Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: http://www.tandfonline.com/loi/rquf20

News, volatility and jumps: the case of natural gas futures

Svetlana Borovkova & Diego Mahakena

To cite this article: Svetlana Borovkova & Diego Mahakena (2015) News, volatility and jumps: the case of natural gas futures, Quantitative Finance, 15:7, 1217-1242, DOI: 10.1080/14697688.2014.986513

To link to this article: http://dx.doi.org/10.1080/14697688.2014.986513

Quantitative Finance, 2015
Vol. 15, No. 7, 1217–1242, http://dx.doi.org/10.1080/14697688.2014.986513

Routledge
Taylor & Francis Group

News, volatility and jumps: the case of natural gas futures

SVETLANA BOROVKOVA*† and DIEGO MAHAKENA‡

†Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
‡Flow Traders, Jacob Bontiusplaats 9, 1018 LL Amsterdam, The Netherlands

(*Received 20 May 2014; accepted 30 October 2014*)

We investigate the impact of news sentiment on the price dynamics of natural gas futures. We propose a Local News Sentiment Level model, based on the Local Level model of Durbin and Koopman [*Time Series Analysis by State Space Methods*, 2001], to construct a running series of news sentiment from irregularly observed news items' sentiments. We construct several return and variation measures to proxy for the fine dynamics of the natural gas futures prices. We employ event studies, Granger causality tests and several state-of-the-art volatility models to assess the effect of news on the returns, price jumps and the volatility. We find significant relationships between news sentiment and the dynamic characteristics of natural gas futures prices. Our findings are, among others, that the arrival of news in non-trading periods causes overnight returns, that news sentiment is Granger caused by negative semi-volatility and that news sentiment is more sensitive to negative than to positive jumps. In addition to that we find strong evidence that news sentiment severely Granger causes price jumps and conclude that market participants trade futures as some function of aggregated news. We augment volatility models with news sentiment measures and conduct an out-of-sample volatility forecasting study. The first class of models is the generalized autoregressive conditional heteroskedasticity models the second class is the high-frequency-based volatility models of Shephard and Shephard [*J. Appl. Econ.*, 2010, **25**, 197–231] and Noureldin *et al.* [*J. Appl. Econ.*, 2012, **27**(6), 907–933]. We adapt both models to account for asymmetric volatility, leverage and time to maturity effects. By augmenting all models with news sentiment variables, we find that including news sentiment in volatility models significantly improves volatility forecasts.

Keywords: News sentiment; Natural gas futures; State space modelling; Kalman filter; Realized variance; Bipower variation; Event study; Granger causality; GARCH; HEAVY; Volatility forecasting

JEL Classification: C58, G14, G17, C22, C53

1. Introduction

Over the last decade, the rise of algorithmic trading catalyzed the IT revolution in global financial markets. Algorithmic trading, and more specifically high-frequency trading, has urged the need for more refined data to analyse security price dynamics. Traditionally, the analysis of security price dynamics was concentrated on responses to quantitative or *hard* measures, such as price-derived data, corporate fundamentals, macroeconomic statistics or other variables intended to proxy for qualitative characteristics. Nowadays, powerful computers in combination with sophisticated linguistic analysis techniques allow one to process huge bulks of digitalized text to quantify *soft* qualitative information such as sentiment. As a result, there has been a surge in automated language processing tools, applied to news or social media, which 'read' news or social media content and interpret these in terms of relevance, sentiment, novelty and other characteristics. For instance, business data provider Thomson Reuters recently has introduced the Thomson Reuters News Analytics (TRNA). This engine is based on powerful linguistic analysis techniques and conducts a computerized real-time analysis of millions of news articles to determine whether the news reflect a positive, negative or neutral sentiment and the relevance for a specific security.

Behavioural economics and finance literature observes that sentiment affects the behaviour and decision-making of individuals, see Smith (2003) and Nofsinger (2005). However, the amount of research on news sentiment in relationship to security price dynamics is rather limited and confined to equities. The article of Tetlock (2007) studies the relationship between daily Dow Jones Industrial Average (DJIA) returns and sentiment measured by Harvard IV-4 Psychology Dictionary (HPSD) in the Wall Street Journal. The same author studied the relationship between Standard & Poors 500 (S&P 500) companies and HPSD-based news sentiment,

*Corresponding author. Email: s.a.borovkova@vu.nl

see Tetlock *et al.* (2008). In both articles, he finds that news sentiment helps to predict price dynamics. The more recent study by Bollen *et al.* (2011) employs Granger causality analysis and Self-Organizing Fuzzy Neural Networks to investigate the predictability of DJIA returns by sentiment derived from daily text content of Twitter. They find 86.7% accurate predictions of daily up and down changes in daily DJIA returns. The article by Gross-Klussmann and Hautsch (2011) analyses high-frequency response of stock prices to news and finds that high-positive or negative sentiment has significant effect on returns, volatilities and trading volume.

Commodities have attracted a lot of attention in the past 6–8 years, from investors and other finance professionals. As a response to this growing interest, in 2009, Thomson Reuters extended its news analytics engine to commodities. But despite growing interest in commodities, there is practically no academic literature on news effects in commodity markets, except for Borovkova (2011).

In this paper, we investigate the impact of news sentiment on the dynamics of intraday and daily natural gas futures prices traded on the New York Mercantile Exchange (NYMEX). After crude oil, natural gas is the largest and most liquid commodity market. In contrast to equities, there is a constant stream of news about energy commodities, and in particular, about oil and gas. On average, there are 250 gas-related news items per day, as opposite to equities, where even for the most liquid stocks there are on average 40 news items per week. Moreover, the price dynamics of natural gas futures is very rich, incorporating high (and rapidly changing) volatility and jumps, which are reported by many authors (Deng (1999), Geman and Roncoroni (2006), Deng and Oren (2006), Crosby (2008), Marfe (2010), Mason and Wilmot (2014) and others). The reason for jumps in gas prices is the limited (expensive and technically challenging) storability of gas, which makes it difficult to rapidly respond to spikes in demand. This is not the case for oil, which can be stored very cheaply and efficiently and where the presence of jumps has not been confirmed. So the combination of the rich price dynamics and a large volume of news makes natural gas the ideal asset for our analysis.

Like other (energy) commodities, natural gas is traded in the form of futures contracts with monthly maturities that stretch several years into the future. Therefore, most research based on (energy) commodities investigates the forward curve of futures prices, see for example Borovkova (2004) and Borovkova and Geman (2006). However, Samuelson (1965) states that futures contracts with shorter time-to-maturity should be more volatile. Because of that the first month futures contract is the most liquid and thus news-sensitive contract. Therefore, in this paper, we consider the nearest-to-expiry futures contract.

For the news data, we use the historical database of TRNA for commodities, in particular, for natural gas. This data archive contains news articles with measures of positive, neutral and negative sentiment, which are intended to be interpreted as probabilities that the article conveys a positive, neutral or negative outlook on a specific security price (natural gas in our case). We map the natural gas-tagged TRNA output into an aggregated news sentiment on a 5-min time grid from the beginning of 2006 to the end of 2010.

News arrives at irregular intervals, clustered in time and with possible large gaps between subsequent news items, which represent a challenge for our analysis. Literature on behavioural processes states that animals† virtually discount future pay-offs in the form of exponential or hyperbolic decay, see for example Brannon *et al.* (2001), Nickerson (2009), Uttal (2008) and Reilly *et al.* (2012). Given that economic agents are a kind of 'animals', assumed to be at least as intelligent as the animals used in these studies, we can postulate that economic agents aggregate observed news sentiment as a function of some time decay. This gives a reason to assume that the real unobserved sentiment can be modelled as an autoregressive time series model. Therefore, we utilize the state space modelling framework of Durbin and Koopman (2001) to model the real unobserved autoregressive news sentiment. Specifically, we apply the Local Level (LL) model specification to news sentiment probabilities recorded on the 5-min time grid. We name this the Local News Sentiment Level (LNSL) model. By means of the Kalman filter, we are able to filter out the unobserved states of sentiment series for each day. By doing this, we construct *autocorrelated* news sentiment series, which convey positive, neutral or negative outlook on natural gas prices, based on a 5-min time grid from the beginning of 2006 to the end of 2010.

We construct several return and variation measures based on different time frequencies to proxy for the dynamics of natural gas futures prices. Specifically, we construct daily (squared) returns based on close-to-close (CtC), close-to-open (CtO) and open-to-close (OtC) prices. Additionally, we make several daily-realized measures based on intra-daily data on a 1- and 5-min time grid. Specifically, we construct the realized variance, the more robust realized kernel of Barndorff-Nielsen *et al.* (2008a) and Barndorff-Nielsen *et al.* (2008b) and the jump robust bipower variation of Barndorff-Nielsen and Shephard (2004), to estimate the daily quadratic variance.

The work of Black (1976), Nelson (1991), Engle and Ng (1993) and Glosten *et al.* (1993) indicated the importance of asymmetric returns as a driver of conditional variance, also known as a *leverage effect*. In short, this effect, often observed in stock markets, arises when high-negative returns lead to a bigger increase in volatility than positive returns of the same magnitude. In some commodity markets, the so-called *inverse leverage effect* is observed, when high-positive return lead to a greater increase in volatility than negative returns. So we augment our volatility models to include leverage effect and investigate its sign. We also construct realized semivariance and several jump variation variables to proxy for (asymmetric) jumps, based on the work of Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen *et al.* (2008).

To analyse the impact of news sentiment on daily natural gas futures returns, we employ event studies. Specifically, we set up event studies as described in MacKinlay (1997) and used in Tetlock *et al.* (2008) and Borovkova (2011). We find that the front month natural gas futures price shows a mean reverting effect around days which are characterized by extreme positive sentiment. Also, we find that the price evolution around extreme negative sentiment days shows strong and

†Most of these studies are based on experiments with animals like pigeons, rats and mice.

negative price momentum which precedes the event day and continues long after it. From this, we conclude that there is a significant relationship between news sentiment and the natural gas futures returns.

We employ the Granger causality test of Granger (1969), also used in similar context by Tetlock (2007) and Bollen *et al.* (2011). We construct vector autoregression models, including all news sentiment measures and all constructed price dynamic measures. We investigate whether the news sentiment Granger causes the price dynamic measures and vice versa. We find that the arrival of news in non-trading periods causes overnight returns and volatilities and that news sentiment is Granger caused by volatility (in particular, negative semi-volatility) in the market. Also, we find that the news sentiment is more sensitive to negative than by positive price jumps. Finally, we find strong evidence that news sentiment severely Granger causes jumps. From this, we conclude that market participants trade as some function of aggregated news. More specifically, market participants seem to hard sell or hard buy natural gas futures contracts when news sentiment is high in an absolute sense.

We conduct an out-of-sample volatility forecasting study, where we compare the one-step-ahead forecasting performance of two types of volatility models. The first is the generalized autoregressive conditional heteroskedasticity (GARCH) of Engle (1982) and Bollerslev (1986) and the second is the high-frequency-based volatility (HEAVY) models of Shephard and Sheppard (2010) and Noureldin *et al.* (2012). The main difference between these two classes of models is that GARCH models are driven by squared daily returns and HEAVY models are driven by daily estimates of the quadratic variance such as the realized variance or the realized kernel. Put differently, GARCH models are daily return (low frequency) driven and HEAVY models are intraday (high frequency) driven.

We use GARCH and HEAVY-type models with several error distributions, including models that allow for asymmetric returns based on the work of Glosten *et al.* (1993) (GARCH type) and Barndorff-Nielsen *et al.* (2008) (HEAVY-r type). We extend all volatility models with a time-to-maturity variable to account for the Samuelson effect, as suggested by Borovkova and Geman (2006) and Baillie *et al.* (2007).

The set-up of our forecasting study is similar to the work of Andersen *et al.* (1999), Martens (2002), Hansen and Lunde (2005) and Koopman *et al.* (2005). The essential difference is that we are not interested in the forecasting performance of a particular model. Instead, we are interested in the hypothesis whether including news sentiment in volatility models results in superior volatility forecasts. Specifically, we follow the work of Hansen and Lunde (2005) and Koopman *et al.* (2005) and conduct Superior Predictive Ability tests of Hansen (2005) to test our hypothesis. We find significant evidence that including news sentiment measures in volatility models indeed results in superior volatility forecasts.

This paper is organized as follows. Section 2 describes the construction of the LNSL model. The natural gas futures price data and constructed low- and high-frequency measures are described in section 3. Sections 4 and 5 consider the volatility models used in this study and the volatility forecasting methodology. All results, including the event studies and Granger causality tests, are presented in section 6. Section 7 contains concluding remarks.

2. Modelling news sentiment

2.1. Average news sentiment

We define the news sentiment of a news item X_n as a triple $(s^{pos}_{X,n}, s^{neu}_{X,n}, s^{neg}_{X,n})$ for all $n = 1, 2, \ldots, N$. Here, $s^{pos}_{X,n}$, $s^{neu}_{X,n}$ and $s^{neg}_{X,n}$ represent the probability that news item X_n conveys a positive, neutral or negative outlook on the news item, respectively. Given that the sum of these three probabilities adds up to 1 and $s^{neu}_{X,n} = 1 - s^{pos}_{X,n} - s^{neg}_{X,n}$, the news sentiment of a news item can be seen as a draw from a (time-varying) trinomial distribution with parameters $s^{pos}_{X,n}$, $s^{neg}_{X,n}$ and N.

News items arrive not equally spaced in time. To gauge news sentiment with respect to time, we propose a function f which maps the news sentiment of all the N_d news items X_{n_d} observed at $(d-1, d]$ to d, such that $\sum_{d=1}^{D}\sum_{n_d=1}^{N_d} n_d = N$, which implies $\sum_{d=1}^{T} N_d = N$. This results in an aggregated news item $\bar{X}_d$ given by

$$\bar{X}_d := f(w_d, X, d), \quad \text{for} \quad d = 1, 2, \ldots, D, \tag{1}$$

where w_d is a N_d-dimensional vector of weights and X the N-dimensional vector of news items. In this research, we use a simple weighted average for $f(w_d, X, d)$, such that the aggregated news sentiment $(s^{pos}_{X,n}, s^{neu}_{X,n}, s^{neg}_{X,n})$ is given by

$$\bar{s}^{p}_{d} := N_d^{-1} \sum_{n_d=1}^{N_d} w_{n_d} s^{p}_{X,n_d}, \quad \forall p \in \{pos, neu, neg\}, \tag{2}$$

where $0 \leq N_d$, $N_d \neq N_s$ and $N_d = N_r$ where $d \neq s$, $d \neq r$ $\forall d, s, r \in \{1, 2, \ldots, D\}$. The averaged probabilities $\bar{s}^{p}_{d}$ are normalized by

$$s^{p}_{d} = \frac{\bar{s}^{p}_{d}}{\bar{s}^{pos}_{d} + \bar{s}^{neu}_{d} + \bar{s}^{neg}_{d}}. \tag{3}$$

When nothing is observed at $(d-1, d]$, N_d is equal to zero. In this case, we define $\bar{s}^{p}_{d}$ and thus, s^{p}_{d} as missing.

2.2. Absolute news sentiment

As mentioned earlier, the news sentiment of a news item $\bar{X}_d$ consist three probabilities. $(\bar{s}^{pos}_{d}, \bar{s}^{neu}_{d}, \bar{s}^{neg}_{d})$. By definition, this is a relative measure.† We introduce a new variable which we define as the absolute news sentiment. This variable is given by

$$A\bar{S}_d := \left|\bar{s}^{pos}_{d} - \bar{s}^{neg}_{d}\right| (1 - \bar{s}^{neu}_{d}). \tag{4}$$

The absolute news sentiment variable can be interpreted as a measure of news sentiment regardless of whether the news sentiment is positive or negative. This reduces the information about the news sentiment but it does indicate the news sentiment in an absolute positive or absolute negative sense.

†If for example $\bar{X}_d$ is defined as $(\bar{s}^{pos}_{d} = 0.5, \bar{s}^{neu}_{d} = 0.4, \bar{s}^{neg}_{d} = 0.1)$ and $\bar{X}_e$ as $(\bar{s}^{pos}_{e} = 0.5, \bar{s}^{neu}_{e} = 0.1, \bar{s}^{neg}_{e} = 0.4)$ the values of $\bar{s}^{pos}_{d}$ and $\bar{s}^{pos}_{e}$ are equal in value but are different in interpretation.

2.3. *Local news sentiment level model*

We define S_d^p as the real unobserved probability on news sentiment at time d. Literature on behavioural processes states that animals† virtually discount future pay-offs as a function of their seemingly exponential or hyperbolic decay, see for example Brannon *et al.* (2001), Nickerson (2009), Uttal (2008) and Reilly *et al.* (2012). Given that economic agents are in fact animals and assumed to be at least as intelligent as the animals used in these studies, we can state that economic agents aggregate news sentiment as a function of some time decay. This gives reason to assume that the real probability S_t^p contains information on future probabilities and implies that S_t^p can be modelled as an autoregressive time series model.

Unfortunately, S_d^p is unobserved. As a proxy for S_d^p, we use the observed probability s_d^p as defined in (3). A simple but effective way of modelling an unobserved autoregressive time series is the Local Level (LL) described in Durbin and Koopman (2001).‡ Applied to the observed and unobserved probabilities s_d^p and S_d^p, we define the Local News Sentiment Level (LNSL) model as follows

$$\begin{aligned} S_{d+1}^{*p} &= S_d^{*p} + \eta_d^p \quad \eta_d^p \xrightarrow{d} \mathcal{IID}(0, \sigma_{\eta^p}^2), \\ s_{d+1}^{*p} &= S_d^{*p} + \epsilon_d^p \quad \epsilon_d^p \xrightarrow{d} \mathcal{IID}(0, \sigma_{\epsilon^p}^2), \end{aligned} \tag{5}$$

where s_{d+1}^{*p} and S_{d+1}^{*p} are the logit transformed equivalents of s_d^p and S_d^p.§ The first equation in (5) is the state equation which describes the evolution of the unobserved state of S_d^{*p}. The second equation represents the signal equation. Obviously, the signal equation describes the evolution of the observed s_d^{*p}. Both equations are modelled as random walks, such that the LNSL model is equivalent to the LL model.¶ Modifications and extensions of the LL, and thus the LNSL model, such as an Autoregressive Moving Average (ARMA) model for the state equation, can be found in the work of Durbin and Koopman (2001).

2.3.1. The Kalman filter.

We are interested in the unobserved state of S_d^{*p} for all d. Because the states cannot be observed the LNSL model represents a system of many unknowns. However, by assuming stochastic processes for the evolution of both the unobserved state and the signal, the dimensionality of this system is reduced to the characteristic parameters of these stochastic processes. Because of that these parameters are also called *hyperparameters*.

