UNIVERSITY OF CAMBRIDGE
Faculty of Economics

Cambridge Working Paper in Economics

Cambridge Working Paper in Economics: 1606

SPLINE-DCS FOR FORECASTING TRADE VOLUME IN HIGH-FREQUENCY FINANCIAL DATA

Ryoko Ito
(University of Oxford)

We introduce the spline-DCS model for forecasting the high-frequency trade volume of selected equity and foreign currency exchange pairs. The empirical application illustrates that spline-DCS captures salient features of the data and is robust to the choice of sampling frequency or sampling period. The predictive performance of the model is compared with the state-of-the-art volume forecasting model, named the component-MEM, of Brownlees et al. (2011). The model out-performs the component-MEM in minimizing a common robust loss function and the slicing loss function of Brownlees et al. (2011) for Volume-Weighted Average Price trading.

Spline-DCS for Forecasting Trade Volume in High-Frequency Finance

Ryoko Ito[1]

Cambridge Working Papers in Economics CWPE1606
This version: May 24, 2016

Abstract

We develop the spline-DCS model and apply it to trade volume prediction, which remains a highly non-trivial task in high-frequency finance. Our application illustrates that the spline-DCS is computationally practical and captures salient empirical features of the data such as the heavy-tailed distribution and intra-day periodicity very well. We produce density forecasts of volume and compare the model's predictive performance with that of the state-of-the-art volume forecasting model, named the component-MEM, of Brownlees et al. (2011). The spline-DCS significantly outperforms the component-MEM in predicting intra-day volume proportions.

Keywords: order slicing; price impact; robustness; score; VWAP trading
JEL Classification: C22, C51, C53, C58, G12

1 Introduction

A key objective of execution algorithms in high-frequency trading is to minimize the price impact of a given order by slicing it into smaller transaction sizes and spreading the timing of transactions throughout the day. This reduces the risk of slippage in price, which is the difference between the expected price of a trade and its actual traded price. Accurate intra-day volume prediction can help investors optimize the size and the timing of orders in this sense since the level of market liquidity and trade intensity change throughout the day. It also helps investors achieve the execution price of transactions for the day to be near the Volume-Weighted Average Price (VWAP) benchmark.[2] It is a measure widely-used for a range of purposes, such as assessing the performance of a given trading

[1]Nuffield College and the Department of Economics, Oxford University. Email: ryoko.ito@economics.ox.ac.uk. I would like to thank Andrew Harvey for his thoughtful comments and guiding my research. I would also like to thank Jamie Walton and Zhangbo Shi for giving me the opportunity to undertake this project. I am also grateful to Philipp Andres, Michele Caivano, Adam Clements, Oliver Linton, Donald Robertson, Mark Salmon, Stephen Thiele, and the participants of the Score Workshop in 2013 at Tinbergen Institute, especially Andre Lucas, for their helpful comments. Finally, I would like to thank the Institute for New Economic Thinking at the Oxford Martin School, the International Monetary Fund, the Cambridge Trust, the Royal Economic Society, the Keynes Fund, and the Stevenson Fund (of the Faculty of Economics at Cambridge University) for providing various forms of funding.

[2]For a given asset class or an order, it is the average transaction price weighted by the size of each transaction.

strategy in minimizing the price impact, or as a guarantee to clients that their orders will be executed at the VWAP target.

Despite the importance of volume prediction in high-frequency trading, it remains a highly non-trivial task due to the statistically complex features of trade volume. Until the seminal work by Brownlees et al. (2011), there had been no well-established methodology for forecasting high-frequency trade volume. Brownlees et al. (2011) introduced the components multiplicative error model (MEM), and showed that the model can outperform some of the existing common methods for volume prediction.

We introduce the spline-DCS model and use it to forecast high-frequency trade volume of popular assets in the equity and foreign currency exchange (FX) markets. We show that the model captures salient empirical features such as intra-day periodic patterns, autocorrelation, and heavy-tail, and outperforms the component-MEM in minimizing the slicing loss function proposed by Brownlees et al. (2011) for assessing the optimality of trading strategies in achieving the VWAP target. The performance of our model is robust to the choice of sampling frequency and sampling period.

The spline-DCS is an extension of the dynamic conditional score (DCS) model, which is a new observation-driven model formally introduced by Creal et al. (2011, 2013) and Harvey (2013).[3] The time-varying parameter in DCS is driven by the score of the conditional distribution of the data. Recent empirical studies find that the score driven models capture heavy-tails well and outperform existing common forecasting methodologies including GARCH-type models of comparable specifications in a range of literature.[4] DCS is also extended to time-varying copula functions, non-negative distributions, and multivariate distributions in applications including inflation forecasting in macroeconomics, forecasting value-at-risk, modeling credit or sovereign-default risk, modeling mixed-measurement and mixed-frequency panel data, and dynamic location modeling.[5]

Key aspects of our analysis can be summed up in two. First, we produce both density and level forecasts of volume using the spline-DCS and illustrate the model's practicality and good in-sample and out-of-sample performance in the context of both equity and FX. The estimation results are robust to the choice of initial parameters, sampling frequency, and sampling period. The sampling frequency we consider ranges between 30 seconds and 10 minutes, which is high in the volume prediction literature.

Second, we highlight the computational and practical appeal of the spline-DCS that stem from the use of the method of maximum likelihood (ML) and the spline. A typical analysis in high-frequency finance deals with a very large set of data, which makes

[3]It is also called the generalized autoregressive score (GAS) model.

[4]See, for instance, Harvey and Sucarrat (2014), Janus et al. (2014), Harvey and Lange (2015), Gao and Zhou (2016), Lucas and Zhang (2016), Creal et al. (2011), Avdulaj and Barunik (2015), and Salvatierra and Patton (2015).

[5]See, for instance, Creal et al. (2014), Harvey and Luati (2014), Lucas et al. (2014), and Caviano and Harvey (2014), as well as the above references.

forecasting computationally very intensive. But the spline-DCS is found to be remarkably easy and quick to estimate, even when the sample size is large. With the sample size of our data ranging between about 5,000 and 20,000, the spline-DCS was estimated in about 5 minutes, as opposed to hours for the component-MEM (hereafter, c-MEM), which was estimated by the generalized method of moments (GMM). This feature of the spline-DCS means that the model can be regularly re-estimated at little computational cost. The c-MEM model is an observation-driven model akin to GARCH. The periodic component adopts the Fourier series, which is commonly used to approximate intra-day periodic patterns. The periodic component of the spline-DCS is a cubic spline function; it falls in the category of smoothing splines, which are successfully applied to modeling regular patterns in electricity demand, money supply, and yield curves.[6] The spline beautifully captures smooth intra-day periodic patterns with few parameters, and is estimated simultaneously with all other components of the model by ML. Our spline in this paper assumes that the pattern of periodicity is the same every day. Ito (2013) challenges this standard assumption in the literature using the spline-DCS with a dynamic spline, which allows the pattern of periodicity to evolve, and show some empirical merit of this generalization.

The plan of this paper is as follows. Section 2 describes the characteristics of our data and motivates the construction of our model. Section 3 describes the spline-DCS and the estimation method. Sections 4 and 5 report the in-sample and out-of-sample results for the spline-DCS. Section 5.2 compares the predictive performance of competing models. Section 5.3 discusses the aforementioned computational and practical aspects of the spline-DCS. Section 6 concludes.

2 Data characteristics

The equity trade volume we consider is the number of shares of IBM stock traded on the New York Stock Exchange (NYSE) during the market opening hours (9.30am-4pm in the New York local time) between Monday 28 February and Friday 31 March 2000, which includes 25 trading days and no public holidays. In order to explore the sensitivity of our model to the choice of sampling frequency, we consider two sampling frequencies for IBM of 30 seconds and 1 minute. There are 780 observations per trading day if the aggregation interval is 30 seconds, and 390 observations if 1 minute. We refer to the aggregated series as IBM30s if the aggregation interval is 30 seconds, and IBM1m if 1 minute. IMB30s aggregates to IBM1m over any 1-minute interval.

We also consider the trade volume of two of the most popular currency exchange pairs

[6]See, for instance, Harvey and Koopman (1993), Hendricks et al. (1979), Harvey et al. (1997), Hyndman et al. (2005), Bowsher and Meeks (2008) and Jungbacker et al. (2014).

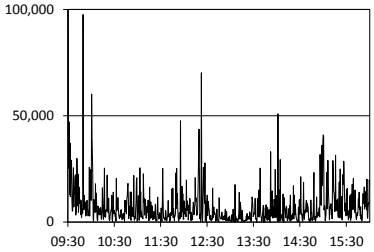

Figure 1. IBM30s on Wednesday 22 March 2000 between 9.30am and 4pm (in the New York local time).

	In-sample (3 weeks)		Out-of-sample (2 weeks)	
Window #	From	To	From	To
1	Mon 06-Jan-14	Sun 26-Jan-14	Mon 27-Jan-14	Sun 09-Feb-14
2	Mon 27-Jan-14	Sun 16-Feb-14	Mon 17-Feb-14	Sun 02-Mar-14
3	Mon 17-Feb-14	Sun 09-Mar-14	Mon 10-Mar-14	Sun 23-Mar-14
4	Mon 10-Mar-14	Sun 30-Mar-14	Mon 31-Mar-14	Sun 13-Apr-14
5	Mon 31-Mar-14	Sun 20-Apr-14	Mon 21-Apr-14	Sun 04-May-14
6	Mon 21-Apr-14	Sun 11-May-14	Mon 12-May-14	Sun 25-May-14
7	Mon 12-May-14	Sun 01-Jun-14	Mon 02-Jun-14	Sun 15-Jun-14
8	Mon 02-Jun-14	Sun 22-Jun-14	Mon 23-Jun-14	Sun 06-Jul-14
9	Mon 23-Jun-14	Sun 13-Jul-14	Mon 14-Jul-14	Sun 27-Jul-14
10	Mon 14-Jul-14	Sun 03-Aug-14	Mon 04-Aug-14	Sun 17-Aug-14
11	Mon 04-Aug-14	Sun 24-Aug-14	Mon 25-Aug-14	Sun 07-Sep-14
12	Mon 25-Aug-14	Sun 14-Sep-14	Mon 15-Sep-14	Sun 28-Sep-14
13	Mon 15-Sep-14	Sun 05-Oct-14	Mon 06-Oct-14	Sun 19-Oct-14
14	Mon 06-Oct-14	Sun 26-Oct-14	Mon 27-Oct-14	Sun 09-Nov-14
15	Mon 27-Oct-14	Sun 16-Nov-14	Mon 17-Nov-14	Sun 30-Nov-14
16	Mon 17-Nov-14	Sun 07-Dec-14	Mon 08-Dec-14	Sun 21-Dec-14

Table 1. Sub-sampling windows for the FX trade volume data. The overall sampling period between Mon 6 Jan and Sun 21 Dec 2014 is split into sixteen sub-sample windows.

in FX: euro-dollar (EURUSD) and dollar-yen (USDJPY).[7] The definition of volume here is the traded units of the left hand currency (e.g. euro for EURUSD and dollar for USDJPY), which is priced in the units of the right hand currency of the same pair (e.g. the US dollar for EURUSD and yen for USDJPY).[8] The sampling frequency we consider here is 10 minutes. We have 144 observations per trading day for the FX data. For a given sampling frequency, the sample size per day for the FX data is more than 3 times larger than for the equity data. This is because FX transactions take place 24 hours during the weekdays and so the data is collected 24 hours every day. The sampling period is between Monday 6 January and Sunday 21 December 2014, which is split into sixteen rolling sub-sample windows for the in-sample and out-of-sample analysis. See Table 1.

