SFB 649 Discussion Paper 2011-044

Predicting Bid-Ask Spreads Using Long Memory Autoregressive Conditional Poisson Models

Axel Groß-Klußmann*
Nikolaus Hautsch*

* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Predicting Bid-Ask Spreads Using Long Memory Autoregressive Conditional Poisson Models *

Axel Groß-Klußmann†
Humboldt-Universität zu Berlin

Nikolaus Hautsch‡
Humboldt-Universität zu Berlin, CASE, CFS

This version: July 2011

Abstract

We introduce a long memory autoregressive conditional Poisson (LMACP) model to model highly persistent time series of counts. The model is applied to forecast quoted bid-ask spreads, a key parameter in stock trading operations. It is shown that the LMACP nicely captures salient features of bid-ask spreads like the strong autocorrelation and discreteness of observations. We discuss theoretical properties of LMACP models and evaluate rolling window forecasts of quoted bid-ask spreads for stocks traded at NYSE and NASDAQ. We show that Poisson time series models significantly outperform forecasts from ARMA, ARFIMA, ACD and FIACD models. The economic significance of our results is supported by the evaluation of a trade schedule. Scheduling trades according to spread forecasts we realize cost savings of up to 13 % of spread transaction costs.

Keywords: Bid-ask spreads, forecasting, high-frequency data, stock market liquidity, count data time series, long memory Poisson autoregression.

---

*For helpful comments and discussions we thank the participants of workshops at Humboldt-Universität zu Berlin. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) via the Collaborative Research Center 649 "Economic Risk".

†Corresponding author. Institute for Statistics and Econometrics, Humboldt-Universität zu Berlin. Email: axel.gross-klussmann@wiwi.hu-berlin.de. Address: Spandauer Str. 1, D-10178 Berlin, Germany.

‡Institute for Statistics and Econometrics and Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin and Center for Financial Studies (CFS), Frankfurt. Email: nikolaus.hautsch@wiwi.hu-berlin.de. Address: Spandauer Str. 1, D-10178 Berlin, Germany.

JEL classification: G14, C32

1 Introduction

Bid-ask spreads reflect the fundamental costs of immediate trading, i.e., the cost of constantly guaranteeing a counterparty for trades to market participants. They are an important determinant of liquidity on stock markets and thus play a dominant role in the literature on market microstructure and stock trading. Theoretical models on market making along the lines of Copeland and Galai (1983), Kyle (1985), Glosten and Milgrom (1985) and Easley and O'Hara (1992) emphasize the adverse selection component in bid-ask spreads indicating information asymmetry among market participants. Based on game-theoretical dynamic models, Foucault (1999) and Foucault et al. (2005) argue that the bid-ask spread is *the* dominant parameter for the decision between different order types on stock markets. Traders can either be patient and submit limit orders or cross the spread and pay the bid-ask spread. Empirical studies confirm that limit and market order submission strategies indeed depend strongly on quoted spreads, see Biais et al. (1995), Harris and Hasbrouck (1996), Ranaldo (2004), Anand et al. (2005), Hall and Hautsch (2006) and Pascual and Veredas (2009).

Despite the importance of bid-ask spreads in trading applications and the relevance of spread predictions for the reduction of transaction costs, the question of how to statistically model bid-ask spreads has not been systematically addressed yet. Our paper is the first contribution establishing a concise econometric methodology for modeling and forecasting bid-ask spreads on a high frequency.

Due to the discreteness of prices, bid-ask spreads are multiples of minimum price changes and hence form a time series of count variables. We observe that spreads reveal a pronounced seasonality pattern and are strongly serially dependent. Geweke and Porter-Hudak (1983) (GPH) tests indicate that the bid-ask spread time series exhibit long range dependence. To capture these empirical properties we introduce a novel count data model - the long memory autoregressive conditional Poisson (LMACP) model. Discussing empirical and theoretical properties we show that the model is suitable for the analysis and prediction of strongly persistent discrete time series.

Traditional approaches, such as Glosten and Harris (1988), George et al. (1991),

Huang and Stoll (1997) and Bollen et al. (2004), decompose bid-ask spreads into their adverse selection, inventory holding and transaction components but are silent regarding their dynamic properties. Conversely, more recent time series approaches typically model bid-ask spreads implicitly in models for bid-ask quotes. Engle and Patton (2004) and Hautsch and Huang (2010), for instance, estimate error correction models for bid and ask quotes with the bid-ask spread serving as cointegration relation. Hasbrouck (2000) models the dynamics of bid- and ask quotes separately but does not explicitly focus on the spread. Based on a vector autoregression framework, Taylor (2002) is the only approach deriving spread forecasts.

Our study contributes to the recent literature on high frequency liquidity forecasting. See, e.g., Brownlees et al. (2010) on volume forecasts and Härdle et al. (2009) on forecasts of limit order book curves. Furthermore, the proposed LMACP model contributes to the literature on dynamic count data models as in Rydberg and Shephard (1999), Heinen (2003), Fokianos et al. (2009) and Ferland et al. (2006) among others. Finally, our study is related to count data predictions as in Sutradahar (2008), Jung et al. (2006) and Freeland and McCabe (2004).

Our forecast study is carried out based on representative stocks from the mid cap sector of the US Russell 3000 universe. We report rolling-window forecasts for quoted spreads on a 30 second frequency. The forecast evaluation of point and direction forecasts shows that LMACP models outperform competitors such as autoregressive moving average (ARMA), autoregressive fractionally integrated moving average (ARFIMA), autoregressive conditional duration (ACD) and fractionally integrated autoregressive conditional duration (FIACD) models. Four main results emerge from the analysis. First, we show the importance of explicitly addressing the discrete nature of bid-ask spreads. In particular, approaches based on continuous distributions are outperformed by Poisson models in terms of point, density and direction forecasts. Second, long memory specifications widely perform better than their short memory counterparts in terms of the root mean squared error and the directional accuracy. Third, additional predictors motivated from market microstructure theory, such as trading volume, volatility, first level depth and order imbalance improve forecasts. Fourth, an economic evaluation of a simple trading scheme reveals significant cost savings of up to 13 % when the trading schedules take bid-ask spread forecasts into account.

The remainder of the paper is organized as follows. Section 2 gives descriptive statistics. In Section 3, we outline the econometric model. Section 4 describes the forecasting setup and corresponding evaluation criteria. In Section 5, the forecasting results are presented. Finally, Section 6 concludes.

2 Properties of Bid-Ask Spreads

2.1 The Bid-Ask Spread as a Liquidity Measure

Trading on financial markets requires the presence of a counterparty for trades. According to theoretical models along the lines of Copeland and Galai (1983) and Glosten and Milgrom (1985), liquidity suppliers like market makers, dealers or participants submitting limit orders act as intermediaries and mitigate the search costs by offering immediate trade execution.[1] Empirical studies like Glosten (1987), Glosten and Harris (1988) and Huang and Stoll (1997) suggest that quoted bid-ask spreads are measures of the costs of order processing, inventory holding and adverse selection incurred by these liquidity providers. Liquidity suppliers recoup their own costs in time t by purchasing at the bid price B_t and selling at a higher ask price A_t. As a measure of immediate trading and hence liquidity costs, the quoted bid-ask spread S_t in t is thus given by $S_t \coloneqq A_t - B_t$, where the quotes A_t and B_t are given as multiples of 0.01 (price ticks). A closely related measure is the effective spread, given as $S_t^B \coloneqq P_t - B_t$ for buyer-initiated trades and $S_t^A \coloneqq A_t - P_t$ for seller-initiated trades, where P_t is the transaction price (in multiples of 0.01).

The importance of bid-ask spreads as liquidity measures for practitioners is reflected in limit order submission strategies employed by market participants to reduce trading costs. Limit and market order submission strategies depend strongly on quoted spreads and quoted depth as outlined in Biais et al. (1995), Harris and Hasbrouck (1996), Griffiths et al. (2000), Anand et al. (2005), Parlour (1998), Ranaldo (2004) and Pascual and Veredas (2009).

[1] See Bessembinder and Venkataraman (2010) for an overview on spread-related trading costs.

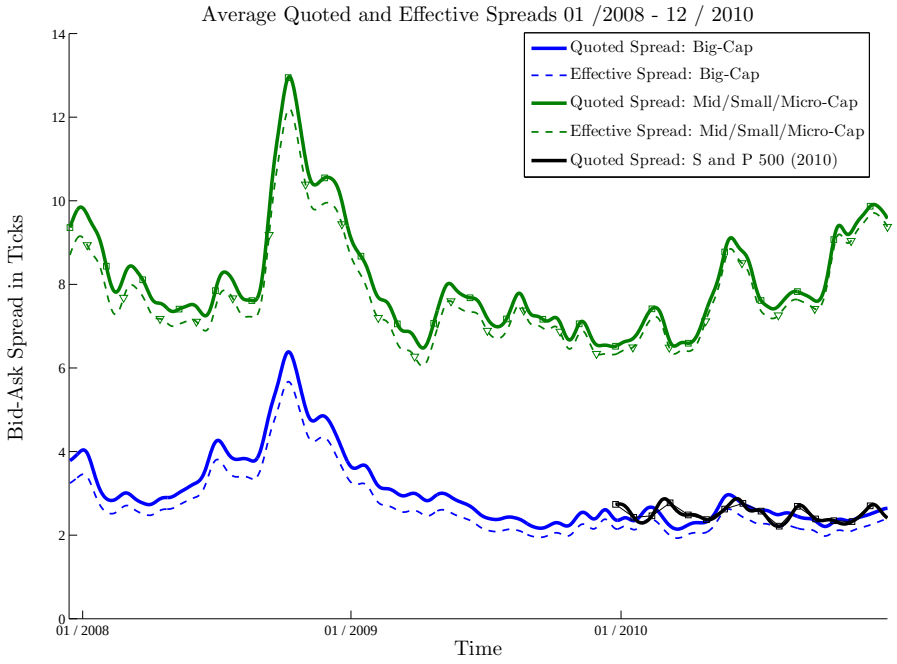

Figure 1: Evolution of the average quoted and average effective bid-ask spreads for 30 big-cap stocks and 80 mid-, small- and micro-cap stocks of the Russell 3000 as well as average quoted spreads of the S&P 500 for 2010. Smoothed via kernel regression

2.2 Empirical Properties

Histograms of spread distributions substantially vary over the Russell 3000 cross-section of the market at NYSE and NASDAQ. We show histograms for a large cross-section for January to February 2008 in the web appendix to the paper, `http://amor.cms.hu-berlin.de/~grosskla/index.html`. In our empirical analysis, we exclude stocks revealing "trivial" spread distributions with spreads being virtually constant to one tick. Rather, we focus on stocks with an average spread of more than two ticks revealing more dispersed distributions. The latter still cover a large fraction of the Russell 3000 index. Figure 1 shows that the average quoted spread for the constituents of the S&P 500 in 2010 is above 2. Moreover, as shown in Figure 1, average bid-ask spreads significantly vary over time. Particularly during the peak of the financial crisis in fall 2008, average spreads nearly doubled compared to the level before.

We employ national best bid and offer (NBBO) quotes and transaction data from

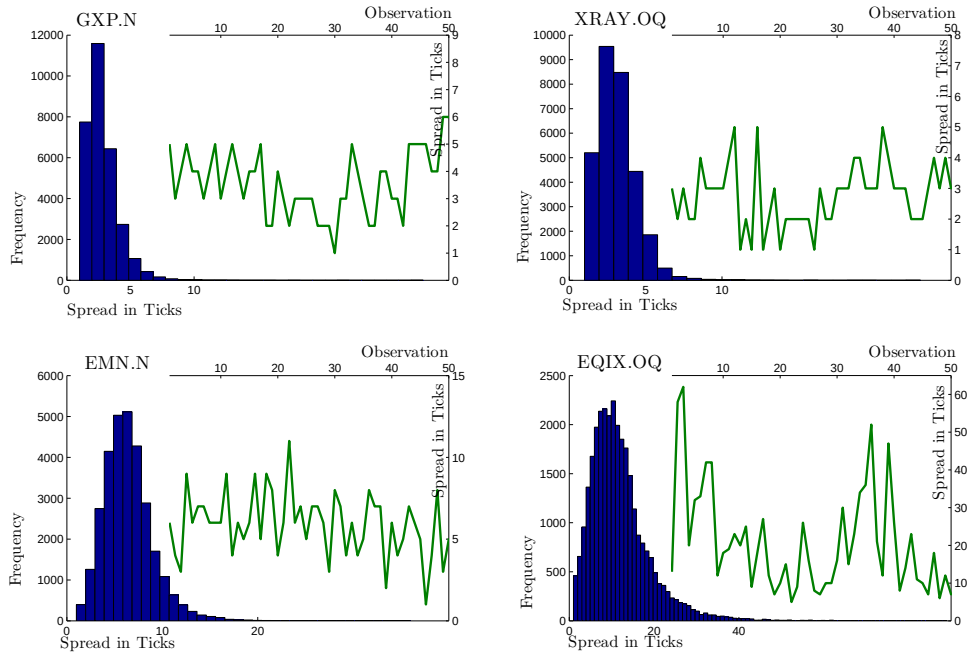

Figure 2: Histogram and typical pattern of the 30s bid-ask spreads for the four selected mid-cap stocks (histogram for Jan. and Feb. 2008)

the Trade and Quote (TAQ) database of the NYSE.[2] Bid-ask spreads are computed as end-point spreads based on a 30 second frequency. We omit the first and last ten minutes of the trading day to reduce the impact of trading starts and stops.

