Haug, Czado:

Mixed effect model for absolute log returns of ultra high frequency data

Sonderforschungsbereich 386, Paper 440 (2005)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Mixed effect model for absolute log returns of ultra high frequency data*

Stephan Haug †‡ Claudia Czado†

September 5, 2005

Summary

Considering absolute log returns as a proxy for stochastic volatility, the influence of explanatory variables on absolute log returns of ultra high frequency data is analysed. The irregular time structure and time dependency of the data is captured by utilizing a continuous time ARMA(p,q) process. In particular we propose a mixed effect model for the absolute log returns. Explanatory variable information is used to model the fixed effects, whereas the the error is decomposed in a non-negative Lévy driven continuous time ARMA(p,q) process and a market microstructure noise component. The parameters are estimated in a state space approach. In a small simulation study the performance of the estimators is investigated. We apply our model to IBM trade data and quantify the influence of bid-ask spread and duration on a daily basis. To verify the correlation in irregularly spaced data we use the variogram, known from spatial statistics.

Keywords: ultra high frequency, CARMA, mixed effect model, state space, Kalman filter, variogram

*This research was supported by the German Science Foundation, Sonderforschungsbereich 386.

†Center for Mathematical Sciences, Munich University of Technology, Boltzmannstr. 3, D-85747 Garching, Germany. Email: haug@ma.tum.de, cczado@ma.tum.de.

‡Correspondence to: Stephan Haug, Phone: ++49/89/28917040, Fax: ++49/8928917035.

1 Introduction

Efficient estimation of stochastic volatility is vital for risk management and option pricing. We are interested in providing such estimates using all available data, allowing for explanatory variables and accounting for market micro structures. For this we use ultra high frequency (uhf) financial data. The term uhf data was defined by Engle [1]. He calls financial data uhf data, if they consist of all transactions and quotes recorded during the trading day. The recorded transactions of course do not take place at regularly spaced time points, i.e. we have to analyse irregularly spaced time series. One way would be to sample it at a given frequency, but this results in a loss of information. Therefore we setup a model directly dealing with this irregular time spacing. Our object of interest will be the absolute log return, which is a proxy for the unobservable instantaneous standard deviation σ_{t_i}, where t_i is the time of the i-th trade, of the log price $S_{t_i} = \log(P_{t_i})$. By modeling the mean of the absolute log returns, we get a model-based estimate of the instantaneous standard deviation. This could then be used for example, like in Jungbacker and Koopman [2], to estimate actual volatility of the interval $[t_i, t_j]$, $j > i$, given by

$$\sigma^{*2}(t_i, t_j) = \int_{t_i}^{t_j} \sigma_t^2 dt$$

based on all available information. Here it is important to account for microstructure noise, when dealing with ultra high frequencies. The problem of market microstructure noise at this frequency is for example explained in Aït-Sahalia, Mykland and Zhang [3]. It is more common to account for microstructure effects on the return level, while we will account for these effects on the absolute log return scale. This is more appropriate in the context of the regression setup we follow for the absolute log returns. The absolute log-return $|S_{t_i} - S_{t_{i-1}}|$ will be modeled in this paper given the past information $\mathcal{G}_{t_{i-1}} = \sigma(S_{t_j}, d_{t_j}\,; j \leq i-1)$ and current duration $d_{t_i} = t_i - t_{i-1}$. Since the duration process is a stochastic process itself one also needs a model for this regularly spaced (measured in tick time) time series. A popular model for the durations given the past information, called *Autoregressive Conditional Duration* (ACD) model, has been proposed by Engle and Russell [4]. There are a number of modifications of the ACD model, which are described for example in Bauwens, Giot, Gramming and Veredas [5].

To cope with the problem of unequally spaced data, we will assume a continuous time parameter price process. The absolute log returns will be the response in a

regression framework with the current duration as one of the explanatory variables and correlated residuals. They have the correlation structure of a continuous time ARMA process. The estimation of correlation for unequally spaced time series is problematic, since e.g. the sample autocorrelation function can not be estimated directly. We compute the sample variogram, which is defined in terms of increments and therefore adequate for irregularly spaced observations. We have already said, that the absolute log return is viewed in this paper as a noisy measure of instantaneous volatility. It can be decomposed into a fixed effect, a random effect and a measurement error. The fixed effect describes the time dependent mean of the data, whereas the random effect specifies the correlation structure. Since the fixed effect is a function of time varying explanatory variables it allows for time of day effects (see for example Bauwens and Giot [6]). The measurement error accounts for the market microstructure noise on this absolute return level. The presence of microstructure effects also allows us to assume the mean function to be a continuous variable, despite the fact that the prices are multiple of one hundreds of a dollar. The return of irregularly spaced transaction data is also modeled as a continuous variable for example in Meddahi, Renault and Werker [7], whereas Engle and Russell [8] or Liesenfeld and Pohlmeier [9] assume that it takes on only countably many values. The influence of the explanatory variables will be modeled in a parametric way, which allows us to compute predictions based on past information and current duration in a very easy way. By using the mean squared error as scoring rule, we are able to quantify the loss in predictive power, when duration is not used as a explanatory variable. Here we would like to mention, that initially we are interested in detecting certain dependencies between the response and the explanatory variables. In a further step one could think about additionally applying an ACD model to compute predictions in real applications. Visualisation of the explanatory variable effect on the absolute log returns on a daily basis is also possible. Renault and Werker [10] studied the instantaneous causality effect from transaction durations to price volatility and found significant empirical evidence for it. There are also further regression models with measures of volatility as response. Corsi [11], Anderson, Bollerslev and Diebold [12] and Ghysels, Santa-Clara and Valkanov [13] have setup different kinds of linear regression models with for example realized volatility (see Barndorff-Nielsen and Shepard [14]) as response. An overview over these three models can be found in Forsberg and Ghysels [15]. As we have already mentioned, Jungbacker and Koopman [2] estimated actual volatility of ultra-high frequency data in a model-based approach.

They considered a state space model for the return process, which is defined for every second. This leads to a missing values problem. We also used a state space approach, but rather prefer to work with time dependent matrices, to account for the irregular time spacing, than to deal with a large number of missing values per day.
The paper is organized as follows. In Section 2 we will setup our model for absolute log-returns. The estimation of the model parameters will be explained in Section 3. The performance of the estimates from Section 3 will be tested in a simulation study in Section 4. Section 5 shows an application of our model to IBM transaction data from the NYSE. The last section gives a summary and draws conclusions.

2 A mixed effect regression model for irregularly spaced data

The main characteristic of the data we deal with is that we have observations at irregularly spaced time points. Therefore we think it is natural to assume, that these observations are observations from a continuous time model. It is common practice to model the volatility of high frequency data as a continuous time linear process (see for example Barndorff-Nielsen and Shepard [16] or Jungbacker and Koopman [2]). Since the absolute log return is a measure of the instantaneous standard deviation, we will model them in such a way, that they have the autocorrelation structure of a continuous time linear process. To be precise, we assume the autocorrelation structure of a continuous time ARMA(p,q) process, henceforth called CARMA(p,q) process.

2.1 Second order Lévy driven CARMA(p,q) process

A second order Lévy driven CARMA(p,q) process $Y := (Y_t)_{t\geq 0}$ is defined (see Brockwell and Marquardt [17]) in terms of the following state-space representation of the formal equation,

$$a(D)Y_t = b(D)DL_t, \qquad t \geq 0, \tag{2.1}$$

in which D denotes differentiation with respect to t, $L := (L_t)_{t\geq 0}$ is a Lévy process (see for example Applebaum [18]) with $\text{var}(L_1) < \infty$,

$$\begin{aligned} \textit{autoregressive polynomial:} \quad & a(z) := z^p + a_1 z^{p-1} + \cdots + a_p, \\ \textit{moving-average polynomial:} \quad & b(z) := 1 + b_1 z + \cdots + b_{p-1} z^{p-1}, \end{aligned}$$

and the coefficients b_j satisfy $b_q \neq 0$ and $b_j = 0$ for $q < j < p$. It is assumed that $a(z)$ and $b(z)$ have no common factors. The state-space representation consists of the

$$\textit{observations equation:} \quad Y_t = b^T \mathbf{W}_t, \tag{2.2}$$

and

$$\textit{state equation:} \quad d\mathbf{W}_t - A\mathbf{W}_t dt = \mathbf{1}_p dL_t, \tag{2.3}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_p & -a_{p-1} & -a_{p-2} & \cdots & -a_1 \end{bmatrix}, \ \mathbf{1}_p = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} 1 \\ b_1 \\ \vdots \\ b_{p-2} \\ b_{p-1} \end{bmatrix}.$$

The state equation is therefore a system of linear stochastic differential equations (see for example Applebaum [18] for details on stochastic differential equations).

Definition 2.1.

If the real part of the roots $\lambda_1, \ldots, \lambda_p$ of the autoregressive polynomial $a(z)$ is negative and $\mathbf{W}_0$ is independent of the driving Lévy process L, with $\mathbb{E}(L_1^2) < \infty$, then the process

$$Y_t = b^T \mathbf{W}_t,$$

where

$$\mathbf{W}_t = e^{At}\mathbf{W}_0 + \int_0^t e^{A(t-u)}\mathbf{1}_p dL_u,$$

i.e.

$$Y_t = b^T e^{At}\mathbf{W}_0 + \int_0^t b^T e^{A(t-u)}\mathbf{1}_p dL_u, \tag{2.4}$$

is called CARMA(p,q) process with finite second moment.

Remark 2.1.

(i) The exponential matrix e^{Au} *is defined by* $e^{Au} := \sum_{k=0}^{\infty} \frac{(Au)^k}{k!}$.

(ii) If $\mathbf{W}_0$ *has the same distribution as* $\int_0^\infty e^{Au}\mathbf{1}_p dL_u$*, then the CARMA(p,q) process (2.4) is a strictly stationary process.*

(iii) The CARMA(p,q) process $(Y_t)_{t\geq 0}$ *is a weakly stationary process, if* $\mathbf{W}_0$ *has the mean and covariance matrix of* $\int_0^\infty e^{Au}\mathbf{1}_p dL_u$*. The mean and autocovariance function of a weakly stationary CARMA(p,q) process* $(Y_t)_{t\geq 0}$ *are*

$$\mathbb{E}(Y_t) = -\mathbf{b}^T A^{-1}\mathbf{1}_p \mathbb{E}(L_1) \tag{2.5}$$

and

$$\operatorname{cov}(Y_t, Y_{t+h}) \;=\; \operatorname{var}(L_1)\mathbf{b}^T e^{Ah}\Sigma\mathbf{b}, \tag{2.6}$$

where $\Sigma := \int_0^\infty e^{As}\mathbf{1}_p\mathbf{1}_p^T e^{A^T s} ds$.

(iv) For a proof of (ii) and (iii) see Brockwell and Marquardt [17].

