

Andreas Fuest

---

# Econometric Modeling of Ultra-High Frequency Volatility-Liquidity Interactions

---



Dissertation an der Fakultät für Mathematik, Informatik und Statistik der  
Ludwig-Maximilians-Universität München

Eingereicht am 18. Juni 2015



Andreas Fuest

---

# **Econometric Modeling of Ultra-High Frequency Volatility-Liquidity Interactions**

---

Dissertation an der Fakultät für Mathematik, Informatik und Statistik der  
Ludwig-Maximilians-Universität München

Eingereicht am 18. Juni 2015

Erster Gutachter: Prof. Stefan Mittnik, PhD  
Ludwig-Maximilians-Universität München

Zweiter Gutachter: Prof. Dr. Joachim Grammig  
Eberhard Karls Universität Tübingen

Tag der Disputation: 17. August 2015





Over the years I have challenged many audiences to tell me about one complicated model that works well in explanation and prediction and have not heard of a single one.

*Arnold Zellner*



# Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der dynamischen ökonomischen Modellierung hochfrequenter Finanzdaten, welche börsentäglich in großem Umfang von elektronischen Handelssystemen, sogenannten *Limit Order Books* (LOBs), generiert werden. Zusätzlich zu (Handels-) Preisen, die Gegenstand der meisten Modelle der Finanzökonomie und -mathematik sind, liefern solche Daten Informationen zur Liquidität eines Wertpapiers: Sie geben zu jedem Zeitpunkt Auskunft über sämtliche an einem Markt angebotenen und nachgefragten Mengen. Aus Modellierungssicht handelt es sich dabei um einen hochdimensionalen Vektor, der sich in sehr hoher sowie schwankender Frequenz ändert. Die Herausforderung bei der Modellierung besteht, neben der praktischen Komplikation der schier Größe der Datensätze, in einer sinnvollen Reduktion dieser Preis/Mengen- bzw. zeitlichen Dimension.

Die zentrale Idee dieser Arbeit ist es, die beobachtete Liquidität als Realisation eines funktionalen, d.h. *kurvenwertigen* stochastischen Prozesses aufzufassen. Auf diese Weise ist es möglich, einige der wichtigsten Modelle der Finanzökonomie auf innovative Weise um den Liquiditätsaspekt zu erweitern. Bestehende statistische Ansätze aus dem Gebiet der funktionalen Datenanalyse (FDA) spielen hierbei eine entscheidende Rolle.

Basierend auf einem dynamischen semiparametrischen Faktoransatz für Liquiditätskurven sind dies zum einen *multiplikative Fehlermodelle* (MEMs) mit funktionalen Liquiditätseffekten, GARCH- und ACD-FunXL. Zum anderen werden, analog zur Realized-Volatility-Literatur, Ex-post-Maße der Liquidität und Liquiditätsunsicherheit konstruiert. Die realisierte Variation des Preises wird sodann im Rahmen eines linearen Zeitreihenmodells mit realisierten Liquiditätseffekten, HAR-FunXL, modelliert. Die Markov-Eigenschaft des HAR-FunXL-Prozesses (und Markov'scher Versionen der MEMs) ermöglicht dessen Einbettung in die Klasse der generalisierten additiven funktionalen Regressionsmodelle, für deren Schätzung auch einige Alternativen zum faktorbasierten Ansatz diskutiert werden.

Anwendungen all dieser Modelle auf Daten des XETRA-Systems der Deutschen Börse zeigen, dass die vorgeschlagenen Verfahren das Zusammenspiel von Liquidität und Volatilität besser zu verstehen helfen. Ferner wird gezeigt, dass Liquiditätsinformationen in vielen Situationen Volatilitätsprognosen zu verbessern vermögen.



## Summary

This thesis is concerned with the dynamic econometric modeling of high-frequency financial data. Such data are generated by electronic trading systems, so-called *Limit Order Books* (LOBs), each trading day. Apart from security prices, for which most models in financial econometrics and financial mathematics are designed, such data reveal information on the liquidity of a security. At each point in time during the trading day, supply and demand in the market can be reconstructed from LOB data. From a modeling perspective, the LOB is a high-dimensional vector subject to random changes at high and irregular (random) frequency. Apart from the size of LOB data, the major challenge in modeling such a process is a sensible reduction of both the price/size dimension of supply and demand, and the temporal dimension.

The focal idea of this thesis is to view observed liquidity as realization of a functional, i.e., *curve-valued* stochastic process. Some of the most important models of financial econometrics are extended by adding the liquidity aspect in an entirely new way. These extensions heavily draw on statistical methodology developed in the young statistical field of functional data analysis (FDA).

Firstly, multiplicative error models (MEMs) with functional liquidity impact, namely GARCH-FunXL and ACD-FunXL, are proposed based on a semiparametric factor model of liquidity curves. Secondly, in analogy to the realized volatility approach, ex-post measures of liquidity and liquidity uncertainty are introduced. To investigate the impact of realized liquidity on the price process, the HAR-FunXL model is proposed. The HAR-FunXL process, in contrast to MEMs, is a finite-order Markov process. It is shown that this model (as well as Markovian versions of some MEMs) can be embedded into the class of generalized additive functional regression models, facilitating the use of alternative estimation strategies to the factor-based approach.

Applications of the models to LOB data for the German XETRA system show that the proposed methods are able to uncover connections between liquidity and volatility. Moreover, it is shown that liquidity information is valuable for forecasting price volatility in many situations.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Limit order books as functional time series</b>	<b>5</b>
2.1	Introduction . . . . .	5
2.2	Limit order books . . . . .	6
2.2.1	LOB-implied liquidity . . . . .	9
2.2.2	The XETRA LOB data . . . . .	11
2.3	Functional time series . . . . .	12
2.3.1	The basic setting . . . . .	12
2.3.2	Estimation . . . . .	13
2.3.3	Dynamic dimension reduction . . . . .	16
2.4	The LOB as functional time series . . . . .	17
2.4.1	Identification of eigenfunctions . . . . .	20
2.4.2	Uncertainty assessment . . . . .	23
2.5	Functional dynamic modeling of liquidity . . . . .	25
2.5.1	Semiparametric factor model . . . . .	26
2.5.2	Functional autoregressive model . . . . .	27
2.5.3	Empirical results . . . . .	30
2.6	Conclusion . . . . .	31
<b>I</b>	<b>Multiplicative Error Models with Liquidity Impact</b>	<b>37</b>
<b>3</b>	<b>Liquidity impact on volatility: The GARCH-FunXL model</b>	<b>41</b>
3.1	Introduction . . . . .	41
3.2	Liquidity . . . . .	43
3.2.1	Limit order book information . . . . .	43
3.2.2	Diurnal patterns . . . . .	44
3.2.3	Liquidity as functional time series . . . . .	45
3.2.4	Empirical results . . . . .	46

3.3	GARCH-FunXL . . . . .	52
3.3.1	The model . . . . .	52
3.3.2	Estimation . . . . .	54
3.3.3	Liquidity impact . . . . .	58
3.4	Estimation uncertainty in GARCH-FunXL models . . . . .	59
3.5	Modeling XETRA returns . . . . .	61
3.5.1	Estimation results . . . . .	61
3.5.2	Liquidity impact . . . . .	62
3.5.3	Forecasting . . . . .	66
3.6	Conclusion . . . . .	66
<b>4</b>	<b>Functional liquidity and low-latency volatility: The ACD-FunXL model</b>	<b>71</b>
4.1	Introduction . . . . .	71
4.2	Low-latency volatility . . . . .	73
4.3	Low-latency liquidity in a limit order book . . . . .	75
4.3.1	Latency of orders . . . . .	75
4.3.2	Liquidity measure . . . . .	76
4.3.3	Functional time series methods at high frequency . . . . .	78
4.3.4	Diurnal patterns of liquidity . . . . .	78
4.3.5	Diurnal duration patterns . . . . .	80
4.4	Econometric model . . . . .	81
4.4.1	The model . . . . .	81
4.4.2	Estimation . . . . .	83
4.4.3	Liquidity impact . . . . .	86
4.5	In-sample evaluation . . . . .	86
4.5.1	Information criteria . . . . .	87
4.5.2	Results . . . . .	88
4.5.3	The functional parameter . . . . .	93
4.5.4	Liquidity impact . . . . .	93
4.6	Forecasting price durations . . . . .	94
4.7	Conclusion . . . . .	96
<b>II</b>	<b>Realized Volatility and Liquidity Impact</b>	<b>99</b>
<b>5</b>	<b>Realized measures of liquidity</b>	<b>103</b>
5.1	Introduction . . . . .	103
5.2	Realized volatility . . . . .	104
5.3	A continuous-time pure-jump model of liquidity . . . . .	108
5.3.1	The liquidity model . . . . .	108
5.3.2	Properties of the liquidity model . . . . .	110
5.3.3	Realized measures . . . . .	113
5.3.4	Properties of realized measures of liquidity . . . . .	115

5.4	The NHPP liquidity model: Estimation and simulation . . . . .	117
5.4.1	Estimation . . . . .	118
5.4.2	Simulation . . . . .	119
5.5	Empirical results . . . . .	123
5.6	Conclusion . . . . .	131
<b>6</b>	<b>Realized volatility and realized liquidity: The HAR-FunXL model</b>	<b>133</b>
6.1	Introduction . . . . .	133
6.2	The HAR-FunXL model . . . . .	134
6.3	Estimation . . . . .	135
6.3.1	The functional part . . . . .	136
6.3.2	Inference for the HAR-FunXL model . . . . .	138
6.4	Empirical results . . . . .	140
6.4.1	Explaining realized volatility . . . . .	140
6.4.2	Forecasting realized volatility . . . . .	146
6.5	Generalizations . . . . .	149
6.5.1	Families of conditional distributions . . . . .	149
6.5.2	Conditional heteroskedasticity and higher-order moments . . . . .	150
6.5.3	Multivariate models . . . . .	150
6.6	Conclusion . . . . .	151
<b>7</b>	<b>Conclusions and perspectives</b>	<b>161</b>
	<b>Bibliography</b>	<b>163</b>
<b>A</b>	<b>Additional empirical results</b>	<b>175</b>
A.4	Chapter 4: Price durations and the ACD-FunXL model . . . . .	175
A.4.1	Goodness-of-Fit . . . . .	175
A.4.2	Forecast results . . . . .	184
A.5	Chapter 5: Realized measures of liquidity . . . . .	192



## List of Figures

2.1	Sketch of a limit order book . . . . .	8
2.2	Diurnal liquidity patterns for Commerzbank, Linde, and MunichRe . .	18
2.3	Raw and de-diurnalized ask curves, Commerzbank, and SACF for sev- eral $d$ . . . . .	19
2.4	Eigenanalysis for Commerzbank liquidity curves . . . . .	21
2.5	Modes of variation for Commerzbank ask liquidity curves . . . . .	22
2.6	Stationary bootstrap results for the eigenfunctions of Commerzbank ask liquidity curves . . . . .	24
2.7	Estimated kernels of FAR(1) models, ask curves, Linde . . . . .	32
3.1	Snapshot of a limit order book at some time during the trading day . .	42
3.2	Daily closing prices and log returns for Commerzbank, Linde, and Mu- nichRe . . . . .	47
3.3	Diurnal volatility and liquidity patterns for Commerzbank, Linde, and MunichRe . . . . .	49
3.4	Functional time series of ask and bid liquidity curves for two exemplary trading days . . . . .	50
3.5	Eigenvalues and eigenfunctions of Commerzbank's liquidity series . . .	51
3.6	Estimated functional parameters for the GARCH-FunXL model . . . . .	63
3.7	Estimated liquidity impact for GARCH-FunXL models with two func- tional liquidity processes . . . . .	64
3.8	SACF of estimated liquidity impact . . . . .	65
3.9	Cumulative forecast gain, QLIKE(pure log-GARCH) - QLIKE(log-GARCH- FunXL), of liquidity-driven GARCH models . . . . .	67
4.1	State of a limit order book at several times during a trading day . . . .	72
4.2	Limit order submissions and estimated unconditional survivor func- tions of buy and sell limit orders, Commerzbank and Linde . . . . .	77
4.3	De-diurnalized ask liquidity curve for the Commerzbank stock and its FPC approximation . . . . .	79

4.4	Estimated eigenfunctions of liquidity curves for Commerzbank and Linde	80
4.5	Diurnal duration pattern for Commerzbank 5-tick price durations . . . . .	81
4.6	Ljung-Box statistics (10 lags) for data and residuals . . . . .	89
4.7	Estimates of the persistence parameter . . . . .	90
4.8	Volatility of Commerzbank ask quotes in 10/2010 . . . . .	92
4.9	Optimal liquidity dimension for the purpose of forecasting instantaneous volatility . . . . .	92
4.10	Functional parameter estimates for bid and ask liquidity impact, Commerzbank data ( $\delta = 5$ ticks) . . . . .	94
4.11	Liquidity impact on instantaneous volatility for the Commerzbank stock in 10/2010 . . . . .	95
5.1	Illustration of the properties of a nonhomogeneous Poisson process by means of simulation . . . . .	111
5.2	Estimated nonhomogeneous Poisson intensities . . . . .	120
5.3	Distribution of liquidity innovations: Histogram vs. exponential and lognormal fits . . . . .	121
5.4	Simulated trajectories of the nonhomogeneous Poisson liquidity process	122
5.5	Results of the simulation study on realized variance of liquidity for alternative sampling schemes . . . . .	124
5.6	Mean cumulative realized variance for Linde's ask curves . . . . .	126
5.7	Functional realized measures of liquidity for Linde . . . . .	127
5.8	Eigenfunctions of functional realized measures of liquidity, Linde . . . . .	128
5.9	Functional principal components scores for Linde's ask curves . . . . .	129
5.10	Cumulative relative eigenvalues as a function of $K$ , for Linde, MunichRe, and Commerzbank . . . . .	130
6.1	Estimates of the autoregressive parameters for the HAR-FunXL model, realized mid-price variation . . . . .	144
6.2	Estimates of the autoregressive parameters for the HAR-FunXL model, realized transaction price variation . . . . .	145
6.3	Log realized variance and relative liquidity impact on realized quote variation . . . . .	147
6.4	Log realized variance and relative liquidity impact on realized transaction price variation . . . . .	148
A.1	Functional realized measures of liquidity for MunichRe . . . . .	193
A.2	Eigenfunctions of functional realized measures of liquidity, MunichRe . . . . .	194
A.3	Functional principal components scores for MunichRe's ask curves . . . . .	195
A.4	Functional realized measures of liquidity for Commerzbank . . . . .	196
A.5	Eigenfunctions of functional realized measures of liquidity, Commerzbank	197
A.6	Functional principal components scores for Commerzbank's ask curves	198

# List of Tables

2.1	One-step-ahead out-of-sample forecasts, FAR(1) vs. SPF-AR(1) models .	34
2.2	Three-step-ahead out-of-sample forecasts, FAR(1) vs. SPF-AR(1) models	35
3.1	Contemporaneous sample correlations between FPC score series of bid and ask curves, Commerzbank . . . . .	52
3.2	MSE of liquidity impact for simulated GARCH-FunXL models . . . . .	68
3.3	AIC and BIC for GARCH-FunXL models with bid and ask liquidity impact . . . . .	69
3.4	Negative out-of-sample likelihoods for one-step-ahead forecasts of GARCH-FunXL models . . . . .	69
4.1	Number of quote changes of specific size for Linde and Commerzbank	75
4.2	AIC of ACD-FunXL models, Commerzbank price durations . . . . .	93
4.3	Out-of-sample forecast performance of ACD-FunXL for Commerzbank price durations . . . . .	96
5.1	Descriptive statistics: Number of orders and mean order sizes . . . . .	118
6.1	AIC and BIC of the in-sample fits for the HAR-FunXL model with <i>contemporaneous</i> realized liquidity impact, $CRVL_I$ , RV based on the mid-price	152
6.2	AIC and BIC of the in-sample fits for the HAR-FunXL model with <i>lagged</i> realized liquidity impact, $CRVL_I$ , RV based on the mid-price . . . . .	153
6.3	AIC and BIC of the in-sample fits for the HAR-FunXL model with <i>contemporaneous</i> realized liquidity impact, $CRVL_I$ , RV based on transaction prices . . . . .	154
6.4	AIC and BIC of the in-sample fits for the HAR-FunXL model with <i>lagged</i> realized liquidity impact, $CRVL_I$ , RV based on transaction prices . . . . .	155
6.5	AIC and BIC of the in-sample fits for the HAR-FunXL model with <i>contemporaneous</i> realized liquidity impact, $CRVL_{II}$ , RV based on the mid-price . . . . .	156

6.6	AIC and BIC of the in-sample fits for the HAR-FunXL model with lagged realized liquidity impact, $CRVL_{II}$ , RV based on the mid-price . . . . .	157
6.7	RMSE of one-step-ahead out-of-sample forecasts for the HAR-FunXL model with lagged realized liquidity impact . . . . .	158
6.8	RMSE of one-step-ahead out-of-sample forecasts for the HAR-FunXL model with lagged realized liquidity impact . . . . .	159
6.9	RMSE of one-step-ahead out-of-sample forecasts for the HAR-FunXL model with lagged realized liquidity impact . . . . .	160
A.1	AIC for EACD-FunXL, Commerzbank. . . . .	175
A.2	BIC for EACD-FunXL, Commerzbank. . . . .	177
A.3	AIC for EACD-FunXL, Linde. . . . .	180
A.4	BIC for EACD-FunXL, Linde. . . . .	182
A.5	1-step forecast MSEs, EACD-FunXL, bid-quote durations, Commerzbank, for days 16 to 21 in the sample. . . . .	184
A.6	1-step forecast negative log-likelihoods, EACD-FunXL, bid-quote durations, Commerzbank. . . . .	186
A.7	1-step forecast MSEs, EACD-FunXL, bid-quote durations, Linde. . . . .	188
A.8	1-step forecast negative log-likelihoods, EACD-FunXL, bid-quote durations, Linde. . . . .	190

## Introduction

This thesis develops econometric models which aim at exploiting the full ultra-high frequency information available in electronic trading systems in order to measure and forecast financial risk.

Such electronic trading systems do not only reveal information on prices of financial securities, whose statistical analysis has a long tradition tracing back to Bachelier (1900), but also on liquidity present in the market place. According to Amihud et al. (2013),

*Liquidity and its converse, illiquidity, are elusive concepts: You know it when you see it, but it is hard to define.*

In this thesis, however, we argue that liquidity can not only be accurately defined, but even *directly observed* using limit order book (LOB) data. This claim resembles somewhat the claim that *volatility* can be observed based on high-frequency data, as it has been lodged by the founders of the realized volatility revolution (Andersen et al., 2000). As we will show, liquidity, i.e., limit order schedules for a given security can be considered as a bivariate curve-valued or *functional* stochastic process in continuous time. However, as individual order submissions or revisions typically do not change the overall state of the LOB much, it is sensible to employ time series methodology on data generated by a discrete sampling scheme. Which sampling scheme — for example: which sampling frequency — is appropriate depends on the specific application.

The major contribution of this thesis is to augment some of the paramount models of financial econometrics by the liquidity aspect in an entirely new way, building the class of *FunXL* (*Functional eXogeneous Liquidity*) models. In particular, these are

- The *GARCH-FunXL model*, which aims to capture the impact of limit order schedules on intraday dynamics of the conditional return distribution's *variance* at "low" intraday frequencies, i.e., in situations where a continuous return distribution is appropriate.
- The *ACD-FunXL model*, designed to capture functional liquidity impact in situations where interest focuses on the *location* parameter of conditional distributions of time-varying LOB variables with positive support, such as (price, volume, trade) durations or realized measures of volatility.

- The *HAR-FunXL model*, mapping realized measures of liquidity (RML) to realized measures of volatility (RMV) in a dynamic setting. To this end, a first step towards a theory of RML is made. The model belongs to a greater class of Markovian generalized autoregressive FunXL models.

In order to capture the FunXL aspect of LOB data, we draw on recent advances in functional time series methodology and on functional regression techniques on the one hand. On the other hand, we contribute to and draw on the recently active literature on GARCH-X models and their relatives.

### Guideline through the thesis

The thesis can roughly be divided into an introductory Part (Chapters 1 & 2), and two Parts (Chapters 3 & 4, Chapters 5 & 6) presenting the major contributions of this work.

More specifically, Chapter 2 sets the stage for the new econometric methodology introduced in Chapters 3 to 6. Our measure of LOB-implied liquidity is motivated, introduced, and discussed against the background of other liquidity measures put forth in the literature. Its basic statistical properties are presented based on equidistant LOB snapshots. However, we still hesitate to call these properties *stylized facts* as they may vary somewhat from marketplace to marketplace due to different technological and institutional settings. We then use the framework of *functional time series analysis* (FTSA), which allows to model liquidity as a discrete-time stochastic process of curve-valued dynamic objects, to capture liquidity by means of a statistical model. Finally, we use the functional generalization of the AR process, the FAR model due to Bosq (2000), and a semiparametric dynamic factor model to investigate the predictability of liquidity.

Chapters 3 and 4 constitute *Part I*, introducing multiplicative error models (MEMs) with functional liquidity impact. The basic idea is to include liquidity as exogenous information in the conditional variance (mean) equation of GARCH (ACD) models. The methodological challenge lies in mapping liquidity curves to the scalar time-varying parameter of the (univariate) conditional distribution of financial returns, durations or, possibly, other non-negative quantities of interest. The task, which is an ill-posed inverse problem, is accomplished using the dimension reduction techniques introduced in Chapter 2, namely, functional principal component (FPC) decompositions of the curves along with a basis expansion of a functional parameter which exploits the orthonormality of FPCs. We propose a two-step quasi-ML estimation procedure, conjecture its statistical properties, and investigate its finite-sample performance by means of simulations. An extensive empirical application to three DAX constituents traded on XETRA shows that liquidity matters to explain price variation in-sample, and it is shown to augment prediction of volatility in an out-of-sample exercise. Most parts of Chapter 3 and some parts of Chapter 2 are based on the working paper *Fuest, A. and S. Mittnik (2015): Modeling Liquidity Impact on Volatility: A GARCH-FunXL Approach*, which is not yet published, but was selected for the peer-reviewed 8th Annual Society for Financial Econometrics Conference.

While in Part I, LOB *snapshots* are modeled, in Part II, snapshots are *aggregated* over time. Chapters 5 and 6 constitute *Part II*. In Chapter 5, ex-post “realized” measures of liquidity and liquidity variation are introduced. Liquidity at each price level represented in the LOB is here viewed as a realization of a continuous-time compound nonhomogeneous Poisson process, i.e., a generalization of the celebrated compound homogeneous Poisson model of Press (1967), originally proposed to model prices. We show that accurate estimation is still possible even if only a rather small number of intraday LOB snapshots is used. By collecting realized measures at all price levels, we construct (again) *functional* realized measures of liquidity.

In Chapter 6, we then model the impact of these realized measures on price volatility in a HAR framework with functional exogeneous liquidity impact, based on dimension reduction techniques already used in Chapters 3 and 4. The ability of realized liquidity to help explain and predict price variation is investigated in-sample and out-of-sample. We also show that several Markovian models with functional liquidity impact, HAR-FunXL being a special case, can be viewed as (functional, additive) generalized linear regression models. This fact allows to import different approaches from the respective literature, namely penalized functional regression (PFR), FPC regression, and FPLS regression (Goldsmith et al., 2011; Reiss and Ogden, 2007; Goldsmith and Scheipl, 2014; Reiss et al., 2015), providing estimates of the functional parameter based on explicit roughness penalties. We compare one of these approaches, PFR, to our original FPC approach with respect to their in- and out-of-sample performance, finding that the benefits of roughness penalization are disputable in our context.

Chapter 7 concludes the thesis and discusses some directions of further research.

## Some remarks about notation

Although all variables are introduced and defined in the text where it is most suitable with regard to the subsequent explanations, it seems appropriate to introduce some notational principles.

**Deterministic vs. random** We do *not* follow the convention of using capital letters for random quantities and small ones for deterministic quantities as we found it impossible to sustain and, in particular, to avoid collisions with other notational conventions explained below. Whether or not a quantity is random is always stated clearly.

**Exogeneous vs. endogeneous** The FunXL class of models in Parts I and II contains endogeneous quantities, denoted by  $y$ , and exogeneous quantities, denoted by  $x$ . As this distinction is crucial, at times we pay the price of using many sub- and superindexes to consequently stick to this convention.

**Dimension of objects** We have to denote random quantities (which can be either scalars, vectors, matrices, or infinite-dimensional objects), their observed realizations,

## 1 Introduction

and parameters appearing as scalars, vectors, matrices, and functions. Vectors (matrices) are denoted by bold small (bold capital) letters, and vectors are always column vectors if not explicitly stated otherwise. Almost all functions appearing in this thesis map  $\mathbb{R}^p$  or  $[0, 1]$  to  $\mathbb{R}$ , where either  $p = 1$  or  $p = 2$ . Where not obvious from the context, functions are often denoted along with their argument(s), for instance  $x(d)$  or  $\Sigma(c, d)$ . Sometimes functional quantities are put together in constructs like  $x(d)$ , which denotes a set of functions (argument in rows) stacked in a column vector, which we call a “vector” (although it could, very informally, also be described as a  $\#$  rows  $\times$   $\infty$  “matrix”). At some instances,  $x(d)$  represents functions joined together at the ends of their (rescaled) domains.

**Time** In Part I, models almost exclusively live in discrete time, however, a daily and an intraday clock are distinguished at times. For example,  $x_{t,i}$  may denote some  $x$  at intraday time  $i$  on day  $t$ , which means there are two simultaneous discrete-time clocks. In the context of duration models, there are quantities like  $x_{\ell,t,i}$ , i.e. *counted* ( $i$ ) event times in continuous time  $t$ , observed over trading days  $\ell$ . However, this can more compactly written as  $x_{\ell,i}$  without a loss of (essential) information.

In Part II, we introduce continuous-time processes whose properties change from day to day, i.e., in discrete time. Here,  $t$  is always the continuous clock, and  $\ell$  is the discrete clock, as for example in  $x_{\ell,t}$ .

**Price as “space”** The functional quantities introduced are functions of the (continuous) relative price  $d$  to be defined below, so there are quantities like  $x_{t,i}(d)$ , i.e. functional random variables (or their realizations) evolving in (discrete or continuous) time.

**Market side** We always distinguish between supply (ask side) and demand (bid side) in a market, which is denoted by a superindex ( $s$ ),  $s \in \{\text{bid}, \text{ask}\}$ .

## Limit order books as functional time series

### 2.1 Introduction

In this Chapter, we introduce the functional approach to liquidity measurement. As has already been stressed in the introductory Chapter, liquidity is hard to define. An informal definition, *ex negativo*, may be that *illiquidity* of an asset is the effort that has to be made to trade it. More comprehensively, Kyle (1985) states

*[...] a liquid market is a continuous market, in the sense that almost any amount of stock can be bought or sold immediately, and an efficient market, in the sense that small amounts of stock can always be bought and sold very near the current market price, and in the sense that large amounts can be bought or sold over long periods of time at prices that, on average, are near the current market price.*

In much of the econometric literature, a liquid security is simply one which is heavily traded. However, stating an equivalence of trading volume and liquidity confuses things: There may be situations where the market is perfectly liquid but there is no trading, or there may be heavy trading but low liquidity. The paramount example for the latter situation is the flash crash of May 6, 2010, where heavy trading occurred during a transitory drop and subsequent recovery of US stock markets which was due to a *lack* of liquidity (Kirilenko et al., 2014; Easley et al., 2011).

A simple yet more convincing alternative is Kyle's Lambda, which has been developed with a continuous auction model framework in view (Kyle, 1985). The idea is to relate the size of price changes to trading volume. For an interval  $[t - \Delta, t]$  during the trading day, it is defined as

$$\lambda_t = \frac{|\text{Price}_t - \text{Price}_{t-\Delta}|}{\text{Volume}_{[t-\Delta, t]}}$$

where volume is measured in terms of the monetary value of the traded shares, and not by the number of traded shares. Note that the prominent illiquidity measure of Amihud (2002) is almost equivalent. High liquidity corresponds to a small Lambda in

this notion. For a given trading volume, higher price changes therefore always imply lower liquidity, even if they are driven by fundamental news arrivals.