One way to filter the LNS model in (5) is by means of a Kalman Filter.‖ Applied to the LNSL model it updates our knowledge of the unobserved state when a new observation s_d^{*p} becomes available. That is, it updates the mean $\widetilde{S}_d^{*p} = \mathbb{E}[S_d^{*p}|\mathcal{F}_{d-1}]$ and variance $P_d = \mathbb{V}ar[S_d^{*p}|\mathcal{F}_{d-1}]$, where $\mathcal{F}_{d-1}$ contains $\{s_1^{*p}, s_2^{*p}, \ldots, s_{d-1}^{*p}\}$. The state moments $\widetilde{S}_d^{*p} = \mathbb{E}[S_d^{*p}|\mathcal{F}_{d-1}]$ and $P_d = \mathbb{V}ar[S_d^{*p}|\mathcal{F}_{d-1}]$ can be computed by recursively solving the following equations

$$\begin{aligned} &v_d = s_d^{*p} - S_1^{*p} && F_d = P_d + \sigma_\epsilon^2, \\ &K_d = P_d F_d^{-1}, \\ &\widetilde{S}_{d+1}^{*p} = \widetilde{S}_d^{*p} + K_d v_d && P_{d+1} = P_d (1 - K) + \sigma_\eta^2, \\ &\widetilde{S}_1^{*p} = s_1^{*p} && P_1 = 1e^7, \end{aligned}$$

where v_d is the prediction error, F_d the prediction error variance, K_d the Kalman gain and $d = 1, 3, \ldots, D$. The Kalman filter estimates the state at time d by exponentially weighting previous states. More details and generalizations can be found in Durbin and Koopman (2001).

2.3.2. The Kalman smoother.

The Kalman Smoother considers the estimation of the state S_{d+1}^{*p} conditional on $\mathcal{F}_D$. Here, $\mathcal{F}_D$ contains $\{s_1^{*p}, s_2^{*p}, \ldots, s_D^{*p}\}$. The conditional density of $(S_d^{*p}|\mathcal{F}_D)$ is $\mathcal{N}(\hat{S}_d^{*p}, V_d)$, where $\hat{S}_d^{*p} = \mathbb{E}[S_d^{*p}|\mathcal{F}_D]$ and $V_d = \mathbb{V}ar[S_d^{*p}|\mathcal{F}_D]$. These quantities can be computed by solving the following backward recursion equations

$$\begin{aligned} &r_{d-1} = \tfrac{v_d}{F_d} + L_d r_d && N_{d-1} = F_d^{-1} Z_d + L_d^2 N_d, \\ &L_d = 1 - K_d, \\ &\widehat{S}_d^{*p} = \widetilde{S}_d^{*p} + P_d r_{d-1} && V_d = P_d - P_d^2 N_{d-1}, \end{aligned}$$

where $r_D = 0$ and $d = 1, 2, \ldots, D$. Since the Kalman smoother makes use of both forward (Kalman filter) and backward recursions, it effectively estimates the state at time d by exponentially weighting the states around it. More details and generalizations can be found in Durbin and Koopman (2001).

2.3.3. Missing observations and forecasting.

As mentioned before s_d^p can be missing in case of no observed news. An advantage of applying the Kalman filter and smoother to the LNSL model is that it handles such missing observations with great ease.

In case s_d^p is missing, $\widetilde{S}_{d+1}^{*p}$ can be computed by setting v_d and K_d to zero in the Kalman filter equations. This yields to $\widetilde{S}_{d+1}^{*p} = \widetilde{S}_d^{*p}$ and $P_{d+1} = P_d + \sigma_\eta^2$. Effectively, the expectation of the state at time $d+1$ is equal to the expectation of the state at time d. The state variance grows since the uncertainty about the state in case of a missing observation becomes bigger.

Likewise, $\widehat{S}_{d+1}^{*p}$ can be computed by also setting $L_d = 1$ in case s_d^p is missing. The difference is that the uncertainty about the state might be growing slower since future observations can be in $\mathcal{F}_D$, hence, provide information about future states and hence, provide information about the uncertainty of the state at time d.

The $h-$step ahead forecast $\mathbb{E}[S_{D+h}^{*p}|\mathcal{F}_D]$ can be found by simply handling $d = (D+1, D+2, \ldots, D+h)$ as a missing observations. Notice that, since $r_D = 0$, both $\widehat{S}_{D+h}^{*p}$ and $\widetilde{S}_{D+h}^{*p}$ are equal.

†Most of these studies are based on experiments with animals like pigeons, rats and mice.

‡As mentioned earlier, the news sentiment can also be seen as a time-varying trinomial distribution with parameters S_d^{pos}, s_d^{neg} and N. For such distributions, Durbin and Koopman (2001) propose a non-Gaussian state space model instead of the LL. Therefore, we actually propose a quasi-LL model.

§For $x \in [0, 1]$, the logit transformation of x is given by $x^* = \ln(x) - \ln(1-x)$ such that x^* is a real value. The logit transformations of the probabilities s_d^p and S_d^p are required by the LNSL model because both equations are defined on $\mathbb{R}$.

¶The LNS model is a model representation for the Exponential Weighted Moving Average (EWMA) model, see Durbin and Koopman (2001).

‖The Kalman Filter was first derived by Kalman (1960).

2.4. Parameter estimation

Let Ψ be the vector of parameters representing the unknowns in the LNSL model specified in equation (5). The Kalman filter recursions construct prediction errors v_t and prediction error variances F_d subject to Ψ. We assume that the prediction errors are independent and identically distributed with mean zero and finite variance. If we assume normality for the prediction errors, we can use the Gaussian likelihood function $L(y, \theta)$ for the LNSL model. The Gaussian for the LNSL model log-likelihood is specified as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{s}, \Psi) &:= \log L(y, \theta) \\ &= -\frac{D}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{D}|F_d| - \frac{1}{2}\sum_{d=1}^{D} v_d' F_d^{-1} v_d, \end{aligned} \quad (6)$$

where $\mathbf{s}$ represents the news sentiment data. Maximum likelihood (ML) estimation can be used to estimate Ψ. The ML estimation procedure involves maximization of the log-likelihood function in (6) subject to Ψ. The ML estimate $\hat{\Psi}$ maximizes the log-likelihood function in (6). Since we assume normality for the prediction errors the ML estimate $\hat{\Psi}$ is actually a Quasi maximum likelihood estimate (QML).

2.5. News sentiment data

The news items we use in this research are provided by the Thomson Reuters News Analytics engine (TRNA) for commodities. The data-set contains news items from the beginning of 2006 until the end of 2010, and are time-flagged to the millisecond. Each news item is provided with a positive, neutral and negative sentiment measurement intended to be interpreted as a probability of a positive, neutral and negative outlook on the news item. Since all items are tagged by product name, the probabilities can also be interpreted as a positive, neutral and negative outlook on the commodity price of the tagged product. Additionally, the TRNA output also provides a relevance indicator, item-type and several other variables. The relevance indicator is a measure represented by the probability that the news item is relevant for the tagged product. The item-type variable shows if the news item represents an 'ALERT', 'ARTICLE' or other types of news items.

We are interested in the impact of news sentiment on the price dynamics of natural gas futures prices. Therefore, we filter the TRNA data-set for news items tagged as 'NGS', which are natural gas related. We consider the resulting 310 614 news items as X_n^* for $n = 1, \ldots, 310\,614$. From this data-set, we only consider the 'ALERT' and 'ARTICLE' item types because they convey actual news. Also we remove news items for which the relevance indicator is smaller than 0.3 and items for which $|s_{X,n}^{pos} - s_{X,n}^{neg}| \leq 0.05$. This cleaning procedure resulted in 185 982 news items such that X_n for $n = 1, \ldots, 185\,982$.

We aggregate the cleaned news item data-set on a 5-min time grid by applying equations (2) and (3) to every probability in X_n. From the beginning of 2006 to the end of 2010, this results in 841 536 5-min intervals of which 126 789 are estimated as $\bar{X}_d$.

In table 1, some descriptive statistics are presented for X_n^*, X_n and $\bar{X}_d$. Interesting are the time difference statistics for the retained data-set X_n. Here, we see that new news arrives within 23 min on average and according to a median of less than 6 min, most of the times even faster. This is due to less news item arrivals in weekends and after trading hours. The time distribution of the estimated intervals or aggregated news items $\bar{X}_t$ are equal by definition. The weekday frequencies in figure 1 clearly show the weekday-dependent news item arrival rate. The month frequencies show a small decreasing peak from September to November. This can be related to the beginning of the natural gas season.†

Finally, figure 2 shows the sample autocorrelation functions (SACF) based for the probabilities in $\bar{X}_d$. The SACFs are constructed with 95% confidence intervals based on heteroskedasticity robust standard errors of White (1980). The SACFs show clear autocorrelation for the probabilities. This strengthens the statement that economic agents aggregate news sentiment as a function of some time decay and hence the use of the LNSL model.

3. Natural gas futures returns, volatility and jumps

We investigate the price dynamics of front month natural gas futures contracts traded on New York Mercantile Exchange (NYMEX). Since the first-month-ahead contract is more liquid than futures contracts with higher time-to-maturity, we do not consider multi-month-ahead futures contracts. We used a sample of 415 371 last quotes from 3 January 2006 to 31 December 2010 on a 1-min time-frame, based on 1257 trading days. We performed the data cleaning procedure described in Barndorff-Nielsen *et al.* (2008b): Specifically, we deleted all quotes equal to or smaller than zero, all quotes higher than the median mid-point of the last 10 mid-points, all quotes where either the bid or ask was zero, all quotes where the spread was larger than 10 times the spread on that day. Each of the deleted quotes was replaced by the previous 'reliable' quote. From the cleaned data-set, we constructed a 1-min and a 5-min time grid from 9:00 to 14:30 EST, corresponding to 331 and 67 quotes per trading day. Finding the closest quotes before or equal to each grid point in time resulted in a total of 416 067 and 84 219 quotes for the 1-min and 5-min time grid, respectively.

From the 5-min time grid quotes, we constructed mid-quote prices $P_{t,i}$, where $i = 1, \ldots, 331$ in trading day t. From these mid-quote prices, we constructed 5-min $r_{t,i}$, close-to-close (CtC) $r_{t,CtC}$, close-to-open (CtO) $r_{t,CtO}$ and open-to-close (OtC) $r_{t,OtC}$ continuously compounded returns.‡ Where $r_{t,CtC}$, $r_{t,CtO}$ and $r_{t,OtC}$ are daily return measures denoting the log returns on $(P_{t-1,331}, P_{t,331})$, $(P_{t-1,331}, P_{t,1})$ and $(P_{t,1}, P_{t,331})$, respectively. Obviously, this results in 1257 $r_{t,OtC}$ and only 1,256 $r_{t,CtC}$ and $r_{t,CtO}$ returns.

Following Barndorff-Nielsen *et al.* (2008), we assume the log price process to be represented by a Brownian semimartingale ($\mathcal{BSM}$)

$$Y_t = \int_0^t a_s ds + \int_0^t \sigma_s dW_s, \quad t \geq 0, \quad (7)$$

†The natural gas season starts around September–October and more news items can arrive due to projected supply- and demand-related news for the winter months.

‡A continuously compounded or log return for price P_i is defined as $\log\left(\frac{P_i}{P_{i-1}}\right)$.

Table 1. Descriptive statistics – Raw X_n and 5-minute aggregated $\widehat{X}_t$ news sentiment.

Total dataset	Retained dataset						
	$s^{pos}_{X^*,n}$	$s^{neu}_{X^*,n}$	$s^{neg}_{X^*,n}$		$s^{pos}_{X,n}$	$s^{neu}_{X,n}$	$s^{neg}_{X,n}$
Observations	310614	310614	310614	Observations	185982	185982	185982
Mean	0.402	0.236	0.362	Mean	0.421	0.209	0.370
Median	0.357	0.156	0.320	Median	0.390	0.152	0.324
Std.Dev.	0.245	0.205	0.228	Std.Dev.	0.251	0.164	0.234
Skewness	0.280	1.481	0.617	Skewness	0.171	1.440	0.557
Kurtosis	1.636	4.373	2.157	Kurtosis	1.529	4.421	1.974
Minimum	0.030	0.014	0.038	Minimum	0.030	0.014	0.038
Maximum	0.786	0.880	0.830	Maximum	0.786	0.824	0.830
Time differences				Time differences			
	Minutes	Days			Minutes	Days	
Mean	13.540	0.009		Mean	22.614	0.016	
Median	3.800	0.003		Median	5.967	0.004	
Std.Dev.	59.515	0.041		Std.Dev.	93.909	0.065	
Skewness$^{1/3}$	31.177	2.761		Skewness$^{1/3}$	29.073	2.575	
Kurtosis$^{1/4}$	31.556	5.123		Kurtosis$^{1/4}$	27.908	4.530	
Minimum	0.000	0.000		Minimum	0.000	0.000	
Maximum	4457.367	3.095		Maximum	4865.350	3.379	
Aggregated news sentiment – 5-min							
	$\bar{s}^{pos}_t$	$\bar{s}^{neu}_t$	$\bar{s}^{neg}_t$				
Intervals	841536	841536	841536				
Estimated intervals	126789	126789	126789				
Mean	0.423	0.199	0.377				
Median	0.412	0.152	0.346				
Std.Dev.	0.231	0.146	0.219				
Skewness	0.145	1.533	0.580				
Ex.Kurtosis	−1.305	2.133	−0.840				
Minimum	0.030	0.014	0.038				
Maximum	0.030	0.014	0.038				

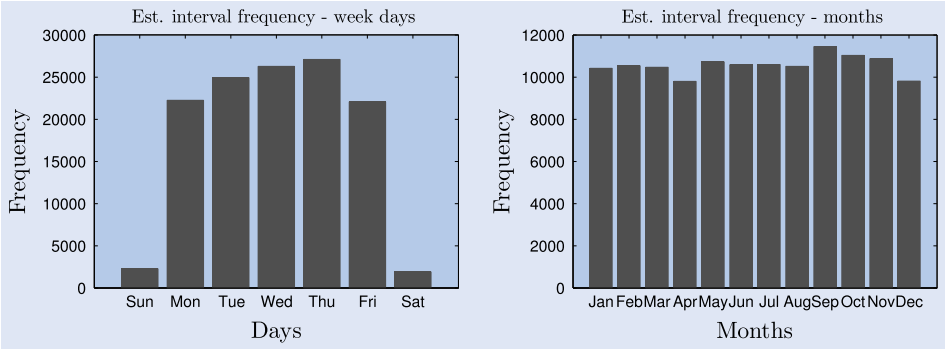

Figure 1. Aggregated interval frequencies.

where a is a locally bounded predictable drift process and σ is a *càdlàg* volatility process, adapted to some common filtration $\mathcal{F}_t$, allowing for leverage effects. The quadratic variance (QV) is given by

$$[Y]_t = \int_0^t \sigma_s^2 ds, \tag{8}$$

and thus,

$$d[Y]_t = \sigma_s^2 dt, \tag{9}$$

tell us everything we can know about the *ex-post* variation of Y. In case of $r_{t,i}$ for $i = 1, \ldots, I$, the squared realized volatility or realized variance estimator is a consistent estimator for QV and is given by

$$RV_t = \sum_{i=1}^{I} r_{t,i}^2. \tag{10}$$

An econometric formalization can be found in Andersen *et al.* (2001) and Barndorff-Nielsen and Shephard (2002). However, as pointed out by Hansen and Lunde (2006), the

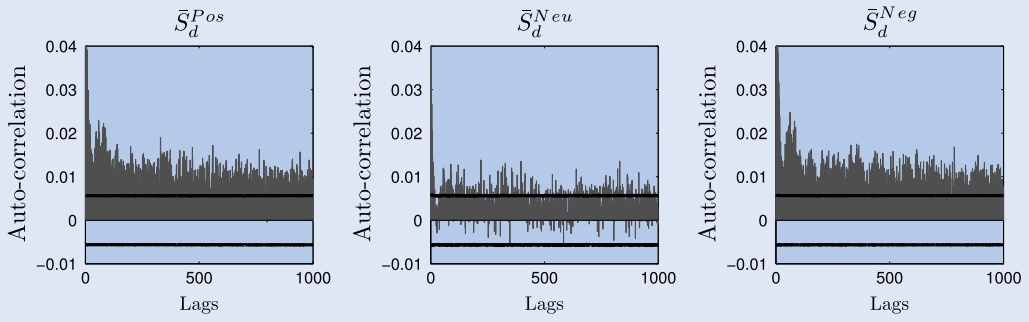

Figure 2. Aggregated interval autocorrelation functions with 95% confidence intervals based on heteroskedasticity robust standard errors of White (1980).

impact of microstructure noise severely influences RV_t. More microstructure noise robust estimators are: pre-averaging of Jacod *et al.* (2009), multiscale of Zhang (2006), the realized kernel of Barndorff-Nielsen *et al.* (2008a) and nearest neighbour truncation of Andersen *et al.* (2012).

We consider the latter two sets of robust estimators, that is, the realized kernel RK_t with a Parzen weight function of Barndorff-Nielsen *et al.* (2008a) and the median realized variance $MEDRV_t$ nearest neighbour truncation estimator of Andersen *et al.* (2012). From these two estimators, we will choose the more robust estimator with respect to our data.

The RK_t estimator is given by

$$RK_t = \sum_{i=-H}^{H} k\left(\frac{h}{H+1}\right)\gamma_h, \quad \gamma_h = \sum_{j=|h|+1}^{I} r_{j,t}r_{j-|h|,t}, \tag{11}$$

where $k(x)$ is the Parzen kernel function

$$k(x) = \begin{cases} 1-6x^2+6x^2 & \text{if } 0 \le x \le 1/2 \\ 2(1-x)^3 & \text{if } 1/2 < x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The $MEDRV_t$ estimator is given by

$$MEDRV_t = \frac{\pi}{6-4\sqrt{3}+\pi}\sum_{i=1}^{I}\left(|r_{t,i}|, |r_{t,i-1}|, |r_{t,i-2}|\right)^2. \tag{12}$$

To consistently estimate the quadratic variance, it is necessary for H to increase with the sample size. Since the degree of microstructure noise, and thus the size of H, can differ for different t, we used the 1-min time grid futures prices to allow for higher H in case of more noisy days. We refer to the work of Barndorff-Nielsen *et al.* (2008a) and Barndorff-Nielsen *et al.* (2008b) for the details of the bandwidth choice of H, since we used the exact same implementation. To be consistent, the $MEDRV_t$ is also based on 1-min time grid futures prices.

Table 2 shows an overview of four sample moment estimations for both RK_t and $MEDRV_t$ over t. Additionally, the same estimations are shown for the logarithmic and normalized logarithmic transformed quantities $\log(RK_t)$, $\log(MEDRV_t)$ and $\overline{\log(RK_t)}$, $\overline{\log(MEDRV_t)}$, respectively.† These transformations map the distributions of both quantities on $\mathbb{R}$ (logarithmic transformation) and more suitable for cross-analysis (normalization). The plots in figure 3 show the same quantities including empirical distribution functions of the normalized logarithmic transformed quantities $\overline{\log(RK_t)}$ and $\overline{\log(MEDRV_t)}$.