Figure 1 gives a snapshot of the equity data. The volume fluctuates a lot throughout the day. There is a diurnal U-shaped pattern in trading activity. Morning transactions

[7]We have more recent and larger samples for the FX data compared to the equity data. This is simply due to the data availability, as we did not have access to more recent volume data for equity at the level of sampling frequency we desired to study. We study both the equity and FX data to illustrate the usefulness of our model in these two disperate applications.

[8]For confidentiality reasons, we divided the original FX trade volume series by some arbitrarily chosen constant number to hide the actual level of volume. This pre-estimation transformation only shifts the intercept parameter, ω, (defined in Section 3) of the dynamic equation.

Series	Obs.	Mean	S.D.	Skew	Max.	Max-99% Q	Zero freq.
IBM30s	19,500	10,539	26,071	29	1,652,100	1,591,073	0.47%
IBM1m	9,750	21,297	39,114	18	1,652,100	1,532,175	0.06%

Series	Obs.	Mean	S.D.	Skew	Max	Max-99%Q	Zero freq.
EURUSD (10 mins)	50,112	136	316	9.5	10,017	8,668	2.1%
USDJPY (10 mins)	50,112	121	250	8.8	7,379	6,322	2.0%

Table 2. Sample statistics of IBM trade volume (top) and EURUSD and USDJPY trade volume (bottom). Sampling period is Mon 28 Feb - Fri 31 Mar 2000 for IBM and Mon 6 Jan - Fri 19 Dec 2014 for the FX data. The skewness statistics must be interpreted with care as the theoretical skewness may not exist.

Window	1	2	3	4	5	6	7	8
EURUSD (10 mins)	0.1%	0.1%	0.6%	0.8%	2.6%	2.2%	2.2%	2.1%
USDJPY (10 mins)	0.2%	0.0%	0.4%	0.9%	1.9%	2.4%	2.1%	1.9%

Window	9	10	11	12	13	14	15	16
EURUSD (10 mins)	2.7%	1.9%	1.8%	1.3%	1.0%	0.8%	0.3%	0.7%
USDJPY (10 mins)	2.7%	1.9%	1.5%	1.3%	0.9%	0.7%	0.4%	0.0%

Table 3. The percentage of samples for each window that are zero-valued. The sixteen sub-sample windows are listed in Table 1.

are driven by overnight news. The activity level bottoms out at around 1pm, but picks up again in the afternoon as traders re-balance their positions before the market closes. (See Hautsch (2012, p.41).)

Figure 2 shows that the FX trade volume also exhibits intra-day periodic patterns. The intra-day percentage distribution of volume appears to have a bimodal pattern for EURUSD and a trimodal pattern for USDJPY. Volume peaks at around 8am and again at around 2pm, but stays low in the evening in GMT. For USDJPY, volume peaks also at around 1am in GMT. These modes come roughly when trading activity in major markets around the world is high for the day.[9] The patterns here are very different from the U-shaped one for the equity data because the dynamics of the equity data is dominated by the overnight effect, whereas the dynamics of the FX data is dominated by the timing of peaks in trade intensity around the world.

Figure 3 shows that the FX volume also fluctuates a lot throughout the day. On Friday 5 September 2014, there is an extreme spike in USDJPY volume at around 1.30pm in GMT. A similar spike was observed for EURUSD. They coincide with the release of non-farm payroll data in the US, which is one of the most important events in FX. This highlights the importance of the announcement effect (see, for instance, Andersen and

[9]In GMT outside the daylight saving period, trading is very active in London between 8am and 4pm, in New York between 1pm and 9pm, and in Tokyo between 11pm and 7am. However, trading is not restricted to these hours; for instance, many traders in London trade between 7am and 5pm in the London local time. For EURUSD, volume is particularly high between 1pm and 4pm when the active period in London and New York overlaps. The London and New York markets attract high volume since the bid-ask spread tends to be tighter there for popular currency pairs than in Asia.

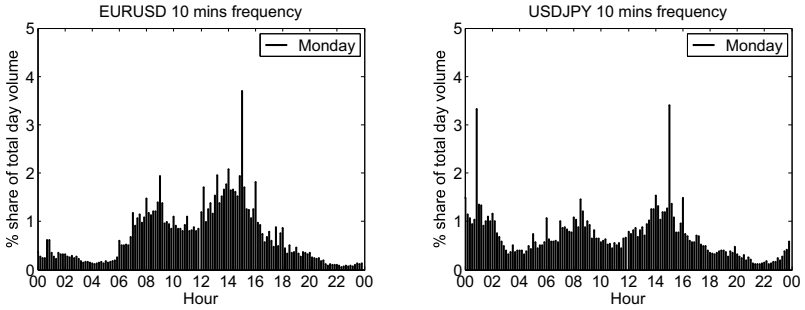

Figure 2. The percentage of total day volume attributed to each intra-day bin. The x-axis is intra-day hours in GMT on Monday. The series are obtained by computing the average trade volume at each intra-day bin on Mondays, and dividing it by the total trade volume of that weekday. The series sum to 100% each day. This uses the data between Monday 6 January and Friday 19 December 2014.

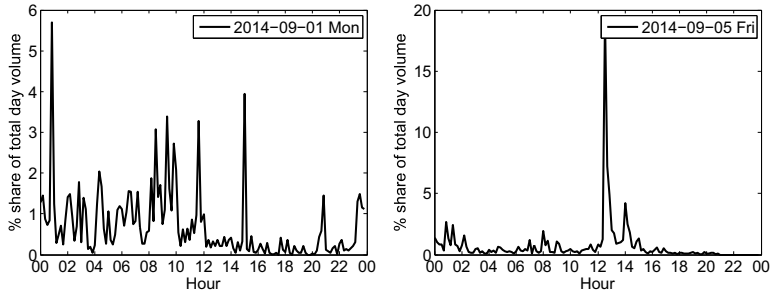

Figure 3. The percentage of total day volume of USDJPY attributed to each intra-day 10-minute bin. The series sum to 100% in each picture. Left: Monday 1 September 2014. Right: Friday 5 September 2014. Each day covers the 24-hour period (in GMT). Intra-day time on the x-axis.

Bollerslev (1998) and Lo and Wang (2010)).

Figure 4 suggests that the empirical distributions exhibit fat-tails. The size of the upper-tail is also suggested by the difference between the maximum and the 99% sample quantile in Table 2. The right column of Figure 4 shows the dynamic structure of our data. The autocorrelation decays slowly for IBM30s. The autocorrelation of the FX data peaks at the 144th lag, reflecting daily periodicity. Our data contains a non-negligible number of zero-valued observations. (See Table 2.) Table 3 shows the percentage of samples for each sub-sampling window that are zero-valued for the FX data. These numbers can be compared with the estimated parameter value of p (defined in Section 3), which is the probability mass of zero-valued observations in the spline-DCS model.

The discussion so far suggests that our model needs a periodic component to capture the intra-day periodic patterns. The non-periodic component should allow for highly persistent dynamics. This can be captured by a combination of autoregressive components. The presence of a non-negligible number of zero-valued observations can be captured by decomposing the distribution of the data to place a discrete probability mass at zero.

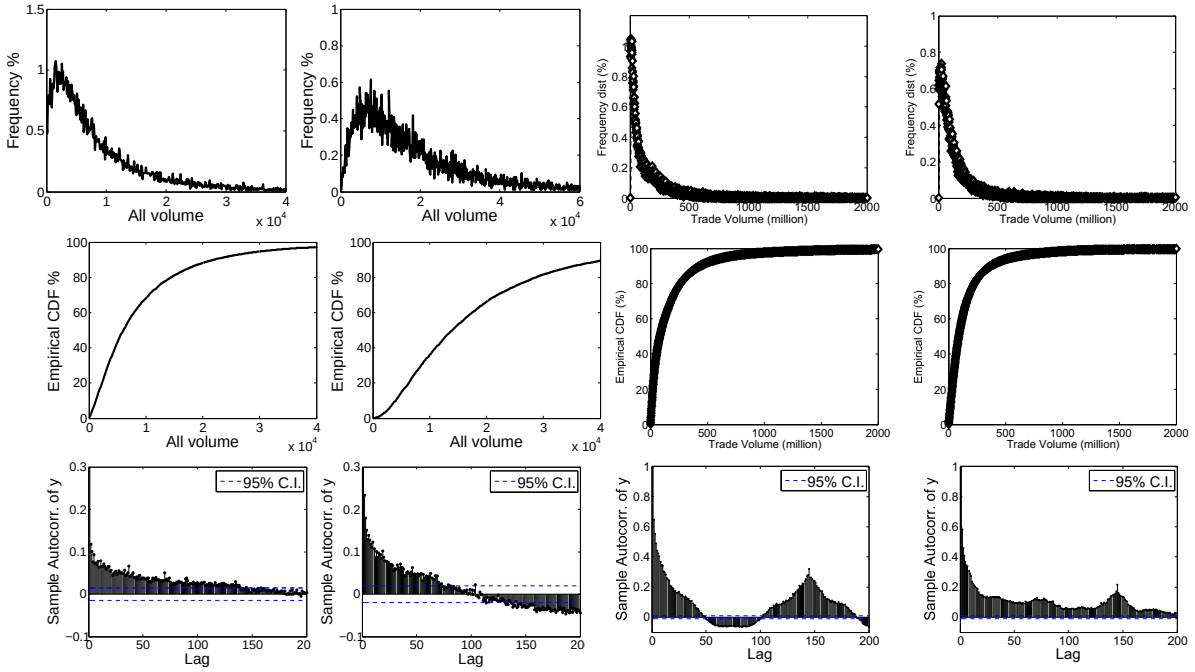

Figure 4. The empirical frequency distribution (top), the empirical cumulative distribution function (middle), and the sample autocorrelation (bottom). IBM30s (left), IBM1m (center left), EURUSD (center right), and USDJPY (center right). The sampling period is Mon 28 Feb - Fri 31 Mar 2000 for the equity data and Mon 6 Jan - Fri 19 Dec 2014 for the FX data. The 200th lag corresponds approximately to 1.5 hours prior for IBM30s, 3 hours prior for IBM1m, and 1.4 days prior for the FX data.