Both liquidity measures, trading volume and Kyle's Lambda, have in common that they are computed based on trade information. If we characterize liquidity as the *potential* cost of trading, then such trading-based measures obviously are biased as they neglect liquidity that may be present even when no trading occurs. In contrast, modern electronic trading systems provide information not only on trades, but also on the costs of potential trades. Therefore, it can be argued that these trading systems make liquidity *observable* — in very much the same way as volatility has been claimed to be made observable by the availability of high-frequency data on trades and quotes, 15 years ago (Andersen et al., 2000).<sup>1</sup> It may still be criticized, however, that only liquidity implied by submitted limit orders is observable, whereas a full picture of liquidity also consists of unobservable trading intentions of market participants.

In this Chapter, we introduce measures of liquidity that are based on full information about limit order schedules. As it turns out, such measures are large-dimensional but smooth, which is why we choose to view them as realizations of a *functional* stochastic process. Moreover, liquidity can in principle be observed at each point in time during a trading day, giving rise to *continuous-time* methods. However, most changes of liquidity typically only occur locally, leading to strong serial dependence at high frequency. It depends on the specific purpose of the analysis which sampling scheme and frequency are appropriate. We advocate dimension reduction techniques for *discrete-time* stochastic processes with short memory, for which theoretical justifications have become available recently (Hörmann and Kokoszka, 2010).

The Chapter proceeds as follows. After a short introduction of limit order books (LOBs) in general, the structure and dynamics of LOBs, especially the German XETRA LOB, are presented. We then introduce liquidity *curves* which can be constructed from order book data, and use the framework of functional time series analysis (FTSA), to capture the curves statistically. This idea allows to model liquidity curves as a discrete-time stochastic process of curve-valued objects. Two models that have been proposed in the literature, namely the functional AR model (Bosq, 2000) and a semi-parametric factor model (Hyndman and Shang, 2009; Aue et al., 2015), are used to characterize the dynamics and, in particular, to assess the predictability of the process.

## 2.2 Limit order books

We introduce the basic rules and principles of LOBs that we need for subsequent analyses. As many exchanges for financial securities exist around the world, and even the trading rules for a given exchange are time-varying, we confine ourselves to the description of the most basic mechanisms, thereby neglecting things like iceberg orders, midpoint orders etc. For an excellent, comprehensive, and timely survey of empirics and theory for LOBs see Gould et al. (2013).

---

<sup>1</sup>At least this was what Andersen et al. (2000) *expected*, as the Dickensian title of their paper suggests.

Market participants can submit *limit* and *market* orders. For both order types, they have to specify the *order size*, which must be a positive integer-valued multiple of the *lot size*, which typically (but not necessarily) amounts to 1 share. A sell (buy) market order is the offer (request) to sell (buy) a certain number of shares of a financial security at an *unspecified* price. A sell (buy) market order is therefore executed against the demand (supply) on the opposite side of the market. Supply and demand stem from the submission of *limit orders*. A limit order is not only characterized by its direction (buy or sell) and size, but additionally by an execution price which has to be specified as a positive integer-valued multiple of the *tick size*  $\delta$ . Limit orders which meet a suitable limit order on the opposite market side are immediately executed and called *effective market orders*. All other limit orders are made visible to the other market participants, and called *active* limit orders until they are either executed against a suitable order from the other side of the market, or cancelled. The set of all active limit orders at some time during the trading day is called the *limit order book* (LOB).

Figure 2.1 shows a stylized example of a LOB at some time during the trading day. The number of offered (requested) shares implied by the active limit orders are depicted in red (blue). We see that prices live on a discrete grid defined by the tick size. The lowest (highest) price at which shares are being offered (requested) is called the ask (bid) price, in the example  $P_0 + 2\delta$  and  $P_0$ , respectively. These prices are also called the *quotes*. The difference between the two is called the bid-ask *spread*, which always amounts to at least  $\delta$ . A limit order submission inside the spread (here: at  $P_0 + \delta$ ) is called an *aggressive* limit order.

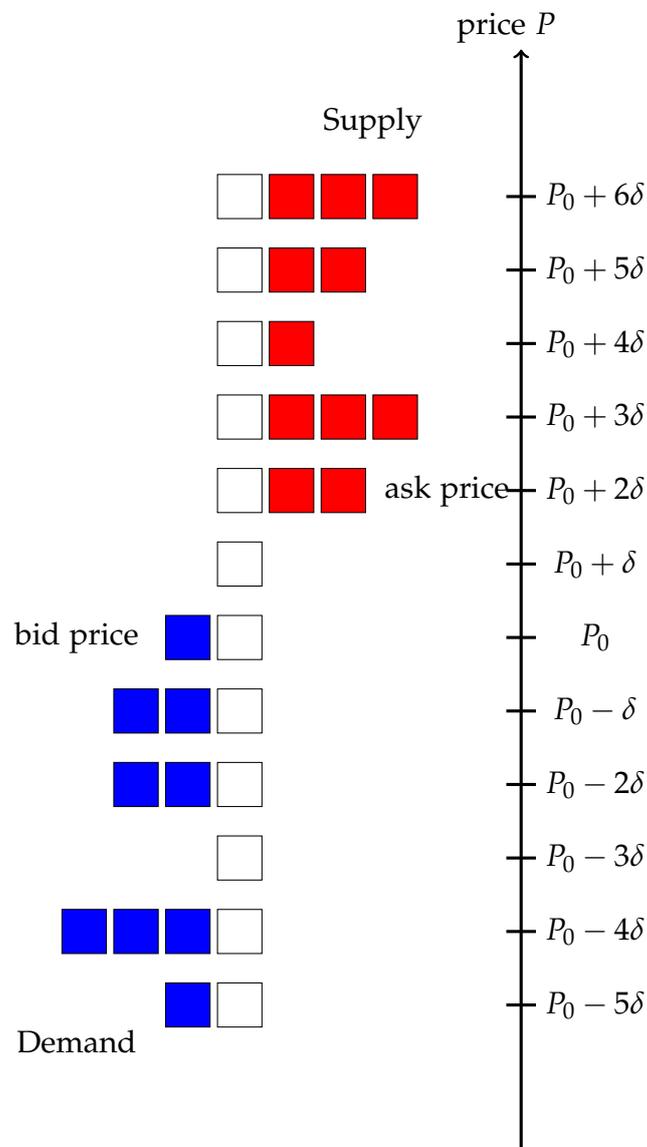
A market buy order of, say, size 5 would absorb the outstanding shares at  $P_0 + 2\delta$  and  $P_0 + 3\delta$ , increasing both ask price and spread by  $2\delta$ . A market sell order of the same size would absorb the best three price levels' outstanding shares, and (since there is no share at  $P_0 - 3\delta$  outstanding) both decrease the bid price and increase the spread by  $4\delta$ .

We now turn to introducing some notation.  $LOB_t$ , the state of the limit order book at some time  $t$  during a trading day, can be characterized by  $P_t^{(s)}$ ,  $s \in \{bid, ask\}$ , the quotes (measured in ticks), and  $v_t^{(s)}(d)$ , the outstanding number of shares on market side  $s$  and at a price distance  $d$  (in ticks) from the respective quote. In the following, we call  $d$  the *relative price* or *distance*. As limit sell orders may, in principle, be posted at any integer number, we write  $d_t^{max} = \max\{d | v_t^{(ask)}(d) > 0\}$  for supply at the highest relevant price to characterize the *dimension* of  $LOB_t$ . Limit buy orders can be posted at any price level between (but excluding) zero and the ask quote.

Then, the LOB at time  $t$  is given by the  $D_t = (P_t^{(bid)} + d_t^{max} + 3)$ -dimensional vector

$$LOB_t := [P_t^{(bid)}, P_t^{(ask)}, v_t^{(bid)}(P_t^{(bid)} + 1), \dots, v_t^{(bid)}(0), v_t^{(ask)}(0), \dots, v_t^{(ask)}(d_t^{max})]'$$

$D_t$  is typically very large, and the pattern of active orders (i.e., all tuples {price, #



**Figure 2.1:** An example limit order book. The price  $P_0$  is arbitrarily set as an anchor. Blue (red) boxes stand for outstanding requested (offered) shares. The tick size is  $\delta$ . The ask price (lowest offer price) is  $P_0 + 2\delta$ , the bid price (highest price of requests) equals  $P_0$ , thus the spread is  $2\delta$ . A market buy order of, say, size 5 would absorb the outstanding shares at  $P_0 + 2\delta$  and  $P_0 + 3\delta$ , increasing both ask price and spread by  $2\delta$ . A market sell order of the same size would absorb the best three price levels' outstanding shares, and (since there is no share at  $P_0 - 3\delta$  outstanding) both decrease the bid price and increase the spread by  $4\delta$ . The picture is similar to the one in Preis (2010).

shares, market side}) highly irregular.

### 2.2.1 LOB-implied liquidity

Adding up the demand (supply) in the market at a given relative price, we obtain the cumulative volume.

**Definition 2.1** (Cumulative volume, cumulative imbalance). Let  $v_t^{(s)}(d)$ ,  $d = 0, 1, 2, \dots$  be the volume in the book at a distance of  $d$  ticks from the best quote on market side  $s \in \{\text{bid}, \text{ask}\}$ . The cumulative volume (CV) at side  $s$ , tick  $d$  and time  $t$  is defined by

$$x_t^{(s)}(d) = \sum_{k=0}^d v_t^{(s)}(k),$$

and the cumulative imbalance at tick  $d$  and time  $t$  by

$$x_t^{(imb)}(d) = x_t^{(ask)}(d) - x_t^{(bid)}(d).$$

Cumulative volume on the sell (buy) side of the market is a strictly non-increasing (strictly non-decreasing) step function with respect to the price. Steep cumulative volume curves with respect to  $d$  imply high liquidity since a high number of shares can be bought or sold without moving the price very much. A perfect market may be characterized by infinite liquidity, i.e., by a situation where the available number of shares at the quotes can be matched with market orders of any size. The two CV functions for buy and sell orders plus bid and ask quotes contain the complete available information about the LOB at a given time  $t$ .

If the volumes in the book are weighted by their prices, the resulting quantity is called *depth*. Therefore, the depth of an order book is a hybrid between a liquidity measure and a measure of the price of an asset. Over short periods of time during which the price stays almost constant, CV and depth contain the same information.

Another, equally informative liquidity measure is the price of instantaneously buying (selling)  $n$  shares (Bowsher, 2004).

**Definition 2.2** (Average cost per share). The average cost per share of a market order of size  $n$  at time  $t$  is given by

$$AC_t^{(s)}(n) = \frac{1}{n} \left[ \left( n - x_t^{(s)}(c) \right) (P_t^{(s)} + c + 1) + \sum_{k=0}^c (P_t^{(s)} + k) v_t^{(s)}(k) \right],$$

where  $P_t^{(s)}$  is either the bid or the ask quote, and

$$c := \sup \{ d : x_t^{(s)}(d) \leq n, d = 0, 1, 2, \dots \},$$

provided that  $n$  is larger than the volume at  $d = 0$  (otherwise,  $c := -1$ ).

Note that  $s = \text{bid (ask)}$  gives the price of a sell (buy) market order. There is a one-to-one relationship between CV and the average cost per share as a function of the offered/requested market order volume  $n$ , which has been analyzed in Gouriéroux et al. (1998) and Bowsher (2004).

Liquidity is high if AC curves are flat and low if they are steep. Moreover, the average cost per share can be rewritten as

$$AC_t^{(s)}(n) = P_t^{(s)} + \frac{c+1}{n} \left( n - \sum_{k=0}^c v_t(k) \right) + \sum_{k=0}^c kv_t(k).$$

Just as depth does, AC depends on the (nonstationary) price process and is therefore a hybrid between a liquidity measure and the price. If we are interested in liquidity only, measured as trading cost per share apart from its (bid or ask) price  $P_t^{(s)}$ , we may measure AC in ticks, i.e. simply drop  $P_t^{(s)}$  in the above expression.

Closely related to AC, Gomber and Schweickert (2002) define the *exchange liquidity measure* (XLM). We note that our version slightly differs from theirs.

**Definition 2.3** (Exchange liquidity measure). *The exchange liquidity measure of two market orders (one buy, one sell) of equal size  $n$  at time  $t$  (a “roundtrip trade”) is given by*

$$XLM_t(n) = XLM_t^{(buy)}(n) + XLM_t^{(sell)}(n),$$

where

$$XLM_t^{(buy)}(n) = \frac{AC_t^{(ask)}(n) - P_t^{(ask)}}{P_t^{(ask)}},$$

$$XLM_t^{(sell)}(n) = \frac{AC_t^{(bid)}(n) - P_t^{(bid)}}{P_t^{(bid)}}$$

are the average costs for buy (sell) orders relative to the respective quotes.

Some recent studies make use of these intrinsically curve-valued liquidity measures. However, instead of accounting for the functional nature of the data, each of the liquidity curves is evaluated at one specific *arbitrary* value of their argument. Examples are Engle et al. (2012), where liquidity of treasury bills is observed at five price tiers behind the quotes, and van Kervel (2015), where measures called *DepthAsk* and *DepthBid* are used which are quite similar to CV. Gomber et al. (2015) use the XLM in the same manner. Gouriéroux et al. (1998), Bowsher (2004) in what would later become Bowsher and Meeks (2008), and Härdle et al. (2012b) capture full LOB-implied liquidity as we do. However, they focus solely on the curves’ dynamics, while we are the first to model both the curves and their impact on the price process.

Our measure of choice for LOB-implied liquidity will be cumulative volume throughout this thesis. As compared to cumulative depth, it measures the number of supplied

shares (liquidity) and not a hybrid between liquidity and the price. This is especially important as we seek to disentangle the variation of the price and the variation of liquidity. Moreover, in contrast to the price, CV can be expected to be stationary, as the number of shares stays (at least approximately) constant over time. A major disadvantage is the incomparability of liquidity curves between different stocks.<sup>2</sup> Finally, as the price is roughly constant over short periods of time, depth and volume may be expected to contain roughly the same information in many situations.

It has already been explained that CV contains the same information as AC. More precisely, CV corresponds to *AC measured in ticks*, whereas cumulative depth corresponds to *AC measured in monetary units*. A major advantage of CV when it comes to modeling is that it is defined for every relative price  $d \geq 0$ . At locations where the book is empty, i.e., for large  $d$ , it simply amounts to the cumulative number of shares in the book. In contrast,  $AC(n)$  is not defined (or may be defined to amount to infinity) for  $n$  larger than the volume in the book. Therefore, although containing the same information in principle, CV is preferable from a modeling perspective.

### 2.2.2 The XETRA LOB data

For all our empirical analyses, we make use of historical LOB data recorded by the XETRA system of the German stock exchange (Deutsche Börse AG), and concentrate on constituents of the DAX30 index, which can be expected to be quite liquid as compared to non-constituents of DAX30. Our sample covers the period from November 2008 to December 2010. In almost all analyses, we use data for three companies: MunichRe, Commerzbank, and Linde. At one instance, we also use data for the Deutsche Bank stock. As will be discussed in subsequent Chapters, the three stocks do not only represent different industries, they also differ in their trend and volatility profile during the observed period. Moreover, they have different tick sizes, which depend on the price. The data contain buy and sell *limit orders* and *trades* (i.e., market orders executed against active limit orders), both time-stamped to the microsecond ( $10^{-6}$  s) from the beginning of the sample to mid-2010, and even to the nanosecond ( $10^{-9}$  s) until the end of the sample. The available information for limit orders is

- The price of the limit order.
- The time of submission,  $t.start$ .
- The time of either cancellation or execution,  $t.end$ . Note that the information whether a limit order was cancelled or executed against a market order is not given directly.
- The size (in number of shares) of the order.

For trades, the same information is available, but by construction  $t.start$  equals  $t.end$ . From these data, the state of the LOB,  $LOB_t$ , can in principle be reconstructed

---

<sup>2</sup>However, comparability of depth between stocks is equally questionable.

for any time  $t$  during a trading day.  $LOB_t$  is given by the collection of all *active* limit orders, i.e., all orders for which  $t.start \leq t$  and  $t.end > t$  holds. For the stocks in our sample, the number of changes of  $LOB_t$  is typically a 6-digit number, the number of transactions a high 4- or low 5-digit number.

Which sampling scheme and, possibly, which preprocessing of  $LOB_t$  is chosen depends on the goal of the specific analysis. In the present Chapter, we use data from the whole sample period, taking *equidistant* snapshots every 20 minutes, that is, at a rather low frequency. The same version of the data is used in Chapter 3. In Chapter 4, we only use one month of data, taking snapshots at *random, non-equidistant* times determined by the frequency of price changes (*price durations*), obtaining up to roughly 70,000 snapshots during that month. Finally, in Chapters 5 and 6, we construct summary statistics (*realized measures*) for each trading day, based on equidistant snapshots taken every 1 or 5 minutes. Thus, in all cases we only use a *subset* of the full information that is provided by the data.

In the next Section, we formalize the notion of liquidity curves observed over time in the language of discrete-time functional stochastic processes.

## 2.3 Functional time series

### 2.3.1 The basic setting

In the following, assume that relative prices, originally observed on a tick grid, are rescaled to lie in  $[0, 1]$ ,  $0 = d_1 < \dots < d_J = 1$ , without loss of generality. Dropping the superscript “(s)” for this exposition, we assume the (possibly de-seasonalized) liquidity curves for each market side to be generated by a functional stochastic process in discrete time,  $(x_t)_{t \in \mathbb{Z}}$ , whose observations are elements of the Hilbert space  $L^2([0, 1])$  with inner product  $\langle x, y \rangle := \int_0^1 x(s)y(s)ds$ , so that  $x_t$  is square integrable.

The liquidity process exhibits a mean function,  $\mu(d) := E[x_t(d)]$ , and a (contemporaneous) covariance operator  $C(z)[\langle x - \mu, z \rangle(x - \mu)]$  with covariance kernel  $\Sigma(d, m) = \text{Cov}(x_t(d), x_t(m))$ . Mean and covariance kernel are constant over time. The covariance operator has the form

$$C(z)(d) = \int_0^1 \Sigma(d, m)z(m)dm,$$

describing the contemporaneous linear dependence of different locations (relative prices) of a liquidity curve. The quantities  $\mu$  and  $\Sigma$  can be viewed as the functional time series analogues to the unconditional mean vector and the lag-zero autocovariance matrix in vector autoregression. The covariance operator admits the spectral representation

$$C(z) = \sum_{j=1}^{\infty} \lambda_j \langle \phi_j, z \rangle \phi_j, \tag{2.1}$$

where the  $\lambda_j$  are the (strictly decreasing) eigenvalues and the  $\phi_j$  are the corresponding orthonormal eigenfunctions of  $C$ , i.e.,  $\int_0^1 \phi_j^2(m) dm = 1$  and  $\int_0^1 \phi_k(m) \phi_j(m) dm = 0$ ,  $k \neq j$  holds. The  $\zeta_{j,t} := \langle \phi_j, x_t \rangle$  are called the scores or loadings of the  $j$ -th eigenfunction on liquidity at time  $t$ .

Based on the spectral representation, liquidity curves can then be represented via the Karhunen-Loève (KL) decomposition,

$$x_t(d) = \mu(d) + \sum_{j=1}^{\infty} \zeta_{j,t} \phi_j(d),$$

which is also called the functional principal component (FPC) representation. The eigenvalues  $\lambda_j$  of the spectral representation are equal to the unconditional variances of the FPC scores  $\zeta_{j,t}$ . As the eigenvalues are strictly decreasing, the FPCs are sorted by their contribution to the  $x_t$ 's (unconditional) variation. This gives rise to the  $K$ -truncated FPC representation

$$x_t(d) = \mu(d) + \sum_{j=1}^K \zeta_{j,t} \phi_j(d) + v_t(d),$$

where  $v_t(d) = \sum_{j=K+1}^{\infty} \zeta_{j,t} \phi_j(d)$  is the truncation error.

In practice, we are interested in approximating the curves using such a truncation. The smallest number of components,  $K$ , necessary to explain a certain proportion (say 99%) of the curves' total variation is called the *effective dimension* of the liquidity process.

In many applications of functional data analysis, either the observation grid is irregular or even sparse, or some additional measurement error is present in the observed  $x_t$ . In all these cases,  $x_t$  can not directly be observed but must be estimated, usually in a nonparametric way where smoothness assumptions are imposed. For our LOB data, however, none of these problems is present: The relative price grid is dense and even equidistant, there are no missing values, and the measurement can be considered to be exact. Nevertheless, it may pay to assume intrinsically smooth underlying curves and employ a corresponding regularization approach in order to reduce estimation variance.

### 2.3.2 Estimation

At this point, the question arises how mean and eigenfunctions can be estimated, and how the empirical FPC scores  $\hat{\zeta}_{j,t}$  can be obtained. In practice, we observe discrete versions of the curves,  $x_t$ , on a grid  $d_1, \dots, d_J$  such that the realization of each curve is a  $J$ -dimensional vector. Raw estimates of mean function and covariance kernel are then given by

$$\hat{\mu}(d_j) = \frac{1}{T} \sum_{t=1}^T x_t(d_j), \quad d_j \in [0, 1],$$

and, with  $\hat{\boldsymbol{\mu}} := [\hat{\mu}(0) \hat{\mu}(d_2) \cdots \hat{\mu}(d_{J-1}) \hat{\mu}(1)]'$  and  $\mathbf{X}^c = [\mathbf{x}_1 - \hat{\boldsymbol{\mu}} \cdots \mathbf{x}_T - \hat{\boldsymbol{\mu}}]'$ ,

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T} \mathbf{X}^{c'} \mathbf{X}^c.$$

The eigenvalues and eigenfunctions of the raw covariance kernel  $\hat{\boldsymbol{\Sigma}}$  can then in practice be computed using standard software for singular value decomposition. Using these estimates, the empirical FPC scores  $\hat{\xi}_{j,t} = \int_0^1 (x_t(m) - \hat{\mu}(m)) \hat{\phi}_j(m) dm$  can be obtained by numerical integration. An estimator for the covariance operator itself is then given by

$$\hat{C}(z) = \frac{1}{T} \sum_{t=1}^T \langle x_t - \hat{\mu}, z \rangle (x_t - \hat{\mu}).$$

Although the observed liquidity curves are step functions by construction, it may be convenient, especially with a view to the desired  $K$ -dimensional approximation, to interpret them as realizations of a process generating intrinsically smooth curves. Given such an assumption, functional observations are noisy versions of smooth underlying curves,  $x_t(d) = \tilde{x}_t(d) + e_t(d)$  say, where  $\tilde{x}_t(d)$  is the true underlying realization at  $t$ , and  $e_t$  is a measurement error with mean zero at all curve locations, independent of measurement errors for other curves (so-called ‘‘classical’’ measurement error, Carroll et al. (2006)), and independent between different locations of the same curve. The KL representation of  $x_t(d)$  then reads

$$x_t(d) = \mu(d) + \sum_{k=1}^{\infty} \xi_{k,t} \phi_k(d) + e_t(d),$$

that is, a sum of  $\tilde{x}_t(d)$ 's KL representation plus measurement error.

The measurement error does not affect consistency of the estimator of the mean curve, the variance of the estimator, however, increases. In contrast, the raw covariance estimator introduced above,  $\hat{\boldsymbol{\Sigma}} = T^{-1}(\mathbf{X}^c)' \mathbf{X}^c$ , has mean

$$E[\hat{\boldsymbol{\Sigma}}] = \begin{cases} \text{Var}(\tilde{x}_t(d_j)) + \text{Var}(e_t(d_j)) & \text{on the diagonal,} \\ \text{Cov}(\tilde{x}_t(d_j), \tilde{x}_t(d_l)), j \neq l, & \text{i.e., off the diagonal,} \end{cases}$$

where  $\text{Var}(e_t(d)) =: \sigma_e^2(d)$  is the variance function of the measurement error. If the intrinsic curves  $\tilde{x}_t(d)$  are assumed to be smooth, this translates to a regularizing smoothness assumption for the covariance kernel and, therefore, its eigenfunctions.

Our strategy is to use the raw mean function, which we find to be already quite smooth in all of our applications, see Chapters 3 to 6. In case of the covariance operator, several estimation methods have been proposed in the literature. We adopt the recent approach of Xiao et al. (2013), Xiao et al. (2013) which proceeds as follows. Stacking the entries of the raw covariance matrix into a vector, we define  $\text{vec}(\hat{\boldsymbol{\Sigma}}) =: \hat{\boldsymbol{\sigma}}$ . Then we use univariate smoother matrices of the form

$$\mathbf{S}_i = \mathbf{B}_i(\mathbf{B}_i' \mathbf{B}_i + \lambda_i \mathbf{D}_i' \mathbf{D}_i)^{-1} \mathbf{B}_i', \quad i = 1, 2,$$

where  $B_i$  is a matrix of B-spline basis functions evaluated at the spline's knots,  $D_i$  is a differencing matrix of order  $i$ , and  $\lambda_i \geq 0$  is a smoothing parameter. This is the P-spline approach due to Eilers and Marx (1996). Using these ingredients, the smoothed, stacked version of the raw covariance function can be expressed as

$$\hat{\sigma}^{\text{smooth}} = (\mathbf{S}_2 \otimes \mathbf{S}_1) \hat{\sigma},$$

i.e., by a tensor product of the univariate directions of the covariance. As the covariance function is symmetric, the same set of knots and degree of smoothing can be used for both directions, and there is only one smoothing parameter  $\lambda$  to be chosen. This is accomplished by using generalized cross validation (GCV) which enjoys favorable asymptotic properties in this setting which are comparable to the usual alternative used in the FDA literature, local linear smoothing (Yao et al., 2005b). At the same time, the sandwich smoother is very fast computationally. A further alternative would be kernel smoothing (Staniswalis and Lee, 1998). Moreover, we found that in our applications eigenfunction estimates based on smoothed and raw covariance kernel estimates do not differ much, so that the smoothing step is not indispensable.

The estimation procedure described so far was originally designed for the case of *iid* data. Recently, Hörmann and Kokoszka (2010) introduced a stationarity concept for functional time series, called  $L^p$ - $m$ -approximability.

**Definition 2.4** ( $L^p$ - $m$ -approximability, Hörmann and Kokoszka (2010)). *A sequence  $(x_t)_t \in L^p$  is called  $L^p$ - $m$ -approximable if each  $x_t$  admits the nonlinear MA( $\infty$ ) representation*

$$x_t = f(\varepsilon_t, \varepsilon_{t-1}, \dots), \quad (2.2)$$

where  $\varepsilon_i$  are i.i.d. mean zero functional innovations in some space  $\mathcal{E}$ , and  $f$  maps the innovations from  $\mathcal{E}^\infty$  to  $L^p([0, 1])$ .

Let the sequence  $(\varepsilon'_i)_i$  be an independent copy of  $(\varepsilon_i)_i$ , and replace the  $\varepsilon_i$  in (2.2) by  $\varepsilon'_i$  for  $i \geq m$ , forming the representation

$$x_t^{(m)} = f(\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-m+1}, \varepsilon'_{t-m}, \varepsilon'_{t-m-1}, \dots). \quad (2.3)$$

Then,

$$\sum_{m=1}^{\infty} \left( E \|x_m - x_m^{(m)}\|^p \right)^{1/p} < \infty \quad (2.4)$$

holds.

The interpretation of  $L^p$ - $m$ -approximability is that the impact of innovations in the distant past on  $x_t$  is so small that they can be replaced by different innovations with the same stochastic properties. While being more general, this is basically the same property as the MA( $\infty$ ) representation of causal (V)AR(MA) processes in classical time series analysis. In other words, it holds for models with sufficiently *short* memory. Moreover, being defined in terms of moments, it is a notion of *weak* stationarity.

Based on the assumption of  $L^p$ - $m$ -approximability, Hörmann and Kokoszka (2010) are able to prove that  $\hat{\mu}$  and  $\hat{C}(z)$  are  $\sqrt{T}$ -consistent estimators of the true mean function and covariance operator, respectively — provided the data-generating process is assumed to be  $L^2$ - $m$ -approximable in case of  $\hat{\mu}$ , and  $L^4$ - $m$ -approximable in case of  $\hat{C}(z)$ , respectively. Throughout, we assume our liquidity process(es) to be stationary in this sense.