Both Table 2 and figure 3 show that RK_t has a higher mean and variance than $MEDRV_t$. However, if we look at the transformed quantities, the difference in mean becomes smaller and the variance even lower. Both the untransformed and transformed $MEDRV_t$ quantities show much more skewness and kurtosis. Therefore, we find that the RK_t is a more robust estimator with respect to our data and hence choose this estimator as our robust estimator.

In case of jumps in the log-price process, the assumption of a $\mathcal{BSM}$ is not sufficient. Now we assume a Brownian semimartingale plus jump process ($\mathcal{BSMJ}$) given by

$$Y_t = \int_0^t a_s ds + \int_0^t \sigma_s dW_s + J_t, \quad t \ge 0, \tag{13}$$

where J is a jump process. If we write jumps in Y as $\Delta Y_t = Y_t - Y_{t-1}$, then

$$[Y]_t = \int_0^t \sigma_s^2 ds + \sum_{s\le t}(\Delta Y_s)^2, \tag{14}$$

and

$$d[Y]_t = \sigma_s^2 dt + \Delta Y_t. \tag{15}$$

The work of Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen and Shephard (2008) introduced the [1, 1]-order Bipower variation process, defined as

$$\{Y\}_t^{[1,1]} = \sum_{i=3}^{[t/\delta]} |Y_{t,i} - Y_{t,i-1}|\,|Y_{t,i-1} - Y_{t,i-2}|, \tag{16}$$

for $\delta \to 0$. They also showed that if Y is a $\mathcal{BSMJ}$, with zero drift and σ independent of W then

$$\{Y\}_t^{[1,1]} = \mu_1^2 \int_0^t \sigma_s^2 ds, \tag{17}$$

where $\mu_1 = \mathbb{E}|u| = \sqrt{2/\pi} \simeq 0.79788$ and $u \xrightarrow{d} \mathcal{N}(0,1)$. Hence, $\mu_1^{-2}\{Y\}_t^{[1,1]} = \int_0^t \sigma_s^2 ds$. They found that this estimator for $\int_0^t \sigma_s^2 ds$ is quite robust to jumps. This implies the following equality

$$[Y]_t - \mu_1^{-2}\{Y\}_t^{[1,1]} = \sum_{s\le t}(\Delta Y_s)^2. \tag{18}$$

Thus, Bipower variation allows us to robustly estimate the jump variance $\sum_{s\le t}(\Delta Y_s)^2$.

†A normalized vector $\overline{x}$ is defined as $\overline{x_i} = \frac{x_i - \mu(x)}{\sigma(x)}$, where $\mu(x)$ and $\sigma(x)$ represent the first and second sample moments of x respectively.

Table 2. Sample moment estimations for realized kernel RK_t and median realized variance $MEDRV_t$ quantities.

	RK_t	$MEDRV_t$	$\log(RK_t)$	$\log(MEDRV_t)$	$\overline{\log(RK_t)}$	$\overline{\log(MEDRV_t)}$
Mean	3,25E−04	2,59E−04	−8,316	−8,580	4,77E−15	7,87E−15
Var	9,91E−08	2,90E−07	0,529	0,525	1,000	1,000
Skewness	4,427	26,254	0,271	0,429	0,271	0,429
Kurtosis	35,243	827,541	3,149	4,211	3,149	4,211

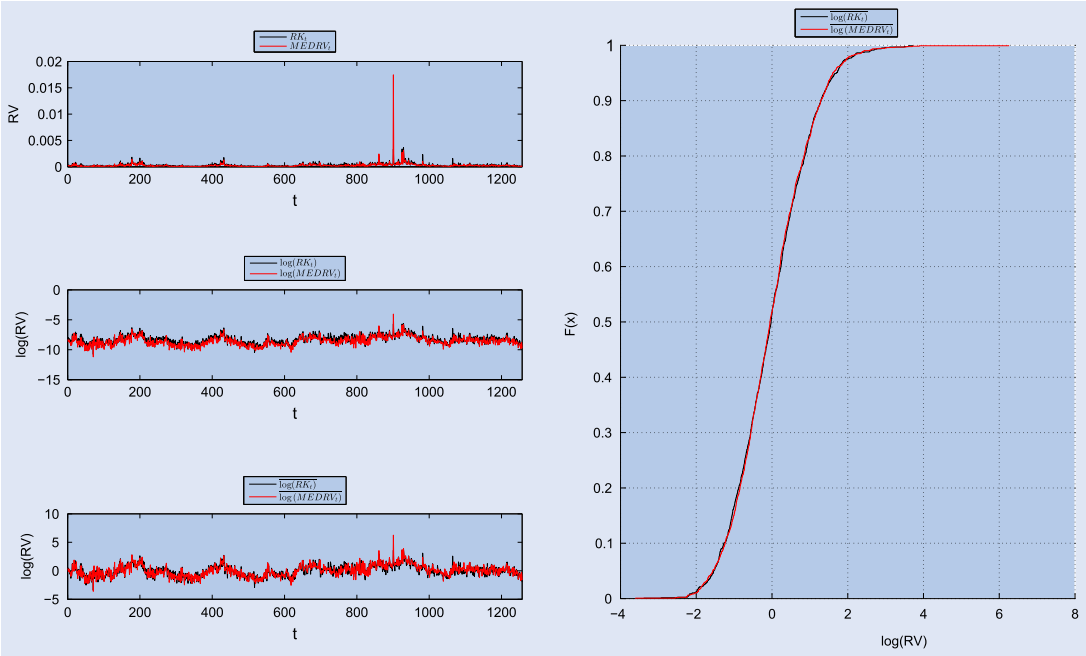

Figure 3. realized kernel RK_t and median realized variance $MEDRV_t$ quantities.

Applied to our 5-minute returns $r_{t,i}$ we estimate the Bipower variation BPV_t process as

$$BPV_t = \mu_1^{-2} \sum_{i=3}^{i \le t} \left| r_{t,i} - r_{t,i-1} \right| \left| r_{t,i-1} - r_{t,i-2} \right|. \qquad (19)$$

For simplicity, we already multiplied the estimator for (16) by μ_1^{-2}.

To test for the presence of jumps in the price process, one can use the procedure introduced by Jacod *et al.* (2009). Many authors have tested and documented the existence of jumps in natural gas prices: see e.g. Deng (1999), Geman and Roncoroni (2006), Deng and Oren (2006), Crosby (2008), Marfe (2010), Mason and Wilmot (2014) and works by many other authors. So we do not repeat such analysis, but instead simply assume that jumps are present in the price series and use the so-called *Bipower jump variation*, introduced by Barndorff-Nielsen, as a tool to isolate the jumps and to estimate their variation.

The Bipower jump variation $BPJV_t$ for $r_{t,i}$ is given by

$$BPJV_t = RV_t - BPV_t, \qquad (20)$$

where RV_t is the simple QV estimator given in (10) and BPV_t the scaled Bipower variation estimator given in (19). Notice that since BPV_t is a jump robust estimator for QV, $BPJV_t$ is asymptotically equal to zero in case of $Y_t \in \mathcal{BSM}$, i.e. when the log-price process does not contain jumps. In the case of non-zero $BPJV_t$, we have $Y_t \in \mathcal{BSMJ}$ implying a jump process in the log-price evolution. Therefore, the $BPJV_t$ is a neat proxy for jump variation and jumps over time. In other words, we do not ask the question 'what is a jump?', but rather we 'define' jumps by the equation (20), so consider as jumps all price movements that are not accounted for by the bipower variation.

The research of Black (1976), Nelson (1991), Engle and Ng (1993) and Glosten *et al.* (1993) indicated the importance of asymmetric returns as a driver of conditional variance, also known as a *leverage effect*. Barndorff-Nielsen *et al.* (2008) introduce the Realized semivariance (RSV) which allows us to estimate the negative and positive part of the QV and hence, the negative and positive part of jump variation. The downside realized semivariance applied to $r_{t,i}$ is estimated by

$$RSV_t^- = \sum_{i=1}^{I} r_{t,i}^2 \mathbf{1}_{r_{t,i} \le 0}, \qquad (21)$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function. Equivalently, the upside realized semivariance is given by

$$RSV_t^+ = \sum_{i=1}^{I} r_{t,i}^2 \mathbf{1}_{r_{t,i} \ge 0}. \qquad (22)$$

Table 3. Summary statistics for daily NG first month trade data based on 1257 trading days from the beginning of 2006 to the end of 2010. The realized kernel RK_t is based on 1-minute price data time grid. All other realized measures RV_t, RSV_t^+, RSV_t^-, BPV_t, $BPJV_t$, $BPUV_t$ and $BPDV_t$ are based on the 5-min price data time grid. The $SACF(l)$ statistic represent the sample autocorrelation function for lag l. Bold $SACF(l)$ statistics are significant at a 5% significance level based on heteroskedasticity robust standard errors described in White (1980). PP + drift statistic represents the Phillips Perron unit-root test statistic where a drift term is assumed in the model, see Phillips and Perron (1988) for details. ADF(l) + drift is the Augmented Dickey Fuller test statistic where up to l lags and a drift term are assumed in the model, see Said *et al.* (1984) for details. Bold (and italic) Phillips Perron and Augmented Dickey Fuller test statistics represent rejection of an unit-root based on a 5% (1%) significance level. The lowertriangle matrix represents the correlation matrix.

	$r_{CtC,t}$	$r_{CtO,t}$	$r_{OtC,t}$	$r^2_{CtC,t}$	$r^2_{CtO,t}$	$r^2_{OtC,t}$	RV_t	RK_t	RSV_t^-	RSV_t^+	BPV_t	$BPJV_t$	$BPDV_t$	$BPUV_t$
mean $1e^3$	−0.71	0.65	−1.36	1.23	0.93	0.32	0.35	0.32	0.18	0.17	0.28	0.06	0.04	0.03
median $1e^3$	0.03	−0.33	−1.15	0.44	0.28	0.13	0.24	0.24	0.12	0.11	0.20	0.02	0.01	0.01
std.dev. $1e^3$	35.05	30.44	17.93	2.94	3.10	0.55	0.36	0.31	0.21	0.21	0.28	0.18	0.16	0.14
skewness	0.63	1.15	−0.06	12.31	15.99	3.86	4.08	4.42	5.40	5.18	4.46	7.95	9.42	7.51
ex.kurtosis	3.79	9.11	0.85	220.61	323.53	21.00	25.33	32.21	51.79	36.97	34.60	89.56	149.56	83.32
SACF(1)	−0.02	−0.05	0.02	**0.02**	0.02	**0.13**	**0.57**	**0.60**	**0.37**	**0.45**	**0.64**	0.03	0.01	0.06
SACF(5)	−0.03	−0.04	0.01	**0.08**	**0.04**	**0.12**	**0.48**	**0.46**	**0.29**	**0.31**	**0.51**	**0.32**	0.03	0.04
SACF(10)	−0.03	−0.02	0.03	**0.08**	0.04	**0.09**	**0.41**	**0.39**	**0.35**	**0.21**	**0.44**	**0.21**	**0.11**	−0.02
SACF(20)	0.01	0.05	0.02	0.05	0.05	0.05	**0.27**	**0.27**	**0.20**	**0.17**	**0.32**	0.06	0.02	0.02
PP + drift	***−36.09***	***−37.23***	***−34.76***	***−34.57***	***−34.59***	***−31.04***	***−18.59***	***−17.72***	***−23.89***	***−21.81***	***−16.65***	***−34.29***	***−34.91***	***−33.21***
ADF(1) + drift	***−24.75***	***−25.21***	***−23.87***	***−24.10***	***−24.28***	***−20.94***	***−13.25***	***−13.49***	***−16.39***	***−14.68***	***−12.75***	***−22.36***	***−24.22***	***−21.53***
ADF(12) + drift	***−8.80***	***−9.06***	***−9.32***	***−6.26***	***−6.91***	***−6.26***	***−4.90***	***−4.70***	***−4.60***	***−5.72***	***−4.69***	***−7.27***	***−8.40***	***−9.36***
ADF(24) + drift	***−7.67***	***−6.57***	***−6.37***	***−4.26***	***−4.49***	**−4.01**	**−3.31**	**−3.29**	**−3.36**	**−3.83**	**−3.42**	***−4.55***	***−5.90***	***−5.76***
$r_{CtC,t}$	–													
$r_{CtO,t}$	0.86	–												
$r_{OtC,t}$	0.50	−0.02	–											
$r^2_{CtC,t}$	0.24	0.31	−0.05	–										
$r^2_{CtO,t}$	0.30	0.36	−0.03	0.92	–									
$r^2_{OtC,t}$	−0.08	−0.02	−0.13	0.16	0.04	–								
RV_t	0.05	0.05	0.02	0.22	0.16	0.56	–							
RK_t	0.07	0.05	0.06	0.25	0.20	0.50	0.90	–						
RSV_t^-	−0.14	0.05	−0.36	0.20	0.13	0.56	0.84	0.72	–					
RSV_t^+	0.22	0.02	0.40	0.17	0.14	0.37	0.84	0.80	0.41	–				
BPV_t	0.09	0.05	0.08	0.24	0.20	0.44	0.86	0.93	0.68	0.77	–			
$BPJV_t$	−0.04	0.01	−0.09	0.08	0.01	0.42	0.64	0.36	0.61	0.47	0.16	–		
$BPDV_t$	−0.27	0.02	−0.57	0.07	0.00	0.38	0.37	0.16	0.76	−0.13	0.03	0.68	–	
$BPUV_t$	0.26	−0.02	0.53	0.03	0.02	0.14	0.43	0.30	−0.05	0.77	0.18	0.56	−0.22	–

This implies

$$RV = RSV^- + RSV^+. \tag{23}$$

Barndorff-Nielsen *et al.* (2008) show that in probability the RSV^- and RSV^+ are both half the RV. Following this result allows us to estimate negative and positive equivalents of $BPJV_t$, see (20). Specifically, the Bipower downward variance $BPDV_t$ is given by

$$BPDV_t \cong RSV_t^- - \frac{1}{2}BPV_t. \tag{24}$$

Equivalently, the Bipower upward variance $BPUV_t$ is given by

$$BPUV_t \cong RSV_t^+ - \frac{1}{2}BPV_t. \tag{25}$$

These variables are proxies for negative and positive jump variation. Also the $BPDV_t$ and $BPUV_t$ can be negative for small samples, since 24 and 25 hold asymptotically, see Barndorff-Nielsen *et al.* (2008).

For our data-set, table 3 reports the basic summary statistics. It is interesting to note that the realized measure estimates RK_t and RV_t are much smaller than the variance of squared daily returns $r^2_{t,CtC}$. Notice that $r_{t,CtC}$ is the sum of $r_{t,CtO}$ and $r_{t,OtC}$. Given that $r_{t,OtC}$ is constructed by the first and last price of day t used by the realized measures, it is not strange that $r^2_{t,OtC}$ is of the same order as RK_t and RV_t. This implies that the more noisy† $r^2_{t,CtO}$ severely influences the size of $r^2_{t,CtC}$. The summary statistics clearly show strong autocorrelation for most realized measures. The squared returns show much less-significant autocorrelation. The correlation between the realized measures are high. Interesting are the correlations between the jump variation measures and the returns. This implies evidence of leverage effects. Also, RK_t shows a higher correlation with the jump robust BPV_t then RV_t. This implies that RK_t is more robust to jumps than RV_t and hence superior representation of *ex-post* variance.

†The CtO measures are more noisy because markets are not open 24 h a day and seven days a week. Additionally, extra noise is related to roll-effects, since the first month futures contract is rolled over into the second month contract at the end of every month.

4. Volatility models

In this section, we discuss modelling of time-varying volatility. Specifically, we discuss extensions of two classes of so-called historical volatility models. The first is the generalized autoregressive conditional heteroskedasticity (GARCH) of Engle (1982) and Bollerslev (1986) and the second the high-frequency-based (HEAVY) volatility models of Shephard and Sheppard (2010) and Noureldin *et al.* (2012). The main difference between both models is that GARCH models are squared daily return driven and HEAVY models are driven by daily estimates of the quadratic variance such as the realized variance or the realized kernel. Put differently, GARCH models are daily return (low frequency) driven and HEAVY models are intraday (high frequency) driven.

For both model classes, we assume the following model for the daily returns

$$r_{t,CtC} = e^{\frac{1}{2}h_t}\epsilon_t, \quad t = 1, \ldots, T, \tag{26}$$

where ϵ_t is either Normal (0, 1) or Student-t with parameter ν or Hansen Skewed-t with parameters (λ, ν).† Here, λ is related to the skewness and ν to the degrees of freedom, where we require $\nu > 2$. The Student-t and Hansen Skewed-t distributions allow for fatter tails and, in the case of the Hansen Skewed-t distribution, positive or negative skewness, which are both very common in commodity return distributions.

The work of Borovkova and Geman (2006) implies a strong relationship between time-to-maturity TM and the variance of natural gas futures returns. Specifically, the returns of futures contract with shorter time-to-maturity showed a higher variance. This time-to-maturity effect can be related to the Samuelson hypothesis which states that futures prices should exhibit higher volatility for shorter time-to-maturity, see Samuelson (1965). Baillie *et al.* (2007) acknowledge the same relationship with other commodity futures returns and estimate time-to-maturity augmented volatility models. For that reason, we augment all our volatility models with a time-to-maturity variable TM_t to adjust for the time-to-maturity effect. Here, TM_t represents the number of days left at time t until the futures contract matures.

4.1. GARCH models

The GARCH(P,Q) model for (36) is given by

$$\begin{aligned}\mathbb{V}ar[r_{CtC,t+1}|\mathcal{F}_t] &= \exp(h_{t+1}) \\ &= \exp\left(\alpha_0 + \sum_{p=1}^{P}\alpha_p r^2_{CtC,t-p+1} + \sum_{q=1}^{Q}\beta_q h_t\right)\end{aligned} \tag{27}$$

where $\beta_q \in (0, 1)$ for all q and $\mathcal{F}_t$ contains the set of returns up to time t. In this research, we only consider $Q = 1$ and $P = 1$. Including the time-to-maturity term TM_t, we define our GARCH model as

$$h_{t+1} = \alpha_0 + \alpha_1 r^2_{CtC,t} + \beta h_t + \tau TM_t. \tag{28}$$

For the GARCH model, we assume ϵ_t in (36) to be normally distributed. We name the GARCH model with Student-t and Skewed-t distributions for ϵ_t as GARCH-t and GARCH-skewt, respectively. Notice that the model in (34) allows to forecast the volatility for time $t + 1$ at time t. For all GARCH models, we use $h_1 = (\alpha_0 + \hat{\sigma}^2(r^2_{CtC,t}))/1 - \beta$ as initial value, where $\hat{\sigma}^2(r^2_{CtC,t})$ is the sample variance of $r^2_{CtC,t}$.