3 The spline-DCS model

We divide each trading day into $I \in \mathbb{N}_{>0}$ intra-day bins. $T \in \mathbb{N}_{>0}$ and $H \in \mathbb{N}_{>0}$ denote the number of in-sample and out-of-sample days, respectively. $y_{t,\tau}$ denotes the observation of trade volume at the τ-th intra-day bin on the t-th trading day. We set $y_{t,I} = y_{t+1,0}$ for all $t \in \mathbb{N}_{>0}$, so that $\tau = 1$ is the location of the first aggregated observation each day. We denote in-sample estimates by $\widehat{\cdot}$ and forecast quantities by $\widetilde{\cdot}$. We use the set notations, $\Psi_{T,\,I} = \{(t,\tau) \in \{1,2,\ldots,T\} \times \{1,2,\ldots,I\}\}$ and $\Psi_{T,I>0} = \{\Psi_{T,\,I} : y_{t,\tau} > 0\}$. The set of all information available at time $(t,\tau) \in \Psi_{T,\,I}$ is denoted by $\mathcal{F}_{t,\tau}$. Any variable at time $(t,\tau) = (1,1)$ is constant. The model is given by:

$$
\begin{aligned}
&y_{t,\tau} = \varepsilon_{t,\tau} \exp(\lambda_{t,\tau}), \qquad \varepsilon_{t,\tau} \sim \text{i.i.d. } F(\varepsilon;\boldsymbol{\theta}), \qquad \lambda_{t,\tau} = \omega + \mu_{t,\tau} + \eta_{t,\tau} + s_{t,\tau} + e_{t,\tau},\\
&\mu_{t,\tau} = \mu_{t,\tau-1} + \kappa_\mu u_{t,\tau-1}, \qquad \eta_{t,\tau} = \eta^{(1)}_{t,\tau} + \eta^{(2)}_{t,\tau}, \qquad e_{t,\tau} = \phi_e e_{t,\tau-1} + \boldsymbol{\kappa}_e^\top \boldsymbol{d}_{t,\tau},\\
&\eta^{(1)}_{t,\tau} = \phi^{(1)}_1 \eta^{(1)}_{t,\tau-1} + \phi^{(1)}_2 \eta^{(1)}_{t,\tau-2} + \kappa^{(1)}_\eta u_{t,\tau-1} + \kappa^{(1)}_{\eta,\mathrm{a}} \mathrm{sign}(-r_{t,\tau-1})(u_{t,\tau-1} + \nu\xi),\\
&\eta^{(2)}_{t,\tau} = \phi^{(2)}_1 \eta^{(2)}_{t,\tau-1} + \kappa^{(2)}_\eta u_{t,\tau-1} + \kappa^{(2)}_{\eta,\mathrm{a}} \mathrm{sign}(-r_{t,\tau-1})(u_{t,\tau-1} + \nu\xi),\\
&\boldsymbol{d}_{t,\tau} = (d_{t,\tau,1},\ldots,d_{t,\tau,m}), \quad d_{t,\tau,i} = \mathbb{1}_{\{\text{type } i \text{ event at time } (t,\tau)\}}, \quad i = 1,\ldots,m.
\end{aligned}
\tag{1}
$$

for $(t,\tau) \in \Psi_{T,I}$ and $\omega \in \mathbb{R}$. The non-periodic components are $\mu_{t,\tau}$ and $\eta_{t,\tau}$. $e_{t,\tau}$ is the event component. The periodic component, $s_{t,\tau}$, is defined first in Section 3.1. Then we

define the distribution function, $F(\cdot;\cdot)$, and the score variable, $u_{t,\tau}$. Then we define the rest of the components.

3.1 The cubic spline

The periodic component, $s_{t,\tau}$, captures the pattern of intra-day periodicity. The notations and specification follow Poirier (1973) and Harvey and Koopman (1993). We refer to the version of the spline studied here as the *static* (cubic) spline, which assumes that the pattern of periodicity does not change over time. This is a standard assumption in the existing literature.[10] Some of the technical details are omitted in the following sections, but we give the complete mathematical construction in Appendix B of the supplementary material.

3.1.1 Static daily spline

The cubic spline is termed a *daily* spline if the periodicity is complete over one trading day. The static daily spline assumes that the shape of intra-day periodic patterns is the same every day. The daily spline is a continuous piecewise function of time and connected at $k+1$ knots for some $k \in \mathbb{N}_{>0}$ such that $k < I$. The coordinates of the knots along the time axis are denoted by $\tau_0 < \cdots < \tau_k$, where $\tau_0 = 1$, $\tau_k = I$, and $\tau_j \in \{2, \ldots, I-1\}$ for $j = 1, \ldots, k-1$. The set of the knots is also called *mesh*. The y-coordinates (height) of the knots are denoted by $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_k)^\top$. The static daily spline ($s_{t,\tau} = s_\tau$) is defined as

$$s_\tau = \sum_{j=1}^{k} \mathbb{1}_{\{\tau \in [\tau_{j-1}, \tau_j]\}} \, \boldsymbol{z}_j(\tau) \cdot \boldsymbol{\gamma}, \qquad \tau = 1, \ldots, I, \tag{2}$$

where $\boldsymbol{z}_j : [\tau_{j-1}, \tau_j]^{k+1} \to \mathbb{R}^{k+1}$ for $j = 1, \ldots k$ is a $(k+1)$-dimensional vector of deterministic functions that conveys all information about the *polynomial order*, *continuity*, and *zero-sum conditions* of the spline. The zero-sum condition and setting $\gamma_k = -\sum_{i=0}^{k-1} w_{*i}\gamma_i / w_{*k}$ ensure that the parameters in $\boldsymbol{\gamma}$ are identified. See Appendix B of the supplementary material for the derivation of $\boldsymbol{z}_j(\tau)$ and $\boldsymbol{w}_* = (w_{*0}, \ldots, w_{*k})^\top$.

For the equity data, we capture the overnight effect that arises from the regular overnight market closure by relaxing the *periodicity condition* of the spline, and allowing for a discrepancy in the spline between the end and the beginning of any two consecutive trading days (i.e. $(\tau_k, \gamma_k) \neq (\tau_0, \gamma_0)$). The definition of $\boldsymbol{z}_j(\tau)$ given above is this case, and it is different from the one defined by Harvey and Koopman (1993). See Appendix B.1 of the supplementary material. Harvey and Koopman (1993) impose the periodicity condition, since their hourly electricity demand data is collected 24 hours a day. The spline we use for the FX data maintains the periodicity condition (i.e. $(\tau_k, \gamma_k) \neq (\tau_0, \gamma_0)$), and it is defined in Appendix B.2 of the supplementary material. (2) is the version for

[10] Ito (2013) challenges this assumption by introducing a generalized spline-DCS specification with the dynamic version of the cubic spline. The empirical merit of such a generalization in high-frequency finance is illustrated in the paper.

the equity data and capture the overnight effect.

3.1.2 Location of daily knots and overnight effect

The location of knots, $\tau_1, \ldots, \tau_{k-1}$, and the size of k depend on the empirical shape of periodicity and the number of intra-day observations. Increasing k does not necessarily improve the fit of the model, and using too many knots deteriorates the speed of computation. We give tips on how to select the location and the number of knots in Section 5.3.

For the FX data, based on the empirical observations we made in Section 2 and from Figure 2, we find that placing knots along the intra-day time axis (in 24-hour format) at 1hr, 2hr, 3.30hr, 5hr, 6hr, 7hr, 8hr, 9.30hr, 11hr, 12hr, 13hr, 14hr, 15hr, 16hr, 17.30hr, 19hr, 20hr, 21hr, 22hr, 23hr, and 24hr works well. The period between Friday 10pm and Sunday 10pm is treated as the regular weekend period of missing data. Although there is no decisive moments on Friday and Sunday at which transactions end and begin for the week, we omit data points over the above weekend period for simplicity. Then, in the static daily spline for FX, we impose the periodicity condition on the knots at Friday 10pm and Sunday 10pm.

For the equity data, we find that placing knots along the intra-day time axis (in 24-hour format) at 9.30hr, 11hr, 12.30hr, 14.30hr, and 16hr works well. The shape of the spline up to 12.30pm captures the busy trading hours in the morning, between 12.30pm and 2.30pm captures the quiet lunch hours, and after 2.30pm captures any acceleration in trading activities before the market closes. There is little to no improvement in the goodness of fit of the model to the data when the number of knots per day increases from these specifications.

3.2 Distribution: dealing with zero-valued observations

The cumulative distribution function (c.d.f.), $F : \mathbb{R}_{\geq 0} \to [0, 1]$, with the constant parameter vector $\boldsymbol{\theta}$ of a standard random variable $X \sim F$ is defined as

$$\mathbb{P}_F(X = 0) = p \in (0, 1), \quad \mathbb{P}_F(X > 0) = 1 - p, \quad \mathbb{P}_F(X \leq x | X > 0) = F^*(x; \boldsymbol{\theta}^*).$$

for any $x > 0$. $F^* : \mathbb{R}_{\geq 0} \to [0, 1]$ is the c.d.f. of some conventional standard continuous random variable with the time-invariant parameter vector $\boldsymbol{\theta}^*$. We write $\boldsymbol{\theta} = (\boldsymbol{\theta}^{*\top}, p)^\top$ and use the notations, f and f^*, to denote the probability density function (p.d.f.) of F and F^*, respectively.

The distribution, F, captures the probability mass of zero-valued observations. The unconditional n-th moment of X is well-defined as long as it is well-defined for F^* because $\mathbb{E}[X^n] = (1 - p) \int_0^\infty x^n f^*(x) dx$. F^* is chosen parametrically, and the quality of its fit to the empirical distribution of data is tested using the standardized observations $\widehat{\varepsilon}_{t,\tau} \equiv y_{t,\tau} / \exp(\widehat{\lambda}_{t,\tau})$. The properties of this type of distributions are studied formally in Hautsch et al. (2014). This decomposition technique is akin to the ones studied by McCulloch

and Tsay (2001) and Rydberg and Shephard (2003).

