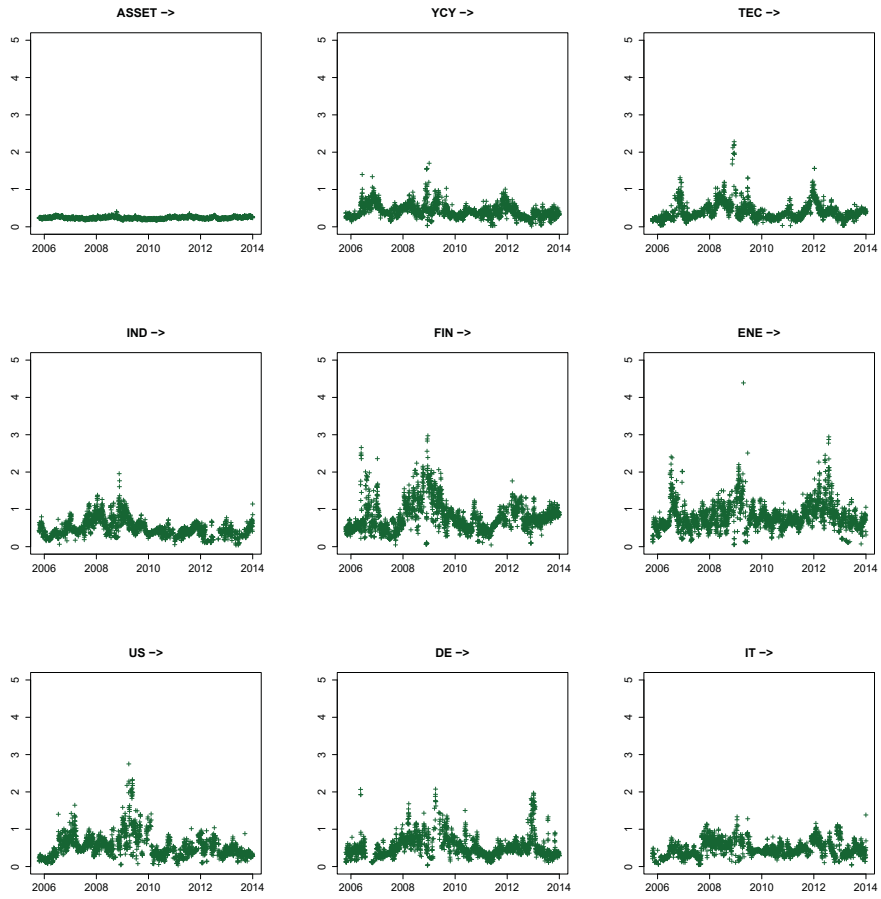


Figure 2.8 The overall strength of the 1 day lags of the news sentiments on the selected US sectors and countries for the US companies ranging from Jan. 1, 2005 to Dec. 31, 2014.

market, the strength of the financial news rises before the crisis and reaches a maximum in 2008; see Figure 2.8 and Appendix 2.C. However, the pattern of the strength of the energy news is fluctuating in contrast to the US market. This might be explained by the fact that, in contrast to the US, the European energy market is less independent and concerns over the prices spread out to the other sectors. Full detailed results of the analysis can be obtained from the authors upon request.

It can be concluded that the news sentiments related to several important sectors provide additional information about the future stock prices. For each market, several important sectors can be defined. The news spillover effects coming from these influential sectors seem to be at least as important as the direct effects. Moreover, the relevance of these spillover effects increases during periods characterized by general economic instability and/or financial market turbulence.

2.6 An empirical illustration

The results of the analysis presented in the previous section show that the sentiment data of several sectors have a significant information component, which can be used to improve the prediction of the assets' returns. This section illustrates the impact of the various sentiment series on the conditional mean and the conditional variance equations of ARMA-GARCH models. It is important to note that the purpose of this section is to give an insight into possible practical applications of the results presented earlier rather than draw general conclusions about sentiment augmented time series models.

Further on, a parsimonious ARMA$(1,1)$-GARCH$(1,1)$ model with Gaussian innovations is applied to the series of log returns of several firms. The choice of the model is due to the fact that lower order GARCH models are used in most applications. Moreover, this model has been selected according to the Akaike information criterion in the majority of cases. This benchmark model is compared to extensions in which the lag of the sentiment index is included in the model as an exogenous variable. The considered model has the following specification:

$$\begin{aligned} Y_t &= \alpha_0 + \alpha_1 Y_{t-1} + \beta_1 \varepsilon_{t-1} + \gamma_1 S_{t-1} + e_t, \\ \operatorname{Var}(e_t|\mathcal{F}_{t-1}) &= a_0 + a_1 \sigma_{t-1}^2 + b_1 e_{t-1}^2 + g_1 S_{t-1}, \end{aligned} \tag{2.13}$$

where $e_t = \sigma_t \varepsilon_t$, with ε_t being a sequence of iid random variables with mean 0 and variance 1, Y_t is the series of the log returns, S_{t-1} is the sentiment index of the pre-defined sector, $t = 1, \ldots, T$, and α_0, α_1, β_1, γ_1, a_0, a_1, b_1, g_1 are the parameters to be estimated. For further details on time series models, we refer to Tsay (2005), Andersen et al. (2009), Francq and Zakoian (2011) and original work on the GARCH model by Bollerslev (1986).

The specification (2.13) results in three competing models. If $\gamma_1 = 0$ and $g_1 = 0$, the ARMA(1,1)-GARCH(1,1) model without exogenous sentiment variables is considered. If $g_1 = 0$, the sentiment variable is included in the conditional mean equation. If both γ_1 and g_1 need to be estimated, the sentiment index is included in the conditional mean and the conditional variance equations.

The analysis performed in the previous sections has shown that the sentiment spillover effects dominate the direct effect during periods of financial turbulence. It has been empirically shown that the sentiment data of the financial and energy sectors drive the market during certain periods. In order to validate this conclusion, we separately estimate the models with the firm-specific sector's sentiment and the sentiment data coming from the financial and energy sectors.

	MSE	DM p-value	MSPE	DM p-value	MSPE	DM p-value
	1.11.2008 - 1.11.2009		1.11.2009 - 1.11.2010		1.11.2010 - 1.11.2011	
IFF (Basic Materials)						
$\gamma_1 = 0, g_1 = 0$	$4.18 \cdot 10^{-5}$		$5.89 \cdot 10^{-5}$		$3.06 \cdot 10^{-5}$	
$g_1 = 0, S =$ MPTRXFIN	$4.96 \cdot 10^{-5}$	$3.76 \cdot 10^{-1}$	$9.70 \cdot 10^{-6}$	$8.05 \cdot 10^{-8}$	$4.29 \cdot 10^{-5}$	$1.51 \cdot 10^{-1}$
$S =$ MPTRXFIN	$5.01 \cdot 10^{-5}$	$3.46 \cdot 10^{-1}$	$9.70 \cdot 10^{-6}$	$8.16 \cdot 10^{-8}$	$4.06 \cdot 10^{-5}$	$1.64 \cdot 10^{-1}$
$g_1 = 0, S =$ MPTRXMAT	$8.94 \cdot 10^{-5}$	$6.00 \cdot 10^{-3}$	$1.47 \cdot 10^{-4}$	$3.12 \cdot 10^{-1}$	$6.01 \cdot 10^{-5}$	$1.02 \cdot 10^{-3}$
$S =$ MPTRXMAT	$6.70 \cdot 10^{-5}$	$3.00 \cdot 10^{-3}$	$1.44 \cdot 10^{-4}$	$3.30 \cdot 10^{-1}$	$5.64 \cdot 10^{-5}$	$3.10 \cdot 10^{-3}$
WMB (Basic Materials)						
$\gamma_1 = 0, g_1 = 0$	$1.05 \cdot 10^{-4}$		$1.39 \cdot 10^{-4}$		$8.76 \cdot 10^{-5}$	
$g_1 = 0, S =$MPTRXFIN	$7.91 \cdot 10^{-5}$	$3.17 \cdot 10^{-1}$	$6.93 \cdot 10^{-5}$	$4.67 \cdot 10^{-3}$	$6.73 \cdot 10^{-5}$	$1.72 \cdot 10^{-1}$
$S =$ MPTRXFIN	$7.53 \cdot 10^{-5}$	$2.46 \cdot 10^{-1}$	$6.94 \cdot 10^{-5}$	$4.78 \cdot 10^{-3}$	$6.75 \cdot 10^{-5}$	$1.75 \cdot 10^{-1}$
$g_1 = 0, S =$MPTRXMAT	$6.80 \cdot 10^{-5}$	$1.74 \cdot 10^{-1}$	$1.69 \cdot 10^{-4}$	$3.43 \cdot 10^{-1}$	$2.17 \cdot 10^{-5}$	$1.26 \cdot 10^{-9}$
$S =$MPTRXMAT	$7.75 \cdot 10^{-5}$	$3.18 \cdot 10^{-1}$	$1.83 \cdot 10^{-4}$	$1.93 \cdot 10^{-1}$	$2.14 \cdot 10^{-5}$	$1.23 \cdot 10^{-9}$

Table 2.3 MSPE of (2.13) (p-values of DM test compared to the model with $\gamma_1 = g_1 = 0$).

After estimating the models based on 200-day rolling windows and obtaining the one-day-ahead forecast series, the models are compared in terms of their predictive power. Their predictive power is expressed in terms of the mean squared prediction error (MSPE). The two-sided test of Diebold and Mariano (1995) (DM test) is applied to the squared errors to check whether the model provides a statistically significant improvement in comparison to the ARMA(1,1)-GARCH(1,1), which is always used as the benchmark.

Four companies, belonging to two industrial sectors (Basic Materials and Healthcare), are used for the illustrative purposes of this section: International Flavors & Fragrances Inc. (IFF), Williams Companies Inc. (WMB), Chevron Corporation (CVX), and Cigna Corporation (CI). The selection of the time intervals is motivated by Figure 2.9. It is evident that the financial news has a strong effect on the companies related to Basic Materials in 2010 and is less important for the returns of this sector in 2008 and 2009. The energy news is of high importance for both Basic Materials and Healthcare at the beginning of the analysed time period (2007) and is less influential at the end of the period (2013).

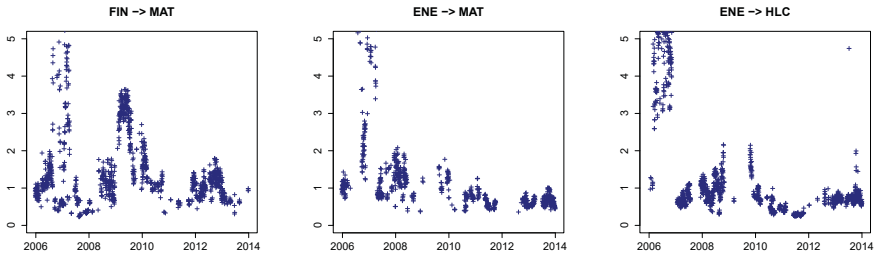


Figure 2.9 The strength of 1 day lags of the news sentiments of the FIN and ENE sectors for the MAT and HLC US companies ranging from the 1st of Jan., 2005 to Dec. 31, 2014.

The mean squared prediction errors and the p-values of the two-sided DM test for the log returns of IFF and WMB are presented in Table 2.3. The comparison is always made with respect to the ARMA(1,1)-GARCH(1,1) model, i.e. model (2.13) with $g_1 = \gamma_1 = 0$. It is evident that including the exogenous variable of financial news in the conditional mean equation improves the predictive power of the ARMA-GARCH model from November 1, 2009 to November 1, 2010, which corresponds to the period of strong influence of the financial sector. An improvement is not observed during the year before and the year after the above mentioned period. The results for the conditional variance equation are not unambiguous and further research should be addressed to this question. In the majority of the analysed cases, exogenous news indices do not provide any further improvement of the predictive power of the model. Including sector related news (Basic Materials) in the model does not significantly improve the predictive power from November 1, 2009 to November 1, 2010, when the financial news was driving the market. The influence of the sector-specific news during other periods is more evident, but not always statistically significant.

Similar results are observed for CVX and CI, see Table 2.4. It is evident that including energy-related news in the conditional mean equation improves the predictive power of the ARMA-GARCH model in 2007 and is not relevant in 2014. The MSPE is not significantly reduced when the sentiment data are included in the conditional variance model. Moreover, sector-related sentiment data do not reduce the MSPE significantly in the period when the market was driven by the news coming from the Energy sector.

The results for the considered firms support the conclusion of the previous section and suggest that sentiment spillover effects dominate the direct effects in periods when the market is driven by the news coming from a different sector. Therefore, the time series models which are augmented by including the news of influential sectors might show better predictive power for returns. The periods when the sentiment indices of particular industries are more informative can be found by applying the procedure described in Section 2.4. The illustrative examples discussed here support the results of Section 2.5, which suggests the importance of the information contained in sentiment data for asset pricing theory.

Conclusion

The goal of this paper is to investigate the cross-industry patterns of the news and stock returns, and in particular to analyze how the news about one industry influences the stock returns in the other industries. For this purpose, the graphical Granger model has

	MSPE	DM p-value	MSPE	DM p-value
	1.12.2006 - 1.12.2007		1.12.2013 - 1.12.2014	
CVX (Basic Materials)				
$\gamma_1 = 0, g_1 = 0$	$1.82 \cdot 10^{-4}$		$2.47 \cdot 10^{-5}$	
$g_1 = 0, S =$MPTRXENE	$1.13 \cdot 10^{-4}$	$2.96 \cdot 10^{-2}$	$2.73 \cdot 10^{-5}$	$5.21 \cdot 10^{-1}$
$S =$MPTRXENE	$1.11 \cdot 10^{-4}$	$2.21 \cdot 10^{-2}$	$2.85 \cdot 10^{-5}$	$3.47 \cdot 10^{-1}$
$g_1 = 0, S =$MPTRXHLC	$1.60 \cdot 10^{-4}$	$5.59 \cdot 10^{-1}$	$4.09 \cdot 10^{-5}$	$1.23 \cdot 10^{-2}$
$S =$MPTRXHLC	$2.19 \cdot 10^{-4}$	$3.70 \cdot 10^{-1}$	$2.37 \cdot 10^{-5}$	$7.99 \cdot 10^{-1}$
CI (Healthcare)				
$\gamma_1 = 0, g_1 = 0$	$9.58 \cdot 10^{-5}$		$1.21 \cdot 10^{-5}$	
$g_1 = 0, S =$MPTRXENE	$3.89 \cdot 10^{-5}$	$4.9 \cdot 10^{-3}$	$9.00 \cdot 10^{-6}$	$2.70 \cdot 10^{-1}$
$S =$MPTRXENE	$3.88 \cdot 10^{-5}$	$4.8 \cdot 10^{-3}$	$9.50 \cdot 10^{-6}$	$3.50 \cdot 10^{-1}$
$g_1 = 0, S =$MPTRXMAT	$1.60 \cdot 10^{-4}$	$3.9 \cdot 10^{-2}$	$4.09 \cdot 10^{-5}$	$2.29 \cdot 10^{-6}$
$S =$MPTRXMAT	$9.77 \cdot 10^{-5}$	$9.5 \cdot 10^{-1}$	$3.70 \cdot 10^{-6}$	$1.12 \cdot 10^{-5}$

Table 2.4 MSPE of (2.13) (p-values of DM test compared to the model with $\gamma_1 = g_1 = 0$).

been applied to the Kalman smoothed sentiment data on 10 US and 10 non-US industries and on the excess returns of 78 US and 78 European companies. The sentiment data on several countries have been included in the analysis as control variables. The adaptive lasso procedure has been applied to estimate the return-news networks. The network-based measures reflecting the relevance and the strength of each news source have been proposed and analyzed over a period of 10 years by employing a rolling window approach.

We found empirical evidence that the relevance of the news coming from different sectors shows a fluctuating behavior and spreads evenly among the industries. Moreover, our results show that the strength of the influence of the news on some sectors grows just before periods of economic and financial instability and reaches a maximum during crises. Interesting patterns are observed in the causality of the financial and energy sentiments. These sentiments can be seen as the most influential, and the spillover effects from the sectors dominate the direct effects. Estimation results show that the overall connectedness among the stock returns and the news is stronger for the US market than for the European market. The importance of sentiment spillover effects has been empirically illustrated. We showed that the sentiment indices of specific industries can be successfully used to improve the predictive power of time series models for returns.

In future research we plan to relax the assumption of predefined lags in the graphical Granger model and to apply the same adaptive lasso methodology to test the significance of arbitrary lags. This could be especially interesting to show the persistence of the news coming from different sectors. Moreover, the study of the nonlinear or quantile cross-industry dependencies among the news could yield further insights into the analysis of the impact of direct sentiment effects as well as spillover sentiment effects on companies' excess returns.

Appendices

Appendix 2.A TRMI sentiment indices

Asset Code	Description
MPTRXUSENE/MPTRXENE	US/non-US Energy (ENE)
MPTRXUSMAT/MPTRXMAT	US/non-US Basic Materials (MAT)
MPTRXUSIND/MPTRXIND	US/non-US Industrials (IND)
MPTRXUSYCY/MPTRXYCY	US/non-US Cyclical Consumer Goods and Services (YCY)
MPTRXUSNCY/MPTRXNCY	US/non-US Non-Cyclical Consumer Goods and Services (NCY)
MPTRXUSFIN/MPTRXFIN	US/non-US Financials (FIN)
MPTRXUSHLC/MPTRXHLC	US/non-US Healthcare (HLC)
MPTRXUSTEC/MPTRXTEC	US/non-US Technology (TEC)
MPTRXUSCOM/MPTRXCOM	US/non-US Telecommunications Services (COM)
MPTRXUSUTL/MPTRXUTL	US/non-US Utilities (UTL)
CN	China
DE	Germany
GR	Greece
IT	Italy
US	USA

Table 2.A.1 Thomson Reuters MarketPsych sentiment indices used for the analysis.