### 2.3.3 Dynamic dimension reduction

An implication of dimension reduction using the KL-representation, i.e., based on the contemporaneous covariance structure of the curves, is the fact that the multivariate score series  $(\xi_{1,t}, \xi_{2,t}, \dots)_t$  is contemporaneously uncorrelated,  $\text{Cov}(\xi_{j,t}, \xi_{k,t}) = 0$ ,  $j \neq k$ , but may be cross-correlated at leads and lags,  $\text{Cov}(\xi_{j,t}, \xi_{k,s}) \neq 0$ ,  $t \neq s$ . In other words, FPCA as defined above is a static procedure which ignores the temporal dependence of curves when constructing the FPC weights or *scores*. As a consequence, dynamic factor models built upon the KL representation, presented in more detail in Section 2.5, in general must employ multivariate time series methodology to account for all dynamic properties.

Drawing on ideas first proposed in Brillinger (1981) for the vectorial case, Hörmann et al. (2015) put forth an alternative, termed *dynamic functional principal components analysis* (DFPCA).<sup>3</sup> DFPCA uses the whole spectrum of the functional stochastic process to obtain score series which are mutually uncorrelated at all leads and lags. In time-domain language, this amounts to exploiting the information provided by the entire set of autocovariance operators,

$$C_h(z)(d) = \int_0^1 \Gamma_h(d, m) z(m) dm,$$

$h \in \mathbb{Z}$ , where  $h = 0$  gives the contemporaneous covariance operator introduced above,  $\Gamma_0(d, m) = \Sigma(d, m)$ . This idea facilitates the use of *univariate* time series models to represent the curves' dynamics. In the vectorial case, the model is known as the *generalized dynamic factor model* (GDFM) due to Forni et al. (2000). For practical purposes, an analogue of the one-sided GDFM of Forni et al. (2005), which only employs *past* observations to construct the score series, would be more appropriate. However, these indicated generalizations are yet to be developed in the context of FTSA.

In spite of the sketched advantages of DFPCA, we expect the approach to have some drawbacks. The estimation of the functional spectrum, put in time-domain language, amounts to an implicit estimation of all (or, in practice, a large number of) autocovariance operators from a finite number of observations. This induces a trade-off between additional estimation variance and the potential benefits of accounting for serial dependence. Moreover, as each DFPCA score does not only depend on  $x_t$ , but also on temporally adjacent liquidity curves, not only estimation but also *interpretation*

<sup>3</sup>See also Panaretos and Tavakoli (2013), moreover Hu and Tsay (2014) for a similar idea in the context of conditional covariance matrices.

becomes less straightforward.

For these reasons, we exclusively make use of the *static* version of FPCA, the KL representation, in the remainder of the thesis.

## 2.4 The LOB as functional time series

To illustrate FPCA for liquidity curves, we consider data for MunichRe, Commerzbank, and Linde, observed every 20 minutes during continuous trading in the period 2008-11-03 to 2010-12-30. Raw liquidity curves for market side  $s$ , day  $t$ , and intraday time  $i$ ,  $\tilde{x}_{t,i}^{(s)}$ , which are strictly non-decreasing step functions, are diurnally adjusted via

$$x_{t,i}^{(s)}(d) = \frac{\tilde{x}_{t,i}^{(s)}(d)}{\hat{v}_i^{(s)}(d)},$$

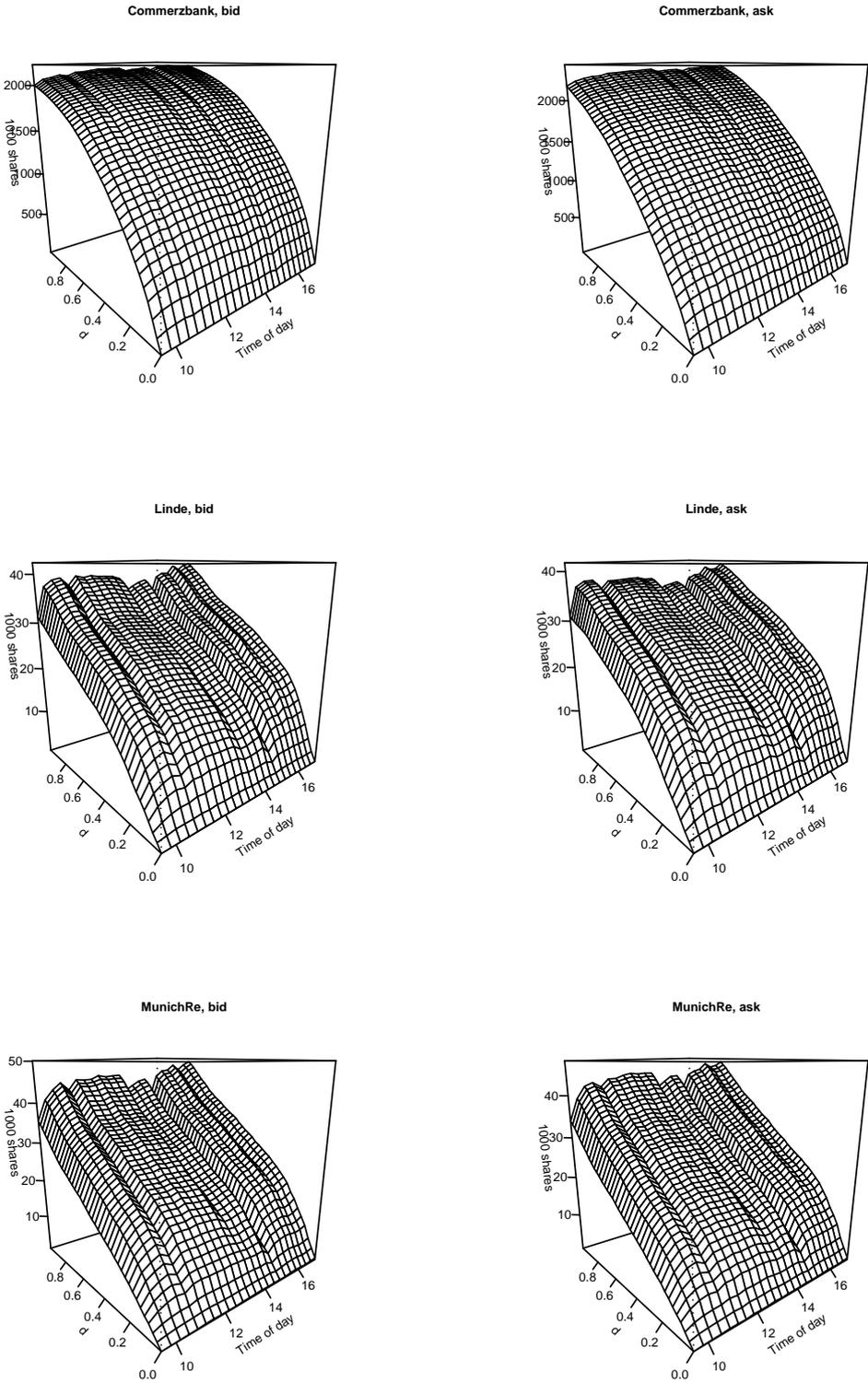
where  $\hat{v}_i^{(s)}(d)$  is the estimated diurnal pattern at the relative price  $d$ . The diurnal pattern is assumed to be deterministic as it is standard practice in modeling high-frequency trading dynamics (Hautsch, 2012). The same data are used in Chapter 3. A motivation of the sampling frequency as well as a detailed description of the data and their preprocessing can be found in Section 3.2.4 where it is put in a more suitable context.

The estimated diurnal surfaces are shown in Figure 2.2. They behave quite smoothly over the trading day. However, there is (i) lower-than-average liquidity at the beginning of the trading day, and a slight dip after the midday auction at 1pm, moreover (ii) a tendency of the LOB-implied amount of liquidity to increase over the day, especially *deep* ( $d \gg 0$ ) in the book.

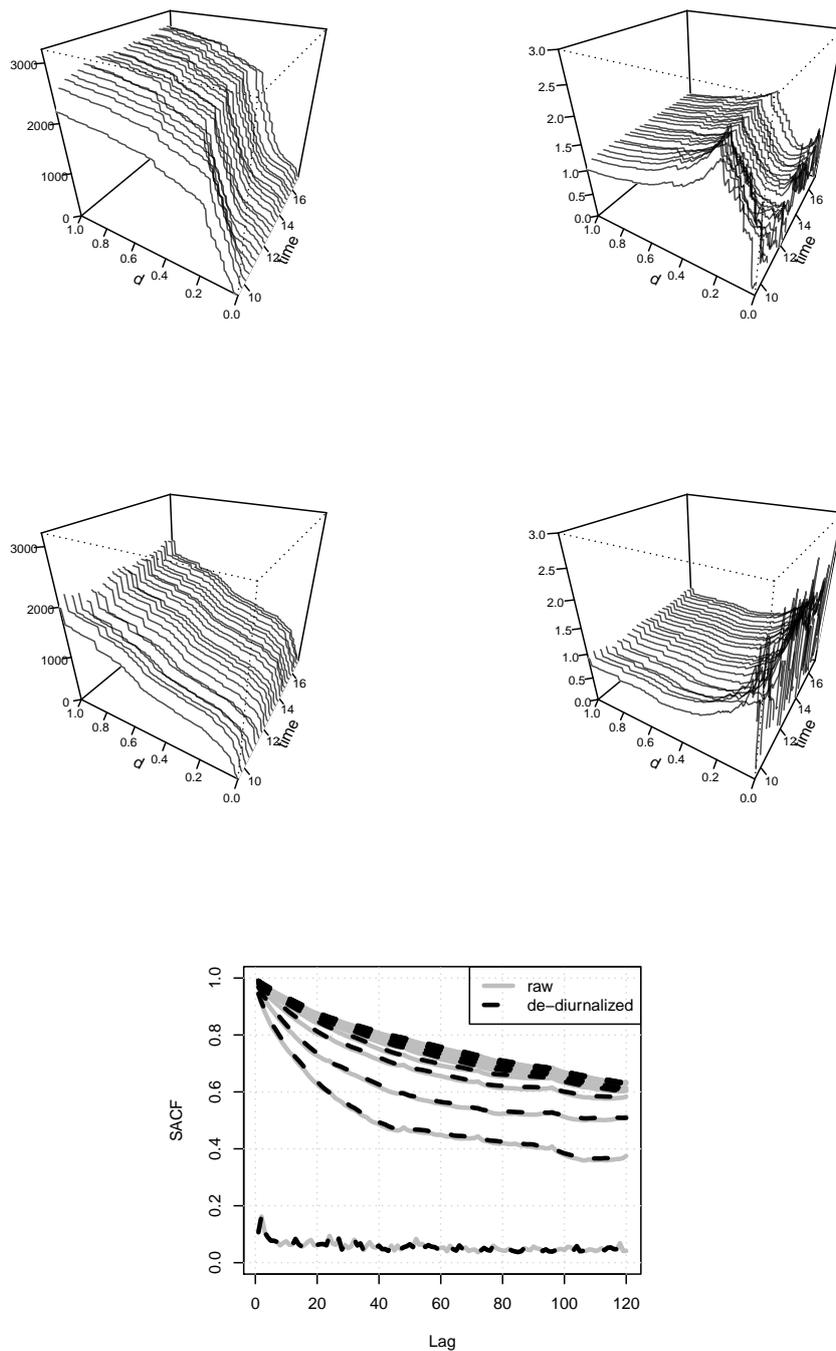
In the following, we model diurnally adjusted curves,  $x_{t,i}^{(s)}$ , but could in principle easily transform the curves back to their original scale. The autocorrelation structure of curves is only slightly affected by the transformation. The modes of variation, however, change distinctly: The original curves' variation increases with  $d$  by construction, whereas de-diurnalized curves measure the level of liquidity *relative* to the unconditionally expected level. Figure 2.3 contrasts raw and de-diurnalized curves for two exemplary trading days. Moreover, it is shown that the autocorrelation structure of liquidity is preserved.

Figures 2.4 and 2.5 show the results of a FPCA for the full sample exemplarily for the Commerzbank stock. The results for the other stocks are essentially the same. We find that the eigenfunctions (which are identifiable only up to the sign) are very similar for both market sides. The same holds for the *complexity* or effective dimension, represented by the eigenvalues, as in both cases three FPCs are needed to capture roughly 90 percent of the curves' variation. The question of *symmetry* of the market sides is addressed in Chapter 3. Figure 2.5 depicts the different modes of variation exemplarily for Commerzbank's ask curves. The findings for the bid side are quite similar. The straight horizontal line in all plots is the estimated mean curve,  $\hat{\mu}^{(ask)}(d)$ , of de-diurnalized ask liquidity, which amounts to 1 for all  $d$  by construction. In the

2 Limit order books as functional time series



**Figure 2.2:** Diurnal liquidity patterns for bid (left panel) and ask (right panel), and (top to bottom) Commerzbank, Linde, and MunichRe, measured in thousands of shares. The diurnal pattern is a function of intraday time and relative price  $d$ .



**Figure 2.3:** Top: Functional time series of ask liquidity curves for April 3, 2009 (top) and February 1, 2010 (bottom). In the left panel, the raw data (cumulative number of shares, measured in thousands, within a range of €0 to €2 from the quotes) are depicted, in the right panel the de-diurnalized versions of the same curves are shown. Bottom: Sample autocorrelations of raw and de-diurnalized ask curves at  $d = 0, 20, 40, \dots, 180, 200$  Cents from the quote, computed over the full sample. The maximum number of lags, 120, corresponds to five trading days. The lowest SACF trajectory corresponds to  $d = 0$ . The higher  $d$  is, the slower is the decay of the SACFs. The SACFs are virtually identical for raw and de-diurnalized data. The fact that we take 24 snapshots per day can be seen from slight heaps at  $h = 25, 49, \dots$  lags.

left panel,  $\hat{\mu}^{(ask)}(d)$  is shown along with  $\hat{\mu}^{(ask)}(d) \pm 1.96\sqrt{\hat{\lambda}_k^{(ask)}\hat{\phi}_k^{(ask)}(d)}$ ,  $k = 1, 2, 3, 4$ , where  $\hat{\lambda}_k^{(ask)}$  is the  $k$ th eigenvalue, and  $\hat{\phi}_k^{(ask)}(d)$  the  $k$ th estimated eigenfunction of the ask curves' estimated contemporaneous covariance operator. The rationale is as follows:  $\hat{\lambda}_k^{(ask)}$  equals the estimated variance of the  $k$ th FPC score,  $\zeta_{k,t}^{(ask)}$ . Simplifying, we assume the scores to be multivariate normal.<sup>4</sup> Then,  $\pm 1.96\sqrt{\hat{\lambda}_k^{(ask)}}$  are the estimated 2.5% (97.5%) quantiles of the marginal distribution of  $\zeta_{k,t}^{(ask)}$ . Therefore, the area between the grey curves in the left panel of Figure 2.5 can be interpreted as the respective main area of variation. For  $k = 1$ , it is found that liquidity curves are shifted almost parallelly, i.e., the first FPC can be interpreted as a *level* component. This level component dominates overall variation as already noted in the analysis of eigenvalues. The FPCs  $k = 2$  to 4 capture several aspects of the curves' variation in the left part of the domain of relative prices, i.e., near the quotes.<sup>5</sup> The much smaller areas between the curves in the second to fourth plots in the left panel reflect the smaller relative contributions of these features (smaller eigenvalues).

Exploiting the diagonal covariance structure of the multivariate score distribution, the right panel of Figure 2.5 shows the mean function and the *cumulative* modes of variation of the first  $k$  components,

$$\hat{\mu}^{(ask)}(d) \pm 1.96 \sum_{j=1}^k \sqrt{\hat{\lambda}_j^{(ask)}} \hat{\phi}_j^{(ask)}(d), \quad k = 1, 2, 3, 4.$$

All in all we find that curves (i) mainly vary in their *level*, and (ii) exhibit more and richer modes of variation near the quotes.

### 2.4.1 Identification of eigenfunctions

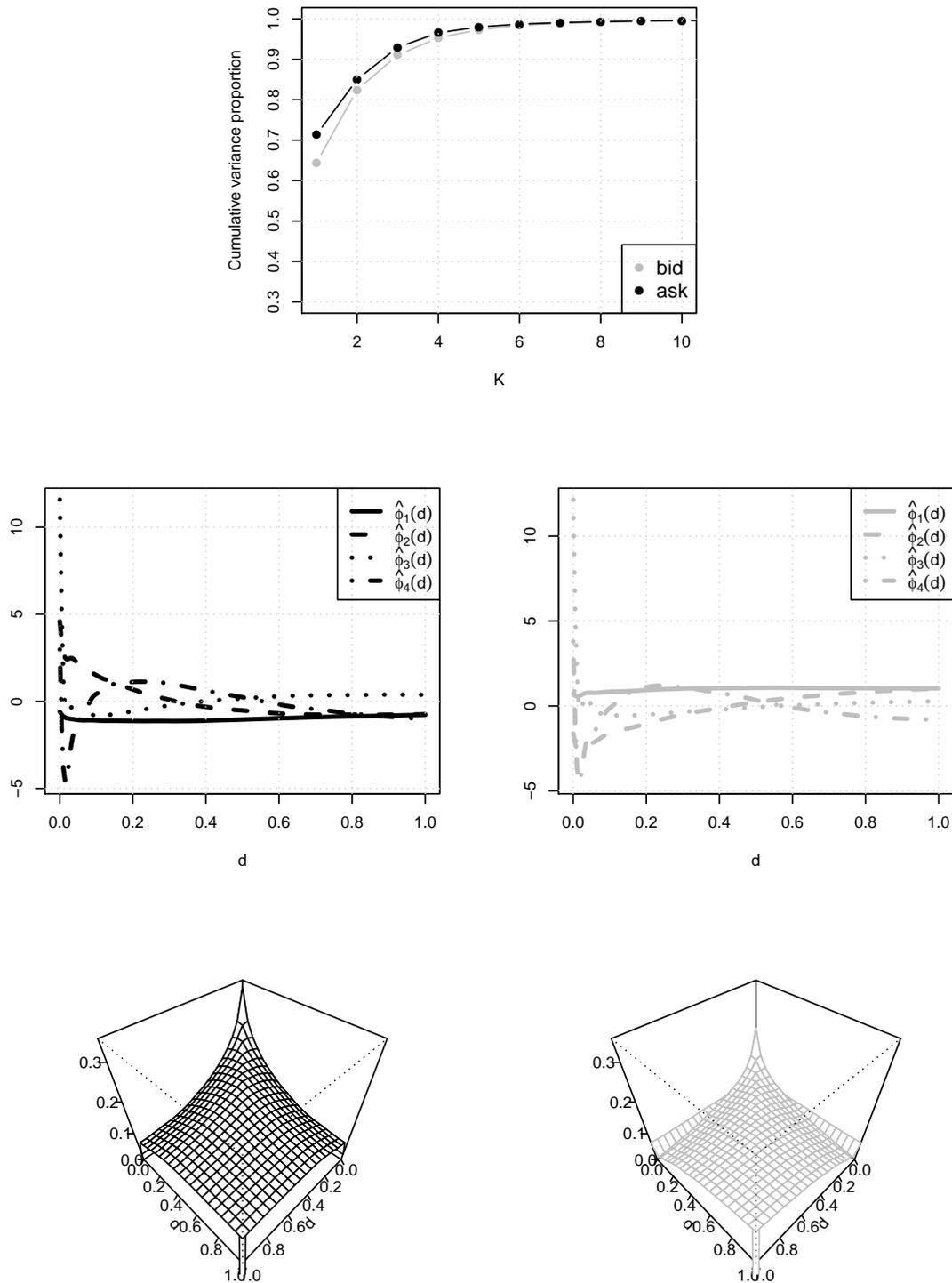
The theoretical results of Hörmann and Kokoszka (2010) explained above testify that dimension reduction based on the contemporaneous covariance structure, frequently used in *i.i.d.* settings, can also be employed in the context of weakly dependent time series. However, it remains to be explored how well the procedure works in finite samples if the factor dynamics are *complex*, i.e., exhibit highly interdependent score processes, and/or *persistent*, i.e., have near unit root dynamics.

We conducted several Monte Carlo experiments to answer these questions, not shown here in detail. The results suggest that in situations where the temporal dependence structure as well as the eigenstructure are chosen realistically, identification of eigenfunctions is very accurate.<sup>6</sup> Moreover, we find that identification of the original eigenfunctions is exacerbated when the eigenvalues decay very slowly with respect

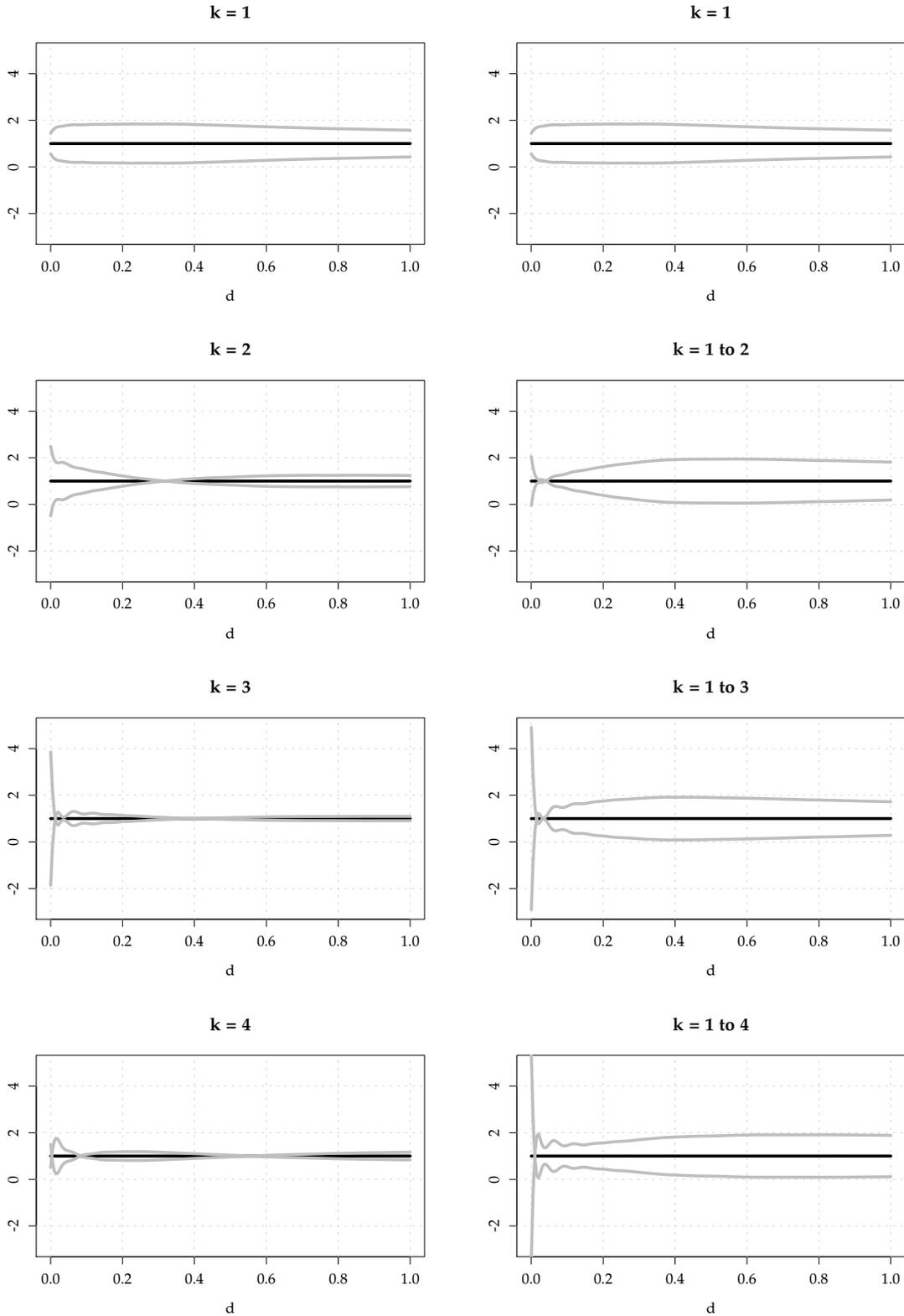
<sup>4</sup>This can be justified either by assuming  $x_t^{(ask)}$  to stem from a Gaussian process or based on asymptotic arguments.

<sup>5</sup>Recall that  $[0, 1]$  corresponds to  $[\in 0, \in 2]$ .

<sup>6</sup>In order to obtain a "realistic" simulation setting, we chose serial dependence and eigenstructure similar to their empirical counterparts. By employing empirical estimates, however, we presuppose that the identification has worked for these estimates, which leads to a somewhat tautological "evidence", as the quality of estimation is exactly what we seek to investigate.



**Figure 2.4:** Top panel: Cumulative eigenvalues, normalized by the sum over the first 30 eigenvalues, for de-diurnalized bid and ask curves of the Commerzbank stock.  $K = 3$  components explain more than 90 percent of curve variation in both cases. Center panel: First four estimated eigenfunctions for ask (left) and bid (right) curves for the Commerzbank data. The eigenstructure is apparently very similar for the two market sides. Bottom panel: Estimated covariance surfaces of bid (right) and ask (left), truncated for better visibility.



**Figure 2.5:** Modes of variation for Commerzbank ask liquidity curves. The straight horizontal line in all plots is the estimated mean curve,  $\hat{\mu}^{(ask)}(d)$ , of de-diurnalized ask liquidity which amounts to 1 for all  $d$  by construction. In the left panel,  $\hat{\mu}^{(ask)}(d)$  is shown along with  $\hat{\mu}^{(ask)}(d) \pm 1.96 \sqrt{\hat{\lambda}_k^{(ask)}} \hat{\phi}_k^{(ask)}(d)$ ,  $k = 1, 2, 3, 4$ . In the right panel, it is shown along with  $\hat{\mu}^{(ask)}(d) \pm 1.96 \sum_{j=1}^k \sqrt{\hat{\lambda}_j^{(ask)}} \hat{\phi}_j^{(ask)}(d)$ ,  $k = 1, 2, 3, 4$ .

to  $k$ , which is in line with the findings of previous studies, typically in *i.i.d.* settings. However, the approximation of curves using these inaccurately estimated basis functions is still quite good. Finally, we somewhat surprisingly find that, given a simple eigenstructure, strong serial dependence does not worsen the results much. Implicitly these issues are taken into account in the simulation study on the liquidity impact in MEM models in Section 3.4.

## 2.4.2 Uncertainty assessment

Finally, we investigate the extent of variation in the eigenfunction estimates. Stable eigenfunction estimates are desirable as will become apparent when dealing with the FunXL models proposed in Parts I and II of the thesis.

We employ the stationary bootstrap of Politis and Romano (1994) to quantify the estimation variability of eigenfunctions. As compared to the standard bootstrap for *i.i.d.* data, where observations are drawn independently, the idea is to draw *blocks* of observations in order to retain the serial dependence inherent in the true data-generating functional stochastic process. Moreover, in contrast to the *standard* block bootstrap, the stationary bootstrap uses random block lengths which are drawn from a geometric distribution. After some experiments, we choose the tuning parameter, the mean block size, to be 10. However, we find that the results are not very sensitive to this choice.