4.2. HEAVY models

Given (36), the HEAVY model of Shephard and Sheppard (2010) is defined as a system of two equations

$$\begin{aligned}\mathbb{V}ar[r_{CtC,t+1}|\mathcal{F}_t^{HF}] &= \exp(h_{t+1}) = \exp(\alpha_{h,0} \\ &\quad + \alpha_{h,1} RM_t + \beta_h h_t)\end{aligned} \tag{29}$$

$$\mathbb{E}[RM_{t+1}|\mathcal{F}_t^{HF}] = \mu_{t+1} = \alpha_{\mu,0} + \alpha_{\mu,1} r_t^2 + \beta_\mu h_t \tag{30}$$

where $\beta_h, \beta_\mu \in [0, 1)$ and $\mathcal{F}_t^{HF}$ contains the set of high-frequency returns up to time t. The second equation can be used to estimate h-step-ahead forecasts $\mathbb{V}ar[r_{t+h}|\mathcal{F}_t^{HF}]$ for $h > 1$. In this research, we are only interested in 1-step-ahead forecasts. Therefore, we only consider the first equation which Shephard and Sheppard (2010) define as the HEAVY-r model. For RM_t, we use the more robust realized kernel RK_t. Including the time-to-maturity term T we define the HEAVY-r model as

$$h_{t+1} = \alpha_{h,0} + \alpha_{h,1} RK_t + \beta_h h_t + \tau TM_t, \tag{31}$$

under the restriction $\beta_h \in (0, 1)$. For the HEAVY-r model, we assume ϵ_t in (36) to be normal distributed. We name the HEAVY-r model with Student-t and Skewed-t distributions for ϵ_t as HEAVY-rt and HEAVY-rskewt, respectively. For all HEAVY-r models, we use $h_1 = (\alpha_{h,0} + \bar{RK})/1 - \beta_h$ as initial value, where $\bar{RK}$ is the sample mean of RK_t. Notice that the (distribution homogeneous) HEAVY-r models are equal to the GARCH models if $r^2_{t,CtC}$ is substituted for RK_t. Comparable to the GARCH model, notice that the model allows to compute the expected volatility for time $t + 1$ at time t.

4.3. Leverage effects

As noted in the Introduction, literature suggests the importance of asymmetric returns as a driver of conditional variance also known as a *leverage effect*, see Black (1976), Nelson (1991), Engle and Ng (1993) and Glosten *et al.* (1993). Leverage effect amounts to an asymmetric response of the conditional variance to negative and positive returns. In stock markets, volatility increases more following large and negative returns than after positive returns of the same magnitude. In commodity markets, an *inverse leverage* effect has been documented in the past, when large positive returns increase volatility more than negative returns. Such an inverse leverage effect is typical of agricultural commodities, as reported in Swaray (2002), Du and Hayes (2009), Stigler (2011), Tothova (2011), Moratoya and Reginaldo (2013) and others. For agricultural commodities, rising prices are still perceived as 'bad' for the economy and mankind. Moreover, rising prices are the sign of depleting inventories, which also leads to increased volatility. For metals and energy commodities, in particular, for crude oil, there has been a significant body of research in this direction (see e.g. Tully and Lucey 2007, Carpantier 2010,

†Details on the Hansen Skewed-t distribution can be found in Hansen (1994).

Hassan 2011, Baur 2012, Salisu and Fasanya 2013), but the evidence is mixed. Some (especially earlier) studies have reported the existence of inverse leverage effect (NG and Pirrong 1994, Giamouridis and Tamvakis 2001). However, in the past few years, this effect seems to have disappeared (at least in oil markets), giving way to the regular leverage effect, as observed by e.g. Cheong (2009), Kristoufek (2014) and in many recent discussions in the professional literature. This is the consequence of the fact that financial players in commodity markets (such as hedge funds and investment funds) have large long positions in crude oil and other energy commodity futures, which makes them behave more like stocks than as typical consumption commodities. This also the case for other commodities traded nowadays primarily for investment purposes, such as copper, but not the case for agricultural commodities, where traditional players are still in majority.

To include a possible leverage (or inverse leverage) effect into our models, we extend our GARCH model in the spirit of Glosten *et al.* (1993). Doing so, (34) becomes

$$h_{t+1} = \alpha_0 + \alpha_1 r^2_{CtC,t} + \beta h_t + \tau TM_t + \gamma r^2_{CtC,t} \mathbf{1}_{(r_{CtC,t}<0)}, \quad (32)$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function. We name this model the GJRGARCH† model. The sign of the estimated coefficient γ will tell us whether the regular or inverse leverage effect is observed for natural gas futures in the period 2003–2010. Analogously to the GARCH case, for the GJRGARCH model we assume ϵ_t in (36) to be normally distributed. We name the GJRGARCH model with Student-t and Skewed-t distributions for ϵ_t as GJRGARCH-t and GJRGARCH-skewt, respectively.

The class of HEAVY models can be extended in the same way as the GARCH models concerning the leverage effect. Shephard and Sheppard (2010) suggest including a realized semivariance measure to (31). Barndorff-Nielsen *et al.* (2008) found log-likelihood improvements by also including the Bipower variation estimate BPV_t. Therefore, we propose to extend (31) with the Bipower downward variation estimate $BPDV_t$ which includes both the realized semivariance estimate RSV_t^- and the Bipower variation estimate BPV_t. The model is given by

$$h_{t+1} = \alpha_{h,0} + \alpha_{h,1} RK_t + \beta_h h_t + \tau TM_t + \gamma_h BPDV_t, \quad (33)$$

We name this the LHEAVY-r model. Analogously to the HEAVY case, for the LHEAVY-r model we assume ϵ_t in (36) to be normal distributed. We name the LHEAVY-r model with Student-t and Skewed-t distributions for ϵ_t as LHEAVY-rt and LHEAVY-rskewt, respectively.

4.4. News sentiment augmented volatility models

We are interested in the effect of news sentiment on the conditional variance $\mathbb{V}ar[r_{t+1,CtC}|\mathcal{F}_t^{HF}]$. For simplicity reasons and to reduce the number of parameters, we augment both the GARCH and HEAVY-r model classes with only the Kalman filtered absolute sentiment variable $\widetilde{AS}_{t|t}$.‡ In the spirit of naming cross-sectional variable augmented GARCH models, see Engle (2002), we name the news sentiment augmented model a GARCHX model.§ We define it as follows

$$h_{t+1} = \alpha_0 + \alpha_1 r^2_{CtC,t} + \beta h_t + \tau TM_t + \phi \widetilde{AS}_{t|t}. \quad (34)$$

Analogously the HEAVYX-r is the news sentiment augmented HEAVY equivalent of the GARCH model. This model is defined as

$$h_{t+1} = \alpha_{h,0} + \alpha_{h,1} RK_t + \beta_h h_t + \tau TM_t + \phi \widetilde{AS}_{t|t}. \quad (35)$$

For both the GARCHX and the HEAVYX-r models, we assume ϵ_t in (36) to be normally distributed. The GARCHX and HEAVYX-r models with Student-t and Skewed-t distributions for ϵ_t are denoted as GARCHX-t, HEAVYX-rt, GARCHX-skewt and HEAVYX-rskewt, respectively. In the same way, the GJRGARCHX and LHEAVYX models represent the news sentiment augmented equivalents of the GJRGARCH (32) and LHEAVY (33) models. Obviously, the GJRGARCHX-t, LHEAVYX-t models assume a Student-t distribution for ϵ_t and GJRGARCHX-skewt, LHEAVYX-skewt a Skewed-t distribution. Based on the models defined in this section, the functional forms of these models are trivial to derive. For that reason, we do not write them down explicitly.

5. Volatility forecasting methodology

In this section, we present the methodology concerning an out-of-sample volatility forecasting study in which we compare the one-step-ahead forecasting performance of the volatility models described in section 4. The set-up of our forecasting study is similar to the work of Andersen *et al.* (1999), Martens (2002), Hansen and Lunde (2005) and Koopman *et al.* (2005). The essential difference is that we are not interested in the forecasting performance of a particular model. Instead, we are interested in the hypothesis if including news sentiment to volatility models results in superior volatility forecasts.

As mentioned in section 4, GARCH models are driven by low frequent daily returns and HEAVY-r models driven by high frequent intraday returns. For that reason, all $K = 24$ volatility models can be divided into two groups. The first group considers the class of GARCH models and the second the class of HEAVY-r models. Additionally, both the GARCH and HEAVY-r volatility model classes can be divided into subgroups of base and news sentiment augmented models. For example, in the case of GARCH models, we have GARCH, GARCH-t, GARCH-skewt et cetera in the base model group and GARCHX, GARCHX-t, GARCHX-skewt et cetera in the news sentiment augmented model group. All the GARCH-type models are denoted as $\mathcal{M}_g$ and all HEAVY-r models are denoted as $\mathcal{M}_h$ for $g = 1, \ldots, G$ and $h = 1, \ldots, H$ where obviously $G = H = 12$ and $G + H = K$.

We estimate all $K = 24$ volatility models 1007 times based on 1007 samples of 250 daily observations, where the first sample is based on an estimation window which starts at 2 January 2006 and ends at 2 January 2007. A one-step-ahead

†The abbreviation GJR stands for Glosten, Jagannathan and Runkle, the authors of Glosten *et al.* (1993).

‡See sections 2 and 6.1 for the exact definition of $\widetilde{AS}_{t|t}$.

§Notice that the inclusion of the (cross-sectional) time-to-maturity term TM_t in the standard GARCH model already is a GARCHX model. But since the inclusion of TM_t is essential when working with commodity futures prices, we do not name our GARCH models GARCHX models.

volatility forecast $\hat{\sigma}_k^2$ is computed for 3 January 2007 based on the estimation window for each model $\mathcal{M}_k$ where $k \in K$. By rolling the estimation window forward by one trading day, we have a second sample of the same size which starts at 3 January 2006 and ends at 3 January 2007. Again, a one-step-ahead volatility forecast $\hat{\sigma}_k^2$ is computed for 4 January 2007 based on the estimation window for each model $k \in K$. More specifically, we estimate $M = 1007$ one-day-ahead forecasts $\hat{\sigma}_{k,m}^2$ where $m \in M$, such that

$$\hat{\sigma}_{k,m}^2 = \mathbb{E}\left[\sigma_m^2 | \mathcal{F}_{m-250,m-1}, \hat{\Psi}\right], \tag{36}$$

where $\mathcal{F}_{m-250,m-1}$ contains all information on interval $[m-250, m-1]$ and $\hat{\Psi}$ is the maximum likelihood estimate of the parameter vector Ψ.

The volatility σ_m^2 is not observable. In section 4 was shown that realized volatility RV_m is a consistent estimator for the latent σ_m^2. However, we will make use of the realized kernel RK_m since it is a more robust estimator of *ex-post* variation. Notice that the latent σ_m^2 represents the variation of close-to-close returns $R_{CtC,t}$ and the realized kernel RK_m is a measure based on open-to-close returns. For that reason we, in some way, have to add the much more noisy variation of close-to-open (overnight) returns $R_{CtO,t}$ to RK_m. Martens (2002), Koopman *et al.* (2005) and Hansen and Lunde (2005) propose similar scaling methods. We follow the method of Hansen and Lunde (2005) who introduce $\tilde{\sigma}_m^2$ as an estimator for σ_m^2. Substituting RK_m for RV_m in their estimator, $\tilde{\sigma}_m^2$ is defined as

$$\tilde{\sigma}_m^2 \equiv \hat{c} RK_m, \quad \text{where} \quad \hat{c} \equiv \left(\frac{T^{-1} \sum_{t=1}^{T} \left(r_{CtC,t} - \hat{\mu}\right)^2}{T^{-1} \sum_{t=1}^{T} RK_t} \right), \tag{37}$$

and where $\hat{\mu} = T^{-1} \sum_{t=1}^{T} r_{CtC,t}$. As mentioned earlier, the less noisy $r_{CtC,t}$ contains the noisy overnight return $r_{CtO,t}$. Hence, by scaling the realized kernel by the variance of $r_{CtC,t}$ we implicitly scale it by the variance of $r_{CtO,t}$ and hence $\tilde{\sigma}_m^2$ is an approximately unbiased estimator for σ_m^2.

Of interest are the volatility forecasts $\hat{\sigma}_{k,m}^2$ for all $k \in K$ models and $\tilde{\sigma}_m^2$. One way to evaluate out-of-sample volatility forecast is in terms of R^2 from a Mincer-Zarnowitz (MZ)-type regression, $\tilde{\sigma}_m^2 = \gamma_0 + \gamma_1 \hat{\sigma}_{k,m}^2 + u_t$. Or the more robust logarithmic version, $\log(\tilde{\sigma}_m^2) = \gamma_0 + \gamma_1 \log(\hat{\sigma}_{k,m}^2) + u_t$ as noted by Pagan and Schwert (1990) and Engle and Patton (2001). However, Hansen and Lunde (2005) note that the R^2 of MZ regressions is not an ideal criterion for comparing volatility models because biased forecasts are not penalized.

Bollerslev *et al.* (1994), Diebold and Lopez (1996) and others, suggest the use of loss functions to determine whether, say, model $\mathcal{M}_k$ outperforms $\mathcal{M}_l$ in forecasting $\tilde{\sigma}_m^2$, where $k \neq l$ and $k, l \in K$. We adopt the same choice of loss functions as considered by Koopman *et al.* (2005) and are given in table 4. Equal and similar loss functions are also considered by Bollerslev *et al.* (1994), Andersen *et al.* (1999), Martens (2002) and Hansen and Lunde (2005).

Following the work of Koopman *et al.* (2005) and Hansen and Lunde (2005), we adopt the robust superior predictive ability (SPA) test of Hansen (2005) to investigate the relative performance of the proposed volatility models in terms of the proposed loss functions.

As mentioned earlier, we split the volatility models into two model groups and perform the SPA test procedure for each group separately. For each group, we have $N+1$ different models $\mathcal{M}_n$ for $n = 0, 1, \ldots, N$.† For each model $\mathcal{M}_n$, we have M volatility forecasts $\hat{\sigma}_{n,m}^2$ for $m = 1, \ldots, M$. For every forecast, we calculate the loss function $L_{i,m,n}$ as given in table 4 for $i = 1, \ldots, 4$. A particular model $\mathcal{M}_0$ is taken as the benchmark model. The loss function of some model $\mathcal{M}_{n \neq 0}$ relative to the benchmark model is defined as

$$X_{n,m} \equiv L_{i,m,0} - L_{i,m,n}, \tag{38}$$

for some loss function i. For $\lambda_n \equiv \mathbb{E}[X_{n,m}]$, model $\mathcal{M}_0$ outperforms all other models, we have $\lambda_n < 0$ for all models $n \neq 0$. Hence, the base model is not outperformed when it accepts the null hypothesis

$$\overset{max}{n \neq 0} \lambda_n \leq 0. \tag{39}$$

Hansen (2005) proposed the associated SPA test statistic

$$T = \overset{max}{n \neq 0} \frac{\sqrt{M}\overline{X}_n}{\hat{\omega}_{n,n}}, \tag{40}$$

where $\hat{\omega}_{n,n}^2$ is a consistent estimate of $\omega_{n,n}^2$, and where $\overline{X}_n = M^{-1} \sum_{m=1}^{M} X_{n,m}$ and $\omega_{n,n}^2 \xrightarrow{a} \mathbb{V}ar[\sqrt{M}\overline{X}_n]$. A consistent estimator of $\omega_{n,n}^2$ and the so-called Hansen consistent p-value of the SPA test statistic T can be found via a bootstrap procedure. Specifically, we apply the stationary bootstrap procedure of Politis and Romano (1994). The procedure consists of constructing new samples for $X_{m,k}$ of length B by concatenation of randomly chosen subsamples of different lengths. The length of the subsamples are independent and are drawn from a geometric distribution with mean q. The subsample lengths are ideally small but sufficiently large to reflect the serial correlation in $X_{m,k}$. After an extensive inspection of the autocorrelation in $X_{m,k}$, we choose a subsample length of five trading days which corresponds to $q = 1/5$. Since we perform the same bootstrap procedure as proposed by Hansen (2005) and performed by Hansen and Lunde (2005) and Koopman *et al.* (2005), we refer to their work for the construction of the Hansen consistent p-value and other details about the SPA test.

6. Results

6.1. *News sentiment index estimates*

In section 2, we introduced the Local News Sentiment Level (LNSL) model. Specifically, the LNSL model is the Local Level model of Durbin and Koopman (2001) applied to the logit transformed probabilities $(\bar{s}_d^{pos}, \bar{s}_d^{neu}, \bar{s}_d^{neg})$ of news item $\bar{X}_d$ for all $d = 1, \ldots, D$. We estimated the LNSL model by means of a quasi maximum likelihood procedure which involves maximization of the log-likelihood described in section 2.‡

†Notice that $N+1 = G$ and $N+1 = H$ in case of the GARCH and HEAVY-r model-type group, respectively.

‡The procedure is implemented in the programme environment Ox of Doornik (2001) and makes use of the SsfPack 2.2 of Koopman *et al.* (1999).

Table 4. Loss function $L_{i,k,m}$ represents the loss of forecast $\hat{\sigma}^2_{k,m}$ of model $\mathcal{M}_k$ based on loss type i for $i \in \{1, \ldots, 4\}$.

Squared error	$L_{1,k,m} = (\tilde{\sigma}^2_{k,m} - \hat{\sigma}^2_{k,m})^2$
Absolute error	$L_{2,k,m} = \lvert\tilde{\sigma}^2_{k,m} - \hat{\sigma}^2_{k,m}\rvert$
Heteroskedasticity adjusted squared error	$L_{3,k,m} = (1 - \tilde{\sigma}^{-2}_{k,m}\hat{\sigma}^2_{k,m})^2$
Heteroskedasticity adjusted absolute error	$L_{4,k,m} = \lvert 1 - \tilde{\sigma}^{-2}_{k,m}\hat{\sigma}^2_{k,m}\rvert$

We estimated the total set consisting of $D = 841,536$ news sentiment articles $\bar{X}_d$ on the 5-min time grid from the beginning of 2006 to the end of 2010. The estimation results are presented in table 5. The estimation results show that the standardized prediction errors $v_d/\sqrt{F_d}$ are clearly not normally distributed. However, the Ljung-Box statistics show no memory in $v_d/\sqrt{F_d}$ which does meet the $\mathcal{IID}$ assumptions made for the LNSL model.

We are interested in the impact of news sentiment on the dynamics of natural gas futures prices. Moreover, we are interested in the hypothesis that news sentiment causes various aspects of the natural gas futures price dynamics. As mentioned in section 3, our data are based on 1257 trading days from 3 January 2006 to 31 December 2010. Therefore, we construct news sentiment levels at closing time C for each trading day t. Specifically, we estimate the news sentiment level up to time d, where d is equal to closing time C of trading day t. This means that we estimate the news sentiment levels 1257 times with an increasing estimation window. Here, we only estimate Kalman filtered news sentiment levels since it is equal to the Kalman smoothed sentiment level at time C.

The Kalman filtered $\widetilde{S}^p_{d|D}$ and Kalman smoothed $\widehat{S}^p_{d|D}$ news sentiment levels are based on the total estimation set D.† The Kalman filtered $\widetilde{S}^p_{C,t|t}$ news sentiment levels are based on the increasing estimation window. Additionally, $\widetilde{AS}_{C,t|t}$ is the absolute news sentiment variable.

