News, volatility and jumps: the case of Natural Gas futures

Svetlana Borovkova [1]
Diego Mahakena [2]

Abstract

We investigate the impact of Thompson Reuters News Analytics (TRNA) news sentiment on the price dynamics of natural gas futures traded on the New York Mercantile Exchange (NYMEX). We propose a Local News Sentiment Level (LNSL) model, based on the Local Level model of Durbin and Koopman (2001), to construct a running series of news sentiment on the basis of the 5-minute time grid. Additionally, we construct several return and variation measures to proxy for the fine dynamics of the front month natural gas futures prices. We employ event studies and Granger causality tests 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 returns. For example, we find that the arrival of news in non-trading periods causes overnight returns, that news sentiment is Granger caused by volatility and that strength of 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 jumps and conclude that market participants trade as some function of aggregated news.

We apply several state-of-the-art volatility models augumented with news sentiment and conduct an out-of-sample volatility forecasting study. The first class of models is the generalized autoregressive conditional heteroskedasticity models (GARCH) of Engle (1982) and Bollerslev (1986) and the second class is the high-frequency-based volatility (HEAVY) models of Shephard and Sheppard (2010) and Noureldin et al. (2011). We adapt both models to account for asymmetric volatility, leverage and time to maturity effects. By augmenting all models with a news sentiment variable, we test the hypothesis whether including news sentiment in volatility models results in superior volatility forecasts. We find significant evidence that this hypothesis holds.

Keywords: news sentiment, natural gas futures, state space modelling, Kalman filter, realized variance, bipower variation, Granger causality, volatility modelling.

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 analyze security price dynamics. Traditionally, the analysis of security price dynamics was concentrated on responses of quantitative or hard measures, such as price derived data, corporate fundamentals, macro economic statistics or other variables intended to proxy for qualitative characteristics. Nowadays, rapid IT developments unable to process huge bulks of digitalized text to quantify soft qualitative information such as sentiment. Recently, the business data provider Thompson Reuters has introduced the Thompson Reuters News Analytics Engine (TRNAE). This engine is based on powerful linguistic analysis techniques and conducts a computerized analysis on millions of news articles to determine whether the news articles reflect a positive, negative or neutral sentiment and the relevance for a specific security.

[1] *Corresponding author, Vrije Universiteit Amsterdam, VU Amsterdam, FEWEB, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Email: s.a.borovkova@vu.nl, tel. +31-20-5982937.*

[2] *Vrije Universiteit Amsterdam, VU Amsterdam, FEWEB, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Email: diego.mahakena@gmail.com.*

Behavioral economics and behavioral finance literature tell us that sentiment can affect the behavior and decision making of individuals, see Smith (2003) and Nofsinger (2005). However, the amount of research on news sentiment and security price dynamics is small. 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, 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 microblog Twitter. They find 86.7% accurate predictions of daily up and down changes in daily DJIA returns.

The aim of our research is to investigate the impact of TRNAE derived news sentiment on the dynamics of daily natural gas futures prices traded on the New York Mercantile Exchange (NYMEX). After crude oil, natural gas is the largest and most liquid energy commodity traded. 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 sensitive contract. Therefore, we only consider first month futures prices.

The TRNAE data 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. We map the natural gas tagged TRNAE data into an aggregated news sentiment on a 5-minute time grid from the beginning of 2003 to the end of 2010.

Literature on behavioral processes states that animals[1] virtually discount future payoffs 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. (2011). 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 unobserved sentiment can be modeled as an autoregressive time series model.

Therefore, we utilize the state space modeling framework of Durbin and Koopman (2001) to model the real unobserved autoregressive news sentiment. Specifically, we apply their Local Level (LL) model specification to the 5-minute time grid news article probabilities. 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 all probabilities for each day. By doing this, we construct *autocorrelated* news sentiment probabilities (news sentiment) which conveys positive, neutral or negative outlook on natural gas prices, based on a 5-minute time grid from the beginning of 2003 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 intradaily data on a 1- and 5-minute 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*. Therefore, 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. (2008-42).

To analyze the impact of news sentiment on the dynamics of daily natural gas futures prices, we employ event studies. Specifically, we setup event studies as described in MacKinlay (1997) and used in Tetlock et al. (2008).

We find that the price evolution of first month natural gas futures contracts shows a mean reverting effect around days which we refer to as extreme positive sentiment days. Also, we find that the price evolution around extreme negative sentiment days shows negative price momentum which strongly continues after the event day before we observe

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

a return to fundamentals. From this we conclude that there is a significant relationship between news sentiment and the evolution of natural gas futures returns.

We find no unit roots in the constructed news sentiment and natural gas futures price dynamic measures. Therefore, we are able to employ Granger causality test as first suggested by Granger (1969) and used by Tetlock (2007) and Bollen et al. (2011). We construct bivariate vector autoregression models which for all news sentiment measures with all constructed price dynamic measures. We perform Granger causality tests on these models to investigate whether news sentiment measures Granger cause the price dynamic measures and vice versa.

We find that the arrival of news in non-trading periods causes effects in overnight returns and that news sentiment is Granger caused by volatility in the market. Also, we find that news sentiment, in an absolute sense, is more sensitive to negative jumps in the market than by jumps in general, including positive jumps. However, 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 in which we compare the one-step-ahead forecasting performance of two types of volatility models 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 volatility (HEAVY) models of Shephard and Sheppard (2010) and Noureldin et al. (2011). 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.

Specifically, we model GARCH and HEAVY-r type models for several error return distributions, including models that allow for asymmetric returns based on the work of Glosten et al. (1993) (GARCH type) and Barndorff-Nielsen et al. (2008-42) (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 setup of our forecasting study is similar to the work of Andersen et al. (1999), Martens (2002), Hansen and Lunde (2005b) 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. Specifically, we follow the work of Hansen and Lunde (2005b) and Koopman et al. (2005) and conduct Superior Predictive Ability tests of Hansen (2005a) to test our hypothesis.

We find significant evidence that including news sentiment to volatility models results in superior volatility forecasts.

This writing 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. Section 4 considers the details on several volatility models and the methodology concerning the forecasting of the volatility models are described in Section 5. All results, including the event studies and Granger causality tests are presented in Section 6. Section 7 contains a summary and suggestions for further research.

2. Modeling 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, most most likely time-varying, trinomial distribution with parameters $s^{pos}_{X,n}$, $s^{neg}_{X,n}$ and N.

News items may arrive non-equispaced and non-sequential over 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

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

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[2]. We introduce a new variable which we define as the absolute news sentiment. This variable is given by

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

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.

2.3. *Local news sentiment level model*

We define S^p_d as the real unobserved probability on news sentiment at time d. Literature on behavioral processes states that animals[3] virtually discount future payoffs 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. (2011). 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^p_t contains information on future probabilities and implies that S^p_t can be modeled as an autoregressive time series model.