(v) Let M *be a second Lévy process independent of* L*, but with the same distribution, and define the following extension of* L*:*

$$L_t^* = L_t\chi_{[0,\infty)}(t) - M_{-t-}\chi_{(-\infty,0)}(t), \qquad -\infty < t < \infty,$$

where M_{t-} *denotes the left limit of* M *at* t *and* χ_A *is the indicator function of the set* A*. Then the process* $Y := (Y_t)_{t\in\mathbb{R}}$ *defined by*

$$Y_t = \int_{-\infty}^{\infty} g(t-u)dL_u^*,$$

where

$$g(t) := \begin{cases} \mathbf{b}^T e^{At}\mathbf{1}_p & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2.7}$$

is a solution to (2.2) and (2.3) (with L *replaced by* L^**). The function* g *is referred to as* the kernel of the CARMA(p,q) process Y*. For more details see Brockwell and Marquardt [17].*

(vi) Discrete time observations $(Y_{t_i}) := (Y_{t_i})_{i=1,\dots,n}$ *follow the discrete time state space model*

$$\begin{aligned} Y_{t_i} &= \mathbf{b}^T \mathbf{W}_{t_i} \\ \mathbf{W}_{t_i} &= e^{A(t_i - t_{i-1})} \mathbf{W}_{t_{i-1}} + \int_{t_{i-1}}^{t_i} \mathbf{b}^T e^{A(t_i - u)} \mathbf{1}_p dL_u. \end{aligned}$$

Example 2.1. *As an example consider the Lévy driven CARMA(2,1) process* Y*, where the driving Lévy process* L *is a compound Poisson process with gamma distributed jumps, i.e.*

$$L_t = \sum_{k=1}^{N_t} X_k.$$

Here (X_k) *are i.i.d. with density* $f(x) = \frac{100^2}{\Gamma(2)} x e^{-100x}$ *and* $N_t \sim Pois(t)$*.*
Since $\mathbb{E}(X_1) = 0.02$ *and* $\mathrm{var}(X_1) = 0.0002$ *we have* $\mathrm{var}(L_1) = \mathbb{E}(X_1)^2 + \mathrm{var}(X_1) = 0.0006$*. As autoregressive and moving-average polynomial of this CARMA(2,1) process we choose*

$$a(z) = z^2 + 8z + 4 \quad \textit{and} \quad b(z) = 1 + z.$$

1000 *observations at integer times of a simulated sample path can be seen in Figure 1.*

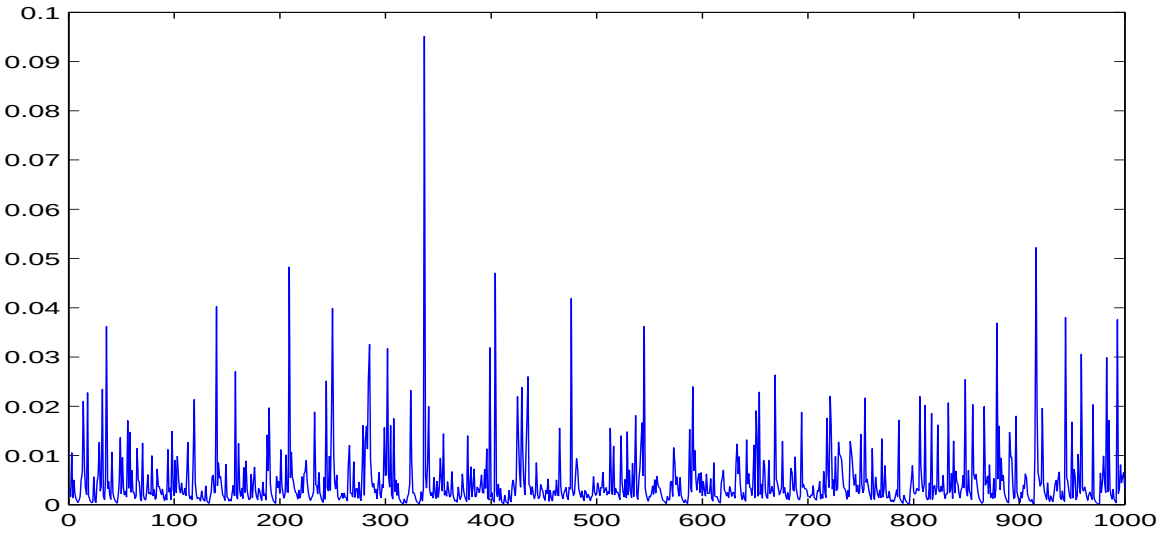

Figure 1: 1000 equidistant observations of the CARMA(2,1) process with $a(z) = z^2 + 8z + 4$ and $b(z) = 1 + z$ from Example 2.1

2.2 Regression mean specification

Ultra high frequency data exhibit some time of day effects (see for example Bauwens and Giot [6]), which result in a nonstationary time series. We try to explain these effects as being influenced by explanatory variables, which have time of day dependent values. In our setup these explanatory variable information is used to model the mean of the data,

$$\mu_{t_i} := \mathbb{E}(|r_{t_i}|),$$

with

$$|r_{t_i}| := |\log(P_{t_i}) - \log(P_{t_{i-1}})| \cdot 100, \qquad i = 1, \ldots, n, \tag{2.8}$$

where P_{t_i} is the stock price observed at time t_i, like in a typcial regression setup. There will be no assumption made about a stock price model, except that we assume, that it is a continuous time process. To assure positivity of the mean we will use a log-link, i.e.

$$\log(\mu_{t_i}) := \mathbf{x}_{t_i}^T \boldsymbol{\beta}, \qquad i = 1, \ldots, n, \tag{2.9}$$

with $\mathbf{x}_{t_i}^T \in \mathbb{R}^{1\times s+1}$ the i-th row of the design matrix

$$X = \begin{pmatrix} \mathbf{x}_{t_1}^T \\ \vdots \\ \mathbf{x}_{t_n}^T \end{pmatrix} \in \mathbb{R}^{n\times s+1}$$

and parameter vector $\boldsymbol{\beta}^T := (\beta_0, \ldots, \beta_s)^T \in \mathbb{R}^{s+1\times 1}$. As can be seen from (2.9), a parametric approach is taken. The specific structure of the design matrix will be discussed in the applications. Potential explanatory variables are

$$\begin{aligned} b_{t_i} &:= \text{the last bid-ask spread before time } t_i \\ d_{t_i} &:= \text{the duration } t_i - t_{i-1} \\ v_{t_i} &:= \text{the volume of the the last trade before time } t_i. \end{aligned}$$

The choice of explanatory variables will be discussed in the applications. The explanatory variable d_{t_i} is unknown before time t_i and has therefore to be estimated, by some *autoregressive conditional duration model*, if the model is used for prediction.

2.3 Correlated residuals

As we have said in the beginning we model the absolute log returns as an autocorrelated process. The question is if autocorrelation is really present in this uhf data. The answer to this question is part of the analysis. The problem with empirically estimating the autocorrelation in uhf data is the irregularly time structure. Therefore the empirical autocorrelation function can not be computed. One way out is to consider the variogram (it will be introduced and discussed in the appendix), which is defined for irregularly spaced data. But the variogram is also not defined for $(|r_{t_i}|)$, because the mean of the increments is not a linear function of the time lag, i.e. $\mathbb{E}(|r_t|-|r_s|) \neq C\cdot(t-s)$, which has to be the case. The variogram is however defined, when we consider the residuals

$$\varepsilon_{t_i} := |r_{t_i}| - \mu_{t_i}, \qquad i = 1, \ldots, n, \tag{2.10}$$

with $\mathbb{E}(\varepsilon_{t_i}) = 0$ and $\text{var}(\varepsilon_{t_i}) =: \sigma_\varepsilon^2$. The ε_{t_i} are autocorrelated because of the following assumption

$$\varepsilon_{t_i} =: Y_{t_i} + \tilde{\varepsilon}_{t_i}, \qquad i = 1, \ldots, n, \tag{2.11}$$

where Y is a CARMA(p,q) process and $(\tilde{\varepsilon}_{t_i})$ is an i.i.d. sequence and uncorrelated with (Y_{t_i}). To motivate (2.11) think of (Y_{t_i}) as the random effect of the absolute log returns, which describes their correlation structure. The mean, as we have already said, will be accounted for by μ_{t_i}. But since we will not observe $\mu_{t_i} + Y_{t_i}$ due to some microstructure noise, like for example the fixed tick size of the log returns, we will make some measurement error $\tilde{\varepsilon}_{t_i}$. To assure that Y is non-negative, the driving Lévy process L of the CARMA(p,q) process Y has to be non-decreasing and the kernel of Y has to be non-negative. By substituting (2.11) into (2.10) we get

$$\tilde{\varepsilon}_{t_i} \;=\; |r_{t_i}| - \mu_{t_i} - Y_{t_i},$$

which leads to

$$\mathbb{E}(\tilde{\varepsilon}_{t_i}) = -\mathbb{E}(Y_{t_i}) = \mathbf{b}^T A^{-1} \mathbf{1}_p \mathbb{E}(L_1).$$

The variance of ε_{t_i} decomposes into

$$\begin{aligned} \sigma_\varepsilon^2 &= \text{var}(Y_{t_i}) + \text{var}(\tilde{\varepsilon}_{t_i}), \\ &=: \text{var}(L_1)\mathbf{b}^T \Sigma \mathbf{b} + \sigma_{\tilde{\varepsilon}}^2, \end{aligned}$$

and the autocovariance function of (ε_{t_i}) is equal to that of (Y_{t_i}), i.e.

$$\text{cov}(\varepsilon_{t_i}, \varepsilon_{t_{i-1}}) = \text{var}(L_1)\mathbf{b}^T e^{A(t_i - t_{i-1})} \Sigma \mathbf{b}.$$

2.4 A generalised regression model with CARMA(p,q) random effects

The above considerations have led us to the model

$$|r_{t_i}| = \exp(\mathbf{x}_{t_i}^T \boldsymbol{\beta}) + Y_{t_i} + \tilde{\varepsilon}_{t_i}, \qquad i = 1, \ldots, n. \tag{2.12}$$

In (2.12) we will understand $\exp(\mathbf{x}_{t_i}^T \boldsymbol{\beta})$ as some fixed effect, Y_{t_i} as some random effect and $\tilde{\varepsilon}_{t_i}$ as a measurement error. The parameters which have to be estimated are

$$\boldsymbol{\theta} := (a_1, \ldots, a_p, b_1, \ldots, b_q, \sigma^2, \beta_0, \ldots, \beta_s, \sigma_{\tilde{\varepsilon}}^2),$$

with $\sigma^2 := \mathrm{var}(L_1)$. This is done by an iterated estimation algorithm, which will be described in the next section.

