Modeling dynamic diurnal patterns in high frequency financial data

Ryoko Ito[1]

[1]Faculty of Economics,
Cambridge University

Tinbergen Institute Amsterdam, January 2013

Introduction

- We want to build a model for high frequency observations of financial data
- Not easy: need to capture stylized features
  - Concentration of zero-observations
  - Diurnal patterns
  - Skewness, heavy tail
  - Seasonality (intra-weekly, monthly, quarterly...)
  - Highly persistent dynamics (long-memory?)
- An extension of DCS model works very well! Several advantages over other existing methods.
- Methods:
  - Distribution decomposition at zero
  - Unobserved components
  - Dynamic cubic spline (Harvey and Koopman (1993))

Data

- *Trade volume* of IBM stock traded on the NYSE. The number of shares traded.
- Period: 5 consequtive trading weeks in February - March 2000
- Sampling frequency: 30 seconds

Empirical features

Diurnal U-shaped patterns

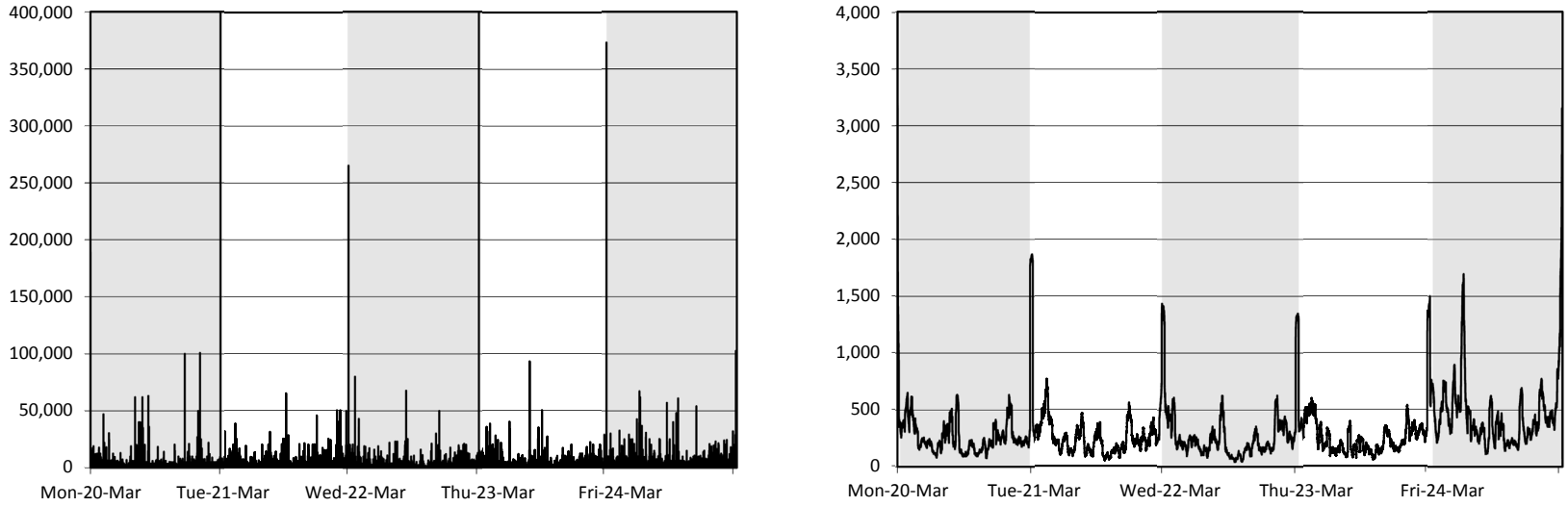

Figure: IBM trade volume (left column) and the same series smoothed by the simple moving average (right column). Time on the x-axis. Monday 20 - Friday 24 March 2000. Each day covers trading hours between 9.30am-4pm (in the New York local time).

Empirical features

Trade volume bottoms out at around 1pm.