In the paper, we present results for four stocks from the mid-cap sector of the Russell index for January and February 2008 (GXP.N, EMN.N (traded at NYSE) and XRAY.OQ, EQIX.OQ (traded at NASDAQ)). The stocks are chosen to be representative for stocks below the big-cap sector of the Russell as well as for stocks of the big-cap sector with an average spread above two ticks. Figure 2 shows the distributions and snapshots of the time series evolution. Additional and robustifying results for a wide cross-section of stocks are provided in the web appendix.

Due to the discreteness of price ticks (as multiples of 0.01), a time-ordered sequence of quoted bid-ask spreads multiplied by 100 forms a time series of count variables. Table 1 gives descriptive statistics for the 30 second spread time series of the selected stocks. We observe that spread distributions can be both over- and underdispersed, i.e., have

[2] According to the US Securities and Exchange Comission Regulation brokers are required to guarantee customers the best quoted prices across US-based exchanges.

	GXP.N	XRAY.OQ	EMN.N	EQIX.OQ
Mean	2.369	2.719	6.081	11.554
Variance	1.742	1.822	7.022	51.438
Max	28.000	23.000	36.000	89.000
Min	1.000	1.000	1.000	1.000
Median	2.000	3.000	6.000	10.000
10% Quantile	1.000	1.000	3.000	4.000
90% Quantile	4.000	4.000	9.000	20.000
LB_{20}	79853.9	71367.0	43658.7	46874.1
Average Mid-Quote	27.632	42.117	63.809	77.500
Relative Spread in %	8.6	6.5	9.5	14.9

Table 1: Descriptive Statistics of the 30s quoted spread in ticks. LB_{20} denotes the Ljung-Box statistic for 20 lags. The relative spread is the spread fraction of the mid-quote price. Sample period: Jan.-Feb. 2008

variance above or below the means. As shown in Figure 3, the autocorrelation functions (ACFs) of bid-ask spread series decay very slowly and indicate long range dependence. In light of the bid-ask spread as a cointegration relation between ask and bid quotes (see Engle and Patton (2004) and Hautsch and Huang (2010)), these results mean that deviations from equilibria are very persistent.

A time series formally is long range dependent if

$$\lim_{j\to\infty} \rho_j/(cj^{-\alpha}) = 1, \tag{1}$$

where $\alpha \in (0,1)$, $c > 0$ and ρ_j denotes the jth order autocorrelation, see, e.g., Beran (1998). An immediate consequence is that autocorrelations are not absolutely summable. Long range dependence is often found in financial market data (see, among many others, Ding and Granger (1996), Andersen et al. (2003) and Corsi (2009) for volatility data, Lux and Kaizoji (2007) for traded volumes and Deo et al. (2010) for trade durations) as well as in macroeconomic time series (see Bhardwaj and Swanson (2006) for an overview). The presence of long range dependence in spreads is supported by Figure 4 showing a convergence rate of the mean slower than $\sqrt{n}$ for two of the four stocks. More formally, we conduct the Geweke and Porter-Hudak (1983) (GPH) test for long memory, which is is based on the spectral regression

$$\ln(I(\omega_\lambda)) = a + b\ln\left(4\sin^2\left(\frac{\omega_\lambda}{2}\right)\right) + n_\lambda, \quad \lambda = 1,..,v, \tag{2}$$

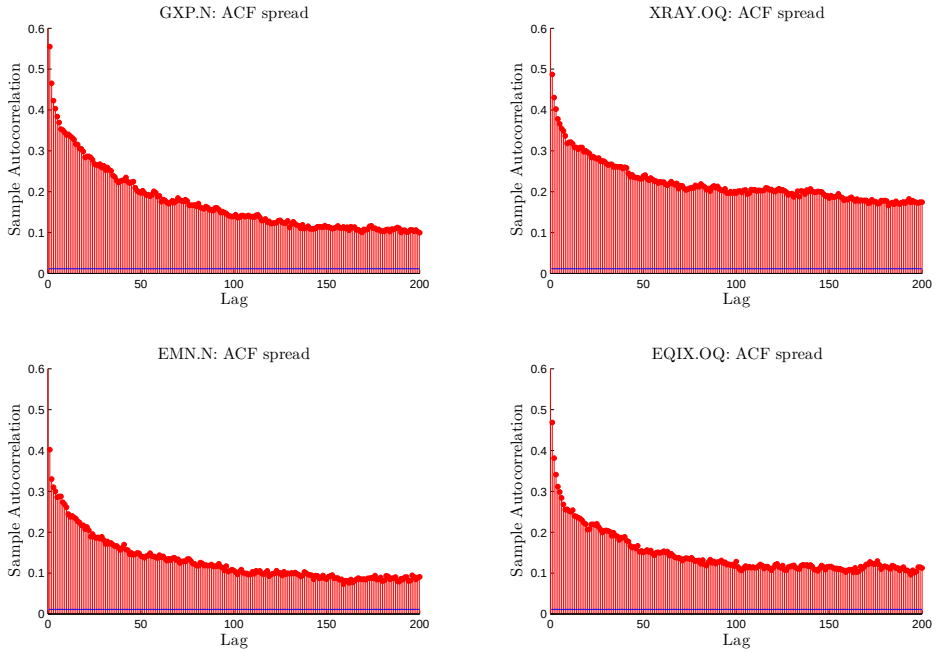

Figure 3: Autocorrelation functions of the 30s spreads for the mid-cap stocks (1-200 lags)

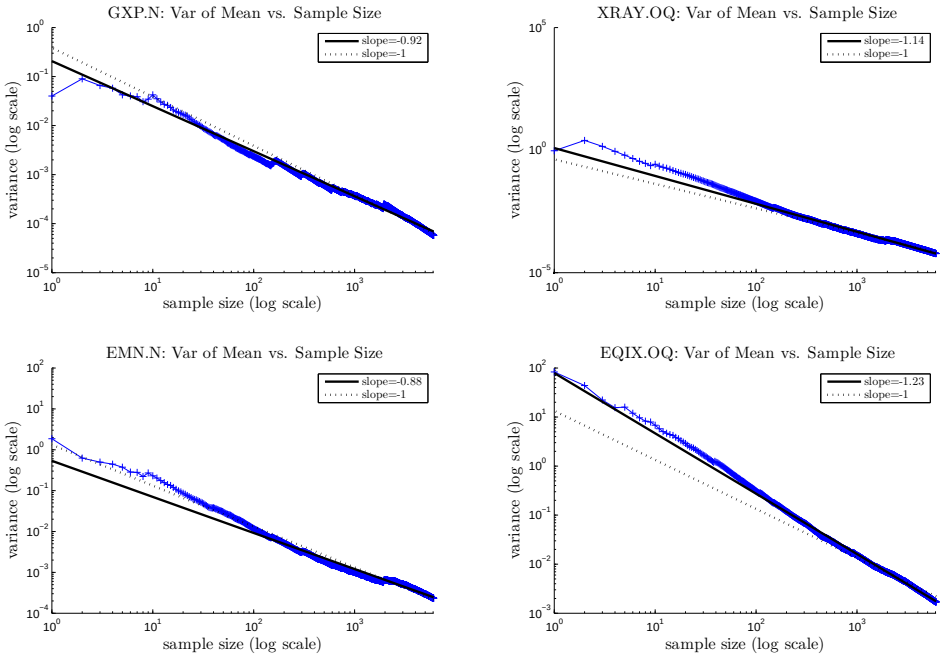

Figure 4: Log-log of variance of mean vs. sample size. The dotted line has slope -1. The thick line is the regression line through the variance of the mean per sample size

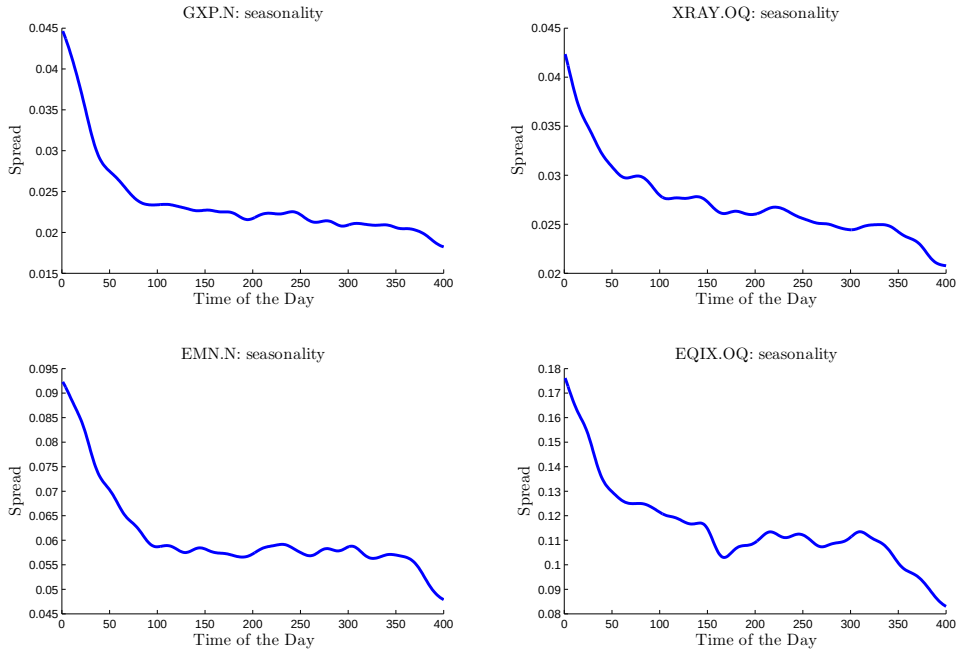

Figure 5: Intraday Periodicities for the 30s bid-ask spread series in January and February 2008 for the mid-cap stocks. Smoothed via kernel regression

	GXP.N	**XRAY.OQ**	**EMN.N**	**EQIX.OQ**
Estimate	0.455	0.439	0.470	0.303
T-Stat	8.868	8.559	9.148	5.901
P-Value	0.000	0.000	0.000	0.000

Table 2: GPH Test of the spread series

where $I(\omega_\lambda)$ is the periodogram of the time series with sample size T at the frequencies $\omega_\lambda = \frac{2\pi\lambda}{T}$. Following Diebold and Rudebusch (1989) we select $v = \sqrt{T}$. Table 2 shows that the estimates of the long memory parameter b are significant and in the range $(0, 0.5)$. This indicates that the series is covariance stationary but long range dependent.

As reflected by Figure 5, we observe pronounced intraday periodicities in line with Chan et al. (1995) as well as Chung et al. (1999). Bid-ask spreads are increased in the beginning of a trading day and decline in the course of the trading session. Such a pattern is explained by a higher adverse selection component in spreads due to the processing of overnight information in the morning.

3 An Econometric Model for Bid-Ask Spread Dynamics

Traditional approaches for time series of count variables are parameter-driven models based on Zeger (1988), Bayesian count data models in the spirit of Harvey and Fernandez (1989), Hidden Markov models (see MacDonald and Zucchini (1997)) or integer autoregressive moving average (INARMA) models (see McKenzie (2003)). Though conceptually elegant, these approaches suffer from tedious estimation procedures which makes them intractable in extensive high-frequency applications. As an alternative, the more recent literature focusses rather on observation-driven models, such as the autoregressive conditional Poisson (ACP) model as introduced by Rydberg and Shephard (1999) and put forward by Heinen (2003), Fokianos et al. (2009) and Ferland et al. (2006). In contrast to the aforementioned models, ACP models are straightforward to estimate and tractable even for a large number of observations. Moreover, in contrast to Hidden Markov Models or the Autoregressive Multinomial Model proposed by Engle and Russell (2005), the ACP does not require to specify the states of the dependent variable prior to the estimation.

3.1 The Autoregressive Conditional Poisson model

Since bid-ask spreads are strictly positive but the Poisson distributions (and extensions thereof) are defined on $\mathbb{N} \cup \{0\}$, we follow Rydberg and Shephard (2003) and shift the spread distributions without loss of generality by one tick to the left. Accordingly, the spread process is re-defined as $S_t \coloneqq \{(\text{spread in number of ticks}) - 1\}$.

Let $\mathcal{P}(\lambda_t)$ denote the Poisson distribution with mean λ_t and let $\mathcal{F}_t$ denote the information available in t. Moreover, let two polynomials α and β be given as $\alpha(B) := \alpha_1 B + \alpha_2 B^2 + ... + \alpha_q B^q$ and $\beta(B) := \beta_1 B + \beta_2 B^2 + ... + \beta_p B^p$, where B is the backshift operator and $\alpha_i > 0$, $i = 1, .., q$, as well as $\beta_i > 0$, $i = 1, .., p$. Then, the autoregressive conditional Poisson process $\{S_t\}_{t \in \mathbb{Z}}$ is given by

$$\begin{aligned} S_t | \mathcal{F}_{t-1} &\sim \mathcal{P}(\lambda_t), \quad \forall\, t \in \mathbb{Z}, \\ \lambda_t &= c + \alpha(B) S_t + \beta(B) \lambda_t, \end{aligned} \tag{3}$$

where $c > 0$. Under (3) the conditional probability mass function of $S_t = s$, $s = 0, 1, 2, ..$

is

$$\mathbb{P}(S_t = s|\lambda_t) = \frac{\lambda_t^s}{s!}e^{-\lambda_t}. \tag{4}$$

As shown by Ferland et al. (2006), the series $\{S_t\}$ is covariance stationary as well as strictly stationary if $\sum_{i=1}^{q}\alpha_i+\sum_{j=1}^{p}\beta_j < 1$. Fokianos et al. (2009) derive the ergodicity conditions for a covariance stationary ACP process in the case $q = p = 1$.