The summary statistics are presented in table 6.

The summary statistics show that the low-frequency Kalman filtered news sentiment levels $\widetilde{S}^p_{C,t|t}$ show similar statistics as their high frequency equivalents $\widetilde{S}^p_{d|D}$. Also, the $\widetilde{S}^p_{C,t|t}$ show very high autocorrelation which imply forecasting abilities.

Figure 4 shows plots of $\widetilde{S}^p_{d|D}$ and $\widehat{S}^p_{d|D}$ from 3 January 2006 to 31 December 2010. Figure 5 shows the same plots but for a shorter period, i.e. from 18 October 2008 to 31 October 2008. Additionally, figures 6 and 7 show plots of $\widetilde{S}^p_{C,t|t}$ and $\widetilde{AS}_{C,t|t}$ together with natural gas futures return and realized measures described in section 3.

6.2. Events and Granger causality

In this section, we perform two methods to analyse the impact of news sentiment on the dynamics of natural gas futures prices. The first method is an event study to investigate the natural gas price evolution around extreme news sentiment events. The second method involves Granger causality tests to determine causal relationships between natural gas price dynamics and news sentiment measures.

†Note that $p \in \{pos, neu, neg\}$.

6.3. Event study

We employ event studies as described in MacKinlay (1997) and used in Tetlock *et al.* (2008) and Borovkova (2011). We define a day in which the news sentiment measure $\widetilde{S}^{pos}_{c,t|t}$ is higher than the q quantile of the total sample of $\widetilde{S}^{pos}_{c,t|t}$ as an extreme positive news event. Analogously, a day in which the news sentiment measure $\widetilde{S}^{neg}_{c,t|t}$ is higher than the q quantile of the total sample of $\widetilde{S}^{neg}_{c,t|t}$ is an extreme negative news event. All days which are not selected as extreme positive or negative news events are defined as neutral news events.

We analyse the return dynamics around the event by computing the cumulative return from 10 days before up to 10 days after the event for all events of the same type. Specifically, we use a threshold quantile q of 80% on the set of 1256 trading days and select 252 extreme positive and extreme negative event days and a total of 753 neutral event days. This results in 252 extreme positive and extreme negative event cumulative return paths and 753 neutral event cumulative return paths. From these paths, we simulate the average cumulative return and 95% bootstrapped confidence intervals as described in Davidson and Mackinnon (2004), for each time in the event window. Simple plots of the average cumulative returns and the confidence intervals for each event type and for each time in the event window present the differences between the event types. This corresponds with the suggested statistics described in MacKinlay (1997).

Figure 8 presents the event study with the natural gas futures price and the monthly event frequencies for each event type. The return evolution is clearly different for the three event types. The extreme positive events show a mean reverting effect in the returns evolution around the event day. For extreme negative news, the returns show strong decline prior to and after the event day. Generally, the negative price move around an extreme negative event is much greater than any price effects surrounding positive news events, as already observed for the oil market in Borovkova (2011). As expected, the evolution of the returns around the neutral days are not different from zero.‡

It is interesting to see that the frequency of extreme positive days decreases in 2009, while the frequency of extreme negative news increases from 2008.§

Another interesting result shown in figure 8 is the possibility to make a significant profit from selling the futures on an extreme positive or negative event day.¶ However, the

‡For a q threshold of 80% neutral events are actually more regular than extreme and thus in probability equal to the sample mean of $r_{CtC,t}$.

§This can be related to the credit crunch of 2008 to 2009.

¶Here, we assume no transaction costs and zero market impact.

Table 5. Quasi maximum likelihood parameter estimates for the LNSL model based on the total set of news items $\bar{X}_d$ from the beginning of 2003 to the end of 2010. The asymptotic standard errors are given in parentheses which are obtained by the delta method since several parameters were transformed for estimation. The AIC represents the Akaike Information Criterion and is defined as $-2(\ln L) + 2p$ where p is the number of coefficients estimated. Q(l) is the Ljung-Box test statistic conducted on the standardized residuals for lag length l and is asymptotically χ^2 distributed with l degrees of freedom. JB represents the Jarque-Bera normality test on the standardized residuals and is asymptotically χ^2 distributed with two degrees of freedom.

	$\bar{s}_d^{pos}$	$\bar{s}^{neu}$	$\bar{s}^{neg}$
Observations	841536	841536	841536
$\sigma_\eta 1e^3$	6.071	0.85768	8.3968
	(0.465)	(0.133)	(0.551)
$\sigma_\epsilon 1e^3$	1106.3	903.82	1037.8
	(2.267)	(1.803)	(2.162)
ln L	−193535	−167199	−185813
AIC	387074	334402	371630
$v_d/\sqrt{F_d}$			
Mean	0.003	−0.007	−0.004
Median	0.000	0.000	0.000
Std.Dev.	3.602	6.574	6.160
Skewness	8.93E+02	−9.14E+02	−9.11E+02
Ex.Kurtosis	8.12E+05	8.38E+09	8.34E+05
Q(1)	0.426	0.005	0.029
Q(12)	0.905	0.009	0.063
Q(68)	2.099	0.016	0.404
Q(135)	4.193	0.297	0.879
JB Stat	2.31E+20	2.46E+20	2.44E+20

Table 6. Summary statistics: the Kalman filtered $\widetilde{S}^p_{d|D}$ and the Kalman smoothed $\widehat{S}^p_{d|D}$ which are based on the total information set D and the Kalman filtered $\widetilde{S}^p_{C,t|t}$ and $\widetilde{AS}_{C,t|t}$ selected at closing time C and based on the information set until t. Here p is in $\{pos, neu, neg\}$. The $SACF(l)$ statistic represent the sample autocorrelation function for lag l. Bold $SACF(l)$ statistics are significant at a 5% significance level based on heteroskedasticity robust standard errors described in White (1980). PP + drift statistic represents the Phillips Perron unit-root test statistic where a drift term is assumed in the model, see Phillips and Perron (1988) for details. ADF(l) + drift is the Augmented Dickey Fuller test statistic where up to l lags and a drift term are assumed in the model, see Said *et al.* (1984) for details. Bold (and italic) Phillips Perron and Augmented Dickey Fuller test statistics represent rejection of an unit-root based on a 5% (1%) significance level.

	$\widetilde{S}^{pos}_{d\|D}$	$\widetilde{S}^{neu}_{d\|D}$	$\widetilde{S}^{neg}_{d\|D}$		$\widetilde{S}^{pos}_{C,t\|t}$	$\widetilde{S}^{neu}_{C,t\|t}$	$\widetilde{S}^{neg}_{C,t\|t}$	$\widetilde{AS}_{C,t\|t}$
Mean	0.400	0.166	0.359	Mean	0.430	0.187	0.383	0.064
Median	0.399	0.167	0.355	Median	0.431	0.187	0.383	0.055
Std.Dev.	0.043	0.011	0.046	Std.Dev.	0.043	0.010	0.043	0.047
Skewness	−0.030	−0.332	0.535	Skewness	−0.085	−0.123	0.194	0.857
Ex.Kurtosis	0.121	0.592	0.604	Ex.Kurtosis	0.130	−0.069	0.043	0.570
Minimum	0.230	0.130	0.152					
Maximum	0.718	0.194	0.582	SACF(1)	**0.700**	**0.810**	**0.665**	**0.515**
				SACF(5)	**0.404**	**0.541**	**0.375**	**0.202**
	$\widehat{S}^{pos}_{d\|D}$	$\widehat{S}^{neu}_{d\|D}$	$\widehat{S}^{neg}_{d\|D}$	SACF(10)	**0.365**	**0.420**	**0.328**	**0.180**
Mean	0.400	0.166	0.357	SACF(20)	**0.308**	**0.309**	**0.266**	**0.150**
Median	0.400	0.167	0.354					
Std.Dev.	0.036	0.010	0.037	PP + drift	***−14.864***	***−11.544***	***−15.857***	***−20.037***
Skewness	−0.095	−0.379	0.551	ADF(1) + drift	***−12.13***	***−9.61***	***−12.86***	***−15.83***
Ex.Kurtosis	0.185	0.427	0.682	ADF(12) + drift	***−4.56***	***−5.20***	***−4.81***	***−5.89***
Minimum	0.249	0.139	0.262	ADF(24) + drift	**−3.08**	***−4.30***	**−3.30**	***−3.79***
Maximum	0.522	0.186	0.541					

robustness and riskiness of the suggested trading strategy should be investigated further. The event study also shows significant discriminating power of news sentiment. This is even more clear for the event studies presented in figure 9. These event studies have the same set-up as the event study in figure 8, but for threshold levels q of 50, 70 and 90%. Clearly, the discriminating power increases for a higher threshold.

From this, we can conclude that there is a significant relationship between news sentiment and the evolution of natural gas futures returns. Specifically, we find that the price evolution shows a mean reverting effect around the extreme positive events. Also, we find that the price evolution around extreme negative events is negative and shows positive autocorrelation. This means that negative price momentum

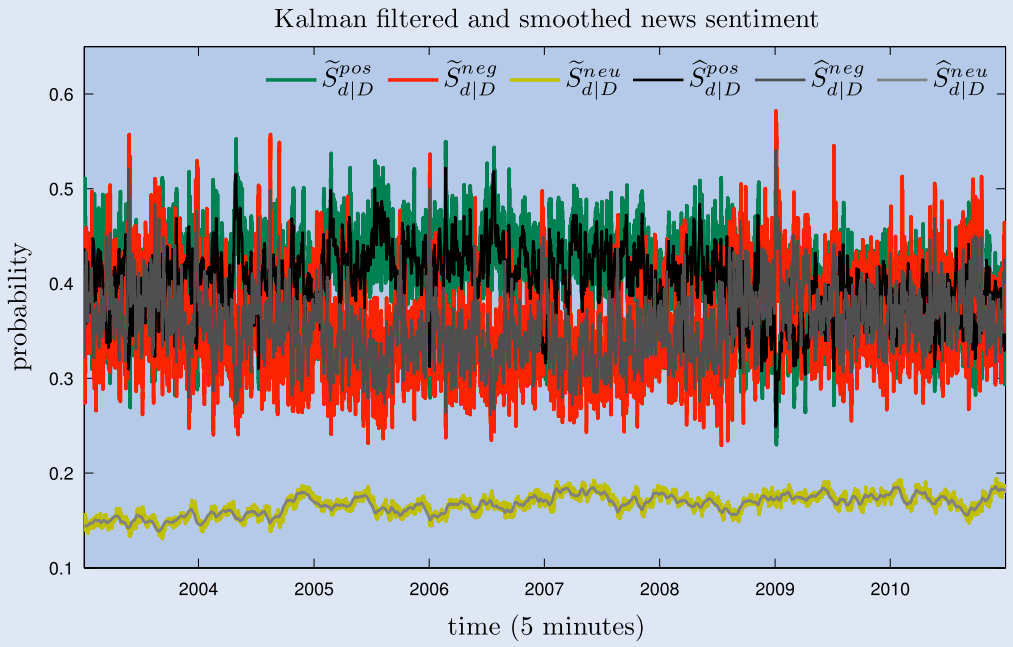

Figure 4. Kalman filtered and smoothed news sentiment levels.

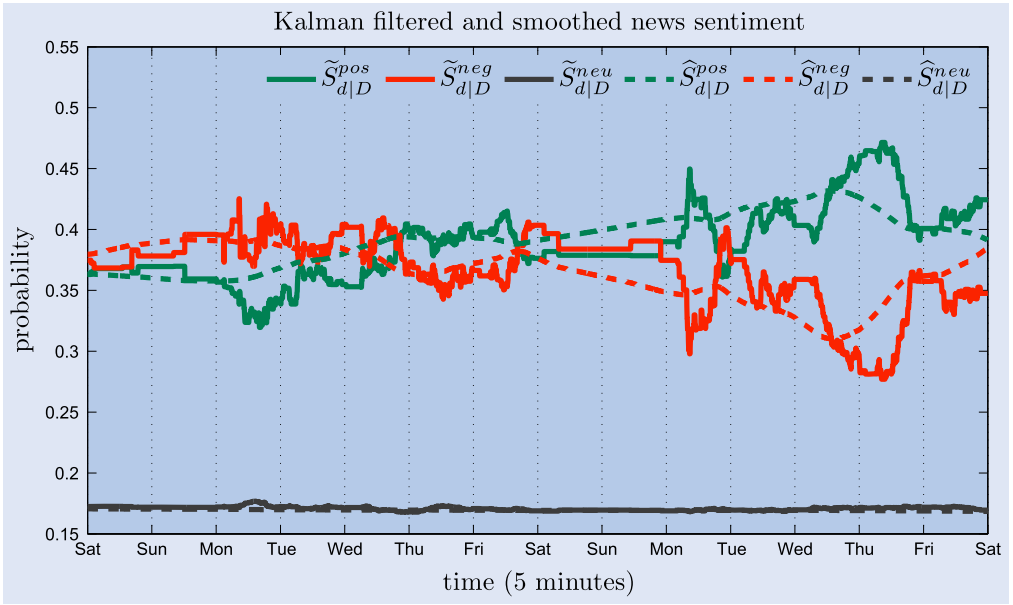

Figure 5. Kalman filtered and smoothed news sentiment levels.

continues after extreme negative days, before we observe any return to fundamentals. The market clearly gives more credence to negative news and reacts much stronger to those—a feature documented in earlier behavioural finance literature.

6.4. Granger causality tests

Both the news sentiment measures and a majority of the natural gas price dynamic measures show significant autocorrelation. This makes it more interesting to determine whether specific features of natural gas futures price dynamics are caused by specific news sentiment measures and/or vice versa. In order to test causality relationships between news sentiment measures and natural gas price dynamic measures, we perform (linear) Granger causality tests as first suggested by Granger (1969). A similar investigation was performed by Tetlock (2007).

For each sentiment measure $\widetilde{S}^{pos}_{c,t|t}$, $\widetilde{S}^{neu}_{c,t|t}$, $\widetilde{S}^{neg}_{c,t|t}$ and $\widetilde{AS}_{c,t|t}$, we construct a bivariate Vector Autoregression (VAR) of order P with each natural gas price dynamic measure described in section 3. A bivariate VAR(P) is defined as

$$\mathbf{Z}_t = \Phi_0 + \sum_{p=1}^{P} \Phi_p \mathbf{Z}_{t-p} + \mathbf{u}_t, \tag{41}$$

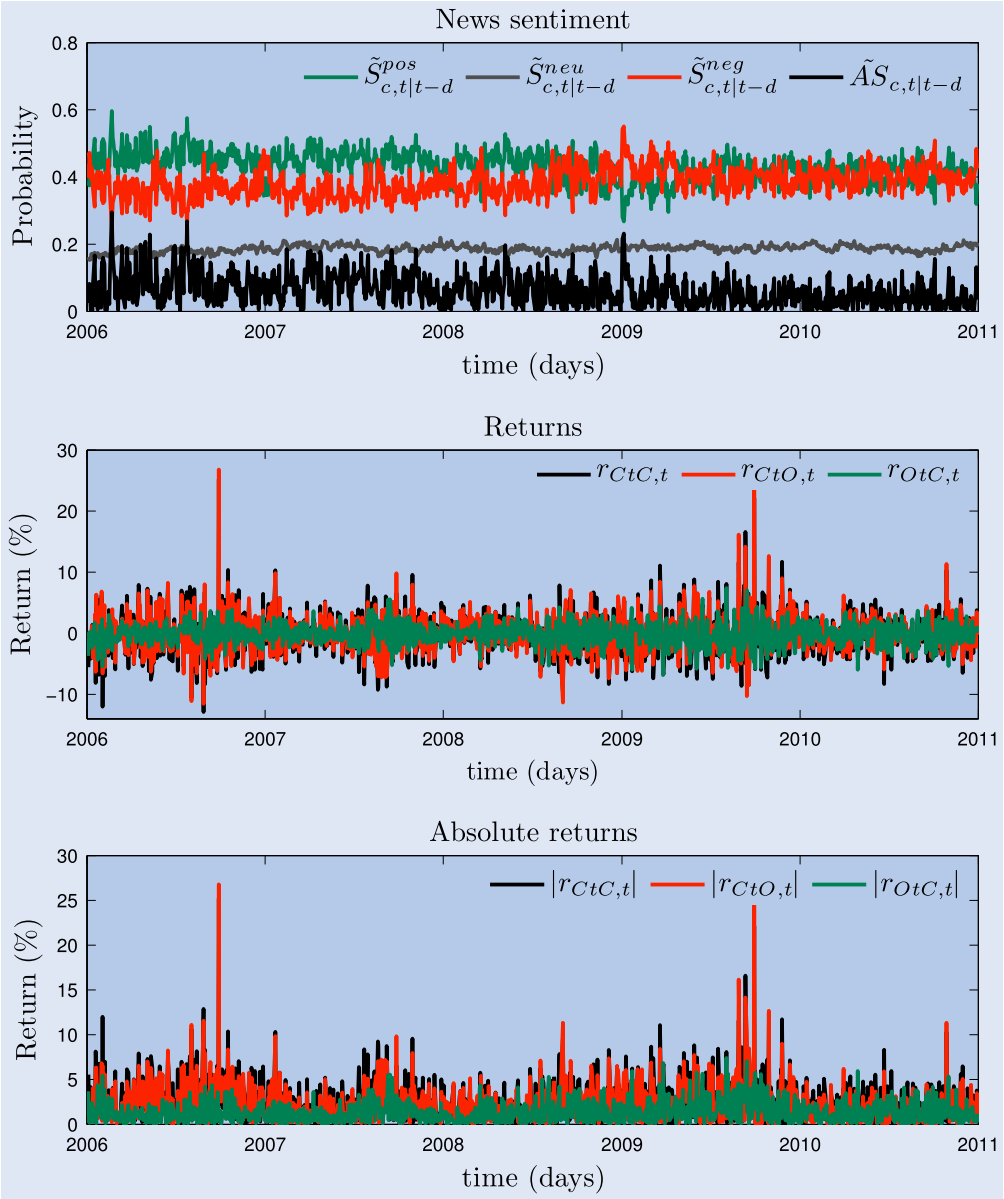

Figure 6. Kalman filtered news sentiment measures and natural gas return measures.

where $\mathbb{E}[\mathbf{u}_t] = 0$, $\mathbb{E}[\mathbf{u}_t\mathbf{u}_t'] = \Omega$, $\mathbf{Z}_t$ is a (2×1) vector, Φ_0 and Φ_p are (2×2) coefficient matrices and Ω a (2×2) covariance matrix. The order of P is determined by a selection procedure based on the Akaike Information Criterion (AIC), described in Lutkepohl (2005) and Gonzalo and Pitarakis (2002), which is a standard procedure for choosing the order in Granger causality tests. For the bivariate VAR of order P case, the AIC is defined as

$$AIC = \log|\widehat{\Omega}| + \frac{2(2^2P + 2)}{T}, \tag{42}$$

where $\widehat{\Omega}$ is the heteroskedasticity and autocorrelation (HAC) robust covariance matrix, see Newey and West (1987), and T the sample size.

