# On the iterative plug-in algorithm for estimating diurnal patterns of financial trade durations 

Yuanhua Feng ${ }^{1}$, Sarah Forstinger and Christian Peitz<br>University of Paderborn


#### Abstract

This paper discusses the detailed performance of an iterative plug-in (IPI) bandwidth selector for estimating the diurnal duration pattern in a recently proposed semiparametric autoregressive conditional duration (SemiACD) model. For this purpose an alternative formula of the asymptotically optimal bandwidth is proposed. A large simulation study was carried out based on this new formula. The effect of different factors, which affect the selected bandwidth is discussed in detail. It is shown that the proposed IPI algorithm works very well in practice and that the SemiACD model in general, is clearly superior to the parametric ACD model, if there is a deterministic trend in the duration data. It is also shown that the quality of the bandwidth selection, the diurnal pattern estimate and the parametric estimation will all be clearly improved, if the sample size is enlarged. Furthermore, according to the goodness-of-fit of the estimated diurnal pattern, a best combination of the above mentioned factors is found.


Keywords: Autoregressive conditional duration, diurnal duration patterns, local linear estimator, iterative plug-in, simulation

JEL Codes: C14, C41

[^0]
## 1 Introduction

Since the introduction of the ACD (autoregressive conditional duration) model by the seminal work of Engle and Russell (1998), the analysis of financial market behaviour based on transaction durations became one of the most important sub-areas of financial econometrics. Numerous extensions of this model are proposed, including the Log-ACD (Bauwens and Giot, 2000), the class of the augmented ACD models (Fernandes and Grammig, 2006) and the threshold ACD (Zhang et al., 2001). For further information on the development in this context we refer the reader to Pacurar (2008), Engle and Russell (2010), and in particular the monograph of Hautsch (2012) and references therein.

A crucial problem faced by the application of an ACD is that intraday trade durations often exhibit a nonstationary deterministic diurnal pattern (or intraday seasonality), $\phi(t)$ say. The estimation of $\phi(t)$ is necessary for further econometric analysis of trade durations using a stationary ACD model. Different approaches are introduced to deal with $\phi(t)$. For instance, in their original work Engle and Russell (1998) proposed the use of a cubic spline. A nonparametric approach is proposed by Bauwens and Giot (2000). Recently, Rodríguez-Poo et al. (2008) proposed to estimate $\phi(t)$ and the ACD parameters jointly using generalized profile likelihood, which results in a transformed kernel estimator of the nonparametric part. Further approaches for estimating $\phi(t)$ are e.g. linear spline (Dufour and Engle, 2000), wavelet (Bortoluzzo et al., 2009) and shrinkage technique (Brownlees and Gallo, 2011).

Most recently, Feng (2013) proposed a semiparametric ACD (SemiACD) model with a local linear estimator for the diurnal pattern and developed an iterative plug-in (IPI) algorithm (Gasser et al., 1991) for selecting the bandwidth. Here, an inflation method is required to calculate the bandwidth for estimating the second derivative in each iteration. Gasser et al. (1991) proposed to use a so-called MIM (multiplicative inflation method). Beran and Feng (2002) proposed a faster EIM (exponential inflation method) with different possible inflation factors. In this paper we first propose to use the asymptotically optimal bandwidth, $b_{\mathrm{A}}$, obtained by minimizing a partially weighted asymptotic MISE (mean integrated squared error), which is design adaptive and hence a stable criterion. Furthermore, for simplicity and to reduce the computing time we propose to calculate two required integrals numerically at just a few equidistant evaluation points, not at all of the observation points. When the number of evaluation points is not smaller than the root of the sample size, this simplification will not affect the rate of convergence of the bandwidth at all. Furthermore, a closed form formula of $b_{\mathrm{A}}$ under $\operatorname{EACD}(1,1)$ (exponential ACD ), is obtained and employed for assessing
the quality of the selected bandwidth in the simulation. However, the IPI algorithm is developed independently of the ACD specification and the sum of all autocovariances of the stationary part in $b_{\mathrm{A}}$ is estimated by a lag-window estimator (Bühlmann, 1996). Hence, the proposed algorithm is applicable under different ACD models.

The main aim of the current paper is to study the practical performance of IPI bandwidth selector in detail. For this purpose, a large simulation is carried out, which is designed based on 12 main cases defined by two diurnal patterns, two $\operatorname{EACD}(1,1)$ models and three different sample sizes. In each of the main cases 400 replications were generated. For each replication the bandwidth is selected by the EIM sub-method with three inflation factors and by the MIM of Gasser et al. (1991) as well. For each sub-method the bandwidth was again selected using 5 different window-widths for estimating $S$, respectively. This leads to a total of 20 selected bandwidths for each replication. To discuss the effect of the nonparametric estimator on the parametric estimation, $\operatorname{EACD}(1,1)$ models are fitted to the original (nonstationary) data and to the standardized durations with each selected bandwidth. The results are then assessed according to the MSE (mean squared error) of $\hat{b}$ with respect to $b_{\mathrm{A}}$, the goodness-of-fit of the estimated diurnal pattern and the quality of the resulting parameter estimation. The analysis confirmed that the IPI bandwidth selector works well in general. In particular, it is found that the larger the sample size, the better the estimation quality following each of the assessment criteria. Some detailed findings are: 1) According to the MSE of $\hat{b}$, the best bandwidth selector changes from case to case. 2) According to the goodness-of-fit of the estimated diurnal patterns, the difference caused by different sub-methods for selecting the bandwidth is not clear. Nevertheless, a combination is found, which works almost overall the best. We will hence suggest the use of this sub-method in practice. 3) The empirical efficiency of the resulting ACD parameters compared to those obtained from the stationary data are quite different for different parameters. It is clear that the larger the sample size, the higher the estimation quality. These efficiencies even achieved $100 \%$ in many cases. On the other hand, the empirical efficiencies of the parameter estimation are very low, if the parameters are estimated from the nonstationary data directly without removing the diurnal pattern. It is particularly found that, for the scale parameter and the parameter of the latent variable, those efficiencies tend to zero, as $N \rightarrow \infty$. Thus, the estimation and adjustment of the diurnal duration pattern is a necessary step before a parametric ACD is fitted.

This paper proceeds as follows. The model and the estimator are defined in Section 2. The bandwidth selector is proposed in Section 3. Section 4 reports the simulation results. Concluding remarks in Section 5 close the paper.

## 2 The semi-ACD model for diurnal durations

Let $T_{o}=t_{0}<t_{1}<\ldots<t_{N}<t_{N+1}=T_{c}$ be the time points at which trades occur, where $N$ is the (random) number of trades on a trading day, and $T_{o}$ and $T_{c}$ denote the opening and closing times of a stock market. Throughout this paper we will assume that $t_{i}$ are rescaled trading time points such that $T_{o}=0$ and $T_{c}=1$. Let $x_{i}=t_{i}-t_{i-1}$ be the durations between two consecutive trades. A commonly used model for $x_{i}$ (Engle and Russell, 1998) is

$$
\begin{equation*}
x_{i}=\phi\left(t_{i-1}\right) \psi_{i} \varepsilon_{i}, \tag{1}
\end{equation*}
$$

where $\phi\left(t_{i}\right)$ is often called a (deterministic) diurnal pattern, $\psi_{i}$ is the conditional expectation of the diurnally adjusted durations, which follows e.g. some stationary ACD model, and $\varepsilon_{i} \geq 0$ are i.i.d. random variables with $E\left(\varepsilon_{i}\right)=1$. Let $y_{i}=\psi_{i} \varepsilon_{i}$. It is assumed that $E\left(y_{i}\right)=1$ so that the model is uniquely defined, i.e. $y_{i}$ follows a unit ACD with $E\left(\psi_{i}\right)=1$. Engle and Russell (1998) propose to specify $\psi_{i}$ following the idea of the GARCH (generalized autoregressive conditional heteroskedasticity, Engle, 1982 and Bollerslev, 1986) model:

$$
\begin{equation*}
\psi_{i}=\omega+\sum_{j=1}^{p} \alpha_{j} y_{i-j}+\sum_{k=1}^{q} \beta_{k} \psi_{i-k} \tag{2}
\end{equation*}
$$

with a standard exponential distribution of $\varepsilon_{i}$. Due to the restriction $E\left(y_{i}\right)=1$ we have $\omega=1-\sum_{j=1}^{p} \alpha_{j}-\sum_{k=1}^{q} \beta_{k}$. Hence, in a SemiACD the scale parameter $\omega$ is no more free.