3 Parameter Estimation

The actual parameter estimation can be done in two ways. The first one (henceforth called direct approach) works directly on the linear regression model approximation to model (2.12), which will be introduced in the following, and the second one (henceforth called state space approach) on the associated state space model with application of the Kalman filter. Both estimation procedures will be explained in Section 3.1 and 3.2, respectively. But first we start by describing the general estimation algorithm. Therefore consider equation (2.10) in vector notation

$$|\mathbf{r}| = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{3.13}$$

with $|\mathbf{r}| = (|r_{t_1}|, \ldots, |r_{t_n}|)^T$, $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ similarly. Since we chose the logarithm as link function, we have the relationship

$$log(\boldsymbol{\mu}) = X\boldsymbol{\beta} =: \boldsymbol{\eta}. \tag{3.14}$$

The covariance matrix of $\boldsymbol{\varepsilon}$ shall be denoted by

$$V(\boldsymbol{\xi}) = \mathrm{cov}(\mathbf{Y}) + \sigma_{\tilde{\varepsilon}}^2 I_n,$$

with $\boldsymbol{\xi} := (a_1, \ldots, a_p, b_1, \ldots, b_q, \sigma^2, \sigma_{\tilde{\varepsilon}}^2)$ and $\mathbf{Y} = (Y_{t_1}, \ldots, Y_{t_n})^T$. Equation (3.13) is just a nonlinear regression model with correlated errors. Therefore the parameters can be estimated by maximizing

$$G(\boldsymbol{\theta}, |\mathbf{r}|) := -(|\mathbf{r}| - \boldsymbol{\mu})^T V(\boldsymbol{\xi})^{-1} (|\mathbf{r}| - \boldsymbol{\mu}). \tag{3.15}$$

Applying the Fisher scoring algorithm to maximize (3.15) leads to an iterative generalised least squares problem. The linear model, occurring in each iteration step, can be constructed as in generalised linear models (McCullagh and Nelder [19] p.40) by applying the link function $g(\cdot) := \log(\cdot)$ to the data $|\mathbf{r}|$ and linearise to the first order. The estimation algorithm, which can also be found e.g. in Schall [20] , is described in the following.

General Estimation Algorithm:

(i) Linearize $\mathbf{g}(|\mathbf{r}|) := (g(|r_{t_1}|), \ldots, g(|r_{t_n}|)^T$ to the first order

$$g(|\mathbf{r}|) = g(\boldsymbol{\mu}) + \left(\frac{\partial}{\partial \boldsymbol{\mu}} g(\boldsymbol{\mu})\right)(|\mathbf{r}| - \boldsymbol{\mu}),$$

where $\left(\frac{\partial}{\partial \boldsymbol{\mu}} g(\boldsymbol{\mu})\right)$ is a diagonal matrix with elements $(\frac{\partial}{\partial \mu_{t_1}} g(\mu_{t_1}), \ldots, \frac{\partial}{\partial \mu_{t_n}} g(\mu_{t_n}))$, and define the new dependent variable

$$\begin{aligned} \mathbf{z} &:= g(\boldsymbol{\mu}) + \left(\frac{\partial}{\partial \boldsymbol{\mu}} g(\boldsymbol{\mu})\right)(|\mathbf{r}| - \boldsymbol{\mu}) \\ &= \boldsymbol{\eta} + \left(\frac{\partial}{\partial \boldsymbol{\mu}} g(\boldsymbol{\mu})\right)\boldsymbol{\varepsilon} \\ &= \boldsymbol{\eta} + \mathbf{e}, \end{aligned}$$

where $\mathbf{e} := \left(\frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{\eta}\right)\boldsymbol{\varepsilon}$. Now we have a linear regression model with correlated errors

$$\mathbf{z} = X\boldsymbol{\beta} + \mathbf{e}, \tag{3.16}$$

where $\mathbb{E}(\mathbf{z}) = X\boldsymbol{\beta}$ and $\mathrm{cov}(\mathbf{e}) = (\frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{\eta}) V(\boldsymbol{\xi}) (\frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{\eta})^T$.

(ii) To get starting values $\hat{\eta}^0, \hat{\mathbf{z}}^0$ we fit a generalised linear model to (3.13) assuming uncorrelated errors, i.e. $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2_{\boldsymbol{\varepsilon}} I_n$.

(iii) Start Iteration $k = 1$

(iv) The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ in (3.16) are then estimated in the direct or state space approach giving parameter estimates

$$\hat{\boldsymbol{\beta}}^k \text{ and } \hat{\boldsymbol{\xi}}^k,$$

respectively.

(v) Construct new estimates of $\boldsymbol{\eta}$, i.e. define

$$\hat{\boldsymbol{\eta}}^k := X\hat{\boldsymbol{\beta}}^k.$$

Check if

$$||\hat{\boldsymbol{\eta}}^k - \hat{\boldsymbol{\eta}}^{k-1}|| < TOL$$

is satisfied. If not set

$$\begin{aligned} \hat{\boldsymbol{\mu}}^k &:= g^{-1}(\hat{\boldsymbol{\eta}}^k) \\ \hat{\mathbf{z}}^k &:= \hat{\boldsymbol{\eta}}^k + \left(\frac{\partial}{\partial \boldsymbol{\mu}}\hat{\boldsymbol{\eta}}^k|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}^k}\right)(|\mathbf{r}| - \hat{\boldsymbol{\mu}}^k) \end{aligned}$$

$k = k + 1$ and go to (iv).

Both estimation approaches will perform *quasi maximum likelihood* (QML) estimation (see for example White [21]) of the parameters, which requires only the knowledge of the first two moments of the model for the data. In particular the quasi maximum likelihood estimate (QMLE) $\widehat{\boldsymbol{\theta}}$ of an arbitrary parameter vector $\boldsymbol{\theta}$ is defined, in this case, to maximizes the QML-estimation criterion

$$Q_n(\boldsymbol{\theta}, \mathbf{z}) := -\frac{1}{n}\left[\log(|\Lambda(\boldsymbol{\xi})|) + (\mathbf{z} - X\boldsymbol{\beta})^T\Lambda(\boldsymbol{\xi})^{-1}(\mathbf{z} - X\boldsymbol{\beta})\right] \tag{3.17}$$

where

$$\Lambda(\boldsymbol{\xi}) := \left(\frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{\eta}\right) V(\boldsymbol{\xi}) \left(\frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{\eta}\right)^T.$$

Therefore

$$\widehat{\boldsymbol{\theta}} := \text{argmax}_{\boldsymbol{\theta}\in\Theta} Q_n(\boldsymbol{\theta}, \mathbf{z}), \tag{3.18}$$

where $\Theta := \tilde{\Theta} \times \mathbb{R}_+ \times \mathbb{R}^{s+1} \times \mathbb{R}_+$, with

$$\begin{aligned} \tilde{\Theta} := \{(a_1, \ldots, a_p, b_1, \ldots, b_q) \quad | \quad & a(z) \neq 0 \text{ if } Re(z) \geq 0; b(z) \neq 0 \text{ if } Re(z) > 0 : \\ & \text{the kernel of } Y \text{ is non-negative } \}. \end{aligned}$$

Conditions for the kernel of Y to be non-negative are given in Tsai and Chan [22].

3.1 Direct approach

The estimation of parameters in (3.16) is a generalised least squares problem. It can be solved in the following way. Since $\Lambda(\boldsymbol{\xi})$ is positive definite there exists a positive definite lower triangular matrix $K(\boldsymbol{\xi})$ with ones on the leading diagonal, and a positive definite diagonal matrix $F(\boldsymbol{\xi})$, such that

$$\Lambda(\boldsymbol{\xi})^{-1} := K(\boldsymbol{\xi})^T F(\boldsymbol{\xi})^{-1} K(\boldsymbol{\xi}).$$

If we transform the data

$$\mathbf{z}^*(\boldsymbol{\xi}) := K(\boldsymbol{\xi})\mathbf{z}, \quad X^*(\boldsymbol{\xi}) := K(\boldsymbol{\xi})X, \quad \mathbf{e}^*(\boldsymbol{\xi}) := K(\boldsymbol{\xi})\mathbf{e},$$

we get the heteroscedastic regression model

$$\mathbf{z}^*(\boldsymbol{\xi}) = X^*(\boldsymbol{\xi})\boldsymbol{\beta} + \mathbf{e}^*(\boldsymbol{\xi}) \quad \text{with} \quad \operatorname{cov}(\mathbf{e}^*) = F(\boldsymbol{\xi}). \tag{3.19}$$

If we assume that $\boldsymbol{\xi}$ is known and fixed, we get the generalised least squares estimate of $\boldsymbol{\beta}$ by solving an ordinary least-squares problem:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(\boldsymbol{\xi}) &= [(F(\boldsymbol{\xi})^{-1/2}X^*(\boldsymbol{\xi}))^T F(\boldsymbol{\xi})^{-1/2}X^*(\boldsymbol{\xi})]^{-1}(F(\boldsymbol{\xi})^{-1/2}X^*(\boldsymbol{\xi}))^T F(\boldsymbol{\xi})^{-1/2}\mathbf{z}^*(\boldsymbol{\xi}) \\ &= [X^T\Lambda^{-1}(\boldsymbol{\xi})X]^{-1}X^T\Lambda^{-1}(\boldsymbol{\xi})\mathbf{z}. \end{aligned} \tag{3.20}$$

Replacing $\boldsymbol{\beta}$ in (3.17) by the above estimate on gets the reduced QML-estimation criterion

$$Q_n(\boldsymbol{\xi}, \mathbf{z}) := \frac{1}{n}\sum_{i=1}^{n}\left[-\log(F_{t_i}(\boldsymbol{\xi})) - \frac{v_{t_i}^2(\boldsymbol{\xi})}{F_{t_i}(\boldsymbol{\xi})}\right], \tag{3.21}$$

with $v_{t_i}(\boldsymbol{\xi}) = z_{t_i}^*(\boldsymbol{\xi}) - x_{t_i}^{*^T}(\boldsymbol{\xi})\widehat{\boldsymbol{\beta}}(\boldsymbol{\xi})$ and $F_{t_i}(\boldsymbol{\xi}) = (F(\boldsymbol{\xi}))_{i,i}$. QMLE of the parameters are therefore obtained by first maximizing (3.21) with respect to $\boldsymbol{\xi}$ to get $\widehat{\boldsymbol{\xi}}$. Afterwards one replaces $\boldsymbol{\xi}$ in $\widehat{\boldsymbol{\beta}}(\boldsymbol{\xi})$ by $\widehat{\boldsymbol{\xi}}$ to get the generalised least squares estimate of $\boldsymbol{\beta}$.

Remark 3.1. *The estimation of the parameters in the direct approach includes the computation of the inverse of $\Lambda(\boldsymbol{\xi})$. In the application, which we have in mind, the dimension of $\Lambda(\boldsymbol{\xi})$ 2000 to 3000. $\Lambda(\boldsymbol{\xi})^{-1}$ will also be a full matrix in comparison to regularly spaced observation, where $\Lambda(\boldsymbol{\xi})^{-1}$ will be sparse (see Jones [23] for details). Computationally it is not efficient to compute this inverse, and therefore we reformulate (3.16) as a state space model and apply the Kalman filter to compute (3.21). The idea to rewrite a regression model in state space form is explained for example in Durbin and Koopman [24] and Jones [23].*

3.2 State space approach

Consider again the linear regression model with correlated errors

$$\mathbf{z} = X\boldsymbol{\beta} + \left(\frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{\eta}\right)\boldsymbol{\varepsilon}.$$

Since $\boldsymbol{\varepsilon} = \mathbf{Y} + \tilde{\boldsymbol{\varepsilon}}$, where $Y_{t_i} = \mathbf{b}^T\mathbf{W}_{t_i}$ is a CARMA(p,q) process, and $\frac{\partial}{\partial \boldsymbol{\mu}}\boldsymbol{\eta} = \text{ diag } (1/\mu_{t_1}, \ldots, 1/\mu_{t_n})$, because of the log-link, we get the following state space representation of (3.16).