Appendix 2.B List of the companies

Name	TRBCEconomicSector
Bangkok Airways PCL	Industrials
Asset Acceptance Capital Corp	Industrials
Apple Inc	Technology
Abcam PLC	Healthcare
Abbott Laboratories	Healthcare
AES Corp	Utilities
Amgen Inc	Healthcare
Amazon.com Inc	Consumer Cyclicals
Anadarko Petroleum Corp	Energy
Avery Dennison Corp	Industrials
American Water Works Company Inc	Utilities
American Express Co	Financials
Boeing Co	Industrials
Bank of America Corp	Financials
Baxter International Inc	Healthcare
BlackRock Inc	Financials
Ball Corp	Basic Materials
Citigroup Inc	Financials
Caterpillar Inc	Industrials
Cigna Corp	Financials
Colgate-Palmolive Co	Consumer Non-cyclicals
CMS Energy Corp	Utilities
CVS Caremark Corp	Consumer Non-cyclicals
Cisco Systems Inc	Technology
CenturyLink Inc	Telecommunication Services
Chevron Corp	Energy
Dow Chemical Co	Basic Materials
DTE Energy Co	Utilities
Devon Energy Corp	Energy
Ford Motor Co	Consumer Cyclicals
Frontier Communications Corp	Telecommunication Services
Corning Inc	Technology
General Motors Co	Consumer Cyclicals
Hasbro Inc	Consumer Cyclicals
Harley-Davidson Inc	Consumer Cyclicals
Hewlett-Packard Co	Technology
International Business Machines Corp	Technology
International Flavors & Fragrances Inc	Consumer Non-cyclicals
Johnson & Johnson	Healthcare
JPMorgan Chase & Co	Financials
Coca-Cola Co	Consumer Non-cyclicals
Eli Lilly and Co	Healthcare
Masco Corp	Consumer Cyclicals
Mattel Inc	Consumer Cyclicals
McDonald's Corp	Consumer Cyclicals
Metlife Inc	Financials
Merck KGaA	Healthcare
Altria Group Inc	Consumer Non-cyclicals
Merck & Co Inc	Healthcare
Morgan Stanley	Financials
Microsoft Corp	Technology
Nike Inc	Consumer Cyclicals
National Oilwell Varco Inc	Energy
Eversource Energy	Utilities
Nucor Corp	Basic Materials
Oracle Corp	Technology
PG&E Corp	Utilities
PepsiCo Inc	Consumer Non-cyclicals
Pfizer Inc	Healthcare
Procter & Gamble Co	Consumer Non-cyclicals
Praxair Inc	Basic Materials
Sealed Air Corp	Basic Materials
Schlumberger NV	Energy
Southern Co	Utilities
AT&T Inc	Telecommunication Services
Target Corp	Consumer Cyclicals
Textron Inc	Industrials
United Parcel Service Inc	Industrials
Vulcan Materials Co	Basic Materials
Verizon Communications Inc	Telecommunication Services
Wells Fargo & Co	Financials
Williams Companies Inc	Energy
Wal Mart Stores Inc	Consumer Cyclicals
Xcel Energy Inc	Utilities
Exxon Mobil Corp	Energy
Yahoo! Inc	Technology
Cambridge Antibody Technology Group PLC	Healthcare
Bell Aliant Inc	Telecommunication Services

Table 2.A.2 The US companies used in the current studies.

Name	TRBCEconomicSector
Anglo American PLC	Basic Materials
Alcatel Lucent SA	Technology
Antofagasta PLC	Basic Materials
Anglo Pacific Group PLC	Basic Materials
Daisy Group PLC	Telecommunication Services
Electricite de France SA	Utilities
Eni SpA	Energy
Glencore PLC	Energy
Iberdrola SA	Utilities
Industria de Diseno Textil SA	Consumer Cyclicals
Nestle SA	Consumer Non-cyclicals
Rio Tinto PLC	Basic Materials
Roche Holding AG	Healthcare
SABMiller PLC	Consumer Non-cyclicals
Swisscom AG	Telecommunication Services
STMicroelectronics NV	Technology
Sirius Minerals PLC	Basic Materials
Syngenta AG	Basic Materials
Telecity Group PLC	Technology
Telenor ASA	Telecommunication Services
United Utilities Group PLC	Utilities
Vitec Group PLC	Technology
Wolfson Microelectronics PLC	Technology
Abengoa Yield PLC	Utilities
ABB India Ltd	Industrials
Aditya Birla Minerals Ltd	Basic Materials
ArcelorMittal SA	Basic Materials
AstraZeneca PLC	Healthcare
Banco Bilbao Vizcaya Argentaria SA	Financials
GlaxoSmithKline PLC	Healthcare
Merck KGaA	Healthcare
Novartis	Healthcare
Banco Santander SA	Financials
Total Energy Services Inc	Energy
Anheuser-Busch Companies LLC	Consumer Non-cyclicals
HSBC Holdings	Financials
Lloyds Banking Group	Financials
BNP Paribas	Financials
Allianz	Financials
UBS Group	Financials
Deutsche Bank	Financials
Logitech International	Technology
Infineon Technologies	Technology
SAP SE	Technology
Siemens	Industrials
Airbus	Industrials
Schneider Electric	Industrials
LINDE	Industrials
Vinci	Industrials
Glencore	Industrials
ThyssenKrupp	Basic Materials
BASF	Basic Materials
Anglo Pacific Group	Basic Materials
EOAN	Utilities
National Grid	Utilities
Enel SPA	Utilities
Engie SA	Utilities
Energie Baden-Wuettenberg	Utilities
DTE Energy	Utilities
Orange	Telecommunication Services
Sanofi	Healthcare
Bayer	Healthcare
Bosch	Consumer Non-cyclicals
Continental	Consumer Non-cyclicals
Man SE	Consumer Non-cyclicals
Peugeot	Consumer Non-cyclicals
Volkswagen	Consumer Non-cyclicals
Daimler	Consumer Non-cyclicals
Sie de Saint-Gobain	Consumer Non-cyclicals
BMW	Consumer Non-cyclicals
Royal Dutch Shell	Energy
British Petroleum	Energy
Unilever	Consumer Non-cyclicals
Sabmiller	Consumer Non-cyclicals
L'oreal	Consumer Non-cyclicals
Moet Hennessy Louis Vuitton SE	Consumer Non-cyclicals
Diageo PLC	Consumer Non-cyclicals

Table 2.A.3 The European companies used in the current studies.

Appendix 2.C Mean strength of the lags of the news sentiments for US and European companies

	FIN	TEC	IND	MAT	UTL	COM	HLC	NCY	ENE	YCY
asset_{t-1}	0.24	0.27	0.25	0.24	0.23	0.24	0.24	0.26	0.23	0.27
asset_{t-5}	0.22	0.23	0.23	0.24	0.21	0.23	0.22	0.22	0.22	0.23
asset_{t-22}	0.19	0.18	0.20	0.18	0.18	0.18	0.19	0.18	0.20	0.23
MPTRXUSFIN_{t-1}	1.33	1.08	1.28	1.53	0.98	1.02	0.83	1.14	1.52	1.35
MPTRXUSFIN_{t-5}	0.84	0.87	1.11	1.32	0.79	1.15	1.21	0.95	1.15	0.88
MPTRXUSFIN_{t-22}	0.90	0.94	0.94	1.28	0.95	1.01	1.07	0.78	1.10	0.87
MPTRXUSTEC_{t-1}	0.29	0.36	0.39	0.39	0.44	0.39	0.37	0.39	0.41	0.45
MPTRXUSTEC_{t-5}	0.41	0.33	0.38	0.49	0.28	0.40	0.26	0.29	0.54	0.46
MPTRXUSTEC_{t-22}	0.30	0.33	0.34	0.39	0.39	0.41	0.41	0.28	0.45	0.38
MPTRXUSIND_{t-1}	0.43	0.54	0.51	0.51	0.49	0.62	0.45	0.54	0.59	0.52
MPTRXUSIND_{t-5}	0.38	0.35	0.53	0.50	0.42	0.50	0.44	0.33	0.51	0.54
MPTRXUSIND_{t-22}	0.57	0.47	0.46	0.64	0.46	0.61	0.43	0.51	0.57	0.52
MPTRXUSMAT_{t-1}	0.35	0.38	0.41	0.55	0.33	0.42	0.37	0.38	0.38	0.35
MPTRXUSMAT_{t-5}	0.37	0.32	0.36	0.45	0.34	0.35	0.32	0.36	0.38	0.30
MPTRXUSMAT_{t-22}	0.32	0.35	0.42	0.45	0.38	0.45	0.41	0.28	0.43	0.29
MPTRXUSUTL_{t-1}	0.45	0.41	0.48	0.60	0.41	0.53	0.46	0.49	0.79	0.56
MPTRXUSUTL_{t-5}	0.53	0.30	0.36	0.56	0.26	0.47	0.28	0.33	0.55	0.47
MPTRXUSUTL_{t-22}	0.48	0.31	0.41	0.47	0.42	0.46	0.39	0.36	0.50	0.36
MPTRXUSCOM_{t-1}	0.45	0.46	0.47	0.41	0.37	0.36	0.41	0.48	0.59	0.42
MPTRXUSCOM_{t-5}	0.32	0.33	0.40	0.37	0.33	0.56	0.31	0.29	0.51	0.39
MPTRXUSCOM_{t-22}	0.33	0.43	0.46	0.63	0.34	0.41	0.32	0.29	0.48	0.44
MPTRXUSHLC_{t-1}	0.41	0.40	0.52	0.47	0.38	0.44	0.38	0.39	0.52	0.32
MPTRXUSHLC_{t-5}	0.41	0.42	0.41	0.49	0.40	0.38	0.39	0.32	0.49	0.43
MPTRXUSHLC_{t-22}	0.40	0.38	0.38	0.49	0.34	0.40	0.48	0.34	0.44	0.33
MPTRXUSNCY_{t-1}	0.29	0.32	0.32	0.36	0.30	0.41	0.30	0.42	0.48	0.39
MPTRXUSNCY_{t-5}	0.21	0.26	0.43	0.37	0.31	0.43	0.34	0.27	0.42	0.28
MPTRXUSNCY_{t-22}	0.30	0.34	0.35	0.35	0.28	0.39	0.26	0.29	0.37	0.36
MPTRXUSENE_{t-1}	0.82	0.94	1.03	2.04	0.92	1.26	1.67	0.83	2.06	0.83
MPTRXUSENE_{t-5}	1.01	0.77	1.12	0.91	0.89	0.93	1.07	0.74	1.33	1.13
MPTRXUSENE_{t-22}	0.78	0.70	0.88	0.88	0.87	1.04	0.87	0.92	0.93	1.21
MPTRXUSYCY_{t-1}	0.59	0.42	0.38	0.77	0.39	0.64	0.34	0.48	0.42	0.64
MPTRXUSYCY_{t-5}	0.36	0.35	0.45	0.37	0.32	0.51	0.33	0.38	0.49	0.44
MPTRXUSYCY_{t-22}	0.39	0.30	0.44	0.54	0.54	0.52	0.52	0.40	0.53	0.39
US_{t-1}	0.71	0.52	0.66	0.80	0.50	0.69	0.48	0.50	0.68	0.53
US_{t-5}	0.57	0.32	0.60	0.50	0.34	0.63	0.44	0.44	0.51	0.54
US_{t-22}	0.41	0.46	0.59	0.68	0.45	0.66	0.67	0.47	0.56	0.41
IT_{t-1}	0.35	0.45	0.38	0.48	0.31	0.38	0.39	0.34	0.41	0.41
IT_{t-5}	0.34	0.34	0.29	0.56	0.32	0.36	0.37	0.31	0.50	0.64
IT_{t-22}	0.34	0.38	0.38	0.58	0.35	0.51	0.63	0.34	0.54	0.43
DE_{t-1}	0.45	0.48	0.49	0.59	0.54	0.62	0.41	0.40	0.48	0.37
DE_{t-5}	0.37	0.38	0.41	0.37	0.47	0.49	0.39	0.38	0.58	0.54
DE_{t-22}	0.47	0.39	0.47	0.43	0.48	0.56	0.43	0.35	0.60	0.44
CN_{t-1}	0.36	0.44	0.38	0.41	0.35	0.42	0.38	0.37	0.41	0.43
CN_{t-5}	0.29	0.26	0.48	0.36	0.37	0.31	0.33	0.23	0.35	0.32
CN_{t-22}	0.34	0.31	0.37	0.55	0.26	0.35	0.34	0.35	0.51	0.34
GR_{t-1}	0.55	0.47	0.57	0.69	0.60	0.62	0.51	0.58	0.78	0.66
GR_{t-5}	0.56	0.41	0.47	0.47	0.54	0.55	0.53	0.53	0.64	0.49
GR_{t-22}	0.51	0.54	0.56	0.51	0.37	0.58	0.48	0.48	0.51	0.41

Table 2.A.4 Mean strength of the lags of the news sentiments on the US sectors and selected countries for the US companies by sector ranging from Jan. 1, 2005 to Dec. 31, 2014.

	FIN	TEC	IND	MAT	UTL	COM	HLC	NCY	ENE	YCY
asset_{t-1}	0.24	0.24	0.23	0.25	0.24	0.27	0.22	0.23	0.25	0.23
asset_{t-5}	0.23	0.23	0.23	0.22	0.22	0.23	0.23	0.23	0.22	0.22
asset_{t-22}	0.19	0.17	0.18	0.18	0.19	0.19	0.17	0.17	0.20	0.17
MPTRXFIN_{t-1}	0.77	0.65	0.98	0.96	0.85	0.61	0.77	0.75	0.84	0.68
MPTRXFIN_{t-5}	0.84	0.58	1.00	1.22	0.76	0.56	0.80	0.75	0.79	0.46
MPTRXFIN_{t-22}	0.61	0.52	0.88	0.83	0.59	0.63	0.65	0.67	0.67	0.39
MPTRXTEC_{t-1}	0.35	0.39	0.36	0.52	0.38	0.35	0.36	0.36	0.53	0.35
MPTRXTEC_{t-5}	0.30	0.36	0.28	0.46	0.43	0.30	0.31	0.34	0.46	0.18
MPTRXTEC_{t-22}	0.46	0.35	0.50	0.51	0.44	0.43	0.45	0.43	0.49	0.22
MPTRXIND_{t-1}	0.47	0.38	0.47	0.57	0.60	0.50	0.55	0.41	0.58	0.26
MPTRXIND_{t-5}	0.36	0.32	0.47	0.44	0.58	0.35	0.48	0.39	0.44	0.22
MPTRXIND_{t-22}	0.45	0.46	0.56	0.64	0.53	0.42	0.40	0.53	0.51	0.31
MPTRXMAT_{t-1}	0.32	0.28	0.29	0.46	0.35	0.25	0.36	0.28	0.42	0.22
MPTRXMAT_{t-5}	0.32	0.28	0.41	0.39	0.32	0.26	0.28	0.25	0.37	0.24
MPTRXMAT_{t-22}	0.34	0.32	0.32	0.37	0.36	0.33	0.33	0.31	0.32	0.19
MPTRXUTL_{t-1}	0.39	0.30	0.66	0.50	0.46	0.40	0.52	0.37	0.53	0.55
MPTRXUTL_{t-5}	0.45	0.37	0.40	0.37	0.37	0.29	0.39	0.30	0.50	0.22
MPTRXUTL_{t-22}	0.38	0.36	0.40	0.44	0.41	0.36	0.42	0.35	0.43	0.23
MPTRXCOM_{t-1}	0.44	0.40	0.46	0.49	0.42	0.38	0.41	0.34	0.48	0.30
MPTRXCOM_{t-5}	0.38	0.31	0.36	0.36	0.47	0.22	0.36	0.31	0.29	0.25
MPTRXCOM_{t-22}	0.33	0.37	0.39	0.38	0.43	0.37	0.43	0.33	0.39	0.21
MPTRXHLC_{t-1}	0.46	0.45	0.67	0.65	0.36	0.37	0.41	0.41	0.53	0.34
MPTRXHLC_{t-5}	0.40	0.39	0.49	0.64	0.41	0.37	0.40	0.41	0.43	0.26
MPTRXHLC_{t-22}	0.40	0.36	0.58	0.55	0.48	0.41	0.49	0.43	0.56	0.28
MPTRXNCY_{t-1}	0.47	0.33	0.68	0.46	0.42	0.33	0.52	0.35	0.46	0.25
MPTRXNCY_{t-5}	0.38	0.34	0.45	0.40	0.46	0.33	0.43	0.40	0.43	0.20
MPTRXNCY_{t-22}	0.34	0.34	0.49	0.47	0.35	0.37	0.47	0.36	0.40	0.20
MPTRXENE_{t-1}	0.87	0.77	1.12	0.84	0.83	0.65	0.77	0.71	0.87	0.62
MPTRXENE_{t-5}	0.57	0.63	0.86	0.80	0.69	0.53	0.72	0.71	1.25	0.38
MPTRXENE_{t-22}	0.68	0.65	0.88	0.87	0.83	0.57	0.87	0.54	0.89	0.49
MPTRXYCY_{t-1}	0.42	0.40	0.46	0.48	0.42	0.39	0.38	0.39	0.44	0.45
MPTRXYCY_{t-5}	0.36	0.33	0.36	0.36	0.41	0.30	0.46	0.30	0.50	0.21
MPTRXYCY_{t-22}	0.30	0.50	0.48	0.59	0.30	0.41	0.38	0.32	0.40	0.21
US_{t-1}	0.71	0.62	0.37	0.55	0.59	0.40	0.46	0.45	0.47	0.50
US_{t-5}	0.57	0.39	0.65	0.54	0.50	0.45	0.55	0.42	0.56	0.25
US_{t-22}	0.58	0.43	0.42	0.47	0.45	0.42	0.54	0.39	0.49	0.33
IT_{t-1}	0.44	0.33	0.41	0.55	0.46	0.52	0.46	0.40	0.46	0.26
IT_{t-5}	0.39	0.29	0.32	0.36	0.41	0.32	0.37	0.37	0.38	0.20
IT_{t-22}	0.38	0.37	0.68	0.49	0.41	0.34	0.38	0.39	0.44	0.20
DE_{t-1}	0.51	0.52	0.72	0.60	0.48	0.47	0.45	0.52	0.43	0.32
DE_{t-5}	0.44	0.37	0.50	0.52	0.46	0.43	0.54	0.43	0.53	0.27
DE_{t-22}	0.45	0.42	0.56	0.56	0.41	0.45	0.44	0.38	0.50	0.21
CN_{t-1}	0.38	0.38	0.40	0.45	0.32	0.36	0.46	0.32	0.38	0.43
CN_{t-5}	0.36	0.29	0.47	0.54	0.40	0.32	0.42	0.32	0.40	0.16
CN_{t-22}	0.47	0.38	0.47	0.44	0.38	0.37	0.40	0.33	0.34	0.26
GR_{t-1}	0.61	0.52	1.13	0.72	0.63	0.62	0.57	0.51	0.65	0.37
GR_{t-5}	0.38	0.41	0.52	0.58	0.54	0.42	0.53	0.40	0.67	0.38
GR_{t-22}	0.54	0.46	0.53	0.59	0.60	0.48	0.55	0.50	0.55	0.31

Table 2.A.5 Mean strength of the lags of the news sentiments on the European sectors and selected countries for the European companies by sector ranging from Jan. 1, 2005 to Dec. 31, 2014.