Note that $\phi\left(t_{i}\right)$ is (approximately) the local mean of $x_{i}$. However, $x_{i}$ and $\phi\left(t_{i}\right)$ depend strongly on $N$. Under regularity conditions we have indeed $x_{i}=O_{p}\left(N^{-1}\right)$. Hence it is more convenient to study the deterministic pattern in the rescaled durations $z_{i}=N x_{i}$, because the local mean of $z_{i}$ is (approximately) a fixed deterministic function. For given $N$, the estimation of the local mean of $z_{i}$ is equivalent to that of $\phi\left(t_{i}\right)$. Furthermore, we assume that trades on a day occur according to some design density $0<f(t)<\infty$ on $t \in[0,1]$ and define $m(t)=1 / f(t)$ and $\phi_{N}(t)=m(t) / N$. According to Feng (2013), it holds $E\left[z_{i} \mid N\right]=m\left(t_{i-1}\right)\left[1+O_{p}\left(N^{-1}\right)\right]$, where the $O_{p}\left(N^{-1}\right)$ term is caused by the randomness of $t_{i}$.

### 2.1 Local linear estimation of the scale function

Note that $x_{i}$ and $z_{i}$ can be rewritten as special nonparametric regression models as follows:

$$
\begin{equation*}
x_{i}=\phi\left(t_{i}\right)+\phi\left(t_{i}\right)\left(y_{i}-1\right) \tag{3}
\end{equation*}
$$

and

$$
\begin{equation*}
z_{i} \approx m\left(t_{i}\right)+m\left(t_{i}\right)\left(y_{i}-1\right) \tag{4}
\end{equation*}
$$

Now, the derivatives $m^{(\nu)}(t)$ can be estimated by minimizing the weighted least squares

$$
\begin{equation*}
Q=\sum_{i=1}^{N}\left\{z_{i}-a_{0}(t)-a_{1}(t)\left(t_{i}-t\right)-\ldots-a_{d}(t)\left(t_{i}-t\right)^{d}\right\}^{2} K\left(\frac{t_{i}-t}{b}\right) \tag{5}
\end{equation*}
$$

where $K(u)$ is a kernel function and $b$ is the bandwidth. We obtain the estimates $\hat{m}^{(\nu)}(t)=$ $\nu!\hat{a}_{\nu}$, for $\nu \leq d$, and accordingly $\hat{\phi}^{(\nu)}(t)=\frac{\nu!\hat{a}_{\nu}}{N}$. If we put $d=1$ and $\nu=0$, this leads to the local linear estimates $\hat{m}(t)=\hat{a}_{0}$ and $\hat{\phi}(t)=\hat{a}_{0} / N$, which will be used in this paper.

The asymptotic properties of $\hat{m}(t)$ and $\hat{\phi}(t)$ are obtained by Feng (2013). Let $\gamma(k)$ denote the autocovariances of $y_{i}$ and $S=\sum \gamma(k)$ be their sum. Furthermore, let $R(K)=\int K^{2}(u) d u$ and $I(K)=\int u^{2} K(u) d u$ for a kernel function $K$. At an interior point $0<t<T$ the asymptotic variance and asymptotic bias of $\hat{m}(t)$ are given by

$$
\begin{equation*}
\operatorname{var}[\hat{m}(t)] \approx \frac{R(K) S}{N b f(t)} m^{2}(t)=\frac{R(K) S}{N b} m^{3}(t) \tag{6}
\end{equation*}
$$

and

$$
\begin{equation*}
B[\hat{m}(t)]=b^{2} \frac{m^{\prime \prime}(t) I(K)}{2} \tag{7}
\end{equation*}
$$

Accordingly, we have var $[\hat{\phi}(t)] \approx \operatorname{var}[\hat{m}(t)] / N^{2}$ and $B[\hat{\phi}(t)] \approx B[\hat{m}(t)] / N$. Based on equations (6) and (7) the asymptotic mean integrated squared error (AMISE), an approximation of $\operatorname{MISE}(\hat{m})=\int E\left\{[\hat{m}(t)-m(t)]^{2}\right\} d t$, is given by

$$
\begin{equation*}
\operatorname{AMISE}(\hat{m})=b^{4} \frac{\int\left[m^{\prime \prime}(t)\right]^{2} I(K)}{4}+\frac{R(K) S \int m^{3}(t) d t}{N b} \tag{8}
\end{equation*}
$$

By minimizing the AMISE we obtain the asymptotically optimal bandwidth

$$
\begin{equation*}
\tilde{b}_{\mathrm{A}}=\left(\frac{R(K) S}{I^{2}(K)} \frac{I\left(m^{3}\right)}{I\left(\left[m^{\prime \prime}\right]^{2}\right)}\right)^{1 / 5} N^{-1 / 5} \tag{9}
\end{equation*}
$$

where $I\left(m^{3}\right)=\int m^{3}(t) d t$ and $I\left(\left[m^{\prime \prime}\right]^{2}\right)=\int\left[m^{\prime \prime}(t)\right]^{2} d t$. One problem with the above formula is that the $I\left(m^{3}\right)$ term may cause unnecessary instability of the selected bandwidth. To solve this problem we propose to use the following formula of the optimal bandwidth

$$
\begin{equation*}
b_{\mathrm{A}}=\left(\frac{R(K) S}{I^{2}(K)} \frac{I\left(m^{2}\right)}{I\left(\left[m^{\prime \prime}\right]^{2}\right)}\right)^{1 / 5} N^{-1 / 5} \tag{10}
\end{equation*}
$$

which minimizes the dominating part of the partially weighted MISE $\int\left\{B[\hat{m}(t)]^{2}+f(t) V[\hat{m}(t)]\right\} d t$. Note that a SemiACD model can also be applied to model other financial variables such as
daily average durations and daily trade volumes. The formula of $b_{\mathrm{A}}$ in (10) is design adaptive, i.e. it is the same for equidistant, non-equidistant fixed design as well as for random design. Hence, an algorithm developed based on this formula works for SemiACD models in all of these cases. This fact also ensures that many known results on the IPI bandwidth selector with dependent errors can be easily adapted to the one developed in the next section. Furthermore, we will see that by means of this idea the computing time can be reduced clearly without affecting the rate of convergence of the proposed bandwidth selector.

To assess the simulation results, we need to calculate $\tilde{b}_{\mathrm{A}}$ or $b_{\mathrm{A}}$ under given design. Note that $R(K)$ and $I(K)$ are two known constants. The terms $I\left(\left[m^{\prime \prime}\right]^{2}\right), I\left(m^{2}\right)$ or $I\left(m^{3}\right)$ can also be calculated easily. However, the formula of the sum of $\gamma(k)$ for a given ACD model is still unknown in the literature. In the simulation in $\operatorname{Section} 4, \operatorname{EACD}(1,1)$ models will be used. In this case, $S$ can be calculated according to the following lemma.

Lemma 1. If $y_{i}$ follow an $\operatorname{EACD}(1,1)$ with $\epsilon_{i} \sim \exp (1)$ and $\psi_{i}=(1-\alpha-\beta)+\alpha y_{i-1}+\beta \psi_{i-1}$, then the sum of all $\gamma(k)$ of $y_{i}$ is given by

$$
\begin{equation*}
S=\left(\frac{1-\beta}{1-(\alpha+\beta)}\right)^{2} \frac{1-(\alpha+\beta)^{2}}{1-(\alpha+\beta)^{2}-\alpha^{2}} \tag{11}
\end{equation*}
$$

The proof of Lemma 1 is given in the appendix. Note that the above formula only holds for an $\operatorname{EACD}(1,1)$. More general results will not be discussed here. For a given diurnal pattern and given $\operatorname{EACD}(1,1)$, it can be shown that the difference between $\tilde{b}_{\mathrm{A}}$ and $b_{\mathrm{A}}$ is quite small. This confirms that the use of $b_{\mathrm{A}}$ is theoretically and practically reasonable.