Rearranging (3), the ACP mean equation can be rewritten as an ARMA$(\max\{p, q\}, p)$ specification of the form

$$(1 - \phi(B))\left(S_t - \frac{c}{(1 - \phi(1))}\right) = (1 - \beta(B))\nu_t, \tag{5}$$

with $\phi(B) := \alpha(B)+\beta(B)$ and $\nu_t := S_t - \lambda_t$ being a martingale difference sequence. As discussed by Ferland et al. (2006), the autocorrelation functions of the representations (5) and (3) are identical which makes it appropriate to interpret (3) as a conditional mean rather than a conditional variance model[3].

While the mean and variance of the ACP process (3) are assumed to be conditionally equal, the variance of the ACP process is unconditionally greater or equal to the unconditional mean. As shown by Heinen (2003), in case of $q = p = 1$ we have

$$\mathbb{E}[S_t] = \frac{c}{1 - (\alpha_1 + \beta_1)}, \qquad \mathrm{Var}[S_t] = \frac{\mathbb{E}[S_t](1 - (\alpha_1 + \beta_1)^2 + \alpha_1^2)}{1 - (\alpha_1 + \beta_1)^2} \geq \mathbb{E}[S_t]. \tag{6}$$

Hence, the ACP specification can generate unconditional overdispersion and the marginal distribution of S_t is no longer Poisson.

To account for the possibility of both conditional as well as unconditional overdispersion and underdispersion, Heinen and Rengifo (2007) propose using the Double Poisson distribution proposed by Efron (1986) and defined by

$$\mathbb{P}(S_t = s|\lambda_t, \gamma) = c(\gamma, \lambda_t) \cdot \gamma^{1/2} e^{-\gamma\lambda_t}\left(\frac{e^{-s}s^s}{s!}\right)\left(\frac{e\lambda_t}{s}\right)^{\gamma s}, \quad \text{s=0,1,2,...,} \tag{7}$$

where the constants $c(\gamma, \lambda_t)$ can be numerically approximated by

$$\frac{1}{c(\gamma, \lambda_t)} = 1 + \frac{1 - \gamma}{12\lambda_t\gamma}\left(1 + \frac{1}{\lambda_t\gamma}\right). \tag{8}$$

[3]Recall, that under the Poisson assumption λ_t equals both the conditional mean and the conditional variance.

The Double Poisson distribution nests the Poisson for $\gamma = 1$.

Accordingly, a so-called Autoregressive Conditional Double Poisson (ACDP) model is given by

$$S_t|\mathcal{F}_{t-1} \sim \mathcal{DP}(\lambda_t, \gamma), \quad \forall t \in \mathbb{Z}, \tag{9}$$

where $\mathcal{DP}$ denotes the Double Poisson distribution and λ_t is parameterized as in (3). The conditional variance of the ACDP model is given by $\text{Var}[S_t|\mathcal{F}_{t-1}] = \lambda_t/\gamma$ with $\gamma > 1$ ($\gamma < 1$) reflecting conditional underdispersion (overdispersion). As shown by Heinen (2003), the Double Poisson assumption can overcompensate the overdispersion generated by the dynamic mean specification resulting in unconditional underdispersion. In particular, in the case $p = q = 1$, the ACDP generates unconditional underdispersion if $\gamma > \frac{1-(\alpha_1+\beta_1)^2+\alpha_1}{1-(\alpha_1+\beta_1)}$.[4]

Explanatory variables are easily included in the given setting using an exponential link function. Let x_t denote a k-dimensional vector of covariates without a constant and let γ denote the corresponding parameter vector. Then, the the conditional mean is re-defined as $\mathbb{E}[S_t|\mathcal{F}_{t-1}] \coloneqq \lambda_t \exp(x_t'\gamma)$, with λ_t given by (3).

AC(D)P models are straightforwardly estimated by maximum likelihood. In case of an ACDP specification the log likelihood function is given by

$$\ln \mathcal{L}(\cdot, \gamma) = \sum_{t=1}^{T} \left(\frac{1}{2} \ln(\gamma) - \gamma\lambda_t + S_t(\ln(S_t) - 1) - \ln(S_t!) + \gamma S_t \left(1 + \ln\left(\frac{\lambda_t}{S_t} \right) \right) \right), \tag{10}$$

where the constants $c(\gamma, \lambda_t)$ are omitted.

3.2 Long Memory Autoregressive Conditional Poisson Models

To account for long range dependence in spread series, we propose two types of long memory autoregressive conditional Poisson models. Both specifications capture hyperbolically decaying autocorrelation functions and are motivated by recent advances in long memory volatility models.

A building block of a long memory model is the fractional differencing operator $(1 - B)^d$ (see Hosking (1981)) which is a polynomial defined in terms of the hypergeometric

[4] An alternative generalization of the Poisson distribution is the Negative Binomial distribution. However, as it can only account for overdispersion in the data, it is less flexible.

function F and can be serially expanded according to

$$(1-B)^d = F(-d,1,1;B) = \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)} B^j. \tag{11}$$

The two types of models considered below differ in the way how $(1-B)^d$ enters the conditional mean specification which has strong implications for the existence of first and second unconditional moments. Practically they differ by providing different forecasts which is analyzed in depth in Section 5.

1) LMACP Type I

The so-called Long Memory ACP (LMACP) type I specification is obtained by augmenting the ARMA representation (5) by $(1-B)^d$ resulting in

$$(1-B)^d(1-\phi(B))(S_t-\omega) = (1-\beta(B))\nu_t, \tag{12}$$

where $\omega \in \mathbb{R}_0^+$. The polynomials ϕ and β are defined as in (5) with the roots of $(1-\phi(B))$ and $(1-\beta(B))$ lying outside the unit circle. In the GARCH case, a corresponding specification has been proposed by Karanasos et al. (2004) based on an ARMA representation of GARCH processes and is closely related to the models by Zaffaroni (2004), Koulikov (2003) and Giraitis et al. (2004).

Expressing (12) in terms of λ_t, a LMACP type I process is obtained by

$$\begin{aligned} &S_t|\mathcal{F}_{t-1} \sim \mathcal{P}(\lambda_t), \quad \forall t \in \mathbb{Z}, \\ &\lambda_t = \frac{(1-\phi(B))(1-B)^d}{(1-\beta(B))}\omega + \Psi(B)S_t = \Psi(B)S_t, \end{aligned} \tag{13}$$

where the polynomial Ψ is given as

$$\Psi(B) \coloneqq 1 - \frac{(1-\phi(B))(1-B)^d}{(1-\beta(B))} = \sum_{i=1}^{\infty} \psi_i B^i \tag{14}$$

and $\frac{(1-\phi(B))(1-B)^d}{(1-\beta(B))}\omega = 0$. Accordingly, the long memory autoregressive conditional Double Poisson (LMACDP) model is given by $S_t|\mathcal{F}_{t-1} \sim \mathcal{DP}(\lambda_t,\gamma)$, $\forall t \in \mathbb{Z}$ with $\lambda_t = \Psi(B)S_t$ and nests the Poisson case for $\gamma = 1$.

Based on the representation

$$\lambda_t = \omega + (\Omega(B) - 1)\,\nu_t, \tag{15}$$

$$\Omega(B) \coloneqq (1 - \Psi(B))^{-1} = \sum_{i=0}^{\infty} \omega_i B^i, \tag{16}$$

we observe that ω corresponds to the unconditional mean and is finite.

Conditions for the non-negativity of the conditional mean λ_t are identical to those for long memory GARCH or fractionally integrated GARCH (FIGARCH) processes (Baillie et al. (1996)). In the case $0 < \beta_1 < 1$ and $p = q = 1$ we have:

Proposition 1. *Let $f_i = \frac{i-1-d}{i}$ for $i = 1, 2, ..$ and let $\phi_1 = \alpha_1 + \beta_1$. The conditional mean of the LMACDP with order $p = q = 1$ is nonnegative a.s. if $0 < \beta_1 < 1$ and either $\psi_1 \geq 0$ and $\phi_1 \leq f_2$ or for $k > 2$ with $f_{k-1} < \phi_1 \geq f_k$ it holds that $\psi_{k-1} \geq 0$.*

Proof. See the proof in Conrad and Haag (2006) for FIGARCH models which directly applies to the mean specification of the long memory (Double) Poisson. □

The next proposition establishes the unconditional variance $\mathrm{Var}[S_t]$. Notably, in contrast to the fourth moment of long memory GARCH models the second moment of S_t exists without imposing additional conditions on the process S_t.

Proposition 2. *The unconditional variance of the long memory autoregressive conditional Double Poisson model type I is given by*

$$\mathrm{Var}[S_t] = \frac{1}{\gamma}\mathbb{E}[\lambda_t] \sum_{j=0}^{\infty} \omega_j^2 < \infty. \tag{17}$$

Proof. See the appendix. □

Accordingly, in the LMACP model ($\gamma = 1$), the dynamic specification induces unconditional overdispersion since $\sum_{j=0}^{\infty} \omega_j^2 \geq 1$ and thus $\mathrm{Var}[S_t] = \mathbb{E}[\lambda_t] \sum_{j=0}^{\infty} \omega_j^2 \geq \mathbb{E}[\lambda_t]$.[5] The overdispersion due to the dynamic specification is dependent on the parameter d via $\omega_j \approx Cj^{d-1}$. However, the parameter γ of the Double Poisson distribution in

[5]Details on why $\sum_{j=0}^{\infty} \omega_j^2 \geq 1$ can be found in the technical appendix.

the LMACDP model can generate underdispersion in case of $\gamma > \sum_{j=0}^{\infty} \omega_j^2$. Likewise overdispersion is captured if $\gamma < \sum_{j=0}^{\infty} \omega_j^2$.

Finally, from the representation $S_t = \omega + \Omega(B)\nu_t$ we straightforwardly get the n-order autocorrelation as

$$\rho_n(S_t) = \frac{\sum_{j=0}^{\infty} \omega_j \omega_{j+n}}{\sum_{j=0}^{\infty} \omega_j^2}. \tag{18}$$

Consequently, the LMACDP type I model is covariance stationary for $0 < d < 0.5$.[6] Using $\omega_j \approx Cj^{d-1}$ for high j we have $\rho_n(S_t) \approx C^* n^{2d-1}$ which in turn implies that $\lim_{n\to\infty} \sum_{k=0}^{n} |\rho_k(S_t)|$ is divergent.

2) LMACP Type II

The LMACP type II model is motivated by the specification of a FIGARCH model for volatility processes and the fractionally integrated autoregressive conditional duration (FIACD) model proposed by Jasiak (1998). It builds on the following representation of the conditional mean,

$$(1 - \phi(B))(1 - B)^d S_t = \omega + (1 - \beta(B))\nu_t, \tag{19}$$

where $\omega \in \mathbb{R}_0^+$. Rearranging, the LMACP type II is then defined as

$$\begin{aligned} &S_t|\mathcal{F}_{t-1} \sim \mathcal{P}(\lambda_t), \quad \forall t \in \mathbb{Z}, \\ &\lambda_t = \frac{\omega}{(1 - \beta(B))} + \Psi(B)S_t, \end{aligned} \tag{20}$$

where Ψ is the polynomial (14). For the LMACDP type II we correspondingly change the conditional distribution assumption to $S_t|\mathcal{F}_{t-1} \sim \mathcal{DP}(\lambda_t, \gamma)$. The non-negativity of the conditional mean is guaranteed as long as the conditions of Proposition 1 are fulfilled (see Conrad and Haag (2006)).

The major difference to the type I specification is that the unconditional mean of the type II model is not finite since for $d < 0.5$ the coefficients of the power expansion of $(1-B)^{-d}$ for $B = 1$ are not summable. This result is analogous to the difference between FIGARCH processes and long memory GARCH specifications proposed by Karanasos

[6] Note that in the case $0.5 \leq d \leq 1$ $\lim_{k\to\infty} \sum_{i=0}^{k} \omega_i^2$ does not converge and thus the LMACDP is no longer covariance stationary.

et al. (2004). Hence, the expectation $\mathbb{E}[S_t] = \mathbb{E}[\Gamma(1)\omega + \Gamma(B)(1-\beta(B))\nu_t] = \mathbb{E}[\Gamma(1)\omega]$ is not defined for $\Gamma(B) \coloneqq (1-\phi(B))^{-1}(1-B)^{-d}$. The type II model is thus not covariance stationary and the long memory condition (1) cannot be directly verified since second moments do not exist. [7]

However, computing impulse response functions we show that the model can still capture long range dependence. The impulse response function of the LMACDP type II is defined in terms of the sequence δ_k, $k = 0, 1, ..,$

$$\delta_k \coloneqq \frac{\partial \mathbb{E}[S_{t+k}|\mathcal{F}_t]}{\partial \nu_t} - \frac{\partial \mathbb{E}[S_{t+k-1}|\mathcal{F}_t]}{\partial \nu_t}. \tag{21}$$

Then, the cumulative impulse response function is given by $\lambda_k \coloneqq \sum_{l=0}^{k} \delta_l$, $k = 0, 1, ...,$ where δ_k can be derived from the first difference in S_t,

$$(1-B)S_t = \frac{\omega}{(1-\phi(B))(1-B)^{d-1}} + \underbrace{(1-B)\Omega(B)}_{=:\Delta(B)}\nu_t, \tag{22}$$

with $\Delta(B) := \sum_{j=0}^{\infty} \delta_j B^j$. The impulse response weights can also be recovered from the cumulative impulse responses by $\Delta(B) = (1-B)\Lambda(B)$, where $\Lambda(B) := \sum_{k=0}^{\infty} \lambda_k B^k$. From (22) we deduce that $\Lambda(B) = \Omega(B)$. Thus, in the long run, shocks to the mean die out because $\Delta(1) = 0$, i.e.,

$$\lim_{k\to\infty} \sum_{l=0}^{k} \delta_l = \lim_{k\to\infty} \lambda_k = \Delta(1) = 0. \tag{23}$$

The shocks exhibit a slow, hyperbolic decay rate dependent on the parameter d since $\lambda_k = \omega_k \approx Ck^{d-1}$ for high k. Since this behavior is present also for $0.5 \leq d \leq 1$, we relax the restriction implied by the type I model and require $0 < d \leq 1$ in the type II specification.