Appendix 2.D Overall relevance and strength of selected sectors and countries for US and European companies

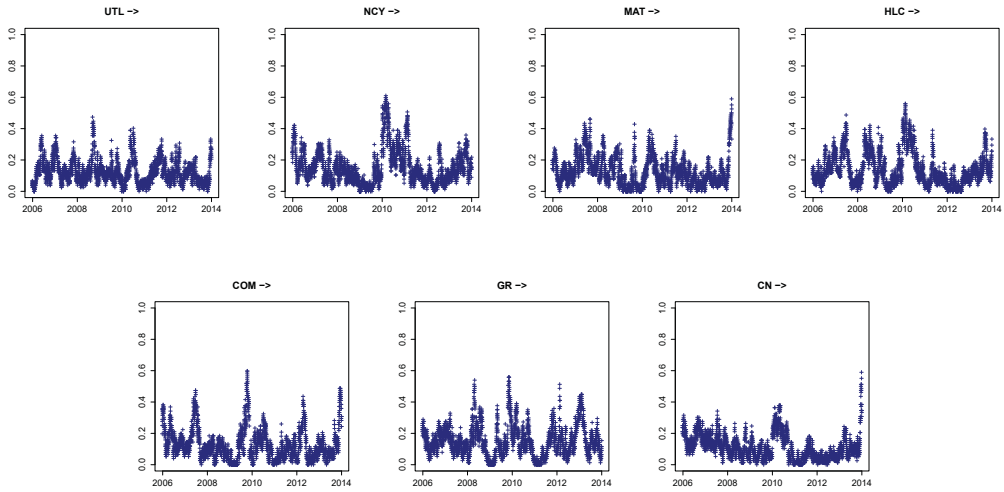


Figure 2.A.1 The overall relevance of the news sentiments of the selected sectors and countries on the US stocks ranging from Jan. 1, 2005 to Dec. 31, 2014.

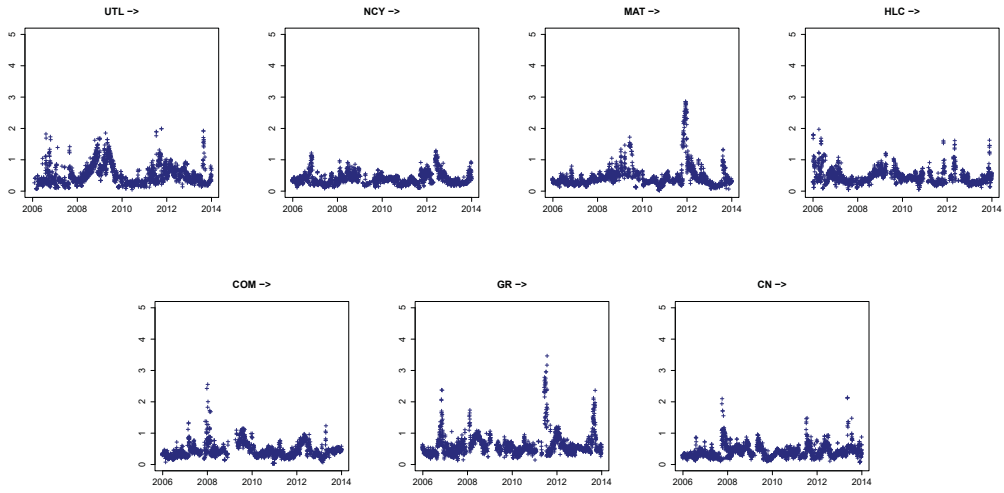


Figure 2.A.2 The overall strength of the news sentiments of the selected sectors and countries on the US stocks ranging from Jan. 1, 2005 to Dec. 31, 2014.

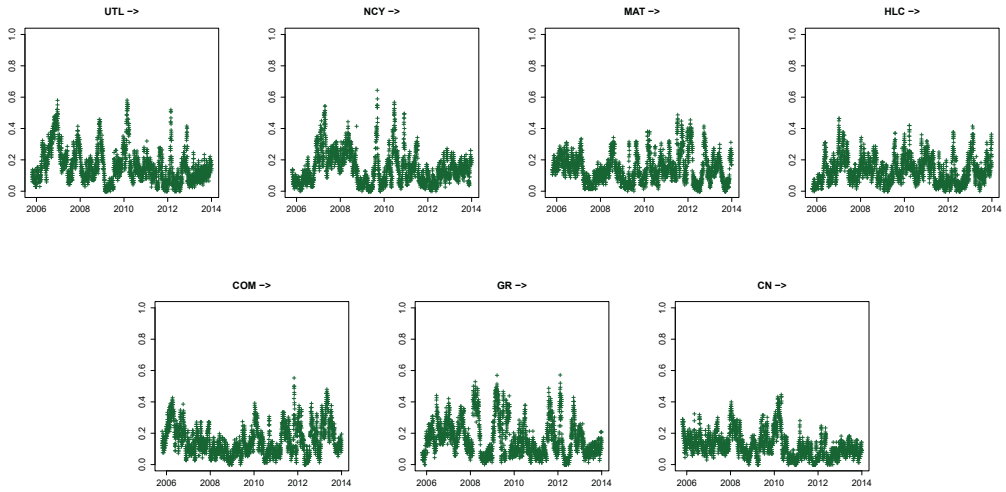


Figure 2.A.3 The overall relevance of the news sentiments of the selected sectors and countries on the European stocks ranging from Jan. 1, 2005 to Dec. 31, 2014.

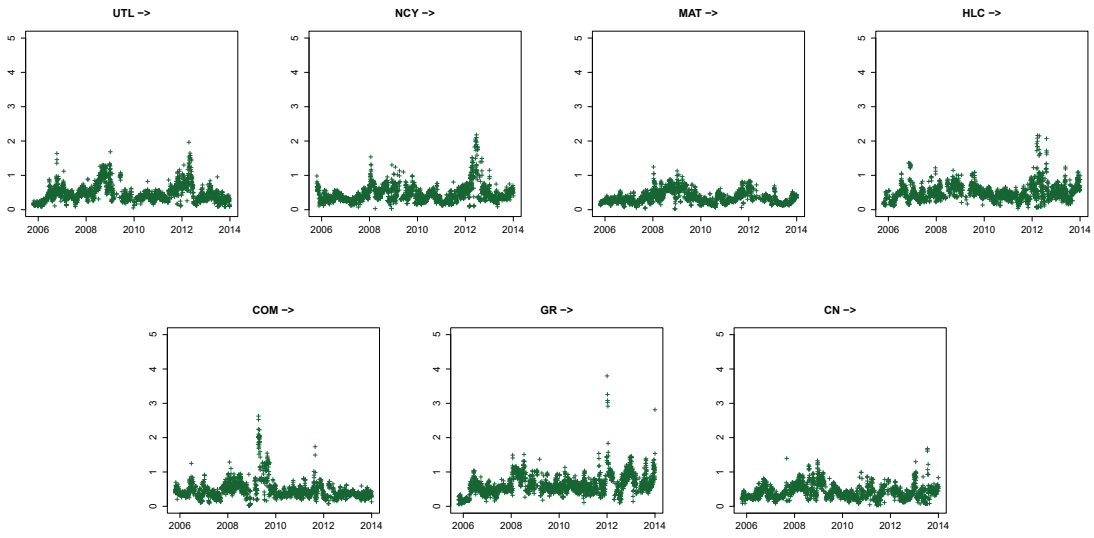


Figure 2.A.4 The overall strength of the news sentiments of the selected sectors and countries on the European stocks ranging from Jan. 1, 2005 to Dec. 31, 2014.

Appendix 2.E Overall relevance and strength of sectors and countries for US companies controlling for the VIX index

	FIN	TEC	IND	MAT	UTL	COM	HLC	NCY	ENE	YCY
asset	0.56	0.68	0.60	0.55	0.50	0.61	0.53	0.57	0.52	0.62
MPTRXUSFIN	0.15	0.18	0.16	0.20	0.17	0.14	0.16	0.18	0.23	0.14
MPTRXUSTEC	0.13	0.15	0.17	0.13	0.23	0.18	0.19	0.16	0.30	0.18
MPTRXUSIND	0.12	0.16	0.16	0.13	0.12	0.17	0.11	0.15	0.19	0.13
MPTRXUSMAT	0.10	0.07	0.16	0.15	0.12	0.10	0.11	0.14	0.21	0.11
MPTRXUSUTL	0.07	0.09	0.14	0.14	0.12	0.17	0.12	0.12	0.24	0.11
MPTRXUSCOM	0.13	0.11	0.10	0.11	0.09	0.13	0.11	0.08	0.18	0.09
MPTRXUSHLC	0.13	0.15	0.19	0.11	0.10	0.13	0.11	0.15	0.14	0.13
MPTRXUSNCY	0.11	0.12	0.18	0.13	0.11	0.15	0.13	0.11	0.15	0.12
MPTRXUSENE	0.08	0.10	0.12	0.08	0.08	0.13	0.13	0.14	0.12	0.18
MPTRXUSYCY	0.08	0.09	0.12	0.10	0.11	0.10	0.08	0.11	0.19	0.11
US	0.13	0.15	0.14	0.10	0.08	0.14	0.15	0.16	0.16	0.12
IT	0.10	0.16	0.10	0.09	0.08	0.11	0.10	0.15	0.12	0.11
GR	0.14	0.13	0.17	0.15	0.14	0.14	0.09	0.17	0.16	0.13
DE	0.14	0.12	0.13	0.10	0.12	0.15	0.11	0.13	0.21	0.13
CN	0.10	0.12	0.14	0.12	0.11	0.10	0.09	0.12	0.15	0.13
VIX	0.89	0.96	0.96	0.84	0.83	0.89	0.88	0.94	0.84	0.87

Table 2.A.6 Mean relevance of the news sentiments of the US sectors and selected countries for the US companies by sector ranging from Jan. 1, 2005 to Dec. 31, 2014 (controlling for the VIX index).

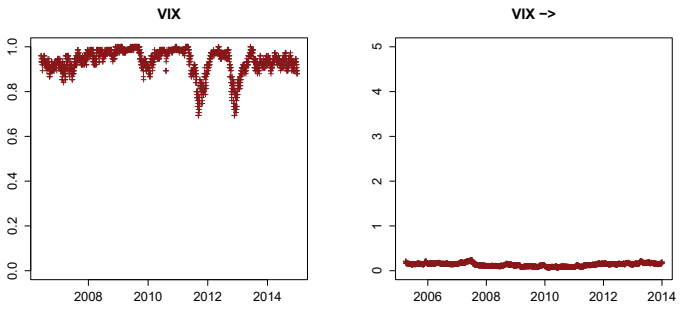


Figure 2.A.5 The overall relevance (left) and strength (right) of the VIX index on the US stocks ranging from Jan. 1, 2005 to Dec. 31, 2014.

	FIN	TEC	IND	MAT	UTL	COM	HLC	NCY	ENE	YCY
asset_{t-1}	0.18	0.21	0.15	0.17	0.18	0.20	0.21	0.20	0.16	0.23
asset_{t-5}	0.19	0.22	0.22	0.22	0.19	0.22	0.22	0.21	0.22	0.23
asset_{t-22}	0.19	0.18	0.19	0.17	0.17	0.18	0.18	0.18	0.20	0.22
MPTRXUSFIN_{t-1}	1.11	1.05	1.26	1.46	1.00	0.64	0.85	0.93	1.24	1.38
MPTRXUSFIN_{t-5}	0.68	0.95	0.77	1.20	0.76	1.19	1.19	0.70	0.99	0.56
MPTRXUSFIN_{t-22}	0.58	0.86	0.88	1.16	0.92	1.11	1.17	0.72	0.92	0.81
MPTRXUSTEC_{t-1}	0.27	0.36	0.35	0.34	0.46	0.38	0.37	0.38	0.44	0.44
MPTRXUSTEC_{t-5}	0.40	0.46	0.35	0.45	0.30	0.44	0.25	0.29	0.52	0.43
MPTRXUSTEC_{t-22}	0.30	0.35	0.28	0.35	0.41	0.48	0.41	0.28	0.44	0.39
MPTRXUSIND_{t-1}	0.39	0.54	0.54	0.50	0.43	0.74	0.39	0.44	0.56	0.54
MPTRXUSIND_{t-5}	0.41	0.20	0.52	0.47	0.39	0.55	0.39	0.25	0.38	0.51
MPTRXUSIND_{t-22}	0.45	0.47	0.40	0.61	0.50	0.54	0.51	0.50	0.54	0.56
MPTRXUSMAT_{t-1}	0.27	0.36	0.38	0.53	0.32	0.37	0.44	0.31	0.35	0.38
MPTRXUSMAT_{t-5}	0.35	0.32	0.39	0.42	0.35	0.39	0.28	0.38	0.22	0.34
MPTRXUSMAT_{t-22}	0.33	0.30	0.41	0.41	0.39	0.56	0.39	0.25	0.39	0.28
MPTRXUSUTL_{t-1}	0.47	0.41	0.45	0.61	0.45	0.57	0.45	0.47	0.78	0.57
MPTRXUSUTL_{t-5}	0.55	0.27	0.17	0.49	0.28	0.53	0.25	0.27	0.52	0.45
MPTRXUSUTL_{t-22}	0.39	0.28	0.38	0.42	0.44	0.29	0.40	0.34	0.39	0.35
MPTRXUSCOM_{t-1}	0.38	0.43	0.44	0.34	0.36	0.13	0.41	0.41	0.57	0.37
MPTRXUSCOM_{t-5}	0.32	0.50	0.40	0.28	0.34	0.52	0.32	0.31	0.43	0.38
MPTRXUSCOM_{t-22}	0.31	0.43	0.39	0.59	0.35	0.41	0.34	0.29	0.46	0.47
MPTRXUSHLC_{t-1}	0.40	0.31	0.53	0.42	0.39	0.44	0.39	0.28	0.53	0.34
MPTRXUSHLC_{t-5}	0.42	0.44	0.30	0.47	0.44	0.45	0.39	0.29	0.50	0.38
MPTRXUSHLC_{t-22}	0.37	0.34	0.36	0.45	0.34	0.29	0.48	0.34	0.34	0.33
MPTRXUSNCY_{t-1}	0.27	0.32	0.27	0.30	0.28	0.42	0.29	0.32	0.49	0.42
MPTRXUSNCY_{t-5}	0.11	0.34	0.39	0.33	0.29	0.44	0.37	0.21	0.42	0.19
MPTRXUSNCY_{t-22}	0.29	0.32	0.35	0.28	0.30	0.36	0.24	0.26	0.29	0.35
MPTRXUSENE_{t-1}	0.78	0.87	1.02	1.98	1.01	1.18	1.71	0.81	2.00	0.87
MPTRXUSENE_{t-5}	0.97	0.69	0.94	0.84	0.84	0.94	1.10	0.60	1.23	1.05
MPTRXUSENE_{t-22}	0.63	0.63	0.72	0.62	0.82	0.91	0.89	0.82	0.72	1.19
MPTRXUSYCY_{t-1}	0.62	0.39	0.41	0.73	0.33	0.67	0.33	0.47	0.37	0.59
MPTRXUSYCY_{t-5}	0.30	0.32	0.43	0.33	0.25	0.47	0.35	0.28	0.42	0.47
MPTRXUSYCY_{t-22}	0.35	0.23	0.42	0.49	0.78	0.22	0.49	0.27	0.39	0.42
US_{t-1}	0.71	0.50	0.66	0.75	0.53	0.73	0.48	0.52	0.66	0.50
US_{t-5}	0.51	0.36	0.59	0.36	0.01	0.55	0.53	0.45	0.51	0.51
US_{t-22}	0.39	0.43	0.56	0.62	0.45	0.56	0.65	0.43	0.50	0.38
IT_{t-1}	0.29	0.42	0.34	0.45	0.27	0.35	0.42	0.06	0.37	0.41
IT_{t-5}	0.32	0.34	0.21	0.51	0.31	0.38	0.32	0.24	0.51	0.61
IT_{t-22}	0.30	0.37	0.33	0.56	0.32	0.49	0.64	0.33	0.48	0.39
GR_{t-1}	0.48	0.45	0.48	0.69	0.57	0.55	0.49	0.58	0.78	0.65
GR_{t-5}	0.44	0.59	0.43	0.46	0.53	0.49	0.57	0.36	0.60	0.53
GR_{t-22}	0.42	0.50	0.58	0.48	0.34	0.54	0.49	0.45	0.36	0.43
DE_{t-1}	0.28	0.41	0.39	0.51	0.52	0.44	0.40	0.42	0.37	0.42
DE_{t-5}	0.35	0.34	0.37	0.32	0.44	0.46	0.42	0.37	0.52	0.51
DE_{t-22}	0.49	0.43	0.45	0.43	0.46	0.70	0.38	0.35	0.55	0.44
CN_{t-1}	0.34	0.46	0.39	0.36	0.38	0.40	0.36	0.33	0.36	0.44
CN_{t-5}	0.27	0.31	0.40	0.31	0.26	0.42	0.34	0.18	0.45	0.30
CN_{t-22}	0.32	0.30	0.36	0.53	0.22	0.38	0.32	0.30	0.47	0.36
VIX_{t-1}	0.14	0.14	0.14	0.14	0.13	0.15	0.11	0.12	0.17	0.10
VIX_{t-5}	0.06	0.07	0.10	0.08	0.07	0.08	0.06	0.08	0.09	0.06
VIX_{t-22}	0.13	0.09	0.08	0.07	0.08	0.14	0.09	0.07	0.09	0.06

Table 2.A.7 Mean strength of the lags of the news sentiments on the US sectors and selected countries for the US companies by sector ranging from Jan. 1, 2005 to Dec. 31, 2014 (controlling for the VIX index).