Unfortunately, S^p_d is unobserved. As a proxy for S^p_d we use the observed probability s^p_d as defined in (3). A simple but effective way of modeling an unobserved autoregressive time series is the Local Level (LL) described in Durbin and Koopman (2001)[4]. Applied to the observed and unobserved probabilities s^p_d and S^p_d we define the Local News Sentiment Level (LNSL) model as follows

$$\begin{aligned} S^{*p}_{d+1} &= S^{*p}_d + \eta^p_d \quad \eta^p_d \xrightarrow{d} \mathcal{IID}(0, \sigma^2_{\eta^p}), \\ s^{*p}_{d+1} &= S^{*p}_d + \epsilon^p_d \quad \epsilon^p_d \xrightarrow{d} \mathcal{IID}(0, \sigma^2_{\epsilon^p}), \end{aligned} \tag{5}$$

where s^{*p}_{d+1} and S^{*p}_{d+1} are the logit transformed equivalents of s^p_d and S^p_d[5]. The first equation in (5) is the state equation which describes the evolution of the unobserved state of S^{*p}_d. The second equation represents the signal equation. Obviously, the signal equation describes the evolution of the observed s^{*p}_d. Both equations are modeled as random walks, such that the LNSL model is equivalent to the LL model[6]. 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]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.

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

[4]As mentioned earlier, the news sentiment can also be seen as a time-varying trinomial distribution with parameters S^{pos}_d, s^{neg}_d 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.

[5]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^p_d and S^p_d are required by the LNSL model because both equations are defined on $\mathbb{R}$.

[6]The LNS model is a model representation for the Exponential Weighted Moving Average (EWMA) model, see 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[7]. 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

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

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

$$\begin{array}{ll} \mathrm{r}_{d-1} = \frac{v_d}{F_d} + L_d r_d & \mathrm{N}_{d-1} = F_d^{-1} Z_d + L_d^2 N_d, \\ \mathrm{L}_d = 1 - K_d, & \\ \widehat{S}_d^{*p} = \widetilde{S}_d^{*p} + P_d r_{d-1} & \mathrm{V}_d = P_d - P_d^2 N_{d-1}, \end{array}$$

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.

[7] 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

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

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 NewsScope Sentiment Engine (RNSE) for historical commodities data. The dataset contains news items from the beginning of 2003 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 RNSE news items also provide a relevance indicator, item-type and several other variables we will not use in this work. 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 RNSE dataset for news items tagged as 'NGS', which are natural gas related. We consider the resulting 310614 news items as X_n^* for $n = 1, \ldots, 310, 614$. From this dataset 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 185982 news items such that X_n for $n = 1, \ldots, 185, 982$.

We aggregate the cleaned news item dataset on a 5-minute time grid by applying equations (2) and (3) to every probability in X_n. From the beginning of 2003 to the end of 2010 this results in 841,536 5-minute 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 dataset X_n. Here we see that new news arrives within 23 minutes on average and according to a median of less than 6 minutes, 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[8].

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.

[8]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.

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-minute			
	$\bar{s}^{pos}_t$	$\bar{s}^{neu}_t$	$\bar{s}^{neg}_t$
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

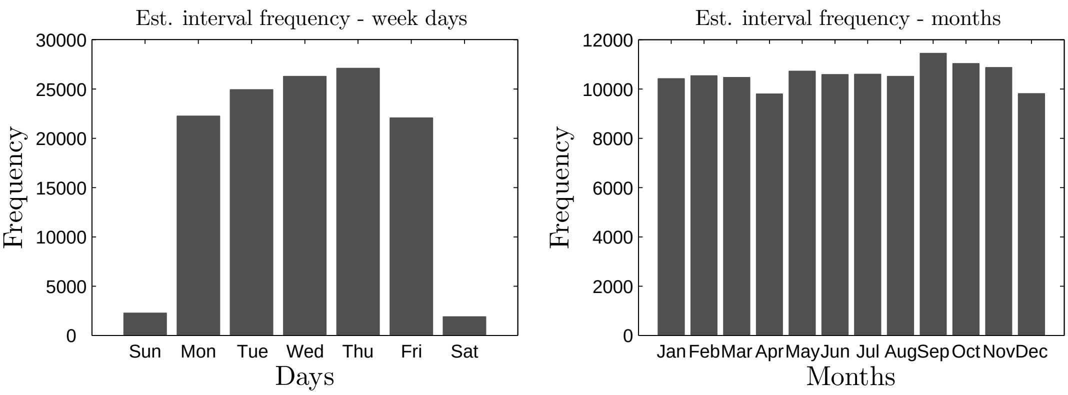

Figure 1: Aggregated interval frequencies.

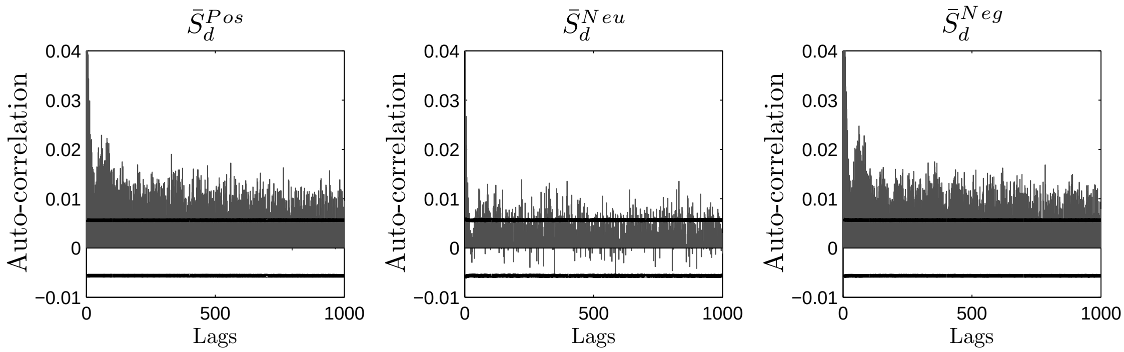

Figure 2: Aggregated interval autocorrelation functions with 95% confidence intervals based on heteroskedasticity robust standard errors of White (1980).

3. Natural gas futures returns, volatility and jumps

We investigate the price dynamics of first 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. Thompson Reuters provided us a sample of 415,371 last quotes from January 3 2006 to December 31 2010 on a 1-minute time-frame, based on 1,257 trading days. After cleaning the sample[9], we constructed a 1-minute and a 5-minute 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-minute and 5-minute time grid respectively.

From the 5-minute 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-minute $r_{t,i}$, 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[10]. 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. (2008-42), we assume the log price process to be represented by a Brownian semimartingale ($\mathcal{BSM}$)

$$Y_t = \int_0^t a_s ds + \int_0^t \sigma_s dW_s, \quad t \geq 0, \tag{7}$$

where a is a locally bounded predictable drift process and σ is a $c\grave{a}dl\grave{a}g$ 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

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

[9]We cleaned the quote data according to the cleaning procedure described in Barndorff-Nielsen et al. (2008b).

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

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 impact of microstructure noise severly influences RV_t. More microstructure noise robust estimators are: pre-averaging Jacod et al. (2009), multiscale Zhang (2006) and the realized kernel Barndorff-Nielsen et al. (2008a). In this research we use the realized kernel with a Parzen weight function which 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

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

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-minute 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.