[7] While Baillie et al. (1996) argue that the strict stationarity holds for FIGARCH models based on the results by Bougerol and Picard (1992), a similar argument does not hold for the LMACDP type II model as the conditional mean cannot be factored out from the Poisson distribution.

4 Forecasting Bid-Ask Spreads

4.1 Computation of Forecasts

We evaluate out-of-sample forecasts based on a rolling window setup where the underlying model is re-estimated every 10 minutes to quickly adapt to potential changes in parameters. In particular, we conduct the following steps for the Jan./Feb. 2008 sample:

(i) *(Estimation)* Estimate the econometric model based on an estimation window corresponding to five trading days of 30 second spread data.

(ii) *(Forecasting)* Using the parameter estimates from (i), derive successive one-step ahead forecasts for the 10 minute horizon ahead of the estimation window.

(iii) *(Rolling forward the windows)* Move estimation and forecast window forward 10 minutes.

The models are estimated based on truncations of the (infinite) power expansion of $(1-B)^d$. Pre-estimation analysis shows that a truncation point of 250 observations is sufficient to obtain reliable estimates which are widely independent of the truncation.

We evaluate point forecasts, $S_{t+1|t}$, and directional forecasts, $D_{t+1|t}$, defined as

$$S_{t+1|t} \coloneqq \left[\widehat{\lambda}_{t+1}\right], \tag{24}$$

$$D_{t+1|t} \coloneqq \mathbb{1}_{\{S_{t+1|t}>S_t\}} - \mathbb{1}_{\{S_{t+1|t}<S_t\}}, \tag{25}$$

where $\widehat{\lambda}_{t+1}$ is the mean forecast for $t+1$ based on the conditional mean specification and $[\cdot]$ rounds its argument to the nearest integer. Hence, $D_{t+1|t} \in \{-1, 0, 1\}$ if spreads increase, are constant and decrease, respectively.

4.2 Additional Predictors based on Market Microstructure Theory

Decomposing the bid-ask spread into its components, Glosten and Harris (1988), George et al. (1991), Huang and Stoll (1997) and Bollen et al. (2004), among others, identify adverse selection and order processing costs as the main factors driving the spread. The adverse selection component of spreads is highly related to the amount of information

asymmetry in the market. To capture states of high uncertainty and imbalances in the market, we include the realized volatility, given as the sum of squared mid-quote returns over each 30s interval and, alternatively, the absolute 30s mid-quote returns. Moreover, we construct measures of relative trade imbalance and relative depth imbalance to account for asymmetries in trading. The relative trade imbalance is given as

$$Timb_t \coloneqq \frac{|\sum_{\tau=t_1}^{t_m} V_\tau \cdot \mathbb{1}_{\{q_\tau=-1\}} - \sum_{\tau=t_1}^{t_m} V_\tau \cdot \mathbb{1}_{\{q_\tau=1\}}|}{\sum_{\tau=t_1}^{t_m} V_\tau}, \tag{26}$$

where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function and V_1, V_2,...,V_m are the trade sizes corresponding to the time points t_1 to t_m of a 30 second interval. The trade indicator q_t classifies trades into buys (+1) and sells (-1) according to the Lee and Ready (1991) algorithm. The relative depth imbalance is defined as

$$DPimb_t \coloneqq \frac{|Adp_t - Bdp_t|}{Adp_t + Bdp_t}, \tag{27}$$

where Adp_t denotes the best ask depth and Bdp_t the best bid depth. As additional predictors we include the overall depth, given as the sum of the order book depth at best bid and ask level as well as the 30s cumulative trading volume serving as proxies for possible adverse selection in the market.

Finally, intraday periodicities in spreads are captured by a flexible Fourier form as proposed by Gallant (1981),

$$s(t) = \delta^s \bar{t} + \sum_{j=1}^{Q} (\delta_{1,j}^s \cos(\bar{t} 2\pi j) + \delta_{2,j}^s \sin(\bar{t} 2\pi j)), \tag{28}$$

where $\delta^s, \delta_{1,j}^s$ and $\delta_{2,j}^s$ are parameters and $\bar{t} \in [0,1]$ is the normalized intraday time defined as the time elapsed from the beginning of a trading day until observation t, divided by the length of the trading day.

In addition to the static inclusion of covariates, we alternatively conduct the following adaptive selection of the covariates in each step to allow for possible structural changes (see Blaskowitz and Herwartz (2009) for a related setup):

(i) Estimate an AR model for spreads with all covariates based on observations within the estimation window. The AR setup is chosen here because (least squares) estimates can be computed in closed form which significantly reduces

the computation burden in the rolling window framework.

(ii) Discard predictors which are insignificant according to heteroscedasticity and autocorrelation consistent standard errors and execute steps (i) to (iii) from the scheme in 4.1 using only the remaining relevant predictors.

4.3 Forecast Benchmarks

To benchmark our approach, we compute forecasts using the following competing models:

(i) A random walk model ("naïve" forecast) given by $S_t = S_{t-1} + \varepsilon_t$, where ε_t is white noise.

(ii) An exponentially weighted moving average (EWMA) given by

$$S_{t+1} = \gamma_0 S_t + \gamma_1 S_{t-1} + \gamma_2 S_{t-2} + \cdots , \tag{29}$$

where the weights are computed according to $\gamma_i = \alpha(1-\alpha)^i$, $0 < \alpha < 1$ and the smoothing coefficient α is selected as the value minimizing the mean squared prediction error of one-step ahead forecasts.

(iii) An ARMA(p,q) model for S_t, defined by the equation

$$(1 - \alpha(B))(S_t - c) = (1 - \beta(B))\varepsilon_t, \qquad t \in \mathbb{Z}, \tag{30}$$

where α and β are lag polynomials as defined above and the errors ε_t are assumed to be normally distributed.

Moreover, we consider the ARFIMA (p,d,q) model put forward by Granger and Joyeaux (1980), Granger (1981) and Hosking (1981), given by

$$(1 - \alpha(B))(1 - B)^d(S_t - c) = (1 - \beta(B))\varepsilon_t, \qquad t \in \mathbb{Z},\ c \in \mathbb{R}. \tag{31}$$

(iv) The autoregressive conditional duration (ACD) model introduced by Engle and Russell (1998) and Engle (2000), which is the workhorse to capture serially dependent positive-valued random variables, given by $S_t = \mu_t \cdot \varepsilon_t$ for $t \in \mathbb{Z}$ with conditional mean μ_t,

$$\mu_t = \omega + \alpha(B)S_t + \beta(B)\mu_t, \quad \omega > 0. \tag{32}$$

The errors are assumed to be Weibull distributed, $\varepsilon_t|\mathcal{F}_{t-1} \sim \mathcal{W}(\mu_t, \gamma)$, with parameter γ.

Accordingly, the FIACD proposed by Jasiak (1998) is given by $S_t = \mu_t \cdot \varepsilon_t$ with $\varepsilon_t|\mathcal{F}_{t-1} \sim \mathcal{W}(\mu_t, \gamma)$ and

$$(1 - \phi(B))(1 - L)^d S_t = \omega + (1 - \beta(B))\nu_t, \quad \omega > 0, \tag{33}$$

where $\phi(B) \coloneqq \alpha(B) + \beta(B)$ and $\nu_t \coloneqq S_t - \mu_t$ is a martingale difference.

4.4 Point Forecast and Directional Forecast Evaluation

Let $\varepsilon^i_{t+1|t} \coloneqq S^i_{t+1|t} - S^i_{t+1}$ be the forecast error of model $i \in \{1, 2\}$. To assess the basic forecast performance we report the root mean squared error (RMSE) of a series of forecast errors $\varepsilon^i_{t+1|t}$, $t = 1, .., T$. The predictive accuracy of competing forecast models, $i = 1, 2$, is tested using the test of Diebold and Mariano (1995) (DM), based on the loss differential

$$d_t \coloneqq \left(\varepsilon^1_{t+1|t}\right)^2 - \left(\varepsilon^2_{t+1|t}\right)^2. \tag{34}$$

To test for differences in forecast performances, we test the null $H_0 : \mathbb{E}[d_t] = 0$. In the case of one-step ahead forecasts, the DM test statistic takes the form $DM := \bar{d}/\sqrt{\widehat{\mathrm{Var}}(\bar{d})}$, where $\bar{d}$ is the average of the d_t.

A minor modification of the DM test is necessary if nested models are compared. This is the case when we augment the models by additional predictors which inflate the RMSE due to additional estimation errors. Clark and West (2007) propose a test that explicitly accounts for the nested model structure. Let model 2 nest model 1 and let

$$\bar{a} \coloneqq \frac{1}{T}\sum_{t=1}^{T}\left(S^1_{t+1|t} - S^2_{t+1|t}\right)^2. \tag{35}$$

A suitable test statistic of the null is then given by

$$CW \coloneqq \frac{\bar{d} - \bar{a}}{\sqrt{\widehat{\mathrm{Var}}\left(\bar{d} - \bar{a}\right)}}. \tag{36}$$

As the distribution of CW is non-standard, simulated critical values based on Clark and McCracken (2001) have to be used.

Note that Diebold-Mariano and Clark-West type tests can only compare two models. According to White (2000), a problem of such a sequential testing of competing models is that standard p-values may become invalid because of a possible spurious selection of the best model due to data snooping. Hansen (2005) proposes a test for superior predictive accuracy (SPA) that can account for this problem. Let $i = 1$ denote the benchmark model and let $d_t^i \coloneqq \left(\varepsilon_{t+1|t}^1\right)^2 - \left(\varepsilon_{t+1|t}^i\right)^2$ be the loss differential to the rival model $i \in \{2, 3, .., m\}$. The null hypothesis of the SPA test is

$$H_0 : \ \mathbb{E}[d_t^i] \leq 0 \qquad \forall i \in \{2, .., m\}. \tag{37}$$

Hence, H_0 is rejected whenever at least one of the competing models generates significantly better forecasts. The null can be tested based on the statistic

$$SPA \coloneqq \max \left\{ \max_i \left\{ \widehat{\mathrm{Var}}(\bar{d}^i)^{-1/2} \bar{d}^i \right\}, 0 \right\}, \tag{38}$$

where $\widehat{\mathrm{Var}}(\bar{d}^i)$ denotes the estimated variance of $\bar{d}^i$. The distribution of the SPA statistic has to be bootstrapped since the real distribution is nonstandard. Details of the stationary bootstrap procedure can be found in Hansen (2005).

In case of directional forecasts, the tests outlined above are straightforwardly applied based on the directional forecast errors $\varepsilon_{t+1|t} \coloneqq D_{t+1|t} - D_{t+1}$, where $D_{t+1} \coloneqq \mathbb{1}_{\{S_{t+1} > S_t\}} - \mathbb{1}_{\{S_{t+1} < S_t\}}$ is the realized direction of the spread movement. In addition to the SPA, DM and Clark-West tests for the squared directional error series, we report the directional accuracy of the forecasts which is given as

$$DA \coloneqq \frac{\#\{t \ : \ D_{t+1} = D_{t+1|t}\}}{T}. \tag{39}$$

5 Results

5.1 Estimation Results and RMSE Performance

Model selection for all models is conducted by minimizing the RMSEs for one-step ahead forecasts of the January 2008 data. In this respect, we globally identify an ACP(1,1), ARMA(4,2), ACD(1,1) and Double ACP(1,1) as the best performing specifications across the stocks considered. To restrict the computational burden, the

fractionally integrated and long memory specifications are restricted to $p = q = 1$. Diagnostics in terms of Probability Integral Transforms (PIT) based on probability mass function forecasts and autocorrelation functions are given in the web appendix, `http://amor.cms.hu-berlin.de/~grosskla/index.html`. The diagnostics show that the LMACDP type II yields the best model fit in terms of capturing the spread distribution and dynamics.