Figure 2.6 shows some exemplary results, again for Commerzbank's ask curves. We find that the eigenfunction estimates do not show much variation. The same holds for  $\hat{\mu}^{(s)}$  (not shown). Let us for one moment neglect this estimation uncertainty which we found to be small anyway. This amounts to pretending to know the true mean,  $\mu^{(s)}(d)$ , and eigenfunctions,  $\phi_k^{(s)}(d)$ ,  $k = 1, 2, \dots$ . Then, the (estimated) scores,

$$\hat{\xi}_{k,t}^{(s)} = \int_0^1 \left( x_t^{(s)}(m) - \hat{\mu}^{(s)}(m) \right) \hat{\phi}_k^{(s)}(m) dm,$$

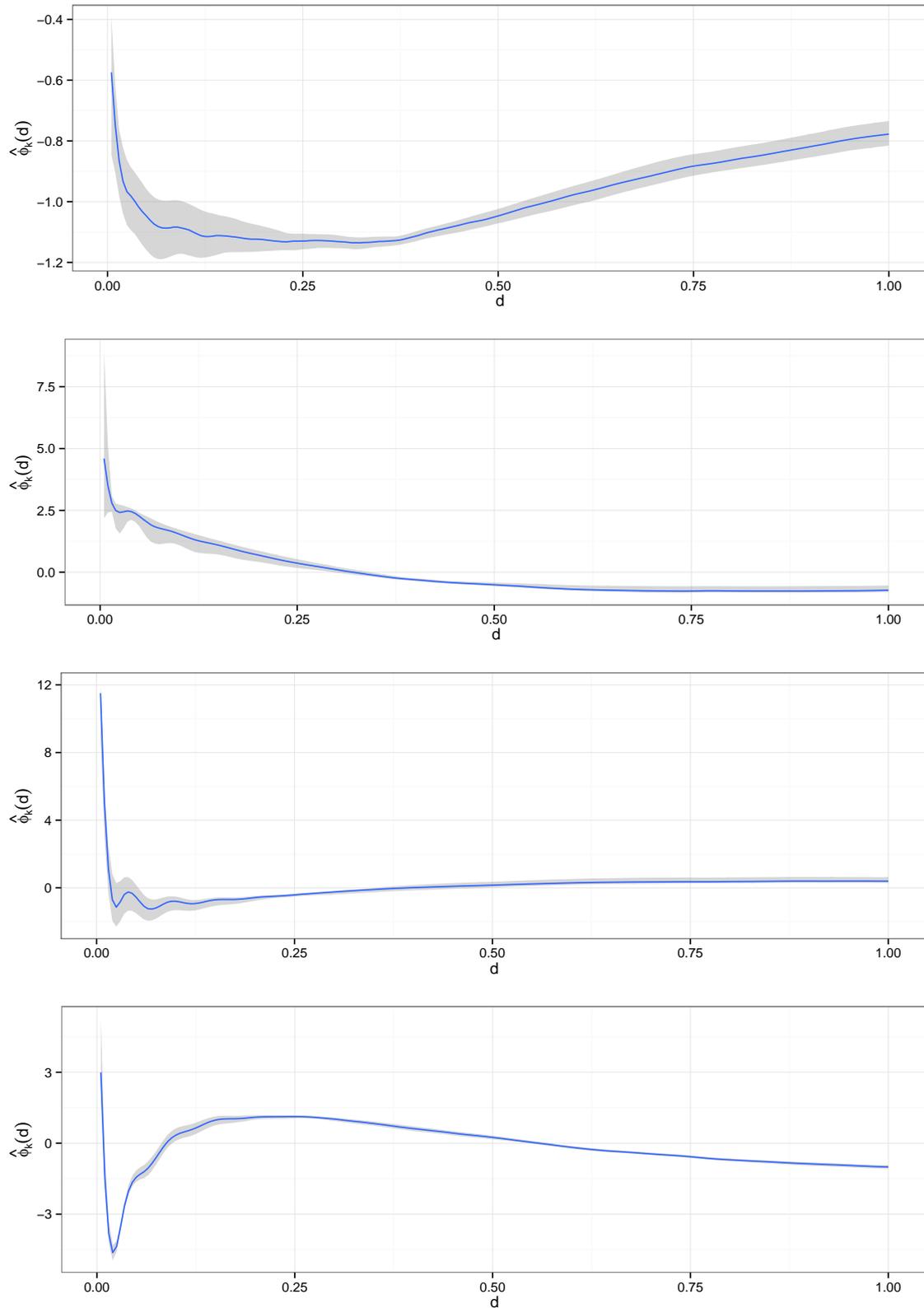
computed using the observed curves  $x_t^{(s)}(m)$ , can be treated as if they were directly observed. We return to this point in Chapters 3, 4, and 6 as it is crucial for estimation and model choice of FunXL models.

We also experimented with manually selected blocks, for example all observations during one month of trading, finding that eigenfunction estimates are also quite stable over time. Therefore, the assumption of a stationary underlying liquidity process appears to be appropriate, so that we turn to the application of stationary functional dynamic models in Section 2.5.<sup>7</sup>

---

<sup>7</sup>In addition to this rather informal investigation, a formal test of stationarity of functional time series could be performed, which has recently been proposed in Horváth et al. (2014).

## 2 Limit order books as functional time series



**Figure 2.6:** Uncertainty assessment for the estimated eigenfunctions of Commerzbank's ask curves, based on  $B = 100$  stationary bootstrap samples with mean block length of 10, results for  $k = 1$  to  $k = 4$ , from top to bottom.

## 2.5 Functional dynamic modeling of liquidity

Although this thesis is mainly concerned with the impact of liquidity on the price process, the question of the dynamics of liquidity itself arises naturally. In the following, the two existing approaches to autoregressive modeling of functional time series are introduced briefly and then applied to the limit order book snapshot data.

We distinguish two basic model classes, semiparametric factor (SPF) models and functional autoregressive (FAR) models. The basic idea of SPF models is to first project the curve-valued objects onto a finite-dimensional subspace and then model the dynamics of the projection. Forecasts can be obtained by inverting the projection, i.e. by reconstructing curves from vectors or scalars. The SPF approach considered here is based on FPCs. In contrast, FAR models exhibit an infinite-dimensional parameter, the autoregressive operator, which acts directly on the curve-valued objects of interest. However, the FAR model as introduced in Bosq (2000) employs a finite number of FPCs to represent the operator, meaning that the same kind of dimension reduction is used as in FPC-based SPF models, but at a different stage.

A prominent example of a SPF model is the dynamic Nelson-Siegel model of Diebold and Li (2006). In contrast to the FPC-based approach, the factors (and, hence, the number of factors) are predetermined — level, slope, and curvature. The model is estimated either by a two-step procedure — weights in the first step, dynamics in the second —, or in a joint state space approach. The latter is either accomplished by the Kalman filter (if Gaussianity is assumed) or by Bayesian techniques which allow for more flexible measurement error and innovation distributions. Another example with pre-determined factors, also in a state space setting, is the functional signal-plus-noise model of Bowsher and Meeks (2008), which was applied to AC curves in Bowsher (2004). This is to the best of our knowledge the first dynamic model for liquidity curves.

SPF models with unknown factor structure are developed in Fengler et al. (2007) and Park et al. (2009). An application of this model to liquidity curves can be found in Härdle et al. (2012b). There, the factor structure plus weights on the one hand, and the dynamics on the other hand, are estimated separately. The FPC-based approach propagated in this thesis can first be found in Hyndman and Shang (2009). Accompanied by a sound theoretical grounding, the model is re-introduced in Aue et al. (2015). To our knowledge, there are only two examples of models where the unknown factor structure (basis functions plus loadings) and the dynamics are estimated *jointly* in one step. The first example is Hays et al. (2012), where an EM algorithm is used, the second, using Bayesian estimation techniques, is Kowal et al. (2014). In contrast to SPF models, the literature on the FAR approach, which is the direct analogue to (V)AR models, can be overseen more easily. The FAR model was pioneered in Bosq (2000), further references are Kargin and Onatski (2008) and Besse et al. (2000). They have never been applied to liquidity curves before.

### 2.5.1 Semiparametric factor model

We have already indicated that many different SPF models have been proposed. To construct a semiparametric factor model, consider the FPC representation of the random curve  $x_t$ ,

$$\begin{aligned} x_t(d) &= \mu(d) + \sum_{j=1}^{\infty} \zeta_{j,t} \phi_j(d) \\ &\approx \mu(d) + \sum_{j=1}^K \zeta_{j,t} \phi_j(d). \end{aligned} \quad (2.5)$$

We assume the truncation parameter  $K$  to be chosen generously large, rendering the truncation error  $\sum_{j=K+1}^{\infty} \zeta_{j,t} \phi_j(d)$  negligible. The curves' dynamics are solely captured by the time-varying weights of the eigenfunctions which are collected in the  $K$ -dimensional vector  $\zeta_t = [\zeta_{1,t} \ \cdots \ \zeta_{K,t}]'$ .

For the dynamics of the FPC scores or *factors*  $\zeta_{1,t}, \dots, \zeta_{K,t}$  we use a VAR model,

$$\zeta_t = a_0 + \sum_{j=1}^p A_j \zeta_{t-j} + v_t, \quad (2.6)$$

where  $v_t$  is multivariate Gaussian white noise. All in all, we have a nonparametric first step, the dimension reduction, and a parametric second step, the dynamics. This is the same model as in Aue et al. (2015). Park et al. (2009) and Härdle et al. (2012b) use the same dynamics, but extract the factor series slightly differently, while Hyndman and Shang (2009) also use FPC-based factors, but only univariate autoregressive factor dynamics. We note that, although dimension reduction in this setting is a "linear" procedure (contemporaneously uncorrelated factors based on the covariance structure), the dynamics of the conditional mean may be allowed to be nonlinear. Another feature that may be desirable is conditional heteroskedasticity of  $v_t$ . However, it is unclear how such features translate back to the original, infinite-dimensional function space of liquidity curves.

#### Estimation and prediction

Estimation of the semiparametric dynamic factor model proceeds in two steps.

1. Estimate the factor structure as outlined in Section 2.3, yielding a  $K$ -variate time series of FPC scores or *factors*,  $\left(\hat{\zeta}_t\right)_{t=1,\dots,T}$ .
2. Fit a VAR model to the score series  $\left(\hat{\zeta}_t\right)_{t=1,\dots,T}$  by OLS.

We represent the resulting estimate as VAR(1), which also encompasses higher-order VAR models if  $\zeta_t$ ,  $A$ ,  $v_t$  are redefined appropriately (Lütkepohl, 2005). Moreover, we drop the *hats* stemming from estimation step 1, notationally pretending to

observe the FPC scores directly, while retaining those stemming from step 2, i.e., for all other quantities, yielding

$$\boldsymbol{\zeta}_t - \hat{\boldsymbol{\mu}}_{\boldsymbol{\zeta}} = \hat{\mathbf{A}} \left( \boldsymbol{\zeta}_{t-1} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\zeta}} \right) + \hat{\boldsymbol{v}}_t, \quad (2.7)$$

where  $\hat{\boldsymbol{\mu}}_{\boldsymbol{\zeta}} = \mathbf{0}$ , i.e., the estimated unconditional mean vector of the FPC scores is zero by construction of the scores.

Then, the usual recursive minimum MSE predictor can be applied to obtain  $h$ -step-ahead forecasts of the process,

$$E[\boldsymbol{\zeta}_{t+h} | \boldsymbol{\zeta}_t] = \hat{\mathbf{A}}^h \boldsymbol{\zeta}_t.$$

Using the estimated eigenfunctions and mean function for the liquidity curves, a  $h$ -step-ahead forecast for the original liquidity curves is then obtained by

$$E[x_{t+h} | x_t] = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\phi}}' \left( \hat{\mathbf{A}}^h \boldsymbol{\zeta}_t \right),$$

where  $\hat{\boldsymbol{\phi}}$  is the  $K$ -dimensional “vector” stacking the  $K$  estimated eigenfunctions  $\hat{\phi}$ .

If the homoskedastic normality assumption for the innovations is taken literally, we also obtain prediction intervals for the scores, which can in turn be used to construct prediction intervals for the liquidity curves — thereby ignoring estimation uncertainty for mean and eigenfunctions, i.e., from step 1.

## 2.5.2 Functional autoregressive model

The direct analogue of an autoregressive model of order one is the linear FAR(1) model,

$$x_t(d) - \mu(d) = \int_0^1 \Psi(c, d) (x_{t-1}(c) - \mu(c)) dc + \varepsilon_t(d), \quad (2.8)$$

where  $\varepsilon_t(d)$  is a functional white noise process, i.e., which has mean zero for all  $d$ , and is serially uncorrelated. Additionally, but not necessarily,  $\varepsilon_t(d)$  may be assumed to have a homokedastic and/or diagonal contemporaneous covariance operator.<sup>8</sup> Moreover,

$$\boldsymbol{\Psi}(x) = \int_0^1 \Psi(c, d) x(c) dc$$

---

<sup>8</sup>The model is often used to describe a continuous-time process which is cut into smaller pieces, for example cumulative intraday return trajectories sampled each trading day. Then a natural innovation process is the (contemporaneously heteroskedastic) Brownian bridge as it is deterministically zero at  $d = 0$  and  $d = 1$ .

denotes a linear autoregressive operator such that  $E[x_t|x_{t-1}] = \Psi(x_{t-1})$  provided that  $x_t$  is de-meaned. The model can then be more compactly rewritten as

$$x_t - \mu = \Psi(x_{t-1} - \mu) + \varepsilon_t. \quad (2.9)$$

We emphasize that  $\mu$  is a functional parameter, and  $x_t, \varepsilon_t$  are functional random processes. This model is due to Bosq (2000). It is easily seen that it is the direct analogue to the scalar AR(1) and the vectorial VAR(1) model. Accordingly, in analogy to the VAR case FAR models of order  $p$  can be constructed using the FAR(1) representation,

$$\mathbf{x}_t - \boldsymbol{\mu} = \Psi(\mathbf{x}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t, \quad (2.10)$$

where  $\mathbf{x}_t, \boldsymbol{\mu}$ , and  $\boldsymbol{\varepsilon}_t$  are constructed by sticking together  $p$  functional quantities,

$$\begin{aligned} \mathbf{x}_t &= (x_t \ x_{t-1} \ \cdots \ x_{t-p+1}) \\ \boldsymbol{\mu} &= (\mu \ \mu \ \cdots \ \mu) \\ \boldsymbol{\varepsilon}_t &= (\varepsilon_t \ 0 \ \cdots \ 0) \end{aligned} \quad (2.11)$$

and the autoregressive kernel operator is given by

$$\Psi = \begin{bmatrix} \Psi_1 & \Psi_2 & \cdots & \Psi_{p-1} & \Psi_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix}. \quad (2.12)$$

where  $\mathbf{I}$  denotes the identity operator (Bosq, 2000). Note that  $x_t(\mu, \varepsilon_t)$  do not denote vectors of functions as before, but one function mapping  $[0, 1] \rightarrow \mathbb{R}$ . The domain of  $d, [0, 1]$ , is thereby partitioned into  $p$  sections of equal length and the elements of  $\mathbf{x}_t(\mu, \varepsilon_t)$  are rescaled appropriately. In the following, however, we confine ourselves to the FAR(1) model.

To be stationary and causal, the autoregressive operator must satisfy  $\|\Psi\| < 1$ , where

$$\|\Psi\|^2 = \int_0^1 \int_0^1 \Psi^2(c, m) dc dm.$$

### Estimation and prediction

To estimate  $\Psi$ , Bosq (2000) proposes a Yule-Walker-type estimator. The subsequent presentation follows Hörmann and Kokoszka (2012) closely. Throughout, we assume de-meaned  $x_t$ , i.e.,  $E[x_t(d)] = 0$  for all  $d$ .

In the univariate scalar AR(1) case,  $x_t = \psi x_{t-1} + \varepsilon_t$ , where  $\varepsilon_t$  is white noise and  $|\psi| < 1$ , the Yule-Walker estimator of  $\psi$  is given in terms of the estimated autocovariance function at lags 0 and 1,  $\hat{\psi} = \gamma_1/\gamma_0$ . In the functional case, the estimator looks just the same in principle, but the autocovariances are replaced by autocovariance operators  $C_0$  (formerly simply written as  $C$ ) and  $C_1$ , where

$$C_h(z) := E[\langle x_t, z \rangle x_{t+h}], \quad (2.13)$$

so that the functional Yule-Walker equation of the FAR(1) model can be written

$$C_1 = \Psi C_0. \quad (2.14)$$

Thus, just as in the scalar or vectorial case, the estimator is obtained by inverting the contemporaneous covariance,

$$\Psi = C_1 C_0^{-1}. \quad (2.15)$$

Using the spectral representation (2.1) and exploiting the orthonormality of the eigenfunctions, the inverse operator is obtained by simply inverting the eigenvalues,

$$C_0^{-1}(z) = \sum_{j=1}^{\infty} \lambda_j^{-1} \langle \phi_j, z \rangle \phi_j \quad (2.16)$$

However, as the eigenvalues go to zero for  $j \rightarrow \infty$ ,  $C_0^{-1}$  is unbounded. Therefore, to render estimation feasible in practice, we truncate the representation, using only the first  $K$  eigencomponents. By plugging in estimated eigenfunctions and eigenvalues, we arrive at the empirical inverse contemporaneous covariance operator

$$\hat{C}_0^{-1}(z) = \sum_{j=1}^K \hat{\lambda}_j^{-1} \langle \hat{\phi}_j, z \rangle \hat{\phi}_j \quad (2.17)$$

while using the estimator

$$\hat{C}_1(z) = \frac{1}{T-1} \sum_{t=1}^{T-1} \langle x_t, z \rangle x_{t+1} \quad (2.18)$$

for the order 1 autocovariance operator. In the final estimator of  $\psi$  due to Bosq (2000), the curves  $x_t$  are approximated by their first  $K$  FPCs,  $x_t \approx \sum_{k=1}^K \hat{\zeta}_{k,t} \hat{\phi}_k$ , leading to

$$\hat{\psi}_K(z) = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{k=1}^K \sum_{j=1}^K \hat{\lambda}_k^{-1} \langle z, \hat{\phi}_k \rangle \hat{\zeta}_{k,t} \hat{\zeta}_{j,t+1} \hat{\phi}_j \quad (2.19)$$

with kernel

$$\hat{\psi}_K(c, d) = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{k=1}^K \sum_{j=1}^K \hat{\lambda}_k^{-1} \hat{\zeta}_{k,t} \hat{\zeta}_{j,t+1} \hat{\phi}_k(c) \hat{\phi}_j(d). \quad (2.20)$$

All quantities needed to estimate the autoregressive kernel,  $\hat{\psi}_K(c, d)$ , are obtained via FPCA. The serial dependence of the curves is basically captured by means of the covariance of the FPC scores which is measured by the product  $\hat{\zeta}_{k,t} \hat{\zeta}_{j,t+1}$ , as can be seen from (2.20). Moreover, these contributions are weighted by the inverse eigenvalues of the components. As a consequence, choosing  $K$  large assigns (potentially too) much weight to components with small explained variance.

Turning to forecasting, we note that in our application the curves are observed on a regular grid of length  $J = 201$ ,  $\mathbf{x}_t = (x_t(0) \ x_t(d_2) \cdots x_t(d_{J-1}) \ x_t(d_J))'$ . The same holds for the estimated mean,  $\hat{\boldsymbol{\mu}}$ . We use the same grid for evaluating  $\hat{\psi}_K(c, d)$ , which is then a  $J \times J$  matrix. Denote this matrix by  $\hat{\boldsymbol{\psi}}_{K,J}$ . Then,  $h$ -step-ahead forecasts are basically obtained in the same way as for VAR models, rescaled by the number of grid points  $J$ ,

$$E[\mathbf{x}_{t+h} | \mathbf{x}_t] = \hat{\boldsymbol{\mu}} + J^{-2} \hat{\boldsymbol{\psi}}_{K,J}^h (\mathbf{x}_t - \hat{\boldsymbol{\mu}}). \quad (2.21)$$

### 2.5.3 Empirical results

We fit FPC-based SPF models and FAR models to the XETRA data for Commerzbank, MunichRe, and Linde, using  $K = 1, \dots, 8$  functional principal components, respectively. As this analysis is only supposed to be a preliminary one with the new models in Parts I and II in view, we confine ourselves to models with  $p = 1$ , although a higher-order model will certainly be slightly more appropriate. Concerning the SPF models, we find the score dynamics to be pretty diagonal. For instance, for a VAR(1) model of Linde's ask curves and  $K = 5$ , we obtain

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.87 & 0.28 & 0.20 & 0.19 & -0.17 \\ 0.04 & 0.35 & 0.19 & 0.13 & -0.14 \\ 0.01 & 0.07 & 0.55 & 0.23 & -0.01 \\ 0.01 & 0.01 & 0.13 & 0.38 & -0.21 \\ 0.00 & -0.03 & 0.00 & -0.13 & 0.40 \end{bmatrix}.$$

The result suggests that it is reasonable to confine to univariate AR(1) score dynamics in the SPF models, resembling the original FPC-based SPF model (Hyndman and Shang, 2009). We abbreviate these models SPF-AR(1). Note that all SPF models that we fitted imply clearly stationary dynamics.

The estimated autoregressive kernels for Linde (ask) and  $K = 1, \dots, 4$  are shown in Figure 2.7. For  $K = 1$ , it is positive everywhere and relatively flat, while for  $K \geq 1$ , the positive autocorrelation near the quotes becomes more pronounced — even to such a degree that the implied model is no more stationary.

Nevertheless, we evaluate all estimated models, SPF-AR(1) and FAR(1) with  $K = 1, \dots, 8$ , in an out-of-sample forecast exercise, regardless of possibly nonstationary parameter constellations. As the evaluation of the FAR(1) kernel at the  $J^2$  grid points is computationally intense, the models are estimated only once. We use two training periods, 2008-11-03 to 2009-12-30 and 2008-11-03 to 2010-06-30, and compute forecasts for (the remainder of) all observations in 2010. As the results are qualitatively the same for both situations, we report only the first one. We employ the root mean integrated squared error (RMISE) as our loss function. Denoting the discretized point forecast by  $\hat{x}_{t+h} := E[x_{t+h}|x_t]$ , it is defined as

$$RMISE_h = \left( I^{-1} \sum_{i=1}^I \int_0^1 (x_{i+h}(m) - \hat{x}_{i+h}(m))^2 dm \right)^{1/2} \quad (2.22)$$

$$\approx \left( I^{-1} J^{-1} \sum_{i=1}^I \sum_{j=1}^J (x_{i+h}(m_j) - \hat{x}_{i+h}(m_j))^2 \right)^{1/2}, \quad i \in \mathcal{I}, \quad (2.23)$$

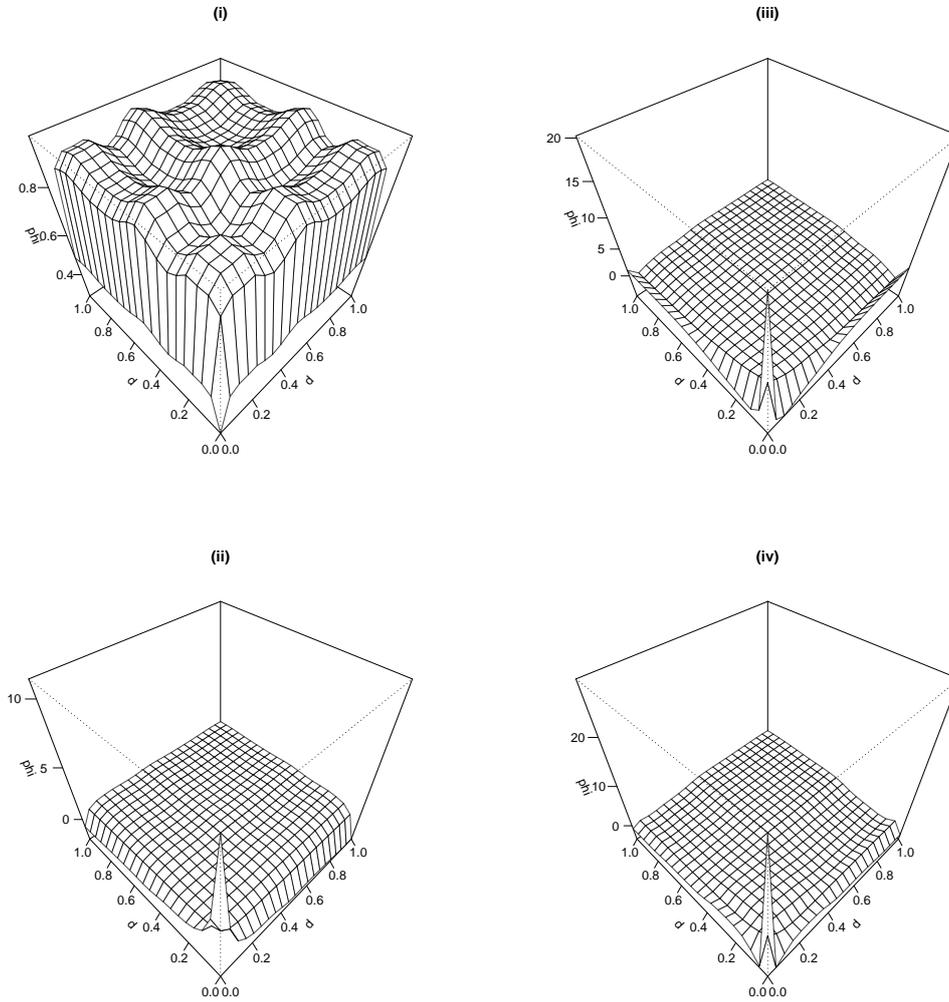
where  $\mathcal{I}$  is the index set of forecast origins and  $I = |\mathcal{I}|$ . That is, we approximate the integrated error using  $J = 201$  grid points for each curve.

The results of the exercise are shown in Tables 2.1 ( $h = 1$ ) and 2.2 ( $h = 3$ ). The forecast horizons correspond to 20 minutes and 1 hour, respectively. We find that both models perform similarly well. Models with rather many components perform fairly well, even in case of FAR, which is a bit surprising given the reciprocal eigenvalue problem explained above. However, the often-cited study which emphasizes this issue, Didericksen et al. (2012), evaluates the estimation performance based on the estimation error of the autoregressive kernel,  $\psi(c, d)$ , which is evaluated in simulation studies, and not based on the forecast performance. Returning to our empirical result, we find that even the occasionally nonstationary FAR(1) specifications perform considerably well in many situations. Between FAR and SPF, there is no clear favorite. Both are able to outperform the no-change forecast (RW) for some  $K$  in most cases. However, the variation of forecast is quite large for all models (including RW), so that they do not differ at all in the light of Diebold-Mariano tests.<sup>9</sup>

## 2.6 Conclusion

The present Chapter has introduced cumulative volume curves as *functional measures of liquidity*, as well as the framework of *weakly dependent functional time series*, due to

<sup>9</sup>We use the DM test somewhat naively, and only pairwise, as an explorative tool. For a discussion of the DM assumptions and alternatives see the very enlightening survey Diebold (2015).



**Figure 2.7:** Estimated autoregressive kernels of a functional autoregressive model of order 1, fitted to de-diurnalized ask curves of the Linde stock. Plots (i) to (iv) depict the estimates based on  $K = 1, \dots, 4$  functional principal components, evaluated on a grid. The second to fourth components introduce strong autocorrelation for liquidity near the ask quote. The implied norm,  $\|\hat{\psi}\| = (\int \int \hat{\Psi}_K^2(t,s) dt ds)^{1/2}$ , equals 0.87 (0.98, 1.16, 1.27) for  $K = 1$  (2,3,4), meaning that the estimated process is nonstationary for  $K \geq 3$ , presumably due to the spike near 0,0.

Hörmann and Kokoszka (2010), as a statistical framework for modeling such curves. Two model classes have been presented that can be used to model the curves' dynamics. For LOB snapshots taken every 20 minutes during trading, we find that both autocorrelation and persistence of liquidity curves are substantial.