$u_{t,\tau}$ is the score of the conditional distribution of the data given by $u_{t,\tau} = \partial \log\left(\exp(-\lambda_{t,\tau}) f^*(\varepsilon_{t,\tau}; \boldsymbol{\theta}^*)\right) / \partial \lambda_{t,\tau}$ for all $y_{t,\tau} > 0$. Here, $\exp\left(-\lambda_{t,\tau}\right) f^*(\varepsilon_{t,\tau}; \boldsymbol{\theta}^*) = f_y^*(y_{t,\tau}; \boldsymbol{\theta}^*, \lambda_{t,\tau})$ is the likelihood of a single positive observation. We set $u_{t,\tau} = \inf_{s \in \Omega} u_{t,\tau}(s)$ whenever $y_{t,\tau} = 0$.

Our choice of $F^*(\cdot; \boldsymbol{\theta}^*)$ is GB2, which is found to capture the empirical distribution of the data well. It nests or relates to several well-known distributions including log-logistic, Burr, Pareto, and Weibull. (See Appendix A of the supplementary material and Kleiber and Kotz (2003).) In this case, $f_y^*(y_{t,\tau}; \boldsymbol{\theta}^*, \lambda_{t,\tau})$ is

$$\log(\nu) - \nu\xi\lambda_{t,\tau} + (\nu\xi - 1)\log(y_{t,\tau}) - \log B(\xi, \zeta) - (\xi + \zeta)\log\left(\left(y_{t,\tau}\exp\left(-\lambda_{t,\tau}\right)\right)^\nu + 1\right),$$

where $\boldsymbol{\theta}^* = (\nu, \xi, \zeta)^\top > \mathbf{0}$ are the shape parameters and $B(\cdot, \cdot)$ is the beta function. Then the score is $u_{t,\tau} = \nu(\xi + \zeta)(y_{t,\tau}\exp\left(-\lambda_{t,\tau}\right))^\nu / \left(\left(y_{t,\tau}\exp\left(-\lambda_{t,\tau}\right)\right)^\nu + 1\right) - \nu\xi$ for $y_{t,\tau} > 0$, and we set $u_{t,\tau} = -\nu\xi$ whenever $y_{t,\tau} = 0$. GB2 is formally defined in Appendix A.1 of the supplementary material.

p may change throughout the day because the probability of observing a trade must change with the level of trading activity. Rydberg and Shephard (2003) and Hautsch et al. (2014) independently study decomposition models that allow p to change over time. A natural extension of our model would also consider this generalization. However, in the context of this paper, we assume p to be constant for simplicity. This is inconsequential for us as the fraction of zero-valued observations is small.

3.3 Non-periodic components

The stationary component is $\eta_{t,\tau}$. $\eta_{t,\tau}$ consists of two stationary components, $\eta_{t,\tau}^{(1)}$ and $\eta_{t,\tau}^{(2)}$. This structure allows us to capture highly persistent dynamics similar to long memory.[11] $\eta_{t,\tau}^{(1)}$ and $\eta_{t,\tau}^{(2)}$ are stationary if $-\phi_1^{(1)} + \phi_2^{(1)} < 1$, $\phi_2^{(1)} > -1$, $0 < \phi_1^{(1)} + \phi_2^{(1)} < 1$, and $0 < \phi_1^{(2)} < 1$ (see, for instance, Harvey (1993, p.19)). The non-stationary component, $\mu_{t,\tau}$, captures the slowly changing movements that is non-periodic. The estimation results in Section 4 suggest that this component can do a good job in capturing the low-frequency dynamics of our data.

The role of each component is such that $\mu_{t,\tau}$ should be less sensitive to changes in $u_{t,\tau-1}$ than $\eta_{t,\tau}^{(1)}$, which should be, in turn, less sensitive than $\eta_{t,\tau}^{(2)}$. That is, we should typically expect $|\kappa_\mu| < |\kappa_\eta^{(1)}| < |\kappa_\eta^{(2)}|$ (although this condition is not imposed during the estimation). Moreover, the scale of trade volume should increase in the wake of positive news. Thus we would expect $\kappa_\mu > 0$. We set $\eta_{1,1}^{(1)} = \eta_{1,1}^{(2)} = 0$ as we have

[11] A generalization of $\eta_{t,\tau}$ is $\eta_{t,\tau} = \sum_{j=1}^{J} \eta_{t,\tau}^{(j)}$, where $\eta_{t,\tau}^{(j)} = \phi_1^{(j)}\eta_{t,\tau-1}^{(j)} + \phi_2^{(j)}\eta_{t,\tau-2}^{(j)} + \cdots + \phi_{m^{(j)}}^{(j)}\eta_{t,\tau-m^{(j)}}^{(j)} + \kappa_\eta^{(j)} u_{t,\tau-1}$ for $(t,\tau) \in \Psi_{T,I}$, $j = 1, \ldots, J$, and $J \in \mathbb{N}_{>0}$. We assume that $m^{(j)} \in \mathbb{N}_{>0}$ and $\eta_{t,\tau}^{(j)}$ is stationary for all $j = 1, \ldots, J$. $J = 2$ works well for our application.

$\mathbb{E}[\eta^{(1)}_{t,\tau}] = \mathbb{E}[\eta^{(2)}_{t,\tau}] = 0$.[12] Since $\mathbb{E}[\mu_{t,\tau}] = \mu_{1,1}$, we assume $\mu_{1,1} = 0$ so that ω is identified. The identification conditions of the parameters in $s_{t,\tau}$ are as laid out in Section 3.1.

3.3.1 Asymmetric effect

For the equity data, analogously to the well-documented leverage effects in equity return volatility, we can test for asymmetric effects in volume related to the direction of price change by testing the significance of the coefficients, $\kappa^{(1)}_{\eta,\mathrm{a}}$ and $\kappa^{(2)}_{\eta,\mathrm{a}}$. $\kappa^{(1)}_{\eta,\mathrm{a}} > 0$ (or $\kappa^{(1)}_{\eta,\mathrm{a}} < 0$) gives an increase (decrease) in the scale of volume when price falls (i.e. when the return, $r_{t,\tau}$, which is the log-difference in price, is negative). We use the sign function to capture the asymmetric effect of price change in both the positive and negative directions. That is, the sign function with $\kappa^{(i)}_{\eta,\mathrm{a}} > 0$ (or $\kappa^{(i)}_{\eta,\mathrm{a}} < 0$) gives a decrease (increase) in the scale of volume when price increases for $i = 1, 2$. How to model leverage effects in the DCS models is discussed in Harvey (2013).

In FX, testing for the asymmetric effect or its interpretation is less straight forward than equity. For instance, a fall in the price of the US dollar per euro is an increase in the price of euro per dollar. A sudden sizable strengthening of one currency does not necessarily trigger panic reactions or asymmetric effects, unless the change was against a strong market-wide expectation. Thus, we set $\kappa^{(i)}_{\eta,\mathrm{a}} = 0$ for $i = 1, 2$ for the FX data for simplicity.

3.3.2 Announcement effect

The event component, $e_{t,\tau}$, captures the effect of anticipated macroeconomic events. Its dynamics are assumed to be deterministic. Any deviation of market response from the deterministic pattern at each event is assumed to be captured by other non-deterministic components. We set $e_{1,1} = 0$. $e_{t,\tau}$ reverts to zero if $|\phi_e| < 1$.

Since there are many events per day that can impact our FX data, we categorize events by the anticipated size of the impact using the information provided in the Forex Economic Calendar by DailyFX (www.dailyfx.com) as a benchmark. Then we assign a dummy variable $(d_{t,\tau,i}, i = 1, \ldots, m)$ to each category. The small events tabulated in Table 4 include scheduled releases of retail sales data, manufacturing data, home sales data, and various indicators of house prices. The intermediate events include some central bank announcements and other data releases (e.g. GDP data, employment figures, and consumer price data) in relevant currency areas. The release of US non-farm payroll data on the first Friday of each month is assigned its own category since it is the most important event for the dollar. The elements of $\boldsymbol{d}_{t,\tau}$ are ordered by the anticipated size of the impact. The size of m varies across the sub-sample windows. The first element of $\boldsymbol{d}_{t,\tau}$ correspond to the event category with the largest anticipated impact for that window. If multiple events of the same category occur simultaneously, they are treated as one event

[12] With the asymmetry terms, this assumes that $r_{t,\tau}$ and $y_{t,\tau}$ are independent for any $(t, \tau) \in \Psi_{T,I}$, and $\mathbb{E}[r_{t,\tau}] = 0$.

Currency	EURUSD		USDJPY	
Event category	Count	Frequency (%)	Count	Frequency (%)
Small	1024	93%	543	90%
Intermediate	65	6%	51	8%
US non-farm payroll	12	1%	12	2%

Table 4. Event schemes between Monday 6 January - Sunday 21 December 2014.

of that category.

For the equity data, events that matter include the company's quarterly or annual earnings and dividend announcements, the earnings and dividend announcements by the competitors (e.g. Accenture, Hewlett-Packard, and Microsoft), and important news in the technology industry. To our knowledge, IBM did not make any earnings or dividend related announcements during the sampling period. We could not find the exact timing of news releases in intra-day hours for all of the companies mentioned here. Due to information limitation, we exclude the event component, $e_{t,\tau}$, for the equity data for simplicity.

3.4 The estimation method

All of the parameters of the model are estimated by ML. The joint log-likelihood function given by $F(\cdot;\boldsymbol{\theta})$ for the set of observations $(y_{t,\tau})_{(t,\tau)\in\Psi_{T,I}}$ is

$$\log L = A\log(1-p) + (T\times I - A)\log(p) + \sum_{(t,\tau)\in\Psi_{T,I>0}} \log f_y^*(y_{t,\tau};\boldsymbol{\theta}^*,\lambda_{t,\tau}),$$

where $A = |\Psi_{T,I>0}|$. It is easy to check that the ML estimator (MLE) of p is $\widehat{p} = (T\times I - A)/(T\times I)$. How to compute analytical standard errors is outlined in Appendix D of the supplementary material. We simulate the asymptotic distribution of MLE in the spline-DCS and check its large-sample behavior in Appendix C of the supplementary material. The results suggest consistency and validate standard statistical inference for model selection using t-statistics of this estimator at our sample sizes.

4 In-sample estimation results

For the FX data, using GB2 as the error distribution (F^*) achieved a good fit to the empirical distribution of the data. Burr, which is a special case of GB2 with $\xi = 1$, was found to fit the empirical distribution of the equity data well.[13] Figure 5 illustrates the impressive fit of GB2. The empirical c.d.f. of non-zero $\widehat{\varepsilon}_{t,\tau}$ appears to overlap the theoretical c.d.f. of GB2$(\widehat{\nu},\widehat{\xi},\widehat{\zeta})$ or Burr$(\widehat{\nu},\widehat{\zeta})$. The closeness of the fit can be also checked by inspecting the the probability integral transform (PIT) of non-zero $\widehat{\varepsilon}_{t,\tau}$ computed under the assumption that it is from $F^*(\cdot;\widehat{\boldsymbol{\theta}}^*)$. The empirical c.d.f. of the PIT values lies

[13] GB2 did not fit the equity data well since the discreteness of volume due to market microstructure dominated the empirical distribution when the sampling frequency is as high as 1 minute. Restricting distribution parameters and testing nested distributions worked.