### 2.2 Estimation of the ACD parameters

Having estimated the scale function and diurnally adjusted the original duration series, an ACD model can be fitted to the diurnally adjusted durations. Let $\theta$ denote the vector of the unknown $\operatorname{ACD}(\mathrm{p}, \mathrm{q})$ parameters, $\theta=\left(\omega, \alpha_{1}, \ldots, \alpha_{p}, \beta_{1}, \ldots, \beta_{q}\right)^{\prime}$. Assume that $\hat{m}(t)$ and $\hat{\phi}(t)$ are consistent estimates of $m(t)$ and $\phi(t)$, then $\theta$ can be estimated from $\hat{y}_{i}=x_{i} / \hat{\phi}\left(t_{i}\right)$ using the QML (quasi maximum likelihood) method under an $\operatorname{EACD}(p, q)$ assumption as proposed by Engle and Russell (1998) and Engle (2000). If the type of the distribution of $\epsilon_{i}$ is assumed, fully efficient ML estimates of $\theta$ can also be employed. For a detailed description on these topics we refer the reader to Chapter 5.3 of Hautsch (2012) and references therein. The resulting parameter estimate will be denoted by $\hat{\theta}$. Now, assume that $y_{i}=\psi_{i} \epsilon_{i}$ were
observable. The parameter vector $\theta$ could also be estimated from $y_{i}$ using the same method. Denote this (practically unavailable) estimate by $\tilde{\theta}$. It is well known that $\tilde{\theta}$ is $\sqrt{N}$-consistent and asymptotically normal. According to the similarity between the GARCH and the ACD models, consistency and asymptotic properties of $\hat{\theta}$ can be obtained following the ideas of Lemma 1 and Theorem 3 in Feng (2004). These results indicate that $\hat{\theta}$ is also $\sqrt{N}$-consistent and asymptotically normal up to a bias term. Following the proof of Theorem 3 in Feng (2004), we can see that this bias term is the same for kernel and local linear estimates of $m(t)$. Moreover, it is easy to see that this conclusion does not depend on $N$. Hence we have $B(\hat{\theta})=E[\hat{\theta}-\tilde{\theta}]=O\left[b^{2}+(N b)^{-1}\right]$, where the $O\left(b^{2}\right)$ term is due to the integral of the bias $E[\hat{m}(t)-m(t)]$ and the $O\left[(N b)^{-1}\right]$ term is caused by the variance of $\hat{m}(t)$. If a bandwidth $b=O\left(n^{-a}\right)$ with $1 / 4<a<1 / 2$ is used, this bias term is asymptotically negligible. If a bandwidth of the optimal order $O\left(b_{\mathrm{A}}\right)$ is used, we have $B(\hat{\theta})=O\left(N^{-2 / 5}\right)$. Furthermore, if $x_{i}$ follow a parametric ACD model with $\phi(t)$ to be a constant, then $\hat{\theta}$ is $\sqrt{N}$-consistent and asymptotically normal, if $b$ is of a larger order than $O\left(N^{-1 / 2}\right)$. This is particularly true, when $b$ is selected by the proposed data-driven algorithm in the next section. This means that the SemiACD model also works well in the case when the data do not have a diurnal pattern, but with some loss of the efficiency. Proofs of those results are omitted to save space.

For the practical implementation we propose to fit an $\operatorname{EACD}(1,1)$ or another suitable ACD model to $\hat{y}_{i}$ using the fACD package in R. Other available ACD packages in the literature can also be employed for this purpose. As in the parametric case, model selection using the AIC or BIC can also be applied to $\hat{y}_{i}$.

## 3 The bandwidth selection procedure

The IPI bandwidth selector to be proposed extends the original idea of Gasser et al. (1991) in different ways. Let $b_{0}$ denote the starting bandwidth. In the $j$-th iteration, $m^{\prime \prime}(t)$ will be estimated using the bandwidth $b_{2 j}$ calculated from $b_{j-1}$, the selected bandwidth in the j -th iteration. The formula for calculating $b_{2 j}$ from $b_{j-1}$ is called the inflation method. Gasser et al. (1991) propose to use the following MIM inflation form

$$
\begin{equation*}
b_{2 j}=b_{j-1} N^{1 / 10} \tag{12}
\end{equation*}
$$

On the other hand, Beran and Feng (2002) proposed to use a faster EIM inflation form

$$
\begin{equation*}
b_{2 j}=b_{j-1}^{\lambda}, \tag{13}
\end{equation*}
$$

where $0<\lambda<1$ denotes the inflation factor, which determines the rate of convergence of $\hat{b}_{A}$.
Assume that the MIM or the EIM with a suitable value of $\lambda$ is used and that $\hat{S}_{j}$ is calculated from $\hat{\gamma}(k)$ using the Bartlett window $w_{k}=1-k /(L+1)$ with $L=c_{f} N^{1 / 3}$. Let $\sqrt{N}<M<N$ be an odd integer. Define $t_{r}^{*}=(r-1) /(M-1)$ to be $M$ equidistant evaluation points, and $m_{1}=[0.05 *(M-1)]$ and $m_{2}=[0.95 * M]$, where $[\cdot]$ denotes the integer part. The proposed IPI algorithm processes as follows:

Step 1a. In the j-th iteration estimate $\hat{m}_{j}\left(t_{r}^{*}\right), r=1, \ldots, M$, by $b_{j-1}$. Calculate $\hat{I}_{j}\left(m^{2}\right)=$ $\left\{\sum_{r=m_{1}}^{m_{2}}\left[\hat{m}_{j}\left(t_{r}^{*}\right)\right]^{2}\right\} /\left(m_{2}-m_{1}+1\right)$ and $\hat{y}_{j i}=x_{i} / \hat{\phi}_{j}\left(t_{i}\right), i=1, \ldots, N$. Then calculate $\hat{\gamma}_{j}(k)$ from $\hat{y}_{j i}$ and obtain $\hat{S}_{j}=\sum_{|k|<K} w_{k} \hat{\gamma}_{j}(k)$.

Step 1b. Calculate $b_{2 j}$ using the chosen method, estimate $\hat{m}_{j}^{\prime \prime}\left(t_{r}^{*}\right)$ by local cubic regression and calculate $\hat{I}_{j}\left(\left[m^{\prime \prime}\right]^{2}\right)=\left\{\sum_{r=m_{1}}^{m_{2}}\left[\hat{m}_{j}^{\prime \prime}\left(t_{r}^{*}\right)\right]^{2}\right\} /\left(m_{2}-m_{1}+1\right)$.

Step 2. Insert the values of $\hat{I}_{j}\left(m^{2}\right), \hat{S}_{j}$ and $\hat{I}_{j}\left(\left[m^{\prime \prime}\right]^{2}\right)$ into (10) to obtain $b_{j}$.
Step 3. Increase $j$ by one and repeatedly carry out Steps 1 and 2 until convergence or until a given number of iterations is reached. Put $\hat{b}=b_{j}$.

The idea to estimate $\hat{m}_{j}(t)$ and $\hat{m}_{j}^{\prime \prime}(t)$ only at $M$ evaluation points will reduce the computing time very clearly. In particular note that $\hat{m}_{j}^{\prime \prime}(t)$ is a 3rd order local polynomial estimator, which has to be carried out in each iteration. This simplification will not affect the rate of convergence of $\hat{b}$, if $M>\sqrt{N}$, because, the highest rate of convergence of an IPI bandwidth selector in the current context is of the order $O\left(N^{-2 / 7}\right)$. Our empirical experience shows that bandwidths selected by different $M$ values are almost the same. In the simulation in the next section, $M=201$ is fixed to ensure that the large simulation can be finished in a adequate time. Note that even for the smallest sample size there, i.e. $N=8000, M$ is just about $2.5 \%$ of the whole observation time points. Our simulation results show that this simplification works very well in practice. Although it is well known that local polynomial regression has automatic boundary correction, the curve estimation quality at a boundary point is still worse than that at an interior point. This problem was dealt with in two ways. Firstly, at any boundary point, the total bandwidth used is kept to be the same as at an interior point. For instance, for a given bandwidth $b$, the estimation at a point $t<b$ is carried out with all observations within the interval $t_{i} \in[0,2 b]$. Secondly, the integrals $\hat{I}_{j}\left(m^{2}\right)$ and $\hat{I}_{j}\left(\left[m^{\prime \prime}\right]^{2}\right)$ are calculated without the $5 \%$ estimates at each boundary to avoid their effect on the bandwidth selection.