The functional principal component or Karhunen-Loève representation of curves is a powerful tool for dimension reduction, i.e., for simple yet informative representations of the data. It is therefore instrumental in developing the new econometric models in subsequent Chapters of this thesis.

model	number of FPCs								RW
	1	2	3	4	5	6	7	8	
<b>Commerzbank</b>									
bid									
FAR	3.6118	2.7486	2.5955	2.4787	2.4224	2.3987	2.3802	2.3157	2.6851
SPF	3.6118	2.7561	2.6760	2.5573	2.5013	2.4846	2.4682	2.4595	2.6851
ask									
FAR	3.3520	3.0424	2.6834	2.5023	2.4395	2.4002	2.3773	2.3572	2.4678
SPF	3.3520	3.0451	2.8210	2.5961	2.5262	2.4942	2.4736	2.4594	2.4678
<b>MunichRe</b>									
bid									
FAR	6.7148	6.4169	6.6628	6.7535	6.7471	6.7120	6.7056	5.2044	5.8243
SPF	6.7145	6.1348	5.7408	5.7208	5.7192	5.7101	5.7082	5.7051	5.8243
ask									
FAR	7.1981	6.9839	7.1020	7.1046	7.0946	7.0663	7.0686	7.1081	6.6205
SPF	7.1961	6.6753	6.3582	6.3454	6.3440	6.3341	6.3320	6.3267	6.6205
<b>Linde</b>									
bid									
FAR	4.0104	3.9217	3.8506	3.8498	3.8150	3.7572	3.7570	3.5478	4.0168
SPF	4.0104	3.8802	3.8220	3.8131	3.7661	3.7404	3.7289	3.7218	4.0168
ask									
FAR	4.0879	3.9727	3.9003	3.8530	3.8664	3.8335	3.8201	3.8037	4.1558
SPF	4.0879	3.9802	3.9051	3.8722	3.8636	3.8341	3.8260	3.8163	4.1558

**Table 2.1:** RMISE ( $\times 1000$ ) of one-step-ahead out-of-sample forecasts of FAR(1) and SPF-AR(1) models. The models are estimated only once, using data from November 3, 2008 to December 30, 2009. The evaluation period is the whole year 2010.

model	number of FPCs								RW
	1	2	3	4	5	6	7	8	
<b>Commerzbank</b>									
bid									
FAR	3.7480	3.0138	2.8763	2.8042	2.7710	2.7518	2.7389	2.6944	3.1384
SPF	3.7479	3.0445	3.0394	2.9744	2.9421	2.9340	2.9261	2.9215	3.1384
ask									
FAR	3.4878	3.3551	2.9484	2.8269	2.7760	2.7473	2.7329	2.7210	2.9417
SPF	3.4878	3.3554	3.3129	3.1828	3.1538	3.1381	3.1282	3.1221	2.9417
<b>MunichRe</b>									
bid									
FAR	7.4940	8.0714	8.0268	8.0823	8.0919	8.1224	8.1095	6.2088	6.3611
SPF	7.4914	7.3235	7.2728	7.2712	7.2707	7.2704	7.2703	7.2703	6.3611
ask									
FAR	8.5128	9.0851	8.9640	9.0026	8.9956	9.0081	8.9931	9.0454	7.3497
SPF	8.5012	8.3610	8.3109	8.3080	8.3077	8.3072	8.3071	8.3070	7.3497
<b>Linde</b>									
bid									
FAR	4.3927	4.3902	4.3583	4.3473	4.3591	4.3013	4.3201	4.0948	4.5407
SPF	4.3929	4.3831	4.3611	4.3606	4.3523	4.3445	4.3434	4.3429	4.5407
ask									
FAR	4.4885	4.4343	4.3670	4.2952	4.2989	4.3268	4.3107	4.3194	4.6892
SPF	4.4888	4.4808	4.4660	4.4599	4.4591	4.4545	4.4522	4.4516	4.6892

**Table 2.2:** RMISE ( $\times 1000$ ) of three-step-ahead out-of-sample forecasts of FAR(1) and SPF-AR(1) models. The models are estimated only once, using data from November 3, 2008 to December 30, 2009. The evaluation period is the whole year 2010.



## **Part I**

# **Multiplicative Error Models with Liquidity Impact**



# Multiplicative Error Models with Liquidity Impact

*Part I* consists of Chapters 3 and 4. In Chapter 3, the GARCH model with *Functional eXogeneous Liquidity*, GARCH-FunXL, is introduced. The idea is to capture the impact of the (functional) shape of the LOB on the conditional variance of the return process of an asset by using a (functional, nonparametric) linear predictor. A two-step Gaussian QML estimation procedure in a log-GARCH-X framework is proposed. Estimation relies on the FPC expansion of both liquidity curves *and* the infinite-dimensional parameters which are needed to map the curves to the conditional variance. Monte Carlo evidence suggests that the model works well in finite samples. In an application to intraday log-returns of three stocks traded on the XETRA LOB, including an out-of-sample forecast exercise, the importance of liquidity impact for price variation is documented.

In Chapter 4, the logarithmic ACD-FunXL model is introduced. As Gaussian and exponential QML inference are equivalent, the estimation procedure of Chapter 3 is invoked. In an application to price durations for two DAX stocks traded on XETRA, the liquidity impact on price dynamics at ultra-high frequency is investigated. The model performs favorably.

Both models, GARCH-FunXL and ACD-FunXL, can be viewed as special cases of multiplicative error models, constituting the class of MEM-FunXL models. Note that both Chapters originally were prepared as two standalone manuscripts. Thus, they contain brief explanations of FTS methods that have already been discussed in depth in Chapter 2. However, we believe that these pieces improve the overall readability of these Chapters.



## Liquidity impact on volatility: The GARCH-FunXL model

### 3.1 Introduction

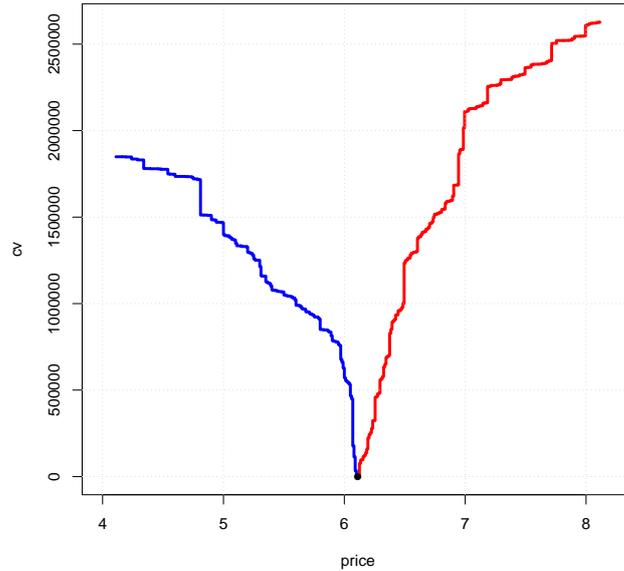
In recent years, the availability of high-frequency financial data has helped to deepen our understanding of the microstructure of financial markets. At the same time, the microstructure itself has changed as electronic order-driven trading platforms have become the de-facto standard for trading of financial securities. In financial econometrics, starting in the 1990s, two major strands of literature have emerged which make use of high-frequency data. The first is concerned with the dynamic properties of the trading process as a whole, not only of prices but also of trading volume, bid-ask spreads etc. as well as their interactions. Financial duration models pioneered by Engle and Russell (1998) and Engle (2000) and their many applications are the most prominent examples of this line of research. The second strand can be termed *realized volatility*. It aims at using high-frequency information on *prices* to improve the measurement of price volatility and, in a second step, modeling its dynamics. In other words, the realized volatility project has the same goals as for example GARCH modeling for, say, daily data, but hopes to improve accuracy by increasing the sampling frequency.

The present Chapter contributes to the literature by combining a traditional dynamic GARCH volatility model with the complete (in a cross-sectional sense) “micro-state of the market” for a financial security as implied by its limit order book (LOB).

To sketch the idea, consider the following linear GARCH(1,1) specification for an intradaily, de-seasonalized log-return,  $r_{t,i} = \log P_{t,i} - \log P_{t,i-1}$ , on the mid-quote,  $P_{t,i}$ , at intra-daily time  $i$  and trading day  $t$ ,

$$r_{t,i} = \sigma_{t,i} \varepsilon_{t,i}, \quad \varepsilon_{t,i} \stackrel{iid}{\sim} (0, 1), \quad \sigma_{t,i}^2 = \omega + \alpha r_{t,i-1}^2 + \beta \sigma_{t,i-1}^2, \quad (3.1)$$

as it has for instance been applied by Engle and Sokalska (2012) to 10-minute returns on NYSE. The LOB however does not only reveal information on prices, but also on the size of limit orders around the quotes, the LOB inventories. Figure 3.1



**Figure 3.1:** State of a limit order book at some time during the trading day, consisting of (i) prices, i.e. bid and ask quotes whose mean (the mid-quote) is represented by a dot, and (ii) liquidity, represented by the available requested (left) and offered (right) cumulative number of shares as a function of the price.

illustrates the state of an LOB at some specific time during a trading day.

We denote the left curve, the bid curve, by  $x_{t,i}^{(bid)}(P)$  and the right curve, the ask curve, by  $x_{t,i}^{(ask)}(P)$ , and view them as functions of the price. Now consider the case where order book inventories,  $x_{t,i}^{(ask)}(P)$  say, have an impact on the price. We extend the model introduced above by capturing possible effects of these function-valued objects on the variation of the price by specifying the conditional variance in terms of

$$\sigma_{t,i}^2 = \omega + \alpha r_{t,i-1}^2 + \beta \sigma_{t,i-1}^2 + \int_{\mathcal{P}} \gamma(m) x_{t,i-1}^{(ask)}(m) dm, \quad (3.2)$$

where  $\mathcal{P}$  is the interval of possible prices.

Model (3.2) is now of the GARCH-X type but, in contrast to existing specifications of this kind, the exogenous variable is a curve-valued, i.e. infinite-dimensional quantity. In fact,  $\int_{\mathcal{P}} \gamma(m) x_{t,i-1}^{(ask)}(m) dm$  is the limit case of a linear predictor, where  $\gamma(m)$  is a functional rather than a vectorial parameter, mapping liquidity at all price levels in  $\mathcal{P}$  to scalar conditional volatility.

Early studies on the connection of liquidity and volatility (e.g. Gallant et al. (1992), Jones et al. (1994)) employ daily transactions data to measure liquidity. More recently, the structure (Gouriéroux et al., 1998) and dynamics (Härdle et al. (2012b); Bowsher (2004) in an early version of Bowsher and Meeks (2008)) of liquidity as implied by an order-driven market have been studied. However, to the best of our knowledge the proposed GARCH-FunXL approach is the first attempt to model volatility dynamics using the full LOB.

The remainder of the Chapter is organized as follows: Section 3.2 explains in detail how liquidity curves, sampled intradaily at a constant frequency, are constructed from LOB data and parsimoniously represented in the framework of functional time series analysis. Basic empirical properties of these curves are evaluated for three liquid stocks traded on the German XETRA system. Parts of these results are also shown in Chapter 2.

Section 3.3 introduces the GARCH-FunXL model. A two-step QML estimation procedure for the model's parameters, especially the functional parameter  $\gamma(m)$ , is developed and the different sources of estimation uncertainty are discussed. Section 3.4 provides a simulation study investigating the finite-sample performance of the estimation strategy. In an empirical application presented in section 3.5, the model is fitted to the three XETRA stocks. Both in-sample results and out-of-sample forecast evaluations underline the relevance of liquidity for explaining price variation. Section 3.6 concludes the Chapter.

## 3.2 Liquidity

Limit order books carry dynamic information on price and liquidity of an asset. Strictly speaking, the price is an implication of the liquidity present. Nevertheless, as we explain in the following, it is possible to measure both phenomena separately.

### 3.2.1 Limit order book information

Regardless of the specific market design (limit order book or competing market makers), information on price and liquidity of a given stock is given by the requested (demand) and offered (supply) volume of shares around the quotes. The latter are an implication of the offers and requests: The bid quote is the highest supply price, the ask quote is the lowest demand price. The difference of the two is always at least one tick.

$LOB_t$ , the state of the limit order book at some time  $t$  during a trading day, can be characterized by  $P_t^{(s)}$ ,  $s \in \{bid, ask\}$ , the quotes (measured in ticks), and  $v_t^{(s)}(d)$ , the outstanding number of shares on market side  $s$  and at a price distance  $d$  (in ticks) from the respective quote. The LOB at time  $t$  is given by the vast-dimensional and irregularly populated vector

$$LOB_t := [P_t^{(bid)}, P_t^{(ask)}, v_t^{(bid)}(P_t^{(bid)} + 1), \dots, v_t^{(bid)}(0), v_t^{(ask)}(0), \dots]'$$

Adding up the demand (supply) in the market at a given relative price, we obtain the cumulative volume,

$$x_t^{(s)}(d) = \sum_{k=0}^d v_t^{(s)}(k),$$

see Definition 2.1.

There is a one-to-one relationship between CV and the average price per share as a function of the offered/requested market order volume as has been analyzed in Gouriéroux et al. (1998) and Bowsher (2004). If the volumes in the book are weighted by their prices, the resulting quantity is called *depth*. Therefore, the depth of an order book is a hybrid between a liquidity measure and a measure of the price of an asset. As we seek to analyze liquidity impact on the price process, we will use CV as our liquidity measure. A possible drawback of this approach is that a comparison of the liquidities of different stocks gets more complicated.

In the following, we call the mid-quote,  $P_t := \left( P_t^{(bid)} + P_t^{(ask)} \right) / 2$ , the *price* of the asset. Assuming an equidistant sampling scheme at frequency  $1/\Delta$ , the log-return of the price is given by  $r_t = \log P_t - \log P_{t-\Delta}$ . Confining ourselves to only a single constant sampling frequency, we set  $\Delta := 1$  without loss of generality.

Cumulative volume curves are termed (bid or ask) *liquidity* in the remainder of the paper.

### 3.2.2 Diurnal patterns

Both price volatility and liquidity exhibit certain regularities during a trading day that can be treated as being deterministic. In case of volatility, this is the well-known U-shape over the trading day. For liquidity, the pattern is a bivariate function of both time of day and relative price whose shape will be shown in brief.

#### Volatility pattern

As we are interested in asset returns at high frequency, we introduce a second time index or clock, so that  $r_{t,i}$  denotes the  $i$ -th of  $I$  raw intraday returns on day  $t$ ,  $t = 1, \dots, T$ . Following Andersen and Bollerslev (1997), Andersen and Bollerslev (1998) and Engle and Sokalska (2012), we assume that the raw return is given by the product of a stochastic component,  $y_{t,i}$ , and a deterministic diurnal component,  $s_i$ , i.e.,

$$r_{t,i} = y_{t,i}s_i.$$

Our interest centers on the conditional variance of the stochastic part,  $y_{t,i}$ . The deterministic diurnal pattern  $s_i$  can be estimated as the mean squared return at the specific intraday interval,  $\hat{s}_i = T^{-1} \sum_{t=1}^T r_{t,i}^2$  (Engle and Sokalska, 2012), or a smoothed version thereof. Andersen and Bollerslev (1997), for instance, use the flexible Fourier functional form proposed by Gallant (1981), where smoothness of the fitted pattern is implicitly imposed through the choice of constant and cyclical components.

Similar to this second approach, we use a cubic smoothing spline to fit the scatterplot of squared intraday returns vs. (intraday) time, where the smoothing parameter is chosen via generalized cross validation. Results are shown in Section 3.2.4.

### Liquidity pattern

Curve-valued liquidity can be decomposed in an analogous way. Here, for each market side, the diurnal pattern is itself a deterministic function of the relative price  $d$  and intraday time  $i$ . Raw liquidity on market side  $s$ ,  $\tilde{x}_{t,i}^{(s)}$ , is given by

$$\tilde{x}_{t,i}^{(s)}(d) = v_i^{(s)}(d)x_{t,i}^{(s)}(d),$$

where  $v_i(d)$  is the deterministic diurnal liquidity surface and  $x_{t,i}^{(s)}(d)$  the stochastic liquidity component, which is of primary interest in our analysis.

As the pattern is typically less pronounced than the volatility pattern, we do not smooth the diurnal pattern (which could in principle easily be done, for example by using a tensor product spline) but rather use simple averaging. The estimator is then given by

$$\hat{v}_i^{(s)}(d_j) = T^{-1} \sum_{t=1}^T \tilde{x}_{t,i}^{(s)}(d_j),$$

where  $d_1, \dots, d_J$  is the observation grid along the price axis. Empirical results are shown in Section 3.2.4.

### 3.2.3 Liquidity as functional time series

Let us briefly review the results from Section 2.3 that are relevant here. We have demonstrated that liquidity curves can be viewed as realization of a stationary functional stochastic process in discrete time. This is exactly the setting we consider here. A crucial ingredient we need in order to build the GARCH model with liquidity impact is the Karhunen-Loève decomposition or FPC representation,

$$x_t(d) = \mu(d) + \sum_{j=1}^{\infty} \zeta_{j,t} \phi_j(d).$$

The eigenvalues  $\lambda_j$  of the spectral representation are equal to the unconditional variances of the FPC scores  $\zeta_{j,t}$ . As the eigenvalues are strictly decreasing, the FPCs are sorted by their contribution to the (unconditional) variation of  $x_t$ . This gives rise to the  $K$ -truncated FPC representation

$$x_t(d) = \mu(d) + \sum_{j=1}^K \zeta_{j,t} \phi_j(d) + v_t(d),$$

where  $v_t(d) = \sum_{j=K+1}^{\infty} \zeta_{j,t} \phi_j(d)$  is the truncation error. In practice, we are interested in approximating the curves using such a truncation. The smallest number of components,  $K$ , necessary to explain a certain proportion (say 99%) of the curves' total variation is called the *effective dimension* of the liquidity process.

To empirically obtain the decomposition, the mean is estimated and extracted first. Discretized versions of the eigenfunctions  $\phi_k$  are obtained as the (rescaled) eigenvectors of the covariance matrix of the de-meaned curves. However, we smooth the

covariance matrix first using the FACE algorithm (Xiao et al., 2013), leading to smooth eigenfunction estimates  $\hat{\phi}_k$ . In our well-behaved setting with a dense and regular relative price grid  $d_1, \dots, d_J$ , the empirical FPC scores,  $\hat{\xi}_{k,t} = \int_0^1 (x_t(m) - \hat{\mu}(m)) \hat{\phi}_k(m) dm$ , can then be computed by simple numerical integration methods.

The procedure, originally developed in an *i.i.d.* setting, is valid for dependent functional data provided they are  $L^p$ - $m$ -approximable in the sense of Hörmann and Kokoszka (2010). Then  $\hat{\mu}$  and  $\hat{C}(z)$  are  $\sqrt{T}$ -consistent. Throughout, we assume the liquidity process(es) to be stationary in this sense.

### 3.2.4 Empirical results

Let us turn to the task of estimating the diurnal patterns and the FPCs for the data at hand. Before doing so, we provide a description of the data.

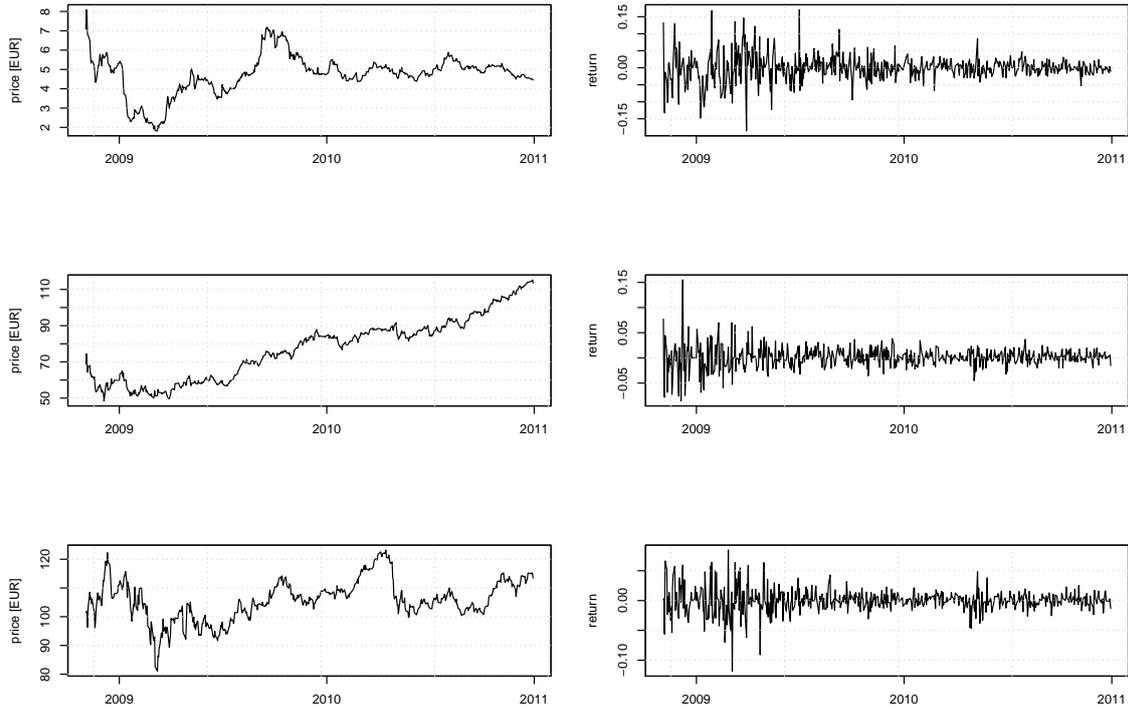
#### Data

We use historical limit order book data from the German XETRA system from November 3, 2008 to December 31, 2010, obtained from the Deutsche Börse AG. The data set covers 531 trading days and contains information on all limit order submissions, revisions and cancellations and all trades (i.e., market orders) at all permissible price levels for three DAX constituents: Linde, an industrial company, the Commerzbank, and the (re-)insurance company MunichRe.<sup>1</sup>

The sample starts in midst of the global financial crisis and ends in more tranquil times. To get a first impression of the “longer-run” behavior during these 26 months, consider the daily closing prices and the corresponding returns shown in Figure 3.2. The price levels of the three stocks behave quite differently during this period: Linde experienced a positive trend, MunichRe a sideways market, and Commerzbank stabilizes in the second half after a considerable turmoil in early 2009. In contrast, the dispersion patterns over time are quite similar for all three stocks.

The permissible price levels are the positive multiples of the tick size. The tick size on XETRA depends on the price. For most of our sample period (from January 2009 on), it is €0.001 if the instrument’s price is under €10, €0.005 for prices in the interval [€10, €50), 1 cent for prices from €50 to €100, and 5 cent otherwise. The data set allows us, in principle, to reconstruct  $LOB_t$  for any  $t$  in a given trading day. Continuous trading on XETRA starts after a 5-minute opening auction at 9am, is interrupted by another auction of this type at 1pm, and ends just before the closing auction at 5:30pm. In our application, we take snapshots of  $LOB_t$  every 20 minutes sampled during continuous trading. Specifically, we avoid auction effects by sampling at 9:09am, 9:29am, . . . , 12:49pm, 1:09pm, . . . , 5:29pm.

<sup>1</sup>For Commerzbank (MunichRe, Linde), 530 (526, 529) full trading days have been observed, with 25 daily snapshots each, amounting to 13250 (13150, 13225) observations, from which we compute 24 intraday returns per day. After removing the diurnal volatility pattern as explained in the present section, we subtract the mean from each series, so that no zero returns remain. Then we remove 2 (3, 0) outliers. This leaves us with 12719 (12621, 12696) observations, respectively. For the bid and ask liquidity curves, no missing values or outliers are present.



**Figure 3.2:** Daily closing prices (left panel) and log returns (right panel) for Commerzbank (upper panel), Linde (center panel), and MunichRe (bottom panel).

Order book data are typically available only up to a certain “level”, which means that only the volumes at the first 10 or 20 best prices are provided. Therefore, the actual price range covered depends on both the tick size and on how densely the orders are posted. We, in contrast, have information on *all* admissible price levels. However, in our analyses, we only consider volumes posted at the nearest €2 around the quotes on each market side, a range we expect to suffice to observe all relevant aspects of the LOB. For all three stocks, we record cumulative volumes in increments of one cent, so that the snapshots are of the form

$$\begin{aligned}
 LOB_t := & [P_t^{(bid)}, P_t^{(ask)}, \\
 & v_t^{(bid)}(\text{€}0), v_t^{(bid)}(\text{€}0.01), \dots, v_t^{(bid)}(\text{€}2.00), \\
 & v_t^{(ask)}(\text{€}0), v_t^{(ask)}(\text{€}0.01), \dots, v_t^{(ask)}(\text{€}2.00)]'
 \end{aligned}$$

Moreover, our data set not only contains the information that was available to the market participants, it also provides a full picture of hidden liquidity as, for example, iceberg orders are included.

## Diurnal patterns

The estimated diurnal volatility patterns for the three considered stocks are shown in the left panel of Figure 3.3. They basically exhibit the familiar U-shape. In the second half of the day, volatility typically rises after 3:29pm (US markets open at 3:30pm CET) and peaks at 4:09pm before decreasing somewhat until the end of trading at 5:30pm.

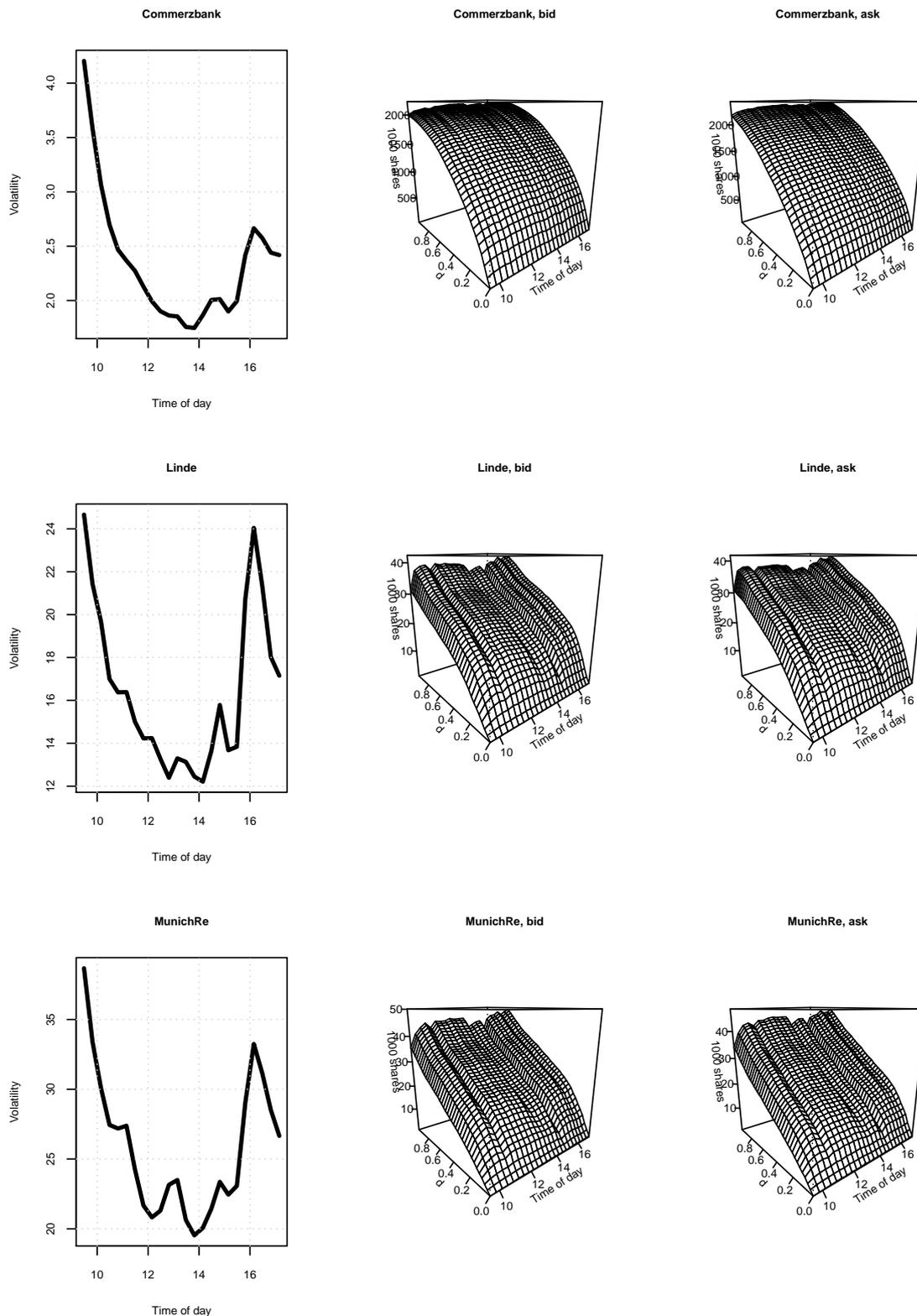
The center and right panels of Figure 3.3 show the corresponding liquidity patterns known from Chapter 2. They are quite similar for the two market sides. Especially at greater distance from the quotes, there is a slight upward trend which may be attributed to the fact that some of these limit orders remain only in the book because they are unlikely to be executed. Still, more than 95 percent of the total limit order volume is cancelled instead of being executed. The median lifetime of a limit order is less than one second, while the mean lifetime increases as the distance to the quotes increases.