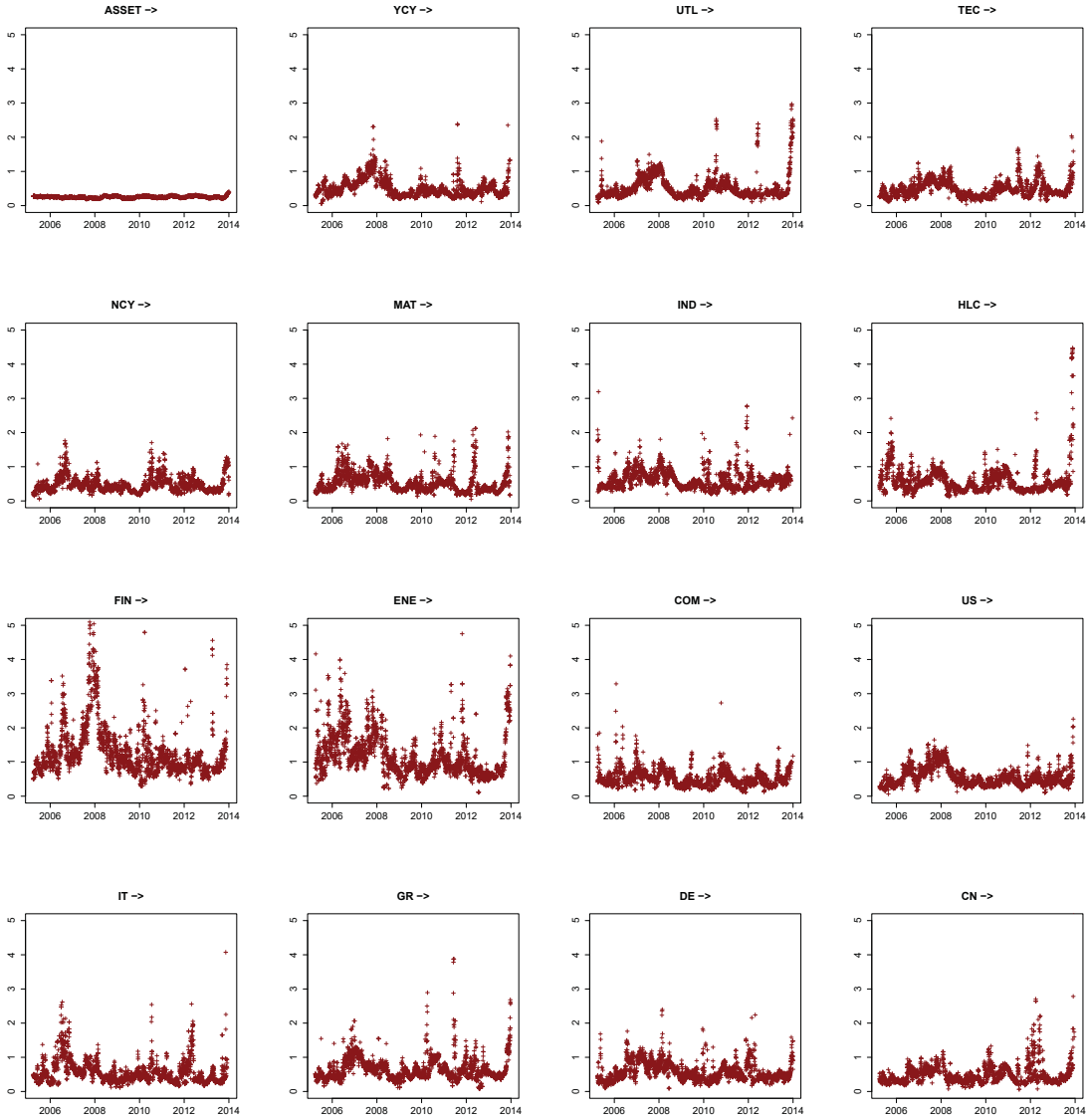


Figure 2.A.6 The overall strength of the news sentiments of sectors and countries on the US stocks ranging from Jan. 1, 2005 to Dec. 31, 2014 (controlling for the VIX index).

Chapter 3

Do financial companies communicate to one another in the news? (Application of multivariate Hawkes graphs to uncover Granger causality of financial news)

Anastasija Tetereva [1]

[1]Chair of Mathematics and Statistics, University of St Gallen, Bodanstrasse 6, 9000 St Gallen, Switzerland, anastasija.tetereva@unisg.ch

Abstract

A considerable amount of current research in finance addresses the influence of news and social media on stock returns and volatility. Although news data are used in many applications, the mutual relationship among public announcements remains unclear. Moreover, the majority of studies are conducted using aggregated data, which are less effective in detecting causal links than observations of higher frequency. This paper provides evidence of self and mutual triggering of news announcements in the financial sector. It is proposed that the news arrival times be modelled as a multivariate Hawkes process to test the Granger causality of company-specific news and to detect the most influential companies. Based on this information, a novel method of constructing a composite news intensity index (NII) is presented. The NII demonstrates the ability to timeously describe the uncertainty in financial markets. The proposed measure Granger causes VIX at 6-month lag and can therefore be used to diagnose the health of a financial system.

3.1 Introduction

The increasing availability of news and social media data has attracted attention regarding the sentiment models in the field of financial econometrics during the last decade. Special attention has been drawn to the possibility of improving the prediction power of the models by augmenting them with information contained in public announcements. Several studies suggested that news data contain additional information that might be useful to explain investors' behaviour. Prior work in this field focused primarily on the influence of investor sentiment on stock returns, for example, Neal and Wheatley (1998), Lee et al. (1991), Brown and Cliff (2004). More recent studies by Baker and Wurgler (2006) and Baker and Wurgler (2007) attempted to quantify investor sentiment and measure the impact of news announcements on stock returns. Some attempts were made with the purpose of explaining the influence of public announcements on the volatility of stock returns, see Wang et al. (2006), Ho et al. (2013) and Lee et al. (2002) for more details. Much of the current debate revolves around the construction of news-based indices that would be able to predict the future movements of the market and provide early signals of economic instability. The first investigations in this direction were made by Shefrin (2007) and Borovkova et al. (2017).

Traditionally, the focus has always been on the direct effects of news announcements about an asset on its returns and volatility. Audrino and Tetereva (2017) were the first to investigate sentiment spillover effects. The authors demonstrated that returns in all sectors are driven by sentiment from a few industries. In particular, the importance of financial news during periods of financial instability was demonstrated. The question that remains unanswered in the study by Audrino and Tetereva (2017) is whether financial news is correlated with time and among companies. The debate on the contribution of an individual company to the risk of the whole system is not new in the field. Many attempts were made to measure the contribution of financial companies to systemic risk in terms of returns, for example Brechmann et al. (2013), Hautsch et al. (2014) and Härdle et al. (2016). One issue that needs to be raised is whether financial institutions are connected in terms of public announcements, and in particular, whether there is information diffusion in the media. Cerchiello et al. (2017) made the first attempt to measure the contagion among financial news. In other words, their study attempted to address the question of the mutual causality of financial news. While the authors found significant evidence of news contagion among countries, the main limitation of their study was the low level of granularity due to insufficient data coverage. Their analysis was performed on a monthly level, which could have led to a deterioration in causal effects. Therefore, there is still considerable

uncertainty with regard to news contagion among financial institutions on a daily or even a high-frequency level.

In recent decades, research in the field of financial econometrics has provided ample support for the assertion that one can benefit from analysing financial data at a transaction level; more details can be found in preliminary works that Hasbrouck (1991) and Engle and Russell (1998) carried out in the 1990s. With regard to the stock prices, the irregular occurrences of transactions can be seen as a point process. Moreover, it was empirically observed that transaction events are clustered over time, and durations are positively autocorrelated. Engle and Russell (1998) made the first attempts to model durations by means of a conditional autoregressive model; more details on duration models can be found in Bauwens and Giot (2001) and Bauwens and Hautsch (2009). Although the autoregressive conditional duration models by Engle and Russell (1998) and their modifications discussed in Zhang et al. (2001) and Fernandes and Grammig (2006) gained prominent interest in the field, a continuous time intensity based setting was shown to provide a more flexible and powerful tool for modelling multivariate point processes. For this reason, the current study takes advantage of multivariate Hawkes processes.

This paper contributes to the considerable amount of news sentiment-related literature by studying the mutual excitation of the news in the financial sector at a high-frequency level. It is proposed to uncover Granger causality in financial news data by means of multivariate Hawkes graphs. This approach allows one to better understand the properties of high-frequency news data and to measure the influence of individual companies on the whole system in terms of news announcements. Based on the information extracted from multivariate Hawkes graphs, the construction of a composite NII (news intensity index) is proposed that can potentially be used to describe the health of a financial system. In contrast to prior studies, this approach allows one to construct a valid index based purely on news intensity information, and it avoids the need to compute sentiment scores. This makes the index robust to measures such as relevance and the novelty of a topic. Moreover, there is no need to use financial dictionaries to translate the text into numerical measures. The performance of the index is tested by means of Granger causality, and its influence on some real financial markets' measures is analysed by means of an impulse response function. The ability of the index to predict uncertainty among investors at a 6-month lag is demonstrated.

The paper is organized as follows: Section 3.2 provides a brief overview of mutually exciting point processes and introduces Granger causality for multivariate Hawkes graphs. Section 3.3 describes the data and examines the properties of high-frequency observations

that are specific for RavenPack sentiment data. The results are presented in Section 3.4. Section 3.5 describes the way in which the results of Section 3.4 can be used to construct the NII, and some conclusions are drawn in the final section.

3.2 Hawkes process

To measure the financial news contagion, the current study considers publishing times as mutually exciting point processes. Before introducing the model, the essential definitions of the considered framework are recalled. Further on, a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is assumed, and a sequence of non-negative random variables - $(t_i)_{i\in\ ^\star}$ such that $\forall i \in\ ^\star$, $t_i < t_{i+1}$ - is considered on this probability space. The process $(t_i)_{i\in\ ^\star}$ is called a point process on $_+$. In this paper, t_i represents the times of occurrence of news announcements. The right continuous process is called the counting process associated with the point process $(t_i)_{i\in\mathbb{N}^\star}$:

$$N(t) = \sum_{i\in\mathbb{N}^\star} \mathbb{1}_{t_i<t}.$$

The left continuous intensity process is defined as follows:

$$\lambda\left(t\middle|\mathcal{F}_t^N\right) = \lim_{h\downarrow 0} \frac{1}{n} \mathrm{P}\left(N(t+h) - N(t) > 0\middle|\mathcal{F}_t^N\right).$$

In the simplest case of a point process, the probability of occurrences of an event in $(t, t+h]$ does not depend on the history of the process; this type of point processes is the well-known Poisson process. A generalization of the Poisson process is a linear self-exciting process:

$$\lambda(t) = \lambda_0(t) + \int_{-\infty}^{t} h(t-s)\,\mathrm{d}N_s = \lambda_0(t) - \sum_{t_i<t} h(t-t_i), \tag{3.1}$$

where λ_0 is a baseline intensity and $h(\cdot)$ is an excitement function that represents the influence of the historical events on the current intensity process. Such point process is called a Hawkes or self-exciting point process. (3.1) can be further extended to a case when, apart from self-excitement, events of different types trigger each other. Such a generalization is referred to as a multivariate Hawkes process. The conditional intensity of a d-variate Hawkes process is defined as

$$\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}_0 + \int_{-\infty}^{t} H(t-s)\boldsymbol{N}\mathrm{d}s, \tag{3.2}$$

where $\mathbf{N} = \left(N^1, N^2, \ldots, N^d\right)^\top$ is a d-dimensional point process, $\boldsymbol{\lambda}_0$ is the vector of baseline intensities, and $H = (h_{i,j})_{1 \le i,j \le d}$ is a measurable $d \times d$ matrix-valued excitement function. The component $h_{i,j}(t)$ of the matrix $H(t)$ is called a kernel; it represents the effect of events in component j on the intensity of the component i. The parametric kernels, and exponential kernels in particular, have attracted considerable interest in the literature due to the advantages of numerical computations and intuitive interpretation. Moreover, this kernel function makes the dynamics of the Hawkes process Markovian; more details on Hawkes processes with exponential kernels can be found in the work of Farajtabar et al. (2014), Rasmussen and Williams (2006), Zhou et al. (2013b), Hall and Willett (2014), and Yan et al. (2015). The power-law kernel was applied in Zhao et al. (2015). The nonparametric estimation of the triggering kernels is desirable when the form of $h_{i,j}(t)$ is not known a priori; this type of estimation was considered in Kirchner (2016), Xu et al. (2016) and Eichler et al. (2017).

The triggering kernels play a major role in uncovering Granger causality for Hawkes processes. The first definition of Granger non-causality for a Hawkes process appeared in Eichler et al. (2017). The authors suggested that a type-i event does not Granger cause a type-j event if $h_{i,j}(t) = 0$ for $t \in [0, \infty)$, $i, j = 1, \ldots, d$. This definition is valid under the assumption that $\mathrm{d}\mathbf{N}(t-s) > 0$ for $0 \le s \le t$. The mutual Granger causality of a multivariate Hawkes process can be visualized by making use of a causality graph with the set of vertices $[d] = \{1, 2, \ldots, d\}$ representing event types and the set of directed edges $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ demonstrating the causality. The strength of Granger causality is usually associated with the branching matrix A, where

$$A_{i,j} = \int \left(h_{i,j}(t)\mathrm{d}t\right)_{1 \le i,j \le d}. \tag{3.3}$$

According to Embrechts and Kirchner (2016a), the effect of a branching matrix manipulation of the form (3.3) can be summarized in the set of cascade and feedback coefficients, which are defined as follows:

$$c_{i_0} = \frac{\lambda_{i_0} \sum_{j=1}^{d} e_{i_0,j}}{\sum_{i=1}^{d} \lambda_i \sum_{j=1}^{d} e_{i_0,j}}, \; i_0 \in [d] \tag{3.4}$$

and

$$f_j = \frac{\lambda_j e_{j,j}}{\sum_{i=1}^{d} \lambda_i e_{i,j}}, \; j \in [d] \tag{3.5}$$

with $e_{i,j} = (\boldsymbol{I} - A)^{-1}$, where A is the branching matrix from (3.3) and $\boldsymbol{I}$ is a $d \times d$ matrix with ones on the main diagonal and zeros elsewhere. The cascade coefficients (3.4) measure the fraction of events in the system stemming from type-i events, and they are important from a systemic point of view. Embrechts and Kirchner (2016a) pointed out that event types with $c_i > \frac{1}{d}$ have large impacts on the whole system. The feedback coefficients (3.5) indicate how much of the total intensity experienced by a type-j event is due to its own past activity.

There are two possible ways in which to estimate the branching matrix (3.3) and consequently the cascade and feedback coefficients. The most well-known approach is to estimate the baseline intensities and the excitement function based on (3.2). More recent studies, for example, Achab et al. (2016) and Clements et al. (2017), proposed avoiding the estimation of the whole multidimensional Hawkes process and, instead, estimating the branching matrix (3.3) directly by matching the integrated cumulants of the process that Jovanović et al. (2015) presented. The latter method is often preferred when the causality relationships between the different types of events of the process are of particular interest. However, this approach does not prevent the kernels from having negative values. Some numerical experiments demonstrated that the obtained estimates are often not feasible from a theoretical point of view. Therefore, the current study focuses on the full estimation of a multivariate Hawkes process.

Initial work in the field of the estimation of Hawkes processes, for example, Ogata and Akaike (1982), focused primarily on the maximum likelihood estimator. In this case, the parametric form of the kernel function needs to be specified in advance; however, this might be too restrictive in practical applications. As mentioned above, the usual choice of kernel is the exponential function due to its remarkable advantage in computational costs. Some attempts were made with the purpose of replacing triggering kernel functions with their nonparametric versions. For example, Zhou et al. (2013c) suggested estimating the triggering kernels from the data by replacing the parametric kernels with a linear combination of the base kernels. The usual choices for the base kernel discussed in the literature are the exponential and Gauss kernels. Moreover, the nonparametric kernel method by Zhou et al. (2013c) was further developed in Zhou et al. (2013a) by including the penalty terms in the likelihood function, and Xu et al. (2016) implemented the penalized nonparametric likelihood function to define and estimate the Granger causality in multivariate Hawkes graphs.

The current work implements another nonparametric procedure to estimate the Granger causality in the multivariate Hawkes process introduced by Embrechts and Kirchner (2016a).

The authors demonstrated that the Hawkes process can be represented as an integer-valued autoregressive (INAR) time series process. It is shown that for a piecewise continuous function $h : \mathbb{R} \to \mathbb{R}_0^+$ with $h(t) = 0$, $t \leq 0$, satisfying $\int h(t)\mathrm{d}t < 1$, constants $\delta > 0$ and $\widetilde{K} < 1$ exist such that

$$K^{(\Delta)} = \Delta \sum_{i=1}^{\infty} h(k\Delta) \leq \widetilde{K} < 1 \text{ for any } \Delta \in (0, \delta). \tag{3.6}$$

Applying the result (3.6), the authors demonstrate that for $\Delta \to 0$,

$$N^{(\Delta)}(A) = \sum_{k:k\Delta \in A} X_k^{(\Delta)} \xrightarrow{w} N,\ A \in \mathcal{B},\ \Delta \in (0, \delta), \tag{3.7}$$

where N is a Hawkes process with immigration intensity λ and reproduction intensities h, $\left(X_n^{(\Delta)}\right)$ is the corresponding INAR (∞) sequence, and $\mathcal{B}$ is a Borel set on $\mathbb{R}$.

Expression (3.7) means that the sequence $\left(X_n^{(\Delta)}\right)$ with immigration parameter $\Delta\lambda$ and reproduction coefficients $\Delta h(k\Delta)$ and $k \in \mathbb{N}$ approximates the bin-count sequences of the considered Hawkes process. The proofs of the above-mentioned results can be found in Kirchner (2016).

In practice, the sample from the Hawkes process described in (3.2) on the time interval $(0, T]$ is considered. The timeline is divided into bins of size $\Delta > 0$, and the following sequence is constructed:

$$X_k^{(\Delta)} = N\Big(((k-1)\Delta, k\Delta]\Big)^{\top},\ k = 1, 2, \ldots, n. \tag{3.8}$$

(3.8) represents the number of observations per bin of point process data with $n = \lfloor \frac{T}{\Delta} \rfloor$. Given representation (3.8), the autoregressive parameters of the sequence $\left(X_n^{(\Delta)}\right)$ are estimated as

$$\left(\widehat{\alpha}_0^{(\Delta)}, \widehat{\alpha}_1^{(\Delta)}, \ldots, \widehat{\alpha}_p^{(\Delta)}\right) = \underset{\left(\widehat{\alpha}_0^{(\Delta)}, \widehat{\alpha}_1^{(\Delta)}, \ldots, \widehat{\alpha}_p^{(\Delta)}\right)}{\arg\min} \sum_{k=p+1}^{n} \left(X_k^{(\Delta)} - \widehat{\alpha}_0^{(\Delta)} - \sum_{l=1}^{p} \widehat{\alpha}_l^{(\Delta)} X_{k-l}^{(\Delta)}\right)^2, \tag{3.9}$$

$k = 1, 2, \ldots, p$. The estimates of the immigration and reproduction intensities of the original process (3.2) can be computed from (3.9), namely, $\widehat{\lambda}_0 = \frac{\widehat{\alpha}_0^{(\Delta)}}{\Delta}$ and $\widehat{h}_k = \frac{\widehat{\alpha}_k^{(\Delta)}}{\Delta}$.