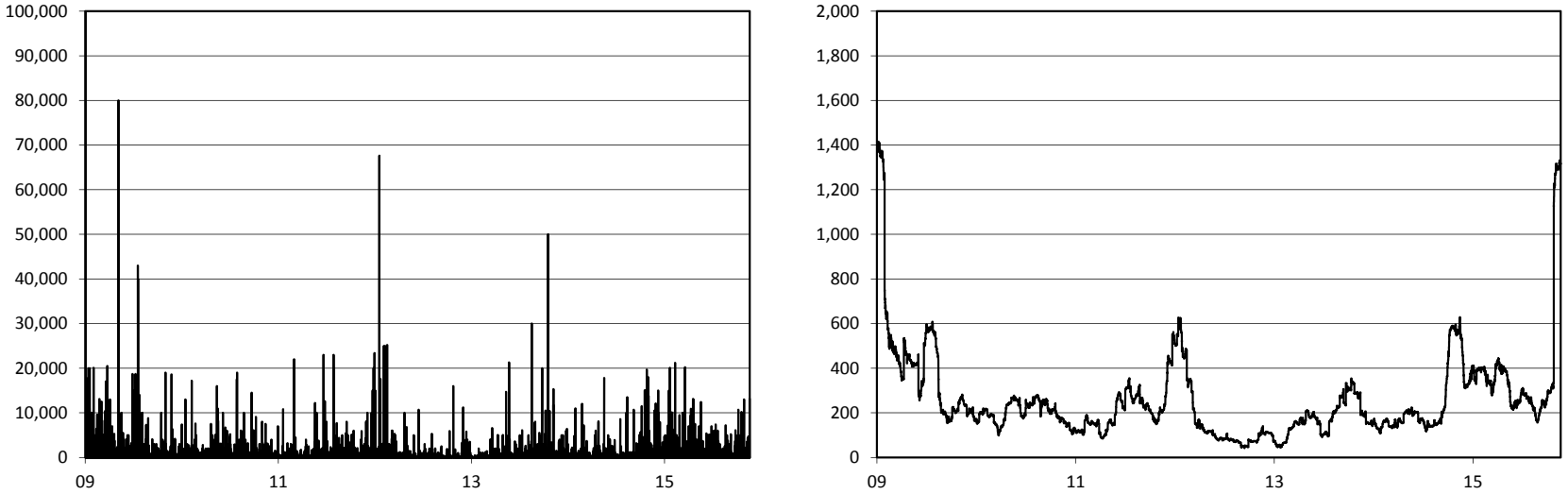

Figure: IBM trade volume (left column) and the same series smoothed by the simple moving average (right column). Time on the x-axis. Wednesday 22 March 2000, covering 9.30am-4pm (in the New York local time).

Empirical features

Sample autocorrelation. Highly persistent.

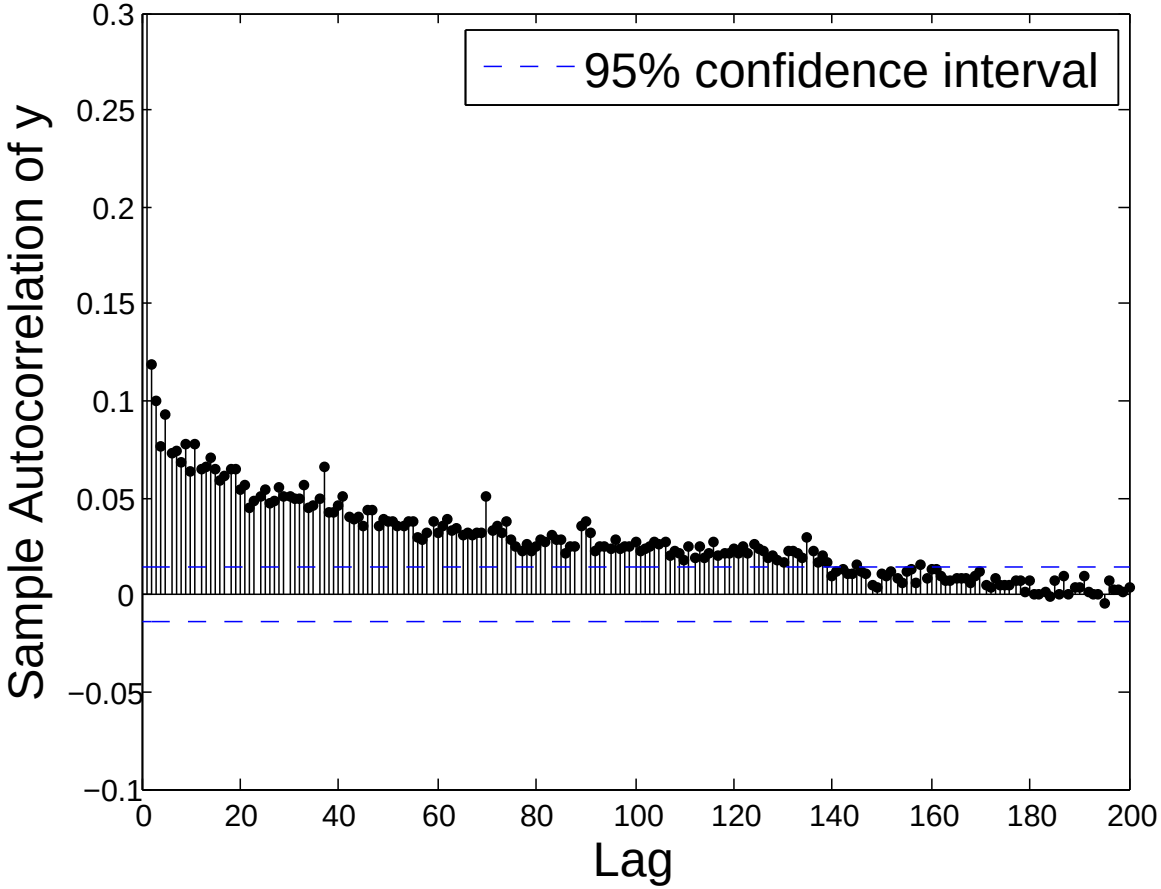

Figure: Sample autocorrelation of IBM trade volume. Sampling period: 28 February - 31 March 2000. The 200th lag corresponds approximately to 1.5 hours prior.

Empirical features

Skewness, long upper-tail.