For calculating the standardized durations in the $j$ th iteration, $\hat{\phi}_{j}\left(t_{i}\right)$ are obtained from $\hat{m}_{j}\left(t_{r}^{*}\right)$ by means of linear interpolation. At the beginning, we propose to fix $b_{0}=1 / 10$, so that a rather large number of observations, i.e. $20 \%$ of the observations, is used for estimating the scale function in the first iteration. In general, the finally selected bandwidth does not depend on $b_{0}$, if it set to any reasonable value, because the IPI algorithm is a fix-point search procedure. It can be shown that with the above starting bandwidth $b_{j}$ will become a consistent estimator of $b_{\mathrm{A}}$ in a few iterations. On the other hand, the choice of the inflation method is more important. In this paper we will mainly consider the use of the EIM. Although, as shown by Beran and Feng (2002), inflation factor $\lambda=5 / 7$ will lead to the highest $O\left(N^{-2 / 7}\right)$ convergence rate, Feng (2013), proposed the use of $\lambda=1 / 2$ so that $\hat{b}$ is most stable but with a lower rate of convergence of the order $O\left(N^{-1 / 5}\right)$, because the variation of the intraday durations is very large. This idea was confirmed by the simulation results in the next section. A further choice of $\lambda$ is $\lambda=5 / 9$ to minimize the MSE of $\hat{m}^{\prime \prime}$ with a rate of convergence of the order $O\left(N^{-2 / 9}\right)$. A bandwidth selected by the MIM of Gasser et al. (1991) is also most stable with the rate of convergence of the order $O\left(N^{-1 / 5}\right)$. The error in $\hat{S}$ will cause an additional error term in $\hat{b} / b_{\mathrm{A}}$ of the order $O_{p}\left(N^{-1 / 3}\right)$, which is asymptotically negligible. This fact is not affected by the choice of $c_{f}$. For practical application, we propose to use $\hat{m}\left(t_{i}\right)$ and $\phi\left(t_{i}\right)$ obtained by using the selected bandwidth $\hat{b}$ at all observation points $t_{i}$ as the final estimates.

## 4 The simulation study

In the simulation study different cases were constructed to examine the practical performance of the bandwidth selector in detailed and to see, whether a relatively better combination of the control parameters exists and how the algorithm can be further improved.

### 4.1 Description of the simulation study

Firstly, two diurnal patterns, $m_{1}(t)$ and $m_{2}(t)$, were chosen, where $m_{1}(t)$ exhibits a typical inverse U-shape and $m_{2}(t)$ shows an atypical duration pattern with long durations in the morning and afternoon and comparatively short durations around noon. These two patterns are displayed in Figure 1, which were indeed designed based on the estimated diurnal duration
patterns of the BMW stocks on two trading days in August 2011. The closed function forms are very complex and are hence omitted.

For each diurnal pattern data were generated using two $\operatorname{EACD}(1,1)$ models with:

$$
\begin{equation*}
\mathrm{ACD}_{1}: \psi_{1 i}=0.04+0.09 x_{i-1}+0.87 \psi_{1 i-1} \tag{14}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathrm{ACD}_{2}: \psi_{2 i}=0.04+0.14 x_{i-1}+0.82 \psi_{2 i-1} \tag{15}
\end{equation*}
$$

with $\omega=1-\alpha-\beta$. The simulation was carried out with three different sample sizes $N_{1}=8000, N_{2}=16000$ and $N_{3}=32000$. These combinations define 12 main cases of the simulation in total. For each main case 400 replications were generated. Here, bandwidth selection using four different inflation methods, i.e. the MIM and the three EIM with $\lambda=$ $5 / 7,5 / 9$ and $1 / 2$, are considered. For each method, $\hat{S}$ was then calculated with five values of $c_{f}$, namely $c_{f}=2,4,6,8$ and 10 , respectively. Hence, for each simulated data set the bandwidth was selected by 20 sub-methods separately. In addition, the bisquare kernel is used in all cases as weight function.

### 4.2 Results of the simulation study

The quality of the bandwidth selection is first discussed according to its bias, variance and MSE, and then assessed by the goodness-of-fit, i.e. the corresponding MSE's of the estimated diurnal patterns using the selected bandwidths. Finally, the simulation results are evaluated by the quality of the estimated ACD parameters in each case.

### 4.2.1 Performance of the selected bandwidth

Tables 1 to 3 show the means, standard deviations as well as the mean squared errors of the bandwidths (multiplied by 100) selected in the 400 replications in the corresponding sub-cases for $N_{1}=8000, N_{2}=16000$ and $N_{3}=32000$, respectively, together with the true values of $b_{\mathrm{A}}$ calculated following Lemma 1 (also multiplied by 100). Firstly, we can see that the MSE decreases strongly, when $N$ increases, which indicates that the proposed bandwidth selector is consistent. It is clear that the performance of $\hat{b}$ depends on the form of the diurnal pattern and the properties of the ACD model very strongly. It is the easiest to select the bandwidth for the second diurnal pattern with the first ACD model, while the bandwidth is
very difficult to select in the combination of the first diurnal pattern with the second ACD model. In the former case, the bandwidth can already be selected very well with $N_{1}=8000$. In the latter case the quality of the selected bandwidth with $N_{2}=16000$ is still not good enough. Furthermore, we can see that, if the bandwidth is difficult to select, the effect caused by increasing the sample size is usually more clear. Another more important question, that was to be addressed is whether an overall superior inflation method for selecting a bandwidth can be identified. If the trend is simple, the results suggest to apply the EIM with $\lambda=1 / 2$ but for large sample sizes the MIM also works well. If the trend is more complicated no clear statement can be made on which inflation method is generally superior to the others, as it seems to depend on the features of the ACD model as well as the number of observations.

Concerning the choice of $c_{f}$, the results are ambiguous, as well. For all cases with first trend and sample size $N_{1}$ the optimal $c_{f}$ is 6 . For $N_{2}$ one optimal $c_{f}$ cannot be clearly identified, however $c_{f}=8$ and $c_{f}=10$ can be ruled out to be optimal. For $N_{3}$ as well as almost of the cases simulated with the second trend the majority of smallest MSE values are achieved by $c_{f}=2$. Thus, it is not possible to find an overall superior choice of $c_{f}$. But it seems that the performance of a moderate $c_{f}$ is more stable. Hence we will propose to use $c_{f}=4,6$ or 8 . If $N$ is large enough, $c_{f}=2$ can also be chosen.

### 4.2.2 Goodness of fit of $\hat{m}(t)$

To assess the goodness-of-fit of the data-driven estimate of the diurnal pattern directly, we will define the RASE (the root of the average of the averaged squared errors) as follows. For a given diurnal pattern and sample size, the ASE for the $j$-th replication is defined by

$$
\begin{equation*}
\mathrm{ASE}_{j}=\frac{1}{0.9 N} \sum_{k=0.05 N+1}^{0.95 N}\left(\hat{m}\left(t_{k}\right)-m\left(t_{k}\right)\right)^{2} \tag{16}
\end{equation*}
$$

where again $5 \%$ estimates at each boundary are not used for calculating this criterion. The RASE is then defined as the root of the average of $\mathrm{ASE}_{j}$ over all 400 replications:

$$
\begin{equation*}
\text { RASE }=\sqrt{\frac{1}{400} \sum_{j=1}^{400} \mathrm{ASE}_{j}} . \tag{17}
\end{equation*}
$$

The obtained results of RASE (multiplied by 100) are displayed in Table 4. An important empirical finding is that these results indicate a clear order of the goodness-of-fit of the four methods for calculating $b_{2 j}$. Now the sub-method EIM with $\lambda=1 / 2$ performs the best
overall. The MIM sub-method is the second best one and the EIM with $\lambda=5 / 7$ is the worst. Furthermore, these results also suggest that $c_{f}=2$ should not be used. For the best sub-method, the difference between the results with the other values of $c_{f}$ is unclear, although $c_{f}=6$ or $c_{f}=8$, or sometimes $c_{f}=10$, is usually the best. Note that the main purpose of nonparametric estimation of the diurnal pattern is to fit $m(t)$ as well as possible. Hence we will suggest the use of the EIM with $\lambda=1 / 2$. Also note that the MIM was proposed to achieve a most stable bandwidth. Our simulation results seem to indicate that the stability of the bandwidth selection is more important than the rate of convergence of the bandwidth itself. The simulation results indicate that, following the RASE criterion, a relatively larger value of $c_{f}$ is more preferable. This shows again that the stability of the selected bandwidth plays a more important role for the goodness-of-fit of the resulting curve estimation. Furthermore, these results indicate that the estimation of $m_{1}(t)$ under $\mathrm{ACD}_{1}$ is the easiest, while the estimation of $m_{2}(t)$ under $\mathrm{ACD}_{2}$ is most difficult. Finally, conclusions obtained following the RASE are quite different to those drawn from the MSE of $\hat{b}$.