	GXP.N	**XRAY.OQ**	**EMN.N**	**EQIX.OQ**
	Median of Estimates for the LMACDP(1,d,1)		**type I** | **type II**	*models*
ω	1.414 | 0.175	1.802 | 0.327	5.298 | 1.084	11.098 | 1.562
ϕ_1	0.373 | 0.000	0.330 | 0.000	0.320 | 0.000	0.361 | 0.000
β_1	0.626 | 0.000	0.669 | 0.120	0.680 | 0.000	0.638 | 0.099
d	0.455 | 0.314	0.498 | 0.260	0.500 | 0.251	0.474 | 0.293
γ	1.512 | 1.565	1.484 | 1.541	1.023 | 1.043	0.341 | 0.359
	Median of Estimates for the ACDP(1,1) *model*			
ω	0.120	0.141	0.304	1.257
ϕ_1	0.261	0.196	0.164	0.245
β_1	0.647	0.714	0.777	0.632
d	1.541	1.528	1.041	0.353

Table 3: Median parameter estimates for the ACDP and LMACDP type I (on the left in each column) and II (on the right). The median is taken over the rolling window iterations. All variables significant at the 10% level in 95 % of the iterations. Notation is as in section 2

Table 3 gives the median parameter estimates of the Double ACDP and LMACDP type II models over all estimates of the rolling window setup. Figure 6 shows the evolution of estimates for the fractional integration parameter d of the LMACDP type II model. We observe that the persistence in bid-ask spread clearly varies over time which makes it necessary to allow for parameter changes in a rolling window setup. The evolution of the estimates of the additional regressor coefficients is given in Figures 8 to 13 in the appendix. The signs of the parameter estimates are in line with economic theory. While volatility and trading volume are positively related to bid-ask spreads, the effects of trade and depth imbalances fluctuate around zero. Moreover, depth coefficients are widely negative, reflecting that deep markets are accompanied by low spreads reflecting periods of high liquidity.

The upper panel of Table 4 gives the one-step ahead RMSEs and the directional accuracy for all models without additional predictors. We observe that a Poisson-based

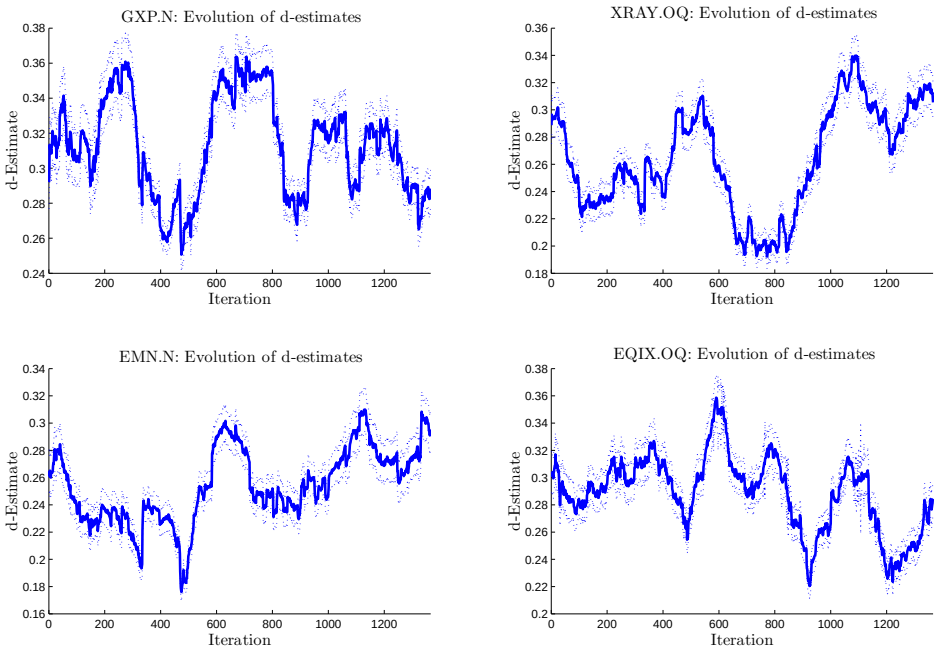

Figure 6: Evolution of the estimates of the fractional integration parameter in the LMACDP type II. 95 % confidence intervals dotted

model always performs best with RMSEs on average 18 % lower than those of the random walk benchmark. Likewise, the directional accuracy of Poisson-based forecasts improves on average by 25 % over the naïve benchmark. Overall, the LMACDP type I and II models outperform FIACD and ARFIMA benchmarks in terms of the RMSE and DA, with differences in the forecast performance being especially high for the directional forecasts. These results indicate that an appropriate distribution as implied by the Double Poisson distribution yields efficiency gains which lead to superior predictions. Moreover, in most cases we find forecasts of LMACDP type II specifications to be (marginally) superior compared to type I specifications.

The lower panel of Table 4 shows the RMSE and DA for the LMACDP type II model augmented by covariates. It turns out that the inclusion of predictors improves both point and direction forecasts. However, including all predictors ("All Predictors") or adaptive selections thereof ("Preselected Pred.") do not necessarily provide better forecasts than including only single predictors. Hence, potential forecast gains by the inclusion of several variables are obviously offset by a higher estimation uncertainty.

RMSE | *DA*	**GXP.N**	**XRAY.OQ**	**EMN.N**	**EQIX.OQ**
	RMSE of Basic Models | *Directional Accuracy of Basic Models*			
Naïve	1.2498 | 0.4756	1.3481 | 0.3740	2.9459 | 0.1950	7.1555 | 0.1444
EWMA	1.2867 | 0.5043	1.2381 | 0.5215	2.5649 | 0.5432	6.7206 | 0.5437
ARMA	1.1667 | 0.5321	1.2122 | 0.5131	2.5500 | 0.5231	6.3682 | 0.5287
ARFIMA	1.1056 | 0.3095	1.1281 | 0.3717	2.3906 | 0.4965	5.9638 | 0.5150
ACD	1.0880 | 0.3192	1.1217 | 0.3752	2.3692 | 0.4984	5.8634 | 0.5209
FIACD	1.0826 | 0.3181	1.1101 | 0.3731	2.3478 | 0.4975	5.8195 | 0.5246
ACP	1.0803 | 0.5516	1.1152 | 0.5369	2.3436 | 0.5430	5.8466 | 0.5400
ACDP	1.0803 | 0.5516	1.1153 | 0.5369	**2.3436** | **0.5433**	5.8466 | 0.5400
LMACDP type I	1.0805 | **0.5625**	1.1153 | 0.5322	2.3440 | 0.5184	5.9553 | **0.5503**
LMACDP type II	**1.0720** | 0.5552	**1.1086** | **0.5389**	2.3448 | 0.5394	**5.8026** | 0.5452
	RMSE | *Directional Accuracy of* ***LMACDP*** *type II model plus additional predictors*			
+All Predictors	**1.0577** | 0.5534	1.1113 | 0.5411	**2.3177** | **0.5458**	5.7996 | **0.5489**
+Preselected Pred.	1.0637 | 0.5549	1.1178 | 0.5380	2.3380 | 0.5416	5.8192 | 0.5484
+Seasonality	1.0678 | 0.5519	**1.1022** | 0.5383	2.3281 | 0.5426	**5.7924** | 0.5461
+Depth	1.0710 | 0.5550	1.1026 | **0.5427**	2.3391 | 0.5420	5.8007 | 0.5443
+Depth Imb.	1.0727 | 0.5553	1.1086 | 0.5383	2.3431 | 0.5392	5.8034 | 0.5437
+Real. Vola	1.0800 | 0.5544	1.1203 | 0.5392	2.3369 | 0.5428	5.8038 | 0.5470
+Absolute Ret.	1.0637 | 0.5546	1.1181 | 0.5390	2.3384 | 0.5407	5.8071 | 0.5483
+Traded Vol.	1.0722 | **0.5556**	1.1098 | 0.5388	2.3405 | 0.5408	5.8036 | 0.5460
+T.Vol. Imb.	1.0725 | **0.5556**	1.1090 | 0.5389	2.3453 | 0.5394	5.8038 | 0.5445

Table 4: Upper panel: RMSE and DA values of models without additional predictors. Best model in terms of RMSE and DA highlighted. Lower panel: Models with additional predictors as defined in section 4.2. All predictors denotes inclusion of all variables. Preselected pred. refers to the preselection scheme outlined in section 4.2. Abbreviations are Imb.for imbalance, Real. Vola. for realized volatility, Ret. for return, Vol. for Volume and T.Vol. Imb. for Trading Volume Imbalance. Seasonality denotes the seasonality component (28) with $Q = 2$. Lowest RMSE and highest DA highlighted

DM and CW statistics	**GXP.N**	**XRAY.OQ**	**EMN.N**	**EQIX.OQ**
DM test for equal forecast performance of Basic Model with lowest RMSE | highest DA				
and naïve model	-19.23 | -7.74	-31.36 | -15.12	-37.71 | 2.58	-31.69 | 14.66
P-Value	(0.00) | (0.00)	(0.00) | (0.00)	(0.00) | (0.00)	(0.00) | (0.00)
and 2nd best Basic model	-3.57 | -3.64	-3.26 | -6.62	0.34 | 5.87	-7.55 | -12.55
P-Value	(0.00) | (0.00)	(0.00) | (0.00)	(0.36) | (0.00)	(0.00) | (0.00)
DM test for equal forecast performance of Model with additional Predictors with lowest RMSE | highest DA				
and 2nd best with add. Predic.	-2.57 | 0.60	-0.18 | -2.72	-2.94 | 1.28	-1.57 | 0.44
P-Value	(0.01) | (0.27)	(0.43) | (0.00)	(0.00) | (0.10)	(0.06) | (0.33)
CW test for equal forecast performance of Model with additional Predictors with lowest RMSE | highest DA				
and the LMACDP type II	2.66* | 0.23	2.24* | 113.57*	4.81* | -27.61*	1.93* | 6.92*

Table 5: Diebold Mariano and Clark-West test results for point and direction forecasts. The $*$ in the last row denotes significance at the 10% level

% P-Values of SPA-Test	**GXP.N**	**XRAY.OQ**	**EMN.N**	**EQIX.OQ**
	SPA test based on squared error series for point forecasts | *direction forecasts*			
Naïve	0.00 | 0.00	0.00 | 0.00	0.00 | 5.25	0.00 | 0.00
EWMA	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00
ARMA	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00
ARFIMA	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00
ACD	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00
FIACD	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00
ACDP	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00	0.00 | 0.00
LMACDP type I	0.00 | 0.75	0.00 | **31.25**	0.00 | 4.50	0.00 | 0.00
LMACDP type II	4.00 | 0.00	0.00 | 18.75	0.00 | 2.50	9.00 | 0.00
LMACDP type II				
+All Pred.	0.75 | 0.00	14.25 | 1.75	0.75 | 2.25	**25.25** | 0.00
+preselected Pred.	1.50 | 0.00	1.50 | 15.75	0.00 | 5.25	3.50 | 0.00
+Seasonality	4.00 | 0.00	**49.00** | 1.00	0.00 | 0.00	23.25 | 0.00
+Depth	**7.00** | 1.25	47.25 | 27.50	0.00 | 4.00	8.25 | 0.00
+Depth Imb.	5.25 | 0.25	0.25 | 13.00	0.00 | 0.50	7.25 | 0.00
+Real Vola.	1.50 | 0.00	1.25 | 11.75	0.00 | 3.00	13.00 | 0.00
+Abs. Return	1.00 | 0.00	7.25 | 2.50	0.00 | **5.50**	21.00 | 0.00
+Traded Vol.	4.00 | 0.00	0.00 | 14.75	0.00 | 1.00	8.75 | 0.00
+T.Vol. Imb.	5.75 | 0.75	0.00 | 18.25	0.00 | 0.25	10.50 | 0.00

Table 6: P-Values of the SPA test for all competing models. Highest value highlighted. Notation for the additional covariates as in Table 4

Table 5 shows Diebold-Mariano and Clark-West tests for forecast comparisons of the best and second best specifications in terms of the RMSE. It turns out that point and directional forecasts of models without predictors are significantly different from each other. Hence, the superior performance of LMACDP models in terms of the RMSE and DA shown in Table 4 is statistically significant. The Clark-West tests indicate that the inclusion of trading characteristics in LMACDP type II models yields a significant improvement of the forecasting power over the basic LMACDP type II. However, among specifications including covariates, differences between squared prediction errors are often insignificant according to the Diebold-Mariano Test. This reflects that most of the covariates carry similar information about the adverse selection costs driving spreads.

To identify the overall best performing model in terms of point and direction forecasts we present the results of the SPA test in Table 6. In three out of the four cases we cannot reject the null at the 10 % level that a LMACDP type II model including covariates provides the best forecast performance. The p-values are widely in accordance with the ordering of models according to the RMSE and DA results tabulated in Table 4.

We summarize the following main findings: First, the efficiency gains implied by (Double) Poisson modeling yield significantly superior forecast results in terms of the RMSE and DA criterion. Second, the forecast performance of the long memory specifications compared to their short-memory counterparts indicate the importance of accounting for the strong persistence in spreads. Third, the inclusion of predictors significantly improves point- and direction forecasts.

5.2 A Trading Schedule based on Spread Forecasts

To evaluate the potential economic gains implied by spread forecasts, we consider quantifying spread costs in trading schemes. The trading schedules are motivated by the fact that transaction costs of large trades can be reduced by splitting orders into smaller trades and are distributed over time. In such strategies, spread forecasts can improve trading algorithms by allowing to intensify trading in periods when spreads are expected to be small.