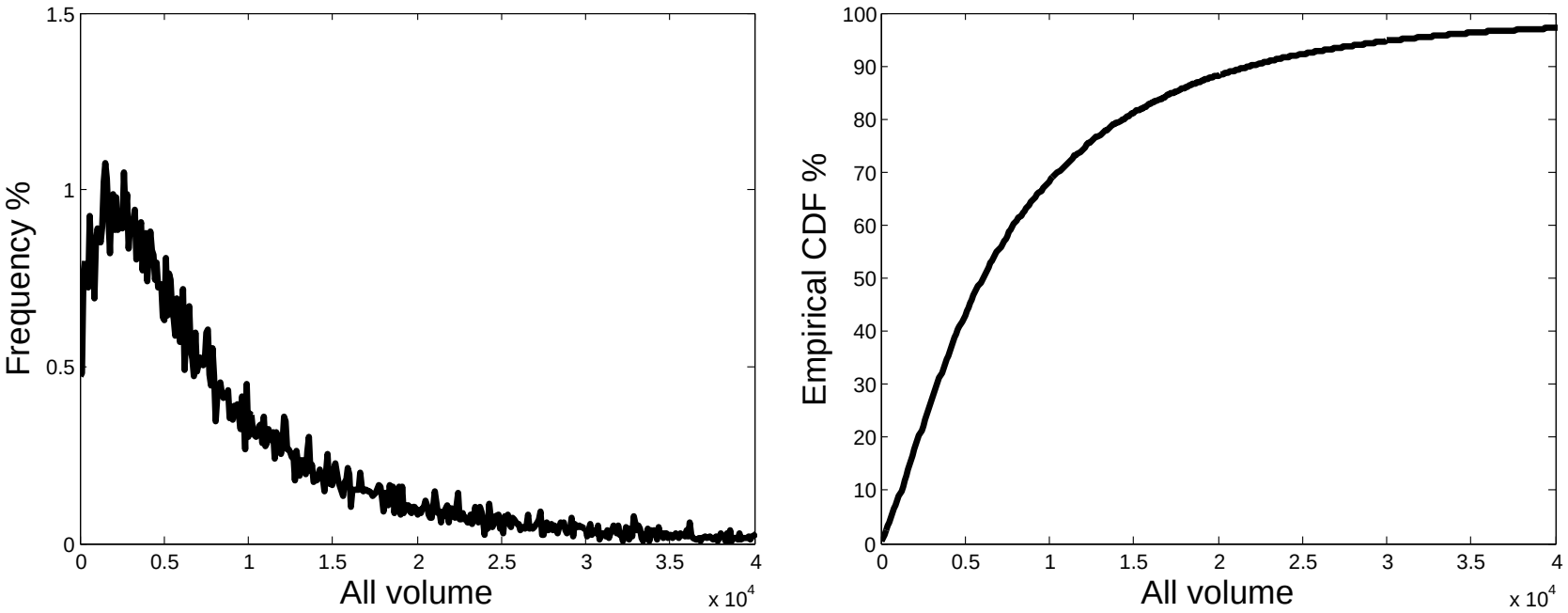

Figure: Frequency distribution (left) and empirical cdf (right) of IBM trade volume. Sampling period: 28 February - 31 March 2000.

Empirical features

Sample statistics.

Observations (total)	19,500
Mean	10,539
Median	6,100
Maximum	1,652,100
Minimum	0
Std. Dev.	26,071
Skewness	29
99.9% sample quantile	293,654
Max - 99.9% quantile	1,358,446
Frequency of zero-obs.	**0.47% (92 obs.)**

Intra-day DCS with dynamic cubic spline

Our model: intra-day DCS with dynamic cubic spline

Time index: intra-day time τ on t-th trading day as $\cdot_{t,\tau}$
$\tau = 0, \ldots, I$ and $t = 1, \ldots, T$.
$\tau = 0$ and $\tau = I$: the moments of market open and close.

Distribution decomposition at zero

- Define CDF $F : \mathbb{R}_{\geq 0} \to [0,1]$ of a standard random variable $X \sim F$
  - the origin has a discrete mass of probability
  - strictly positive support is captured by a conventional continuous distribution
- Formally,

$$\begin{aligned} \mathbb{P}_F(X = 0) &= p & p \in [0,1] \\ \mathbb{P}_F(X > 0) &= 1 - p & \\ \mathbb{P}_F(X \leq x | X > 0) &= F^*(x) & x > 0 \end{aligned} \tag{1}$$

  $F^* : \mathbb{R}_{>0} \to [0,1]$ is the cdf of a conventional standard continuous random variable with constant parameter vector θ^*.

- Decomposition technique: Amemiya (1973), Heckman (1974), McCulloch and Tsay (2001) Rydberg and Shephard (2003), Hautsch, Malec, and Schienle (2010)

- Our p is constant. OK as the number of zero-observations is small.
  - Possible extension: time-varying $p_{t,\tau}$ using logit link [Rydberg and Shephard (2003), Hautsch, Malec, and Schienle (2010)]
- Apply DCS filter *only* to positive observations.
  - irregular term $u_{t,\tau}$: the score of F^*
  - $\lambda_{t,\tau}$ driven by $u_{t,\tau-1}\mathbb{1}_{\{y_{t,\tau-1}>0\}}$

Unobserved components

Assumption 1: periodicity and autocorrelation in data are due to the time-varying scale parameter $\alpha_{t,\tau} = \exp(\lambda_{t,\tau})$. Standardized observations are `iid` and free of periodicity and autocorrelation.

$\Rightarrow$ Let $\lambda_{t,\tau}$ have a component structure to capture each feature of data.