The multivariate Hawkes estimator is consequently defined as

$$\widehat{H}^{(\Delta)} = \frac{1}{\Delta}\left(Z^{\top} Z\right)^{-1} Z^{\top} Y, \tag{3.10}$$

where $Y\left(X_1^{(\Delta)},\ldots,X_n^{(\Delta)}\right)=\left(X_{p+1}^{(\Delta)},X_{p+2}^{(\Delta)},\ldots,X_n^{(\Delta)}\right)^\top$ with $p=\lceil\frac{s}{\Delta}\rceil$ being the order of the considered INAR process and $Z\left(X_1^{(\Delta)},\ldots,X_n^{(\Delta)}\right)$ being the design matrix:

$$Z\left(X_1^{(\Delta)},\ldots,X_n^{(\Delta)}\right)=\begin{pmatrix}\left(X_p^{(\Delta)}\right)^\top & \left(X_{p-1}^{(\Delta)}\right)^\top & \ldots & \left(X_1^{(\Delta)}\right)^\top & 1\\ \left(X_{p+1}^{(\Delta)}\right)^\top & \left(X_p^{(\Delta)}\right)^\top & \ldots & \left(X_2^{(\Delta)}\right)^\top & 1\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ \left(X_{n-1}^{(\Delta)}\right)^\top & \left(X_{n-2}^{(\Delta)}\right)^\top & \ldots & \left(X_{n-p}^{(\Delta)}\right)^\top & 1\end{pmatrix} \tag{3.11}$$

It is important to note that the individual elements of the matrix $\widehat{H}=\left(\widehat{H}_1,\ldots,\widehat{H}_p,\widehat{\lambda}_0\right)^\top$ from (3.10) approximate the excitement functions or triggering kernels, i.e.

$$\widehat{H}_k=\begin{pmatrix}\widehat{h}_{1,1}(k\Delta) & \widehat{h}_{1,2}(k\Delta) & \ldots & \widehat{h}_{1,d}(k\Delta)\\ \widehat{h}_{2,1}(k\Delta) & \widehat{h}_{2,2}(k\Delta) & \ldots & \widehat{h}_{2,d}(k\Delta)\\ \ldots & \ldots & \ldots & \ldots\\ \widehat{h}_{d,1}(k\Delta) & \widehat{h}_{d,2}(k\Delta) & \ldots & \widehat{h}_{d,d}(k\Delta)\end{pmatrix}. \tag{3.12}$$

Kirchner (2016) demonstrated that for a large T, a large p and a small Δ, the elements of $\widehat{H}$ are approximately jointly normally distributed around the true values. Moreover, the covariance matrix of $\mathrm{vec}\left(\widehat{H}^\top\right)$ can be consistently estimated as

$$\widehat{S}^2=\frac{1}{\Delta^2}\left[\left(Z^\top Z\right)^{-1}\otimes\mathbb{1}_{d\times d}\right]W\left[\left(Z^\top Z\right)^{-1}\otimes\mathbb{1}_{d\times d}\right], \tag{3.13}$$

where $W=\sum_{k=p+1}^n w_k w_k^\top$ and

$$w_k=\left(\left(\left(X_{k-1}^{(\Delta)}\right)^\top,\left(X_{k-2}^{(\Delta)}\right)^\top,\ldots,\left(X_{k-p}^{(\Delta)}\right)^\top,1\right)^\top\otimes\mathbb{1}_{d\times d}\right)\cdot\left(X_k^{(\Delta)}-\Delta\widehat{\lambda}-\sum_{l=1}^p\Delta\widehat{H}_l^\top X_{k-l}^{(\Delta)}\right), \tag{3.14}$$

where $k=p+1,p+2,\ldots,n$.

The results presented above enable the estimation of the elements of the branching matrix: $a_{i,j}=\int h_{i,j}(t)\mathrm{d}t$ is estimated as $\widehat{a}_{i,j}=\Delta\sum_{k=1}^p\widehat{h}_{i,j}(k\Delta)$. Moreover, the estimate of the covariance matrix given in (3.13) makes it possible to test whether $\widehat{a}_{i,j}$, $i,j=1,\ldots,d$ is significantly larger than 0 and whether i-type events Granger cause j-type events. This allows one to identify the non-causality of different types of events of a multivariate Hawkes process. As a result, a Hawkes graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, with the adjacency matrix equal to the

branching matrix of the corresponding Hawkes process, can be constructed. If $a_{i,j} > 0$, then an i-type event belongs to the set of the parent nodes of a j-type event, i.e., $i \in \mathrm{PA}(j)$.

The choice of two parameters could potentially influence the quality of the estimator (3.9), namely the choice of the bin width Δ and the order p of the corresponding autoregressive process. Embrechts and Kirchner (2016b) pointed out that the choice of Δ does not heavily influence the estimation of Granger non-causality due to the available testing procedure. In contrast, the quantitative estimation of the branching matrix requires Δ to be small enough and p to be large enough. The authors recommend choosing Δ such that the expected number of observations within the intervals is equal to 1, and they state that the choice of p is less important. In the first step, Embrechts and Kirchner (2016b) recommend choosing quite a large Δ to find the non-zero elements of the branching matrix or the so-called skeleton of the Hawkes graph. The Hawkes skeleton estimator is given by the following set of edges:

$$\widehat{\mathcal{E}} = \{(i,j) \in [d]^2 : \widehat{a}_{i,j} > \widehat{\sigma}_{i,j} z_{1-\alpha}^{-1}\}, \tag{3.15}$$

where $z_{1-\alpha}^{-1}$ is the quantile of a standard normal distribution, and

$$\widehat{\sigma}_{i,j}^2 = \Delta^2 E_{(i-1)d+j,\cdot}^{\top} \widehat{S}^2 E_{(i-1)d+j,\cdot}$$

$\widehat{S}^2$ is given in (3.13) and E_l is the l-th row of matrix $E \in \{0,1\}^{d^2 \times (d^2 p + d)}$, which consists of zeros and ones in row $(i-1)d+j$ at entries $(k-1)d^2 + (i-1)d + j$, $k = 1, 2, \ldots, p$. Advice for the next step, is to choose a much finer Δ and to regress the number of observations per bin only on the past observations of the corresponding parent nodes. In practice, the size of the edge set of the Hawkes skeleton is much smaller than d^2. This makes the procedure computationally tractable in a high-dimensional setting.

3.3 The data

The analysis of this paper focuses on RavenPack News Analytics – Dow Jones Edition. RavenPack is one of the leading providers of real-time news analysis. The data contained in the Dow Jones Edition measures news sentiment and news flow based on Dow Jones Newswires, the Wall Street Journal, Barron's and Marketwatch. The dataset is based on a linguistic analysis of large volumes of articles that contain the world's best business and financial news. In the current study, only news-based sentiment scores were considered; they refer to mainstream media and account for the stories produced by reputable sources.

year	company	min	$q_{0.25}$	median	mean	$q_{0.25}$	max
2005	Morgan Stanley	2.00	30.00	44.00	41.76	54.00	102.00
	HSBC	1.00	18.00	23.00	26.27	30.00	185.00
	Deutsche Bank	1.00	18.00	22.00	22.00	27.00	70.00
	Credit Suisse Group	1.00	12.00	17.00	17.41	22.00	50.00
	Bank of China	1.00	2.00	3.00	3.71	5.00	34.00
2010	Morgan Stanley	2.00	46.00	60.00	63.10	82.00	182.00
	HSBC Holdings	1.00	28.00	38.50	48.54	52.25	902.00
	Deutsche Bank	1.00	21.00	29.00	28.64	37.00	80.00
	Credit Suisse Group	1.00	23.00	31.00	32.62	40.00	175.00
	Bank of China	1.00	1.00	2.00	3.18	4.00	19.00
2015	Morgan Stanley	2.00	72.00	128.00	126.00	172.00	314.00
	HSBC Holdings	1.00	48.00	59.00	60.06	71.00	417.00
	Deutsche Bank	1.00	31.00	42.00	43.41	55.00	130.00
	Credit Suisse Group AG	1.00	15.00	20.00	19.87	26.00	50.00
	Bank of China	1.00	1.00	2.00	2.79	3.00	20.00

Table 3.1 Summary statistics for the number of daily news announcements for selected companies ($q_{0.25}$ and $q_{0.75}$ are the first and the third quartiles correspondingly).

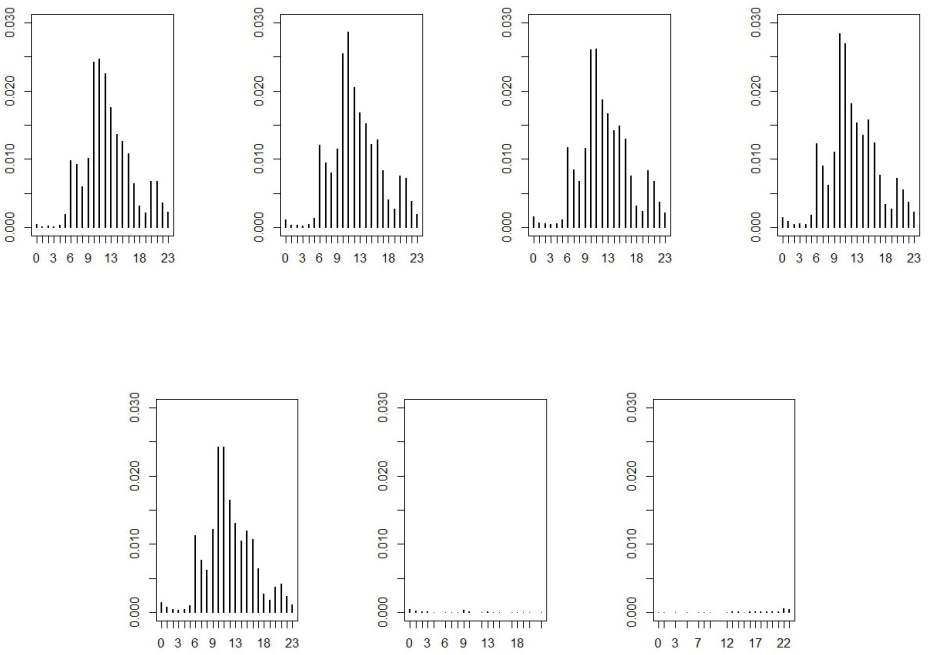


Figure 3.1 The relative frequency of the news announcements.

The quality of social media announcements varies significantly; therefore, this source of information was not used for the analysis. It is important to note that real-time analytics makes it possible to take into account the data of news intensity. Each Dow Jones Edition

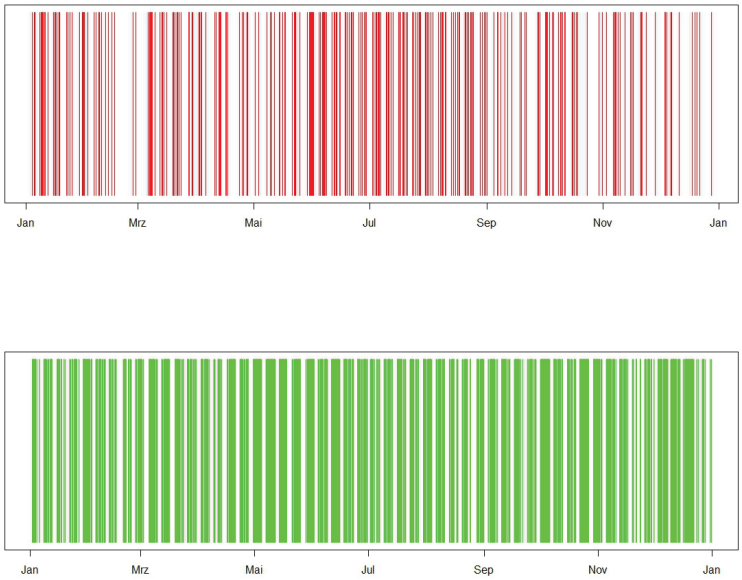


Figure 3.2 Barcode plots of the Deutsche Bank-related negative (upper panel) and positive (lower panel) CSSs from Jan. 1, 2007 to Dec. 31, 2007.

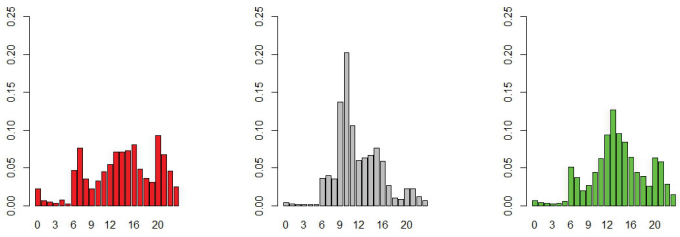


Figure 3.3 The relative frequency of the negative, neutral and positive Morgan Stanley-related news announcements versus time of day.

record contains 48 fields, including a timestamp; company identifiers; scores for relevance, novelty, market impact and sentiment; and unique identifiers for each news story that was analysed. Furthermore, one piece of news can concern several companies.

As Audrino and Tetereva (2017) demonstrated, the financial sector-related news has the strongest impact on the stock returns of all sectors. Therefore, the mutual excitation of news on financial companies is of particular interest. In the studies by Brechmann et al. (2013) and Härdle et al. (2016), the set of influential (in terms of returns) financial institutions was presented. Some companies from this set were selected for the given study, while others

were omitted due to the lack of corresponding news observations – the complete list of included companies is provided in Section 3.4. The analysis focuses on the panel of these companies for the time period between January 1, 2005 and December 31, 2016. In the remaining part of the section, the main characteristics of high-frequency news data and the preparation of the data for further analysis are discussed in greater detail.

The timestamps of the announcements were calculated in milliseconds, and the data are real-time. As the records are based on originally published messages, the frequency of the data strongly depends on the underlying equity. Table 3.1 presents the summary statistics of the number of daily news announcements. It is apparent that the frequency of news differs from company to company and demonstrates a steady increase from 2005 to 2015. For example, the mean number of Morgan Stanley announcements increased from 41.76 to 126 over the considered time period. Figures 3.1 and 3.2 demonstrate the characteristics of high-frequency news announcements in more detail. There is a clear presence of weekly and daily seasonality in the data, and the intensity of news announcements over weekends is much lower. Furthermore, periods such as holidays have rates of news intensity that are below the average. On the contrary, an increase in the rate at periods of quarterly releases is observed. The current analysis is based on the timestamps of the RavenPack composite sentiment score (CSS). This score was chosen because of the larger number of observations available, while other scores are provided for significantly smaller numbers of announcements. For example, the event sentiment score (ESS) was calculated only for 3% of all announcements that mentioned Morgan Stanley during the time period from 2005 to 2016. The CSS is a score between 0 and 100 that represents the news sentiment of a given story by combining various sentiment analysis techniques. Observations with score values larger than 50 were considered as positive signals, whereas negative signals were associated with a CSS smaller than 50. News announcements with a CSS equal to 50 were considered to be neutral. The direction of the score was determined by looking at emotionally charged words and phrases. It is important to note, that most announcements of a given data set are neutral. For example, 63% of all Morgan Stanley-related announcements correspond to a CSS equal to 50.

In contrast to other studies – for example, Hsuan (2017) – the CSS data do not exhibit more positive scores in the morning. Hsuan (2017) stated that news sentiment is mostly positive at the beginning of the trading day, which means that investors start the new trading day with optimistic expectations. Figure 3.3 demonstrates the hourly relative frequencies of negative, neutral and positive scores for Morgan Stanley for the time period between 2005 and 2016, and it suggests that a change in the CSS occurs mostly at noon.

This pattern is more pronounced for the positive and neutral CSSs and less evident for negative ones.

Figure 3.2 illustrates the arrival times of the positive and negative news for Deutsche Bank for the year 2007. A line represents each event; the vertical axis in this plot is for visualization only and has no further meaning. It is possible to assess the visible clusters of news arrival times from the plot – some of them are due to seasonality, which is discussed in the rest of this section, while the others provide convincing evidence of the contribution of previous events to the current intensity. On the basis of the empirical evidence available, it seems fair to suggest that Hawkes processes might be useful tools for better understanding and modelling of the news arrival processes.

One way in which to handle the observed weekly seasonality is to exclude all weekend news from the analysis. However, the objective of this paper is to study the mutual excitation of news announcements, and some important releases could potentially appear outside the usual business hours. For transaction data, it is well known that the frequency of trading is higher near the open and the close of market, and slightly longer durations are usually observed around noon. Such effects are not found in news announcement data. From Figure 3.1, it can be seen that the intensity of news releases is higher around noon, and another minor increase in intensity appears at approximately 8 PM.

As mentioned in Engle and Russell (1997), in the context of the autoregressive conditional duration models, the expected durations can be decomposed into deterministic and stochastic components. The stochastic component can then be considered as a proxy for the relative news activity. Engle and Russell (1997) proposed seasonally or diurnally adjusting the data first and then fitting the model to the data, and Grammig and Maurer (2000) and Zhang et al. (2001) mentioned the same approaches. Durations are divided by the estimated daily cyclical component, and the seasonality adjustment is commonly made by estimating the seasonal component using nonparametric regression methods, for example, splines, a Fourier series or Nadaraya-Watson regression; more details on nonparametric regression methods can be found in the work of Wasserman (2006). As is evident from both Figure 3.1 and the barcodes presented in Figure 3.2, there are two main types of seasonality in the news time series: daily and weekly. Following Hautsch (2004), the durations of news announcements were adjusted by applying cubic splines. The deterministic component was captured by the diurnal factor function, which is obtained by fitting cubic splines. The resulting diurnally adjusted duration was thus obtained by dividing the original duration by the diurnal factor associated with the time point at which the news occurred. Due to the dual nature of seasonality in the news time series, the adjustment was made in two steps.

First, the adjustment for the daily component was made. Thereafter, the weekly seasonality was removed. More details on seasonality adjustment for duration models can be found in Kwok et al. (2009).

3.4 Results

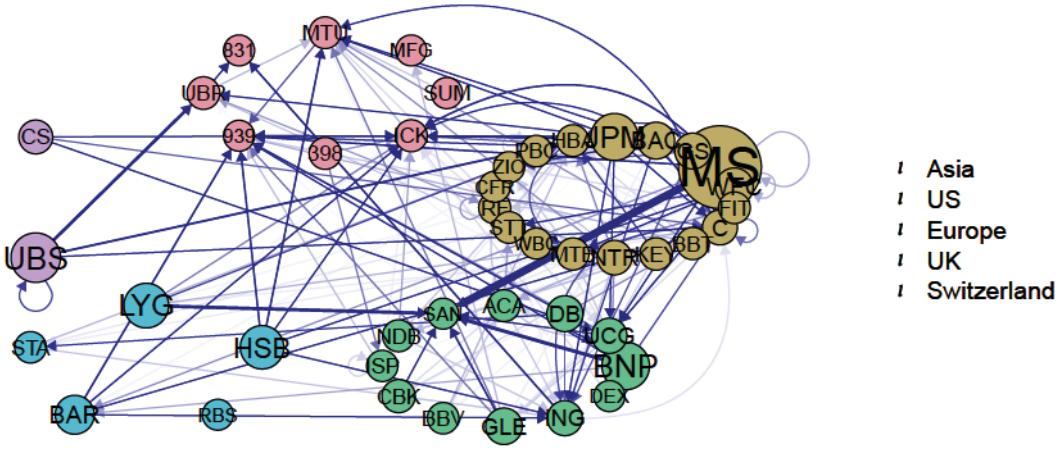


Figure 3.4 The estimated skeleton of Hawkes process for Oct. 11, 2007.