The summary statistics of the natural gas price dynamics and the news sentiment measures show that none of the variables contain a unit-root up to a lag length of 24, see tables 3 and 6. Therefore, we are able to test the null hypothesis of no Granger causality for each time series in each bivariate VAR model. For details concerning the Granger causality test we refer to Granger (1969) and Lutkepohl (2005).

The Granger causality test results show many significant Granger causality relationships. First of all, absolute news sentiment (which measures the disagreement about the sentiment) Granger causes close-to-close and close-to-open returns $R_{CtC,t}$ and $R_{CtO,t}$. The same applies to open-to-close returns $R_{OtC,t}$, but only at a 10% significance level. If we look at the squared returns, we see a similar pattern for all news sentiment measures. Specifically, highly significant Granger causal relationships are shown for news sentiment measures and again, close-to-close and close-to-open squared returns $R^2_{CtC,t}$ and $R^2_{CtO,t}$, but hardly any for open-to-close squared returns $R^2_{OtC,t}$.

Recall that CtC returns are sums of CtO and OtC returns, so the above results imply that the arrival of news during non-trading periods has a large effect on the overnight (close-to-open) returns.

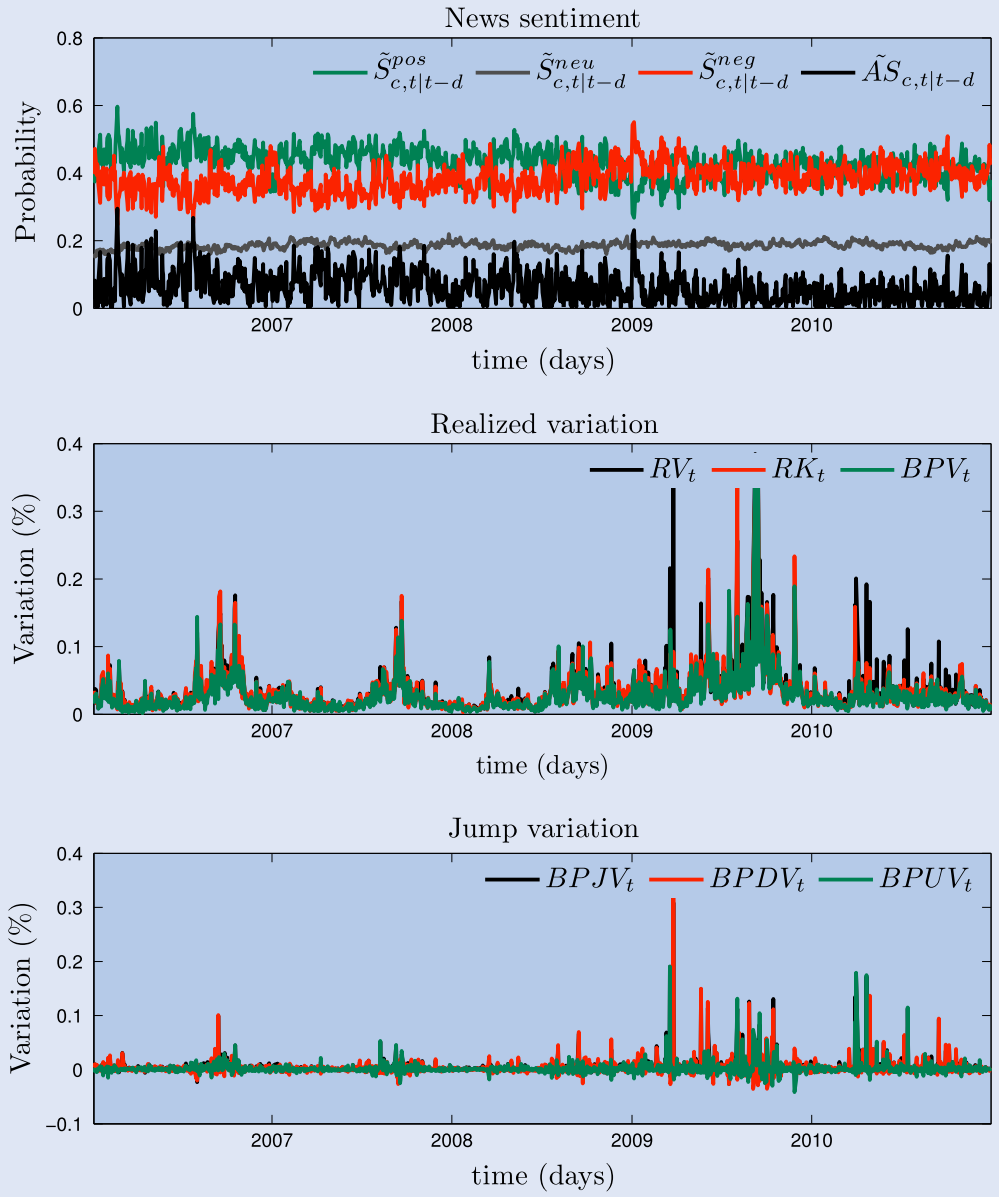

Figure 7. Kalman filtered news sentiment measures and natural gas realized measures.

Second, if we look at the realized variation measures, we see cross causality relationships between news sentiment measures and the realized kernel RK_t as well as the positive semivariance RSV_t^+. Also we see that the realized variance RV_t is significantly Granger caused by negative news sentiment and news sentiment itself - by negative semivariance RSV_t^-. Furthermore, we see that the jump variation robust Bipower variation BPV_t is significantly Granger caused by news sentiment.

Finally, if we look at the jump variation measures $BPJV_t$, $BPDV_t$ and $BPUV_t$, the Granger causality test show significant relationships between jump variation and news sentiment. There are some cross-correlation relationships, but the jump variation measure $BPJV_t$ is caused only by news sentiment measures and it only Granger causes the absolute news sentiment measure (disagreement). The same applies to $BPUV_t$, but it only shows Granger causal effects on the positive and negative news sentiment measures.

This difference between $BPJV_t$ and $BPUV_t$ shows that the disagreement (absolute news sentiment) is caused by jumps and, in particular, by negative jumps. This corresponds with the significant Granger causal effect of the realized negative semivariance RSV on the absolute news sentiment measure.

From all of the above, we can conclude that the volatility, and especially the negative semi-volatility, Granger causes the news sentiment. The opposite also holds: the news sentiment also Granger causes volatility. This implies that news cause and is caused by volatility in the market, but also that market participants trade as some function of aggregated news.

Granger causality relations between (absolute) news sentiment and the jump variation measures show the importance of asymmetric returns and jumps on news. Specifically, news

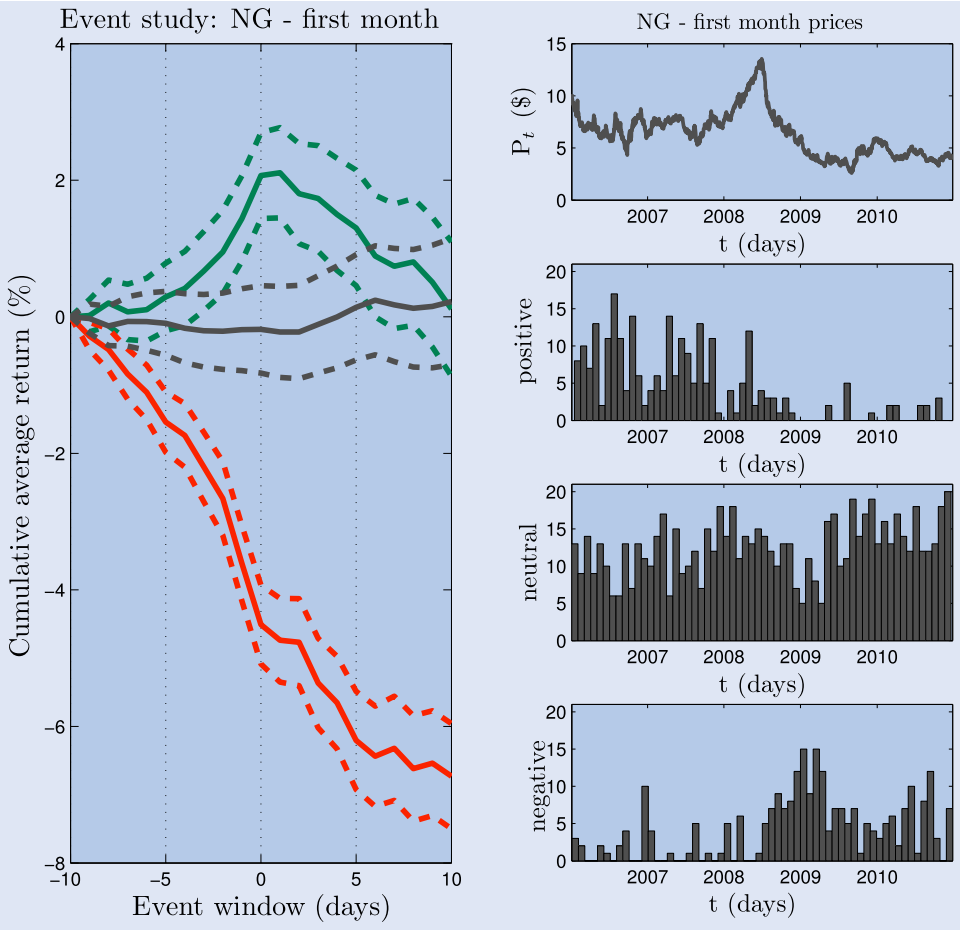

Figure 8. Cumulative return evolutions around extreme positive, extreme neutral and extreme negative event days are plotted based on a 10-day event window around the event day and a threshold q of 80%. Also the natural gas futures prices are plotted from the beginning of 2006 to the end of 2010. Furthermore, the monthly event frequencies are shown for all different event types.

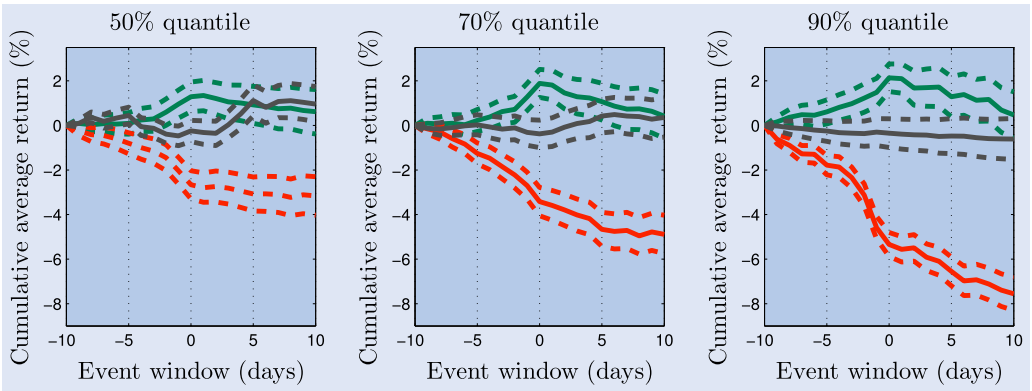

Figure 9. Cumulative return evolutions around extreme positive (green), extreme neutral (grey) and extreme negative (red) event days are plotted based on a 10 days event window around the event day and threshold q of 50, 90 and 90%.

sentiment is more sensitive to negative than to positive price jumps. We have also observed that news sentiment severely Granger causes jumps. That is, market participant seem to hard sell or hard buy natural gas futures contracts when news sentiment is high (in absolute value).

It is likely that there are also significant non-linear relationships between news, volatility and jump measures. For example, when negative sentiment rises by factor 2, volatility and jump measures could rise by a higher factor, given that market participants tend to overreact to negative news. These effects are not accounted for by means of linear Granger causality, but non-linear Granger causality tests of Hiemstra and Jones (1994) can reveal such relationships. However, since Granger causality analysis is not the main focus of this paper,

Table 7. Granger causality test results are presented based on bivariate vector autoregressions of order P (VAR(P)) for each news sentiment measure with all individual natural gas return measures. The results can be divided into four subtables which correspond with Granger causality tests based on bivariate VAR(P) of the news sentiment measures and return, squared return, realized variation and realized jump variation measures. Each subtable presents the LM test statistic of the null hypothesis that the measure of column j does not Granger cause the measure in row i. The LM test statistic is χ^2 distributed with P degrees of freedom. Italic, bold and italic+bold LM test statistics represent rejection of the null hypothesis at a 10%, 5% and 1% significance level, respectively. The value under the LM test statistic represents the lag order P of the bivariate VAR(P).

	$r_{t,CtC}$	$r_{t,CtO}$	$r_{t,OtC}$	$\widetilde{S}^{pos}_{c,t\|t}$	$\widetilde{S}^{neg}_{c,t\|t}$	$\tilde{A}S_{c,t\|t}$
$r_{t,CtC}$	–	–	–	21.526 24	23.798 24	**38.840** 24
$r_{t,CtO}$	–	–	–	31.390 24	*35.283* 24	**36.605** 24
$r_{t,OtC}$	–	–	–	24.015 24	25.136 24	*34.513* 24
$\widetilde{S}^{pos}_{c,t\|t}$	31.044 24	22.561 24	***48.534*** 24	–	–	–
$\widetilde{S}^{neg}_{c,t\|t}$	32.501 24	22.734 24	***47.556*** 24	–	–	–
$\tilde{A}S_{c,t\|t}$	16.938 24	23.285 24	30.015 24	–	–	–

	$r^2_{t,CtC}$	$r^2_{t,CtO}$	$r^2_{t,OtC}$	$\widetilde{S}^{pos}_{c,t\|t}$	$\widetilde{S}^{neg}_{c,t\|t}$	$\tilde{A}S_{c,t\|t}$
$r^2_{t,CtC}$	–	–	–	**38.374** 23	**35.704** 23	***42.787*** 23
$r^2_{t,CtO}$	–	–	–	***46.852*** 22	***45.607*** 22	**39.262** 22
$r^2_{t,OtC}$	–	–	–	28.767 24	*35.151* 24	26.398 24
$\widetilde{S}^{pos}_{c,t\|t}$	23.995 23	24.280 22	29.329 24	–	–	–
$\widetilde{S}^{neg}_{c,t\|t}$	23.170 23	26.679 22	26.954 24	–	–	–
$\tilde{A}S_{c,t\|t}$	24.522 23	**39.494** 22	*34.266* 24	–	–	–

	RV_t	RK_t	RSV^-_t	RSV^+_t	$\widetilde{S}^{pos}_{c,t\|t}$	$\widetilde{S}^{neg}_{c,t\|t}$	$\tilde{A}S_{c,t\|t}$
RV_t	–	–	–	–	33.137 24	**37.703** 24	28.115 24
RK_t	–	–	–	–	**38.718** 24	***43.132*** 24	25.981 24
RSV^-_t	–	–	–	–	31.585 23	31.072 23	22.349 23
RSV^+_t	–	–	–	–	***57.590*** 24	***61.144*** 24	29.465 24
$\widetilde{S}^{pos}_{c,t\|t}$	27.311 24	***45.243*** 24	*32.361* 23	**39.437** 24	–	–	–
$\widetilde{S}^{neg}_{c,t\|t}$	27.570 24	***46.060*** 24	*34.874* 23	***37.528*** 24	–	–	–
$\tilde{A}S_{c,t\|t}$	27.019 24	21.807 24	**41.587** 23	17.621 24	–	–	–

	BPV_t	$BPJV_t$	$BPDV_t$	$BPUV_t$	$\widetilde{S}^{pos}_{c,t\|t}$	$\widetilde{S}^{neg}_{c,t\|t}$	$\tilde{A}S_{c,t\|t}$
BPV_t	–	–	–	–	*36.019* 24	**39.379** 24	30.695 24
$BPJV_t$	–	–	–	–	***55.806*** 23	***58.887*** 23	***60.226*** 23
$BPDV_t$	–	–	–	–	***53.357*** 23	***55.907*** 23	***52.710*** 23
$BPUV_t$	–	–	–	–	***47.164*** 23	***51.225*** 23	***56.344*** 23
$\widetilde{S}^{pos}_{c,t\|t}$	30.488 24	24.448 23	***42.076*** 23	**39.350** 23	–	–	–
$\widetilde{S}^{neg}_{c,t\|t}$	32.423 24	29.170 23	***47.528*** 23	**38.519** 23	–	–	–
$\tilde{A}S_{c,t\|t}$	28.013 24	**36.601** 23	**39.943** 23	22.300 23	–	–	–

but rather a side issue, we postpone such an investigation to a subsequent paper. Moreover, for causality relationships to be used in trading, they should preferably be linear, because only then automated trading strategies can be devised (we cannot trade squares or exponentials of futures). For trading applications, any discovered non-linear relationships should be linearized. Hence, for trading applications, linear Granger causality analysis is appropriate.

6.5. Volatility forecasting results

6.5.1. Parameter estimation results. The estimation results from the GARCH-type volatility models described in section 4 are presented in table 8. Likewise, the estimation results from the HEAVY-r-type models are described in table 9. These estimation results are based on the total data-set from the beginning of 2006 to the end of 2010, see section 3 for specifics. All programmes for estimating the parameters are written in Ox, the programming environment of Doornik (2001). Implementation details are described in the work of Tsay (2005), Greene (2003) and Hansen (1994). The latter is used for the implementation of the Hansen skewed-t log-likelihood. All volatility models indicate high persistency. However, the HEAVY-r models are much stronger driven by RK_t than the GARCH models are driven by $R^2_{t,CtC}$. This indicates that the expected variance of $R_{t,CtC}$ is stronger driven by lagged RK_t than lagged $R^2_{t,CtC}$, as noted by Shephard and Sheppard (2010). The time-to-maturity parameter τ is significant and indicates and confirms the Samuelson hypothesis described in section 4. The estimates of the leverage parameter γ are all highly significant and positive for all GJRGARCH- and LHEAVY-type volatility models, confirming the hypothesis that the (regular) leverage effect is present in recent natural gas prices and not the inverse leverage. This is due to the fact that our data sample starts in 2006, when energy commodities started playing role of investment assets, due to the arrival of financial players in the commodity markets.

If we look at different assumed distributions for the return errors, we see that the student-t distribution is preferred over the normal distribution and the Hansen skewed-t distribution preferred over the student-t distribution. That is, in terms of a higher log-likelihood and of lower AIC and BIC criterion values. Although the skewness parameter λ is not significant for any model with Hansen skewed-t distributed return errors, allowing for non-zero skewness does increase the estimated degrees of freedom parameter ν with respect to models

Table 8. Quasi Maximum likelihood parameter estimation results for the GARCH-type volatility models described in section 4. The asymptotic standard errors are given in parentheses which are obtained by the delta method since several parameters were transformed for estimation. The AIC represents the Akaike Information Criterion and is defined as $-2(\ln L) + 2p$ where p is the number of coefficients estimated. The BIC represents the Bayesian Information Criterion and is defined as $-2(\ln L) + p \ln T$ where p is the number of coefficients estimated and T the sample size. The A-LM(l) and Q(l) statistics are the ARCH-LM test statistic of Engle (1982) and the Ljung-Box test statistic conducted on the standardized residuals for lag length l. Both statistics are asymptotically χ^2 distributed with l degrees of freedom. JB represents the Jarque-Bera normality test on the standardized residuals and is asymptotically χ^2 distributed with two degrees of freedom.