### 4.2.3 Performance of the ACD parameter estimation

For each of the 400 repetitions of a main case three $\operatorname{EACD}(1,1)$ models were fitted to the duration data simulated without a trend, $y_{i}$, the duration data simulated with a trend, $x_{i}$, and the diurnally adjusted durations, $\hat{y}_{i}=x_{i} / \hat{\phi}\left(t_{i}\right)$. Let $\theta$ denote the true parameter vector $\theta=(\omega, \alpha, \beta)^{\prime}$. Denote by $\tilde{\theta}, \hat{\theta}^{x}$ and $\hat{\theta}^{\hat{y}}$ the estimated parameter vector based on $y_{i}, x_{i}$ and $\hat{y}_{i}$, respectively. For assessing the quality of the parameter estimation, the relative efficiencies (REFF) of $\hat{\theta}^{x}$ and $\hat{\theta}^{\hat{y}}$ with respect to $\tilde{\theta}$ are defined as follows:

$$
\begin{equation*}
\operatorname{REFF}\left(\hat{\theta}^{\hat{y}}\right)=\frac{\operatorname{MSE}(\tilde{\theta})}{\operatorname{MSE}\left(\hat{\theta}^{\hat{y}}\right)} * 100(\%) \text { and } \operatorname{REFF}\left(\hat{\theta}^{x}\right)=\frac{\operatorname{MSE}(\tilde{\theta})}{\operatorname{MSE}\left(\hat{\theta}^{x}\right)} * 100(\%) \tag{18}
\end{equation*}
$$

These results are listed in Tables 5 to 7 for the three parameters, respectively. Theoretically, if there is a deterministic trend in the data $x_{i}, \hat{\theta}^{x}$ is obtained under misspecification and is hence inconsistent. As indicated before, $\hat{\theta}^{\hat{y}}$ is however consistent. These facts can be seen clearly from $\operatorname{REFF}\left(\hat{\theta}^{\hat{y}}\right)$ and $\operatorname{REFF}\left(\hat{\theta}^{x}\right)$ and the comparison between them will indicate the gain in parameter estimation by means of the SemiACD.

Some general findings which we can draw from these results are as follows: 1) The larger $N$, the higher the REFF of the estimated parameters from $\hat{y}_{i}$ but the lower the REFF of those obtained from $x_{i}$. As $N \rightarrow \infty, \hat{\theta}^{\hat{y}}$ will tend to $100 \%$ but $\operatorname{REFF}\left(\hat{\theta}^{x}\right)$ will however tend to zero. This fact can be seen more clearly, if the estimation of $\beta$ is considered. See Table 7. 2) The
quality of $\hat{\omega}^{\hat{y}}$ is the poorest, because $\omega$ is the scale parameter and $\phi(t)$ is the scale function. Indeed, the proposed SemiACD model can be asymptotically rewritten as an ACD with only one time varying scale parameter, while its $\alpha$ and $\beta$ are constant, as in a parametric ACD. 3) The highest REFF's are achieved by $\hat{\alpha}^{\hat{y}}$, where these efficiencies are about $100 \%$ in most cases. Now, the REFF's of $\hat{\alpha}^{x}$ are also high, because $\alpha$ reflects the short term dependence and is not affected by the diurnal pattern so much.

Furthermore, the quality of the parameter estimation based on $\hat{y}_{i}$ depends on the combination of the diurnal pattern and the ACD model very strongly. The case, where the estimation of $\omega$ is the easiest seems to be the combination of $m_{1}(t)$ with $\mathrm{ACD}_{2}$. By the combination of $m_{2}(t)$ and $\mathrm{ACD}_{1}, \omega$ is very difficult to estimate. Now the REFF of $\hat{\omega}^{\hat{y}}$ for $N_{1}=8000$ using any sub-method is clearly smaller than $50 \%$. Similar conclusions can be drawn for $\hat{\beta}^{\hat{y}}$. The difference is only that the REFF's of $\hat{\beta}^{\hat{y}}$ are usually clearly higher than those of $\hat{\omega}^{\hat{y}}$ in corresponding cases.

Concerning the difference caused by the sub-methods for the bandwidth selection we can find that the EIM with $\lambda=1 / 2$ performs usually the best, except for the combination of $m_{2}(t)$ and $\mathrm{ACD}_{2}$. In this case the EIM with $\lambda=5 / 7$ performs slightly better than the other methods. However, we will suggest the use of the EIM with $\lambda=1 / 2$ again, because it seems to be more stable. Note in particular that by the combination of $m_{2}(t)$ and $\mathrm{ACD}_{1}$, the EIM with $\lambda=5 / 7$ performs clearly poorer than all of the other methods. This submethod is hence not a suitable choice. When the sub-method EIM with $\lambda=1 / 2$ is chosen, the difference caused by the choice of $c_{f}$ is usually unclear. In general, all of the $c_{f}$ values perform well. However, we will still suggest the use of $c_{f}=6$ or $c_{f}=8$, because now the proposed algorithm performs more stable than with the other $c_{f}$ values.

### 4.3 Estimation results for two simulated data examples

In order to further illustrate the performance of the proposed algorithm, two simulated data sets were chosen, for which the fitted results using the proposed best algorithm, i.e. the EIM with $\lambda=1 / 2$ and $c_{f}=6$, are shown in more detail. The first example is the first simulated data set in the case with $m_{1}(t), \mathrm{ACD}_{1}$ and $N_{1}$, called Case 111. The second example is the last simulated data set in the case with $m_{2}(t), \mathrm{ACD}_{2}$ and $N_{3}$, called Case 223. The left panels of Figure 2 display the simulated data without trend $y_{i}$, the simulated data with trend, $x_{i}$, the true trend, $m_{1}\left(t_{i}\right)$, together with the estimated trend, $\hat{m}_{1}\left(t_{i}\right)$, and the standardized
duration series, $\hat{y}_{i}$ for Case 111. Those for Case 223 are shown in the right panels of Figure 2 , where all data are displayed against the cumulative sum of $x_{i}$.

Figures 2(e) and 2(f) show that the estimated trend fits the true trend well in both cases. For Case 223, the selected bandwidth seems to be a little small for the second peak as it is not completely smooth at that point. At the beginning and the end, the fit is nearly perfect, though, which also holds for Case 111. This indicates the drawback of the use of a global bandwidth. However, the choice of a local bandwidth would be too complex and is hence not discussed in the current paper. The figures for the diurnally adjusted data further show, that both estimations of the trend are good, as after the removal of the estimated trends, the standardized series in figures $2(\mathrm{~g})$ and $2(\mathrm{~h})$ seem to be quite stationary and look very similar to the originally simulated data without trend in Figure 2(a) and 2(b). The points in the series where the estimated trends did not fit the true trends perfectly are also visible in the standardized duration series. For example, for Case 111 where the estimated trend is above the true trend, the standardized durations are smaller at that point than the simulated data without trend as a trend larger than the true one, was removed at that point. Finally, the fitted ACD models in the two cases are with $\psi_{1 i}=0.038+0.082 x_{i-1}+0.880 \psi_{1 i-1}$ and $\psi_{2 i}=0.040+0.139 x_{i-1}+0.820 \psi_{2 i-1}$, respectively. We see, the selected examples show that the proposed algorithm works very well in practice.

## 5 Conclusion

In this paper a data-driven estimation of the diurnal pattern in a recently proposed SemiACD model is discussed. Detailed results on the bandwidth selection are obtained. A large simulation was carried out to discuss the practical performance of the proposed bandwidth selector in different cases. The results are then assessed in three ways. It is shown that the IPI bandwidth selector works well in general. One of the sub-methods using the EIM inflation form, an inflation factor $\lambda=1 / 2$ and a coefficient $c_{f}=6$ for calculating the lagwindow estimator of the sum of all autocovariances seems to work better than the others in most of the cases, in particular if the performance is assessed using the goodness-of-fit of the estimated diurnal pattern. The results of the parameter estimation further showed that if a significant daily pattern is not removed from the data, the fitted ACD model is inconsistent. Hence, in practice the SemiACD not the stationary parametric ACD should be used.