Suppose a benchmark trading schedule is based on trades occurring at the end

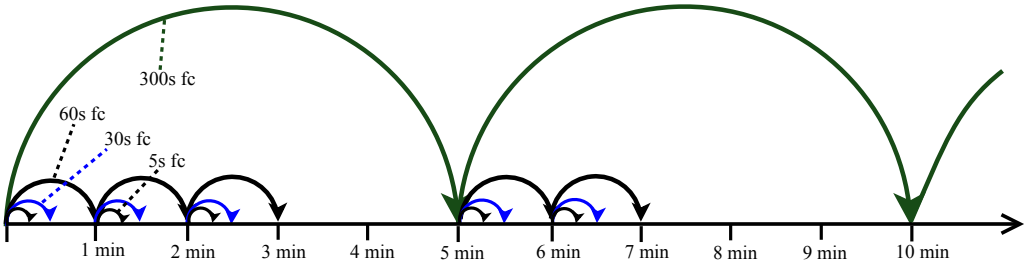

Figure 7: The principle of the trading schedule. 5s fc, 30s fc, 60s fc and 300s fc stand for the forecasts based on different data aggregation frequencies

of each 5 minute interval. In an alternative trading schedule the time of the trades is flexible within each interval and can be chosen in accordance with corresponding spread forecasts. Then, the resulting transaction costs serve as a measure of the implied economic gains.

To obtain a fine grid of spread forecasts within each 5 minute interval, we construct bid-ask spread forecasts on a 5, 30, 60 and 300 second frequency employing the LMACDP type II specification with seasonal component. Let fc_5^x, fc_{30}^x, fc_{60}^x and fc_{300} denote spread forecasts on a 5, 30, 60 and 300 second data aggregation frequency, respectively. Moreover, let $x \in \{1, 2, 3, 4, 5\}$ indicate the corresponding 1 minute subinterval withing the 5 minute interval. Then, the timing of trades in the flexible schedule is chosen as follows.

- Starting from the left interval boundary, we search in the x successive subintervals for the smallest forecast out of $\{fc_5^x, fc_{30}^x, fc_{60}^x, fc_{300}\}$ until we arrive at the right boundary or a minimum is found. Once a minimum is found for $x \in \{1, 2, 3, 4, 5\}$, we do not consider subintervals $x^* > x$ since these are not known by a trader using one-step ahead forecasts on 5s, 30s, 60s and 300s frequencies. In case fc_{30}^x or fc_{60}^x is a new optimum, we optimize the time of trading within the corresponding subinterval using the forecasts on higher frequencies. Figure 7 illustrates the procedure.

- In case of equal forecasts, $\min\{fc_5^x, fc_{30}^x, fc_{60}^x\}=fc^{300}$, we choose the timing of

Schedule	**GXP.N**	**XRAY.OQ**	**EMN.N**	**EQIX.OQ**
Fast	7.74	7.47	8.17	11.60
PMF Info	11.85	10.51	10.58	12.90
Slow	14.42	8.22	10.51	11.90

Table 7: Spread cost savings as percentage of the benchmark schedule when employing forecasts in the schedules. Schedule names refer to the three rules in case of equal forecasts

trades according to one of the following three options.

(i) Motivated by traders' tendency to trade as fast as possible, we choose the time of trading to be closest to the left boundary of the interval and stop searching for further minima. The resulting schedule is labelled "fast".

(ii) We choose the time of trading to be closest to the right boundary and successively optimize the time of trading using forecasts on higher frequencies until the time of the new minimum forecast. The resulting schedule is labelled "slow".

(iii) We use the information from the probability mass function and weight the equal forecasts with the assigned forecasted probabilities $f(S_{t+1|t}, \widehat{\lambda}_{t+1}, \widehat{\gamma})$, where f denotes the Double Poisson probability mass function. Then, we choose the most probable forecast to be the new optimal one and optimize the trading in the subintervals until we arrive at another optimum. This schedule is labelled "PMF info".

After the timing of trades is chosen we sum up the incurred transaction costs (induced by crossings of the market) for the benchmark strategy and the three alternative schedules. Table 7 reports the percentage spread cost savings over the benchmark strategy. We find that the strategy exploiting the information from the pmfs of the Double Poisson assumption yields the highest average gains: spread costs reduce by 11.46% of the costs of the benchmark schedule. The "slow" schedule saves us 11.26% and the "fast" strategy still 8.74%. The better performance of the strategies "slow" and "PMF info" is obviously induced by the use of more optimization steps due to the rule specification in case of equal forecasts.

5.3 Robustness of the Results

Several robustness checks underscore the relevance of our study. First, we can confirm the reported RMSE results for a larger cross-section of stocks from the mid-cap segment of the Russell 3000. Our web appendix `http://amor.cms.hu-berlin.de/~grosskla/index.html` shows descriptive statistics, point forecast and trading schedule evaluations for 46 stocks ordered according to their average spread. We find that the RMSE improvement of the LMACDP type II over the naïve benchmark model is 19.51% on average across stocks with average spread ticks $\in (2, 4)$ and 18.90% for stocks with mean spread ticks greater than 4. Interestingly, forecast gains are also possible in case of nearly constant stocks: RMSE improvements over the naïve model are on average 13.55% for stocks with mean spread ticks < 2.

Second, the spread cost savings from the trading schedules tend to increase with the size of average spreads. Moreover, the specification of a flexible probability mass function in the Poisson models becomes more and more important with increasing spread sizes. While the "PMF info" trading schedule does not generate higher percentage cost savings than the alternative two schedules for stocks with spreads smaller than 4 ticks, the opposite is true when average spreads become larger than 4 ticks. We conclude that the flexible Double Poisson modelling is the more useful the larger the spreads and the more dispersed the spread series.

Third, the results are not only relevant for quoted spreads but also for alternative spread measures like, e.g., effective spreads. Effective spreads are closely related to quoted spreads (see Figure 1) and reveal very similar time series properties.

6 Conclusions

Motivated by the relevance of bid-ask spreads in trading decisions and market microstructure modelling this study is the first one systematically analyzing forecasts of quoted bid-ask spreads. To capture the empirical features of spread time series for Russell 3000 mid cap stocks traded at NYSE and NASDAQ we introduce a novel long memory autoregressive conditional Poisson (LMACP) model. The LMACP can accomodate highly persistent time series of count data and is thus suitable for modeling persistent discrete time series which are often found in high-frequency data applica-

tions.

We find that autoregressive conditional Poisson (ACP) models and their long memory extension well capture the distributional and dynamic properties of quoted bid-ask spreads. Generallizations of the Poisson distribution, such as the Double Poisson distribution, are able to account for both under- and overdispersion found in the data and underscore the good fit of the proposed model. Forecasting bid-ask spreads in a rolling window out-of-sample framework shows that long memory ACP models outperform competing benchmarks like ARFIMA, ARMA, ACD, FIACD and exponential moving average models in terms of the root mean squared error, directional accuracy and density forecasts. Implementing the spread forecasts in a simple trading algorithm we find that spread forecast can induce transaction cost savings of up to 12%.

References

Anand, A., S. Chakravarty, and T. Martell (2005): "Empirical evidence on the evolution of liquidity: Choice of market versus limit orders by informed and uninformed traders," *Journal of Financial Markets*, 8, 289–309.

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2003): "Modeling and forecasting realized volatility," *Econometrica*, 71, 579–625.

Baillie, R. T., T. Bollerslev, and H. O. Mikkelsen (1996): "Fractionally integrated generalized autoregressive conditional heteroscedasticity," *Journal of Econometrics*, 74, 3–30.

Beran, J. (1998): *Statistics for long-memory processes*, Boca Raton, Florida: Chapman and Hall/ CRC.

Bessembinder, H. and K. Venkataraman (2010): "Bid-ask spreads: Measuring trade execution costs in financial markets," in *Encyclopedia of Quantitative Finance*, London: John Wiley and Sons.

Bhardwaj, G. and N. R. Swanson (2006): "An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series," *Journal of Econometrics*, 131, 539–578.

Biais, B., P. Hillion, and C. Spatt (1995): "An empirical analysis of the limit order book and the order flow in the Paris bourse," *Journal of Finance*, 50, 1655–1689.

Blaskowitz, O. and H. Herwartz (2009): "Adaptive Forecasting of the EURIBOR swap term structure," *Journal of Forecasting*, 28, 575–594.

Bollen, N. P. B., T. Smith, and R. E. Whaley (2004): "Modeling the bid/ask spread: Measuring the inventory holding premium," *Journal of Financial Economics*, 9, 97–141.

Bougerol, P. and N. Picard (1992): "Stationarity of GARCH processes and of some nonnegative time series," *Journal of Econometrics*, 52, 115–127.

Brownlees, C. T., F. Cippolini, and G. M. Gallo (2010): "Intra-daily Volume Modeling and Prediction for Algorithmic Trading," *Journal of Financial Econometrics*, 9, 1–30.

Chan, K. C., W. G. Christie, and P. H. Schultz (1995): "Market structure and the intraday pattern of bid-ask spreads for NASDAQ securities," *Journal of Business*, 68, 35–60.

Chung, K. H., B. F. Van Ness, and R. A. Van Ness (1999): "Limit orders and the bid-ask spread," *Journal of Financial Economics*, 53, 255–287.

Clark, T. E. and M. W. McCracken (2001): "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105, 85–110.

Clark, T. E. and K. D. West (2007): "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 138, 291–311.

Conrad, C. and B. R. Haag (2006): "Inequality constraints in the fractionally integrated GARCH model," *Journal of Financial Econometrics*, 4, 413–449.

Copeland, T. E. and D. Galai (1983): "Information effects on the bid/ask spread," *Journal of Finance*, 38, 1457–1469.

Corsi, F. (2009): "A simple approximate long-memory model of realized volatility," *Journal of Financial Econometrics*, 7, 174–196.

Deo, R., M. Hsieh, and C. M. Hurvich (2010): "Long memory in intertrade durations, counts and realized volatility of NYSE stocks," *Journal of Statistical Planning and Inference*, 12, 3715–3733.

Diebold, F. X. and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253–265.

Diebold, F. X. and G. D. Rudebusch (1989): "Long memory and persistence in aggregate output," *Journal of Monetary Economics*, 24, 189–209.

Ding, Z. and C. W. J. Granger (1996): "Modeling volatility persistence of speculative returns: A new approach," *Journal of Econometrics*, 73, 185–215.

Easley, D. and M. O'Hara (1992): "Time and the process of security price adjustment," *Journal of Finance*, 47, 577–605.

Efron, B. (1986): "Double exponential families and their use in generalized linear regression," *Journal of the American Statistical Association*, 81, 709–721.

Engle, R. F. (2000): "The econometrics of ultra-high frequency data," *Econometrica*, 68, 1–22.

Engle, R. F. and A. J. Patton (2004): "Impact of trades in an error-correction model of quote prices," *Journal of Financial Markets*, 7, 1–25.

Engle, R. F. and J. R. Russell (1998): "Autoregressive conditional duration: a new model for irregularly spaced transaction data," *Econometrica*, 66, 1127–1162.

——— (2005): "A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times," *Journal of Business and Economic Statistics*, 23, 166–180.

Ferland, R., A. Latour, and D. Oraichi (2006): "Integer-valued Garch process," *Journal of Time Series Analysis*, 27, 923–942.

Fokianos, K., A. Rahbek, and D. Tjostheim (2009): "Poisson Autoregression," *Journal of the American Statistical Association*, 12, 1430–1439.

Foucault, T. (1999): "Order flow decomposition and trading costs in a dynamic limit order market," *Journal of Financial Markets*, 2, 99–134.

Foucault, T., O. Kahan, and E. Kandel (2005): "Limit order book as a market for liquidity," *Review of Financial Studies*, 18, 1171–1217.

Freeland, R. K. and B. P. M. McCabe (2004): "Forecasting discrete valued low count time series," *International Journal of Forecasting*, 20, 427–434.

Gallant, R. A. (1981): "On the bias in flexible functional forms and an essential unbiased form: The Fourier flexible form," *Journal of Econometrics*, 15, 211–245.

George, T. J., G. Kaul, and M. Nimalendran (1991): "Estimation of the bid-ask spread and its components: A new approach," *Review of Financial Studies*, 4, 623–656.

Geweke, J. and S. Porter-Hudak (1983): "The estimation and application of long memory time series models," *Journal of Time Series*, 4, 221–238.

Giraitis, L., P. M. Robinson, and D. Surgailis (2004): "LARCH, leverage and long memory," *Journal of Financial Econometrics*, 2, 177–210.

Glosten, L. R. (1987): "Components of bid-ask spread and statistical properties of transaction prices," *Journal of Finance*, 42, 1293–1307.

Glosten, L. R. and L. E. Harris (1988): "Estimating the components of the bid/ask spread," *Journal of Financial Economics*, 21, 123–142.

Glosten, L. R. and P. Milgrom (1985): "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of Financial Economics*, 14, 71–100.

Granger, C. W. J. (1981): "Some properties of time series data and their use in econometric model specification," *Journal of Econometrics*, 16, 121–130.

Granger, C. W. J. and R. Joyeaux (1980): "An introduction to long memory time series models and fractional differencing," *Journal of Time Series Analysis*, 1, 15–39.

Griffiths, M., B. Smith, A. Turnbull, and R. White (2000): "The costs and determinants of order aggressiveness," *Journal of Financial Economics*, 56, 65–88.

Hall, A. D. and N. Hautsch (2006): "Order aggressiveness and order book dynamics," *Empirical Economics*, 30, 973–1005.

