Modeling dynamic diurnal patterns in high frequency financial data

Ryoko Ito

Faculty of Economics, Cambridge University
Email: ri239@cam.ac.uk
Website: www.itoryoko.com

This paper: Cambridge Working Papers in Economics CWPE1315

OxMetrics User Conference, Sept 2014
We want to model **daily periodic patterns** (diurnal patterns) in high-frequency financial data.

Popular ways to capture periodicity:

- Fourier flexible form approximation
- Compute (re-scaled) sample moments for each intra-day bins

**The pattern of periodicity is fixed over time.** (e.g. Andersen and Bollerslev (1998), Engle and Russell (1998), Shang et al. (2001), Campbell and Diebold (2005), Engle and Rangel (2008), Brownlees et al. (2011), Engle and Sokalska (2012).)

**Contribution:** **dynamic cubic spline** to model periodicity (c.f. Harvey and Koopman (1993)). Advantages:

- Parsimonious. One-step estimation with all other coefficients.
- Dynamic periodicity.
- Fits the empirical distribution of our data well (including the upper extreme quantiles).
Also need to capture other stylized features

- Concentration of zero-observations
- Non-normality, heavy tail
- Highly persistent dynamics (long-memory?)

Methods:

- Distribution decomposition at zero
- **Dynamic Conditional Score** (Harvey (2013)) to capture non-normality, heavy-tail
- Unobserved components
Data (short sampling period): empirical features

- **Trade volume** of IBM stock traded on the NYSE. The number of shares traded.
- Period: 5 consecutive trading weeks in February - March 2000
- Aggregation interval: 30 seconds (15 seconds – 1 minute also in the paper)

**Figure**: IBM30s (left column) and the same series smoothed by the simple moving average (right column). Time on the x-axis. Monday 20 - Friday 24 March 2000. Each day covers trading hours between 9.30am-4pm (in the New York local time).
Empirical features (short sampling period)

Diurnal U-shaped patterns.
Trade volume “bottoms out” at around 1pm.

Figure: IBM30s (left column) and the same series smoothed by the simple moving average (right column). Time on the x-axis. Wednesday 22 March 2000, covering 9.30am-4pm (in the New York local time).
Empirical features

Sample autocorrelation. Highly persistent.

Heavy, long upper-tail.

Figure: Sample autocorrelation of IBM30s. Sampling period: 28 February - 31 March 2000. The 200th lag corresponds approximately to 1.5 hours prior.

Figure: Frequency distribution (top) and empirical cdf (bottom) of IBM30s. Sample: 28 February - 31
Data (long sampling period): empirical features

- **Trade volume** of IBM stock traded on the NYSE. The number of shares traded.
- In-sample period: January 2007 - December 2010 (4 years)
- Aggregation interval: 10 minutes

**Figure:** Left: IBM10m between Mon 7 Jan - Fri 11 Jan 2008. Each day covers trading hours between 9.30am-4pm (in the New York local time). Right: autocorrelation of IBM10m, Jan 2007 - Dec 2010.
The model

- **Spline-DCS model.**

  \[ y_{t,\tau} = \varepsilon_{t,\tau} \exp(\lambda_{t,\tau}), \quad \varepsilon_{t,\tau} | \mathcal{F}_{t,\tau-1} \sim \text{i.i.d. } F(\varepsilon; \theta) \]

  \[ \lambda_{t,\tau} = \delta + s_{t,\tau} + \mu_{t,\tau} + \eta_{t,\tau} \]

- **\( s_{t,\tau} \): periodic component** capturing diurnal patterns

- **\( \mu_{t,\tau} \): low-frequency nonstationary component.**

  \[ \mu_{t,\tau} = \mu_{t,\tau-1} + \kappa_{\mu} u_{t,\tau-1} \]

- **\( \eta_{t,\tau} \): stationary component.** A mixture of AR to capture behavior similar to long-memory.

  \[ \eta_{t,\tau} = \eta_{t,\tau}^{(1)} + \eta_{t,\tau}^{(2)}, \quad \eta_{t,\tau}^{(1)} = \phi_{1}^{(1)} \eta_{t,\tau-1}^{(1)} + \phi_{2}^{(1)} \eta_{t,\tau-2}^{(1)} + \kappa_{\eta}^{(1)} u_{t,\tau-1} \]

  \[ \eta_{t,\tau}^{(2)} = \phi_{1}^{(2)} \eta_{t,\tau-1}^{(2)} + \kappa_{\eta}^{(2)} u_{t,\tau-1} \]

- **\( u_{t,\tau} \): the score** of distribution of \( y_{t,\tau} \) (i.e. \( \partial f_{y}(y_{t,\tau})/\partial \lambda_{t,\tau} \)).