(i) Observation equation:

$$z_{t_i} = \mathbf{x}_{t_i}^T\boldsymbol{\beta} + G_{t_i}\boldsymbol{\alpha}_{t_i} + \frac{1}{\mu_{t_i}}\tilde{\varepsilon}_{t_i}, \tag{3.22}$$

where

$$G_{t_i} := \frac{1}{\mu_{t_i}}\mathbf{b}^T \text{ and } \boldsymbol{\alpha}_{t_i} := \mathbf{W}_{t_i}.$$

with $\mathbf{x}_{t_i}^T$ the i-th row of $X \in \mathbb{R}^{n\times s+1}$.

(ii) State equation:

$$\boldsymbol{\alpha}_{t_{i+1}} = T_{t_i}\boldsymbol{\alpha}_{t_i} + \boldsymbol{\zeta}_{t_i}, \tag{3.23}$$

where

$$T_{t_i} := e^{A(t_{i+1}-t_i)} \text{ and } \boldsymbol{\zeta}_{t_i} := \int_{t_i}^{t_{i+1}} e^{A(t_{i+1}-u)}\mathbf{1}_p dL_u.$$

One standard assumption for state-space models is the zero mean of the noise processes. This assumption is not fullfilled in (3.22) and (3.23). But we can construct a second state-space model, which has the same first and second moment structure for the observations as the first model. Since we will use a quasi-likelihood approach to estimate the parameters $\boldsymbol{\xi}$, only the first two moments are required. Because of the assumption $\mathbb{E}(\tilde{\varepsilon}_{t_i}) = -\mathbb{E}(Y_{t_i})$, a zero mean CARMA(p,q) process $(Y_t^*)_{t\geq 0} = (\mathbf{b}^T\mathbf{W}_t^*)_{t\geq 0}$, with $\text{cov}(Y_t^*, Y_s^*) = \text{cov}(Y_t, Y_s)$, together with an i.i.d. noise sequence$(\tilde{\varepsilon}_{t_i}^*)$, with $\mathbb{E}(\tilde{\varepsilon}_{t_i}^*) = 0$, $\text{var}(\tilde{\varepsilon}_{t_i}^*) = \sigma_{\tilde{\varepsilon}}^2$ and uncorrelated with Y^*,will lead to the same first and second order structure of z_{t_i}. Let L^* be a Lévy process with $\mathbb{E}(L_1^*) = 0$ and $\text{var}(L_1^*) = \text{var}(L_1)$. Then we get the state-space model:

(i) Observation equation:

$$z_{t_i} = \mathbf{x}_{t_i}^T\boldsymbol{\beta} + G_{t_i}\boldsymbol{\alpha}_{t_i}^* + \frac{1}{\mu_{t_i}}\tilde{\varepsilon}_{t_i}^*, \tag{3.24}$$

where

$$G_{t_i} = \frac{1}{\mu_{t_i}}\mathbf{b}^T \text{ and } \boldsymbol{\alpha}^*_{t_i} := \mathbf{W}^*_{t_i}.$$

with $\mathbf{x}^T_{t_i}$ the i-th row of $X \in \mathbb{R}^{n\times s+1}$.

(ii) State equation:

$$\boldsymbol{\alpha}^*_{t_{i+1}} = T_{t_i}\boldsymbol{\alpha}^*_{t_i} + \boldsymbol{\zeta}^*_{t_i}, \tag{3.25}$$

where

$$T_{t_i} = e^{A(t_{i+1}-t_i)} \text{ and } \boldsymbol{\zeta}^*_{t_i} := \int_{t_i}^{t_{i+1}} e^{A(t_{i+1}-u)}\mathbf{1}_p dL^*_u.$$

An augmented Kalman filter (see e.g. Durbin and Koopman [24]) will be applied to (3.24) and (3.25). The idea of this filter is to apply the Kalman filter with observation matrix G_{t_i} and state matrix T_{t_i} to the variables $z_{t_i}, x^T_{t_i,1}, \ldots, x^T_{t_i,s+1}$ consecutively. $x^T_{t_i,k}$ is the k-th element of the row vector $\mathbf{x}^T_{t_i}$. For each of the variables $x^T_{t_i,1}, \ldots, x^T_{t_i,s+1}$ a new state vector $\boldsymbol{\alpha}^k_{t_i}$, $k = 1, \ldots, s+1$ is taken, but the variance elements in the Kalman filter are the same as for z_{t_i}. The Kalman filter computes best linear predictions $\widehat{z}_{t_i}, \widehat{x}^T_{t_i,1}, \ldots, \widehat{x}^T_{t_i,s+1}$ based on all past observations $\{z_{t_j}, x^T_{t_j,1}, \ldots, x^T_{t_j,s+1}; 1 \le j < i\}$. In each step of the filter we store the one-step forecast errors $z^*_{t_i}(\boldsymbol{\xi}) := z_{t_i} - \widehat{z}_{t_i}, x^{*^T}_{t_i,1}(\boldsymbol{\xi}) := x^T_{t_i,1} - \widehat{x}^T_{t_i,1}, \ldots, x^{*^T}_{t_i,s+1}(\boldsymbol{\xi}) := x^T_{t_i,s+1} - \widehat{x}^T_{t_i,s+1}$. These forecast errors can then be used to calculate the generalised least square estimates $\widehat{\boldsymbol{\beta}}$, given by

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\xi}) := \left(\sum_{i=1}^{n} X^{*^T}_{t_i}(\boldsymbol{\xi})F^{-1}_{t_i}(\boldsymbol{\xi})X^*_{t_i}(\boldsymbol{\xi})\right)^{-1} \sum_{i=1}^{n} X^{*^T}_{t_i}(\boldsymbol{\xi})F^{-1}_{t_i}(\boldsymbol{\xi})z^*_{t_i}(\boldsymbol{\xi}), \tag{3.26}$$

where $\mathbf{x}^{*^T}_{t_i}(\boldsymbol{\xi}) := (x^{*^T}_{t_i,1}(\boldsymbol{\xi}), \ldots, x^{*^T}_{t_i,s+1}(\boldsymbol{\xi}))$ and $F_{t_i}(\boldsymbol{\xi}) := \mathrm{var}(z^*_{t_i}(\boldsymbol{\xi}) - \mathbf{x}^{*^T}_{t_i}(\boldsymbol{\xi})\boldsymbol{\beta})$. To see that (3.26) is equal to (3.20) one has to recall that

$$\Lambda^{-1}(\boldsymbol{\xi}) = K^T(\boldsymbol{\xi})F^{-1}(\boldsymbol{\xi})K(\boldsymbol{\xi}). \tag{3.27}$$

Inserting (3.27) into (3.20) yields

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\xi}) = [(K(\boldsymbol{\xi})X)^T F^{-1}(\boldsymbol{\xi})K(\boldsymbol{\xi})X]^{-1}(K(\boldsymbol{\xi})X)^T F^{-1}(\boldsymbol{\xi})K(\boldsymbol{\xi})\mathbf{z}.$$

Since the Kalman filter performs the Cholesky decomposition (3.27) (Harvey [25]), we see that applying the Kalman filter is equivalent to the multiplication by the matrix $K(\boldsymbol{\xi})$. For more details on the augmented Kalman filter see Durbin and Koopman [24] or Harvey [25].

The procedure to estimate the parameters is then exactly the same as in the direct approach. First $\boldsymbol{\xi}$ is estimated by maximizing

$$\begin{aligned} Q_n(\boldsymbol{\xi}, \mathbf{z}) &= \frac{1}{n}\sum_{i=1}^{n}\left[-\log(F_{t_i}(\boldsymbol{\xi})) - \frac{(v^*_{t_i}(\boldsymbol{\xi}) - X^*_{t_i}(\boldsymbol{\xi})\widehat{\boldsymbol{\beta}}(\boldsymbol{\xi}))^2}{F_{t_i}(\boldsymbol{\xi})}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[-\log(F_{t_i}(\boldsymbol{\xi})) - \frac{v^2_{t_i}(\boldsymbol{\xi})}{F_{t_i}(\boldsymbol{\xi})}\right] \end{aligned}$$

with respect to $\boldsymbol{\xi}$. This estimate is denoted by $\widehat{\boldsymbol{\xi}}$. Afterwards $\boldsymbol{\xi}$ in (3.26) is replaced by $\widehat{\boldsymbol{\xi}}$ to get the generalised least squares estimate of $\boldsymbol{\beta}$.

4 Simulation results

The performance of the QML estimator using the state space approach is going to be analysed in a small simulation study. The parameters are estimated in two setups. One with regularly spaced observations and the other with irregularly spaced ones. For the regularly spaced observations we created 2000 equidistant time points in the interval $(0, 400)$. In case of irregularly sampling the durations are exponentially distributed, with a mean value of 0.2, to assure that time points are also in the interval $(0, 400)$.

In each of the 100 simulations the sample size was 2000. As a explanatory variable we took real bid ask spreads from the IBM stock. The regression coefficient β was taken equal to 0.3. We did not include an intercept in the regression. The correlation was simulated by a CARMA(1,0) process with parameter $a = 0.8$. As driving Lévy process L wwe chose a compound Poisson process with jumps (X_k) i.i.d. expo(100) $(\mathbb{E}(X_k) = 0.01, \mathrm{var}(X_k) = 0.0001)$ and $N_t \sim Pois(3t)$. The jump rate of the Poisson process N was taken equal to 3. The mean and variance of L_1 are then 0.0375 and $\sigma^2 = 0.0006$, respectively. The choice of the parameter values was motivated by similar parameter values obtained in the application presented later. The measurement noise $\tilde{\varepsilon}$ was simulated as a Gaussian i.i.d. noise with mean -0.0375 and variance $\sigma^2_{\tilde{\varepsilon}} = 0.0001$, respectively.

For the resulting estimates we computed estimates of mean, bias, mean absolute error (MAE), mean squared error (MSE) and the estimated standard errors of these estimates. The results can be seen in Table 1 and 2 showing satisfying performance for both settings.