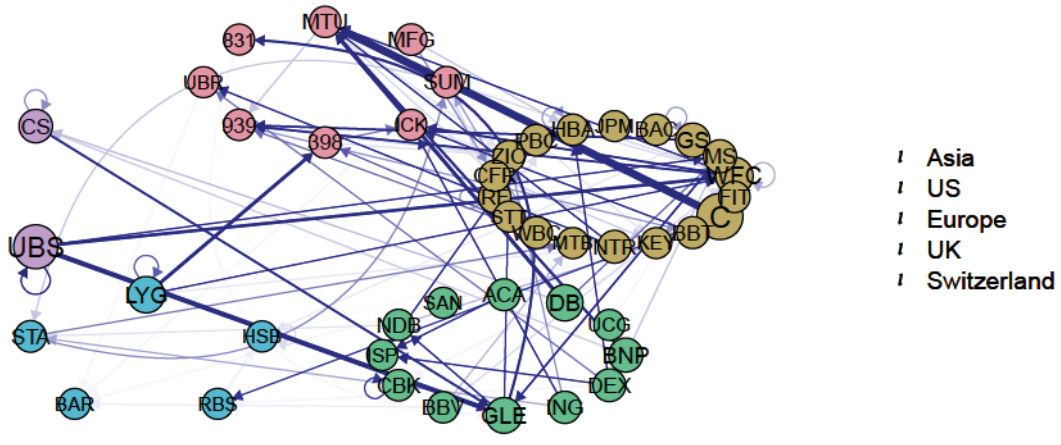


Figure 3.5 The estimated skeleton of Hawkes process for Mar. 18, 2014.

The data set contains news observations for 45 potentially influential financial companies for a time period of 12 years. The selection of companies was motivated by the previous

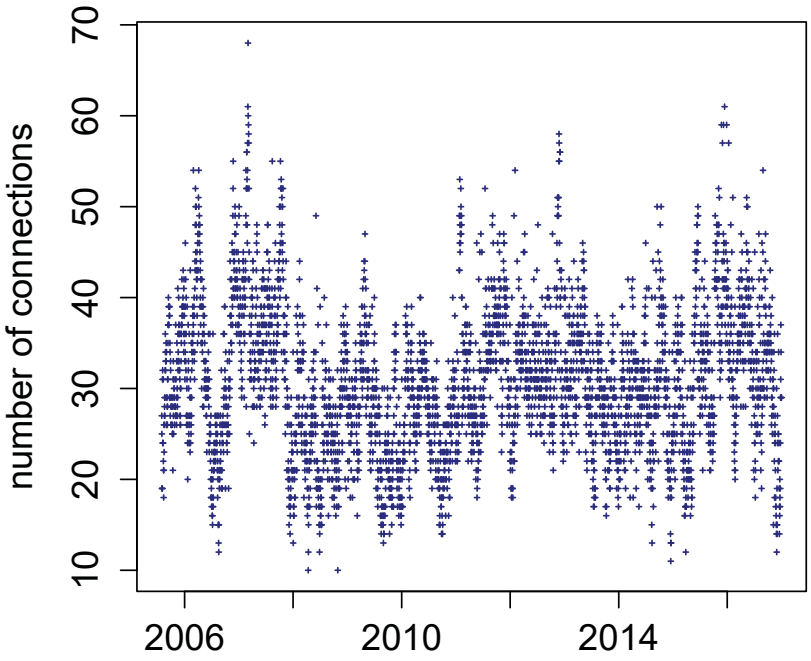


Figure 3.6 The total number of outgoing edges corresponding to the nodes representing Royal Bank of Scotland, Citigroup, Wells Fargo, UBS Group, Credit Suisse, Deutsche Bank.

studies on contagion effects in stock returns, for example, Brechmann et al. (2013) and Härdle et al. (2016). For such a long time period, it can not be assumed that the systemic contribution of each individual company remains unchanged over time. For this reason, a rolling window analysis was performed. The size of the window was chosen to be 200 days, which is standard for the dynamic modelling literature. Moreover, the chosen size of the window allows for ample observations for the estimation procedure for each company.

To assess the influence and evaluate the news diffusion among financial companies, the methodology described in Section 3.2 was used. Furthermore, to approximate the Hawkes process with the INAR(p) process and consequently learn the Granger causality for a Hawkes graph, one needs to select the size of the bin Δ and the order of the autoregressive process. As suggested in Embrechts and Kirchner (2016b), Δ was chosen to ensure that the mean value of the observations in a bin is equal to 1. This parameter was fixed for the first and second steps of the estimation procedure. The first step of the estimation involves the estimation of the Hawkes skeleton. In other words, in the first step, one is

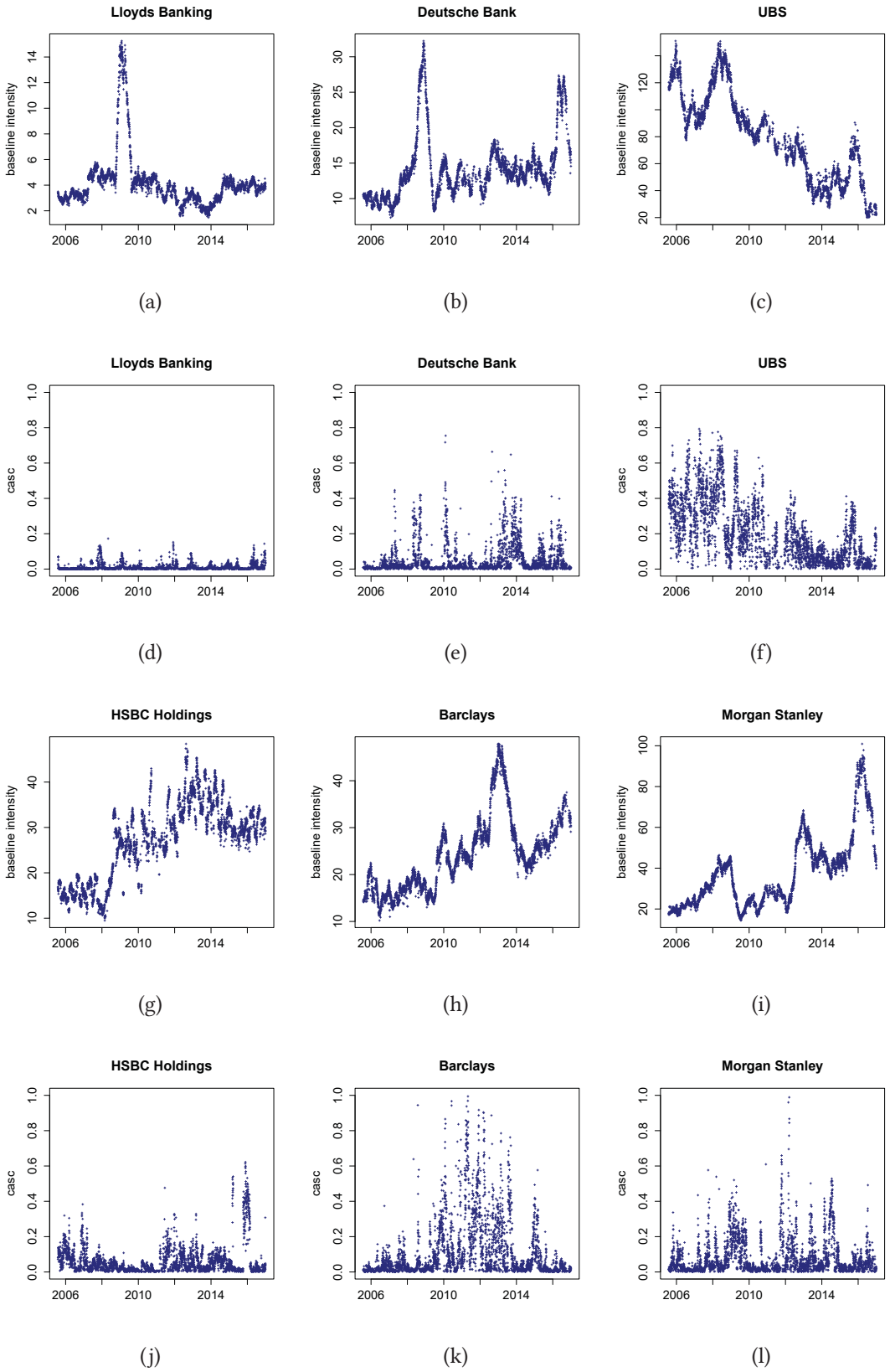


Figure 3.7 The estimated baseline intensities and estimated cascade coefficients for selected companies.

company	ticker	$\overline{\text{casc}}$	sd(casc)	$\overline{\text{feed}}$	sd(feed)
Barclays	BARC	**0.106**	0.169	0.283	0.260
Citigroup	C	**0.104**	0.155	0.299	0.276
UBS	UBS	**0.077**	0.179	0.349	0.309
Morgan Stanley	MS	**0.071**	0.106	0.274	0.265
HSBC Holdings	HSBC	**0.055**	0.084	0.431	0.362
Wells Fargo	WFC	**0.053**	0.084	0.227	0.237
Bank of America	BA	**0.046**	0.089	0.239	0.245
Deutsche Bank	DB	**0.044**	0.078	0.247	0.251
Credit Suisse	CS	**0.038**	0.076	0.240	0.247
JPMorgan Chase	JPM	**0.030**	0.059	0.252	0.256
Goldman Sachs	GS	**0.025**	0.045	0.225	0.240
Credit Agricole	ACA	**0.025**	0.044	0.220	0.269
ING	ING	0.019	0.039	0.261	0.311
State Street	STT	0.018	0.035	0.217	0.290
Bank of China	3988	0.018	0.029	0.263	0.315
Sumitomo	SUMA	0.016	0.031	0.214	0.284
Nordea Bank	NDB	0.016	0.038	0.193	0.238
Societe Generale	GLE	0.016	0.032	0.219	0.275
Mitsubishi UFJ	MTUN	0.015	0.043	0.183	0.266
Dexia	DEXB	0.015	0.040	0.236	0.303
UniCredit	UCGR	0.013	0.028	0.248	0.313
Standard Chartered	STAN	0.011	0.021	0.271	0.323
Zions	ZION	0.011	0.016	0.279	0.356
Fifth Third Bancorp	FITB	0.010	0.027	0.204	0.275
Westpac Banking	WBC	0.009	0.019	0.231	0.285
Commerzbank	CBK	0.009	0.023	0.243	0.302
BB&T	BBT	0.009	0.023	0.189	0.262
Banco Bilbao	BBVA	0.009	0.019	0.207	0.263
Lloyds Banking	LYG	0.009	0.018	0.200	0.269
Intesa Sanpaolo	ISP	0.008	0.019	0.235	0.324
China Construction Bank	UBRA	0.008	0.015	0.239	0.306
BNP Paribas	BNP	0.007	0.014	0.198	0.231
Industrial & Commercial Bank of China	ICK	0.007	0.013	0.236	0.327
Huntington Bancshares	HBAN	0.006	0.018	0.170	0.289
Royal Bank of Scotland	RBS	0.006	0.012	0.269	0.336
M&T Bank	MTB	0.005	0.009	0.234	0.336
Regions Financial	RF	0.005	0.010	0.199	0.288
Mizuho Financial Group	MFG	0.005	0.012	0.236	0.295
Banco Santander	SAN	0.005	0.024	0.165	0.332
KeyCorp	KEY	0.004	0.009	0.227	0.336
Northern Trust	NTRS	0.004	0.008	0.183	0.217
Sumitomo Mitsui Banking	8316	0.002	0.005	0.391	0.476
Cullen/Frost Bankers	CFR	0.002	0.004	0.187	0.342
BOK Financial	939	0.001	0.003	0.309	0.435
People's United Financial	PBCT	0.000	0.003	0.190	0.344

Table 3.2 Mean values and standard deviations of the cascade coefficients (casc) (3.4) and the feedback coefficients (feed) (3.5).

interested in detecting the set of parent nodes for each node, and this allows the strength of the connections among the nodes to be misspecified. In the current study, s was fixed at a value equal to 3 days in the first step, and p was consequently computed as $\lceil \frac{s}{\Delta} \rceil$. Here, it

is important to note, that the set of parent nodes was found by applying the (3.13)-based testing procedure described in Section 3.2. As mentioned in Embrechts and Kirchner (2016b), this procedure is robust to the choice of p.

The skeletons containing 45 nodes were estimated using 200-day rolling windows. The minimum number of non-zero edges corresponded to March 18, 2014, while the most connected skeleton was estimated for October 11, 2007. The corresponding skeletons are displayed in Figures 3.4 and 3.5. The size of the node is proportional to the number of outgoing edges. From the estimated Hawkes graphs, it can be seen that US banks are interconnected the most. Other outgoing edges are mostly generated by UK banks, the Deutsche Bank and UBS. The company names corresponding to the tickers used in these figures are presented in Table 3.2. From the rolling window analysis, it was evident that the number of connections within the skeleton changes significantly over time. For example, Figure 3.6 considers the number of the outgoing edges corresponding to the banks with the largest losses during the 2008 crisis. The chosen set of companies contains mainly European companies. Therefore, it was observed that the number of outgoing connections increased during the European sovereign debt crisis in 2010 and Brexit voting in 2016. The Madoff investment scandal is another possible explanation for the increasing connectivity among financial companies in late 2008.

After estimating the skeleton and reducing the number of parent nodes for each company, it was possible to estimate the weights of the edges more precisely. The value of Δ was left unchanged to ensure that the mean number of observations in a bin is equal to 1. The parameter s was set to 7 days, and p was computed as $\lceil \frac{s}{\Delta} \rceil$. In this way, the potential influence of events within the last week is taken into account. After estimating the branching matrix (3.3), the cascade (3.4) and the feedback (3.5) coefficients could be computed. As mentioned above, the cascade coefficients measure the contribution of each individual company to the whole news intensity of the system; the feedback coefficients demonstrate the proportion of that intensity caused by the past events of the corresponding company itself. Table 3.2 presents the mean values and standard deviations of the cascade and feedback coefficients over a period of 12 years. As stated in Embrechts and Kirchner (2016b), the events with cascade coefficients greater than $\frac{1}{d}$, where d is the dimension of the data set, can be considered to be systemically important. In the case of $d = 45$, companies with a cascade coefficient greater that 0.022 significantly influence the news intensity of the whole system. Such companies are marked in bold in Table 3.2.

Generally speaking, evidence of systemic influence in terms of news was found for US and UK financial institutions, and the highest mean value of the cascade coefficient

corresponded to Barclays and Citigroup. Apart from US and UK banks, UBS seemed to influence news diffusion in the whole system. Empirical evidence of the importance of news about Deutsche Bank was found, while no significant patterns were found for the feedback coefficients. However, the mean values of the feedback coefficients for all companies were much smaller than 1. This means that financial institutions are connected to each other in terms of news information diffusion, and the news intensity of a company can be only partially explained by the past news arrivals of the company itself.

It is important to note that the mean values presented in Table 3.2 are taken over a period of 12 years and should therefore be interpreted carefully. Figure 3.7 presents the rolling window results for selected companies. In the remaining part of the section, the permanent importance of some companies and the information flow generated by the group of vulnerable banks are explained. Contrary to intuition, the baseline intensities do not grow constantly over time, despite the constantly increasing volumes of information available. Moreover, the increasing baseline intensity does not necessarily lead to greater influence of a company on the whole system. For example, the increase in the baseline intensity of news related to Lloyds Bank corresponds to January 19, 2009 – during that period, after long discussions Lloyds took over HBOS, and this event generated a spike in the baseline intensity of that bank (Figure 3.7(a)). However, the exploding baseline intensity had a minor impact on the corresponding cascade coefficient (Figure 3.7(d)). The Deutsche Bank is another systemically important European bank . The first peak of the baseline intensity of Deutsche Bank news corresponds to the global financial crisis in 2008, and the second peak relates to the breaking news on Brexit (Figure 3.7(b)). Moreover, in this case, the increase in the baseline intensity can be matched to the increasing cascade coefficients (Figure 3.7(e)). Figures 3.7(c) and 3.7(f) depict the estimated baseline intensity and cascade coefficients for UBS. This bank is considered to have suffered much during the global financial crisis, and the slowly decaying cascade coefficient is evidence of this fact. Moreover, the peak of the baseline intensity corresponding to the UBS tax evasion controversy is observed in early 2008. The stabilisation of UBS, starting from 2009, resulted in the decaying baseline intensity and moderate cascade coefficients. Figures 3.7(h) and 3.7(k) offer clear evidence of the intense systemic influence of Barclays. The spike in the corresponding cascade coefficient can be related to LIBOR manipulations by Barclays and the Deutsche Bank in 2012. It is important to note that the cascade coefficient began to grow before Barclays-related troubles. This means that in particular cases, the increasing influence of firm-specific news announcements can be seen as early warning signals of coming instability. The detailed rolling windows analysis for HSBC and Morgan Stanley is presented in Figure 3.7. It is

apparent that the first peak of the cascade coefficient corresponds to the report of record earnings of a British bank, while the second peak might be related to the discussion on Brexit voting. The increase in the influence of Morgan Stanley in 2011 might be explained by the fact that Morgan Stanley had to pay out multiple fines to settle lawsuits initiated by regulators. The detailed analyses for all companies are available from the author upon request.

3.5 Application – building a news intensity index

lag	NII → VIX	NII → price	NII → volume	VIX → NII
1	0.0302	0.0087	0.0326	0.1600
2	0.0101	0.0016	0.0439	0.0889
3	0.0179	0.0062	0.0305	0.0750
4	0.0360	0.0223	0.0038	0.1116
5	0.0365	0.0347	0.0052	0.1135
6	0.0497	0.0116	0.0084	0.0871

Table 3.3 Granger causality between the NII and VIX, S&P 500 price and volume (monthly data).

One of the objectives of financial economics is to construct market indices. On the one hand, such statistical measures are of obvious interest to policy makers. On the other hand, some market indices allow for the diagnosis of the health of a financial system as a whole. The construction of early warning indicators of distress gained more attention in academia and the industry after the global financial crisis. Most research in this area is devoted to systemic risk measures. Among others, measures such as SRISK (a conditional capital shortfall measure of systemic risk) by Brownlees and Engle (2016) and CoVaR (conditional Value at Risk) by Adrian and Brunnermeier (2016) are of particular importance. The composite SRISK index is constructed to describe the overall risk of a financial system, and it measures the total amount of capital that a government would have to provide in the case of financial stress. This measure can be used to monitor the system and detect early signs of financial instability. On the contrary, CoVaR measures the contribution of an individual company to the risk of the whole system in terms of Value at Risk. Another popular index for the stability of financial systems is VIX, which is the expected market variance of the S&P 500. This index is often called a 'fear index' as it reflects uncertainty and variance risk premium.