DCS = dynamic conditional score [Harvey (2013) and Creal, Koopman, and Lucas (2011, 2013)]
Dynamic cubic spline

- $s_{t,\tau}$: dynamic cubic spline (Harvey and Koopman (1993))

$$s_{t,\tau} = \sum_{j=1}^{k} \mathbb{1}_{\{\tau \in [\tau_{j-1}, \tau_j]\}} \mathbb{1}_{\tau} (\tau) \cdot \gamma$$

- **Fixed v.s. dynamic spline**: let $\gamma \rightarrow \gamma_{t,\tau}$ where

$$\gamma_{t,\tau} = \gamma_{t,\tau-1} + \kappa^* \cdot u_{t,\tau-1}$$

**Figure**: Fixed spline (left) and dynamic spline (right).
Dynamic cubic spline (ctd)

- $s_{t,\tau}$: dynamic cubic spline (Harvey and Koopman (1993))

\[ s_{t,\tau} = \sum_{j=1}^{k} \mathbb{1}_{\{\tau \in [\tau_{j-1}, \tau_{j}]\}} z_{j}(\tau) \cdot \gamma \]

- $k$: number of knots
- $\tau_0 < \tau_1 < \cdots < \tau_k$: coordinates of the knots along time-axis
- $\gamma = (\gamma_1, \ldots, \gamma_k)^\top$: y-coordinates (height) of the knots
- $z_{j} : [\tau_{j-1}, \tau_{j}]^{k} \rightarrow \mathbb{R}^{k}$: $k$-dimensional vector of weighting functions. Conveys information about (i) polynomial order, (ii) continuity, (iii) length of periodicity, and (iv) zero-sum conditions.
- Bowsher and Meeks (2008): “special type of dynamic factor model”
- Time-varying spline: let $\gamma \rightarrow \gamma_{t,\tau}$ where

\[ \gamma_{t,\tau} = \gamma_{t,\tau-1} + k^* \cdot u_{t,\tau-1} \]
Why use this dynamic spline?

Alternative options used by many:

- Fourier representation
- Sample moments for each intra-day bins
- Diurnal pattern = deterministic function of intra-day time


So why use this spline?

- Allows for changing diurnal patterns (can improve upper quantile fit)
- No need for a two-step procedure to “diurnally adjust” data
- Allow for the day-of-the-week effect via changes in shape of diurnal patterns as well as level shift.
  - Unlike the alternative: seasonal dummies. Test for level differences. Used by many (e.g. Andersen and Bollerslev (1998), Lo and Wang (2010))
Core assumption: $\hat{\epsilon}_{t,\tau} = y_{t,\tau}/\exp(\hat{\lambda}_{t,\tau})$ has to be free of autocorrelation.
Satisfied - no autocorrelation in $\hat{\epsilon}_{t,\tau}$.

Figure: IBM30s: sample autocorrelation of trade volume (top), of $\hat{\epsilon}_{t,\tau}$ (left). The 95% confidence interval is computed at $\pm 2$ standard errors.
Estimation results (short sample period)

$F \sim \text{GB2}$ distribution fits very well. ($\text{GB2} = \text{generalized beta distribution of the second kind}$.)

PIT: $F(\hat{\varepsilon}_{t,\tau}) \sim \text{U}[0,1]$. Fit seems to be the best when our spline is time-varying.

Figure: Empirical cdf of $\hat{\varepsilon}_{t,\tau} > 0$ against cdf of GB2 ($\hat{\nu}, \hat{\zeta}, \hat{\xi}$) (left). Empirical cdf of the PIT of $\hat{\varepsilon}_{t,\tau} > 0$ computed under $F(\cdot; \theta) \sim \text{GB2} (\hat{\nu}, \hat{\zeta}, \hat{\xi})$ (right).
Compare with log-normal distribution

- Log-normal distribution popular. Often used in literature. (e.g. Alizadeh, Brandt, Diebold (2002))
- But log-normal inferior to GB2.
- PIT of $\hat{\varepsilon}_{t,\tau}$ far from $U[0,1]$. Why?

**Figure:** \textbf{Log(trade volume)}: The frequency distribution (left) and the QQ-plot (right). Using non-zero observations re-centered around mean and standardized by one standard deviation.
### Estimated coefficients