## Appendix: Proof of Lemma 1

Assume that the true scale functions and ACD model parameters $\omega, \alpha$ and $\beta$ are known. Let $\eta_{i}=y_{i}-\psi_{i}$ be a martingale difference sequence by construction. Following Engle and Russell $(1998)$, the $\operatorname{EACD}(1,1)$ model can be represented as an $\operatorname{ARMA}(1,1)$ model: $y_{i}=\omega+(\alpha+\beta) y_{i-1}-\beta \eta_{i-1}+\eta_{i}$. Based on well known results on the sum of all autocovariances of an $\operatorname{ARMA}(1,1)$ model we have

$$
\begin{equation*}
S=\left(\frac{1-\beta}{1-(\alpha+\beta)}\right)^{2} \sigma_{\eta_{i}}^{2} \tag{A.1}
\end{equation*}
$$

Straightforward calculation leads to

$$
\begin{equation*}
\sigma_{\eta_{i}}^{2}=\operatorname{var}\left(y_{i}\right)+\operatorname{var}\left(\psi_{i}\right)-2 \operatorname{cov}\left(y_{i}, \psi_{i}\right) \tag{A.2}
\end{equation*}
$$

Following Engle and Russell (1998) we have

$$
\operatorname{var}\left(y_{i}\right)=\frac{1-\beta^{2}-2 \alpha \beta}{1-\beta^{2}-2 \alpha \beta-\alpha^{2}}
$$

Bauwens and Giot (2000) showed that $\operatorname{var}\left(\psi_{i}\right)=\frac{\alpha^{2}}{1-\beta^{2}-2 \alpha \beta}$ and

$$
\begin{align*}
\operatorname{cov}\left(y_{i}, \psi_{i}\right) & =E\left[\left(y_{i}-E\left(y_{i}\right)\right)\left(\psi_{i}-E\left(\psi_{i}\right)\right)\right] \\
& =E\left[y_{i} \psi_{i}\right]-E\left[y_{i}\right] E\left[\psi_{i}\right] . \tag{A.3}
\end{align*}
$$

Under the weakly stationarity assumption we have $E\left(y_{i}\right)=E\left(\psi_{i}\right)$. Furthermore, following Bauwens and Giot (2000), we have $E\left[y_{i} \psi_{i}\right]=E\left[\psi_{i}^{2}\right]$. This leads to $\operatorname{cov}\left(y_{i}, \psi_{i}\right)=E\left[\psi_{i}^{2}\right]-\mu^{2}=$ $\operatorname{var}\left(\psi_{i}\right)$ and

$$
\begin{equation*}
\sigma_{\eta_{i}}^{2}=\frac{1-(\alpha+\beta)^{2}}{1-(\alpha+\beta)^{2}-\alpha^{2}} \tag{A.4}
\end{equation*}
$$

Inserting (A.4) into equation (A.1) gives

$$
\begin{equation*}
S=\left(\frac{1-\beta}{1-(\alpha+\beta)}\right)^{2} \frac{1-(\alpha+\beta)^{2}}{1-(\alpha+\beta)^{2}-\alpha^{2}} \tag{A.5}
\end{equation*}
$$

Lemma 1 is proved.


Figure 1: A typical trend and an atypical trend used in the simulation.
(a) Simulated data without trend for Case 111

(c) Simulated data with trend for Case 111

(e) Simulated (black) and estimated (red) trend

(g) Standardized durations for Case 111

(b) Simulated data without trend for Case 223

(d) Simulated data with trend for Case 223

(f) Simulated (black) and estimated (red) trend

(h) Standardized durations for Case 223


Figure 2: Estimation results for the selected examples Case 111 (left) and Case 223 (right).

| $m(t)$ | ACD | $c_{f}$ | $b_{A}$ | EIM, $\lambda=5 / 7$ |  |  | EIM, $\lambda=5 / 9$ |  |  | EIM, $\lambda=1 / 2$ |  |  | MIM |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | mean | SD | MSE | mean | SD | MSE | mean | SD | MSE | mean | SD | MSE |
| 1 | 1 | 2 | 15.98 | 10.51 | 2.24 | 34.85 | 13.95 | 1.69 | 6.94 | 14.40 | 1.50 | 4.70 | 13.58 | 2.15 | 10.36 |
|  |  | 4 |  | 11.33 | 2.37 | 27.21 | 14.71 | 1.79 | 4.82 | 15.17 | 1.58 | 3.14 | 14.54 | 2.17 | 6.79 |
|  |  | 6 |  | 11.49 | 2.45 | 26.12 | 14.91 | 1.84 | 4.54 | 15.40 | 1.62 | 2.96 | 14.75 | 2.20 | 6.33 |
|  |  | 8 |  | 11.43 | 2.51 | 26.95 | 14.94 | 1.89 | 4.62 | 15.44 | 1.67 | 3.07 | 14.77 | 2.25 | 6.52 |
|  |  | 10 |  | 11.24 | 2.58 | 29.07 | 14.88 | 1.95 | 5.00 | 15.37 | 1.72 | 3.34 | 14.69 | 2.32 | 7.03 |
| 1 | 2 | 2 | 18.86 | 10.46 | 2.87 | 78.77 | 15.07 | 2.69 | 21.55 | 15.84 | 2.45 | 15.14 | 14.49 | 3.20 | 29.38 |
|  |  | 4 |  | 11.39 | 3.10 | 65.40 | 16.01 | 2.76 | 15.73 | 16.69 | 2.58 | 11.36 | 15.71 | 3.24 | 20.44 |
|  |  | 6 |  | 11.52 | 3.19 | 64.04 | 16.26 | 2.83 | 14.74 | 16.96 | 2.74 | 11.11 | 15.96 | 3.29 | 19.24 |
|  |  | 8 |  | 11.49 | 3.28 | 65.06 | 16.32 | 2.90 | 14.84 | 17.03 | 2.81 | 11.23 | 16.00 | 3.36 | 19.47 |
|  |  | 10 |  | 11.36 | 3.36 | 67.62 | 16.29 | 2.94 | 15.25 | 17.01 | 2.87 | 11.64 | 15.95 | 3.43 | 20.19 |
| 2 | 1 | 2 | 6.83 | 6.22 | 0.63 | 0.76 | 6.60 | 0.76 | 0.63 | 7.11 | 0.91 | 0.90 | 6.52 | 0.72 | 0.61 |
|  |  | 4 |  | 6.56 | 0.70 | 0.57 | 6.93 | 0.94 | 0.89 | 7.56 | 1.19 | 1.93 | 6.83 | 0.87 | 0.76 |
|  |  | 6 |  | 6.60 | 0.83 | 0.75 | 6.96 | 0.95 | 0.92 | 7.63 | 1.21 | 2.11 | 6.87 | 0.92 | 0.84 |
|  |  | 8 |  | 6.53 | 0.89 | 0.88 | 6.91 | 1.01 | 1.02 | 7.60 | 1.29 | 2.25 | 6.80 | 0.94 | 0.88 |
|  |  | 10 |  | 6.42 | 0.97 | 1.10 | 6.81 | 1.00 | 1.00 | 7.52 | 1.41 | 2.47 | 6.72 | 1.04 | 1.10 |
| 2 | 2 | 2 | 8.06 | 7.08 | 1.27 | 2.59 | 8.34 | 2.58 | 6.72 | 9.16 | 2.75 | 8.79 | 7.85 | 2.35 | 5.55 |
|  |  | 4 |  | 7.60 | 1.72 | 3.17 | 8.84 | 2.86 | 8.81 | 9.73 | 2.94 | 11.41 | 8.34 | 2.81 | 7.99 |
|  |  | 6 |  | 7.73 | 2.00 | 4.10 | 9.03 | 3.22 | 11.33 | 9.93 | 3.24 | 13.99 | 8.49 | 3.08 | 9.67 |
|  |  | 8 |  | 7.71 | 2.13 | 4.66 | 9.06 | 3.42 | 12.65 | 9.97 | 3.48 | 15.75 | 8.46 | 3.20 | 10.39 |
|  |  | 10 |  | 7.64 | 2.32 | 5.55 | 9.05 | 3.64 | 14.18 | 9.93 | 3.61 | 16.48 | 8.41 | 3.28 | 10.90 |