All the above-mentioned indices are based on the observations originating from the asset market. However, recent findings in the sentiment literature illustrate that stock prices

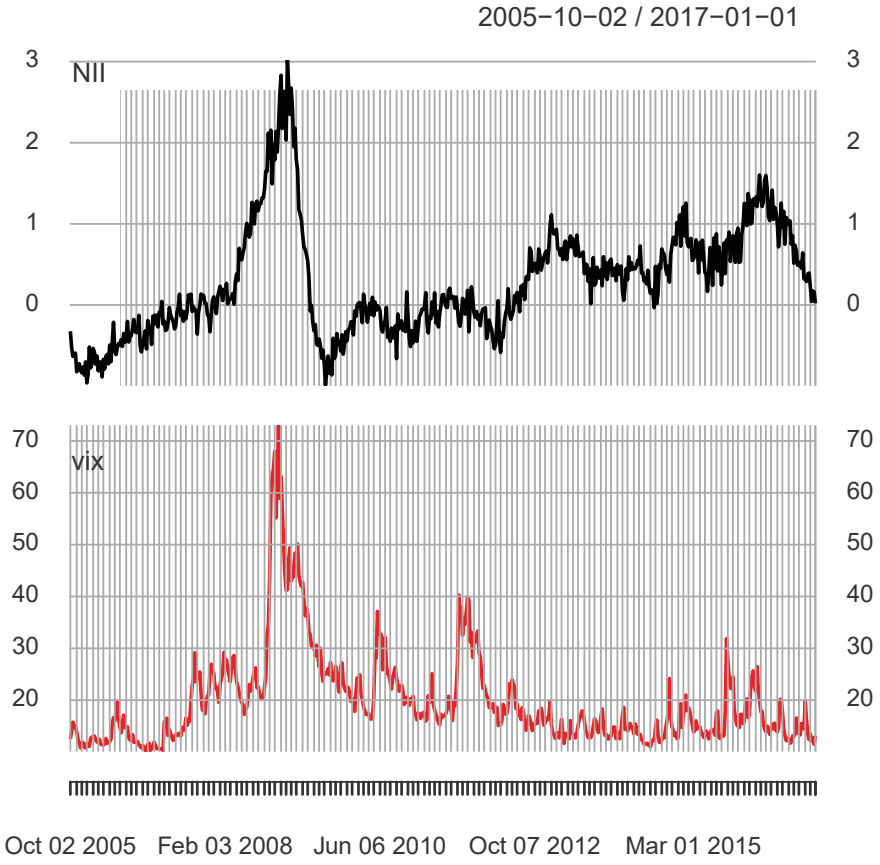
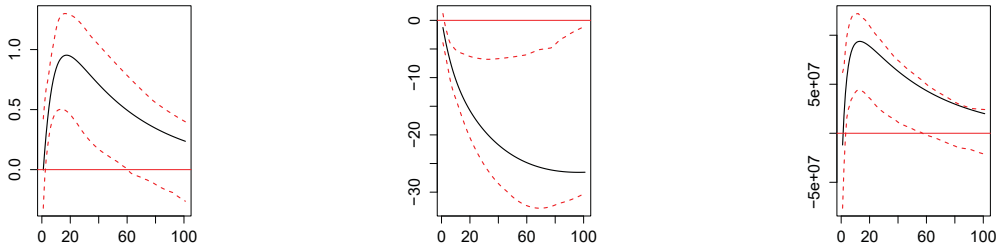
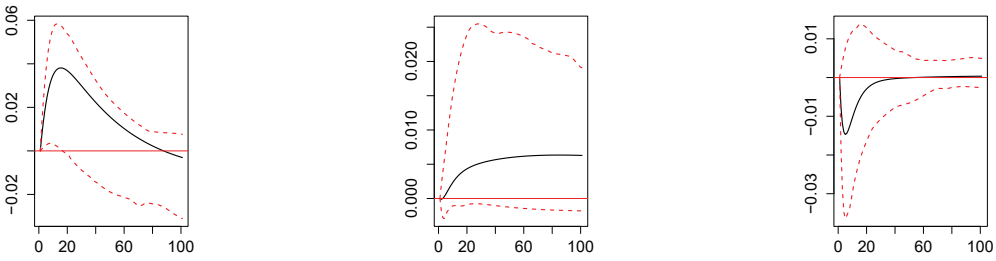


Figure 3.8 NII vs. VIX (weekly data).



(a) NII → VIX (b) NII → price (c) NII → volume

Figure 3.9 The impulse response function (VAR[1], weekly data) – NII on VIX and S&P500 price and volume.



(a) VIX → NII (b) price → NII (c) volume → NII

Figure 3.10 The impulse response function (VAR[1], weekly data) – VIX and S&P 500 price and volume on the NII.

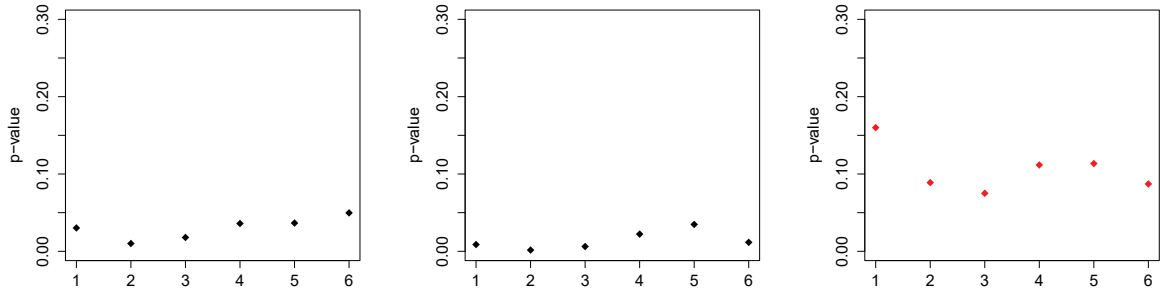


Figure 3.11 Granger causality test – p-values for NII against VIX (left), NII against S&P 500 price (center) and VIX against NII (right) (monthly data).

are often driven by information rather than reality. Therefore, news- and media-based indicators could be of particular importance for predicting the stability of a financial system. One of the attempts to construct a news-based index is referred to as sentiment-based systemic risk indicator (SenSR) by Borovkova et al. (2017). It is a composite index, which accounts for the sentiments of important financial companies. The main advantage of this index is its Granger causality to other indices, such as SRISK and VIX, and its ability to detect early signals of financial stress. The main limitation of SenSR is that its performance heavily depends on the selection of the weights for individual companies in the composite index. Weights that are proportional to the leverage, market value or debt have been proposed by the authors. Moreover, the sentiment indices for individual companies should

be calculated by taking into account the novelty, relevance and other characteristics of firm-specific news announcements.

Apart from the usefulness of the sentiment indices, the predictive power of news intensity is widely discussed in the literature. For example, Sidorov et al. (2013) demonstrated that the GARCH model, augmented by the daily number of press releases on a stock, has more predictive power. Moreover, as has been demonstrated in Section 3.4, the health of a financial system could be alternatively measured in terms of the news intensity. It has been observed that the intensity of firm-specific announcements is significantly higher around stress events. Therefore, the weighted average baseline intensity can be seen as a proxy for panic among traders. However, the increasing intensity of firm-specific announcements does not necessarily lead to the spread of distress to other financial institutions. The influence of a company in the system can be described by means of cascade coefficients (3.5). It is important to note that, in general, cascade coefficients are not proportional to the baseline intensity of firm-specific announcements or the size of the company. In this paper, the construction of a news intensity index (NII) is suggested as the weighted average of the baseline intensities, with the weights being equal to the cascade coefficients:

$$\text{NII} = \sum_{i \in I} c_i \cdot \widehat{\lambda}_i^{\text{std}}, \tag{3.16}$$

where $I = \{i : c_i > \frac{1}{d}\}$, $\widehat{\lambda}_i^{\text{std}}$ is the standardised estimated baseline intensity, and d is the number of companies. Only the systemically important companies with cascade coefficients larger than $1/d$ are taken into account when the index is constructed. The advantage of the proposed news index (3.16) is that it considers only news arrival times, and it is still valid for characterising and predicting the health of a financial system. This allows one to avoid the calculation of the firm-specific sentiment index of each announcement.

For illustrative purposes, the NII for the US market is considered. Country-specific or industry-specific news intensity based indices can be analogically constructed. The upper panel of Figure 3.8 displays the NII for the US financial sector, which has been constructed based on the news intensities of US companies from Table 3.2. The lower panel of Figure 3.8 presents the closest benchmark of the proposed news index in terms of the mood of the traders. It is observed that the NII had more predictive power during the global financial crisis. Moreover, it was able to mimic the uncertainty of the market participants around Brexit voting and during the debt problems of European banks.

To test the predictive power of the NII for VIX, S&P 500 volume and S&P 500 price, several formal tests are conducted. First, the impulse response functions are studied. The

results are presented in Figure 3.9. It can be concluded that NII has a significant positive impact on VIX and S&P 500 volume, and it negatively influences S&P 500 price. The influence of the NII on VIX and S&P 500 volume is significant for approximately a one-year period. The confidence intervals are constructed according to Hafner and Herwartz (2009), which allows one to correct for potential heteroscedasticity and autocorrelations. In contrast, VIX, S&P 500 price and volume do not significantly affect the proposed NII; this can be seen as proof that the NII contains novel information, affording more accurate and timely signals of market instability.

In addition to the analysis of the impulse response function, the Granger causality test is conducted. In this section, the original definition of the Granger causality introduced by Granger (1969) is used. Table 3.3 presents the p-values of the corresponding F-test for the monthly data. Lower granularity is used here to test for the Granger causality of a longer time period. It is concluded that the proposed NII measure Granger causes VIX and the price and volume of the S&P 500 at a 5% significance level at time lags up to six months. The opposite is not true; past values of VIX do not contain information that helps to predict NII values beyond the information contained in past values of the NII alone. For illustrative purposes, the corresponding p-values are graphically presented in Figure 3.11. It is observed that the predictive power of the NII is slowly decaying; however, it is still significant at a 6-month lag. Black points indicate the lags that are significant at a 5% level, and red points correspond to non-significant lags. In summary, the introduced NII contains timely information about the mood of market participants, and it can be used as an early warning signal of distress in other indicators.

Conclusion

This paper highlighted the importance of studying news data at higher frequencies than weekly or daily. It was suggested to model the news arrival times by means of multivariate Hawkes processes. This approach provided a new method for assessing the systemic importance of an individual company in the news diffusion process.

RavenPack sentiment data from January 1, 2005 to December 31, 2016 have been used to study the nature of real-time news. Empirical evidence has been found that news arrival times are triggered by the past values of both a company itself and other companies. This empirical finding adds to a growing body of literature on sentiment analysis by indicating that financial companies are contagious to each other in terms of news.

On average, it has been found that US and UK financial companies are spreading news to other countries. The strong influence of UBS and Deutsche Bank is also identified. The most remarkable result to emerge from the data is that several companies drive the news diffusion process. With a few exceptions, the obtained results have reveal that the mutual Granger causality of the news arrival process dominates the autoregressive component. In other words, the news arrival process of a company is triggered by news about other companies. Another interesting finding is that in analogue to stock returns, mutual dependencies of news arrival times are higher during times of instability.

This finding might have broad applications. The present study suggests using the obtained cascade coefficients as the weights to construct the composite news intensity index (NII) for the US market. The advantage of the proposed measure is that it uses only the intensity of the news as an input and avoids the step of measuring the relevance or sentiment of announcements. This makes the index robust with regard to the choice of dictionary and the computations of novelty and relevance.

The relevance of this index was supported by analysing its Granger causality to VIX and S&P 500 price and volume. The NII's predictive power has been demonstrated for these indicators at a 6-month lag. Therefore, the proposed index provides timely information about the mood of traders and the health of the financial system. Moreover, this result might be of obvious interest to policy makers as it provides an early warning signal of the future changes in the market.

There is much room for further progress in constructing news intensity indices for different countries and industries. More research is required to better understand whether the proposed index can be improved by distinguishing between positive and negative signals originating from the news. Moreover, it would be interesting to include the constructed index into time series models or to study whether the use of the index can enhance trading strategies.

Bibliography

Achab, M., Bacry, E., Gaïffas, S., Mastromatteo, I. and Muzy, J.-F. (2016). Uncovering Causality from Multivariate Hawkes Integrated Cumulants. ArXiv preprint arXiv:1607.06333.

Adrian, T. and Brunnermeier, M. K. (2016). CoVaR, *The American Economic Review* **106**(7): 1705–1741.

Akyildirim, E., Altarovici, A. and Ekinci, C. (2015). Effects of Firm-Specific Public Announcements on Market Dynamics: Implications for High-Frequency Traders, *Handbook of High Frequency Trading* .

Allen, D. E., McAleer, M. J. and Singh, A. K. (2015). Machine News and Volatility: The Dow Jones Industrial Average and the TRNA Real-Time High-Frequency Sentiment Series, *The Handbook of High Frequency Trading*, Academic Press San Diego, pp. 327–344.

Andersen, L. and Sidenius, J. (2004). Extensions to the Gaussian copula: Random recovery and random factor loadings, *Journal of Credit Risk Volume* **1**(1): 29–70.

Andersen, T., Bollerslev, T., Diebold, F. and Labys, P. (2002). Modeling and forecasting realized volatility, *Econometrica* **71**(2): 579–625.

Andersen, T. G., Davis, R. A., Kreiss, J.-P. and Mikosch, T. V. (2009). *Handbook of financial time series*, Springer Science & Business Media.

Appel, G. (2003). Become Your Own Technical Analyst: How to Identify Significant Market Turning Points Using the Moving Average Convergence-Divergence Indicator or MACD, *The Journal of Wealth Management* **6**(1): 27–36.

Arnold, A., Liu, Y. and Abe, N. (2007). Temporal causal modeling with graphical granger methods, *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 66–75.

Audrino, F. and Camponovo, L. (2015). Oracle properties and finite sample inference of the adaptive lasso for time series regression models. ArXiv preprint arXiv:1312.1473.

Audrino, F. and Corsi, F. (2010). Modeling Tick-by-Tick Realized Correlations, *Computational Statistics & Data Analysis* **54**(11): 2372–2382.

Audrino, F. and Knaus, S. D. (2016). Lassoing the HAR model: A model selection perspective on realized volatility dynamics, *Econometric Reviews* **35**(8-10): 1485–1521.

Audrino, F. and Tetereva, A. (2017). Sentiment spillover effects for US and European companies. Available at SSRN: https://ssrn.com/abstract=2957581.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns, *The Journal of Finance* **61**(4): 1645–1680.

Baker, M. and Wurgler, J. (2007). Investor sentiment in the stock market, *The Journal of Economic Perspectives* **21**(2): 129–151.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2004). *Regular and modified kernel-based estimators of integrated variance: The case with independent noise*, University of Oslo.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. and Shephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise, *Econometrica* **76**(6): 1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps, *Journal of Financial Econometrics* **2**(1): 1–37.

Bauer, G. H. and Vorkink, K. (2011). Forecasting multivariate realized stock market volatility, *Journal of Econometrics* **160**(1): 93–101.

Bauwens, L. and Giot, P. (2001). The moments of first-order Log-ACD models. Unpublished Paper, Université Catholique de Louvain.

Bauwens, L. and Hautsch, N. (2009). Modelling financial high frequency data using point processes, *Handbook of financial time series* pp. 953–979.

Bauwens, L., Storti, G., Violante, F. et al. (2012). Dynamic conditional correlation models for realized covariance matrices. CORE DP 2012/60.

Becker, B. (2016a). Sentiment May Have Bottomed For Volkswagen, http://seekingalpha.com/article/3957529-sentiment-may-bottomed-volkswagen. Accessed: 2017-12-01.

Becker, B. (2016b). Starbucks' Bitter Rewards – Good For Investors?, http://lipperalpha.financial.thomsonreuters.com/2016/05/starbucks-bitter-rewards-good-for-investors/. Accessed: 2017-12-01.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines, *Annals of Mathematics and Artificial intelligence* **32**(1): 245–268.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**(3): 307–327.

Bollerslev, T., Patton, A. J. and Quaedvlieg, R. (2016). Modeling and Forecasting (Un)Reliable Realized Covariances for More Reliable Financial Decisions. Available at SSRN: https://ssrn.com/abstract=2759388.

Borovkova, S. (2015). The Role of News in Commodity Markets. Available at SSRN: https://ssrn.com/abstract=2587285.

Borovkova, S., Garmaev, E., Lammers, P. and Rustige, J. (2016). SenSR: A Sentiment-Based Systemic Risk Indicator. Available at SSRN: https://ssrn.com/abstract=2759289.

Borovkova, S., Garmaev, E., Lammers, P. and Rustige, J. (2017). SenSR: A Sentiment-Based Systemic Risk Indicator. Available at SSRN: https://ssrn.com/abstract=2951036.

Borovkova, S. and Lammiman, A. (2010). The impact of news sentiment on energy futures returns. Vrije Universiteit Amsterdam, https://sbe.vu.nl/en/Images/paper_borovkova_tcm258-204330.pdf.

Borovkova, S. and Mahakena, D. (2015). News, volatility and jumps: the case of natural gas futures, *Quantitative Finance* **15**(7): 1217–1242.

Brechmann, E. C., Hendrich, K. and Czado, C. (2013). Conditional copula simulation for systemic risk stress testing, *Insurance: Mathematics and Economics* **53**(3): 722–732.

Brenner, M. and Galai, D. (1989). New financial instruments for hedge changes in volatility, *Financial Analysts Journal* **45**(4): 61–65.

Breymann, W., Dias, A. and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance, *Quantitative Finance* **3**(1): 1–14.

Brown, G. W. and Cliff, M. T. (2004). Investor sentiment and the near-term stock market, *Journal of Empirical Finance* **11**(1): 1–27.

Brownlees, C. and Engle, R. F. (2016). SRISK: A conditional capital shortfall measure of systemic risk, *The Review of Financial Studies* **30**(1): 48–79.

Brownlees, C. T. and Engle, R. F. (2015). SRISK: A conditional capital shortfall index for systemic risk measurement, *Department of Finance, New York University* .

Brownless, C. and Gallo, G. (2006). Financial econometric analysis at ultra-high frequency: Data handling concerns, *Computational Statistics & Data Analysis* **51**(4): 2232–2245.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Cahan, R., Jussa, J. and Luo, Y. (2009). Breaking news: How to use news sentiment to pick stocks. Macquarie US Equity Research.

Cerchiello, P., Nicola, G. et al. (2017). Assessing News Contagion in Finance, *Technical report*, University of Pavia, Department of Economics and Management.

Cherubini, U., Mulinacci, S., Gobbi, F. and Romagnoli, S. (2011). *Dynamic Copula methods in finance*, Vol. 625, John Wiley & Sons.

Chiriac, R. and Voev, V. (2011). Modelling and forecasting multivariate realized volatility, *Journal of Econometrics* **26**(6): 922–947.