<table>
<thead>
<tr>
<th></th>
<th>IBM30s</th>
<th>IBM1m</th>
<th>IBM30s</th>
<th>IBM1m</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa_{\mu}$</td>
<td>0.006 (0.001)</td>
<td>0.007 (0.002)</td>
<td>$\gamma_{0;1,0}$</td>
<td>1.273 (0.106)</td>
</tr>
<tr>
<td>$\phi_{1}^{(1)}$</td>
<td>0.557 (0.136)</td>
<td>0.377 (0.093)</td>
<td>$\gamma_{1;1,0}$</td>
<td>0.078 (0.058)</td>
</tr>
<tr>
<td>$\phi_{2}^{(1)}$</td>
<td>0.410 (0.135)</td>
<td>0.567 (0.096)</td>
<td>$\gamma_{2;1,0}$</td>
<td>-0.469 (0.070)</td>
</tr>
<tr>
<td>$\kappa_{\eta}^{(1)}$</td>
<td>0.049 (0.007)</td>
<td>0.045 (0.008)</td>
<td>$\gamma_{3;1,0}$</td>
<td>-0.227 (0.047)</td>
</tr>
<tr>
<td>$\phi_{1}^{(2)}$</td>
<td>0.688 (0.041)</td>
<td>0.621 (0.057)</td>
<td>$\omega$</td>
<td>9.146 (0.174)</td>
</tr>
<tr>
<td>$\kappa_{\eta}^{(2)}$</td>
<td>0.092 (0.008)</td>
<td>0.069 (0.008)</td>
<td>$\nu$</td>
<td>1.631 (0.016)</td>
</tr>
<tr>
<td>$\kappa_{0}^{*}$</td>
<td>0.003 (0.002)</td>
<td>0.003 (0.002)</td>
<td>$\zeta$</td>
<td>1.486 (0.045)</td>
</tr>
<tr>
<td>$\kappa_{1}^{*}$</td>
<td>0.001 (0.001)</td>
<td>0.000 (0.001)</td>
<td>$p$</td>
<td>0.0047 (0.0005)</td>
</tr>
<tr>
<td>$\kappa_{2}^{*}$</td>
<td>-0.002 (0.001)</td>
<td>-0.002 (0.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\kappa_{3}^{*}$</td>
<td>0.000 (0.001)</td>
<td>0.000 (0.001)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parametric assumptions, identifiability requirements satisfied.

$\eta_{t,\tau}$ stationary.

$\hat{p}$ is consistent with sample statistics.
Estimation results (long sample period): fit of distribution

\[ \varepsilon_{t,\tau} \sim \text{GB2} \text{ fits very well. PIT: } F(\hat{\varepsilon}_{t,\tau}) \sim U[0,1]. \]

Figure: Empirical cdf of \( \hat{\varepsilon}_{t,\tau} > 0 \) against cdf of \( \text{GB2}(\hat{\nu}, \hat{\zeta}, \hat{\xi}) \) (left). Empirical cdf of the PIT of \( \hat{\varepsilon}_{t,\tau} > 0 \) computed under \( F(\cdot; \theta) = \text{GB2}(\hat{\nu}, \hat{\zeta}, \hat{\xi}) \) (right).
Estimation results (long sample period): check autocorrelation in $\hat{u}_{t,\tau}$

Should have $\hat{u}_{t,\tau} \sim \text{i.i.d.}$ [and Beta distributed.]

$s_{t,\tau}$ is **Fourier (left)** and our **fixed spline (right)**. The number of coefficients in $s_{t,\tau}$ are the same.

**Figure:** IBM10m: sample autocorrelation of $\hat{u}_{t,\tau}$. The periodic component $s_{t,\tau}$ is Fourier (left) and our fixed spline (right). The 95% confidence interval is computed at $\pm 2$ standard errors.
Estimated dynamic cubic spline, $\hat{s}_{t,\tau}$ (long sample period)

$\hat{s}_{t,\tau}$: dynamic cubic spline. Reflects diurnal patterns that evolve over time.

**Figure:** IBM10m: $\exp(\hat{s}_{t,\tau})$. Sampling period is Jan 2007 - Dec 2010. Trading time between 9.30am-4pm.
Estimated dynamic cubic spline, $\hat{s}_{t, \tau}$

Reflects diurnal patterns that evolve over time.

Figure: $\hat{s}_{t, \tau}$ of Model 2 for IBM30s. Over 6 - 31 March 2000 (left). $\hat{s}_{t, \tau}$ of a typical day, Tuesday 14 March, from market open to close (right). Time along the x-axes.

Day-of-the-week effect? Do we need dynamic periodicity?
Out-of-sample performance (long sample period)

One-step ahead forecasts, $\tilde{\varepsilon}_{t,\tau} = y_{t,\tau}/\exp(\tilde{\lambda}_{t,\tau})$ for 50 days without re-estimating parameters. Forecast horizon: January - March 2011.

Figure: Left: PIT of forecast $\tilde{\varepsilon}_{t,\tau}$, Dynamic Spline. Right: QQ-plot of forecast $\tilde{\varepsilon}_{t,\tau}$, Dynamic Spline (blue) and Fixed Spline (red).
Out-of-sample performance

- Our model and parameter estimates are stable
- One-step ahead density forecasts (without re-estimation): very good for at least 20 days ahead.
- Multi-step ahead density forecasts: very good (i.e. PIT approx. iid $\sim U[0,1]$) for one complete trading-day ahead (equivalent of 780 steps for IBM30s).

More discussions in the paper.
Intuition, future direction

- Dynamic spline can reflect changes in the pattern of morning trading activity (i.e. how we standardize large-sized morning observations).
- Important feature when the amount (or nature) of overnight news can change morning trading patterns.

Still to do:
- Further investigate the performance of the model at the upper-tail.
- Multi-variate version: price and volume.
- Model for higher-frequency: 1 second?
- Application to panel data (using composite likelihood?)
- Asymptotic properties of MLE when DCS non-stationary.
- etc.