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

$$Y_t = \int_0^t a_s ds + \int_0^t \sigma_s dW_s + J_t, \quad t \ge 0, \tag{12}$$

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

$$[Y]_t = \int_0^t \sigma_s^2 ds + \sum_{s \le t} (\Delta Y_s)^2, \tag{13}$$

and

$$d[Y]_t = \sigma_s^2 dt + \Delta Y_t. \tag{14}$$

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

$$\{Y\}_t^{[1,1]} = \sum_{i=3}^{[t/\delta]} |Y_{t,i} - Y_{t,i-1}| \, |Y_{t,i-1} - Y_{t,i-2}|, \tag{15}$$

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

$$\{Y\}_t^{[1,1]} = \mu_1^2 \int_0^t \sigma_s^2 ds, \tag{16}$$

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

$$[Y]_t - \mu_1^{-2}\{Y\}_t^{[1,1]} = \sum_{s \le t} (\Delta Y_s)^2. \tag{17}$$

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

Table 2: **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-minute 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	-

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

$$BPV_t = \mu_1^{-2} \sum_{i=3}^{i \leq t} |r_{t,i} - r_{t,i-1}| \, |r_{t,i-1} - r_{t,i-2}| \, . \tag{18}$$

For simplicity we already multiplied the estimator for (15) by μ_1^{-2}. We estimate the Bipower jump variation $BPJV_t$ for $r_{t,i}$ by

$$BPJV_t = RV_t - BPV_t, \tag{19}$$

where RV_t is the simple QV estimator given in (10) and BPV_t the scaled Bipower variation estimator given in (18). Notice that since BPV_t is a jump robust estimator for QV, $BPJV_t$ is asymptotically equal to zero in case of $Y_t \in \mathcal{BSM}$, which means that the log-price process does not contain a jump process. In the case of nonzero $BPJV_t$ 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.

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. (2008-42) 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

$$RSV_t^- = \sum_{i=1}^{I} r_{t,i}^2 \mathbf{1}_{r_{t,i} \leq 0}, \tag{20}$$

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

$$RSV_t^+ = \sum_{i=1}^{I} r_{t,i}^2 \mathbf{1}_{r_{t,i} \geq \leq 0}. \tag{21}$$

This implies

$$RV = RSV^- + RSV^+. \tag{22}$$

Barndorff-Nielsen et al. (2008-42) 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 (19). Specifically, the Bipower downward variance $BPDV_t$ is given by

$$BPDV_t \cong RSV_t^- - \frac{1}{2} BPV_t. \tag{23}$$

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

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

These variables are proxies for negative and positive jump variation. Also the $BPDV_t$ and $BPUV_t$ can be negative for small samples, since 23 and 24 hold asymptotically, see Barndorff-Nielsen et al. (2008-42).

For our dataset Table 2 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 $r_{t,CtC}^2$. 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 $r_{t,OtC}^2$ is of the same order as RK_t and RV_t. This implies that the more noisy[11] $r_{t,CtO}^2$ severely influences the size of $r_{t,CtC}^2$. 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.

[11]The CtO measures are more noisy because markets are not open 24 hours a day and 7 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 modeling 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. (2011). 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

$$r_{t,CtC} = e^{\frac{1}{2}h_t}\epsilon_t, \quad t = 1, \ldots, T, \tag{25}$$

where ϵ_t is either: $\epsilon_t \xrightarrow{d} \mathcal{N}(0,1)$ (Normal) or $\epsilon_t \xrightarrow{d} \sqcup(\nu)$ (Student-t) or $\epsilon_t \xrightarrow{d} \mathcal{S}\|\rceil\sqsupseteq - \sqcup(\lambda, \nu)$ (Hansen Skewed-t[12]). 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 (35) is given by

$$\mathbb{V}ar[r_{CtC,t+1}|\mathcal{F}_t] = \exp(h_{t+1}) = \exp\left(\alpha_0 + \sum_{p=1}^{P}\alpha_p r^2_{CtC,t-p+1} + \sum_{q=1}^{Q}\beta_q h_t\right) \tag{26}$$

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

$$h_{t+1} = \alpha_0 + \alpha_1 r^2_{CtC,t} + \beta h_t + \tau TM_t. \tag{27}$$

For the GARCH model we assume ϵ_t in (35) 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 (33) 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 (35), the HEAVY model of Shephard and Sheppard (2010) is defined as a system of two equations

$$\mathbb{V}ar[r_{CtC,t+1}|\mathcal{F}_t^{HF}] = \exp(h_{t+1}) = \exp(\alpha_{h,0} + \alpha_{h,1}RM_t + \beta_h h_t) \tag{28}$$

$$\mathbb{E}[RM_{t+1}|\mathcal{F}_t^{HF}] = \mu_{t+1} = \alpha_{\mu,0} + \alpha_{\mu,1}r_t^2 + \beta_\mu h_t \tag{29}$$

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

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

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

$$h_{t+1} = \alpha_{h,0} + \alpha_{h,1} RK_t + \beta_h h_t + \tau TM_t, \tag{30}$$

under the restriction $\beta_h \in (0,1)$. For the HEAVY-r model we assume ϵ_t in (35) 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 earlier, 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). We extend our GARCH model in the spirit of Glosten et al. (1993). Doing so, (33) becomes

$$h_{t+1} = \alpha_0 + \alpha_1 r^2_{CtC,t} + \beta h_t + \tau TM_t + \gamma r^2_{CtC,t} \mathbf{1}_{(r_{CtC,t}<0)}, \tag{31}$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function. We name this model the GJRGARCH[13] model. Analogously to the GARCH case, for the GJRGARCH model we assume ϵ_t in (35) 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 (30). Barndorff-Nielsen et al. (2008-42) found log-likelihood improvements by also including the Bipower variation estimate BPV_t. Therefore, we propose to extend (30) with the Bipower downward variation estimate $BPDV_t$ which includes both the realised semivariance estimate RSV_t^- and the Bipower variation estimate BPV_t. The model is given by

$$h_{t+1} = \alpha_{h,0} + \alpha_{h,1} RK_t + \beta_h h_t + \tau TM_t + \gamma_h BPDV_t, \tag{32}$$

We name this the LHEAVY-r model. Analogously to the HEAVY case, for the LHEAVY-r model we assume ϵ_t in (35) 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 $\widetilde{AS}_{t|t}$[14]. In the spirit of naming cross-sectional variable augmented GARCH models, see Engle (2002), we name the news sentiment augmented model a GARCHX model[15]. We define it as follows

$$h_{t+1} = \alpha_0 + \alpha_1 r^2_{CtC,t} + \beta h_t + \tau TM_t + \phi \widetilde{AS}_{t|t}. \tag{33}$$

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

$$h_{t+1} = \alpha_{h,0} + \alpha_{h,1} RK_t + \beta_h h_t + \tau TM_t + \phi \widetilde{AS}_{t|t}. \tag{34}$$

For both the GARCHX and the HEAVYX-r models we assume ϵ_t in (35) 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,

[13] The abbreviation GJR stands for Glosten Jagannathan Runkle, the authors of Glosten et al. (1993).

[14] See Sections 2 and 6.1 for the exact definition of $\widetilde{AS}_{t|t}$

[15] 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.