Unobserved components

- **Unobserved components**:

$$y_{t,\tau} = \varepsilon_{t,\tau} \exp(\lambda_{t,\tau}), \quad \varepsilon_{t,\tau} | \mathcal{F}_{t,\tau-1} \sim \texttt{iid} F$$
$$\lambda_{t,\tau} = \delta + \mu_{t,\tau} + \eta_{t,\tau} + s_{t,\tau}$$

- $\mu_{t,\tau}$: low-frequency component. Captures highly persistent nonstationary dynamics

$$\mu_{t,\tau} = \mu_{t,\tau-1} + \kappa_\mu u_{t,\tau-1} \mathbb{1}_{\{y_{t,\tau-1}>0\}}$$

- $\eta_{t,\tau}$: stationary (autoregressive) component. A mixture of AR components captures long-memory.

$$\eta_{t,\tau} = \sum_{j=1}^{J} \eta_{t,\tau}^{(j)}$$
$$\eta_{t,\tau}^{(j)} = \phi_1^{(j)} \eta_{t,\tau-1}^{(j)} + \phi_2^{(j)} \eta_{t,\tau-2}^{(j)} \cdots + \phi_m^{(j)} \eta_{t,\tau-m^{(j)}}^{(j)} + \kappa_\eta^{(j)} u_{t,\tau-1} \mathbb{1}_{\{y_{t,\tau-1}>0\}}$$

  for some $J \in \mathbb{N}_{>0}$ and $m^{(j)} \in \mathbb{N}_{>0}$.

- $s_{t,\tau}$: periodic component capturing diurnal patterns

Dynamic cubic spline

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

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

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

$$\underline{\gamma}^{\dagger}_{t,\tau} = \underline{\gamma}^{\dagger}_{t,\tau-1} + \underline{\kappa}^{*} \cdot u_{t,\tau-1} \mathbb{1}_{\{y_{t,\tau-1} > 0\}}$$

Why use this spline?

Alternative options used by many:

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

(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).)

So why use this spline?

- Allows for changing diurnal patterns
- No need for a two-step procedure to "diurnally adjust" data
- Formal test for the day-of-the-week effect. Compare shape of dirunal patterns.
  - Unlike the alternative: seasonal dummies. Test for level differences. Used by many (e.g. Andersen and Bollerslev (1998), Lo and Wang (2010))

Estimation

Apply spline-DCS model to IBM trade volume data

Estimation results

Assumption 1: $\widehat{\varepsilon}_{t,\tau} = y_{t,\tau}/\widehat{\alpha}_{t,\tau}$ has to be free of autocorrelation.
Satisfied - no autocorrelation in $\widehat{\varepsilon}_{t,\tau}$.

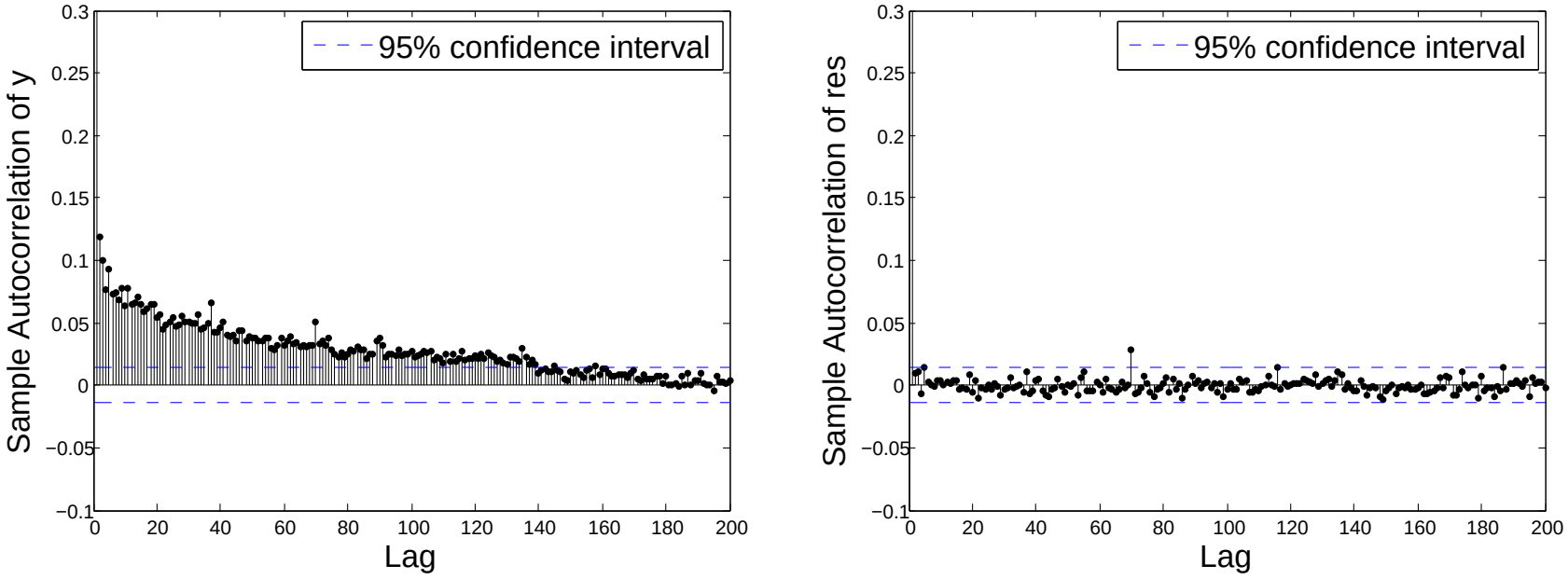

Figure: Sample autocorrelation of trade volume (top), of $\widehat{\varepsilon}_{t,\tau}$ (left). The 95% confidence interval is computed at ± 2 standard errors.