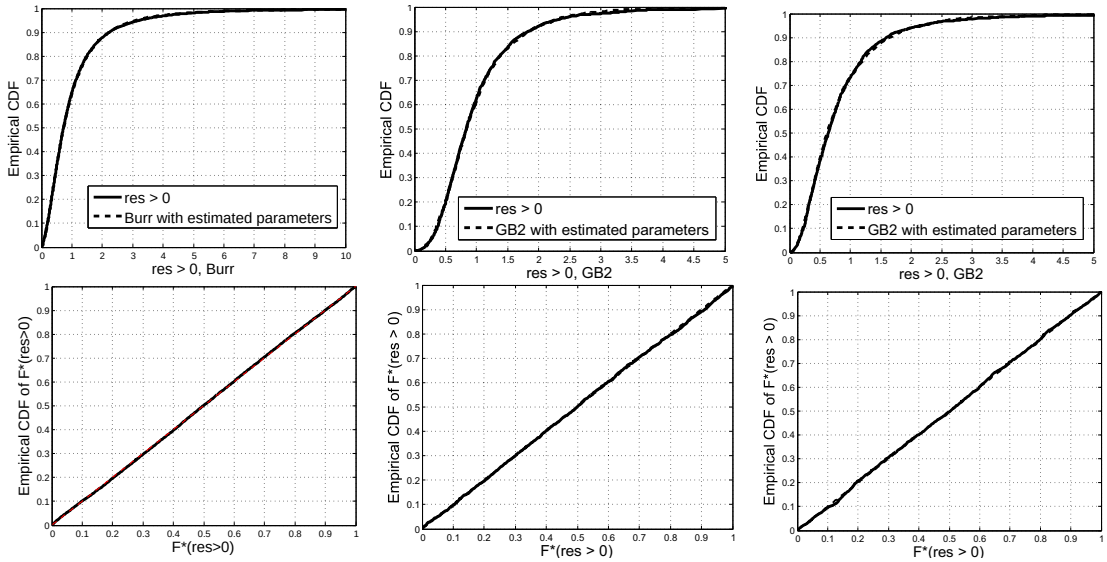

Figure 5. Fitting Burr to IBM30s over 28 February - 31 March 2000 (left) and GB2 to the trade volume of EURUSD (middle) and USDJPY (right) in Window 1. The empirical c.d.f. of positive $\widehat{\varepsilon}_{t,\tau}$ plotted against the theoretical c.d.f. of Burr$(\widehat{\nu},\widehat{\zeta})$ or GB2$(\widehat{\nu},\widehat{\xi},\widehat{\zeta})$ (top). The empirical c.d.f. of the PIT values of positive $\widehat{\varepsilon}_{t,\tau}$ when $F^*(\cdot;\widehat{\boldsymbol{\theta}}^*)$ is Burr$(\widehat{\nu},\widehat{\zeta})$ or GB2$(\widehat{\nu},\widehat{\xi},\widehat{\zeta})$ (bottom). The spline-DCS with the static daily spline was used.

along the diagonal, indicating that the PIT values are close to being standard uniformly distributed (denoted by $U[0,1]$). The results are remarkably similar for IBM1m and other sampling windows for the FX data. The Kolmogorov-Smirnov statistics in Table 5 are outside the 5% rejection region for most cases under the null that the distribution is the estimated Burr or GB2. The model is robust to the choice of sampling frequency and sampling period. The computing time for the ML estimation to converge in all of these cases were generally short (about 5 minutes).

The gamma distribution, which is a more popular non-negative distribution in finance, is a special case of the generalized gamma disribution (GG), and GG is a limiting case of GB2 for when ζ is large. Gamma and GG did not fit the empirical distribution of the data well compared to GB2. This is consistent with the estimated ζ (reported in Table 7), which is far from being large. Since $\nu\zeta$ is the (upper) tail-index of GB2, its estimate, which is far from being large, suggest that our data is heavy-tailed (see Table 7). Since a gamma distribution cannot be heavy-tailed, GB2 may be preferred for heavy-tailed data. Burr and GB2 are flexible since they have two to three shape parameters ((ν,ξ,ζ) for GB2 with $\xi = 1$ for Burr). Also see Harvey (2013, p.12, p.189). In Table 7, we have $2 < \widehat{\nu}\widehat{\zeta} < 3$ for IBM1m and IBM30s, implying that only the first and second moments exist and that the theoretical skewness does not exist under the assumption that F^* is Burr. For the FX data, $3 < \widehat{\nu}\widehat{\zeta} < 4$, so that the moments up to the third exist. (See Appendix A.1 of the supplementary material for the existence of moments.)

Figure 6 shows the estimated daily spline, $\widehat{s}_{t,\tau}$, for IBM30s. The spline beautifully

IBM30s	IBM1m
1.01	1.27

Window	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
EURUSD	1.22	1.23	0.96	1.22	1.32	0.99	1.14	0.95	0.82	1.00	0.83	1.329	0.74	0.85	0.86	0.68
USDJPY	1.53*	1.31	1.326	1.40*	0.84	0.63	1.41*	1.02	0.98	1.28	1.41*	1.02	1.16	1.22	0.82	1.23

Table 5. Kolmogorov-Smirnov statistics computed under the null that positive $\widehat{\varepsilon}_{t,\tau}$ comes from $F^*(\cdot;\widehat{\boldsymbol{\theta}}^*)$ being either Burr$(\widehat{\nu},\widehat{\zeta})$ for the equity data or GB2$(\widehat{\nu},\widehat{\xi},\widehat{\zeta})$ for the FX data. * 5% significance. ** 1% significance.

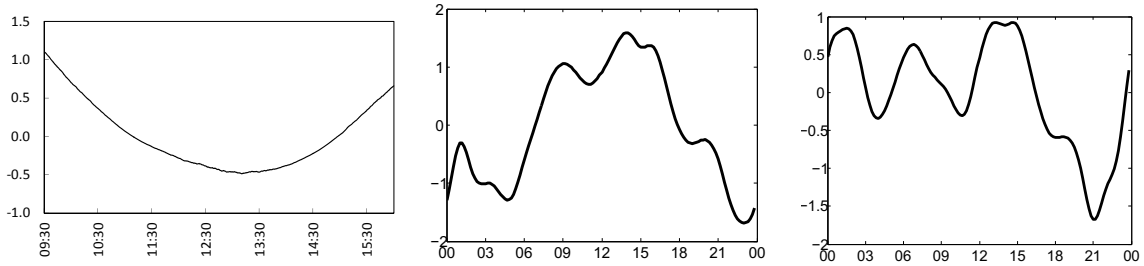

Figure 6. $\widehat{s}_{t,\tau}$ for IBM30s from market open to close (left). $\widehat{s}_{t,\tau}$ for the EURUSD volume (middle, the in-sample window 2) and the USDJPY volume (right, the in-sample window 13). The spline-DCS with the static daily spline. Intra-day hours in the NY local time for IBM30s and in GMT for FX along the x-axes.

captures the intra-day U-shaped pattern in the equity data. The height of the spline at 4pm is different from that at 9.30am, reflecting the overnight effect. Figure 6 also shows the estimated spline for the FX data. The spline successfully captures the bimodal and trimodal patterns we saw in the data in Section 2. $\widehat{s}_{t,\tau}$ is smooth in all of these cases.

The spline-DCS captures the dynamics of the data well. Figure 7 shows that the estimation residuals, $\widehat{\varepsilon}_{t,\tau}$, and the score, $\widehat{u}_{t,\tau}$, exhibit little to no signs of serial correlation. For the FX data, the results are similar for other sampling windows. However, the large sample size makes the Ljung-Box statistics sensitive to small departures from zero autocorrelation. This can be seen in the statistics reported in Table 6, which pick up statistically significant autocorrelation.[14]

In Table 7, we have $\widehat{\kappa}_\eta^{(2)} > \widehat{\kappa}_\eta^{(1)} > \widehat{\kappa}_\mu > 0$, which means that $\eta_{t,\tau}^{(2)}$ is more sensitive to changes in $u_{t,\tau-1}$ than $\eta_{t,\tau}^{(1)}$, which is, in turn, more sensitive than $\mu_{t,\tau}$. The stationarity conditions for $\eta_{t,\tau}^{(1)}$ and $\eta_{t,\tau}^{(2)}$ outlined in Section 3.3 are satisfied by $\widehat{\phi}_1^{(1)}$, $\widehat{\phi}_2^{(1)}$, and $\widehat{\phi}_1^{(2)}$. We have $|\widehat{\phi}_e| < 1$ so that $\widehat{e}_{t,\tau}$ reverts to zero after an event. The estimates of the probability mass at zero are consistent with the sample statistics in Tables 2 and 3 (e.g. $\widehat{p} = 0.0047$ for IBM30s). For the equity data, the estimated asymmetry term, $\widehat{\kappa}_{\eta,a}^{(2)}$, in $\eta_{t,\tau}^{(2)}$ is negative and statistically significant, reflecting the tendency of volume to decrease when price falls. Brownlees et al. (2011) found that the sign of their asymmetry term was positive

[14]Note that the sample autocorrelation of $\widehat{u}_{t,\tau}$ may exhibit stronger serial correlation than that of $\widehat{\varepsilon}_{t,\tau}$, because the score weighs down (and thus it is robust to) the effect of large-sized observations.