GARCHX-skewt and HEAVYX-rskewt respectively. In the same way the GJRGARCHX and LHEAVYX models represent the news sentiment augmented equivalents of the GJRGARCH (31) and LHEAVY (32) 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 we 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 setup of our forecasting study is similar to the work of Andersen et al. (1999), Martens (2002), Hansen and Lunde (2005b) 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 January 2 2006 and ends at January 2 2007. A one-step-ahead volatility forecast $\hat{\sigma}_k^2$ is computed for January 3 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 January 3 2006 and ends at January 3 2007. Again, a one-step-ahead volatility forecast $\hat{\sigma}_k^2$ is computed for January 4 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

$$\hat{\sigma}_{k,m}^2 = \mathbb{E}\left[\sigma_m^2 | \mathcal{F}_{m-250,m-1}, \hat{\Psi}\right], \tag{35}$$

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 (2005b) propose similar scaling methods. We follow the method of Hansen and Lunde (2005b) 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

$$\tilde{\sigma}_m^2 \equiv \hat{c} RK_m, \quad \text{where} \quad \hat{c} \equiv \left(\frac{T^{-1} \sum_{t=1}^{T} (r_{CtC,t} - \hat{\mu})^2}{T^{-1} \sum_{t=1}^{T} RK_t} \right), \tag{36}$$

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 (2005b) 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 $\tilde{\sigma}^2_m$, 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 3. Equal and similar loss functions are also considered by Bollerslev et al. (1994), Andersen et al. (1999), Martens (2002) and Hansen and Lunde (2005b).

Following the work of Koopman et al. (2005) and Hansen and Lunde (2005b), we adopt the robust superior predictive ability (SPA) test of Hansen (2005a) 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

Table 3: **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 Absolut Error	$L_{4,k,m} = \lvert 1 - \tilde{\sigma}^{-2}_{k,m}\hat{\sigma}^2_{k,m}\rvert$

and perform the SPA test procedure for each group group separately. For each group we have $N + 1$ different models $\mathcal{M}_n$ for $n = 0, 1, \ldots, N$[16]. For each model $\mathcal{M}_n$ we have M volatility forecasts $\hat{\sigma}^2_{n,m}$ for $m = 1, \ldots, M$. For every forecast we calculate the loss function $L_{i,m,n}$ as given in Table 3 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

$$X_{n,m} \equiv L_{i,m,0} - L_{i,m,n}, \tag{37}$$

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

$$\overset{max}{n \neq 0}\, \lambda_n \leq 0. \tag{38}$$

Hansen (2005a) proposed the associated SPA test statistic

$$T = \overset{max}{n \neq 0}\, \frac{\sqrt{M} X_n}{\hat{\omega}_{n,n}}, \tag{39}$$

where $\hat{\omega}^2_{n,n}$ is a consistent estimate of $\omega^2_{n,n}$, and where $X_n = M^{-1} \sum_{m=1}^{M} X_{n,m}$ and $\omega^2_{n,n} \xrightarrow{a} \mathbb{V}ar[\sqrt{M} X_n]$. A consistent estimator of $\omega^2_{n,n}$ 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 5 trading days which corresponds to $q = 1/5$. Since we perform the same bootstrap procedure as proposed by Hansen (2005a) and performed by Hansen and Lunde (2005b) 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 $(\bar{s}^{pos}_d, \bar{s}^{neu}_d, \bar{s}^{neg}_d)$

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

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[17].

We estimated the total set consisting of $D = 841,536$ news sentiment articles $\bar{X}_d$ on the 5-minute time grid from the beginning of 2003 to the end of 2010. The estimation results are presented in Table 4. 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

Table 4: **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 2 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

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 natural gas futures price dynamics. As mentioned in Section 3 our data is based on 1,257 trading days from January 3 2006 to December 31 2010. Therefore we construct news sentiment levels at closing time C for each trading day t. Specifically, we estimate the news sentiment level from the beginning of 2003 to time d, where d is equal to closing time C of trading day t. This means that we estimate the the news sentiment levels 1,257 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[18]. 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 5. The summary statistics show that the low frequent Kalman filtered news sentiment levels $\widetilde{S}^p_{C,t|t}$ show similar statistics as their high frequent equivalents $\widehat{S}^p_{d|D}$. Also, the $\widetilde{S}^p_{C,t|t}$ show very high autocorrelation which imply forecasting abilities.

Figure 3 shows plots of $\widetilde{S}^p_{d|D}$ and $\widehat{S}^p_{d|D}$ from January 3 2006 to December 31 2010. Figure 4 shows the same plots but from October 18 2008 to 31 October 2008. Additionally, Figures 5 and 5 show plots of $\widetilde{S}^p_{C,t|t}$ and $\widetilde{AS}_{C,t|t}$ together with natural gas futures return and realized measures which are described in Section 3.

6.2. Events and Granger causality

In this section we perform two methods to analyze 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

[17] The procedure is implemented in the program environment Ox of Doornik (2001) and makes use of the SsfPack 2.2 of Koopman et al. (1999).
[18] Note that $p \in \{pos, neu, neg\}$.

Table 5: **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					

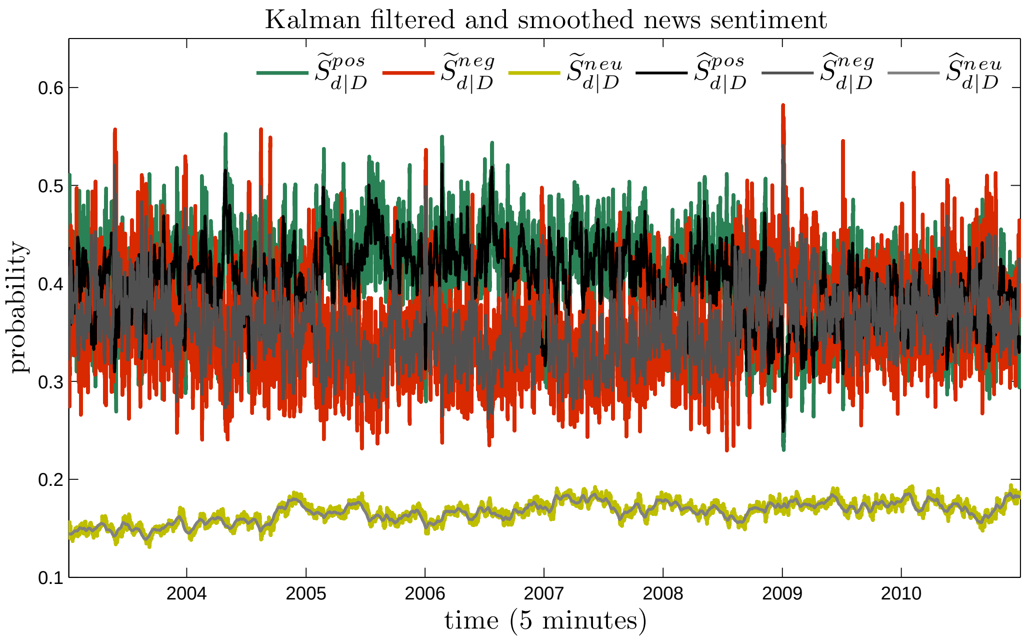

Figure 3: Kalman filtered and smoothed news sentiment levels.