Estimation results

$F^* \sim$ Burr distribution fits very well. PIT: $F^*(\widehat{\varepsilon}_{t,\tau}) \sim$ U[0,1].

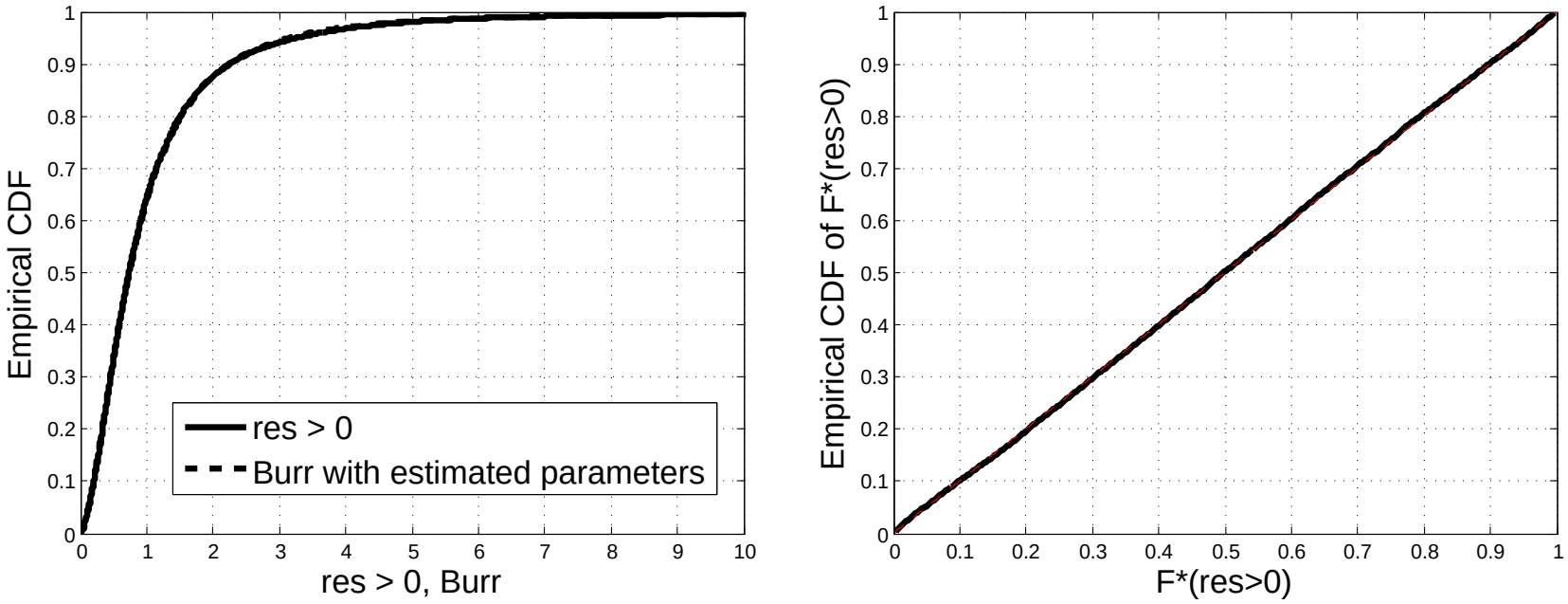

Figure: Empirical cdf of $\widehat{\varepsilon}_{t,\tau} > 0$ against cdf of Burr$(\widehat{\nu}, \widehat{\zeta})$ (left). Empirical cdf of the PIT of $\widehat{\varepsilon}_{t,\tau} > 0$ computed under $F^*(\cdot;\theta^*) \sim$Burr$(\widehat{\nu}, \widehat{\zeta})$ (right).