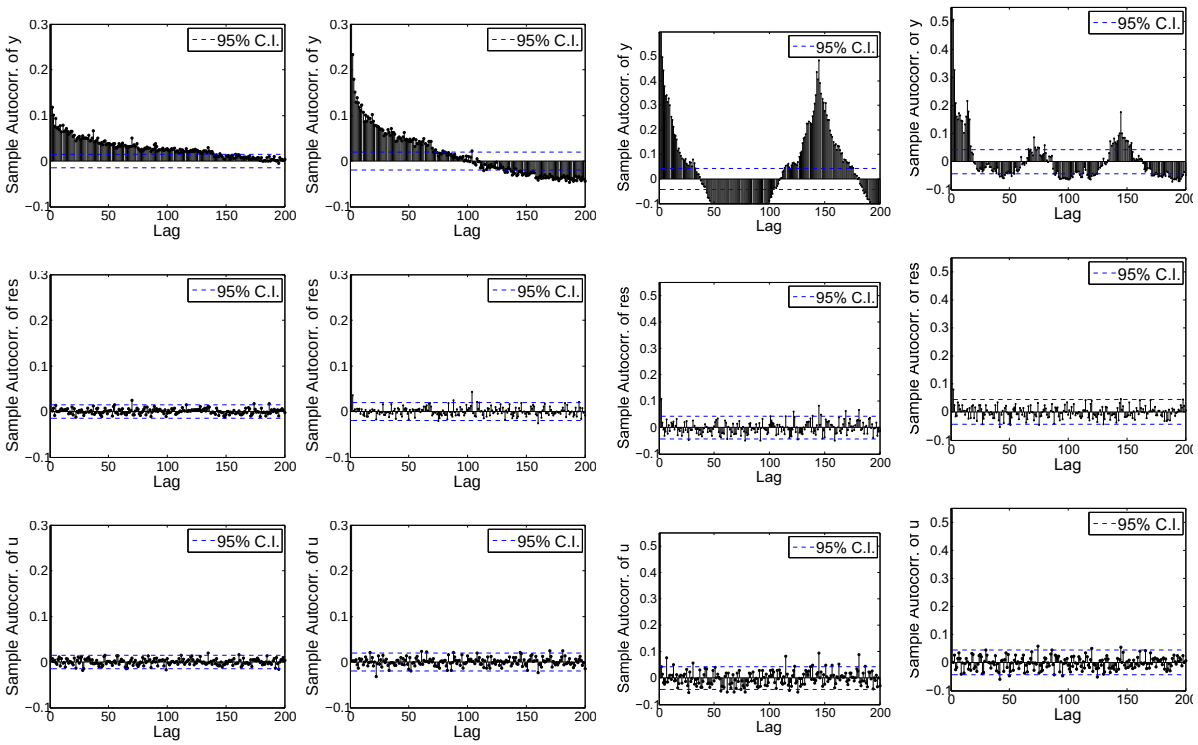

Figure 7. The sample autocorrelation of trade volume (top), $\widehat{\varepsilon}_{t,\tau}$ (middle), and $\widehat{u}_{t,\tau}$ (bottom). IBM30s (left), IBM1m (left center), EURUSD (Window 1, right center), and USDJPY (Window 6, right). The spline-DCS with the static daily spline. The 95% confidence interval is computed at ± 2 standard errors.

so that volume on average increases when price falls. $\widehat{\kappa}^{(1)}_{\eta,a}$ was found to be statistically insignificant for both IBM1m and IBM30s, suggesting that there is no asymmetry effect in the lower frequency component, $\eta^{(1)}_{t,\tau}$.

5 Out-of-sample performance

5.1 One-step ahead forecasts: model stability

We use the predictive c.d.f. based on one-step ahead forecasts to assess the stability of the estimated model and the quality of the forecasts. We introduce the following set notations:

$$\Psi_H = \{(t,\tau) \in \{T+1,\ldots,T+H\} \times \{1,\ldots,I\}\}, \quad \Psi_{H,>0} = \{\Psi_H : y_{t,\tau} > 0\}.$$

We recursively update $\lambda_{t,\tau}$ given a new out-of-sample observation, $y_{t,\tau}$, without re-estimating the model to obtain one-step ahead forecasts, $(\widetilde{\lambda}_{t,\tau})_{(t,\tau)\in\Psi_H}$. Then, we compute $F^*(\widetilde{\varepsilon}_{t,\tau}; \widehat{\boldsymbol{\theta}}^*)$, where $\widetilde{\varepsilon}_{t,\tau} = y_{t,\tau}/\exp(\widetilde{\lambda}_{t,\tau})$ for $(t,\tau) \in \Psi_{H,>0}$. This should be standard uniformly distributed.

Figure 8 shows the empirical c.d.f. of $F^*(\widetilde{\varepsilon}_{t,\tau}; \widehat{\boldsymbol{\theta}}^*)$ for IBM30s and EURUSD in the out-of-sample window 1. The forecast horizons go up to $H = 20$ days ahead for the equity data, and up to $H = 12$ days ahead for the FX data. The length of the out-of-

	$\widehat{\varepsilon}_{t,\tau}$						$\widehat{\varepsilon}^2_{t,\tau}$						$\widehat{u}_{t,\tau}$					
Window	$\widehat{\rho}_1$	$\widehat{\rho}_{day}$	Q_1	Q_{day}	p-val$_1$	p-val$_{day}$	$\widehat{\rho}_1$	$\widehat{\rho}_{day}$	Q_1	Q_{day}	p-val$_1$	p-val$_{day}$	$\widehat{\rho}_1$	$\widehat{\rho}_{day}$	Q_1	Q_{day}	p-val$_1$	p-val$_{day}$
1	0.120	-0.033	31.048	169.116	0.000	0.000	0.080	-0.010	13.961	53.550	0.000	1.000	0.029	-0.031	1.855	155.616	0.173	0.002
2	0.147	-0.018	46.998	188.952	0.000	0.000	0.125	-0.006	33.870	43.940	0.000	1.000	0.007	-0.003	0.101	298.334	0.750	0.000
3	0.187	-0.002	74.982	277.065	0.000	0.000	0.094	-0.006	18.854	68.910	0.000	0.999	0.044	-0.002	4.139	233.940	0.042	0.000
4	0.079	-0.009	13.309	142.910	0.000	0.016	0.069	0.000	10.271	92.462	0.001	0.872	0.004	-0.016	0.037	196.560	0.847	0.000
5	0.069	-0.023	10.246	150.444	0.001	0.005	0.037	-0.009	2.862	84.197	0.091	0.963	0.024	-0.009	1.252	178.331	0.263	0.000
6	0.080	0.009	13.494	157.383	0.000	0.002	0.177	0.008	65.872	191.092	0.000	0.000	0.002	0.001	0.006	183.804	0.936	0.000
7	0.050	-0.001	5.387	130.444	0.020	0.089	0.007	0.007	0.095	16.065	0.758	1.000	0.008	-0.035	0.144	137.420	0.705	0.039
8	0.099	-0.030	20.931	163.575	0.000	0.000	0.156	-0.014	51.776	192.510	0.000	0.000	0.001	-0.022	0.003	133.777	0.957	0.047
9	0.223	-0.019	104.555	268.999	0.000	0.000	0.388	0.003	317.345	361.753	0.000	0.000	0.017	-0.035	0.625	151.152	0.429	0.004
10	0.092	0.030	18.075	174.686	0.000	0.000	0.050	0.016	5.386	201.496	0.020	0.000	0.044	0.024	4.172	150.216	0.041	0.005
11	0.131	-0.017	36.784	182.354	0.000	0.000	0.047	-0.002	4.613	16.008	0.032	1.000	0.012	-0.036	0.326	166.686	0.568	0.000
12	0.084	-0.006	14.891	161.028	0.000	0.001	0.038	-0.002	3.075	98.504	0.079	0.755	0.036	0.000	2.780	188.838	0.095	0.000
13	0.088	-0.024	16.557	171.534	0.000	0.000	0.017	-0.006	0.613	206.254	0.434	0.000	0.018	-0.036	0.732	141.613	0.392	0.017
14	0.128	-0.033	34.979	201.509	0.000	0.000	0.045	-0.004	4.436	29.458	0.035	1.000	0.008	-0.008	0.139	169.272	0.709	0.000
15	0.188	-0.022	75.888	195.670	0.000	0.000	0.090	-0.002	17.539	23.249	0.000	1.000	0.008	-0.022	0.127	170.162	0.721	0.000
16	0.082	-0.036	12.476	129.013	0.000	0.082	0.060	-0.011	6.660	25.878	0.010	1.000	0.025	-0.033	1.160	265.047	0.281	0.000

Table 6. Residual analysis for the spline-DCS with the static daily spline fitted to the trade volume of EURUSD. Q_l is the Ljung-Box statistic to test the null of no autocorrelation up to the l-th lag.

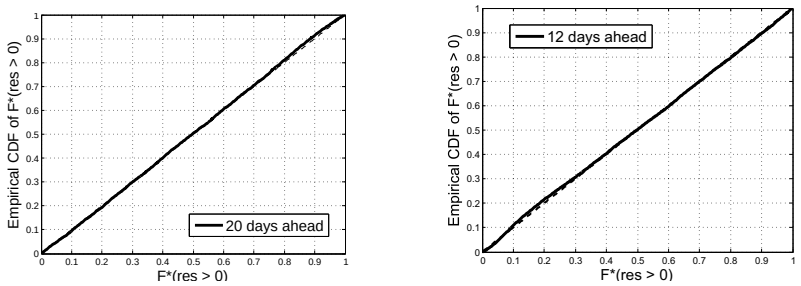

Figure 8. The empirical c.d.f. of $F^*(\widetilde{\varepsilon}_{t,\tau}; \widehat{\boldsymbol{\theta}}^*)$ from one-step ahead forecasts. Left: IBM30s up to 20 out-of-sample days ahead (between 3 - 23 April 2000). Right: EURUSD trade volume up to 12 out-of-sample days ahead of window 1. Computed using the theoretical c.d.f. of Burr$(\widehat{\nu}, \widehat{\zeta})$ for IBM30s and GB2$(\widehat{\nu}, \widehat{\xi}, \widehat{\zeta})$ for EURUSD. The spline-DCS with the static daily spline is used.

sample windows for the FX data is two weeks as tabulated in Table 1. The results are similar for IBM1m, and other out-of-sample windows of EURUSD and USDJPY. The distribution of the PIT values is roughly $U[0, 1]$ for this extended out-of-sample period. The estimated model appears to capture the out-of-sample empirical distribution well, although the non-negligible deterioration in the quality of the fit for some of the cases are reflected in the Kolmogorov-Smirnov statistics in Table 8.

Figure 9 shows the sample autocorrelation of one-step ahead $\widetilde{\varepsilon}_{t,\tau}$ and $\widetilde{u}_{t,\tau}$ for EURUSD in the out-of-sample window 1. The results are similar for the equity data, as well as other out-of-sample period of EURUSD and USDJPY. The one-step ahead forecasts appear to capture the out-of-sample dynamics well over an extended out-of-sample period.