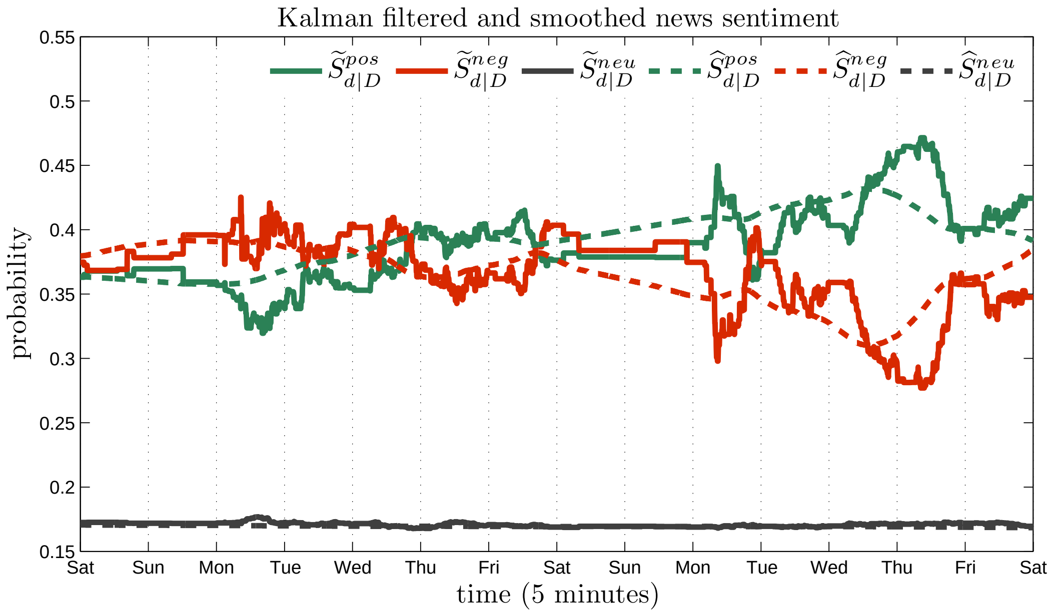

Figure 4: Kalman filtered and smoothed news sentiment levels.

sentiment events. The second method involves Granger causality tests to determine causal relationships between natural gas price dynamics and news sentiment measures.

6.3. Event study

We employ event studies as described in MacKinlay (1997) and used in Tetlock et al. (2008). 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 $\widetilde{S}^{neg}_{c,t|t}$ as an extreme negative news event. All days which are not selected as extreme positive or negative news events are defined as extreme neutral news events.

We analyze 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 1,256 trading days and select 252 extreme postive and extreme negative event days and a total of 753 extreme neutral event days. This results in 252 extreme positive and extreme negative event cumalative return paths and 753 extreme 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 7 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 shows a mean reverting effect in the return evolution around the event day. For extreme negative news the return evolution shows strong positive autocorrelation around the event day. As expected, the evolution of the returns around the extreme neutral days are not different from zero[19].

Interesting to see is that the frequency of extreme positive days decreases to a bottom at 2009 while the frequency of extreme negative news increases from 2008[20].

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

[20]This can be related to the credit crunch of 2008 to 2009.

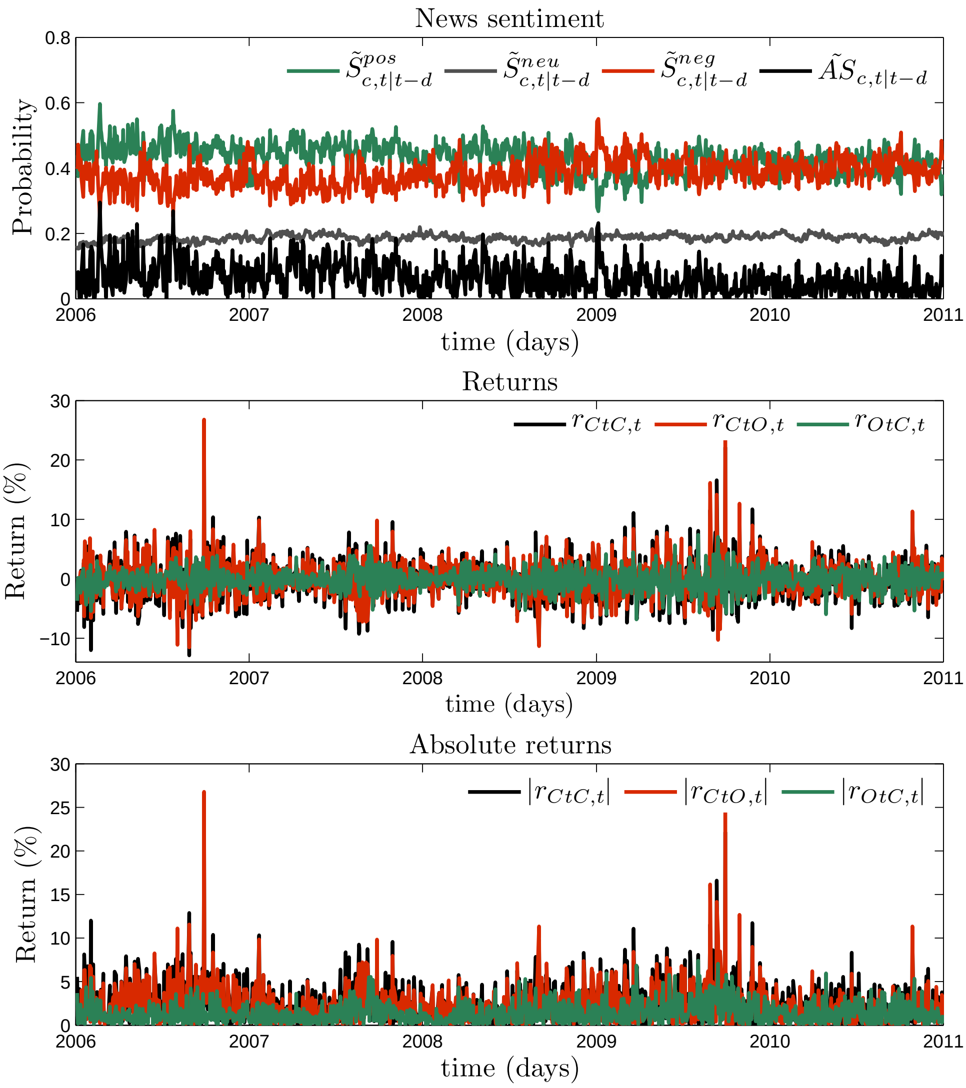

Figure 5: Kalman filtered news sentiment measures and natural gas return measures.