## The structure of liquidity

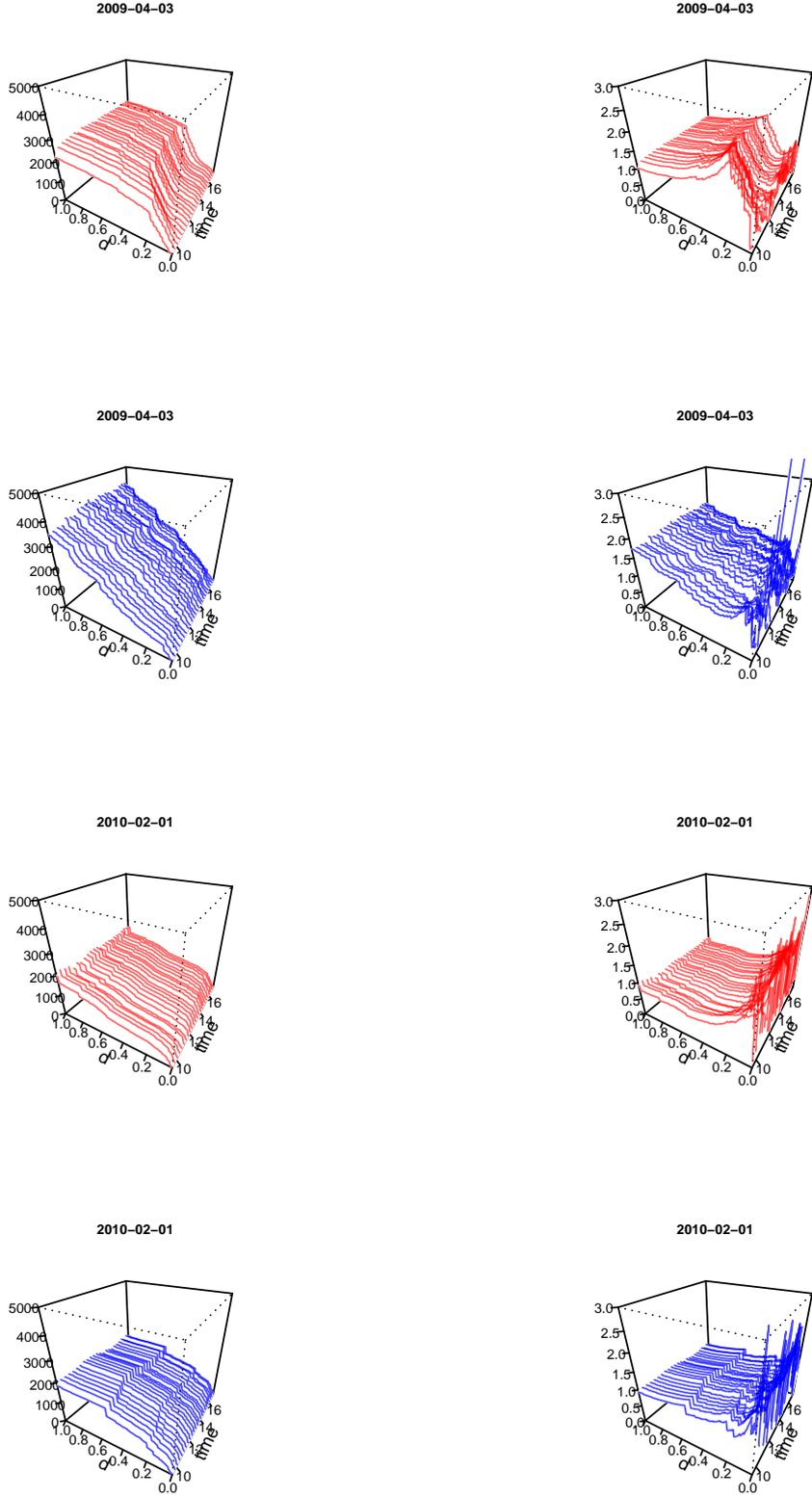
We now turn to analyzing the de-seasonalized liquidity over the full sample. The time series of bid and ask liquidity curves for two exemplary days and the Commerzbank stock are shown in Figure 3.4. Note that by construction the unconditional mean of de-seasonalized liquidity is 1 for all locations, so that values above (below) 1 can be interpreted as high (low) liquidity at a specific time of the day and region of the LOB, respectively.

Turning to the functional principal components of the series, the top panel of Figure 3.5 shows the normalized eigenvalues, i.e., the variances explained by each of the first ten components, both for the bid and the ask side using the whole range of 201 tick levels. For both market sides, four components are needed to capture 95 percent of the variation in liquidity. The bottom panel of Figure 3.5 shows the first four estimated eigenfunctions for the two market sides. Note that the eigenfunctions are unique only up to the sign. With this in mind, we conclude that the factor structure of both sides is quite similar. In both cases, the first eigenfunction is almost horizontal and, therefore, can be interpreted as a “level” factor. The remaining eigenfunctions capture different aspects of liquidity variation, all having in common that deviations near the quotes (in the left part of the domain) are larger than at a greater distance from the quotes. This finding reflects the larger variation of liquidity near the quotes, as could already be seen from the exemplary days shown above.

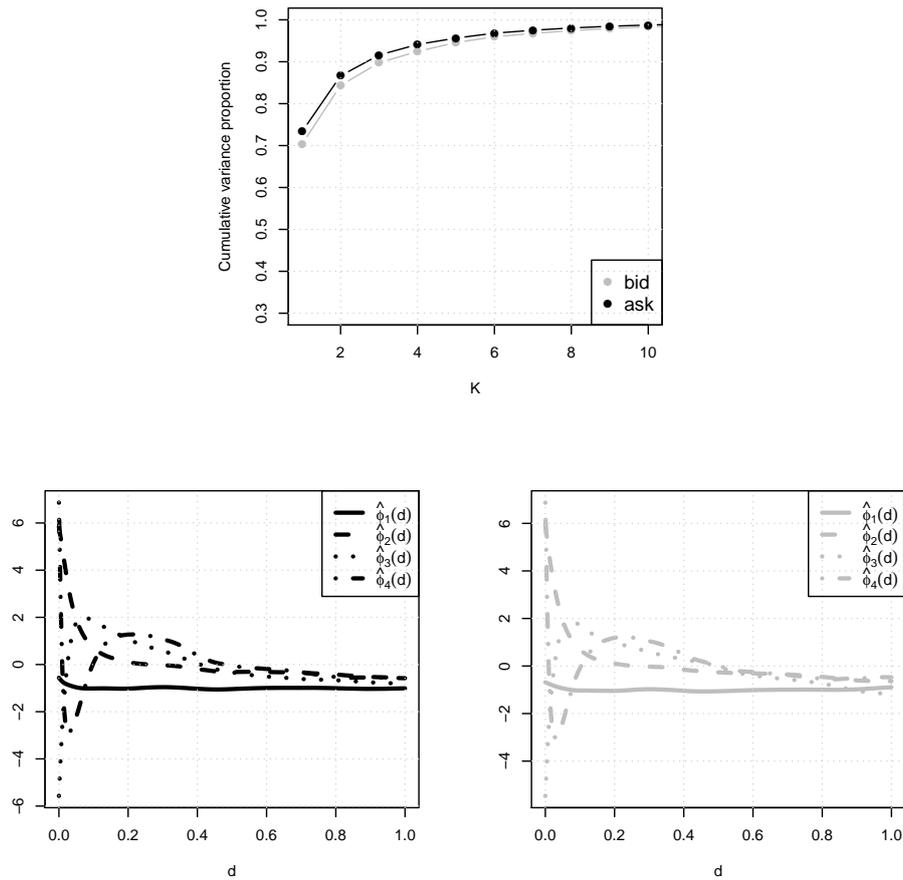
To statistically test for differences in the eigenfunctions, Benko et al. (2009) propose a resampling-based test for equality of eigenfunctions for two-sample situations like ours. We do not apply this test here. Instead we estimate the scores,  $\tilde{\zeta}_{j,t}$ , in the truncated FPC expansion  $\hat{x}_t = \sum_{j=1}^4 \tilde{\zeta}_{j,t} \hat{\phi}_j$ , using OLS regressions of the  $x_t$  against the first four eigenfunctions. We find that eigenfunction estimates based on bid curves only, based on ask curves only, and in a pooled estimation, respectively, yield in all three cases virtually identical score estimates,  $\hat{\zeta}_{j,t}$ , and, hence, approximations  $\hat{x}_t = \sum_{j=1}^4 \hat{\zeta}_{j,t} \hat{\phi}_j$ .



**Figure 3.3:** Diurnal volatility (left panel) and liquidity (center and right panel) patterns for the three stocks, fitted using a smoothing spline. From top to bottom: Commerzbank, Linde, and MunichRe.



**Figure 3.4:** Functional time series of ask (red) and bid (blue) liquidity curves for April 3, 2009 (top) and February 1, 2010 (bottom). In the left panel, the raw data (cumulative number of shares, measured in thousands, within a range of €0 to €2 from the quotes) are depicted, in the right panel the de-seasonalized versions of the same curves are shown.



**Figure 3.5:** Top panel: Cumulative normalized eigenvalues of the estimated liquidity covariance operators. Bottom panel: First four estimated eigenfunctions for ask (left) and bid (right) sides, all for the Commerzbank stock. Note that the eigenfunctions are unique only up to the sign.

In contrast, liquidity itself is far from symmetric. The contemporaneous correlation of the scores between market sides is rather low (see Table 3.1).

	bid.1	bid.2	bid.3	bid.4
ask.1	-0.27	0.07	-0.04	0.34
ask.2	0.07	-0.02	0.09	-0.15
ask.3	0.08	-0.02	0.03	0.10
ask.4	-0.06	0.34	0.10	-0.08

**Table 3.1:** Contemporaneous sample correlations between the first four FPC score series for the two market sides. Not shown: Correlations between scores of the same market side which are orthonormal by construction.

### 3.3 GARCH-FunXL

#### 3.3.1 The model

The raw log-returns  $r_{t,i}$  during intraday times  $i - 1$  and  $i$  on day  $t$ ,  $i = 1, \dots, I$ , are generated by

$$r_{t,i} = y_{t,i}s_i, \quad (3.3)$$

$$y_{t,i} = \sigma_{t,i}\varepsilon_{t,i}, \quad \varepsilon_{t,i} \stackrel{iid}{\sim} (0, 1), \quad (3.4)$$

where  $s_i$  is a deterministic diurnal volatility component, and  $\sigma_t$  is the conditional volatility of the de-seasonalized or *de-diurnalized* returns  $y_{t,i}$ . This setup is as in Andersen and Bollerslev (1997), Andersen and Bollerslev (1998) and Engle and Sokalska (2012), AB and ES henceforth, with the exception that these authors further decompose  $\sigma_t$  into a daily and an intradaily component. ES use commercially available data based on multifactor risk models for the daily component, whereas AB use a GARCH specification. In principle we could also adopt these approaches, but at this stage prefer to keep it simple.

Note that the model generates only intraday returns, and assumes that  $\sigma_{t,i}$  stays constant between the trading hours of subsequent trading days, i.e.,  $\sigma_{t,I} = \sigma_{t+1,0}$ .

The conditional volatility follows a GARCH specification, which is augmented by exogeneous information in terms of the liquidity curves at the beginning of each intraday interval from both market sides, i.e.

$$\sigma_{t,i} = f(y_{t,i-1}, \dots, x_{t,i-1}^{(ask)}, x_{t,i-1}^{(bid)}), \quad (3.5)$$

where  $y_{t,i-1}, \dots$  denotes the entire return history, in the following denoted by  $\mathcal{F}_{t,i-1}$ . As explained in Section 3.2, the liquidity curves are, as the  $y_{t,i}$ , de-diurnalized quantities.

In the following we lighten notation by dropping the  $i$ . Furthermore, we consider a log-GARCH specification whose endogeneous part has order (1,1). We choose the functional liquidities to enter the model in a log-linear fashion as well. Then, the conditional log-variance becomes

$$\begin{aligned} \log \sigma_t^2 &= \omega + \alpha \log y_{t-1}^2 + \beta \log \sigma_{t-1}^2 \\ &\quad + \int_0^1 \gamma^{(ask)}(m) x_{t-1}^{(ask)}(m) dm + \int_0^1 \gamma^{(bid)}(m) x_{t-1}^{(bid)}(m) dm. \end{aligned} \quad (3.6)$$

We denote this larger information set, consisting of past returns plus liquidities at  $t-1$ , by  $\mathcal{F}_{t-1}^L$ .

Collecting all this and the ingredients from Section 3.2, the logarithmic GARCH(1,1)-FunXL model for returns with conditional volatility influenced by  $x_{t-1}^{(s)}$ ,  $s \in \{\text{bid}, \text{ask}\}$ , can be defined as follows.

**Definition 3.1** (Logarithmic GARCH(1,1)-FunXL process). *Let  $x_t^{(ask)}, x_t^{(bid)}$  be drawn from curve-valued exogeneous liquidity processes as specified before. Then,  $y_t$  follows a logarithmic GARCH(1,1)-FunXL process, if*

$$y_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \stackrel{iid}{\sim} (0, 1) \quad (3.7)$$

$$\begin{aligned} \log \sigma_t^2 &= \omega + \alpha \log y_{t-1}^2 + \beta \log \sigma_{t-1}^2 \\ &\quad + \int_0^1 \gamma^{(ask)}(m) x_{t-1}^{(ask)}(m) dm + \int_0^1 \gamma^{(bid)}(m) x_{t-1}^{(bid)}(m) dm, \end{aligned} \quad (3.8)$$

where

$$x_t^{(s)} = \mu^{(s)} + \sum_{k=1}^{\infty} \xi_{k;t}^{(s)} \phi_k^{(s)}. \quad (3.9)$$

We assume that the terms  $\int_0^1 \gamma^{(s)}(m) x_{t-1}^{(s)}(m) dm$ ,  $s \in \{\text{bid}, \text{ask}\}$ , are non-degenerate in the sense that the coefficient functions are finite over  $[0, 1]$  and recall that both liquidity processes are stationary in the sense explained above, especially having finite mean function and covariance operator. Then, if  $|E(\log \varepsilon_t^2)| < \infty$ , (3.7)-(3.9) admits the ARMA(1,1)-(Fun)X representation

$$\begin{aligned} \log y_t^2 &= \pi_0 + \pi_1 \log y_{t-1}^2 + \theta_1 \log u_{t-1} \\ &\quad + g \left( x_{t-1}^{(ask)}, x_{t-1}^{(bid)}; \gamma^{(ask)}, \gamma^{(bid)} \right) + u_t, \end{aligned}$$

where

$$\begin{aligned}\pi_0 &= \omega + (1 - \beta)E[\log \varepsilon_t^2], \\ \pi_1 &= \alpha + \beta, \\ \theta_1 &= -\beta, \\ u_t &= \log \varepsilon_t^2 - E[\log \varepsilon_t^2], \\ g\left(x_{t-1}^{(ask)}, x_{t-1}^{(bid)}; \gamma^{(bid)}, \gamma^{(ask)}\right) &= \int_0^1 \gamma^{(ask)}(m) x_{t-1}^{(ask)}(m) dm + \int_0^1 \gamma^{(bid)}(m) x_{t-1}^{(bid)}(m) dm,\end{aligned}$$

see also Sucarrat et al. (2013). An interesting feature is that the intercept and therefore the autocorrelation function of  $\log y_t^2$  depends on the innovation distribution via  $E[\log \varepsilon_t^2]$ . It follows immediately that  $\log y_t^2$  is stationary if the roots of the AR polynomial lie outside the unit circle, i.e., for a model of order (1,1),  $-1 < \alpha + \beta < 1$ .

We choose the log-GARCH specification primarily to avoid non-negativity constraints for the functional exogenous part of the model. Such constraints would be difficult to impose in presence of an infinite-dimensional parameter and de-diurnalized liquidity curves without positivity constraints. One further attractive feature of log-GARCH specifications is that the log conditional variance has no lower bound (in contrast to the standard GARCH case). A possible drawback could be that the model does not allow for zero returns.<sup>2</sup> For more details see Francq et al. (2013), who use an asymmetric log-GARCH specification (similar to the GJR-GARCH), which could also be adopted here. Another application of the log-GARCH-X, which is in some aspects similar to ours, is the Realized GARCH model of Hansen et al. (2012).

### 3.3.2 Estimation

We recall from Section 3.2 that the liquidity curves can be approximated by their first  $K$  functional principal components,

$$x_t^{(s)}(d) \approx \mu(d)^{(s)} + \sum_{k=1}^K \phi_k^{(s)}(d) \xi_{k,t}^{(s)}.$$

### Assumptions

We make the following assumptions:

---

<sup>2</sup>This is, however, not of practical relevance for de-meaned returns.

(i) For each  $s$ , there is some  $K < \infty$  for which

$$\begin{aligned} & \int_0^1 \sum_{j=K+1}^{\infty} \phi_j^{(s)}(m) x_t^{(s)}(m) dm = 0 \\ \Leftrightarrow & \int_0^1 \sum_{j=K+1}^{\infty} \sum_{i=1}^{\infty} \phi_j^{(s)}(m) \phi_i^{(s)}(m) \zeta_{i,t}^{(s)} dm = 0 \\ \Leftrightarrow & \int_0^1 \sum_{j=K+1}^{\infty} \sum_{i=K+1}^{\infty} \phi_j^{(s)}(m) \phi_i^{(s)}(m) \zeta_{i,t}^{(s)} dm = 0 \end{aligned}$$

holds.

(ii) This  $K$  is the same for both market sides.

This means that only a finite number of liquidity components, and moreover only those that explain liquidity best, have an impact on our quantity of primary interest, the price volatility.

However, this assumption does not rule out dynamic dependencies between the first  $K$  and the remaining components. Consider, for example, the following vector autoregressive (VAR) liquidity dynamics for a process whose covariance has  $K + L$  non-zero eigenvalues,  $L \geq 1$ ,

$$\begin{aligned} \zeta_t^{(s)} &= \begin{bmatrix} \zeta_{1,t}^{(s)} \\ \vdots \\ \zeta_{K+L,t}^{(s)} \end{bmatrix} = \mathbf{v}^{(s)} + \sum_{j=1}^p \mathbf{A}_j^{(s)} \zeta_{t-j}^{(s)} + \mathbf{v}_t^{(s)}, \\ x_t^{(s)}(d) &= \mu^{(s)}(d) + \sum_{j=1}^{K+L} \phi_j^{(s)}(d) \zeta_{j,t}^{(s)}, \end{aligned}$$

i.e., each liquidity curve can be decomposed to  $K + L$  components which are orthonormal ( $\int_0^1 \phi_i(m) \phi_j(m) dm = 1$  for  $i = j$  and zero otherwise).

All univariate score processes are contemporaneously uncorrelated with all the others, but may well depend on lagged values of other processes. This is ruled out if the autoregressive matrix of the full liquidity process is assumed to be block diagonal in the sense that the first  $K$  components' scores do not interact with the remaining components  $K + 1, \dots, K + L$ .

While not explicitly claiming that the scores have such VAR dynamics, we finally assume that the lead and lag effects of components  $K + 1, \dots$  on the first  $K$  components' scores are negligible, which is a reasonable assumption for our data: Fitting VAR models to the empirical FPC scores, we find that autoregressive matrices are nearly diagonal, i.e., each individual score series is mainly driven by its own past.

## Two-step estimation

We estimate the GARCH-FunXL model in two steps.

1. Estimation of the liquidity curves using the orthonormal FPC expansion

$$x_t^{(s)}(d) = \hat{\mu}^{(s)}(d) + \sum_{k=1}^K \hat{\phi}_k^{(s)}(d) \hat{\xi}_{k,t}^{(s)},$$

where the true  $K$ , mean function  $\mu$ , and eigenfunctions  $\phi_k$  are unknown, and the  $\hat{\xi}_{k,t}^{(s)} = \int_0^1 (x_t^{(s)}(m) - \hat{\mu}^{(s)}(m)) \hat{\phi}_k^{(s)}(m) dm$  are computed via numerical integration. This step has been outlined in detail in Section 3.2.

2. QML estimation of the GARCH-FunXL parameters using the scores  $\hat{\xi}_{k,t}^{(s)}$ ,  $k = 1, \dots, K$ ,  $t = 1, \dots, T$  from Step 1 and the return data.

For statistical inference conditional on a truncated  $K$ -component FPC decomposition of  $x_t^{(s)}$ , we employ a Gaussian quasi-likelihood approach to obtain estimates of  $\omega, \alpha, \beta, \gamma^{(bid)}, \gamma^{(ask)}$ . The conditional distribution of the logarithmic GARCH(1,1)-FunXL with Gaussian innovations is given by

$$y_t | \mathcal{F}_{t-1}^L \sim N \left( 0, \exp \left( \omega + \alpha \log y_{t-1}^2 + \beta \log \sigma_{t-1}^2 + \int_0^1 \gamma^{(ask)}(m) x_{t-1}^{(ask)}(m) dm + \int_0^1 \gamma^{(bid)}(m) x_{t-1}^{(bid)}(m) dm \right) \right). \quad (3.10)$$

We do not claim the innovations to be Gaussian, instead we are interested in inference on the latent volatility process only. The Gaussian quasi-log-likelihood is given by

$$l(\mathbf{y}, \mathbf{x}; \omega, \alpha, \beta, \gamma^{(bid)}, \gamma^{(ask)}) = -\frac{1}{2} \sum_{t=2}^T \left( \sigma_t^2 + \frac{y_t^2}{\sigma_t^2} \right),$$

where  $\mathbf{y}$  is the vector of de-diurnalized returns, and  $\mathbf{x}$  the “matrix” of de-diurnalized liquidity curves.

As both the  $x_t^{(s)}$  and the coefficients  $\gamma^{(s)}(\cdot)$  are infinite-dimensional objects, the term  $\int_0^1 \gamma^{(s)}(m) x_t^{(s)}(m) dm$  has to be approximated by some finite-dimensional representation. In our practical application, we use  $K = 1, \dots, 5$ . For all three stocks considered,  $K = 4$  components explain at least 95 percent of the curves’ variation.

Introducing a  $K$ -dimensional parameter vector  $\gamma^{(s)} = [\gamma_1^{(s)} \ \dots \ \gamma_K^{(s)}]$  for each market side, we expand the coefficient function using the same set of  $K$  eigenfunctions that is used to represent the curves themselves,

$$\gamma^{(s)}(d) = \sum_{k=1}^K \gamma_k^{(s)} \hat{\phi}_k^{(s)}(d),$$

so that, plugging in estimated mean, eigenfunctions and scores from the FPCA of the liquidity curves, the integral  $\int_0^1 \gamma^{(s)}(m)x_t^{(s)}(m)dm$  becomes

$$\int_0^1 \sum_{j=1}^K \sum_{k=1}^K \hat{\xi}_{j;t}^{(s)} \hat{\phi}_j^{(s)}(m) \gamma_k^{(s)} \hat{\phi}_k^{(s)}(m) dm = \sum_{k=1}^K \gamma_k^{(s)} \hat{\xi}_{k;t}^{(s)}$$

by orthonormality of the eigenfunctions. This approach is well-known from *functional principal component regression* and its core idea is the same as in PC regression within the usual scalar multiple regression setting. Note that this structural assumption along with the assumption that only a finite number of components affects the conditional variance implies an identification problem: There are infinitely many functions  $\phi_j(d)$  which are orthogonal to the  $K$  functions appearing in either the “true” or the fitted liquidity representation. Thus, each of these  $\phi_j(d)$  could be added to the basis expansion of  $\gamma^{(s)}(m)$  without affecting the model’s goodness of fit.

Defining

$$G_{t-1} := \alpha \log y_{t-1}^2 + \beta \log \sigma_{t-1}^2,$$

we can now write the conditional volatility as

$$\begin{aligned} \log \sigma_t^2 &= \omega + G_{t-1} + \int_0^1 \gamma^{(bid)}(m)x_{t-1}^{(bid)}(m)dm + \int_0^1 \gamma^{(ask)}(m)x_{t-1}^{(ask)}(m)dm \\ &= \omega + G_{t-1} + \sum_s \int_0^1 \gamma^{(s)}(m)x_{t-1}^{(s)}(m)dm \left( \hat{\mu}^{(s)}(m) + \sum_{k=1}^K \hat{\xi}_{t-1;k}^{(s)} \hat{\phi}_k^{(s)}(m) \right) dm \\ &= \omega + G_{t-1} + \underbrace{\sum_s \int_0^1 \gamma^{(s)}(m)\hat{\mu}^{(s)}(m)dm}_{=:\gamma_0^{(s)}} + \sum_{k=1}^K \hat{\xi}_{k;t-1}^{(s)} \underbrace{\int_0^1 \gamma^{(s)}(m)\hat{\phi}_k^{(s)}(m)dm}_{=:\gamma_k^{(s)}} \\ &= \underbrace{\omega'}_{:=\omega+\gamma_0} + G_{t-1} + \sum_s \sum_{k=1}^K \gamma_k^{(s)} \hat{\xi}_{k;t-1}^{(s)}. \end{aligned} \quad (3.11)$$

Doing so, the infinite-dimensional problem boils down to an estimation of  $2K$  additional scalar parameters compared to the original model,  $\omega, \alpha, \beta, \gamma_1^{(bid)}, \dots, \gamma_K^{(bid)}, \gamma_1^{(ask)}, \dots, \gamma_K^{(ask)}$ . However, in practice it may pay to be rather generous in the choice of  $K$ , because the components that explain much of liquidity variation are not guaranteed to have a large impact on price volatility. Conversely, modes of variation that are rather unimportant for liquidity variation may well be of great importance for predicting price volatility.

### Properties of the estimators

Although the inclusion of exogeneous variables (like interest rates or realized volatility) has already been in vogue for some time, theoretical properties of GARCH-X processes and especially QML estimators of their parameters have only been established very recently. Han (2013) and Han and Kristensen (2014) investigate QML estimation for a broad class of possible exogeneous processes, including long-memory and integrated processes, but only for linear specifications and univariate exogeneous processes. In particular, in case of a stationary exogeneous process (as assumed here), QMLEs of linear GARCH processes' parameters retain their favorable properties.

To the best of our knowledge, and as has been documented by Sucarrat et al. (2013), Francq et al. (2013), and Hansen et al. (2012), such asymptotic results for log-GARCH-X models do not exist. The latter however, based on the results of Straumann et al. (2006), conjecture consistency and asymptotic normality of their QMLE in a scalar log-GARCH-X framework without providing a proof. With the above assumptions regarding liquidity effects on volatility, we are basically in a log-GARCH-X setting as well. However, an additional complication is given by the fact that our exogeneous variables, the FPC scores, are generated by a nonparametric procedure in the first place, inducing additional estimation uncertainty. Surprisingly, this fact is rarely addressed in the FDA literature, see for example Yao et al. (2005b). Therefore, even if the step-two parameter estimates could be shown to be consistent and asymptotically normal, ignoring estimation uncertainty from the first step, i.e., for mean functions, eigenfunctions, and scores will result in confidence bands for  $\gamma^{(s)}(\cdot)$  which are too narrow. An alternative, frequently used in FDA, is to choose a bootstrap approach. For our time series setting, the stationary bootstrap proposed by Politis and Romano (1994), where blocks of random lengths are drawn and reassembled to form resamples of the original series, is most suitable.

Simulation results reported in section 3.4 convey an idea of the properties of QML estimators in our log-linear specification.

### 3.3.3 Liquidity impact

The conditional variance of the GARCH-FunXL model can be written as a product of the (endogeneous) GARCH part and the exogeneous liquidity part, i.e.,

$$\sigma_t^2 = \exp \left( \omega + \alpha \log y_{t-1}^2 + \beta \log \sigma_{t-1}^2 \right) \times \exp \left( \int \gamma^{(ask)}(m) x_{t-1}^{(ask)}(m) dm + \int \gamma^{(bid)}(m) x_{t-1}^{(bid)}(m) dm \right).$$

Defining

$$LI_t := \exp \left( \int \gamma^{(ask)}(m) x_{t-1}^{(ask)}(m) dm + \int \gamma^{(bid)}(m) x_{t-1}^{(bid)}(m) dm \right), \quad (3.12)$$

the second term is henceforth called the *liquidity impact*. Note that the sub-index is chosen according to the target variable, the conditional variance. Due to the multiplicative structure of the model, liquidity reduces volatility for  $LI_t < 1$  and increases it for  $LI_t > 1$ .  $LI_t$  can further be splitted up into the contributions of each market side,  $LI_t = LI_t^{(ask)} LI_t^{(bid)}$ .