5.2 Model comparison

We use the FX data to compare the out-of-sample predictive performance of the spline-DCS and the c-MEM defined in Appendix F of the supplementary material. For the FX data, we do not include asymmetry terms or overnight dummies of the type used by Brownlees et al. (2011), but we include an event component to be comparable with the spline-DCS. Brownlees et al. (2011) estimate the c-MEM by the generalized method of

Variable	IBM30s	IBM1m	USDJPY	EURUSD
Window	—	—	2	3
ν	1.632 (0.016)	2.229 (0.033)	1.738 (0.018)	2.002 (0.019)
ξ	—	—	1.738 (0.069)	1.369 (0.048)
ζ	1.484 (0.045)	1.143 (0.044)	2.062 (0.085)	1.501 (0.059)
ω	9.172 (0.199)	9.774 (0.178)	5.521 (0.095)	4.232 (0.068)
κ_μ	0.006 (0.001)	0.007 (0.002)	0.010 (0.003)	0.006 (0.003)
$\phi_1^{(1)}$	0.554 (0.134)	0.391 (0.091)	0.570 (0.017)	0.582 (0.016)
$\phi_2^{(1)}$	0.414 (0.133)	0.555 (0.093)	0.373 (0.017)	0.363 (0.016)
$\kappa_\eta^{(1)}$	0.049 (0.007)	0.047 (0.008)	0.059 (0.007)	0.070 (0.007)
$\phi_1^{(2)}$	0.690 (0.041)	0.610 (0.057)	0.438 (0.083)	0.369 (0.114)
$\kappa_\eta^{(2)}$	0.092 (0.008)	0.067 (0.008)	0.094 (0.009)	0.080 (0.010)
$\kappa_{\eta,a}^{(2)}$	-0.004 (0.004)	-0.008 (0.003)	—	—
p	0.0047 (0.0005)	0.0006 (0.0003)	0.000 (0.000)	0.006 (0.002)
γ_0	1.197 (0.066)	1.119 (0.064)	(omitted)	(omitted)
γ_1	0.061 (0.041)	0.066 (0.041)	(omitted)	(omitted)
γ_2	-0.419 (0.036)	-0.392 (0.036)	(omitted)	(omitted)
γ_3	-0.216 (0.037)	-0.244 (0.037)	(omitted)	(omitted)
ϕ_e	—	—	0.807 (0.047)	0.504 (0.108)
$\kappa_{e,1}$	—	—	2.739 (1.024)	0.998 (0.021)
$\kappa_{e,2}$	—	—	2.457 (1.094)	0.645 (0.097)
$\kappa_{e,3}$	—	—	1.424 (0.021)	1.707 (1.208)

Table 7. The estimated parameter values for the spline-DCS with the static daily spline. The analytical standard errors here uses the outer-product of the first derivative of the joint log-likelihood (see Appendix D of the supplementary material). The coefficients of $s_{t,\tau}$ are omitted here for the FX data to save space. Burr (GB2 with $\xi = 1$) is estimated for the IBM data due to the considerations given in Footnote 13. For the FX data, results for other sampling windows are given in Appendix E of the supplementary material.

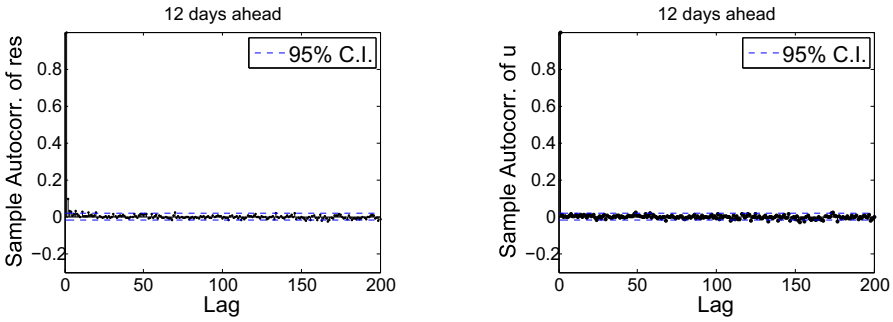

Figure 9. The sample autocorrelation of one-step ahead $\widetilde{\varepsilon}_{t,\tau}$ (left) and $\widetilde{u}_{t,\tau}$ (right) of EURUSD for up to 12 out-of-sample days ahead of window 1. The spline-DCS with the daily spline is used.

moment (GMM) in order to allow for a greater flexibility in the distribution of the error term. We adopt their GMM estimation strategy. The in-sample estimation method and results are given in Appendix G of the supplementary material.

Since the proposed GMM in the c-MEM can be used to forecast the level of volume (hereafter, level forecasts), but not the density, the performance of the competing models are compared using their level forecasts. We produce forecasts over each out-of-sample window listed in Table 1. The loss functions we consider are the daily mean absolute errors (MAE), the daily root mean squared errors (RMSE), and the daily slicing loss

IBM30s	IBM1m
3.05**	1.18

Window	1	2	3	4	5	6	7	8	9	10	11	12
EURUSD	1.35*	2.03**	1.51*	1.13	0.67	1.45*	1.57*	1.38*	0.94	1.12	1.51*	4.65**
USDJPY	3.39**	1.21	1.42*	1.332	1.34*	1.35*	2.00**	1.86**	0.96	1.60*	2.01**	7.10**

Window	13	14	15	16
EURUSD	0.96	1.17	1.51*	10.02**
USDJPY	2.09**	1.95**	0.68	1.56*

Table 8. Kolmogorov-Smirnov statistics computed under the null that positive $\widetilde{\varepsilon}_{t,\tau}$ comes from $F^*(\cdot;\widehat{\boldsymbol{\theta}}^*)$ being either Burr$(\widehat{\nu},\widehat{\zeta})$ for the equity data or GB2$(\widehat{\nu},\widehat{\xi},\widehat{\zeta})$ for the FX data. Windows are out-of-sample. * 5% significance. ** 1% significance.

given by

$$L^{MAE}\left((y_{T+h,\tau},\widetilde{y}_{T+h,\tau})_{\tau=1}^{I}\right) = (I)^{-1}\sum_{\tau=1}^{I}\left|y_{T+h,\tau}-\widetilde{y}_{T+h,\tau}\right|,$$

$$L^{RMSE}\left((y_{T+h,\tau},\widetilde{y}_{T+h,\tau})_{\tau=1}^{I}\right) = \sqrt{(I)^{-1}\sum_{\tau=1}^{I}\left(y_{T+h,\tau}-\widetilde{y}_{T+h,\tau}\right)^2},$$

$$L^{slicing}\left((y_{T+h,\tau},\widetilde{y}_{T+h,\tau})_{\tau=1}^{I}\right) = -\sum_{\tau=1}^{I} w_{T+h,\tau}\log\widehat{w}_{T+h,\tau},$$

where $w_{T+h,\tau}$ is intra-day volume proportion, for $h=1,\ldots,H$ and $\tau=1,\ldots,I$.

Intra-day volume is highly volatile and the value of the daily loss functions are typically dominated by large errors. Thus, we consider not only RMSE, but also MAE, since RMSE is sensitive to extreme observation points. These metrics can be used to assess the quality of forecast volume dynamics. The conditional first moment is theoretically optimal in the sense of minimizing RMSE. The conditional median is theoretically the optimal predictor if the loss function is MAE. The slicing loss is developed by Brownlees et al. (2011) to evaluate VWAP trading strategies. It is a common term determining the ranking of models by the negative multinomial log-likelihood loss and the Kullback-Leibler loss. The forecast slicing weights, $\widehat{w}_{T+h,\tau}$, are computed under the Dynamic VWAP replication strategy outlined by Brownlees et al. (2011).[15]

The daily MAE and RMSE are computed using one-step ahead volume forecasts. The Dynamic VWAP updates Static-VWAP, which is one-day ahead intra-daily volume forecasts, throughout the day using new intra-day data. These loss functions are computed for each $h=1,\ldots,H$ of each out-of-sample window.

With the spline-DCS, formulae for one-step ahead volume forecasts are (conditional)

[15]We do not consider the VWAP-tracking MSE discussed in Bialkowski et al. (2008) as we did not have the price data. Brownlees et al. (2011) prefer the slicing loss, which is less noisy.

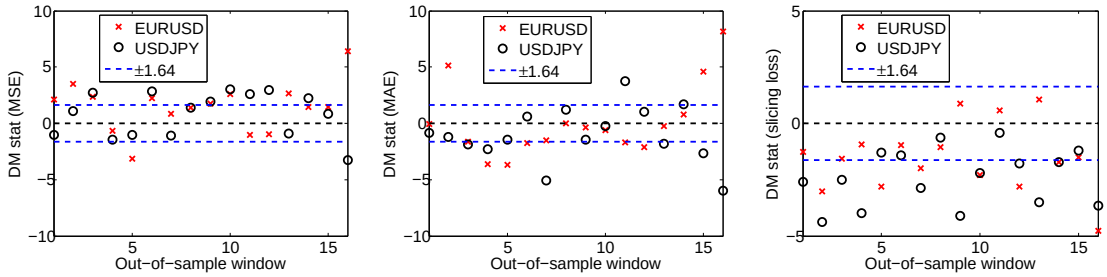

Figure 10. Diebold-Mariano statistics from Table 9.

mean forecasts,

$$\widetilde{y}_{T+h,\tau} \equiv \mathbb{E}[y_{T+h,\tau}|\mathcal{F}_{T+h,\tau-1}] = \exp(\widetilde{\lambda}_{T+h,\tau}) \int_0^\infty x f(x;\widehat{\boldsymbol{\theta}})dx, \tag{3}$$

and (conditional) median forecasts,

$$\widetilde{y}_{T+h,\tau} \equiv \exp(\widetilde{\lambda}_{T+h,\tau}) Q_{0.5}(\widehat{\boldsymbol{\theta}}), \text{ where } \mathbb{P}\left(\varepsilon_{T+h,\tau} \leq Q_{0.5}(\widehat{\boldsymbol{\theta}})|\mathcal{F}_{T+h,\tau-1}\right) = 0.5, \tag{4}$$

for $h = 1, \ldots, H$ and $\tau = 1, \ldots, I$. Since the error distribution is fully given by GB2, the median and the first moment of $F(\varepsilon;\widehat{\boldsymbol{\theta}})$ can be computed analytically. The mean forecasts are used for the MSE and the median forecasts for the MAE. With the c-MEM, we produce mean forecasts only given the minimal GMM assumption for the error distribution.

The proportion forecasts for the Dynamic VWAP utilizes the multi-step predictor of volume given by $\widetilde{y}_{t,\tau} = \mathbb{E}[\varepsilon_{t,\tau}]\,\mathbb{E}[\exp(\lambda_{t,\tau})|\mathcal{F}_{T,I}]$ for $(t,\tau) \in \Psi_H$, but continuously updates this quantity with new intra-day trade data. See Brownlees et al. (2011). Note that this moment quantity is evaluated analytically for the spline-DCS since the moment generating function (m.g.f.) of $u_{t,\tau}$ exists.

Table 9 and Figure 10 show Diebold-Mariano statistics computed from the resulting loss function values. The results based on RMSE are marginally in favor of the c-MEM, as test statistics are in a statistically significantly positive region more frequently. The results based on MAE are marginally in favor of the spline-DCS. These results reflect the fact that the spline-DCS is more robust to large-sized observations than the c-MEM, and that RMSE penalizes models for occasionally throwing up large errors more severely than MAE. In the interest of VWAP replication strategies, the spline-DCS outperforms the c-MEM in minimizing the slicing loss function.