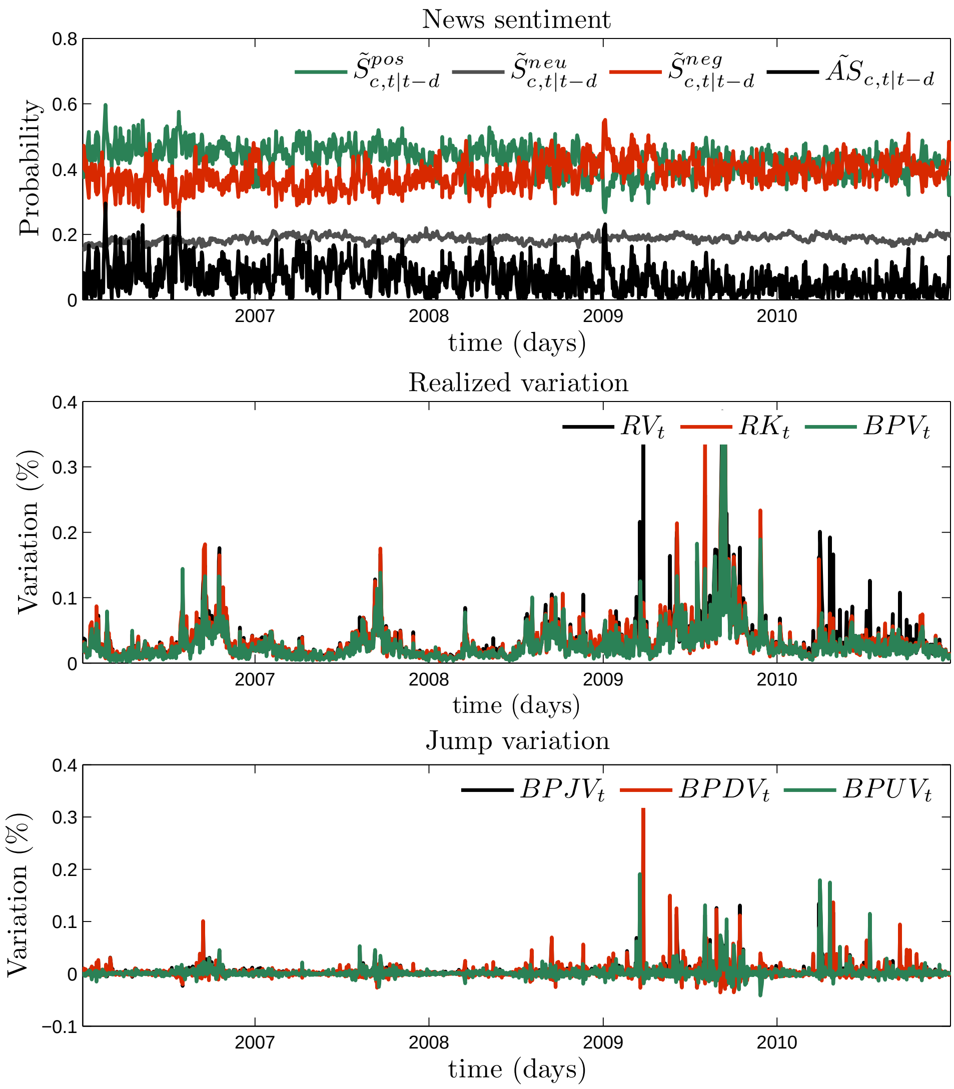

Figure 6: Kalman filtered news sentiment measures and natural gas realized measures.

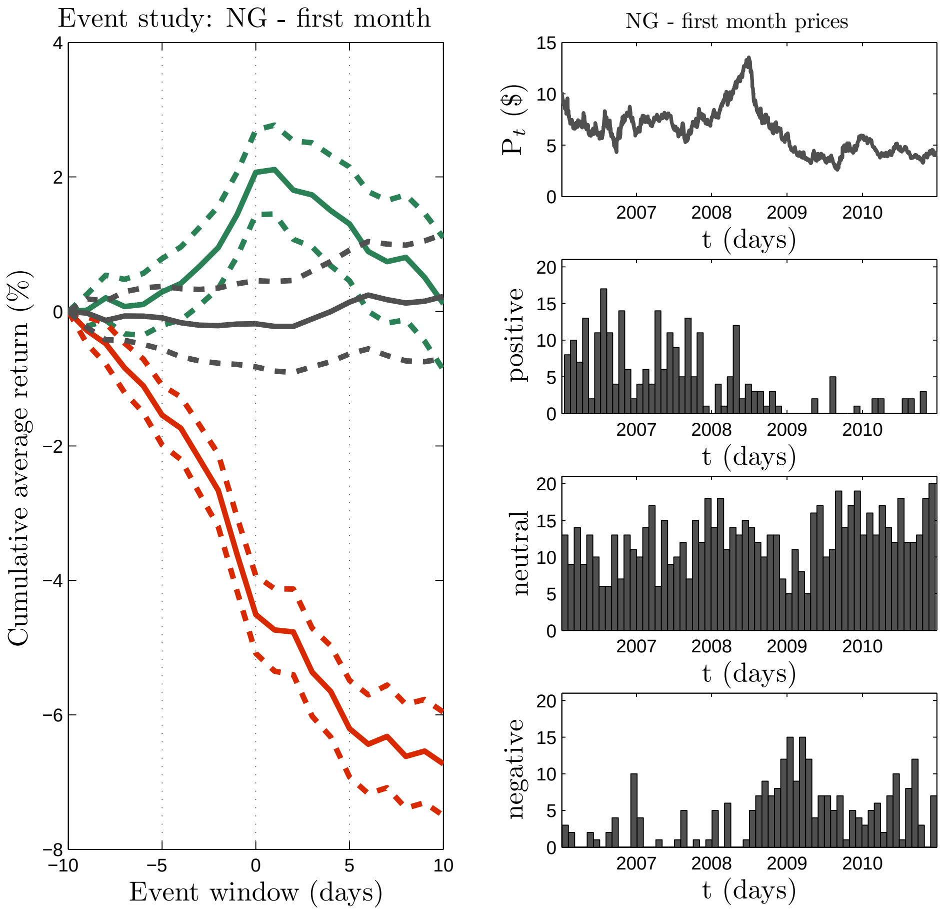

Figure 7: Cumulative return evolutions around extreme positive, extreme neutral and extreme negative event days are plotted based on a 10 day event window around the event day and a threshold q of 80%. Also the natural gas futures prices are plotted from the beginning of 2006 to the end of 2010. Furthermore, the monthly event frequencies are are shown for all different event types.

Another interesting result showed in Figure 7 is the possibility to make a significant profit from selling the futures contract at an extreme positive or negative event day[21]. Unfortunately, the event type frequencies show that extreme events are clustered over time. This questions the robustness of the suggested trading strategy. However, the event study does show significant discriminating power of news sentiment. This is even more clear for the event studies presented in Figure 8. These event studies have the same setup as for the event study in Figure 7 but for a q threshold 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 continues after extreme negative days, before we observe a return to fundamentals.

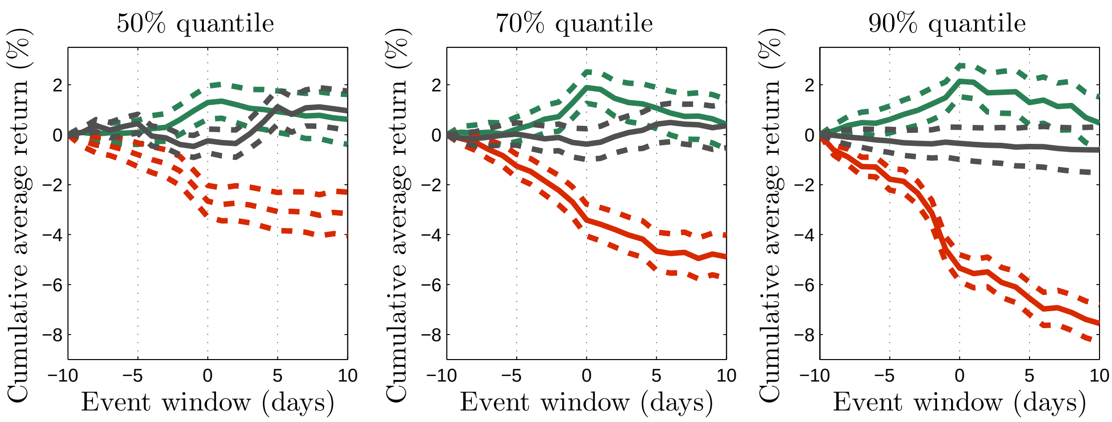

Figure 8: Cumulative return evolutions around extreme positive (green), extreme neutral (grey) and extreme negative (red) event days are plotted based on a 10 days event window around the event day and threshold q of 50%, 90% and 90%.