Moreover, in analogy to the news impact curve, a three-dimensional plot of  $LI_t$  against  $d$  and  $x_{t-1}^{(s)}(d)$  amounts to a *liquidity impact surface* (LIS). For a given market side, the LIS shows the influence of liquidity on conditional variance at all locations,  $d$ , within the LOB. However, as both the functional parameter and the liquidity curves can have quite complex shapes, for instance different signs at different locations of their domain, the LIS's interpretability is limited.

The  $K$ -truncated, estimated version of the liquidity impact is given by

$$\widehat{LI}_t = \exp \left( \hat{\gamma}'^{(ask)} \hat{\xi}_{t-1}^{(ask)} + \hat{\gamma}'^{(bid)} \hat{\xi}_{t-1}^{(bid)} \right).$$

Confidence statements for the liquidity impact are directly linked to estimation uncertainty about the functional parameters discussed above.

### 3.4 Estimation uncertainty in GARCH-FunXL models

We investigate by means of a simulation study how well the two-step estimation works for our model in finite samples. In functional regression, the quantity of interest is typically the parameter. In the present study, however, we focus solely on the liquidity impact (which of course involves estimation of the functional parameter).

In order to simulate from the model, we have to specify a data-generating mechanism for the liquidity process first. We choose the setting of Aue et al. (2015) where the scores follow a VAR model,

$$\begin{aligned} \xi_t^{(s)} &= \begin{bmatrix} \xi_{1,t}^{(s)} \\ \vdots \\ \xi_{K,t}^{(s)} \end{bmatrix} = \mathbf{v}^{(s)} + \sum_{j=1}^p \mathbf{A}_j^{(s)} \xi_{t-j}^{(s)} + \mathbf{v}_t^{(s)}, \\ x_t^{(s)}(d) &= \mu^{(s)}(d) + \sum_{j=1}^K \phi_j^{(s)}(d) \xi_{j,t}^{(s)}. \end{aligned}$$

We simulate from a  $K = 5$ -dimensional score process with  $p = 1$  and  $T = 1000, 5000, 10000$ . Note that empirically, even  $K = 4$  components capture more than 95 percent of liquidity variation for all data sets. We use the eigenfunctions from the FPC representation of Brownian motion (see Ash and Gardner (1975)),

$$\phi_k(d) = \sqrt{2} \sin(k - 0.5)\pi d,$$

two different sets of eigenvalues and also two different serial dependence structures for the scores. As eigenvalues, we use (i) the decay as for Brownian motion,

$\lambda_k = 4/(2k - 1)^2\pi^2$  and (ii) the empirical decay as estimated from Commerzbank's ask curves. As serial dependence structures we use (i) serial independence,  $A_1 = \mathbf{0}_{(5 \times 5)}$ , and (ii)  $A_1$  resembling the dependence found empirically (estimated from Commerzbank's ask curves),

$$A_1 = \begin{bmatrix} 0.98 & 0.08 & 0.09 & -0.09 & 0.08 \\ 0.01 & 0.82 & -0.37 & 0.03 & -0.09 \\ 0.00 & -0.20 & 0.24 & -0.22 & 0.14 \\ 0.00 & 0.02 & -0.10 & 0.81 & 0.26 \\ 0.00 & -0.01 & 0.04 & 0.11 & 0.76 \end{bmatrix},$$

i.e. with high diagonal elements and low off-diagonal elements.

Empirical versions of  $\zeta_t^{(s)}$  from which  $A_1$  can be estimated are obtained as a by-product of the GARCH-FunXL estimation procedure proposed in section 3.3.

The unconditional variance  $\Gamma_\zeta(0)$  in VAR models depends on both the autoregressive matrices  $A_j$  and the covariance matrix  $\Sigma_v$  of the innovation vector. As we seek to discriminate between effects due to serial dependence and effects due to score variation, we simulate from (i) processes with identical innovation covariances and different serial dependencies (implying different unconditional covariance matrices) and (ii) processes with equal unconditional covariance matrices and different serial dependencies (implying different innovation covariance matrices).

We investigate models with only one functional exogenous process, using three different functional parameters, (i)  $\gamma(m) = 0.01\phi_1(m)$ , (ii) a linear combination of all 5 eigenfunctions with weights  $(0.01, 0.004, 0.002, 0.0006, 0.001)$ , and (iii)  $\gamma(m) = 0.002(4 + 5m - 10m^2 + 4\cos(5m))$ . Thus, parameters (i) and (ii) should be identifiable more easily than (iii), which is not constructed from eigenfunctions. The implied true liquidity impacts are similar to those found in the data. All function evaluations are on  $J = 201$  equidistant grid points in  $[0, 1]$ .

A summary of the results is shown in table 3.2. Not surprisingly, estimation accuracy increases substantially with increasing sample size.

Note that estimated models which capture the true liquidity impact accurately do not necessarily exhibit functional parameter estimates that are close to the truth as well. This is due to the identification issue discussed above and especially the case if (i)  $K$ , the number of estimated components used in the basis expansion of the parameter, exceeds the true  $K$  and if (ii) the true parameter can not be approximated well in terms of the eigenfunctions  $\phi_k$  of the liquidity process.

If liquidity dynamics are governed by a stable VAR process and  $T$  is reasonably large, we found that eigenfunctions can be estimated very accurately. Therefore, in situations where the true functional parameter is a linear combination of these eigenfunctions, its estimation is also accurate. Moreover, so are confidence bands constructed based on the conjectured normal asymptotics, despite ignoring the estimation uncertainty about the eigenfunctions.

We finally note that the results gained for GARCH-FunXL also hold for the ACD-FunXL model of Chapter 4 as the latter is estimated by the exponential QML method which is equivalent to Gaussian QML. Both models are part of the broader class of

MEM-FunXL models.

## 3.5 Modeling XETRA returns

We use the GARCH-FunXL process to model 20-minute snapshots for the Commerzbank, MunichRe, and Linde stocks traded on the German XETRA LOB from November 3, 2008 to December 31, 2010.<sup>3</sup>

We fit logarithmic GARCH-FunXL models of GARCH order (1,1) to the data, considering

- models with liquidity impact from only one market side (bid or ask),
- models with illiquidity impact from both market sides,
- models with liquidity imbalance impact only, i.e., with functional regressor  $x_t^{(imb)} = x_t^{(ask)} - x_t^{(bid)}$ .

Moreover, we vary the domain of prices around the quotes taken into account, using  $D = 51$  (101, 151, 201) corresponding to the domains [€0, €0.50] ([€0, €1.00], [€0, €1.50], [€0, €2.00]). By doing this, we judge whether or not it pays to take liquidity far from the quotes into account. Note that  $D$  (the width of the liquidity domain) can be viewed as an additional model parameter.

### 3.5.1 Estimation results

The goodness-of-fit is assessed by the AIC and BIC criteria. Their use in this context is somewhat critical as the number of parameters in the penalty term is  $3 + K$  (models with one FunX component) or  $3 + 2K$  (two components), ignoring the fact that the FPC scores are constructed regressors which have been estimated nonparametrically from the liquidity data.

Keeping this in mind, we find that the model with imbalance impact performs worst for all three stocks and all four domains considered, regardless of the information criterion used. This result indicates that the imbalance measure, by construction, eliminates information on the individual curves which is valuable for predicting the price process. Moreover, the models accounting for both market sides' liquidity outperform those using only one side's liquidity information in virtually all cases.

Table 3.3 shows the two goodness-of-fit criteria for these bid+ask models, fitted using  $K = 0, \dots, 5$  FPCs, where  $K = 0$  corresponds to the pure log-GARCH model. We see a considerable improvement of the fit when allowing for liquidity impact. In many cases the improvement is largest when introducing the second (not the first!) FPC, whose eigenfunction is able to induce stronger liquidity impact near the quotes than deeper in the book, see Figure 3.5. In some cases, even the fifth component improves the fit. Interestingly, the results are not very sensitive with respect to the choice of  $D$ .

<sup>3</sup>For details on construction of the snapshots and removing deterministic patterns, see Section 3.2.

Figure 3.6 depicts the functional parameter estimates for all three stocks considered and the models with both bid and ask liquidity impact. Using the conjecture that the QML estimator of  $\theta = (\omega, \alpha, \beta, \gamma^{(bid)}, \gamma^{(ask)})$ , where  $\gamma^{(s)} = (\gamma_1^{(s)} \dots \gamma_K^{(s)})'$ , is asymptotically normal, see Section 3.3.2, and assuming the estimation error for the eigenfunctions to be negligible, confidence bands are constructed as follows.

We denote the the covariance matrix of  $\theta$  by  $\Sigma_\theta$ .  $\Sigma_\theta$  contains  $K \times K$ -dimensional partitions  $\Sigma_{\gamma^{(s)}}$ , the covariance matrices of  $\gamma^{(s)}$ . We stack the corresponding eigenfunctions,  $\phi_k^{(s)}$ ,  $k = 1, \dots, K$ , in the “vector”  $\Phi^{(s)}$ . The covariance kernel of the functional parameter  $\gamma^{(s)}(m)$  is then given by

$$\Sigma_{\gamma^{(s)}}(d, m) = \Phi'^{(s)}(d) \Sigma_{\gamma}^{(s)} \Phi^{(s)}(m).$$

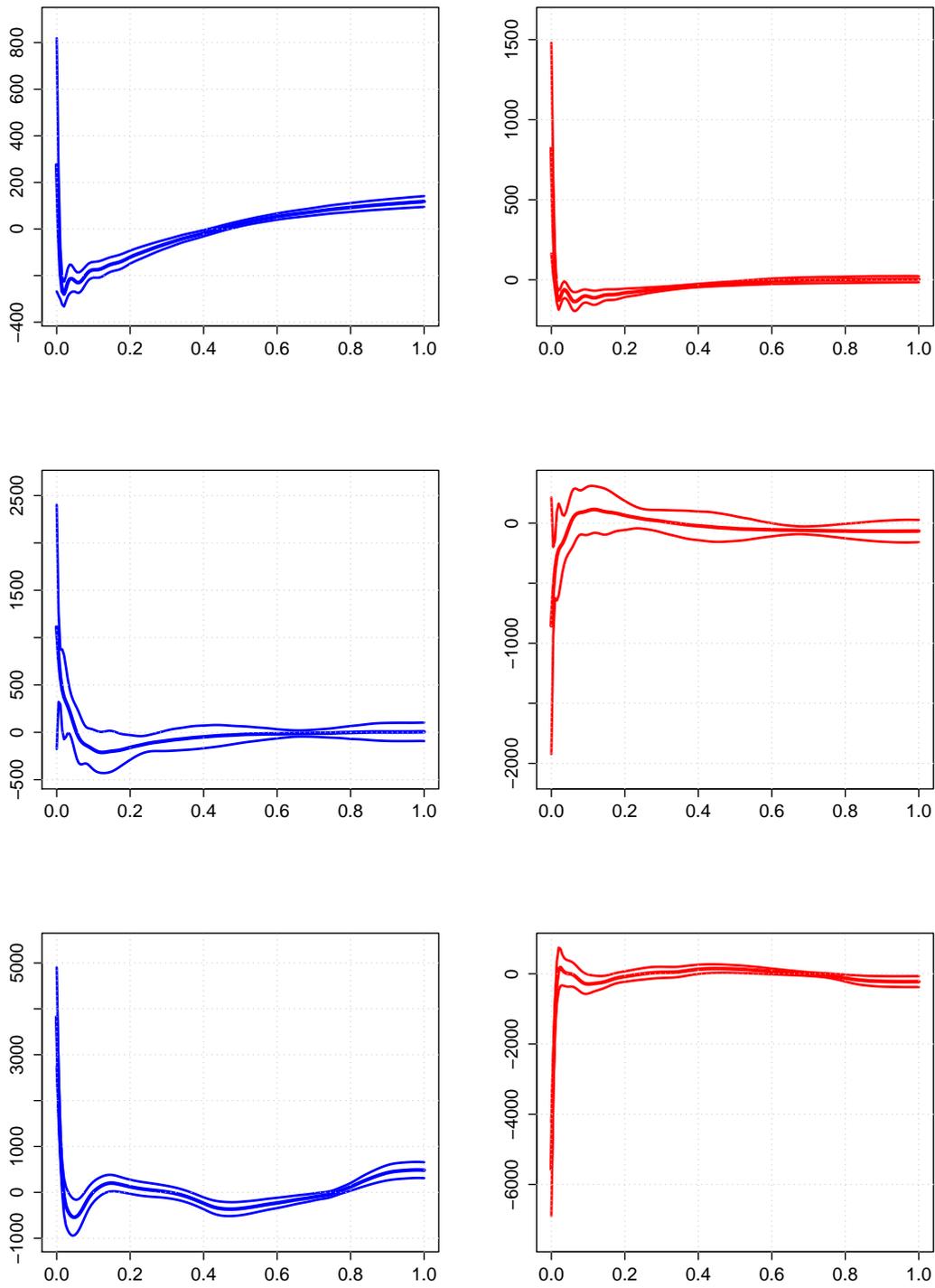
Plugging in estimates of  $\Phi^{(s)}$  and  $\Sigma_{\gamma^{(s)}}$ , standard errors of  $\hat{\gamma}^{(s)}$  are given by the square root of the kernel estimate’s diagonal,  $(\hat{\Sigma}_{\hat{\gamma}^{(s)}}(d, d))^{1/2}$ .

All estimates have in common that both the effect and its uncertainty are largest at the left edge of the domain, i.e., near the quotes. We find that confidence bands based on the stationary bootstrap are virtually identical, which can be attributed to the great stability of eigenfunction estimates as reported in Chapter 2. High bid liquidity near the quotes tends to increase, high ask liquidity tends to decrease volatility (with the exception of Commerzbank). The local impact tends to vanish for liquidity deeper in the book. However, none of the estimates is strictly positive or negative over the entire domain of relative prices, and the effect sizes are hard to interpret and compare anyway. We therefore advocate interpretation of the cumulative impact,  $LI_t$ , instead.

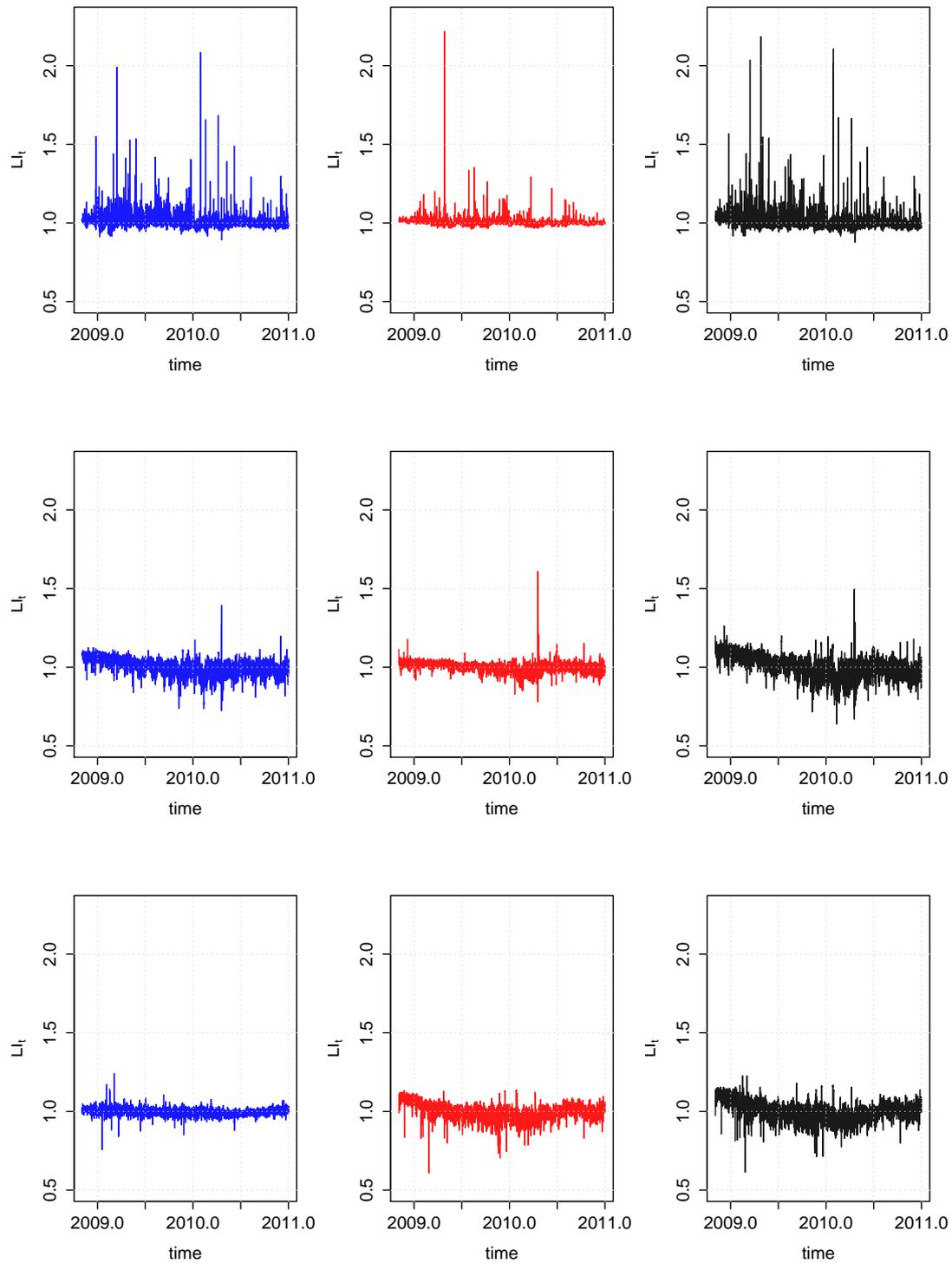
### 3.5.2 Liquidity impact

Liquidity impact is the cumulative effect of liquidity on the conditional variance over the entire domain of relative prices. Figure 3.7 depicts the estimated liquidity impact trajectories implied by the models whose functional parameter estimates have been shown in Figure 3.6. Note that liquidity impact appears to be fairly robust to the specific choice of  $K$ , results remain quite similar for  $K = 2, 4, 5$ . We find the impact not only to be time-varying, but to differ in size and direction, both between stocks and between market sides for the same stock. For instance, for the Commerzbank stock, the liquidity contribution to volatility is typically large as compared to Linde.  $LI_t$ ’s unconditional distribution is heavily skewed to the right for Commerzbank (with many volatility-increasing outbursts), but fairly symmetric or even slightly left-skewed for Linde and MunichRe.

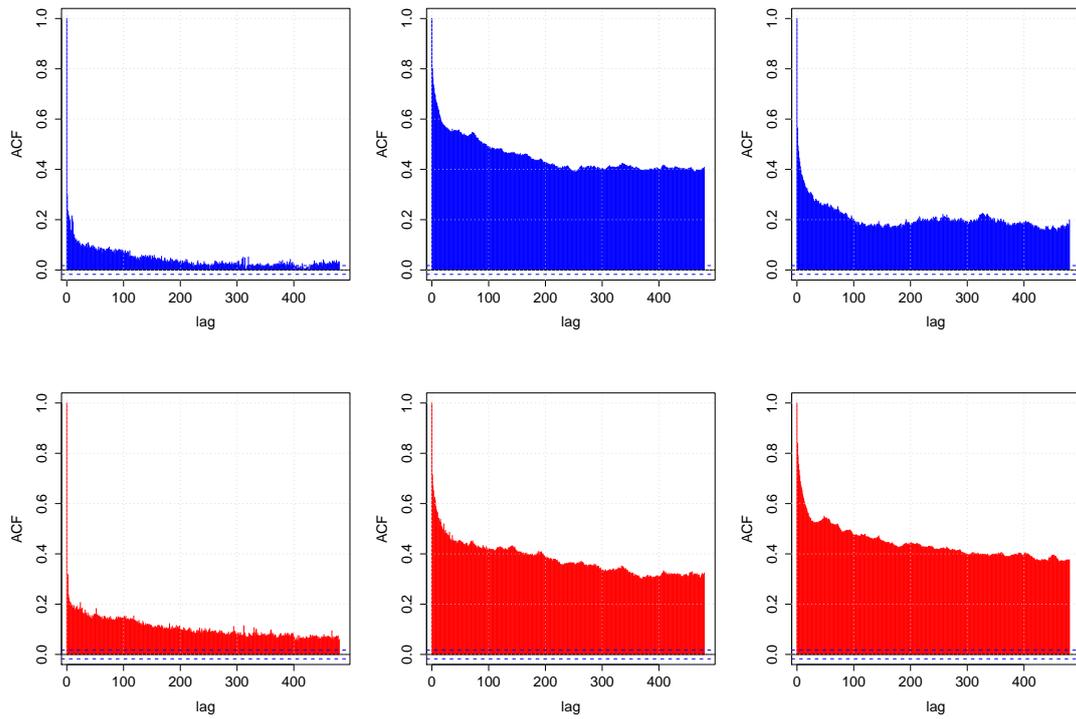
As  $LI_t$  is a linear combination of the FPC scores of the liquidity processes, it inherits its autocorrelation structure from  $\zeta_t^{(s)}$ . As most components are highly persistent, so is  $LI_t$  which is shown in Figure 3.8.



**Figure 3.6:** Estimated functional parameters for Commerzbank (top panel), MunichRe (center panel), and Linde (bottom panel), along with 95 percent confidence bands. Results are for models with both bid (left) and ask (right) liquidity impact, with  $K = 3$  liquidity components, and  $D = 201$  relative price levels. The confidence bands are based on the conjectured asymptotic normality. Estimation uncertainty from step 1 (eigenfunctions) is assumed to be negligible.



**Figure 3.7:** Estimated liquidity impact for the models with two functional liquidity processes,  $K = 3$ ,  $D = 201$ . From left to right: Bid impact (blue), ask impact (red), and cumulative impact (their product, black). From top to bottom: Commerzbank, MunichRe, and Linde.



**Figure 3.8:** Sample autocorrelation functions of the estimated liquidity impact for the models with two functional liquidity processes,  $K = 3$ , and  $D = 201$ , for the first 480 lags (roughly four weeks). Both bid (top, blue) and ask (bottom, red) impacts are highly persistent for all stocks — Commerzbank (left), MunichRe (center), Linde (right) —, and market sides.

### 3.5.3 Forecasting

Is liquidity information able to improve the forecast performance of the model? To answer this question, we conduct an out-of-sample forecast exercise. We follow Engle and Sokalska (2012) and use the negative quasi-log-likelihood (QLIKE) of one-step-ahead volatility forecasts as loss function. An alternative method of forecast evaluation would be the use of realized measures of volatility. However, such measures can be expected to be very noisy in our high-frequency setting, even if subsampling or other measures supposedly alleviating the impact of microstructure noise are used when estimating RV.

Starting with the first two thirds of the available observations,  $1 : \lfloor \frac{2}{3}T \rfloor$ , we reestimate all models considered above for each forecast step, i.e.  $T - \lfloor \frac{2}{3}T \rfloor - 1$  times, in an expanding window scheme. As before, the models accounting for both market sides yield the best results. The sum over all one-step-ahead negative log-likelihoods is shown in Table 3.4.

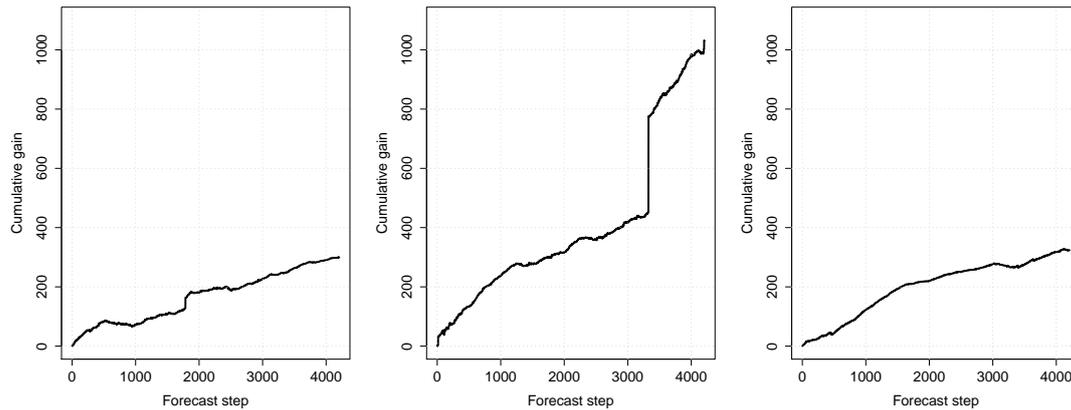
For all stocks considered, most of the GARCH-FunXL models provide highly significant improvements of QLIKE. As in forecast exercises parsimony is typically rewarded, it is noteworthy that in all cases models with  $K = 4$ , i.e., with as many as 11 parameters deliver the best forecasts. There is also strong evidence that models with  $D = 51$  or  $D = 101$  outperform their competitors, indicating that liquidity deep in the book is of limited relevance to volatility prediction. Interestingly, this finding is much more pronounced for Linde than for Commerzbank, although Commerzbank's price is much lower than Linde's throughout the sample, so that a price change by  $D$  cents implies a much higher relative change for Commerzbank than for Linde.

Figure 3.9, depicting QLIKE differences between log-GARCH-FunXL and pure log-GARCH, cumulated forecast-step by forecast-step for models with  $D = 51$  and  $K = 3$ , shows that GARCH-FunXL gains most of the improvement as compared to pure GARCH at a constant pace. Moreover, some large jumps indicate that also extreme returns can be forecast more accurately by using liquidity information.

## 3.6 Conclusion

In the present Chapter, we have put forward a class of semiparametric GARCH-X models with functional exogeneous variables. The model is able to capture the impact of liquidity as implied by a limit order book on asset price volatility. In simulations and applications, we have confined ourselves to a log-GARCH version of the model which conveniently allows for the inclusion of complex, potentially negative functional predictors. In many aspects, linear GARCH models are better understood and more tractable than log-linear versions like log-GARCH or EGARCH. Therefore, an alternative GARCH-FunXL specification could for example use the framework of Amado and Teräsvirta (2013), i.e., a product of a linear endogeneous GARCH part and a suitable transformation of the exogeneous functional variables.

The GARCH-FunXL model has been shown to be successful in predicting intraday volatility in an application to the German XETRA limit order book. While originally



**Figure 3.9:** Cumulative forecast gain,  $QLIKE(\text{pure log-GARCH}) - QLIKE(\text{log-GARCH-FunXL})$ , of liquidity-driven GARCH models ( $K = 3$ ,  $D = 51$ ), computed forecast-step by forecast-step, as compared to models without liquidity impact, for Commerzbank (left), MunichRe (center), and Linde (right).

taylorized for limit order book data, the model may be useful in other fields of application as well. For example, the term structure of interest rates can be viewed as a functional time series. A further extension that we address in Chapter 4 is financial duration modeling with functional exogenous liquidity. The duration analogue of the log-GARCH-FunXL in this Chapter is an extension of the log-ACD of Bauwens and Giot (2000).

$\gamma(m)$	$T = 1000$			$T = 5000$			$T = 10000$			
	ind. (a)	dep. (b)	dep. (c)	ind. (a)	dep. (b)	dep. (c)	ind. (a)	dep. (b)	dep. (c)	
(i)	0.0095 (0.0061)	0.0103 (0.0073)	0.0042 (0.0032)	0.0016 (0.0011)	0.0018 (0.0011)	0.0008 (0.0006)	0.0008 (0.0005)	0.0009 (0.0005)	0.0003 (0.0002)	0.0022 (0.0028)
(ii)	0.0092 (0.0057)	0.0096 (0.0058)	0.0039 (0.0032)	0.0016 (0.0011)	0.0019 (0.0012)	0.0006 (0.0004)	0.0010 (0.0006)	0.0008 (0.0005)	0.0004 (0.0003)	0.0018 (0.0023)
(iii)	0.0097 (0.0072)	0.0108 (0.0067)	0.0036 (0.0028)	0.0018 (0.0010)	0.0020 (0.0012)	0.0006 (0.0005)	0.0009 (0.0005)	0.0008 (0.0006)	0.0003 (0.0003)	0.0004 (0.0003)

**Table 3.2:** Mean (standard deviation) of the MSE of liquidity impact computed over  $B = 100$  simulation runs, for functional parameters (i)-(iii), serially independent (ind.) vs. dependent (dep.) score processes, and unconditional covariance matrices (a) to (c) (with increasing diagonal elements). Innovation covariance matrices are diagonal with elements decaying from top left to bottom right. In case of serial independence, the eigenstructure is the empirical one (Commerzbank ask curves) up to a constant. Dependence is VAR(1) with autoregressive matrix estimated from Commerzbank ask curves' FPC scores.