Compare with log-normal distribution

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

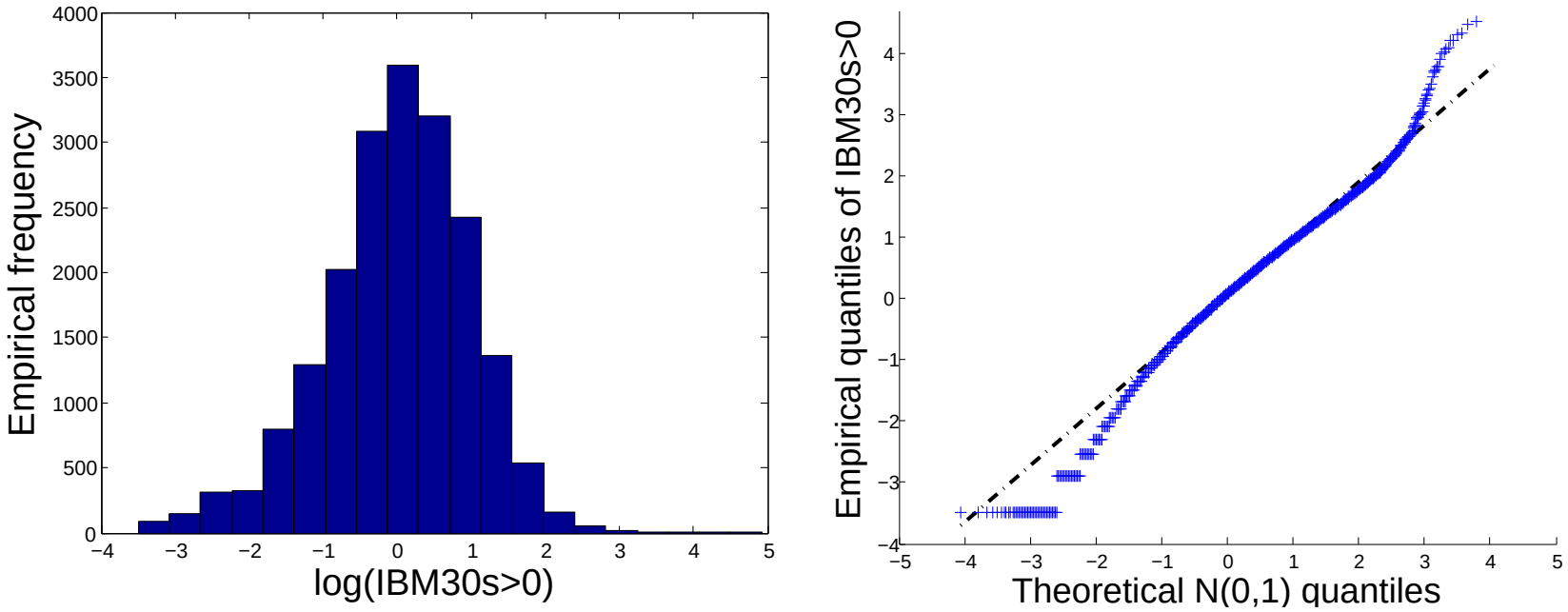

Figure: Log(trade volume): The frequency distribution (left) and the QQ-plot (right). Using non-zero observations re-centered around mean and standardized by one standard deviation.

Estimated spline component

Reflects diurnal patterns that evolve over time.

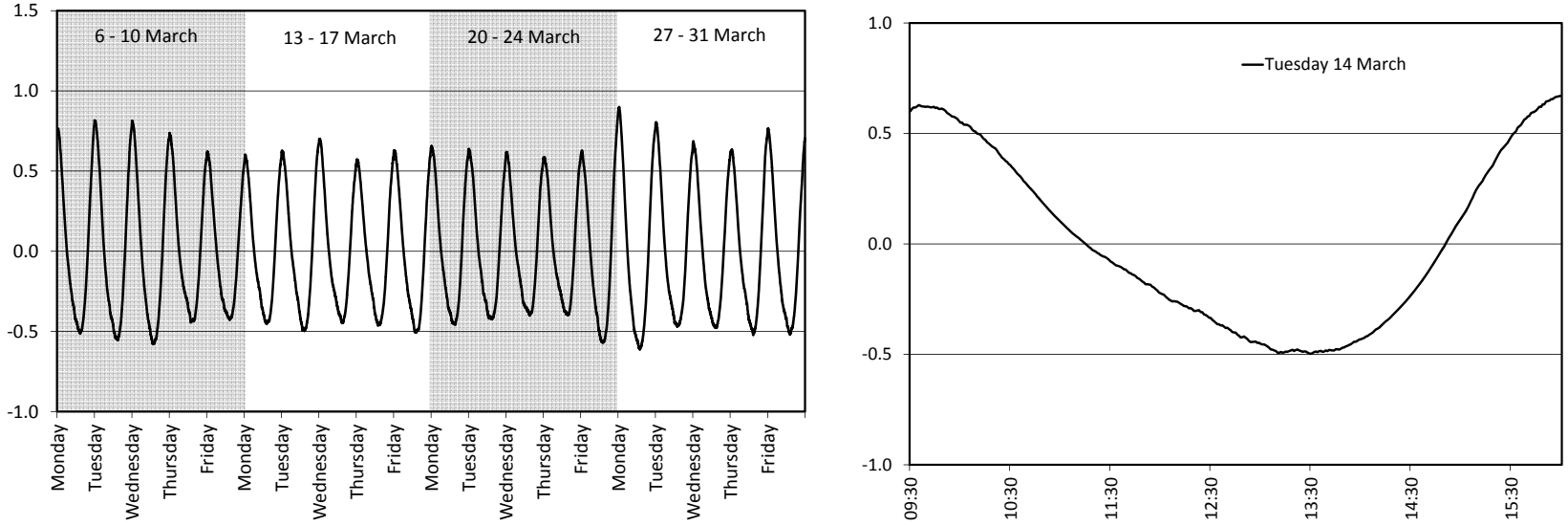

Figure: $\widehat{s}_{t,\tau}$: over 6 - 31 March 2000 (left) and of a typical day, Tuesday 14 March, from market open to close (right).

Formal test for the day-of-the-week effect (a likelihood ratio test)
$\Rightarrow$ there is no statistically significant evidence in data.

Overnight effect

Standard methods:

- Dummy variables. Treat extreme observations as outliers.
- Differentiate day and overnight jumps
- Treat morning events as censored.

(e.g. Rydberg and Shephard (2003), Gerhard and Hautsch (2007), Boes, Drost and Werker (2007).)