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 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 Granger causality tests as first suggested by Granger (1969). A similar investigation was performed by

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

$$\mathbf{Z}_t = \mathbf{\Phi}_0 + \sum_{p=1}^{P} \mathbf{\Phi}_p \mathbf{Z}_{t-p} + \mathbf{u}_t, \tag{40}$$

where $\mathbb{E}[\mathbf{u}_t] = 0$, $\mathbb{E}[\mathbf{u}_t \mathbf{u}_t'] = \Omega$, $\mathbf{Z}_t$ is a (2×1) vector, $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_p$ 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). For the bivariate VAR of order P case, the AIC is defined as

$$AIC = \log|\widehat{\Omega}| + \frac{2(2^2 P + 2)}{T}, \tag{41}$$

[21] Here we assume no transaction costs and zero market impact.

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 2 and 5. 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 significant Granger causality relationships. First of all, absolute news sentiment Granger causes $R_{CtC,t}$ and $R_{CtO,t}$. The same applies to $R_{OtC,t}$ but only for a 10% significance level. If we look at the squared returns we see an equal pattern for all news sentiment measures. Specifically, highly significant Granger causal relationships are shown for news sentiment measures on $R^2_{CtC,t}$ and $R^2_{CtO,t}$ but hardly any on the $R_{OtC,t}$.

Given that the CtC return measures are sums of the CtO and OtC return measures, this implies that the arrival of news during non-trading periods have a large effect on the overnight CtO returns.

Second of all, if we look at the realized variation measures we see cross causality relationships between news measures and the realized kernel RK_t and 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 by negative semivariance RSV_t^-. Furthermore, we see that the jump variation robust Bipower variaton 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 Granger causality relationships between jump variation and news sentiment. Most statistics show cross correlation relationships, but the jump variation measure $BPJV_t$ is only caused by all news sentiment measures and it only Granger causes effects in the absolute news sentiment measure. The same applies to $BPUV_t$ but it does only shows Granger causal effects in the positive and negative news sentiment measures.

This difference between $BPJV_t$ and $BPUV_t$ shows that 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 this we can learn that volatility, and especially negative volatility, Granger causes news sentiment. Furthermore, news sentiment also Granger causes volatility. This implies that news is caused by volatility in the market but also that market participants trade as some function of aggregated news.

Also, the different Granger causality relations between (absolute) news sentiment and the jump variation measures show the importance of asymmetric returns and jumps on news. Specifically, news sentiment in an absolute sense is more sensitive to negative jumps in the market than by jumps in general. However, we can state 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.

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 7. Likewise, the estimation results from the HEAVY-r type models are described in Table 8. These estimation results are based on the total dataset from the beginning of 2006 to the end of 2010, see Section 3 for specifics. All programs 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 γ estimates show the tremendous significance of the so-called leverage effects in the GJRGARCH and LHEAVY type volatility models.

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

Table 6: **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 4 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).**

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

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

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

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

Table 7: **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 2 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

Table 8: **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 2 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

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 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 $\widetilde{AS}_{c,t|t}$. The GARCHX type models show very low significant positive relationships between the volatility forecast and the news sentiment variable. Moreover, the log-likelihood is only slightly increased when including news sentiment to GARCH type models[22].

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 less influenced by news than $r^2_{CtO,t}$. 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 $r^2_{CtC,t}$ and $r^2_{CtO,t}$.

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 $r^2_{CtC,t}$ 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 9 presents the logarithms of the volatility forecasts versus the realized forecast target $\tilde{\sigma}^2_m$ for GARCH and HEAVY-r type models with and without news sentiment[23]. 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 9 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 Absolut 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

[22]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.

[23]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 make the plots more clear.

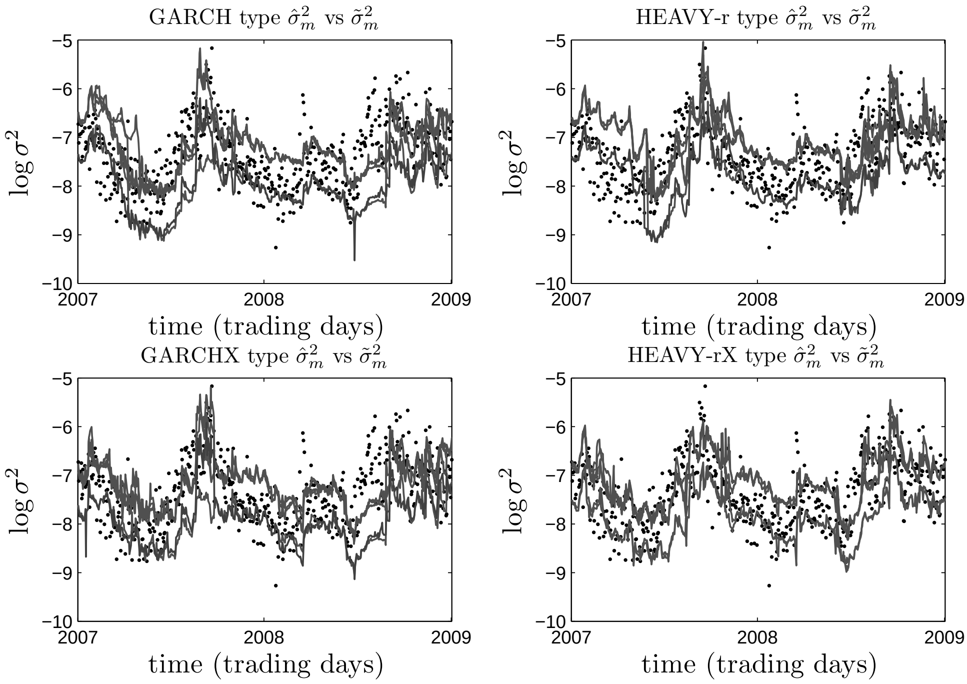

Figure 9: A subset of the forecasts and the forecasting targets are presented from the beginning of 2007 to the end of 2008. The logarithms of the forecasting target $\tilde{\sigma}^2_m$ are plotted as black dots. The logarithms of the forecasts $\hat{\sigma}^2_m$ are plotted as gray lines. Specifically, the forecasts are plotted for GARCH, GARCHX, HEAVY-r and HEAVY-rX model types.

(2005b) 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 (2005b), we perform superior predictive ability (SPA) tests of Hansen (2005a) as described in Section 5.