	$\widehat{a}$	$\widehat{\beta}$
true value	0.8000	3.0000e-01
mean	0.8122 (0.0095)	2.9881e-01 (1.1341e-03)
median	0.8106 (0.0095)	2.9972e-01 (1.1341e-03)
bias	0.0122 (0.0095)	-1.1903e-03 (1.1341e-03)
MAE	0.0781 (0.0056)	8.8016e-03 (7.1971e-04)
MSE	0.0092 (0.0013)	1.2875e-04 (1.8454e-05)
	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_{\tilde{\varepsilon}}$
true value	6.0000e-04	1.0000e-04
mean	6.1091e-04 (6.9019e-06)	9.9395e-05 (7.8563e-07)
median	6.0989e-04 (6.9019e-06)	9.9384e-05 (7.8563e-07)
bias	1.0916e-05 (6.9019e-06)	-6.0439e-07 (7.8563e-07)
MAE	5.7419e-05 (3.9417e-06)	6.2597e-06 (4.7446e-07)
MSE	4.8352e-09 (6.04782e-10)	6.1470e-11 (7.7510e-12)

Table 1: Mean, median, bias, mean absolute error (MAE) and mean squared error (MSE) for $\widehat{a}$, $\widehat{\beta}, \widehat{\sigma}^2$ and $\widehat{\sigma}^2_{\tilde{\varepsilon}}$ together with their estimated standard errors in parentheses in case of *regularly spaced observations*.

	$\widehat{a}$	$\widehat{\beta}$
true value	0.8000	3.0000e-01
mean	0.8015 (0.0092)	2.9844e-01 (9.4843e-04)
median	0.7944 (0.0092)	2.9764e-01 (9.4843e-04)
bias	0.0015 (0.0092)	-1.5541e-03 (9.4843e-04)
MAE	0.0696 (0.0059)	8.1259e-03 (5.0689e-04)
MSE	0.0082 (0.0014)	9.1468e-05 (1.1264e-05)
	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_{\tilde{\varepsilon}}$
true value	6.0000e-04	1.0000e-04
mean	6.0974e-04 (6.9191e-06)	9.8657e-05 (5.4509e-07)
median	6.0357e-04 (6.9191e-06)	9.9198e-05 (5.4509e-07)
bias	9.7488e-06 (6.9191e-06)	-1.3423e-06 (5.4509e-07)
MAE	5.5842e-05 (4.1634e-06)	4.4295e-06 (3.4225e-07)
MSE	4.8344e-09 (6.8064e-10)	3.1220e-11 (4.2670e-12)

Table 2: Mean, median, bias, mean absolute error (MAE) and mean squared error (MSE) for $\widehat{a}$, $\widehat{\beta}, \widehat{\sigma}^2$ and $\widehat{\sigma}^2_{\tilde{\varepsilon}}$ together with their estimated standard errors in parentheses in case of *irregularly spaced observations*.

5 Application

The data,which we will use, comes from the *Trades and Quotes* (TAQ) database of the New York Stock Exchange (NYSE). We will work with IBM trade data from September 30, 2002 up to October 31, 2002. The NYSE market opens 9:30 am and closes at 4:00 pm. Tradings outside these official trading hours have been deleted. Since we want to concentrate on real price changes we also excluded all zero returns and the corresponding explanatory variables. We also eliminated all multiple trades. Trades for the same transaction price were treated as a single trade by adding up the volumes. Different transaction prices were averaged and the volumes totalled. The resulting data set consists of transaction, bid and ask prices (all measured in cents of US dollars), transaction times (measured in seconds) and volumes (measured in the number of shares) realised over the specified time period. No further data manipulations have been carried out. Exemplary the absolute log returns of six trading days have been plotted in Figure 2.

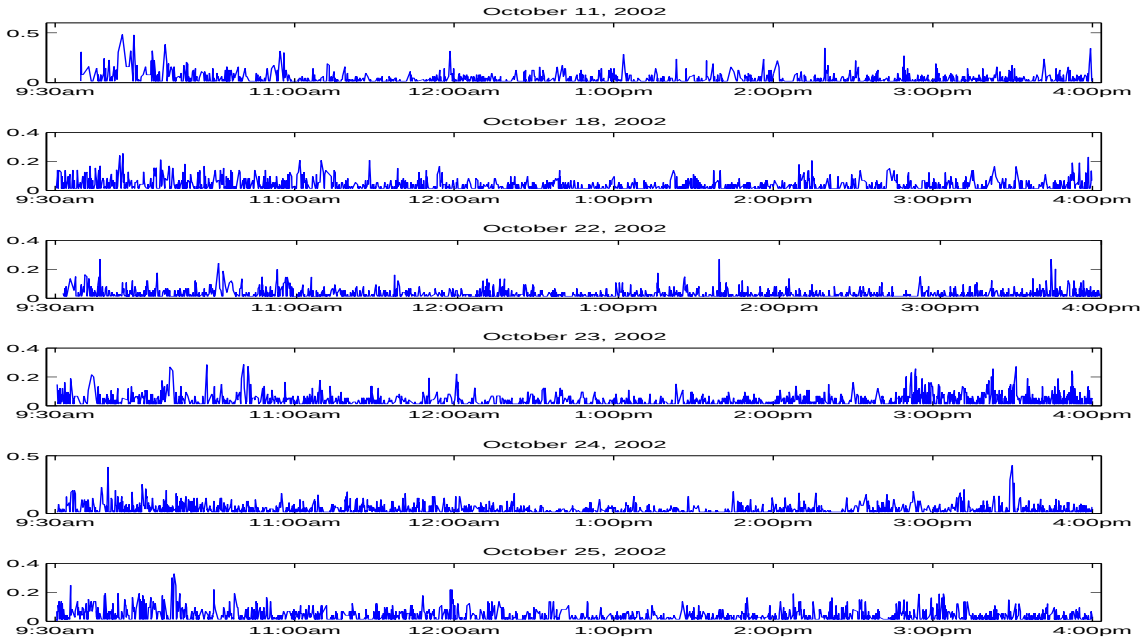

Figure 2: Absolute log returns of the 11th (*first row*), 18th (*second row*), 22nd (*third row*), 23rd (*fourth row*), 24th (*fifth row*) and 25th (*last row*) of October 2002. The time is measured in real time.

In Section 2 we have said, that a parametric approach is used. But up to now we have not specified the parametric setup. To get an idea how the absolute log return may depend on the explanatory variables, we perform some kind of explorative data analysis by fitting a *Generalized Additive Model* (see Hastie and Tibshirani [26]) with uncorrelated errors to the data. The functional relationship displayed by the model, will then be used to set up a parametric model. The aim of the analysis in this section is to fit our model to the data. Then to check if the fitted correlation structure can be justified and investigate the predictive power of the explanatory variables. The one step ahead predictions of the absolute log return for October 14th until October 31st, 2002, will be computed using the information corresponding to each of the following four setups:

(i) the last day

(ii) the last three days

(iii) the last day and the same day one week ago

(iv) the same day one and two weeks ago.

The different forecasts are then compared using the mean squared error as criterion. Exemplary we will present the estimation results for the days needed to predict October 25th, 2002.

5.1 Explorative data analysis

Initially we chose only the bid-ask spread and the duration as explanatory variables. The influence of the volume will be analysed in a further study. Therefore the generalised additive model under consideration is the following one

$$\log(\mu_{t_i}) = s_1(b_{t_i}) + s_2(d_{t_i}),$$

where $s_i(),\ i = 1, 2$, are smoothing splines and b_{t_i} (bid-ask spread) and d_{t_i} (durations) are the explanatory variables. This model is fitted using the Splus function `gam()` under the assumption of uncorrelated errors. The results of this estimation procedure can be seen in Figure 3.

For the bid-ask spread as well as the duration one can recognize a relatively smooth functional relationship. We decided, that a polynomial of third order has enough

Day	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$
October 11, 2002	-4.2726	18.0106	-48.9357	49.1082	1.8313
October 18, 2002	-4.4576	18.0729	-27.0261	-42.8318	2.6261
October 22, 2002	-4.6144	24.0957	-113.1130	253.7693	2.1601
October 23, 2002	-4.3120	17.2861	-38.7623	22.5341	1.3981
October 24, 2002	-4.4129	15.7375	-27.6543	16.4028	2.8124
October 25, 2002	-4.6366	26.6262	-117.5710	228.4430	1.8190
Day	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{a}$	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_{\tilde{\varepsilon}}$
October 11, 2002	-2.1036	0.7714	0.3942	1.1e-03	2.6e-09
October 18, 2002	-4.7253	2.6941	0.5942	7.4e-04	9.1e-13
October 22, 2002	-3.8395	3.2206	0.9886	2.1e-04	4.1e-04
October 23, 2002	-0.2093	-0.7194	0.7301	1.3e-03	2.1e-10
October 24, 2002	-3.4322	-0.0021	0.5253	7.1e-04	4.0e-08
October 25, 2002	-1.4558	0.2407	0.8991	9.8e-04	1.4e-04

Table 3: QMLE based on the augmented Kalman filter.

flexibility to model both explanatory variables. This led us to consider a model with design matrix X, where

$$\mathbf{x}_{t_i}^T\boldsymbol{\beta} := \beta_0 + \beta_1 b_{t_i} + \beta_2 b_{t_i}^2 + \beta_3 b_{t_i}^3 + \beta_4 d_{t_i} + \beta_5 d_{t_i}^2 + \beta_6 d_{t_i}^3,$$

with bid-ask spread b_{t_i} and duration d_{t_i}.

5.2 Estimation results

The application of the augmented Kalman filter, which was described in Section 3.2, and the quasi maximum likelihood estimation of the remaining parameters resulted in the parameter estimates, which can be seen in Table 3. The coefficients $\widehat{\beta}_k$, $k = 4, 5, 6$, correspond to durations measured in one-hundredth of a second, whereas the time was measured in seconds. The plots of the absolute log returns together with their fitted mean values are shown in Figure 4 demonstrating no obvious lack of fit.

The regression coefficients lead to estimates of the two polynomials

$$p_b(b_{t_i}) := \beta_0 + \beta_1 b_{t_i} + \beta_2 b_{t_i}^2 + \beta_3 b_{t_i}^3 \tag{5.28}$$

$$p_d(d_{t_i}) := \beta_4 d_{t_i} + \beta_5 d_{t_i}^2 + \beta_6 d_{t_i}^3. \tag{5.29}$$

The estimated polynomials of the m-th day are denoted by

$$\hat{p}_b^m(x) := \hat{\beta}_0^m(\mathbf{b}^m, \mathbf{d}^m) + \hat{\beta}_1^m(\mathbf{b}^m, \mathbf{d}^m)x + \hat{\beta}_2^m(\mathbf{b}^m, \mathbf{d}^m)x^2 + \hat{\beta}_3^m(\mathbf{b}^m, \mathbf{d}^m)x^3$$

and

$$\hat{p}_d^m(x) := \hat{\beta}_4^m(\mathbf{b}^m, \mathbf{d}^m)x + \hat{\beta}_5^m(\mathbf{b}^m, \mathbf{d}^m)x^2 + \hat{\beta}_6^m(\mathbf{b}^m, \mathbf{d}^m)x^3$$

and the observations on the m-th day by

$$\mathbf{b}^m := (b_{t_1}^m, \ldots, b_{t_{n_m}}^m) \quad \text{and} \quad \mathbf{d}^m := (d_{t_1}^m, \ldots, d_{t_{n_m}}^m)$$

where n_m is the number of observations on day m. These estimated polynomials are shown in Figure 3.