Model Parameter	GARCH	GARCH-t	GARCH-skewt	GJRGARCH	GJRGARCH-t	GJRGARCH-skewt
α_0	−0.324	−0.322	−0.367	−0.264	−0.245	−0.277
	(0.062)	(0.078)	(0.060)	(0.051)	(0.058)	(0.045)
α_1	30.009	25.658	25.829	8.488	7.048	7.085
	(0.005)	(0.006)	(0.004)	(0.005)	(0.006)	(0.004)
β_1	0.946	0.947	0.947	0.959	0.962	0.962
	(0.009)	(0.011)	(0.008)	(0.008)	(0.009)	(0.006)
τ	−0.005	−0.004	−0.003	−0.003	−0.003	−0.003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ν		10.853	24.652		12.348	27.970
		(0.003)	(0.004)		(0.007)	(0.008)
λ			−0.004			−0.003
			(0.019)			(0.019)
γ				46.417	40.531	40.852
				(0.010)	(0.010)	(0.007)
$\ln L$	2488.082	2500.853	8177.584	2498.644	2507.068	8190.323
AIC	−4968.163	−4991.705	−16343.168	−4987.287	−5002.136	−16366.645
BIC	−4947.617	−4966.023	−16312.349	−4961.605	−4971.317	−16330.690
A-LM(1)	0.036	0.104	0.101	0.188	0.060	0.065
A-LM(12)	22.123	22.049	22.051	22.906	22.778	22.782
Q(1)	0.372	0.343	0.344	0.107	0.102	0.101
Q(12)	13.235	13.460	13.452	13.144	13.244	13.239
JB	81.303	108.713	107.609	30.394	40.346	39.764
Model Parameter	GARCHX	GARCHX-t	GARCHX-skewt	GJRGARCHX	GJRGARCHX-t	GJRGARCHX-skewt
α_0	−0.311	−0.310	−0.352	−0.267	−0.249	−0.282
	(0.058)	(0.074)	(0.012)	(0.051)	(0.058)	(0.045)
α_1	28.304	24.172	24.327	8.977	7.639	7.675
	(0.005)	(0.005)	(0.003)	(0.005)	(0.006)	(0.004)
β_1	0.948	0.950	0.950	0.959	0.962	0.962
	(0.009)	(0.011)	(0.006)	(0.008)	(0.009)	(0.006)
τ	−0.005	−0.004	−0.004	−0.004	−0.003	−0.003
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
ν		11.020	24.979		12.509	28.291
		(0.003)	(0.004)		(0.007)	(0.008)
λ			-0.003			-0.003
				(0.025)		(0.019)
γ				45.063	38.909	39.230
				(0.010)	(0.011)	(0.008)
ϕ	0.140	0.136	0.136	0.071	0.066	0.066
	(0.091)	(0.100)	(0.086)	(0.076)	(0.079)	(0.056)
$\ln L$	2491.231	2503.758	8185.962	2502.014	2509.409	8197.561
AIC	−4972.462	−4995.515	−16357.925	−4992.028	−5004.818	−16379.121
BIC	−4946.779	−4964.697	−16321.969	−4961.209	−4968.862	−16338.029
A-LM(1)	0.097	0.193	0.189	0.123	0.026	0.029
A-LM(12)	23.660	23.461	23.468	23.396	23.294	23.306
Q(1)	0.357	0.334	0.335	0.115	0.112	0.111
Q(12)	12.736	12.991	12.982	12.860	12.979	12.971
JB	79.718	106.451	105.394	31.156	41.966	41.369

employing a student-t distribution. However, The Jarque-Bera test statistic increases for non-Gaussian distributions.

The Ljung-Box test statistics show no significant memory in the standardized residuals for each estimated model. Unfortunately, the ARCH-LM test statistics shows significant memory in the squared residuals for 12 lags. This suggests that some of the autocorrelation in the squares of $R_{CtC,t}$ is not captured by any of the models. Since this applies to all models it does not influence our research interest per se. However, it does give reason to add more autoregressive terms to the models or to consider fractionally integrated volatility models, see Koopman *et al.* (2005) and Baillie *et al.* (2007).

As described in section 4, the parameter ϕ is related to the absolute news sentiment variable $\widetilde{AS}_{c,t|t}$, which measures disagreement about the news. The GARCHX type models show very low significant positive relationships between the

Table 9. Quasi Maximum likelihood parameter estimation results for the HEAVY-r-type volatility models described in section 4. The asymptotic standard errors are given in parentheses which are obtained by the delta method since several parameters were transformed for estimation. The AIC represents the Akaike Information Criterion and is defined as $-2(\ln L) + 2p$ where p is the number of coefficients estimated. The BIC represents the Bayesian Information Criterion and is defined as $-2(\ln L) + p \ln T$ where p is the number of coefficients estimated and T the sample size. The A-LM(l) and Q(l) statistics are the ARCH-LM test statistic of Engle (1982) and the Ljung-Box test statistic conducted on the standardized residuals for lag length l. Both statistics are asymptotically χ^2 distributed with l degrees of freedom. JB represents the Jarque-Bera normality test on the standardized residuals and is asymptotically χ^2 distributed with two degrees of freedom.

Model parameter	HEAVY-r	HEAVY-r-t	HEAVY-r-skewt	LHEAVY-r	LHEAVY-r-t	LHEAVY-r-skewt
$\alpha_{h,0}$	−0.703	−0.654	−0.737	−0.709	−0.660	−0.758
	(0.148)	(0.299)	(0.229)	(0.224)	(0.254)	(0.228)
$\alpha_{h,1}$	200.620	170.960	172.200	184.630	162.430	166.080
	(0.034)	(0.069)	(0.048)	(0.054)	(0.059)	(0.047)
$\beta_{h,1}$	0.893	0.900	0.900	0.892	0.900	0.897
	(0.020)	(0.040)	(0.028)	(0.030)	(0.035)	(0.028)
τ	−0.006	−0.005	−0.005	−0.006	−0.005	−0.005
	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)
ν		12.269	27.952		12.605	28.605
		(0.004)	(0.006)		(0.007)	(0.009)
λ			−0.004			−0.005
			(0.019)			(0.019)
γ_h				163.830	100.560	103.070
				(0.029)	(0.008)	(0.002)
ln L	2490.70	2497.94	8172.06	2491.51	2498.43	8172.60
AIC	−4973.40	−4985.87	−16332.12	−4973.02	−4984.86	−16331.20
BIC	−4952.85	−4960.19	−16301.30	−4947.33	−4954.04	−16295.24
A-LM(1)	1.02	0.71	0.73	0.85	0.59	0.65
A-LM(12)	24.68	23.94	23.97	23.76	22.17	23.53
Q(1)	0.16	0.15	0.15	0.17	0.13	0.16
Q(12)	15.27	15.13	15.14	15.51	15.44	15.31
JB	28.33	36.74	36.16	24.08	33.66	31.96
Model parameter	HEAVY-rX	HEAVY-rX-t	HEAVY-rX-skewt	LHEAVY-rX	LHEAVY-rX-t	LHEAVY-rX-skewt
$\alpha_{h,0}$	−0.278	−0.266	−0.300	−0.234	−0.255	−0.288
	(0.068)	(0.070)	(0.021)	(0.052)	(0.064)	(0.050)
$\alpha_{h,1}$	108.980	97.657	98.433	82.477	87.610	88.142
	(0.021)	(0.021)	(0.006)	(0.017)	(0.020)	(0.014)
$\beta_{h,1}$	0.955	0.957	0.957	0.962	0.959	0.958
	(0.010)	(0.010)	(0.001)	(0.008)	(0.009)	(0.006)
τ	−0.006	−0.006	−0.006	−0.006	−0.006	−0.006
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
ν		16.684	37.274		17.653	39.372
		(0.007)	(0.012)		(0.012)	(0.015)
λ						-0.004
						(0.000)
γ_h				143.940	89.001	91.221
				(0.003)	(0.002)	(0.001)
ϕ	0.506	0.463	0.465	0.534	0.464	0.467
	(0.099)	(0.103)	(0.188)	(0.087)	(0.098)	(0.069)
ln L	2504.912	2508.787	8194.215	2506.012	2509.378	8195.467
AIC	−4999.824	−5005.574	−16374.430	−5000.024	−5004.756	−16374.934
BIC	−4974.141	−4974.755	−16338.475	−4969.205	−4968.801	−16333.842
A-LM(1)	0.272	0.248	0.257	0.258	0.181	0.187
A-LM(12)	24.453	25.749	25.797	25.651	25.101	25.129
Q(1)	0.085	0.073	0.073	0.068	0.072	0.072
Q(12)	12.951	12.885	12.885	12.570	13.120	13.126
JB	14.218	20.917	20.549	15.588	17.039	16.648

volatility forecast and the news sentiment variable. Moreover, the log-likelihood is only slightly increased when including news sentiment to GARCH type models.†

The HEAVY-r type models show a much stronger and highly significant positive relationship between the volatility forecast and the news sentiment variable. Also, the news sentiment augmented HEAVY-r-type models show higher log-likelihoods and lower values for both the AIC and BIC criterion functions than regular HEAVY-r-type models. From the Granger causality test results in section 6.4, we learned that $r^2_{OtC,t}$ is

†The AIC criterion values are all slightly lower for news sentiment augmented models. The BIC criterion values are only lower for GARCH models with skewed-t return errors.

Table 10. The Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Heteroskedasticity Adjusted Squared Error (MHASE), Mean Heteroskedasticity Adjusted Absolute Error (MHAAE) are defined as $M^{-1}\sum_{m=1}^{M} L_{i,k,m}\ \forall i$ and for each volatility model. A bold value represents the lowest average loss for the specific loss type. The R_1^2 and R_2^2 represent the goodness-of-fit statistic of the Mincer-Zarnowitz regressions $\tilde{\sigma}_m^2 = \gamma_0 + \gamma_1 \hat{\sigma}_{k,m}^2 + u_t$ and $\log(\tilde{\sigma}_m^2) = \gamma_0 + \gamma_1 \log(\hat{\sigma}_{k,m}^2) + u_t$, respectively. A bold value represents the highest R^2 statistic.

Model $\mathcal{M}_g$	MSE	MAE	MHASE	MHAAE	R_1^2	R_2^2
GARCH	1.57E−06	6.60E−04	1.395	0.751	0.107	0.198
GARCH-t	1.12E−06	6.11E−04	1.217	0.719	0.120	0.206
GARCH-skewt	1.27E−06	6.41E−04	0.334	0.504	0.158	0.237
GJRGARCH	1.22E−06	6.09E−04	1.094	0.665	0.157	0.241
GJRGARCH-t	1.11E−06	5.94E−04	1.042	0.659	0.152	0.240
GJRGARCH-skewt	1.26E−06	6.38E−04	0.340	0.512	0.177	0.261
GARCHX	1.91E−06	6.57E−04	1.158	0.702	0.135	0.222
GARCHX-t	1.57E−06	6.05E−04	1.074	0.675	0.108	0.219
GARCHX-skewt	1.37E−06	6.47E−04	0.332	**0.495**	0.115	0.229
GJRGARCHX	1.13E−06	**5.65E−04**	0.818	0.602	0.181	**0.265**
GJRGARCHX-t	**1.07E−06**	5.73E-04	0.879	0.624	**0.193**	0.259
GJRGARCHX-skewt	1.27E−06	6.29E−04	**0.319**	0.496	0.161	0.254
Model $\mathcal{M}_h$	MSE	MAE	MHASE	MHAAE	R_1^2	R_2^2
HEAVY-r	**9.31E−07**	5.53E−04	1.210	0.693	0.232	0.267
HEAVY-r-t	9.52E−07	5.54E−04	1.227	0.698	0.184	0.254
HEAVY-r-skewt	1.24E−06	6.14E−04	0.329	0.485	0.195	0.265
LHEAVY-r	1.15E−06	6.02E−04	1.303	0.732	0.186	0.251
LHEAVY-r-t	1.05E−06	5.95E−04	1.317	0.735	0.184	0.228
LHEAVY-r-skewt	1.15E−06	6.01E−04	0.327	0.484	0.223	0.270
HEAVY-rX	1.03E−06	5.52E−04	0.879	0.628	0.198	0.322
HEAVY-rX-t	9.32E−07	**5.41E−04**	0.840	0.627	0.222	0.323
HEAVY-rX-skewt	1.17E−06	5.97E−04	**0.272**	**0.455**	**0.237**	**0.354**
LHEAVY-rX	1.07E−06	5.64E−04	0.797	0.620	0.167	0.308
LHEAVY-rX-t	1.02E−06	5.60E−04	0.789	0.620	0.152	0.276
LHEAVY-rX-skewt	1.26E−06	6.20E−04	0.281	0.464	0.169	0.305

Table 11. The table presents the Hansen consistent SPA p-values of Hansen (2005) based on the Squared Error (SE), Absolute Error (AE), Heteroskedasticity Adjusted Squared Error (HASE) and Heteroskedasticity Adjusted Absolute Error (HAAE) loss function for each volatility model. The p-value can be interpreted as the intensity of base model $\mathcal{M}_k$ producing superior forecast. A p-value of <0.001 denotes a number smaller than 0.001.

Base model $\mathbf{M}_g$	SE	AE	HASE	HAAE	Base model $\mathbf{M}_h$	SE	AE	HASE	HAAE
GARCH	0.093	0.010	<0.001	0.001	HEAVY-r	**1.000**	0.553	<0.001	0.001
GARCH-t	0.848	0.153	<0.001	0.001	HEAVY-r-t	0.710	0.570	<0.001	<0.001
GARCH-skewt	0.110	0.084	0.398	0.302	HEAVY-r-skewt	0.022	0.054	0.013	0.008
GJRGARCH	0.099	0.031	0.001	0.006	LHEAVY-r	0.124	0.023	<0.001	<0.001
GJRGARCH-t	0.602	0.281	<0.001	0.006	LHEAVY-r-t	0.114	0.015	<0.001	<0.001
GJRGARCH-skewt	0.269	0.121	0.105	0.038	LHEAVY-r-skewt	0.015	0.056	0.014	0.007
GARCHX	0.160	0.044	<0.001	<0.001	HEAVY-rX	0.218	0.346	<0.001	<0.001
GARCHX-t	0.326	0.145	<0.001	0.001	HEAVY-rX-t	0.951	**1.000**	<0.001	<0.001
GARCHX-skewt	0.016	0.050	0.383	**1.000**	HEAVY-rX-skewt	0.017	0.106	**1.000**	**1.000**
GJRGARCHX	0.846	**1.000**	<0.001	0.012	LHEAVY-rX	0.204	0.272	<0.001	<0.001
GJRGARCHX-t	**1.000**	0.730	<0.001	0.008	LHEAVY-rX-t	0.217	0.329	<0.001	<0.001
GJRGARCHX-skewt	0.250	0.176	**1.000**	0.592	LHEAVY-rX-skewt	0.013	0.045	0.036	0.045

less influenced by news than $r_{CtO,t}^2$. Since HEAVY-r models are RK_t driven, and thus virtually OtC driven, the significant added value of including news sentiment to HEAVY-r models can be related to the strong Granger causal effect of news sentiment on $r_{CtC,t}^2$ and $r_{CtO,t}^2$.

Interesting to see is that the α_1 and $\alpha_{h,1}$ estimates are lower for the models including the news sentiment variable. This implies that news sentiment reduces the dependence of lagged $r_{CtC,t}^2$ and RK_t in the forecasts of future volatility. Hence, the

news sentiment models give less weight to extreme positive or negative volatile days and are more robust to outliers.

6.5.2. Preliminary forecasting results. The volatility forecasts are constructed as described in section 5. Figure 10 presents the logarithms of the volatility forecasts versus the realized forecast target $\tilde{\sigma}_m^2$ for GARCH- and HEAVY-r-type models with and without news sentiment.† For all model types, the errors between the forecasts and the forecasting targets are clearly heteroskedastic. This implies that the non-heteroskedasticity adjusted loss functions do not reflect an unbiased measure of loss. The GARCH-type volatility models do not clearly improve in forecasting performance when including news sentiment. In case of the HEAVY-r-type models, we do see an improvement in forecasting performance. More specifically, the HEAVY-r models including the news sentiment variable are less influenced by extreme high- or low-volatile days. This confirms the earlier mentioned implication that news sentiment augmented volatility models are more robust to outliers.

Table 10 presents the means of all loss functions $M^{-1}\sum_{m=1}^{M} L_{i,k,m}$: $L_{1,k,m}$; the Mean Squared Error (MSE), $L_{2,k,m}$; Mean Absolute Error (MAE), $L_{3,k,m}$; Mean Heteroskedasticity Adjusted Squared Error (MHASE), $L_{3,k,m}$; Mean Heteroskedasticity Adjusted Absolute Error (MHAAE).

All news sentiment augmented models outperform their non-news sentiment equivalents for all loss functions excluding the MSE and MAE statistics. In case of the HEAVY-r-type models, the lowest MSE is estimated for HEAVY-r model with normally distributed return errors. However, the MSE and MAE values are very low and do not differ that much. As mentioned earlier, the forecasting errors are heteroskedastic and because of that the MSE and MAE are not the preferred loss statistics.

The R^2 of the Mincer-Zarnowitz regressions, as described in section 5, show that news sentiment augmented models do help to improve the forecasting performance. As noted by Pagan and Schwert (1990) and Engle and Patton (2001), R_2^2 based on the more robust logarithmic regression is even more decisive. Nevertheless, Hansen and Lunde (2005) mentioned that the R^2 of Mincer-Zarnowitz regression is not ideal as an criterion for comparing volatility models since it does not penalize for a biased forecast.

6.5.3. Superior predictive ability test results. The preliminary forecasting results showed that news sentiment augmented volatility models outperform volatility without the news sentiment variable. However, these results only apply to one selected sample. To obtain more robust outcomes, we want to have the same outcome for similar samples. In the same spirit of Koopman *et al.* (2005) and Hansen and Lunde (2005), we perform superior predictive ability (SPA) tests of Hansen (2005) as described in section 5.

The SPA test results are presented in table 11. Specifically, the Hansen SPA p-values of Hansen (2005) are presented for each base model. The p-value can be interpreted as the intensity of base model $\mathcal{M}_k$ producing superior forecasts with respect to all other models. We do this independently for the GARCH-type and HEAVY-r-type volatility models. For example, the p-values for the GARCH base model represent the intensity of the GARCH model producing superior forecasts with respect to all GARCH-type models. The same applies to HEAVY-r-type volatility models. By doing this, we can analyse the forecasting performance of both news sentiment augmented GARCH- and HEAVY-r-type volatility models separately.