Christoffersen, P. and Pelletier, D. (2004). Backtesting Value-at-Risk: a duration-based approach, *Journal of Financial Econometrics* **2**(1): 84–108.

Clements, A., Hurn, A., Lindsay, K. and Volkov, V. (2017). A semi-parametric point process model of the interactions between equity markets. Working paper.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility, *Journal of Financial Econometrics* **7**(2): 174–196.

Creal, D., Koopman, S. J. and Lucas, A. (2013). Generalized autoregressive score models with applications, *Journal of Applied Econometrics* **28**(5): 777–795.

Czado, C. (2010). Pair-copula constructions of multivariate copulas, *Copula theory and its applications*, pp. 93–109.

Dias, A., Embrechts, P. et al. (2004). Dynamic copula models for multivariate high-frequency data in finance. Manuscript, ETH Zurich 81.

Diebold, F. X. and Mariano, R. (1995). Comparing forecast accuracy, *Journal of Business and Economic Statistics* **13**(3): 253–263.

Ding, X., Zhang, Y., Liu, T. and Duan, J. (2015). Deep learning for event-driven stock prediction, *Proceedings of the 24th International Joint Conference on Artificial Intelligence (ICJAI'15)*, pp. 2327–2333.

Durante, F. and Sempi, C. (2015). *Principles of Copula Theory*, Chapman and Hall.

Durbin, J. and Koopman, S. (2001). Time Series Analysis by Space State Methods, *Nova Iorque: Oxford University Press* .

Eichler, M., Dahlhaus, R. and Dueck, J. (2017). Graphical modeling for multivariate hawkes processes with nonparametric link functions, *Journal of Time Series Analysis* **38**(2): 225–242.

Embrechts, P., Hoeing, A. and Juri, A. (2003). Using Copulae to bound the Value-at-Risk for functions of dependent risks, *Finance & Stochastics* **7**(2): 145–167.

Embrechts, P. and Kirchner, M. (2016a). Hawkes graphs. ArXiv preprint arXiv:1601.01879.

Embrechts, P. and Kirchner, M. (2016b). Hawkes graphs. ArXiv preprint arXiv:1601.01879.

Embrechts, P., McNeil, A. and Straumann, D. (1999). Correlation: Pitfalls and alternatives, *RISK Magazine* pp. 69–71.

Engle, R. (2002). Dynamical conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroscedastic models, *Journal of Business and Economic Statistics* **20**(3): 339–350.

Engle, R. F. and Russell, J. R. (1997). Forecasting the frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model, *Journal of Empirical Finance* **4**(2-3): 187–212.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data, *Econometrica* **66**(5): 1127–1162.

Engle, R. and Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles, *Journal of Business & Economic Statistics* **22**(4): 367–381.

Erawan, S. D. (2015). *Essays on Behavioural Approach in Finance*, PhD thesis, University of St. Gallen.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American statistical Association* **96**(456): 1348–1360.

Fang, L. and Peress, J. (2009). Media coverage and the cross-section of stock returns, *The Journal of Finance* **64**(5): 2023–2052.

Farajtabar, M., Du, N., Rodriguez, M. G., Valera, I., Zha, H. and Song, L. (2014). Shaping social activity by incentivizing users, *Advances in neural information processing systems*, pp. 2474–2482.

Fengler, M. R. and Okhrin, O. (2016). Managing risk with a realized copula parameter, *Computational Statistics & Data Analysis* **100**: 131–152.

Fernandes, M. and Grammig, J. (2006). A family of autoregressive conditional duration models, *Journal of Econometrics* **130**(1): 1–23.

Francq, C. and Zakoian, J.-M. (2011). *GARCH models: structure, statistical inference and financial applications*, John Wiley & Sons.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* **9**(3): 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.

Genest, C., Nešlehová, J. and Ziegel, J. (2011). Inference in multivariate Archimedean copula models, *Test* **20**(2): 223–256.

Genest, C., Nešlehova, J. and Ghorbal, N. B. (2011). Estimators based on Kendall's tau in multivariate copula models, *Australian and New Zealand Journal of Statistics* **53**(2): 157–177.

Górecki, J., Hofert, M. and Holena, M. (2014). On the consistency of an estimator for hierarchical Archimedean copulas, *32nd International Conference on Mathematical Methods in Economics*, pp. 239–244.

Górecki, J., Hofert, M. and Holeňa, M. (2016a). An approach to structure determination and estimation of hierarchical Archimedean Copulas and its application to Bayesian classification, *Journal of Intelligent Information Systems* **46**(1): 21–59.

Górecki, J., Hofert, M. and Holeňa, M. (2016b). On structure, family and parameter estimation of hierarchical Archimedean copulas. ArXiv preprint arXiv:1611.09225.

Grammig, J. and Maurer, K.-O. (2000). Non-monotonic hazard functions and the autoregressive conditional duration model, *The Econometrics Journal* **3**(1): 16–38.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* **37**(3): 424–438.

Granger, C. W. (1980). Testing for causality: a personal viewpoint, *Journal of Economic Dynamics and Control* **2**: 329–352.

Gustafsson, M., Hornquist, M. and Lombardi, A. (2005). Constructing and analyzing a large-scale gene-to-gene regulatory network Lasso-constrained inference and biological validation, *IEEE/ACM Transactions on computational biology and bioinformatics* **2**(3): 254–261.

Hafner, C. M. and Herwartz, H. (2009). Testing for linear vector autoregressive dynamics under multivariate generalized autoregressive heteroskedasticity, *Statistica Neerlandica* **63**(3): 294–323.

Hall, E. C. and Willett, R. M. (2014). Tracking dynamic point processes on networks. ArXiv preprint arXiv:1409.0031.

Hansen, P. R., Lunde, A. and Voev, V. (2014). Realized beta GARCH: a multivariate GARCH model with realized measures of volatility, *Journal of Applied Econometrics* **29**(5): 774–799.

Härdle, W. K., Okhrin, O. and Okhrin, Y. (2013). Dynamic structured copula models, *Statistics & Risk Modeling* **30**(4): 361–388.

Härdle, W. K., Wang, W. and Yu, L. (2016). Tenet: Tail-event driven network risk, *Journal of Econometrics* **192**(2): 499–513.

Hasbrouck, J. (1991). Measuring the information content of stock trades, *The Journal of Finance* **46**(1): 179–207.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Springer.

Hautsch, N. (2004). Modelling irregularly spaced financial data.

Hautsch, N. (2011). *Econometrics of financial high-frequency data*, Springer Science & Business Media.

Hautsch, N., Schaumburg, J. and Schienle, M. (2014). Financial network systemic risk contributions, *Review of Finance* **19**(2): 685–738.

Hayashi, T., Yoshida, N. et al. (2005). On covariance estimation of non-synchronously observed diffusion processes, *Bernoulli* **11**(2): 359–379.

Ho, K.-Y., Shi, Y. and Zhang, Z. (2013). How does news sentiment impact asset volatility? Evidence from long memory and regime-switching approaches, *The North American Journal of Economics and Finance* **26**: 436–456.

Hoeffding, W. (1940). Scale-invariant correlation theory, *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universit at Berlin* **5**(3): 181–233.

Hofert, M. and Scherer, M. (2011). CDO pricing with nested Archimedean copulas, *Quantitative Finance* **11**(5): 775–787.

Hsuan, C. C. Y. (2017). A Continuous-Time Stochastic Volatility Model with Sentiment.

Jacob, L., Obozinski, G. and Vert, J.-P. (2009). Group lasso with overlap and graph lasso, *Proceedings of the 26th annual international conference on machine learning*, ACM, pp. 433–440.

Jaworski, P., Durante, F. and Härdle, W. K. (2013). *Copulae in mathematical and quantitative finance*, Springer.

Jin, X. and Maheu, J. M. (2012). Modeling realized covariances and returns, *Journal of Financial Econometrics* **11**(2): 335–369.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters, *IMS Lecture Notes* **28**: 120–141.

Joe, H. (2014). *Dependence modeling with copulas*, CRC Press.

Jovanović, S., Hertz, J. and Rotter, S. (2015). Cumulants of Hawkes point processes, *Physical Review E* **91**(4): 042802.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems, *Journal of basic Engineering* **82**(1): 35–45.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley.

Kirange, M. D., Deshmukh, R. R., Insaaf, H. M., Morawaka, M., Wanasinghe, W., Fernando, A., Mudugamuwa, M. M. and Dhammearatchi, D. (2016). Sentiment Analysis of News Headlines for Stock Price Prediction.

Kirchner, M. (2016). Hawkes and INAR processes, *Stochastic Processes and their Applications* **126**(8): 2494–2525.

Krämer, N., Brechmann, E. C., Silvestrini, D. and Czado, C. (2013). Total loss estimation using copula-based regression models, *Insurance: Mathematics and Economics* **53**(3): 829–839.

Krupskii, P. and Joe, H. (2013). Factor copula models for multivariate data, *Journal of Multivariate Analysis* **120**: 85–101.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models, *The Journal of Derivatives* **3**(2): 73–84.

Kurowicka, D. (2011). *Dependence modeling: vine copula handbook*, World Scientific.

Kwok, S. S. M., Li, W. K. and Yu, P. L. H. (2009). The autoregressive conditional marked duration model: Statistical inference to market microstructure, *Journal of Data Science* **7**: 189–201.

Lee, C., Shleifer, A. and Thaler, R. H. (1991). Investor sentiment and the closed-end fund puzzle, *The Journal of Finance* **46**(1): 75–109.

Lee, W. Y., Jiang, C. X. and Indro, D. C. (2002). Stock market volatility, excess returns, and the role of investor sentiment, *Journal of Banking & Finance* **26**(12): 2277–2299.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data, *Bioinformatics* **24**(9): 1175–1182.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance* **66**(1): 35–65.

Loughran, T. and McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language, *Journal of Financial Economics* **109**(2): 307–326.

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures, *The Journal of Finance* **69**(4): 1643–1671.

Loughran, T., McDonald, B. and Yun, H. (2009). A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports, *Journal of Business Ethics* **89**(1): 39–49.

Lozano, A. C., Abe, N., Liu, Y. and Rosset, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery, *Bioinformatics* **25**(12): i110–i118.

Lugmayr, A. and Gossen, G. (2013). Evaluation of Methods and Techniques for Language Based Sentiment Analysis for DAX 30 Stock Exchange A First Concept of a "LUGO" Sentiment Indicator, *International SERIES on Information Systems and Management in Creative eMedia* (1): 69–76.

Luss, R. and d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification, *Quantitative Finance* **15**(6): 999–1012.

Narayan, P. K. and Bannigidadmath, D. (2015). Does financial news predict stock returns? New evidence from Islamic and non-Islamic stocks, *Pacific-Basin Finance Journal* **42**: 24–45.

Neal, R. and Wheatley, S. M. (1998). Do measures of investor sentiment predict returns?, *Journal of Financial and Quantitative Analysis* **33**(4): 523–547.

Nelsen, R. B. (1996). Nonparametric measures of multivariate association, *Lecture Notes-Monograph Series* pp. 223–232.

Nelsen, R. B. (2007). *An introduction to copulas*, Springer Science & Business Media.

Nickerson, R. (2011). *Mathematical reasoning: Patterns, problems, conjectures, and proofs*, Taylor & Francis.

Noureldin, D., Shephard, N. and Sheppard, K. (2012). Multivariate high-frequency-based volatility (HEAVY) models, *Journal of Applied Econometrics* **27**(6): 907–933.

Ogata, Y. and Akaike, H. (1982). On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes, *Selected Papers of Hirotugu Akaike*, Springer, pp. 269–274.

Oh, D. H. and Patton, A. J. (2017). Modeling dependence in high dimensions with factor copulas, *Journal of Business & Economic Statistics* **35**(1): 139–154.

Okhrin, O., Okhrin, Y. and Schmid, W. (2013). On the structure and estimation of hierarchical Archimedean copulas, *Journal of Econometrics* **173**(2): 189–204.

Okhrin, O. and Ristig, A. (2014). Hierarchical Archimedean Copulae: The HAC Package, *Journal of Statistical Software* **58**(4).

Okhrin, O., Ristig, A., Sheen, J. R. and Trück, S. (2015). Conditional Systemic Risk with Penalized Copula, *Technical report*, SFB 649 Discussion Paper.

Okhrin, O. and Tetereva, A. (2017). The Realized Hierarchical Archimedean Copula in Risk Modelling, *Econometrics* **5**(2).

Patton, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation, *Journal of Financial Econometrics* **2**(1): 130–168.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models, *Journal of the American Statistical Association* **104**(486): 735–746.

Peress, J. (2014). The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes, *The Journal of Finance* **69**(5): 2007–2043.

Peterson, R. L. (2016). *Trading on Sentiment: The Power of Minds Over Markets*, John Wiley & Sons.

Pooter, M. d., Martens, M. and Dijk, D. v. (2008). Predicting the Daily Covariance Matrix for S&P 100 Stocks Using Intraday Data—But Which Frequency to Use?, *Econometric Reviews* **27**(1-3): 199–229.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, Vol. 1, MIT press Cambridge.

Reilly, M., Posadas-Sanchez, D., Kettle, L. and Killeen, P. (2012). Rats (Rattus norvegicus) and pigeons (Columbia livia) are sensitive to the distanceto food, but only rats request more food when distance increases, *Behavioural Processes* **91**(3): 236–243.

Rodriguez, J. C. (2007). Measuring Financial Contagition: A Copula Approach, *Journal of Empirical Finance* **14**(3).

Salvatierra, I. D. L. and Patton, A. J. (2015). Dynamic copula models and high frequency data, *Journal of Empirical Finance* **30**: 120–135.

Segers, J. and Uyttendaele, N. (2014). Nonparametric estimation of the tree structure of a nested Archimedean copula, *Computational Statistics & Data Analysis* **72**: 190–204.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk, *The Journal of Finance* **19**(3): 425–442.

Shefrin, H. (2007). Behavioral finance: biases, mean–variance returns, and risk premiums, *CFA Institute Conference Proceedings Quarterly*, Vol. 24, CFA Institute, pp. 4–12.

Shiller, R. J. (1999). Results of Surveys about Stock Market Speculation 12/99, http://www.econ.yale.edu/~shiller/data/investor.html. Accessed: 2018-01-20.

Shimamura, T., Imoto, S., Yamaguchi, R. and Miyano, S. (2007). Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data, *Genome Informatics* **19**: 142–153.

Sidorov, S., Date, P., Balash, V. et al. (2013). Using news analytics data in GARCH models, *Applied Econometrics* **29**(1): 82–96.

Sklar, A. (1959). Functions de repartition a n dimensionset leurs marges. Publ. Inst. Statis. Univ. Paris 8.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1): 91–108.

Trivedi, P. K., Zimmer, D. M. et al. (2007). *Copula modeling: an introduction for practitioners*, Vol. 1, Now Publishers, Inc.

Tsay, R. S. (2005). *Analysis of financial time series*, Vol. 543, John Wiley & Sons.

Uyttendaele, N. et al. (2016). On the estimation of nested Archimedean copulas: A theoretical and an experimental comparison.

van der Voort, M. (2007). Factor copulas: External defaults, *The Journal of Derivatives* **14**(3): 94–102.

Wang, Y.-H., Keswani, A. and Taylor, S. J. (2006). The relationships between sentiment, returns and volatility, *International Journal of Forecasting* **22**(1): 109–123.

Wasserman, L. (2006). All of nonparametric statistics.

Xu, H., Farajtabar, M. and Zha, H. (2016). Learning granger causality for hawkes processes, *International Conference on Machine Learning*, pp. 1717–1726.

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., Chu, S. and Yang, X. (2015). On machine learning towards predictive sales pipeline analytics, *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1): 49–67.

Zhang, L., Mykland, P. A. and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data, *Journal of the American Statistical Association* **100**(472): 1394–1411.

Zhang, M. Y., Russell, J. R. and Tsay, R. S. (2001). A nonlinear autoregressive conditional duration model with applications to financial transaction data, *Journal of Econometrics* **104**(1): 179–207.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A. and Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1513–1522.

Zhou, K., Zha, H. and Song, L. (2013a). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes, *Artificial Intelligence and Statistics*, pp. 641–649.

Zhou, K., Zha, H. and Song, L. (2013b). Learning triggering kernels for multi-dimensional hawkes processes, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1301–1309.

Zhou, K., Zha, H. and Song, L. (2013c). Learning triggering kernels for multi-dimensional hawkes processes, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1301–1309.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* **101**(476): 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301–320.

Curriculum vitae

Anastasija Tetereva
Helvetiastrasse 48
9000, St. Gallen
Switzerland

+41 71 224 2183
anastasija.tetereva@unisg.ch

Education

2013 - 2018	PhD in Economics and Finance, University of St. Gallen, St. Gallen, Switzerland
2010 - 2012	Master Degree in Statistics, Humboldt University, Berlin, Germany
2005 - 2010	Diploma in Mathematical Statistics, University of Latvia, Riga, Latvia
2008	Exchange Semester at Technical University of Kaiserslautern, Kaiserslautern, Germany

Work Experience

2013 - 2018	Research Assistant at the Chair of Mathematics and Statistics, University of St. Gallen St. Gallen, Switzerland
2011 - 2012	Consultor at Free University of Berlin, Berlin, Germany
2009 - 2010	Analyst-Statistician at Swedbank, Riga, Latvia
2007 - 2008	Software Engineer at University of Latvia, Riga, Latvia Semiparametric modeling. Comparing the properties of statistical tests.

Teaching experience

Current Feb 2013	Teaching Assistant at University of St. Gallen, St. Gallen, Switzerland Statistics, Microeconomics I and II, Macroeconomics I and II
Current 2014	Lecturer at University of Latvia, Riga, Latvia Teaching Master's level course "Multivariate Statistical Analysis"
2013	Lecturer at University of St. Gallen, St. Gallen, Switzerland
2012 - 2014	Lecturer at Steinbeis-Transfer-Institut, Berlin, Germany Quantitative Methods I, Quantitative Methods II

Conference Talks

2018	SoFiE Conference 2018, Lugano, Switzerland
2017	CFE-CMStatistics, London, UK
2017	Text, Herding and Sentiment conference, Cambridge, UK
2017	European Meeting of Statisticians 2017, Helsinki, Finland (invited session)
2017	SoFiE Conference 2017, New York, USA
2016	CFE-CMStatistics, Seville, Spain (invited session)
2016	Salzburg workshop on dependence models and copulas (invited session)
2016	SoFiE Summer School, Brussels, Belgium

Languages

Mothertongue:	Latvian
Fluent:	English, German, Russian
Basic Knowledge:	French

Computer Skills

Advanced user :	R, Matlab, Stata, Eviews, IBM Statistics, Latex
Intermediate Knowledge:	MySQL, Teradata, Pascal