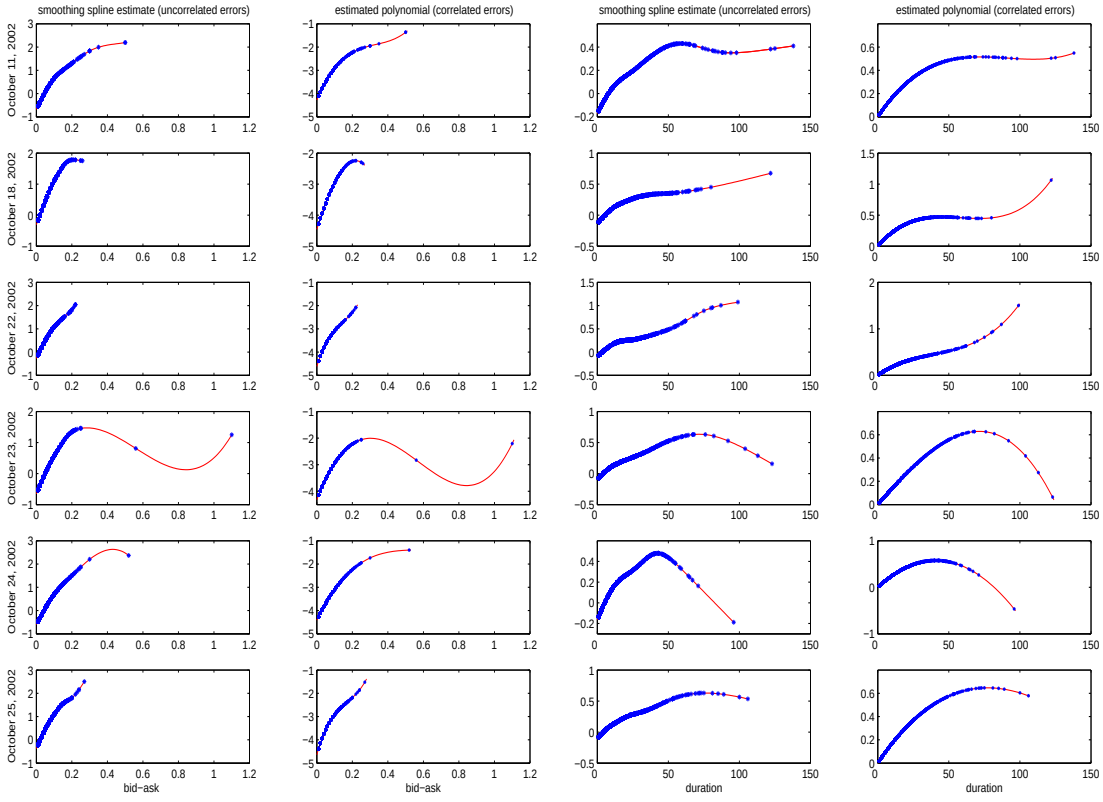

Figure 3: Smoothing spline estimates and estimated bid-ask and duration polynomials $\hat{p}_b^m(\cdot)$ and $\hat{p}_d^m(\cdot)$ for the days 11th (*first row*), 18th (*second row*), 22nd (*third row*), 23rd (*fourth row*), 24th (*fifth row*) and 25th (*last row*) of October 2002. The marks represent the observed values of the explanatory variables.

5.3 Analysis of the correlation structure

In the end we want to take a look at the sample variograms of the residuals, and see if the assumed correlation structure can be justified. The variogram is defined in the appendix, where we also present four examples of sample variograms of simulated CARMA(p,q) processes. Figure 4 contains the sample variograms and variograms of the estimated models for all six residual processes.

The rough structure of the sample variogram is due to the irregularly spaced observations, because the irregular spacing leads to greater changes in the number of observations for consecutive lags. For October 11, 2002 the estimated model proposes stronger correlation than the sample variogram, but despite this fact, the shape of the sample variogram and the variogram based on the estimated model is quite similar. The reason for this might be a numerical imprecison or a misspecified correlation structure, which has to be further analysed. The other days show less correlation in the residuals, which can be seen by the faster increasing variograms. The sample variograms represent the proposed structure of the model variogram quite well. Only for the first few lags we see consistently smaller values of the sample variogram $\widehat{\gamma}(h)$ compared to the model variogram $\gamma(h)$. This may be due to the fact that $\gamma(h) \to \sigma_{\tilde{\varepsilon}}^2$ but $\widehat{\gamma}(h) \to 0$ as $h \to 0$ (see also the appendix). This effect is known in the geostatistics literature as a *nugget effect* and appears because of the superposition of independent noise on an underlying process. The nugget effect can be seen on all six days. Therefore one could try to fit CARMA processes of higher order to the data on October 11th to see, if the fit could be improved. For the remaining days the proposed correlation could be justified.

5.4 Prediction

Since we have shown how to estimated the polynomials, we want to explain now how to predict the mean of the absolute log return of the next trading day. Imagine that we have estimates for $m = 1, \ldots, M$ days. Using these $2M$ polynomials we construct two *mean piecewise polynomials* by averaging over the observed data points

$$\overline{p_b^M}(x) := \frac{1}{|M^b(x)|} \sum_{m \in M^b(x)} \hat{p}_b^m(x) \tag{5.30}$$

$$\overline{p_d^M}(x) := \frac{1}{|M^d(x)|} \sum_{m \in M^d(x)} \hat{p}_d^m(x), \tag{5.31}$$

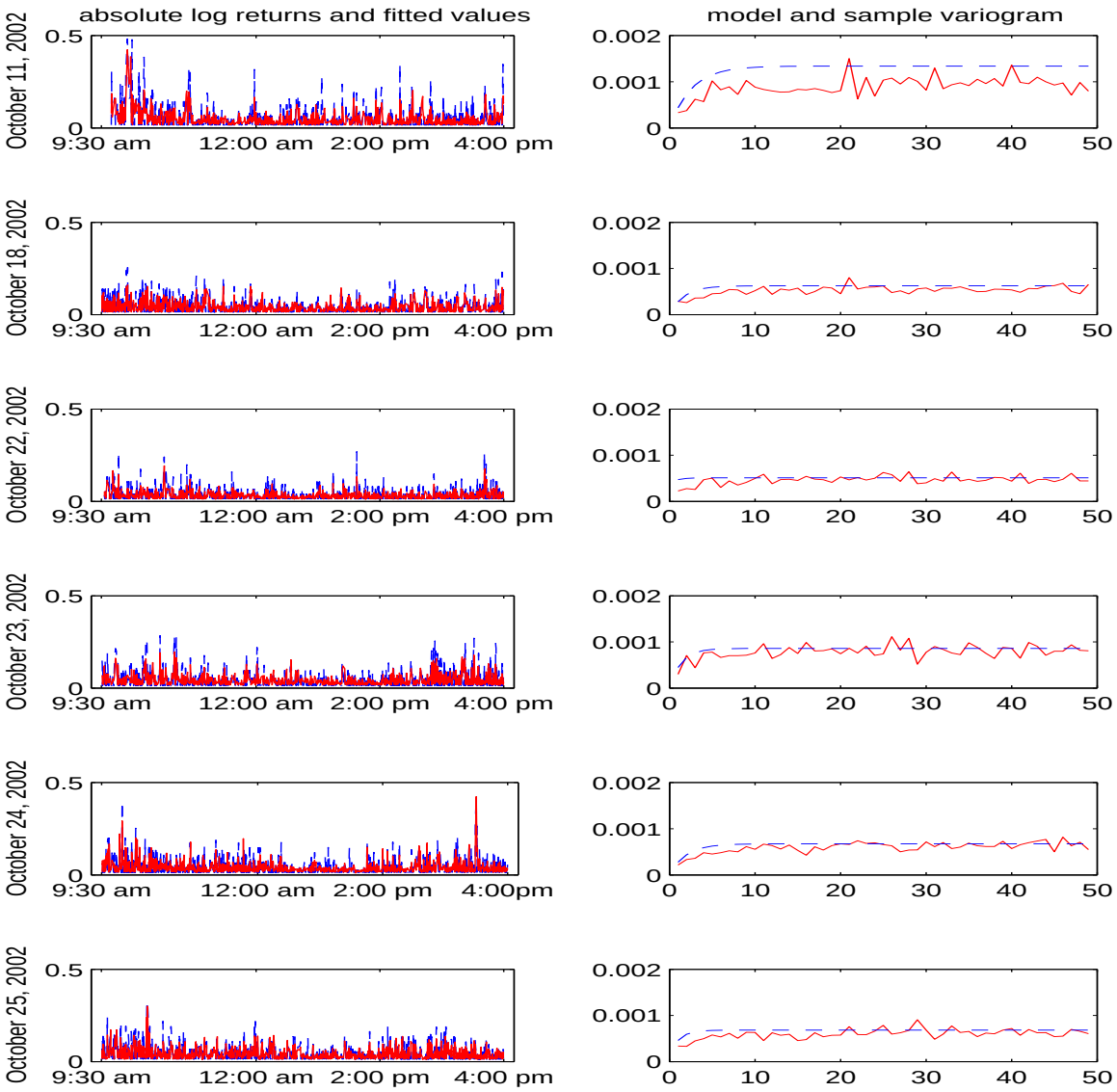

Figure 4: Left column: Absolute log returns (*dashed line*) together with the fitted values (*solid line*) for the days 11th (*top row*), 18th (*second row*), 22nd (*third row*), 23rd (*fourth row*), 24th (*fifth row*) and 25th (*bottom row*) of October 2002. Right column: Model (*dashed line*) and sample variogram of the residuals ε_{t_i} (*solid line*) for the days 11th (*top row*), 18th (*second row*), 22nd (*third row*), 23rd (*fourth row*), 24th (*fifth row*) and 25th (*bottom row*) of October 2002.

where

$$\begin{aligned} M^b(x) &:= \{m \in \{1,\dots,M\} |\ x \in [0, \max_i b^m_{t_i}]\} \\ |M^b(x)| &:= \text{ card } M^b(x) \end{aligned}$$

and

$$\begin{aligned} M^d(x) &:= \{m \in \{1,\dots,M\} |\ x \in [0, \max_i d^m_{t_i}]\} \\ |M^d(x)| &:= \text{ card } M^d(x). \end{aligned}$$

A smoothed version of these two piecewise polynomials for day $M+1$ we get by fitting two smoothing splines at $\overline{p_b^M}(\cdot)$ and $\overline{p_d^M}(\cdot)$ over the intervals $[0, \max_m b^m_{t_{nm}}]$ and $[0, \max_m d^m_{t_{nm}}]$. The smoothing splines $\overline{p_b}(\cdot)$ and $\overline{p_d}(\cdot)$ minimise

$$\sum_{i=1}^{n} \left(\overline{p_b^M}(x^b_{t_i}) - \overline{p_b}(x^b_{t_i})\right)^2 + \lambda_b \int_0^{T_b} \left[\frac{\partial^2 \overline{p_b}(x)}{\partial^2 x}\right]^2 dx, \qquad x^b_{t_i} \in [0, \max_m b^m_{t_{nm}}], \quad (5.32)$$

and

$$\sum_{i=1}^{n} \left(\overline{p_d^M}(x^d_{t_i}) - \overline{p_d}(x^d_{t_i})\right)^2 + \lambda_d \int_0^{T_d} \left[\frac{\partial^2 \overline{p_d}(x)}{\partial^2 x}\right]^2 dx, \qquad x^d_{t_i} \in [0, \max_m d^m_{t_{nm}}] \quad (5.33)$$

respectively, where $\lambda_b, \lambda_d > 0$ are smoothing parameters, $T_b := \max_m b^m_{t_{nm}}$ and T_d similarly. λ_b and λ_d are maximum likelihood estimates. Maximum likelihood estimation of smoothing parameters for spline smoothing is explained in Durbin and Koopman [24].
The predicted mean values of the absolute log returns $|\widehat{r}_{t_i}|$ of the $M+1$-th day are then defined like this

$$P(|r^{M+1}_{t_i}|) := \exp(\overline{p_b}(b^{M+1}_{t_i}) + \overline{p_d}(d^{M+1}_{t_i})). \quad (5.34)$$

Remark 5.1. *Observe that d_{t_i} is unknown up to time t_i. Since we mainly want to investigate the dependence on the explanatory variables, we will assume in a first step, that the durations are known. In a second step an ACD model could be fitted to the durations, to get forecasts also for the durations.*

5.5 Prediction results

As we mentioned at the beginning of this section, the one step ahead predictions of the absolute log return for the days October 14th-31st, 2002, will be computed using the data of:

(i) the last day

(ii) the last three days

(iii) the last day and the same day one week ago

(iv) the same day one and two weeks ago.