Overall, the same conclusions can be made as for the averaged loss function and R^2 statistics presented earlier. First of all it is interesting to see that, especially in the case of HEAVY-r-type models, the p-values for the SE and AE loss functions are high. We see that the GJRGARCHX-t and GJRGARCHX models are not outperformed by any other model GARCH-type models in case of the SE and AE loss functions. The same applies to the p-values based on the SE and AE loss functions for the HEAVY-r-type models. This means that based on the SE and AE loss functions the SPA test results are not decisive. This is related to the heteroskedastic forecasting errors as mentioned earlier. This indecisiveness result of the SPA test strengthens the conclusion made in section 6.5.2 that the MSE and MAE are not the preferred loss statistics. Moreover, this indecisiveness result implies that the SE and AE are not the preferred loss functions and therefore, we base our conclusions on the hetereskedasticity adjusted loss functions: HASE and HAAE.

If we look at the GARCH models, we see that GJRGARCHX-skewt and the GARCHX-skewt are not outperformed by any other model in terms of the HASE and HAAE, respectively. This includes the non-news sentiment augmented model equivalents GJRGARCH-skewt and GARCH-skewt models. However, the p-values of these specific GJRGARCH-skewt and GARCH-skewt models are not significantly outperformed by any other model. Therefore, we can only conclude that news sentiment augmented GARCH-type models is *at least* not outperformed by their non-news sentiment equivalents. This means that the inclusion of the news sentiment variable to a GARCH-type model *at least* results in an equal forecasting performance with respect to a GARCH model without the news sentiment variable.

The HEAVY-r models show a more dramatic result. Here, the HEAVY-rX-skewt model is the absolute winner, both in terms of the HASE and the HAAE. Also, the p-values show that all other models are significantly outperformed. Thus, we can state that news sentiment augmented HEAVY-r models outperform their non-news sentiment equivalents. This means that the inclusion of the news sentiment variable to a HEAVY-r-type model results in a forecasting performance with respect to a HEAVY-r model without the news sentiment variable.

7. Conclusion

7.1. Concluding remarks

We investigated the impact of news sentiment on the dynamics of natural gas futures prices traded on the New York Mercantile Exchange (NYMEX). We implemented the proposed LNSL model and constructed *autocorrelated* series of news

†The plots shows a subset of the forecasts and the forecasting targets from the beginning of 2007 to the end of 2008. This together with the logarithmic transformations makes the plots more clear.

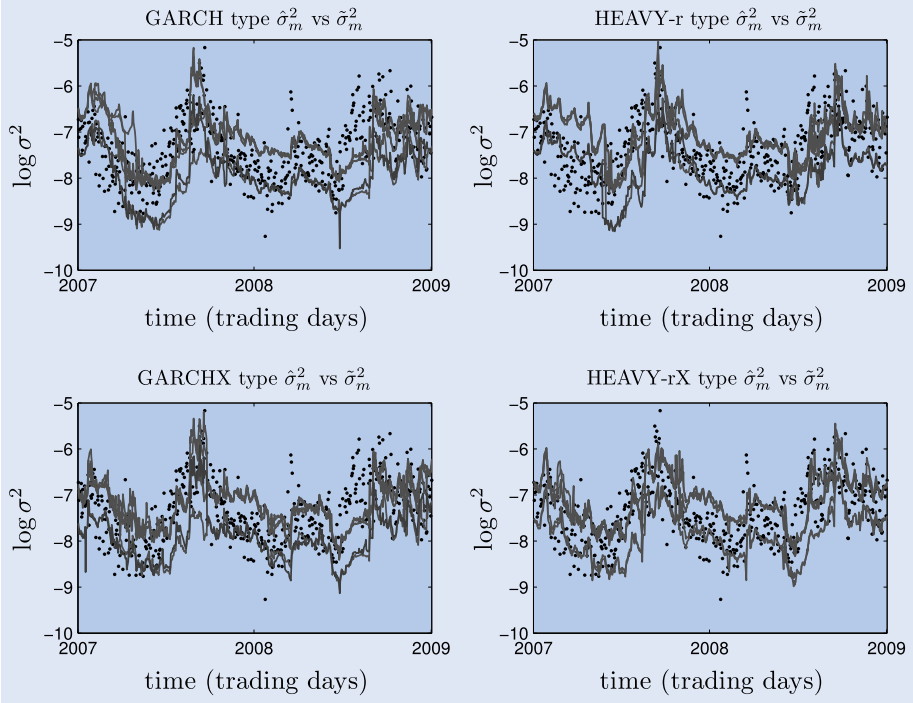

Figure 10. A subset of the forecasts and the forecasting targets are presented from the beginning of 2007 to the end of 2008. The logarithms of the forecasting target $\tilde{\sigma}_m^2$ are plotted as black dots. The logarithms of the forecasts $\hat{\sigma}_m^2$ are plotted as grey lines. Specifically, the forecasts are plotted for GARCH, GARCHX, HEAVY-r and HEAVY-rX model types.

sentiment probabilities which convey a positive, neutral or negative outlook on natural gas prices. Additionally, we constructed several return and variation measures to proxy for the dynamics of first month natural gas futures prices.

We have observed strong negative price trend surrounding the days which we refer to as extreme negative sentiment days, and mean reverting price behaviour after extreme positive sentiment days. From the Granger causality analysis, we found that the arrival of news in non-trading periods causes effects in overnight returns and that news sentiment is Granger caused by volatility (and especially negative semi-volatility). Also we found that news sentiment is caused by negative but not by positive price jumps. On the other hand, we found strong evidence that news sentiment severely Granger causes both positive and negative price jumps. From this, we conclude that market participants trade as some function of aggregated news. More specifically, market participants seem to hard sell or hard buy natural gas futures contracts when the absolute value of the news sentiment is high.

We conducted an out-of-sample volatility forecasting study in which we compared the one-step-ahead forecasting performance of two types of volatility models: the generalized autoregressive conditional heteroskedasticity (GARCH) and the high-frequency-based volatility (HEAVY) models of Shephard and Sheppard (2010) and Noureldin *et al.* (2012). By augmenting all models with news sentiment variables, we have shown that including news sentiment in volatility models results in superior volatility forecasts.

In this research, we assumed a quasi Local level model for the unobserved news sentiment (that is, the Local Level model of Durbin and Koopman (2001), applied to all three news sentiment probabilities separately). As mentioned in section 2, the news article's sentiments can be seen as draws from a trinomial distribution. For further research, we suggest modelling the unobserved news sentiment by modelling the time-varying trinomial distribution, for details see Durbin and Koopman (2001).

Also, it might be interesting to analyse the impact of news sentiment on the entire forward curve. However, the latter is hard to analyse since natural gas is subject to seasonal effects, as shown in Borovkova and Geman (2006), and the forward curve is only liquid up to contracts which mature no longer than a year ahead.

Finally, it is of great interest to conduct a similar volatility forecasting study for more volatility models, especially of the fractional integrated type.

References

Andersen, T., Bollerslev, T., Diebold, F. and Ebens, H., The distribution of realized stock return volatility. *J. Financ. Econ.*, 2001, **61**(1), 43–76.

Andersen, T., Bollerslev, T. and Lange, S., Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon. *J. Emp. Finance.*, 1999, **6**, 457–477.

Andersen, T.G., Dobrev, D. and Schaumburg, E., Jump-robust volatility estimation using nearest neighbor truncation. *J. Econ.* 2012, **169**(1), 75–93. Recent advances in panel data, nonlinear and nonparametric models: A Festschrift in honor of Peter C.B. Phillips. Available online at: http://www.sciencedirect.com/science/article/pii/S0304407612000127.

Baillie, R., Han, Y., Myers, R. and Song, J., Long memory models for daily and high frequency commodity futures returns. *J. Futures Markets*, 2007, **27**(7), 643–668.

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

Barndorff-Nielsen, O., Hansen, P., Lunde, A. and Shephard, N., Realised kernels in practice: Trades and quotes. *Econ. J.*, 2008b, **4**, 1–32.

Barndorff-Nielsen, O., Kinnebrock, S. and Shephard, N., Measuring downside risk – Realised semivariance. CREATES Research Papers 2008–42, School of Economics and Management, University of Aarhus, 2008.

Barndorff-Nielsen, O. and Shephard, N., Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *J. Roy. Stat. Soc. Ser. B*, 2002, **64**, 253–280.

Barndorff-Nielsen, O. and Shephard, N., Power and bipower variation with stochastic volatility and jumps (with discussion). *J. Financ. Econ.*, 2004, **2**, 1–48.

Barndorff-Nielsen, O. and Shephard, N., Econometrics of testing for jumps in financial economics using bipower variation. *J. Financ. Econ.*, 2008, **4**(1), 1–30.

Baur, D.G., Asymmetric volatility in the gold market. *J. Alternative Invest.*, 2012, **14**(4), 26–38.

Black, F., Studies of stock price volatility changes. In *Proceedings of the 1976 Meeting of the Business and Economic Statistics Section*, American Statistical Association, Washington, DC, 1976, 177–181.

Bollen, J., Mao, H. and Zeng, X., Twitter mood predicts the stock market. *J. Comput. Sci.*, 2011, **2**(1), 1–8.

Bollerslev, T., Generalized autoregressive conditional heteroscedasticity. *J. Econ.*, 1986, **69**, 542–547.

Bollerslev, T., Engle, R. and Nelson, D., Arch models. In *Handbook of Econometrics*, Vol. 4, edited by R.F. Engle and D.L. McFadden, pp. 2961–3038, 1994 (Elsevier Science: Amsterdam).

Borovkova, S., The forward curve dynamic and market transition forecasts. In *Modeling Prices in Competitive Electricity Markets* edited by D.W. Bunn, p. 24, 2004 (John Wiley & Sons Ltd).

Borovkova, S., News analytics for energy futures. Technical Report, 2011. Available online at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1719582.

Borovkova, S. and Geman, H., Seasonal and stochastic effects in commodity forward curves. *Rev. Derivatives Res.*, 2006, **9**(2), 167–186.

Brannon, E., Wusthoff, C., Gallistel, C. and Gibbon, J., Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychol. Sci.*, 2001, **12**, 238–243.

Carpantier, J.F., Commodities inventory effect. CORE Discussion Papers 2010040, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), July 2010. Available online at: http://ideas.repec.org/p/cor/louvco/2010040.html.

Cheong, C., Modeling and forecasting crude oil markets using arch-type models. *Energy Policy*, 2009, **37**, 2346–2355.

Crosby, J., A multi-factor jump-diffusion model for commodities. *Quant. Finance*, 2008, **8**(2), 181–200.

Davidson, R. and Mackinnon, J., *Econometrics Theory and Methods*, 2004 (Oxford University Press: New York).

Deng, S., Stochastic models of energy commodity prices and their applications: Mean-reversion with jumps and spikes. Working Paper, Georgia Institute of Technology, 1999.

Deng, S. and Oren, S.S., Electricity derivatives and risk management. *Energy*, 2006, **31**(6), 940–953.

Diebold, F. and Lopez, J., Forecast evaluation and combination. In *Handbook of statistics*. Vol. 14, Statistical methods in Finance, edited by G.S. Maddala and C.R. Rao, pp. 241–268, 1996 (North-Holland: Amsterdam).

Doornik, J., *Ox: Object Orientated Matrix Programming, 3.0*, 2001 (Tmberlake Consultants Press: London).

Du, X., Yu, C.L. and Hayes, D.J., Speculation and volatility spillover in the crude oil and agricultural commodity markets: A bayesian analysis. Working Paper 09-WP 491, Center for Agricultural and Rural Development Iowa State University, 2009.

Durbin, J. and Koopman, S., *Time Series Analysis by State Space Methods*, 2001 (Oxford Statistical Science Series: Oxford).

Engle, R., Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 1982, **50**, 987–1007.

Engle, R., New frontiers for arch models. *J. Appl. Econ.*, 2002, **17**, 425–446.

Engle, R. and Ng, V., Measuring and testing the impact of news on volatility. *J. Finance*, 1993, **48**, 1749–1778.

Engle, R. and Patton, A., What is a good volatility model? *Quant. Finance*, 2001, **1**(2), 237–245.

Geman, H. and Roncoroni, A., Understanding the fine structure of electricity prices. *J. Bus.*, 2006, **79**(3), 1225–1261.

Giamouridis, D.G. and Tamvakis, M.N., The relation between return and volatility in the commodity markets. *J. Alternative Invest.*, 2001, **4**(1), 54–62.

Glosten, L., Jagannathan, R. and Runkle, D., Relationship between the expected value and the volatility of nominal excess return on stocks. *J. Finance*, 1993, **48**, 1779–1802.

Gonzalo, J. and Pitarakis, J., Estimation and model selection based on inference in single and multiple treshold models. *J. Econ.*, 2002, **110**(2), 319–352.

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

Greene, W., *Econometric Analysis*, 2003 (Prentice Hall: Upper Saddle River, NJ).

Gross-Klussmann, A. and Hautsch, N., When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *J. Emp. Finance*, 2011, **18**(2), 321–340.

Hansen, B., Autoregressive conditional density estimation. *Int. Econ. Rev.*, 1994, **35**(3), 705–730.

Hansen, P., A test for superior predictive ability. *J. Bus. Econ. Stat.*, 2005, **23**, 365–380.

Hansen, P. and Lunde, A., A forecast comparison of volatility models: Does anything beat a garch(1,1)? *J. Appl. Econ.*, 2005, **20**, 873–889.

Hansen, P. and Lunde, A., Realized variance and microstructure noise. *J. Bus. Econ. Stat.*, 2006, **24**, 127–161.

Hassan, S., Modeling asymmetric volatility in oil prices. *J. Appl. Bus. Res.*, 2011, **27**(3), 71–78.

Hiemstra, C. and Jones, J.D., Testing for linear and nonlinear granger causality in the stock price-volume relation. *J. Finance*, 1994, **49**(5), 1639–1664.

Jacod, J., Li, Y., Mykland, P., Podolski, M. and Vetter, M., Microstructure noise in the continuous case: The pre-averaging approach. *Stoch. Process. Appl.*, 2009, **119**, 2249–2276.

Kalman, J., A new approach to linear filtering and prediction problems. *Basic Eng. Trans. ASMA Ser. D*, 1960, **82**, 35–45.

Koopman, S., Jungbacker, B. and Hol, E., Forcasting daily variability of the s&p 100 stock index using historical, realised and implied volatility measurements. *J. Emp. Finance*, 2005, **12**, 445–475.

Koopman, S., Shephard, N. and Doornik, J., Statistical algorithms for models in state space using ssfpack 2.2. *Econ. J.*, 1999, **2**, 113–166.

Kristoufek, L., Leverage effect in energy futures. Working paper series, 2014. Available online at: http://arxiv.org/pdf/1403.0064.pdf.

Lutkepohl, H., *New Introduction to multiple Time Series analysis*, 2005 (Oxford Statistical Science Series: Oxford).

MacKinlay, C., Event studies in economics and finance. *J. Econ. Lit.*, 1997, **35**, 13–39.

Marfe, R., Pure jump models for energy prices. *Energy Risk*, 2010, **7**(6), 57–62.

Martens, M., Measuring and forecasting s&p 500 index-futures returns. *J. Futures Markets*, 2002, **22**, 497–518.

Mason, C. and Wilmot, N., Jump processes in natural gas markets. CESifo Working Paper Series No. 4604, 2014.

Moratoya, E. and Reginaldo, S.F., Volatility clustering and leverage effects in soybean prices: A comparison between chicago board of trade and goias. *Proceedings of 3d Risk management and commodity trading conference*, 2013 (brazil).

Stigler, M., Commodity prices: Theoretical and empirical properties. In *Safeguarding Food Security in Volatile Global Markets*, edited by A. Prakash, pp. 25–41, 2011 (Food and Agriculture Organization of the United Nations: Rome). ISBN 978-92-5-106803-8.

Nelson, D., Conditional heteroskedasticity in asset pricing: A new approach. *Econometrica*, 1991, **59**, 347–370.

Newey, W. and West, K., A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 1987, **55**(3), 703–708.

Ng, V. and Pirrong, S., Fundamentals and volatlity: Storage, spreads and the dynamics of metals prices. *J. Bus.*, 1994, **67**(2), 203–230.

Nickerson, R., *Mathematical Reasoning: Patterns, Problems, Conjectures and Proofs*, 2009 (Psychology Press: London).

Nofsinger, J., Social mood and financial economics. *J. Behav. Finance*, 2005, **6**(3), 144–160.

Noureldin, D., Shephard, N. and Sheppard, K., Multivariate high-frequency-based volatility (HEAVY) models. *J. Appl. Econ.*, 2012, **27**(6), 907–933.

Pagan, A. and Schwert, G., Alternative models for conditional volatility. *J. Econ.*, 1990, **45**, 267–290.

Phillips, P. and Perron, P., Testing for a unit root in time series regression. *Biometrica*, 1988, **75**, 335–346.

Politis, D. and Romano, J., The stationary bootstrap. *J. Amer. Stat. Association*, 1994, **89**, 1303–1313.

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

Said, E., David, A. and Dickey, D., Testing for unit roots in autoregressive moving average models of unknown order. *Biometrica*, 1984, **71**, 599–607.

Salisu, A.A. and Fasanya, I.O., Modelling oil price volatility with structural breaks. *Energy Policy* 2013, **52**(C), 554–562.

Samuelson, P., Proof that properly anticipated prices fluctuate randomly. *Indus. Manag. Rev.*, 1965, **6**, 41–49.

Shephard, N. and Sheppard, K., Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. *J. Appl. Econ.*, 2010, **25**, 197–231.

Smith, V., Constructivist and ecological rationality in economics. *Amer. Econ. Rev.*, 2003, **93**, 465–508.

Swaray, R. B., Volatility of primary commodity prices. Discussion paper in economics, No. 2002/06, University of York, 2002.

Tetlock, P., Giving content tot investor sentiment: The role of media in the stock market. *J. Finance*, 2007, **62**(3), 1139–1168.

Tetlock, P., Saar-Tsechansky, M. and Macskassy, S., Quantifying language to measure firms' fundamentals. *J. Finance*, 2008, **63**(3), 1437–1467.

Tothova, M., Main challenges of price volatility in agricultural commodity markets. In *Methods to Analyze Agricultural Commodity Price Volatility*, edited by I. Piot-Lepetit and R. M'Barek, pp. 13–29, 2011 (Springer Science+Business Media: New York.

Tsay, R., *Analysis of Financial Time Series*, 2nd ed., 2005 (John Wiley & Sons, Inc.: Hoboken, NJ).

Tully, E. and Lucey, B.M., A power garch examination of the gold market. *Res. Int. Bus. Finance*, 2007, **21**(2), 316–325.

Uttal, W., *Time, Space and Number in Physics and Psychology*, 2008 (Sloan Publishing: Cornwall-on-Hudson, NY).

White, H., A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrics*, 1980, **48**(4), 817–838.

Zhang, L., Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. *Bernoulli*, 2006, **12**, 1019–1043.