| Table 2: Statistics from the 400 replications for all cases with $N_{2}=16000$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $m(t)$ | ACD | $c_{f}$ | $b_{A}$ | EIM, $\lambda=5 / 7$ |  |  | EIM, $\lambda=5 / 9$ |  |  | EIM, $\lambda=1 / 2$ |  |  | MIM |  |  |
|  |  |  |  | mean | SD | MSE | mean | SD | MSE | mean | SD | MSE | mean | SD | MSE |
|  |  | 2 |  | 10.73 | 1.65 | 12.85 | 13.16 | 1.13 | 1.83 | 13.52 | 0.96 | 1.08 | 13.28 | 1.28 | 2.03 |
|  |  | 4 |  | 11.46 | 1.74 | 9.02 | 13.81 | 1.20 | 1.45 | 14.12 | 1.03 | 1.10 | 13.97 | 1.29 | 1.66 |
| 1 | 1 | 6 | 13.91 | 11.61 | 1.79 | 8.47 | 13.99 | 1.25 | 1.56 | 14.29 | 1.07 | 1.29 | 14.13 | 1.34 | 1.85 |
|  |  | 8 |  | 11.62 | 1.82 | 8.52 | 14.02 | 1.29 | 1.68 | 14.33 | 1.11 | 1.40 | 14.16 | 1.38 | 1.97 |
|  |  | 10 |  | 11.58 | 1.83 | 8.77 | 14.00 | 1.34 | 1.80 | 14.32 | 1.14 | 1.46 | 14.15 | 1.42 | 2.08 |
|  |  | 2 |  | 10.97 | 2.35 | 35.26 | 14.28 | 1.79 | 7.76 | 14.77 | 1.71 | 5.61 | 14.57 | 2.04 | 7.55 |
|  |  | 4 |  | 11.74 | 2.48 | 28.00 | 15.08 | 1.86 | 5.24 | 15.50 | 1.79 | 4.06 | 15.38 | 2.02 | 5.17 |
| 1 | 2 | 6 | 16.42 | 11.94 | 2.56 | 26.55 | 15.29 | 1.91 | 4.94 | 15.71 | 1.84 | 3.89 | 15.56 | 2.05 | 4.91 |
|  |  | 8 |  | 11.95 | 2.59 | 26.68 | 15.33 | 1.95 | 4.98 | 15.76 | 1.87 | 3.95 | 15.59 | 2.09 | 5.03 |
|  |  | 10 |  | 11.90 | 2.63 | 27.36 | 15.32 | 1.98 | 5.12 | 15.74 | 1.90 | 4.07 | 15.59 | 2.14 | 5.27 |
|  |  | 2 |  | 5.59 | 0.35 | 0.26 | 5.82 | 0.20 | 0.05 | 6.07 | 0.24 | 0.07 | 5.93 | 0.23 | 0.05 |
|  |  | 4 |  | 5.85 | 0.38 | 0.15 | 6.04 | 0.25 | 0.07 | 6.39 | 0.57 | 0.51 | 6.17 | 0.44 | 0.25 |
| 2 | 1 | 6 | 5.95 | 5.88 | 0.40 | 0.17 | 6.09 | 0.53 | 0.30 | 6.44 | 0.60 | 0.61 | 6.20 | 0.52 | 0.33 |
|  |  | 8 |  | 5.84 | 0.43 | 0.19 | 6.05 | 0.46 | 0.22 | 6.41 | 0.59 | 0.57 | 6.16 | 0.51 | 0.31 |
|  |  | 10 |  | 5.77 | 0.46 | 0.24 | 6.00 | 0.56 | 0.31 | 6.36 | 0.73 | 0.70 | 6.09 | 0.39 | 0.17 |
|  |  | 2 |  | 6.51 | 1.04 | 1.34 | 7.19 | 1.61 | 2.62 | 7.79 | 1.77 | 3.73 | 7.08 | 1.60 | 2.56 |
|  |  | 4 |  | 6.93 | 1.32 | 1.74 | 7.51 | 1.66 | 3.00 | 8.19 | 1.70 | 4.27 | 7.35 | 1.63 | 2.75 |
| 2 | 2 | 6 | 7.02 | 7.04 | 1.47 | 2.17 | 7.66 | 1.95 | 4.22 | 8.39 | 2.28 | 7.07 | 7.44 | 1.87 | 3.67 |
|  |  | 8 |  | 7.02 | 1.49 | 2.22 | 7.70 | 2.26 | 5.56 | 8.43 | 2.40 | 7.74 | 7.51 | 2.21 | 5.11 |
|  |  | 10 |  | 7.00 | 1.68 | 2.80 | 7.66 | 2.30 | 5.70 | 8.38 | 2.40 | 7.60 | 7.49 | 2.28 | 5.42 |


| Table 3: Statistics from the 400 replications for all cases with $N_{3}=32000$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $m(t)$ | ACD | $c_{f}$ | $b_{A}$ | EIM, $\lambda=5 / 7$ |  |  | EIM, $\lambda=5 / 9$ |  |  | EIM, $\lambda=1 / 2$ |  |  | MIM |  |  |
|  |  |  |  | mean | SD | MSE | mean | SD | MSE | mean | SD | MSE | mean | SD | MSE |
|  |  | 2 |  | 10.11 | 1.20 | 5.42 | 12.02 | 0.75 | 0.57 | 12.34 | 0.61 | 0.42 | 12.26 | 0.72 | 0.54 |
|  |  | 4 |  | 10.62 | 1.24 | 3.75 | 12.53 | 0.78 | 0.78 | 12.83 | 0.64 | 0.92 | 12.78 | 0.69 | 0.93 |
| 1 | 1 | 6 | 12.11 | 10.76 | 1.26 | 3.42 | 12.67 | 0.80 | 0.96 | 12.96 | 0.66 | 1.16 | 12.93 | 0.69 | 1.15 |
|  |  | 8 |  | 10.79 | 1.28 | 3.38 | 12.72 | 0.82 | 1.04 | 13.01 | 0.67 | 1.26 | 12.97 | 0.71 | 1.25 |
|  |  | 10 |  | 10.77 | 1.29 | 3.44 | 12.72 | 0.83 | 1.07 | 13.01 | 0.69 | 1.29 | 12.98 | 0.72 | 1.28 |
|  |  | 2 |  | 10.85 | 2.00 | 15.84 | 13.58 | 1.35 | 2.32 | 13.97 | 1.23 | 1.63 | 13.92 | 1.33 | 1.90 |
|  |  | 4 |  | 11.49 | 2.03 | 11.99 | 14.18 | 1.43 | 2.05 | 14.55 | 1.30 | 1.75 | 14.49 | 1.31 | 1.75 |
| 1 | 2 | 6 | 14.29 | 11.68 | 2.08 | 11.16 | 14.34 | 1.47 | 2.15 | 14.71 | 1.33 | 1.93 | 14.67 | 1.30 | 1.83 |
|  |  | 8 |  | 11.72 | 2.12 | 11.10 | 14.40 | 1.49 | 2.24 | 14.77 | 1.34 | 2.03 | 14.71 | 1.33 | 1.94 |
|  |  | 10 |  | 11.70 | 2.15 | 11.32 | 14.41 | 1.52 | 2.33 | 14.78 | 1.37 | 2.11 | 14.70 | 1.36 | 2.02 |
|  |  | 2 |  | 5.03 | 0.24 | 0.08 | 5.24 | 0.15 | 0.03 | 5.32 | 0.17 | 0.05 | 5.37 | 0.16 | 0.07 |
|  |  | 4 |  | 5.23 | 0.26 | 0.07 | 5.39 | 0.17 | 0.07 | 5.51 | 0.21 | 0.16 | 5.55 | 0.18 | 0.17 |
| 2 | 1 | 6 | 5.18 | 5.26 | 0.28 | 0.08 | 5.41 | 0.19 | 0.09 | 5.55 | 0.23 | 0.19 | 5.57 | 0.19 | 0.19 |
|  |  | 8 |  | 5.24 | 0.29 | 0.09 | 5.40 | 0.20 | 0.09 | 5.54 | 0.25 | 0.20 | 5.56 | 0.21 | 0.19 |
|  |  | 10 |  | 5.21 | 0.31 | 0.09 | 5.37 | 0.21 | 0.08 | 5.52 | 0.26 | 0.18 | 5.53 | 0.20 | 0.17 |
|  |  | 2 |  | 5.83 | 0.51 | 0.34 | 6.16 | 0.70 | 0.49 | 6.47 | 0.68 | 0.60 | 6.23 | 0.73 | 0.55 |
|  |  | 4 |  | 6.14 | 0.75 | 0.56 | 6.38 | 0.73 | 0.60 | 6.80 | 0.87 | 1.22 | 6.43 | 0.72 | 0.62 |
| 2 | 2 | 6 | 6.11 | 6.22 | 0.89 | 0.81 | 6.44 | 0.88 | 0.88 | 6.89 | 1.11 | 1.84 | 6.49 | 0.98 | 1.10 |
|  |  | 8 |  | 6.21 | 0.94 | 0.89 | 6.47 | 1.02 | 1.16 | 6.93 | 1.27 | 2.28 | 6.48 | 0.96 | 1.06 |
|  |  | 10 |  | 6.18 | 0.94 | 0.89 | 6.44 | 1.03 | 1.17 | 6.89 | 1.21 | 2.06 | 6.46 | 1.01 | 1.14 |