Hamilton, J. (1994): *Time Series Analysis*, Princeton New Jersey: Princeton University Press.

Hansen, P. R. (2005): "A test for superior predictive ability," *Journal of Business and Economic Statistics*, 23, 365–380.

Harris, L. and J. Hasbrouck (1996): "Market vs. limit orders: The SuperDOT evidence on order submission strategy," *Journal of Financial and Quantitative Analysis*, 31, 213–231.

Harvey, A. and C. Fernandez (1989): "Time series models for count or qualitative observations," *Journal of Business and Economic Statistics*, 7, 407–417.

Hasbrouck, J. (2000): "The dynamics of discrete bid and ask quotes," *Journal of Finance*, 54, 2109–2142.

Hautsch, N. and R. Huang (2010): "The market impact of a limit order," Working Paper 1677343, Social Science Research Network.

Heinen, A. (2003): "Modelling time series count data: An autoregressive conditional poisson model," Working Paper 1117187, Social Science Research Network.

Heinen, A. and E. Rengifo (2007): "Multivariate autoregressive modeling of time series count data using copulas," *Journal of Empirical Finance*, 14, 564–583.

Hosking, J. R. M. (1981): "Fractional differencing," *Biometrika*, 68, 165–76.

Härdle, W., N. Hautsch, and A. Mihoci (2009): "Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics," Working Paper 1475168, Social Science Research Network.

Huang, R. and H. Stoll (1997): "The components of the bid-ask spread: A general approach," *Review of Financial Studies*, 10, 995–1034.

Jasiak, J. (1998): "Persistence in intertrade duration," *Finance*, 19, 166–195.

Jung, R. C., M. Kukuk, and R. Liesenfeld (2006): "Time series of count data: Modeling, estimation and diagnostics," *Computational Statistics & Data Analysis*, 51, 2350–2364.

Karanasos, M., Z. Psaradakis, and M. Sola (2004): "On the autocorrelation properties of long memory GARCH processes," *Journal of Time Series Analysis*, 25, 265–281.

Koulikov, D. (2003): "Modeling sequences of long memory non-negative stationary random variables," Working Paper 331100, Social Science Research Network.

Kyle, A. S. (1985): "Continuous auctions and insider trading," *Econometrica*, 53, 1315–1336.

Lee, C. M. C. and M. J. Ready (1991): "Inferring trade direction from intraday data," *Journal of Finance*, 46, 733–746.

Lux, T. and T. Kaizoji (2007): "Forecasting volatility and volume in the Tokyo stock market: Long memory, fractality and regime switching," *Journal of Economic Dynamics and Control*, 31, 1808–1843.

MacDonald, I. and W. Zucchini (1997): *Hidden Markov and other models for discrete-valued time series*, London: Chapman and Hall.

McKenzie, E. (2003): "Discrete variate time series," in *Handbook of Statistics*, London: Elsevier.

Parlour, C. A. (1998): "Price dynamics in limit order markets," *Review of Financial Studies*, 11, 789–816.

Pascual, R. and D. Veredas (2009): "What pieces of limit order book information matter in explaining order choice by patient and impatient traders?" *Journal of Quantitative Finance*, 5, 527–545.

Ranaldo, A. (2004): "Order aggressiveness in limit order book markets," *Journal of Financial Markets*, 7, 53–74.

Rydberg, T. H. and N. Shephard (1999): "BIN models for trade-by-trade data. Modeling the number of trades in a fixed interval of time," Working Paper 0740, Nuffield College.

——— (2003): "Dynamics of trade-by-trade movements: Decomposition and models," *Journal of Financial Econometrics*, 1, 2–25.

Sutradahar, B. C. (2008): "On forecasting counts," *Journal of Forecasting*, 27, 109–129.

Taylor, N. (2002): "The economic and statistical significance of spread forecasts: Evidence from the London stock exchange," *Journal of Banking and Finance*, 26, 795–818.

White, H. (2000): "A reality check for data snooping," *Econometrica*, 68, 1097–1126.

Zaffaroni, P. (2004): "Stationarity and memory of ARCH(∞) models," *Econometric Theory*, 20, 147–160.

Zeger, S.-L. (1988): "A regression model for time series of counts," *Biometrika*, 75, 621–629.

7 Appendix

7.1 Technical appendix

Proof of proposition 2:

Proposition 2. *The unconditional variance of the long memory autoregressive conditional Double Poisson model type I is given by*

$$\mathrm{Var}[S_t] = \frac{1}{\gamma}\mathbb{E}[\lambda_t]\sum_{j=0}^{\infty}\omega_j^2 < \infty.$$

Proof. We obtain an expression for the unconditional variance of the errors ν_t of the Double Poisson specification from the following steps. We have

$$\begin{aligned}\mathbb{E}[\nu_t^2] &= \mathbb{E}[(S_t-\lambda_t)^2] = \mathbb{E}[S_t^2] - 2\mathbb{E}[\mathbb{E}[S_t\lambda_t|\mathcal{F}_{t-1}]] + \mathbb{E}[\lambda_t^2] \\ &= \mathbb{E}[S_t^2] - \mathbb{E}[\lambda_t^2],\end{aligned} \tag{40}$$

since λ_t depends only on past values of λ_t and S_t, and

$$\begin{aligned}\mathrm{Var}[S_t] &= \mathbb{E}[S_t^2] - \mathbb{E}[S_t]^2 \\ &= \mathbb{E}[\underbrace{\mathrm{Var}[S_t|\mathcal{F}_{t-1}]}_{=\frac{\lambda_t}{\gamma}}] + \mathrm{Var}[\underbrace{\mathbb{E}[S_t|\mathcal{F}_{t-1}]}_{=\lambda_t}] = \frac{1}{\gamma}\mathbb{E}[\lambda_t] + \mathbb{E}[\lambda_t^2] - \mathbb{E}[\lambda_t]^2.\end{aligned} \tag{41}$$

Solving (41) for $\mathbb{E}[S_t^2]$ and substituting into (40) we get

$$\begin{aligned}\mathbb{E}[\nu_t^2] &= \mathbb{E}[S_t]^2 + \frac{1}{\gamma}\mathbb{E}[\lambda_t] - \mathbb{E}[\lambda_t]^2 + \mathbb{E}[\lambda_t^2] - \mathbb{E}[\lambda_t^2] \\ &= \mathbb{E}[\underbrace{\mathbb{E}[S_t|\mathcal{F}_{t-1}]}_{=\lambda_t}]^2 + \frac{1}{\gamma}\mathbb{E}[\lambda_t] - \mathbb{E}[\lambda_t]^2 + \mathbb{E}[\lambda_t^2] - \mathbb{E}[\lambda_t^2] = \frac{1}{\gamma}\mathbb{E}[\lambda_t].\end{aligned}$$

From the infinite moving average representation

$$S_t = \omega + \Omega(B)\nu_t \tag{42}$$

we obtain using $\text{Cov}(\nu_t, \nu_{t-1}) = 0$

$$\text{Var}[S_t] = \sum_{j=0}^{\infty} \omega_j^2 \text{Var}[\nu_t] = \frac{1}{\gamma} \mathbb{E}[\lambda_t] \sum_{j=0}^{\infty} \omega_j^2.$$

Furthermore, applying Stirling's formula we obtain $\omega_j \approx C j^{d-1}$ for high j (see Hosking (1981)), where C is a positive constant, such that $\lim_{k\to\infty} \sum_{i=0}^{k} \omega_i^2$ converges for $0 < d < 0.5$. □

The following refers to the footnote on page 15.

Footnote 5. *The sum of the squared coefficients of* $\Omega(B)$ *is greater or equal than* 1*,*

$$\sum_{j=0}^{\infty} \omega_j^2 \geq 1.$$

Proof. We have

$$(1 - B)^{-d} = 1 + \delta_1 B + \delta_2 B^2 + ...$$

(see Hosking (1981)) and

$$(1 - \alpha(B))^{-1} = 1 + a_1 B + a_2 B^2 + ...$$

(see, e.g., Hamilton (1994)) such that the first coefficient of $\Omega(B)$ is $\omega_0 = 1$,

$$\begin{aligned} \Omega(B) &= \omega_0 + \omega_1 B + \omega_2 B^2 + ... = (1 - \beta(B))(1 - \alpha(B))^{-1}(1 - B)^{-d} \\ &= 1 + \omega_1 B + \omega_2 B^2 + ... \end{aligned}$$

and hence

$$1 \leq 1 + \sum_{j=1}^{\infty} \omega_j^2 = \sum_{j=0}^{\infty} \omega_j^2.$$

□

7.2 Evolution of Coefficient Estimates

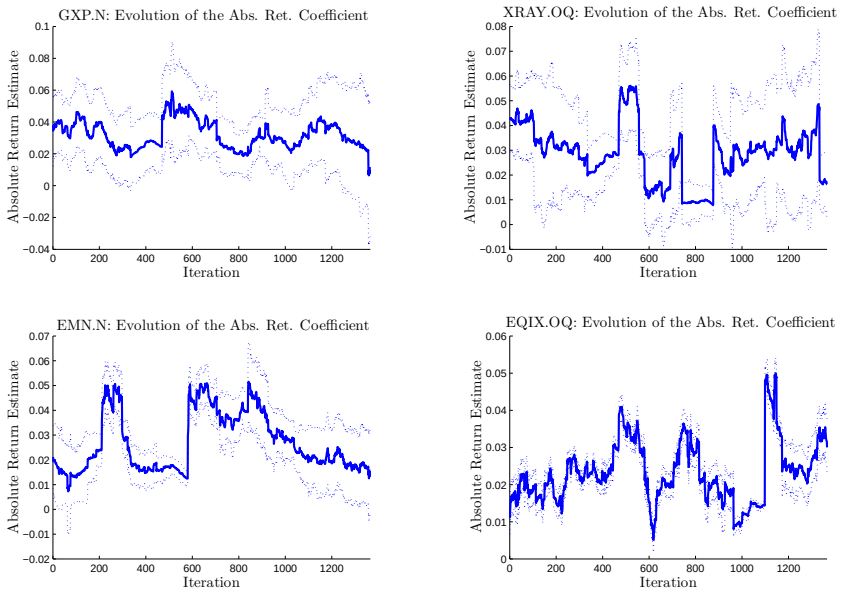

Figure 8: Evolution of the estimates of the absolute return coefficient. 95 % confidence dotted

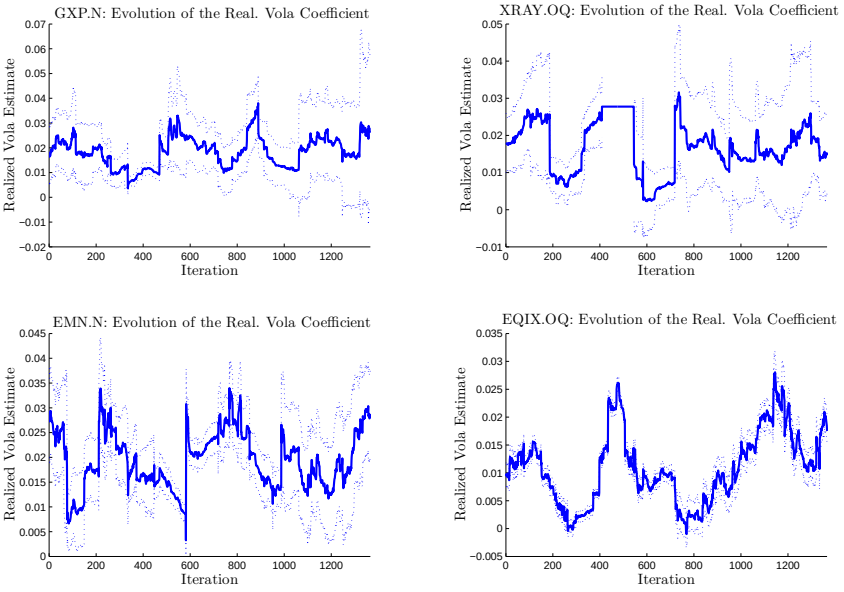

Figure 9: Evolution of the estimates of the Real Vola coefficient. 95 % confidence dotted

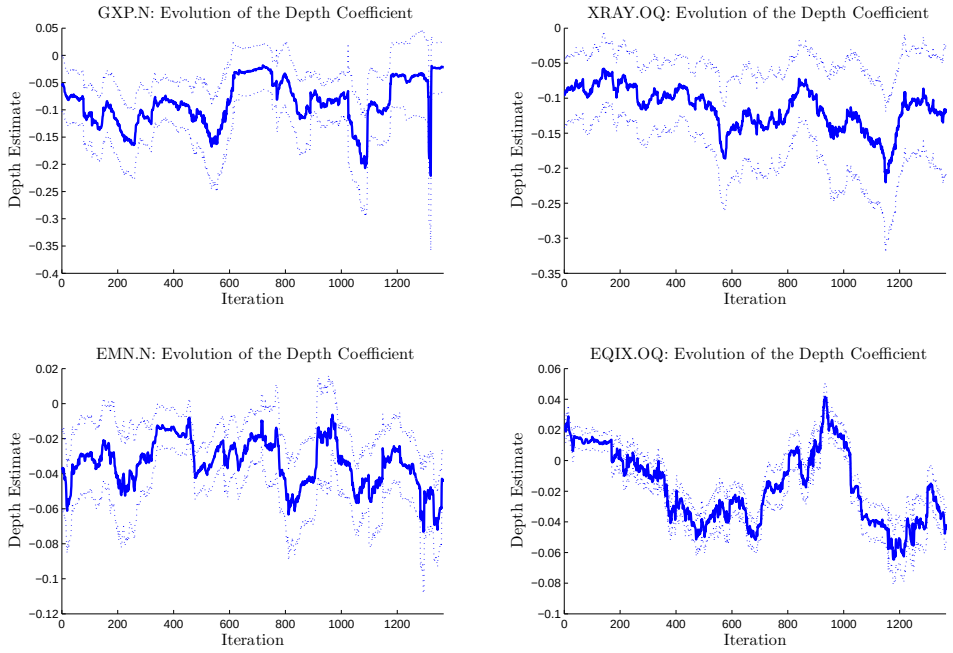

Figure 10: Evolution of the estimates of the Depth coefficient. 95 % confidence dotted