5.3 Discussions

We found MLE in the spline-DCS fast and easy to compute. With the FX data and the same convergence tolerance, the optimization procedure for the spline-DCS converged in about 5 minutes. This is a very attractive feature in high-frequency finance since a typical forecasting exercise deals with a very large sample size, which generally makes estimation computationally intensive.

MSE:

Window	1	2	3	4	5	6	7	8	9	10
EURUSD	2.10**	3.50***	2.35**	**-0.67**	**-3.12*** **	2.22**	0.87	1.39	1.76*	2.61***
USDJPY	**-1.00**	1.07	2.73***	**-1.46**	**-1.00**	2.83***	**-1.10**	1.36	1.90*	3.02***

Window	11	12	13	14	15	16
EURUSD	**-1.05**	**-0.97**	2.65***	1.44	1.47	6.41***
USDJPY	2.57**	2.99***	**-0.94**	2.21**	0.85	**-3.27*** **

MAE:

Window	1	2	3	4	5	6	7	8	9	10
EURUSD	**-0.06**	5.13***	**-1.63**	**-3.61*** **	**-3.72*** **	**-1.75***	**-1.54**	**-0.01**	**-0.34**	**-0.62**
USDJPY	**-0.87**	**-1.20**	**-1.88***	**-2.27** **	**-1.42**	0.61	**-5.07*** **	1.21	**-1.45**	**-0.27**

Window	11	12	13	14	15	16
EURUSD	**-1.72***	**-2.12** **	**-0.26**	0.81	4.60***	8.16***
USDJPY	3.77***	1.00	**-1.83***	1.70*	**-2.69*** **	**-6.00*** **

Slicing loss:

Window	1	2	3	4	5	6	7	8	9	10
EURUSD	**-1.26**	**-3.03*** **	**-1.58**	**-0.95**	**-2.81*** **	**-0.96**	**-1.99** **	**-1.05**	0.88	**-2.29** **
USDJPY	**-2.59*** **	**-4.38*** **	**-2.51** **	**-3.98*** **	**-1.30**	**-1.42**	**-2.88*** **	**-0.63**	**-4.11*** **	**-2.19** **

Window	11	12	13	14	15	16
EURUSD	0.57	**-2.81*** **	1.05	**-1.73**	**-1.50***	**-4.78*** **
USDJPY	**-0.43**	**-1.80***	**-3.49*** **	**-1.71**	**-1.20**	**-3.67*** **

Table 9. Diebold-Mariano statistics to test the null of equal predictive ability against the alternative of different predictive ability, which is a two-sided test with * 10% significance, ** 5% significance, and *** 1% significance. Negative values are in bold font. If the alternative is of a one-sided test, a statistically significant negative (positive) value is in favor of the spline-DCS (c-MEM).

When we were specifying the spline component, increasing the number of knots did not necessarily improve the quality of the fit of the model to the data. We found two rules of thumb that worked well in determining the location and the number of knots to improve the quality of approximation. The first is to place one knot approximately every 1 hour to 1.5 hours. The second is to place relatively more knots when trade intensity changes fast. Such hours correspond to the first and last trading hour of the NYSE for the equity data, and the hours before and after the volume peaks for the FX data. It is useful to sketch how a piecewise function of cubic polynomials can fit the data.

A main objection to ML is that it requires the error distribution to be fully defined. A non-parametric approach may be preferred if no parametric distribution can reasonably describe the empirical distribution of the data. This does not seem to be the case in our application. With a suitable distribution, our model gives more insight into the overall shape of the distribution, the degree of dispersion, and the size of the tail. Quantile forecasts or moment forecasts of different orders can be also produced from density forecasts.

All of these features are useful for volume prediction and risk analysis.

6 Concluding remarks

This paper developed the spline-DCS model for forecasting the dynamics of high-frequency trade volume with intra-day periodic patterns. We showed that it captures salient features of the high-frequency data. Our estimation results are robust to the choice of sampling frequency and sampling period. The out-of-sample analysis showed that the estimation results are stable, and that our model outperforms the c-MEM in minimizing the slicing loss function. The ease of computation is an important advantage of the spline-DCS. Burr and GB2 achieved a very good fit to the empirical distribution of the data. The estimated parameter values indicated that our data is heavy tailed.

The pattern of periodicity was assumed to be the same every day in this paper. Ito (2013) challenges this standard assumption by introducing the spline-DCS with a dynamic spline, and show that the model can improve on the version with the static spline in minimizing the slicing loss function.

The object of our empirical analysis is trade volume, and, as such, this study also contributes to the literature dedicated to the analysis of market activity and intensity. The spline-DCS can be applied to model other variables such as asset returns using a suitable distribution such as Student's t. We studied the movements of volume in complete isolation from price, which is ultimately not satisfactory as return volatility and volume dynamics must interact. A natural extension is to construct multivariate intra-day DCS that jointly models return volatility and volume.

References

Andersen, T. G. and Bollerslev, T. (1998), "Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Annoucements, and Longer Run Dependencies," *Journal of Finance*, 53, 219–265.

Avdulaj, K. and Barunik, J. (2015), "Are Benefits From Oil-Stocks Diversification Gone? New Evidence From a Dynamic Copula and High Frequency Data," *Energy Economics*, 51, 31–44.

Bialkowski, J., Darolles, S., and Le Fol, G. (2008), "Improving VWAP Strategies: A Dynamic Volume Approach," *Journal of Banking and Finance*, 32, 1709–1722.

Bowsher, C. G. and Meeks, R. (2008), "The Dynamics of Economic Functions: Modelling and Forecasting the Yield Curve," *Journal of the American Statistical Association*, 103, 1419–1437.

Brownlees, C. T., Cipollini, F., and Gallo, G. M. (2011), "Intra-Daily Volume Modelling and Prediction for Algorithmic Trading," *Journal of Financial Econometrics*, 9, 489–518.

Caviano, M. and Harvey, A. C. (2014), "Time-Series Models With an EGB2 Conditional Distribution," *Journal of Time Series Analysis*, 35, 558–571.

Creal, D. D., Koopman, S. J., and Lucas, A. (2011), "A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations," *Journal of Business and Economic Statistics*, 29, 552–563.

— (2013), "Generalized Autoregressive Score Models with Applications," *Journal of Applied Econometrics*, 28, 777–795.

Creal, D. D., Schwaab, B., Koopman, S. J., and Lucas, A. (2014), "Observation-Driven Mixed-Measurement Dynamic Factor Models With An Application To Credit Risk," *The Review of Economics and Statistics*, 96, 898–915.

Gao, C.-T. and Zhou, X.-H. (2016), "Forecasting VaR and ES Using Dynamic Conditional Score Models and Skew Student Distribution," *Economic Modelling*, 53, 216–223.

Harvey, A. C. (1993), *Time Series Models*, Harvester: Wheatsheaf, 2nd ed.

— (2013), *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*, Econometric Society Monograph, Cambridge University Press.

Harvey, A. C. and Koopman, S. J. (1993), "Forecasting Hourly Electricity Demand Using Time-Varying Splines," *Journal of the American Statistical Association*, 88, 1228–1236.

Harvey, A. C., Koopman, S. J., and Riani, M. (1997), "The Modeling and Seasonal Adjustment of Weekly Observations," *Journal of Business & Economic Statistics*, 15, 354–68.

Harvey, A. C. and Lange, R.-J. (2015), "Volatility Modeling with a Generalized t-Distribution," Cambridge Working Papers in Economics CWPE1517, University of Cambridge.

Harvey, A. C. and Luati, A. (2014), "Filtering With Heavy Tails," *Journal of the American Statistical Association*, 109, 1112–1122.

Harvey, A. C. and Sucarrat, G. (2014), "EGARCH Models With Fat Tails, Skewness and Leverage," *Computational Statistics and Data Analysis*, 76, 320–338.

Hautsch, N. (2012), *Econometrics of Financial High-Frequency Data*, Springer: Berlin.

Hautsch, N., Malec, P., and Schienle, M. (2014), "Capturing the Zero: A New Class of Zero-Augmented Distributions and Multiplicative Error Processes," *Journal of Financial Econometrics*, 12, 89–121.

Hendricks, W., Koenker, R., and Poirier, D. (1979), "Residential Demand for Electricity," *Journal of Econometrics*, 9, 33–57.

Hyndman, R. J., King, M. L., Pitrun, I., and Billah, B. (2005), "Local Linear Forecasting Using Cubic Smoothing Splines," *Australian and New Zealand Journal of Statistics*, 47, 87–99.

Ito, R. (2013), "Modeling Dynamic Diurnal Patterns in High Frequency Financial Data," Cambridge Working Papers in Economics CWPE1315, University of Cambridge.

Janus, P., Koopman, S. J., and Lucas, A. (2014), "Long Memory Dynamics For Multivariate Dependence Under Heavy Tails," *Journal of Empirical Finance*, 29, 187–206.

Jungbacker, B., Koopman, S. J., and van der Wel, M. (2014), "Smooth Dynamic Factor Analysis With Application to the US Term Structure of Interest Rates," *Journal of Applied Econometrics*, 29, 65–90.

Kleiber, C. and Kotz, S. (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley: New York.

Lo, A. W. and Wang, J. (2010), "Stock Market Trading Volume," in *Handbook of Financial Econometrics*, eds. Aït-Sahalia, Y. and Hansen, L., North-Holland: New York, vol. 2.

Lucas, A., Schwaab, B., and Zhang, X. (2014), "Conditional Euro Area Sovereign Default Risk," *Journal of Business and Economic Statistics*, 32, 271–284.

Lucas, A. and Zhang, X. (2016), "Score-Driven Exponentially Weighted Moving Averages and Value-At-Risk Forecasting," *International Journal of Forecasting*, 32, 293–302.

McCulloch, R. E. and Tsay, R. S. (2001), "Nonlinearity in High-Frequency Financial Data and Hierarchical Models," *Studies in Nonlinear Dynamics and Econometrics*, 5, 1–17.

Poirier, D. (1973), "Piecewise Regression Using Cubic Spline," *Journal of the American Statistical Association*, 68, 515–524.

Rydberg, T. H. and Shephard, N. (2003), "Dynamics of Trade-By-Trade Price Movements: Decomposition and Models," *Journal of Financial Econometrics*, 1, 2–25.

Salvatierra, I. D. L. and Patton, A. J. (2015), "Dynamic Copula Models and High Frequency Data," *Journal of Empirical Finance*, 30, 120–135.