Performing the steps described in Section 5.4 produced for each day the smoothing spline estimates $\overline{p_b}^k(\cdot),\ k = 1, \ldots, 4$ and $\overline{p_d}^k(\cdot),\ k = 1, \ldots, 4$. In the first prediction setup (i) the smoothing splines are equal to the estimated polynomials for the last day, since we have only one polynomial observation in each case. For the 25th of October, the smoothing splines together with the mean piecewise polynomials are shown in Figure 5. The absolute log returns together with corresponding predictions can also be seen.

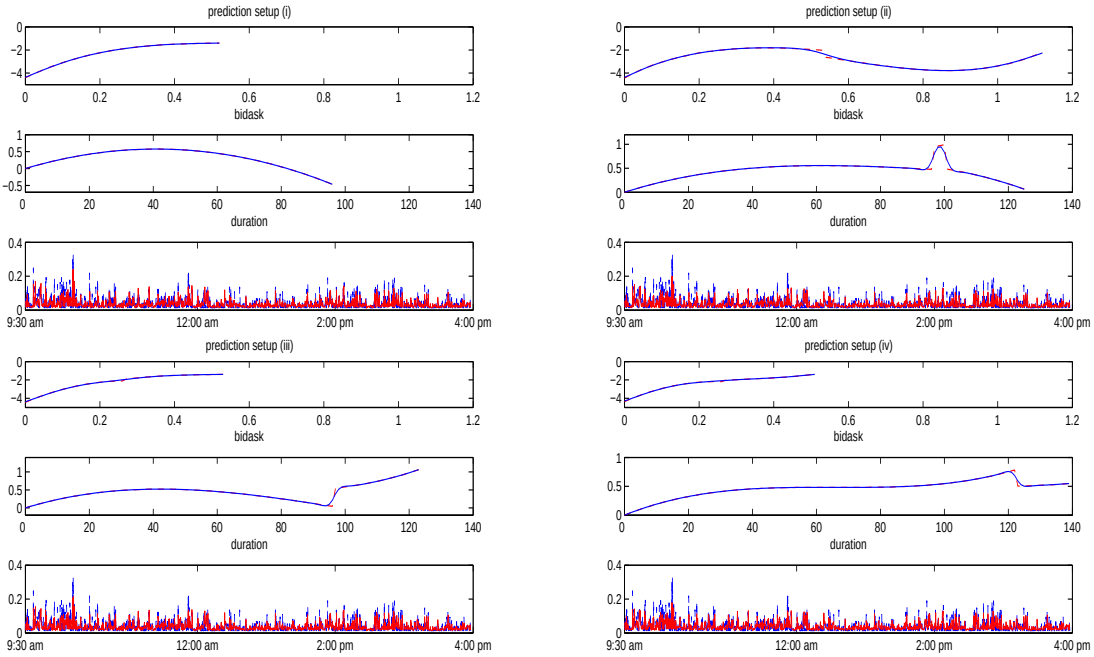

Figure 5: Smoothing spline (*solid line*) and mean piecewise polynomials (*dashed line*) in rows 1,2,4 and 5, absolute log returns (*dashed line*) and mean value predictions (*solid line*) in rows 3 and 6 for the prediction setup (i) (*top left 3 panels*), (ii) (*top right 3 panels*), (iii) (*bottom left 3 panels*) and (iv) (*bottom right 3 panels*).

The different forecast are now compared using the mean squared errors

$$MSE^{k,j} := \frac{1}{n_j}\sum_{i=1}^{n_j}(|r_{t_i}^j| - P^k(|r_{t_i}^j|))^2, \qquad k = 14,\dots,4,\ j \in \mathcal{I},$$

where

$$P^k(|r_{t_i}^j|) := \exp(\overline{p_b}^k(b_{t_i}^j) + \overline{p_d}^k(d_{t_i}^j)), \qquad k = 1,\dots,4,\ j \in \mathcal{I},$$

and $\mathcal{I}$ is the index set of the sample including October 14th to 31st, as criterion. These MSE are shown in Table 5.4. In parentheses one can see the rank of the prediction within each day. For October 14th the random effect could not be de-

Day	setup (i)	setup (ii)	setup (iii)	setup (iv)
October 14	1.3872e-03 (3)	1.3565e-03 (1)	1.3683e-03 (2)	1.3894e-03 (4)
October 15	6.8258e-04 (2)	6.8303e-04 (3)	6.8049e-04 (1)	7.7766e-04 (4)
October 16	9.5416e-04 (1)	9.6394e-04 (2)	9.7708e-04 (3)	1.0133e-03 (4)
October 17	4.5386e-04 (1)	5.1570e-04 (2)	5.6619e-04 (3)	8.8701e-04 (4)
October 18	6.4535e-04 (3)	6.2039e-04 (1)	6.2105e-04 (2)	6.4852e-04 (4)
October 21	5.5981e-04 (1)	5.9419e-04 (3)	5.6657e-04 (2)	8.3658e-04 (4)
October 22	5.2608e-04 (3)	5.2541e-04 (2)	5.2283e-04 (1)	5.9528e-04 (4)
October 23	9.3108e-04 (2)	8.7059e-04 (1)	3.8468e-03 (4)	1.4754e-03 (3)
October 24	7.6446e-04 (4)	7.5806e-04 (3)	7.5375e-04 (2)	7.3782e-04 (1)
October 25	7.0539e-04 (4)	7.0106e-04 (3)	6.9328e-04 (2)	6.9282e-04 (1)
October 28	1.0484e-03 (4)	8.1573e-04 (2)	8.0485e-04 (1)	8.7461e-04 (3)
October 29	8.5291e-04 (2)	8.4212e-04 (1)	8.5606e-04 (3)	8.8279e-04 (4)
October 30	2.6574e-03 (4)	1.8290e-03 (3)	1.2266e-03 (1)	1.3234e-03 (2)
October 31	5.4630e-04 (1)	5.5581e-04 (2)	5.7003e-04 (3)	6.8218e-04 (4)
average rank	(2.50)	(2.07)	(2.14)	(3.28)

Table 4: MSE of the one step ahead predictions on the next trading day for the setup (i), (ii), (iii) and (iv) together with the corresponding rank in parentheses.

scribed by a CARMA(1,0) process. Therefore we fitted a CARMA(2,1) process to the data. To compare the different prediction setups we calculated average ranks over the days. For this data the best strategy would be to use the information of the last three days for prediction. Setup (iii) is the second best strategy and setup (i) and (iv) are third and fourth. This presents a method which allows to empirically

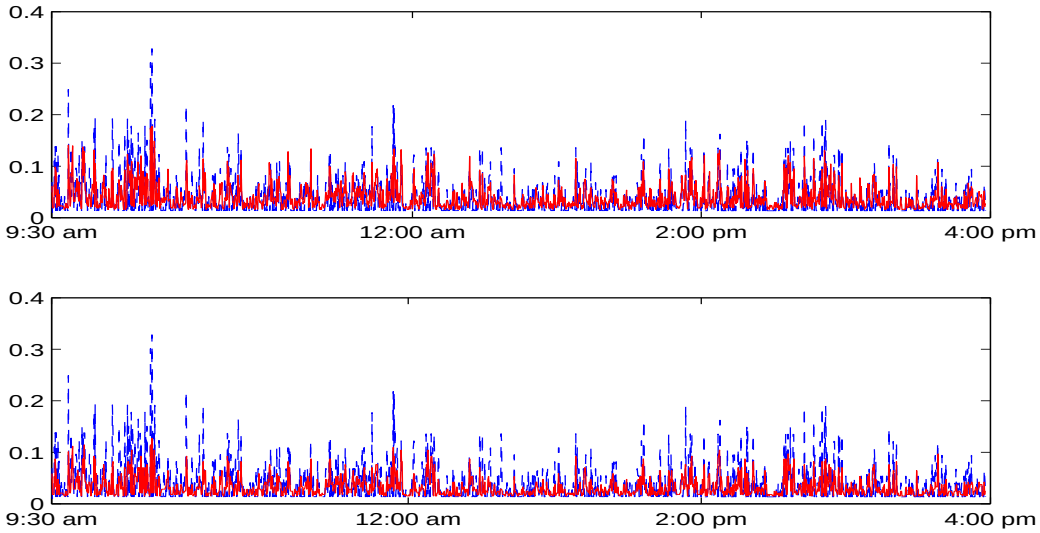

Figure 6: Absolute log returns on October 25th (*dashed line*) and mean value predictions (*solid line*) for prediction setup (iv) using bid ask and duration (*top*) and using only bid ask (*bottom*).

investigate the performance of different prediction strategies. The predictive power of the duration can be seen, when we recompute the predictions for the setup with the smallest MSE without using the duration. We observed an increase in the MSE between 5 and 20 percent. For October 25th the resulting predictions are shown in Figure 6. The mean squared error in this case is equal to $8.1874e-04$, showing a significant increase of about 18 percent.