Issues:

- It may take time for the overnight effect to diminish completely during the day
- Difficult to identify which observations are due to overnight information

Overnight effect

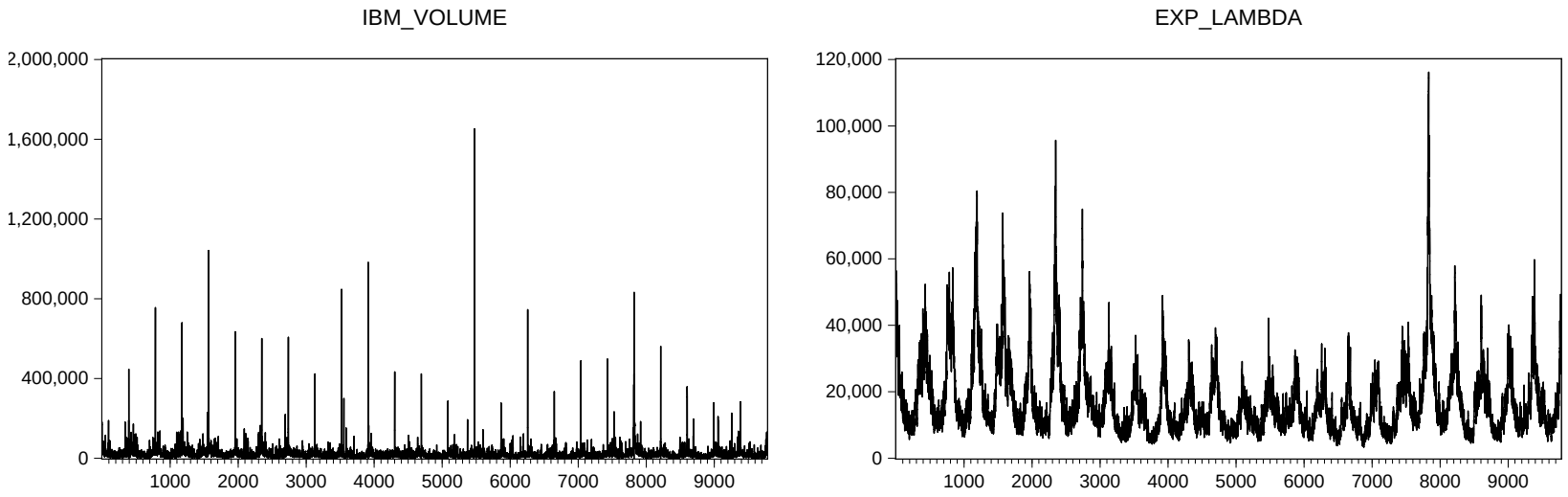

Figure: Capturing overnight effect: trade volume (left) and $\widehat{\alpha}_{t,\tau} = \exp(\widehat{\lambda}_{t,\tau})$ (right).

Periodic hikes in scale parameter $\widehat{\alpha}_{t,\tau}$.
Reflects cyclical surge in market activity.
Overnight information interpreted as the elevated probability of extreme events.
Advantage of the exponential link.

Long memory

Two component specification for $\eta_{t,\tau}$ works well:

$$\eta_{t,\tau} = \eta^{(1)}_{t,\tau} + \eta^{(2)}_{t,\tau}$$
$$\eta^{(1)}_{t,\tau} = \phi^{(1)}_1 \eta^{(1)}_{t,\tau-1} + \phi^{(1)}_2 \eta^{(1)}_{t,\tau-2} + \kappa^{(1)}_\eta u_{t,\tau-1} \mathbb{1}_{\{y_{t,\tau-1}>0\}}$$
$$\eta^{(2)}_{t,\tau} = \phi^{(2)}_1 \eta^{(2)}_{t,\tau-1} + \kappa^{(2)}_\eta u_{t,\tau-1} \mathbb{1}_{\{y_{t,\tau-1}>0\}}$$

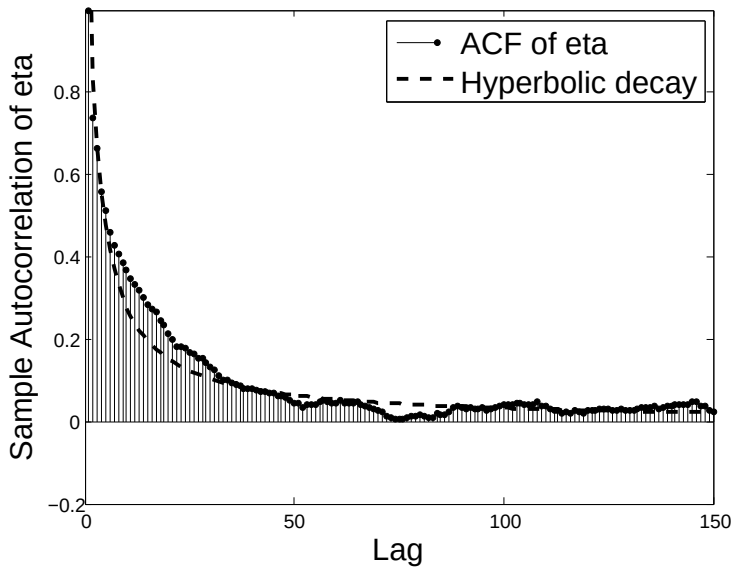

Figure: Autocorrelation function of $\widehat{\eta}_{t,\tau}$.

The degree of fractional integration in $\eta_{t,\tau}$: $\widehat{d} = 0.140$ (s.e. 0.057).
Picking up long-memory in data.

Estimated coefficients

κ_μ	0.007 (0.002)	$\gamma^\dagger_{1;1,0}$	0.122 (0.070)
$\phi_1^{(1)}$	0.561 (0.129)	$\gamma^\dagger_{2;1,0}$	-0.485 (0.079)
$\phi_2^{(1)}$	0.400 (0.129)	$\gamma^\dagger_{3;1,0}$	-0.229 (0.058)
$\kappa_\eta^{(1)}$	0.053 (0.008)	δ	9.254 (0.181)
$\phi_1^{(2)}$	0.676 (0.046)	ν	1.635 (0.016)
$\kappa_\eta^{(2)}$	0.091 (0.009)	ζ	1.467 (0.042)
κ_1^*	0.000 (0.001)	p	0.0047 (0.0005)
κ_2^*	-0.002 (0.001)		
κ_3^*	0.000 (0.001)		

Parametric assumptions, identifiability requirements satisfied.
$\eta_{t,\tau}$ stationary.
$\widehat{p}$ is consistent with sample statistics.