| $\lambda$ | InfM | $c_{f}$ | $(1,1,1)$ | $(1,1,2)$ | $(1,1,3)$ | $(1,2,1)$ | $(1,2,2)$ | $(1,2,3)$ | $(2,1,1)$ | $(2,1,2)$ | $(2,1,3)$ | $(2,2,1)$ | $(2,2,2)$ | $(2,2,3)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5/7 | EIM | 2 | 100.9 | 100.4 | 100.3 | 99.2 | 100.0 | 99.5 | 95.5 | 98.7 | 100.4 | 96.8 | 99.6 | 100.3 |
|  |  | 4 | 101.4 | 100.5 | 100.0 | 99.4 | 100.0 | 99.7 | 95.6 | 99.1 | 100.0 | 95.4 | 99.0 | 100.3 |
|  |  | 6 | 101.4 | 100.3 | 99.2 | 99.4 | 100.1 | 99.7 | 95.3 | 89.9 | 99.8 | 96.7 | 10.2 | 100.5 |
|  |  | 8 | 101.4 | 100.2 | 100.6 | 99.4 | 100.0 | 99.7 | 95.5 | 98.6 | 100.5 | 96.9 | 99.5 | 100.4 |
|  |  | 10 | 101.1 | 100.3 | 100.0 | 99.4 | 99.9 | 99.6 | 94.8 | 98.8 | 100.3 | 96.6 | 99.0 | 100.3 |
| 5/9 | EIM | 2 | 101.0 | 100.5 | 99.3 | 99.3 | 100.3 | 100.0 | 95.6 | 99.1 | 100.2 | 97.5 | 100.2 | 100.1 |
|  |  | 4 | 101.0 | 100.4 | 99.7 | 99.9 | 100.3 | 100.0 | 95.3 | 98.9 | 99.9 | 98.1 | 100.1 | 100.2 |
|  |  | 6 | 101.3 | 100.2 | 99.8 | 99.3 | 100.3 | 100.1 | 92.9 | 98.9 | 99.7 | 98.3 | 100.0 | 100.1 |
|  |  | 8 | 101.2 | 100.1 | 99.4 | 99.2 | 100.2 | 100.0 | 94.4 | 98.6 | 100.6 | 98.2 | 100.2 | 99.9 |
|  |  | 10 | 101.3 | 100.2 | 99.5 | 99.1 | 100.3 | 100.1 | 94.8 | 98.7 | 100.1 | 98.3 | 100.4 | 100.0 |
| 1/2 | EIM | 2 | 101.5 | 100.4 | 98.7 | 98.9 | 99.9 | 100.1 | 95.4 | 98.5 | 100.3 | 99.2 | 99.1 | 100.1 |
|  |  | 4 | 101.5 | 100.4 | 99.2 | 99.9 | 100.3 | 99.9 | 96.1 | 98.7 | 100.2 | 98.8 | 100.1 | 100.2 |
|  |  | 6 | 101.0 | 100.3 | 99.5 | 99.5 | 100.2 | 99.8 | 95.6 | 98.6 | 100.6 | 99.0 | 100.1 | 100.3 |
|  |  | 8 | 101.3 | 100.2 | 99.3 | 99.7 | 100.2 | 99.9 | 96.2 | 98.8 | 100.2 | 99.1 | 99.8 | 100.2 |
|  |  | 10 | 101.3 | 100.2 | 99.8 | 99.7 | 100.2 | 99.9 | 97.0 | 98.6 | 99.8 | 98.7 | 99.7 | 100.4 |
| - | MIM | 2 | 101.1 | 100.1 | 99.6 | 99.2 | 100.1 | 99.6 | 96.3 | 99.0 | 99.8 | 95.7 | 99.7 | 95.9 |
|  |  | 4 | 101.0 | 100.1 | 99.9 | 99.8 | 100.4 | 99.2 | 96.3 | 98.8 | 100.6 | 98.1 | 98.9 | 100.1 |
|  |  | 6 | 101.1 | 100.1 | 99.5 | 99.7 | 100.4 | 99.8 | 95.0 | 98.9 | 100.3 | 98.3 | 100.0 | 100.4 |
|  |  | 8 | 100.9 | 100.3 | 99.5 | 99.7 | 100.1 | 99.9 | 92.1 | 98.8 | 100.3 | 98.4 | 100.5 | 100.2 |
|  |  | 10 | 100.9 | 100.2 | 99.9 | 99.9 | 100.4 | 99.9 | 94.7 | 98.2 | 99.9 | 96.8 | 99.7 | 100.1 |
|  | x |  | 88.7 | 79.9 | 67.4 | 54.2 | 90.5 | 91.1 | 90.6 | 74.5 | 63.5 | 91.7 | 89.8 | 90.2 |



## References

[1] Bauwens, L. and Giot, P. (2000). The Logarithmic ACD model: An Application to the Bid-Ask Quote Process of three NYSE stocks. Annales d'Economie et de Statistique, 60, 117-149.
[2] Beran, J. and Feng, Y. (2002). Iterative plug-in algorithms for SEMIFAR models - definition, convergence and asymptotic properties. J. Computat. Graph. Statist., 11, 690-713.
[3] Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. J. Econometr., 31, 307-327.
[4] Bortoluzzo, A., Morettin, P. and Toloi, C., (2010). Time-varying autoregressive conditional duration model. J. Appl. Statist., 37, 847-864.
[5] Brownlees, C. and Gallo, G. (2011). Shrinkage estimation of semiparametric multiplicative error models. Intern. J. Forecasting, 27, 365-378.
[6] Bühlmann, P. (1996). Locally ataptice lag.window spectral estimation. J. Time Ser. Anal., 17, 247-270.
[7] Dufour, E. and Engle, R. (2000). The ACD Model: Predictability of the time between two consecutive trades. ICMA Centre Discussion Papers in Finance, Henley Business School, Reading University.
[8] Engle, R. (1982). Autoregressive Conditional Heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50, 987-1007.
[9] Engle, R. (2000). The Econometrics of Ultra-high frequency Data. Econometrica, 68, 1-22.
[10] Engle, R. (2002). New frontiers for ARCH models. J Appl. Economet., 17, 425-446.
[11] Engle, R. and Russell, J. (1998). Autoregressive Conditional Duration model: A new model for irregularly spaced transaction data. Econometrica, 66, 1127-1162.
[12] Engle, R. and Russell, J. (2010). Analysis of high frequency data. In Ait-Sahalia, Y.
and Hansen, L.P. (eds): Handbook of Financial Econometrics, Vol. 1, pp. 383-426. Elsevier, Amsterdam.
[13] Feng, Y. (2004). Simultaneously modelling coditional heteroskedasticity and scale change. Econometr. Th., 20, 563-569.
[14] Feng, Y. (2013). Data-driven estimation of diurnal duration patterns. CIE Disscussion Paper, University of Paderborn, revised.
[15] Fernandes, M., Grammig, J. (2006). A family of autoregressive conditional duration models. J. Econometr., 130, 1-23.
[16] Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. J. Amer. Statist. Ass., 86, 43-652.
[18] Hautsch, N. (2012). Econometrics of Financial High-Frequency Data. Springer, Berlin.
[19] Pacurar, M. (2008). Autoregressive conditional duration models in Finance: A survey of the theoretical and empirical literature. J. Econ. Surv., 22, 711-751.
[20] Rodríguez-Poo, J.M., Veredas, D. and Espasa, A. (2008). Semiparametric estimation for financial durations. In: Bauwens, L., Pohlmeier, W. and Veredas, D. (eds): High frequency financial econometrics. Physica-Verlag, Heidelberg, pp. 225-251.
[21] Zhang, M., Russell, J. and Tsay, R. (2001). A nonlinear autoregressive conditional duration model with applications to financial transaction data. J. Econometr., 104, 179-207.


[^0]:    ${ }^{1}$ Corresponding author: University of Paderborn, Warburger Str. 100, 33098 Paderborn, Germany. Email: yuanhua.feng@wiwi.upb.de; Tel. 0049-5251-603379; Fax. 0049-5251-605005.