6 Conclusions and further work

We have proposed a model for ultra high frequency data to investigate the influence of explanatory variables on the mean of the absolute log return. In contrast to other regression analyses of volatility characteristics we worked on a tick-by-tick level. As a result no information is lost in contrast to working with interpolated data of lower frequency. The problem of market microstructure noise of tick-by-tick data will be accounted for on the one hand by the measurement noise and on the other by the fact that we do not accumulate data, but analyse it at every time point. In Section 5 we have seen how to predict the mean value of uhf absolute log returns. To get predictions, which do not depend on unknown explanatory variables, we

could use an autoregressive conditional duration model. Another way of predicting absolute log returns could be to compute some kind of online prediction. This means computing forecasts between two trades for every second, that would display some kind of trend of "inter trade" volatility. These forecasts are then independent of a duration model. One could also think of taking this model as a reference model and trying to replicate the achieved fit with explanatory variables known before the next trade occurs. Here we think of a model with last available bid ask spread, volume of the last trade and the last transaction time as explanatory variables. The MSE as scoring rule has the disadvantage, that it does not take into account the variance of the predictions. Therefore we want to specify the variance of the predictions and use scoring rules like the average ignorance (see for example Gneiting and Raftery [27]), which take into account this variance, to compare predictions.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386, Statisctical Analysis of Discrete Structures.

References

[1] Engle RF. The econometrics of ultra high frequency data. *Econometrica* 2000; **68**: 1-22.

[2] Jungbacker B, Koopman SJ. Model-based measurement of actual volatility in high-frequency data. *Advances in Econometrics*; to appear.

[3] Aït-Sahalia Y, Mykland PA, Zhang L. How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *Review of Financial Studies* 2005; **18**: 351-416.

[4] Engle RF, Russell J. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* 1997; **66**: 1127-1162.

[5] Bauwens L, Giot P, Grammig J, Veredas D. A comparison of financial duration models through density forecasts. *International Journal of Forecasting* 2004; **20**: 589-609.

[6] Bauwens L, Giot P. *Econometric Modelling of Stock Market Intraday Activity.* Kluwer Academic Publishing: Dordrecht, 2001.

[7] Meddahi N, Renault E, Werker BJM. Modelling High-Frequency Data in Continuous Time. Manuscript, University of Montreal 1998. Available at http://www.cireq.umontreal.ca/personnel/meddahi.html

[8] Engle RF, Russell J. A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times: The Autoregressive Conditional Multinomial-Autoregressive Conditional Duration Model. *Journal of Business & Economic Statistics* 2005; **23**: 166-180.

[9] Liesenfeld R, Pohlmeier W. A Dynamic Integer Count Data Model for Financial Transaction Prices. Manuscript, University of Konstanz 2003. Available at http://econometrics.wiwi.uni-konstanz.de/prof/papers.htm

[10] Renault E, Werker BJM. Stochastic volatility models with transaction time risk. Discussion paper 24 Tilburg University 2004. Available at http://center.uvt.nl/staff/werker/preprints/

[11] Corsi F. A Simple Long Memory Model of Realized Volatility. Manuscript, University of Lugano 2004. Available at http://www.nuff.ox.ac.uk/users/shephard/summer.html

[12] Anderson TG, Bollerslev T, Diebold FX. Some Like it Smooth, and Some Like it Rough: Untangling Continuous and Jump Components in Measuring, Modeling, and Forecasting Asset Return Volatility. *CFS Working Paper No. 2003/35.*

[13] Ghysels E, Santa-Clara P, Valkanov R. The MIDAS touch: Mixed Data Sampling Regressions. Manuscript, University of North Carolina and UCLA 2002. Available at http://www.unc.edu/ eghysels/

[14] Barndorff-Nielsen OE, Shephard N. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B* 2002; **64**: 253-280.

[15] Forsberg L, Ghysels E. Why do absolute returns predict volatility so well?. Manuscript, University of North Carolina 2004. Available at http://www.unc.edu/ eghysels/

[16] Barndorff-Nielsen OE, Shephard N. Non-Gaussian Ornstein- Uhlenbeck based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society, Series B* 2001; **63**: 167-241.

[17] Brockwell PJ, Marquardt T. Lèvy-driven and fractionally integrated ARMA processes with continuous time parameter. *Statistica Sinica* 2005; **15**: 477-494.

[18] Applebaum D. *Lévy processes and stochastic calculus*. Cambridge University Press: New York, 2004.

[19] McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London 1989.

[20] Schall R. Estimation in generalized linear models with random effects. *Biometrika* 1991; **78**: 719-727.

[21] White H. *Estimation, Inference and Specification Analysis*. Cambridge University Press: New York 1994.

[22] Tsai H, Chan KS. A Note on Non-negative Continuous-time Processes. Manuscript, University of Iowa 2004. Available at http://www.stat.uiowa.edu/techrep/

[23] Jones RH. *Longitudinal Data with Serial Correlation: A State-space Approach*. Chapman&Hall: London, 1993.

[24] Durbin J, Koopman SJ. *Time Series Analysis by State Space Methods*. Oxford University Press: Oxford, 2001.

[25] Harvey AC. *Forecasting strucural time series models and the Kalman filter*. Cambridge University Press: Cambridge, 1990.

[26] Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall: London, 1990.

[27] Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. Technical Report no. 463, University of Washington. Available at http://www.stat.washington.edu/www/research/reports/2004/

[28] Haslett J. On the sample variogram and the sample autocovariance for non-stationary time series. *The Statistician* 1997; **46**: 475-485.

Appendix

A Variogram for irregularly spaced time series

The variogram is mainly used in geostatistics. Applications for time series data are rare, despite the fact that it has the advantage to be defined for irregularly spaced and even non-stationary time series in comparison to the autocovariance function (see Haslett [28]).

Definition A.1. *(variogram)*
Let $(Z_t)_{0\leq t<\infty}$ be a process, such that

$$\mathbb{E}(Z_{t+h} - Z_t) = Ch,$$

with a constant C, and

$$\text{var}(Z_{t+h} - Z_t) =: 2\gamma(h), \tag{A.1}$$

where $\gamma(h)$ is a conditionally negative definite function. Then $\gamma(h)$ is called the variogram.

Remark A.1. *The requirement that $\gamma(h)$ be conditionally negative definite means that* $\text{var}(\sum_i a_i Y_{t_i})$ *(which is equal to $-\sum_{i,j} a_i a_j \gamma(t_i - t_j)$ when $\sum_i a_i = 0$) be non-negative definite when $\sum_i a_i = 0$.*

For observations $Z_{t_1}, \ldots, Z_{t_n}$, with $C = 0$, the variogram can be estimated through the *sample variogram*

$$\widehat{\gamma}(h) := \frac{1}{2}(n - |N_h|)^{-1} \sum_{(i,j)\in I_h} (Z_{t_i} - Z_{t_j})^2, \tag{A.2}$$

where $N_h := \{(i,j),\ i,j \in \{1,\ldots,n\}| \quad |t_i - t_j| = h\}$.
To compare the sample variogram of the residuals $(\widehat{\varepsilon}_{t_i})$ in (2.12) with the theoretical one, we have to compute the variogram of (ε_{t_i}). It is given by the following expression

$$\gamma_\varepsilon(h) = \text{var}(L_1)\mathbf{b}^T(I_p - e^{Ah})\Sigma\mathbf{b} + \sigma^2_{\widehat{\varepsilon}}. \tag{A.3}$$

Example A.1. *As an example consider a Lévy driven CARMA(p,q) process (Y_t). Here the driving Lévy process (L_t) is chosen to be a compound Poisson process with exponentially distributed jumps, i.e.*

$$L_t = \sum_{k=1}^{N_t} X_k,$$

where (X_k) *i.i.d.with density* $f(x) = 100e^{-100x}$ *and* $N_t \sim Pois(15t)$*. The simulated sample path has 2000 equidistant observations. The variogram* $\gamma(h)$ *and sample variogram* $\widehat{\gamma}(h)$ *for the following parameter sets:*

(i) $p = 1, q = 0, a(z) = z + 0.1, b(z) = 1$

(ii) $p = 2, q = 1, a(z) = z^2 + 0.9z + 0.5, b(z) = 1 + z$

(iii) $p = 2, q = 1, a(z) = z^2 + 0.09z + 0.5, b(z) = 1 + z$

(iv) $p = 3, q = 2, a(z) = z^3 + 1.1z^2 + 2.8174z + 0.2717, b(z) = 1 + 5z + z^2.$

are shown in Figure 7. They are all computed for a maximal lag of 30. Figure 7 shows the flexibility of the CARMA(p,q) process to model a wide variety of correlation structures, represented by a slowly, fast increasing or oscillating variogram.

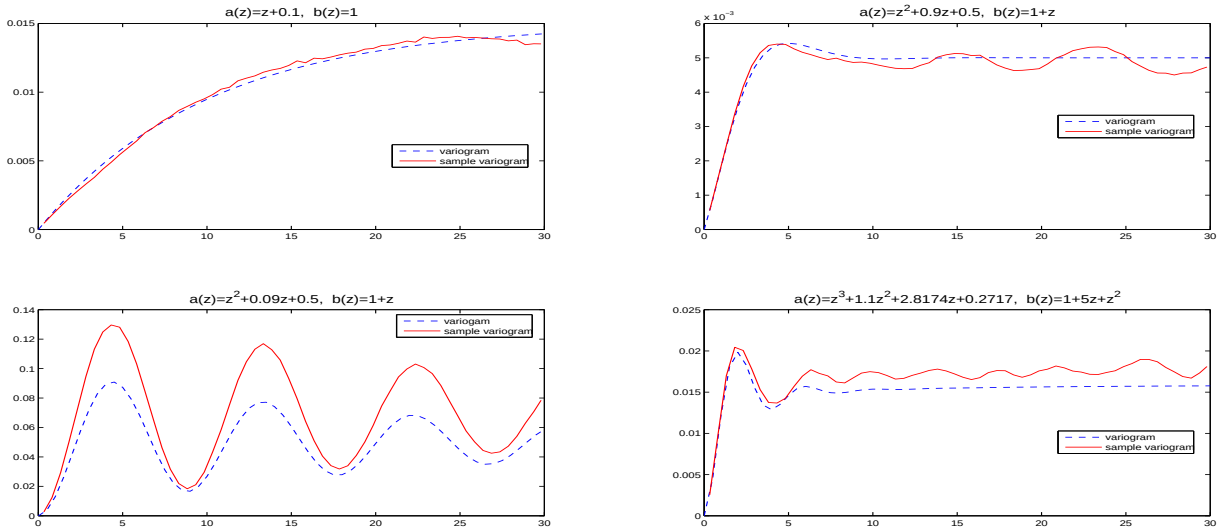

Figure 7: (i) $\gamma(h)$ and $\widehat{\gamma}(h)$ for CAR(1) with $a(z) = z + 0.1$ and $b(z) = 1$ (*top left*), (ii) $\gamma(h)$ and $\widehat{\gamma}(h)$ for CARMA(2,1) with $a(z) = z^2 + 0.9z + 0.5$ and $b(z) = 1 + z$ (*top right*), (iii) $\gamma(h)$ and $\widehat{\gamma}(h)$ for CARMA(2,1) with $a(z) = z^2 + 0.09z + 0.5$ and $b(z) = 1 + z$ (*bottom left*), (iv) $\gamma(h)$ and $\widehat{\gamma}(h)$ for CARMA(3,2) with $a(z) = z^3 + 1.1z^2 + 2.8174z + 0.2717$ and $b(z) = 1 + 5z + z^2$ (*bottom right*)