Out-of-sample performance

- Our model and estimation results are stable
- One-step ahead density forecasts (without re-estimation): very good for 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).

More details and discussions in the paper.

Out-of-sample performance

Multi-step density forecasts

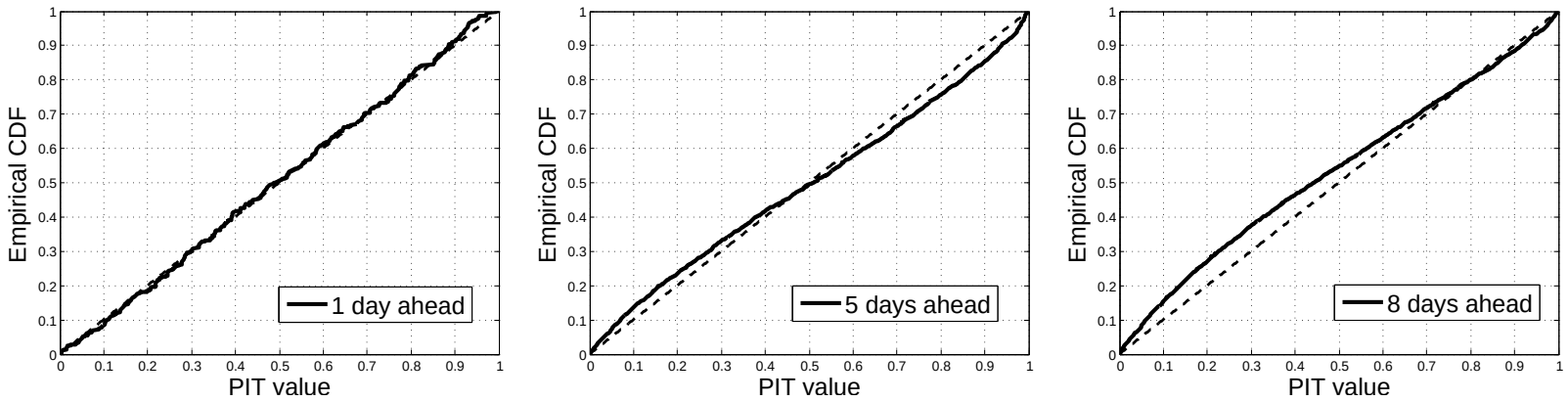

Figure: Empirical cdf of the PIT of multi-step forecasts. Forecast horizons: 1 day ahead (left), 5 days ahead (middle), 8 days ahead (right).

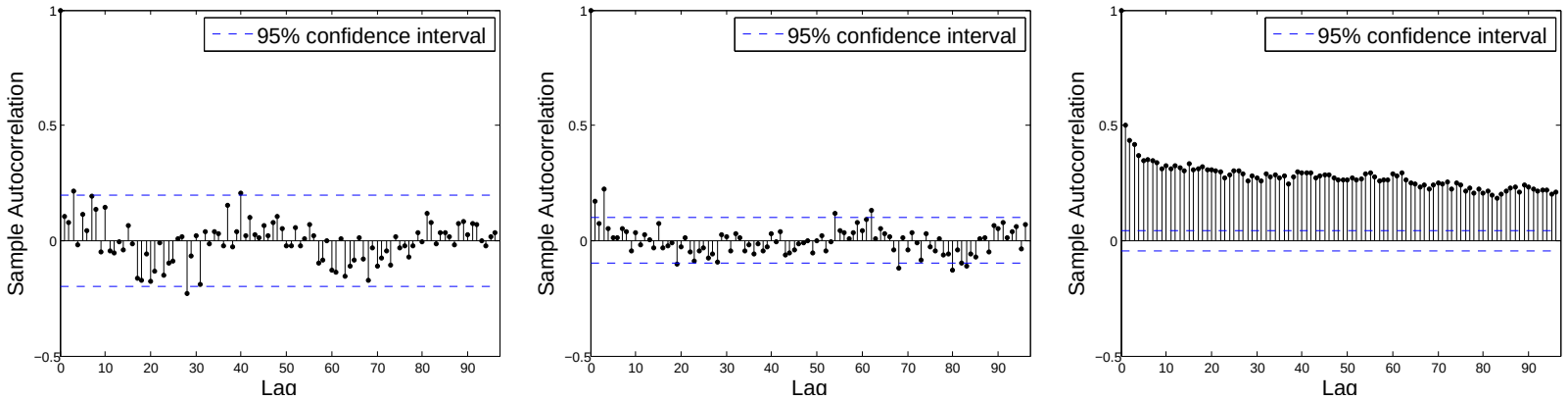

Figure: Autocorrelation of the PIT of multi-step forecasts. Forecast horizons: 1.5 hours ahead (left), one day ahead (middle), five days ahead (right).

Future directions

Model for higher-frequency: 1 second?
Asymptotic properties of MLE when DCS non-stationary
Multi-variate version: price and volume
Application to panel data (using composite likelihood?)
etc. etc.