The SPA test results are presented in Table 10. Specifically, the Hansen SPA p-values of Hansen (2005a) 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 analyze 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.

Table 9: **The Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Heteroskedasticity Adjusted Squared Error (MHASE), Mean Heteroskedasticity Adjusted Absolut 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

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 are *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.

Table 10: **The table presents the Hansen consistent SPA p-values of Hansen (2005a) based on the Squared Error (SE), Absolute Error (AE), Heteroskedasticity Adjusted Squared Error (HASE) and Heteroskedasticity Adjusted Absolut 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 M_g	SE	AE	HASE	HAAE	Base model 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

7. Conclusion

7.1. summary

We investigated the impact of TRNAE derived news sentiment on the dynamics of daily natural gas futures prices traded on the New York Mercantile Exchange (NYMEX). We implemented the proposed LNSL model and constructed *autocorrelated* news sentiment probabilities (news sentiment) which conveys a positive, neutral or negative outlook on natural gas prices, based on a 5-minute time grid from the beginning of 2003 to the end of 2010. Additionally, we constructed several return and variation measures to proxy for the dynamics of first month natural gas futures prices.

To analyze the impact of news sentiment on the dynamics of daily natural gas futures prices, we employed event studies and Granger causality tests.

We found that the price evolution of first month natural gas futures contracts shows a mean reverting effect around days which we refer to as extreme positive sentiment days. Additionally, we found that the price evolution around extreme negative sentiment days shows negative price momentum which strongly continues after the event day before we observe a return to fundamentals. From this we conclude that there is a significant relationship between news sentiment and the evolution of natural gas futures returns.

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 in the market. Also, we found that news sentiment, in an absolute sense, is more sensitive to negative jumps in the market than by jumps in general, including positive jumps. However, we found 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.

Finally, 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 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 volatility (HEAVY) models of Shephard and Sheppard (2010) and Noureldin et al. (2011). By augmenting all models with a news sentiment variable we tested the hypothesis if including news sentiment to volatility models results in superior volatility forecasts. Here we followed the forecasting study setup of of Hansen and Lunde (2005b) and Koopman et al. (2005) and conduct Superior Predictive Ability tests of Hansen (2005a) to test our hypothesis.

We found significant evidence that including news sentiment to volatility models results in superior volatility forecasts.

7.2. Final remarks

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 articles can be seen as draws from a trinomial distribution. For further research, we suggest to model the unobserved news sentiment by modeling the time varying trinomial distribution, for details see Durbin and Koopman (2001).

Also, it might be interesting to analyze the impact of news sentiment on more individual futures contracts or the whole forward curve. However, the latter is hard to analyze since natural gas is subject to seasonal effects, see Borovkova and Geman (2006), and the forward curve is only liquid up to contracts which mature longer than a year from 2006.

Finally, it is of great interest to conduct an equal volatility forecasting study based on a longer forecasting horizon and for more volatility models, especially of the fractional integrated type.

References

Andersen, T., Bollerslev, T., Diebold, F., Ebens, H., 2001. The distribution of realized stock return volatility. Journal of Financial Economics. 61(1), 43–76.

Andersen, T., Bollerslev, T., Lange, S., 1999. Forecasting financial market volatility: sample frequency vis-a-vis forecast horizon. Journal of Empirical Finance. 6, 457–477.

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

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

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

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

Barndorff-Nielsen, O., Shephard, N., 2002. Econpmetric analysis of realised volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society. Series B 64, 253–280.

Barndorff-Nielsen, O., Shephard, N., 2004. Power and bipower variation with stochastic volatility and jumps (with discussion). Journal of Financial Econometrics. 2, 1–48.

Barndorff-Nielsen, O., Shephard, N., 2008. Econometrics of testing for jumps in financial economics using bipower variation. Journal of Financial Econometrics. 4(1), 1–30.

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

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science. 2(1), 1–8.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics 69, 542–547.

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

Borovkova, S., 2004. The forward curve dynamic and market transition forecasts. In D.W. Bunn (Ed.), Modeling prices in competitive electricity markets, John Wiley Sons, Ltd. p. 24.

Borovkova, S., Geman, H., 2006. Seasonal and stochastic effects in commodity forward curves. Review of Derivatives Research, Springer 9 (2), 167–186.

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

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

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

Doornik, J., 2001. Ox: Object orientated matrix programming, 3.0. Tmberlake Consultants Press, London.

Durbin, J., Koopman, S., 2001. Time Series analysis by State Space Methods. Oxford: Oxford Statistical Science Series.

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

Engle, R., 2002. New frontiers for arch models. Journal of Applied Econometrics. 17, 425–446.

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

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

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

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

Granger, C., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 37 (3).

Greene, W., 2003. Econometric Analysis. New Jersey: Prentice Hall.

Hansen, B., 1994. Autoregressive conditional density estimation. International Economic Review. 35(3).

Hansen, P., 2005a. A test for superior predictive ability. Journal of Business and Economic Statistics. 23, 365–380.

Hansen, P., Lunde, A., 2005b. A forecast comparison of volatility models: does anything beat a garch(1,1)? Journal of Applied Econometrics. 20, 873–889.

Hansen, P., Lunde, A., 2006. Realized variance and microstructure noise. Journal of Business Economic Statistics. 24, 127–161.

Jacod, J., Li, Y., Mykland, P., Podolski, M., Vetter, M., 2009. Microstructure noise in the continuous case: the pre-averaging approach. Stochastic processes and their applications. 119, 2249–2276.

Kalman, J., 1960. A new approach to linear filtering and prediction problems. Basic engineering ,Transactions ASMA, Series D 82, 35–45.

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

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

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

MacKinlay, C., 1997. Event studies in economics and finance. Journal of Economic Literature. 35, 13–39.

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

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

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

Nickerson, R., 2009. Mathematical reasoning. patterns, problems, conjectures and proofs. Psychology Press, London.

Nofsinger, J., 2005. Social mood and financial economics. Journal of Behavioral Finance. 6(3), 144–160.

Noureldin, D., Shephard, N., Sheppard, K., 2011. Multivariate high-frequency-based volatility (heavy) models. Journal of Applied Econometrics.

Pagan, A., Schwert, G., 1990. Alternative models for conditional volatility. Journal of Econometrics. 45, 267–290.

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

Politis, D., Romano, J., 1994. The stationary bootstrap. Journal of the American Statistical Association. 89, 1303–1313.

Reilly, M., Posadas-Sanchez, D., Kettle, L., Killeen, P., 2011. Making the trip worthwhile: do rats (rattus norvegicus) and pigeons (columba livia) forage prospectively?. Behavioural Processes (under review).

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

Samuelson, P., 1965. Proof that properly anticipated prices fluctuate randomly. Industrial management review. 6, 41–49.

Shephard, N., Sheppard, K., 2010. Realising the future: Forecasting with high-frequency-based volatility (heavy) models. Journal of Applied Econometrics. 25, 197–231.

Smith, V., 2003. Constructivist and ecological rationality in economics. American Economic Review. 93, 465–508.

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

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

Tsay, R., 2005. Analysis of financial tiem series second edition. John Wiley Sons, Inc. Hoboken, NJ.

Uttal, W., 2008. Time, space and number in physics and psychology. Sloan publishing, Cornwall-on-hudson, NY.

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

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