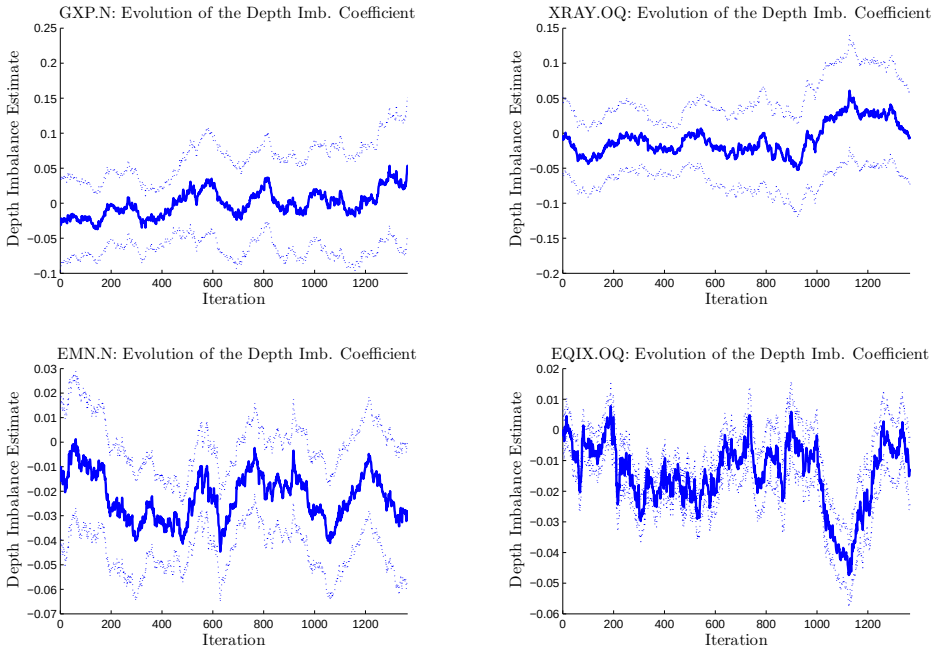

Figure 11: Evolution of the estimates of the Depth Imbalance coefficient. 95 % confidence dotted

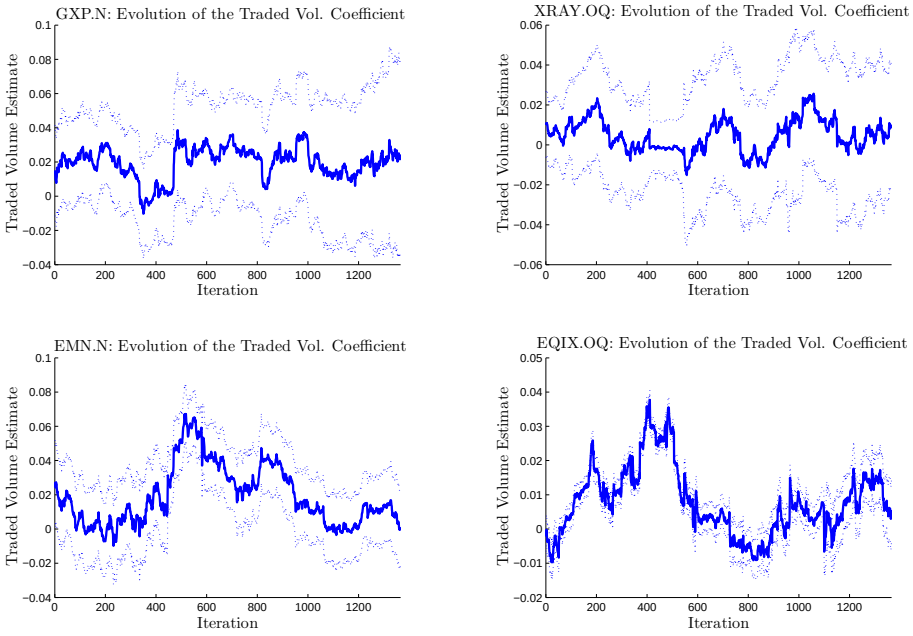

Figure 12: Evolution of the estimates of the Traded Volume coefficient. 95 % confidence dotted

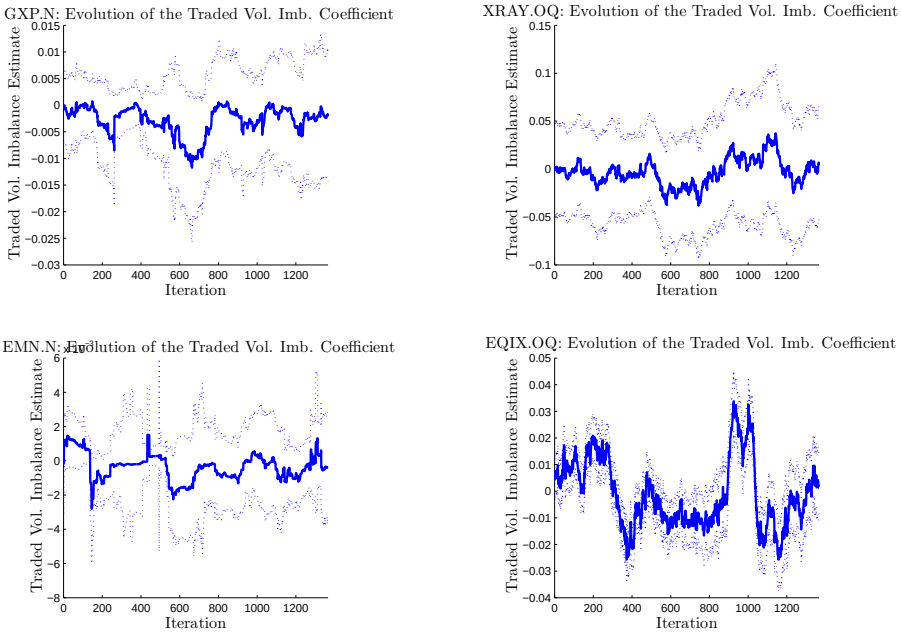

Figure 13: Evolution of the estimates of the Traded Volume Imbalance coefficient. 95 % confidence dotted

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Localising temperature risk" by Wolfgang Karl Härdle, Brenda López Cabrera, Ostap Okhrin and Weining Wang, January 2011.
002 "A Confidence Corridor for Sparse Longitudinal Data Curves" by Shuzhuan Zheng, Lijian Yang and Wolfgang Karl Härdle, January 2011.
003 "Mean Volatility Regressions" by Lu Lin, Feng Li, Lixing Zhu and Wolfgang Karl Härdle, January 2011.
004 "A Confidence Corridor for Expectile Functions" by Esra Akdeniz Duran, Mengmeng Guo and Wolfgang Karl Härdle, January 2011.
005 "Local Quantile Regression" by Wolfgang Karl Härdle, Vladimir Spokoiny and Weining Wang, January 2011.
006 "Sticky Information and Determinacy" by Alexander Meyer-Gohde, January 2011.
007 "Mean-Variance Cointegration and the Expectations Hypothesis" by Till Strohsal and Enzo Weber, February 2011.
008 "Monetary Policy, Trend Inflation and Inflation Persistence" by Fang Yao, February 2011.
009 "Exclusion in the All-Pay Auction: An Experimental Investigation" by Dietmar Fehr and Julia Schmid, February 2011.
010 "Unwillingness to Pay for Privacy: A Field Experiment" by Alastair R. Beresford, Dorothea Kübler and Sören Preibusch, February 2011.
011 "Human Capital Formation on Skill-Specific Labor Markets" by Runli Xie, February 2011.
012 "A strategic mediator who is biased into the same direction as the expert can improve information transmission" by Lydia Mechtenberg and Johannes Münster, March 2011.
013 "Spatial Risk Premium on Weather Derivatives and Hedging Weather Exposure in Electricity" by Wolfgang Karl Härdle and Maria Osipenko, March 2011.
014 "Difference based Ridge and Liu type Estimators in Semiparametric Regression Models" by Esra Akdeniz Duran, Wolfgang Karl Härdle and Maria Osipenko, March 2011.
015 "Short-Term Herding of Institutional Traders: New Evidence from the German Stock Market" by Stephanie Kremer and Dieter Nautz, March 2011.
016 "Oracally Efficient Two-Step Estimation of Generalized Additive Model" by Rong Liu, Lijian Yang and Wolfgang Karl Härdle, March 2011.
017 "The Law of Attraction: Bilateral Search and Horizontal Heterogeneity" by Dirk Hofmann and Salmai Qari, March 2011.
018 "Can crop yield risk be globally diversified?" by Xiaoliang Liu, Wei Xu and Martin Odening, March 2011.
019 "What Drives the Relationship Between Inflation and Price Dispersion? Market Power vs. Price Rigidity" by Sascha Becker, March 2011.
020 "How Computational Statistics Became the Backbone of Modern Data Science" by James E. Gentle, Wolfgang Härdle and Yuichi Mori, May 2011.
021 "Customer Reactions in Out-of-Stock Situations – Do promotion-induced phantom positions alleviate the similarity substitution hypothesis?" by Jana Luisa Diels and Nicole Wiebach, May 2011.

SFB 649, Ziegelstraße 13a, D-10117 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Extreme value models in a conditional duration intensity framework" by Rodrigo Herrera and Bernhard Schipp, May 2011.
023 "Forecasting Corporate Distress in the Asian and Pacific Region" by Russ Moro, Wolfgang Härdle, Saeideh Aliakbari and Linda Hoffmann, May 2011.
024 "Identifying the Effect of Temporal Work Flexibility on Parental Time with Children" by Juliane Scheffel, May 2011.
025 "How do Unusual Working Schedules Affect Social Life?" by Juliane Scheffel, May 2011.
026 "Compensation of Unusual Working Schedules" by Juliane Scheffel, May 2011.
027 "Estimation of the characteristics of a Lévy process observed at arbitrary frequency" by Johanna Kappus and Markus Reiß, May 2011.
028 "Asymptotic equivalence and sufficiency for volatility estimation under microstructure noise" by Markus Reiß, May 2011.
029 "Pointwise adaptive estimation for quantile regression" by Markus Reiß, Yves Rozenholc and Charles A. Cuenod, May 2011.
030 "Developing web-based tools for the teaching of statistics: Our Wikis and the German Wikipedia" by Sigbert Klinke, May 2011.
031 "What Explains the German Labor Market Miracle in the Great Recession?" by Michael C. Burda and Jennifer Hunt, June 2011.
032 "The information content of central bank interest rate projections: Evidence from New Zealand" by Gunda-Alexandra Detmers and Dieter Nautz, June 2011.
033 "Asymptotics of Asynchronicity" by Markus Bibinger, June 2011.
034 "An estimator for the quadratic covariation of asynchronously observed Itô processes with noise: Asymptotic distribution theory" by Markus Bibinger, June 2011.
035 "The economics of TARGET2 balances" by Ulrich Bindseil and Philipp Johann König, June 2011.
036 "An Indicator for National Systems of Innovation - Methodology and Application to 17 Industrialized Countries" by Heike Belitz, Marius Clemens, Christian von Hirschhausen, Jens Schmidt-Ehmcke, Axel Werwatz and Petra Zloczysti, June 2011.
037 "Neurobiology of value integration: When value impacts valuation" by Soyoung Q. Park, Thorsten Kahnt, Jörg Rieskamp and Hauke R. Heekeren, June 2011.
038 "The Neural Basis of Following Advice" by Guido Biele, Jörg Rieskamp, Lea K. Krugel and Hauke R. Heekeren, June 2011.
039 "The Persistence of "Bad" Precedents and the Need for Communication: A Coordination Experiment" by Dietmar Fehr, June 2011.
040 "News-driven Business Cycles in SVARs" by Patrick Bunk, July 2011.
041 "The Basel III framework for liquidity standards and monetary policy implementation" by Ulrich Bindseil and Jeroen Lamoot, July 2011.
042 "Pollution permits, Strategic Trading and Dynamic Technology Adoption" by Santiago Moreno-Bromberg and Luca Taschini, July 2011.
043 "CRRA Utility Maximization under Risk Constraints" by Santiago Moreno-Bromberg, Traian A. Pirvu and Anthony Réveillac, July 2011.

SFB 649, Ziegelstraße 13a, D-10117 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2011

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

044 "Predicting Bid-Ask Spreads Using Long Memory Autoregressive Conditional Poisson Models" by Axel Groß-Klußmann and Nikolaus Hautsch, July 2011.

SFB 649, Ziegelstraße 13a, D-10117 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".