		Commerzbank							
		AIC				BIC			
K		51	101	151	201	51	101	151	201
0		42995	42995	42995	42995	43017	43017	43017	43017
1		42977	42980	42982	42974	43014	43017	43020	43012
2		42650	42807	42874	42837	42702	42859	42926	42889
3		42631	42638	42643	42598	42698	42705	42710	42665
4		42608	42611	42638	42598	42690	42693	42719	42680
5		42551	42603	42612	42509	42648	42699	42709	42606
		MunichRe							
		AIC				BIC			
K		51	101	151	201	51	101	151	201
0		41787	41787	41787	41787	41809	41809	41809	41809
1		41612	41466	41410	41366	41650	41504	41447	41404
2		41119	41039	41061	41057	41171	41091	41113	41109
3		41105	41039	41050	41035	41172	41106	41117	41102
4		41069	41033	41042	41028	41151	41115	41124	41110
5		41062	41034	41031	41029	41159	41130	41128	41125
		Linde							
		AIC				BIC			
K		51	101	151	201	51	101	151	201
0		43517	43517	43517	43517	43539	43539	43539	43539
1		43236	43205	43200	43205	43273	43242	43238	43242
2		43212	43206	43204	43208	43264	43258	43256	43260
3		43210	43202	43196	43188	43277	43269	43263	43255
4		43189	43187	43185	43184	43271	43269	43267	43266
5		43192	43188	43181	43184	43289	43285	43278	43281

**Table 3.3:** Goodness-of-fit measures of the models with bid and ask liquidity impact for the three stocks. The number of price levels apart from the quotes considered when constructing the liquidity curves is  $D = 51$  (101, 151, 201), corresponding to the domains  $[\text{€}0, \text{€}0.50]$  ( $[\text{€}0, \text{€}1.00]$ ,  $[\text{€}0, \text{€}1.50]$ ,  $[\text{€}0, \text{€}2.00]$ ).

		Commerzbank				MunichRe				Linde			
K		51	101	151	201	51	101	151	201	51	101	151	201
0		2912	2912	2912	2912	4066	4066	4066	4066	4700	4700	4700	4700
1		2977	2785	2858	2860	3822	3767	3613	3554	4425	4345	4840	4884
2		2586	2764	2759	2758	3019	3284	3165	3091	4381	4337	4822	4856
3		2608	2515	2567	2573	3035	3290	3103	3020	4377	4328	4824	4856
4		2613	2602	2619	2577	2993	3248	3142	3067	4293	4303	4807	4848
5		2619	2608	2663	2782	3034	3242	3193	3068	4302	4266	4796	4838

**Table 3.4:** Negative out-of-sample likelihoods for one-step-ahead forecasts of GARCH-FunXL models with bid and ask liquidity impacts. The number of price levels apart from the quotes considered when constructing the liquidity curves is  $D = 51$  (101, 151, 201), corresponding to the domains  $[\text{€}0, \text{€}0.50]$  ( $[\text{€}0, \text{€}1.00]$ ,  $[\text{€}0, \text{€}1.50]$ ,  $[\text{€}0, \text{€}2.00]$ ).



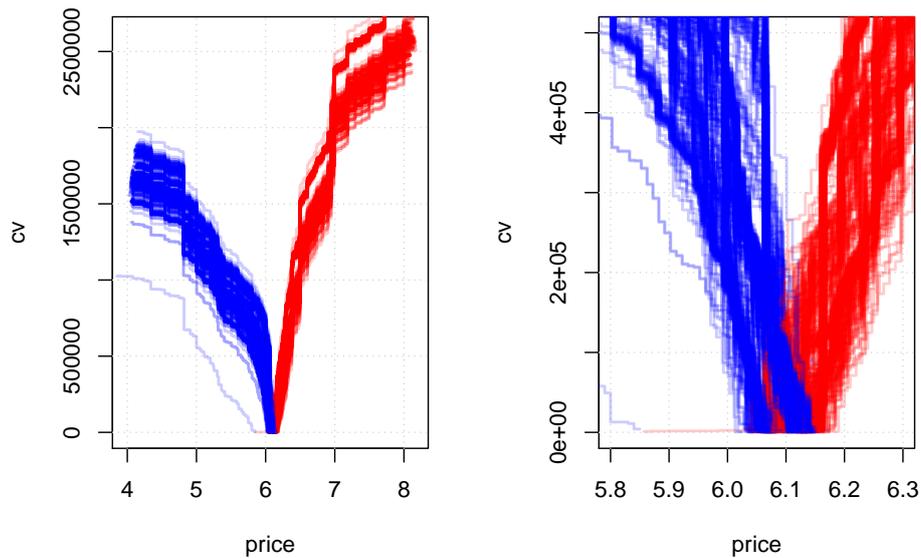
## Functional liquidity and low-latency volatility: The ACD-FunXL model

### 4.1 Introduction

Recent research on securities trading has been mainly devoted to and pointed out the importance of speed in both gathering and processing information, as a major proportion of order submissions and trading volume in electronic limit order books (LOB) can be attributed to (algorithmic) high frequency trading (HFT). Many of these efforts have focused on the analysis of the impact of market *events* — such as market order submissions and cancellations, aggressive limit orders, and so on — on trading activity and volatility (Large, 2007; Hautsch and Huang, 2012; Cont et al., 2014). Further steps into this direction include the identification of algo strategies and their impact on market quality and stability. This identification is difficult, however, as the presence of and, thus, the effect of HFT is typically not directly observable for the researcher (Hasbrouck and Saar, 2013).

Whereas limit order submissions and cancellations near the quotes, occurring at extremely high frequency, drive the small increments of an asset's price, limit order schedules at greater price distance to the quotes may still be worthwhile accounting for. They may eventually reflect the asset's overall liquidity better than the rather noisy bouncing of offers and requests at the quotes. In many situations, information-driven market participants may abstain from silent order submission strategies which they design to avoid a large impact on the price, but whose implementation needs some time. They may rather submit large market orders as long as sufficient liquidity is present in the market. This is possible when limit order schedules can be expected to be quite stable over a trading day. As we will show, this tends indeed to be the case to some extent, in spite of single limit orders being submitted, cancelled, and executed at millisecond frequency.

As in the previous Chapter, we do not focus on the price impact of single liquidity events, but of entire limit order schedules. The GARCH-FunXL model was designed to model volatility at high, but not ultra-high frequency, thereby avoiding microstructure effects. In contrast, in this Chapter we aim at modeling ultra-high frequency



**Figure 4.1:** State of a limit order book over the trading day. Liquidity is represented by the available requested (left) and offered (right) cumulative number of shares as a function of the price. The quantities which may be called *prices*, i.e. bid and ask quotes, are implications of the limit order schedules. The bid quote is the rightmost point of each blue curve, the ask quote is the leftmost point of each red curve. Snapshots of the LOB are depicted for the Commerzbank stock on October 22, 2010. They are taken each minute during the trading day, making 511 snapshots per day. The right plot zooms in to prices between EUR 5.80 and 6.30 and cumulative sizes up to 500,000 shares.

price variation. To this end, we consider the instantaneous volatility of quoted prices on the bid and ask sides of the market. We introduce an econometric model which links curve-valued cumulative volume to instantaneous quote volatility in an autoregressive conditional duration (ACD) framework.

Figure 4.1 shows cumulative volume curves of the Commerzbank stock, sampled over the course of one trading day at a frequency of one minute. In the left plot, which depicts curves over a range of  $\pm 200$  Cents around the quotes, it seems that the basic structure of the curves is preserved throughout the trading day. That is, the curves seem to be only shifted along the price axis due to quote changes while retaining their overall shape. However, the plot in the right panel, zooming in a bit, reveals that the curves exhibit several time-varying features. The question we would like to answer in this Chapter is to what extent the variation of these curves is able to explain quote variation.

In our econometric model, importing ideas from functional data analysis (FDA), these entire curves, capturing the complete LO schedules, are treated as sampled units and used as regressors supposed to explain bid and ask price durations and, thus, the quotes' instantaneous volatility (Engle and Russell, 1998). Using the concept of liquidity impact, we find that LO schedules, including LOs that are placed at some

distance from the quotes, help to explain future moves of quoted prices' volatility to a considerable extent.

The idea to capture all LOs instead of volume at the quotes only is not entirely new, there are several studies implicitly or explicitly using this notion. Engle et al. (2012) analyze liquidity-volatility interactions for US Treasury notes, where liquidity is observed at 5 tiers behind the quotes, and even use a MEM framework as we do — the primary quantity of their interest is the realized variance (see also Chapters 5 and 6 of this thesis). In van Kervel (2015), the measures *DepthAsk* and *DepthBid* are used which are very similar to the cumulative volume (CV) measure used in the present study. Gomber et al. (2015) employ the *exchange liquidity measure (XLM)*, which is the average cost per share of a roundtrip trade and, thus, closely related to CV. However, all these studies evaluate the intrinsically curve-valued liquidity measure at *one specific, arbitrary* value of its argument. They do not account for the shapes of entire curves.

In contrast, Gouriéroux et al. (1998), Bowsher (2004) in what would later become Bowsher and Meeks (2008), and Härdle et al. (2012b) capture full LOB-implied liquidity, but focus solely on the curves' dynamics (as we do in Section 2.5) instead of interactions between liquidity and price dynamics.

The model introduced in the present Chapter allows to answer the following questions: What is the structure of liquidity impact on high-frequency volatility of quoted prices? In particular: Which parts of the LOB (how many price levels in the LOB) are relevant for quote dynamics? What is the role of the tick size? What is the role of the market sides? Is the impact of liquidity on quote variation symmetric or asymmetric?

Apart from the application to price durations, the method may be adapted to analyze further non-negative time-varying quantities of interest like trading volume, or realized measures of volatility.

The Chapter proceeds as follows. Section 4.2 explains the duration-based notion of low-latency volatility and offers a first empirical look at low-latency quote variation. Section 4.3 introduces the functional liquidity concept in a high-frequency setting. In particular, the task of de-diurnalizing non-equidistant curves is addressed. This task is more involved in the present setting than it is in the equidistant setting of Chapters 2 and 3. Section 4.4 introduces the model and discusses its properties. The remaining Sections present the empirical results, i.e., in- and out-of-sample performance, liquidity impact, and functional parameter estimates, and finally conclude.

## 4.2 Low-latency volatility

Following Engle and Russell (1998), starting from the initial price  $P_0$  at the beginning of continuous trading  $\mathcal{T} = [0, T]$ , we measure the waiting times between quote changes of at least  $\delta$ , the *price durations*, for both market sides. The smallest sensible  $\delta$  amounts to one tick. The collection of event times can then be written  $\{t_1, t_2, \dots, t_{N_{T-1}}, t_{N_T}\}$ , where  $N_t$  counts the number of events in  $[0, t]$ . We denote the point process allocating the event times along the time axis by  $(N_t)_{t \in \mathcal{T}}$ . It will be specified more

explicitly below. For the moment it suffices to say that the probabilistic structure of the point process is uniquely determined by its conditional intensity function (Daley and Vere-Jones, 2003). Denoting the past of the point process until (but not including)  $t$  by  $\mathcal{F}_t$ , we define the conditional intensity as

$$\lambda(t|\mathcal{I}_t) = \lim_{\Delta \rightarrow 0} \frac{E[N(t + \Delta) - N(t)|\mathcal{F}_t]}{\Delta}, \quad (4.1)$$

that is, for small  $\Delta$  we can write more instructively,

$$\lambda(t|\mathcal{I}_t)\Delta \approx E[N(t + \Delta) - N(t)|\mathcal{F}_t], \quad \Delta > 0.$$

This means that within a small interval,  $[t, t + \Delta]$ , the conditional intensity can be interpreted as the expected number of events, the expected “local frequency”, of the point process. The conditional intensity can also be expressed in terms of the ratio of the conditional density and the conditional survivor function of the point process.

Let us assume for simplicity that the price process consists of jumps of size  $\delta$  only, upward and downward jumps are equally likely, and denote the price by  $P_t$ . This is certainly a good description of the price process if  $\delta$  amounts to the tick size, and a legitimate simplification otherwise. Then its *instantaneous variance* (Engle and Russell, 1998) can be defined as

$$\sigma^2(t|\mathcal{F}_t) = \lim_{\Delta \rightarrow 0} E \left( \frac{1}{\Delta} \left[ \frac{P_{t+\Delta} - P_t}{P_t} \right]^2 | \mathcal{F}_t \right), \quad (4.2)$$

which can be decomposed into a jump size and a conditional frequency part, leading to the expression

$$\sigma^2(t|\mathcal{F}_t) = \frac{\delta^2}{P_t^2} \lambda(t|\mathcal{F}_t). \quad (4.3)$$

Note that this instantaneous notion of volatility is in contrast to *Part II* of this thesis, in which *realized* concepts of volatility are discussed. In the realized volatility literature, parts of the low-latency volatility concept put forth here would be termed *microstructure noise*, especially when  $\delta$  is small. On the other hand, Tse and Yang (2012) argue that the realized variance may be measured very accurately by means of (integrated) instantaneous variance, given that  $\delta$  is chosen appropriately (typically much larger than 1 tick) as a means to robustify against microstructure noise. Moreover, a growing literature on the estimation of the so-called *spot volatility* has been established recently, which addresses the problem of estimating the latent stochastic variance using high-frequency data and, possibly, in the presence of microstructure noise (Munk et al., 2010; Alvarez et al., 2012). Instantaneous volatility is intimately linked to the concept of spot volatility, with the notable difference that we approach

the problem using an explicit (semi-) parametric model instead of nonparametrically. For further, timely discussions of quote-implied vs. “fundamental” volatility from the perspective of financial economics, see Hasbrouck (2015) and Conrad et al. (2015).

### The XETRA LOB data: Quotes

The permissible price levels are the positive multiples of the tick size. The tick size on XETRA depends on the price. For our sample period, it amounts to EUR 0.001 if the price of the instrument is under EUR 10, EUR 0.005 for prices in the interval [EUR 10, EUR 50), 1 Cent for prices in [EUR 50; EUR 100], and 5 Cents otherwise. The data set allows us, in principle, to reconstruct the book at any time during a given trading day. Continuous trading on XETRA starts after a 5-minute opening auction at 9am, is interrupted by another auction of this type at 1pm, and ends just before the closing auction at 5:30pm.

We confine ourselves to the analysis of continuous trading, considering all orders within [9:05am; 12:59pm] and [1:05pm; 5:29pm]. Moreover, we choose the parameter  $\delta$ , which determines the price duration series, to amount to certain multiples of the tick size. Table 4.1 informs about the frequencies of such events during October 2010 for Commerzbank (with tick size of 0.1 Cents) and Linde (1 Cent). It is interesting to see that, although the stocks’ behaviors as well as the companies they represent differ in many respects, the number of price durations is quite similar with respect to  $\delta$ . We will return to this when analyzing the liquidity impact on price durations.

$\delta$ [Ticks]	Commerzbank					Linde				
	1	2	3	4	5	1	2	3	4	5
bid quote	65030	32278	27735	17359	11601	47855	27366	24433	18565	9382
ask quote	70241	35160	29958	18456	12079	51679	29201	26056	19618	9858

**Table 4.1:** Number of quote changes of at least  $\delta$  ticks in 10/2010 for Linde and Commerzbank. The tick size is 1 Cent for Linde and 0.1 Cent for Commerzbank.

In Section 4.4, we construct an explicit semiparametric model of low-latency variation  $\sigma^2(t|\mathcal{F}_t)$ . Before, we introduce the liquidity part of this model.

## 4.3 Low-latency liquidity in a limit order book

In this Section, we turn to the liquidity aspects of the XETRA LOB data with an emphasis on the high-frequency setting. We then introduce the notion of functional liquidity and analyze liquidity curves sampled at ultra-high frequency empirically.

### 4.3.1 Latency of orders

The analysis of Hasbrouck and Saar (2013) focuses on the speed at which orders are submitted, cancelled, and resubmitted by algorithmic traders in what they call the *millisecond environment*. This is certainly true for the XETRA LOB as well, as Figure 4.2 shows. About half of the limit orders are executed or cancelled within one second

after submission. More than 90 percent of them are cancelled. In the following, we will not focus on this aspect but on the structure and dynamics of the resulting LO inventories.

### 4.3.2 Liquidity measure

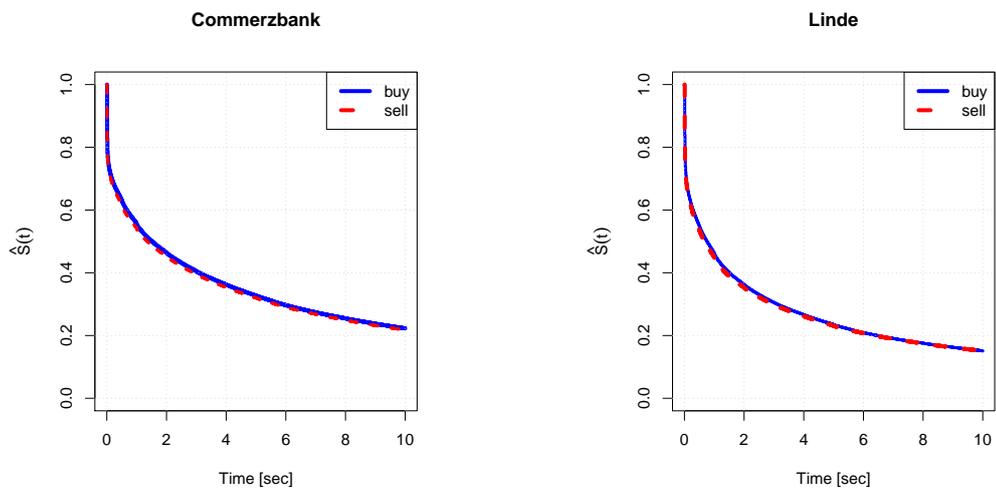
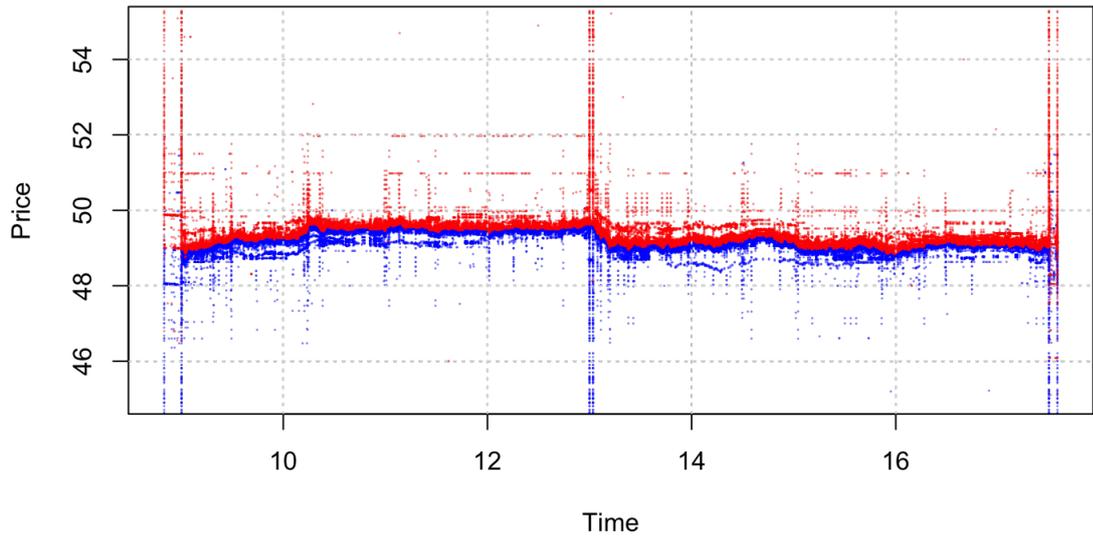
We use cumulative volume as our liquidity measure. The notation is as before, that is, cumulative volume at relative price  $d$ , market side  $s$  and at the end of the  $i$ th duration is denoted by  $x_{t_i}^{(s)}(d)$ . Cumulative volume curves are termed (bid or ask) *liquidity* in the remainder of the Chapter.

One may argue that depth, i.e. economic value assigned to a security, is better comparable between stocks whose prices differ than pure volumes are. However, trading conditions differ in many more respects: For instance, the tick size usually depends on the quoted price itself, restricting the relation between (potential) variation of a price and its level. Apart from that, the basic findings for depth and volume can be expected to be quite similar if the period under study is rather short.

As already indicated before, some studies already make use of this or similar "functional" measures of liquidity, among others van Kervel (2015), Gomber et al. (2015), and Engle et al. (2012). However, whatever specific liquidity function is used, in all these studies it is evaluated at one specific value of the argument (the relative price  $d$  in case of cumulative volume and the measure of van Kervel (2015), the size  $n$  of a roundtrip trade in case of the XLM).

Limit order and trade information is time-stamped to the nanosecond in our data set, such that  $LOB_t$  can be constructed virtually for any  $t$  during the trading day. The nanosecond precision of stamps leads to roughly 30 trillion, potentially different, observable states of the LOB during a single trading day. However, this is obviously neither feasible nor sensible. The maximum speed at which market participants act, even if trading purely algorithmically, is much slower. It depends on the distance to the stock exchange, the quality of the cables transferring the orders, and the computational speed of participants. For instance, consider a market participant who is located at a distance of 1 km away from the stock exchange. Even if she computed and reacted infinitely fast and had a perfect communication with the stock exchange, her reaction would still be limited by the speed of light which travels at "only" 30 centimeters per nanosecond. Therefore, a realistic "speed of the market" is slower at least by a factor of roughly  $10^4$  compared to the nanosecond environment.

The number of LOB updates (i.e. market order submissions, limit order submissions and cancellations) is usually much smaller (typically a 6-digit number for major stocks traded on XETRA) than the number which is technologically possible, and the number of "major updates", to be defined below, even more so. On the other hand, there are periods of high activity where thousands of revisions occur within only a few seconds. We finally note that by the very nature of an electronic market place, we do not only observe intended interactions between market participants, but also those interactions that are due to submission strategies already in place, for example by means of iceberg orders.



**Figure 4.2:** Top panel: All limit orders submitted during an exemplary trading day, buy (sell) orders in blue (red), for the Linde stock. Bottom panel: Unconditional empirical survivor functions of limit buy (solid) and sell (dashed) orders in 10/2010 for Commerzbank (left) and Linde (right) stock, based on roughly 1.24 to 1.28 million orders each.

### 4.3.3 Functional time series methods at high frequency

In the following, let  $(t_i)_i$ ,  $i = 1, \dots, I$ , denote the sequence of sampling times during a trading day. The indexed events defining the sequence are not equidistant (in calendar time). The collections of observed liquidity curves,  $\{x_{t_i}^{(bid)}\}_i$  and  $\{x_{t_i}^{(ask)}\}_i$ , are assumed to be generated by stationary functional stochastic processes. More specifically, although the inter-event times or durations,  $t_i - t_{i-1}$ , are random, we assume that the curves are stationary in the sense of Hörmann and Kokoszka (2010) in operational time  $i$ . In particular, the curves have a constant mean function,  $\mu^{(s)}(d)$ , and a contemporaneous covariance function between different locations on a random curve,  $\text{Cov}(x_t^{(s)}(d_j), x_t^{(s)}(d_k))$ , in common, and exhibit a spectral representation in terms of the contemporaneous covariance operator's eigenvalues and eigenfunctions, all in close analogy to the properties of vector-autoregressive multivariate time series models.

Based on the eigenrepresentation, liquidity curves can then be represented via the Karhunen-Loève decomposition,

$$x_t(d) = \mu(d) + \sum_{j=1}^{\infty} \xi_{j,t} \phi_j(d),$$

or functional principal component (FPC) representation. The eigenvalues  $\lambda_j$  of the spectral representation are equal to the unconditional variances of the FPC scores  $\xi_{j,t}$ . As the eigenvalues are strictly decreasing, the FPCs are sorted by their contribution to the  $x_t$ 's (unconditional) variation. This gives rise to the  $K$ -truncated FPC representation

$$x_t(d) = \mu(d) + \sum_{j=1}^K \xi_{j,t} \phi_j(d) + v_t(d),$$

where  $v_t(d) = \sum_{j=K+1}^{\infty} \xi_{j,t} \phi_j(d)$  is the truncation error.

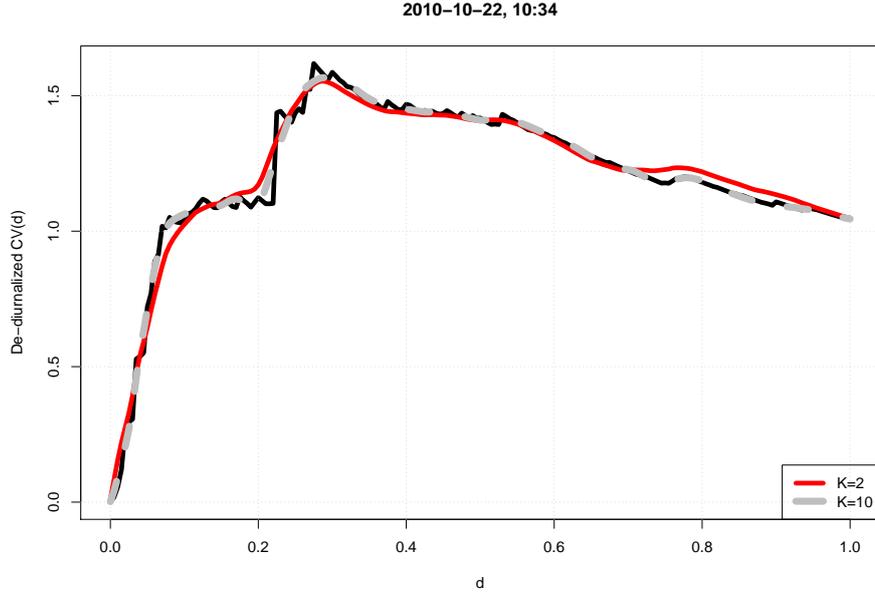
Steep cumulative volume curves can be considered to represent high, flat curves to imply low liquidity.

Figure 4.3 depicts an exemplary liquidity curve along with two approximations in terms of functional principal components. The curve is already de-diurnalized in a way which will be explained in detail in the next Section.<sup>1</sup> We find that the most important modes of variation are explained by only two eigencomponents. Nevertheless, the seemingly unimportant additional features represented by the third to tenth FPC may still be able to explain quote variation to some extent.

### 4.3.4 Diurnal patterns of liquidity

Just as the volatility of quoted prices, liquidity exhibits certain regularities during a trading day that can be treated as deterministic. For liquidity, the pattern is a bivariate

<sup>1</sup>In contrast to de-diurnalization as done for equidistant snapshots, see Chapter 3, a penalized tensor product spline is used, so that the mean of de-diurnalized curves is not constantly at 1 over  $d \in [0, 1]$ .



**Figure 4.3:** De-diurnalized ask liquidity curve for the Commerzbank stock at 10:34 on 2010-10-22, along with its FPC representation using  $K = 2$  (red) and  $K = 10$  (grey) eigenfunctions. Evidently, the mean plus two eigenfunctions provide already a very good approximation, so that adding even eight more components does only lead to a minor further improvement. Estimated eigenfunctions are smooth as their covariance kernel has been smoothed.

function of both time of day and relative price whose shape will be shown in brief.

We assume that each market side has its own pattern. Introducing a third clock,  $\ell$ , and distinguishing between raw and adjusted liquidity, let  $\tilde{x}_{\ell,t}^{(s)}$  denote “raw”, observed liquidity at time  $t$  during day  $\ell$ . It is given by

$$\tilde{x}_{\ell,t}^{(s)}(d) = v_t^{(s)}(d)x_{\ell,t}^{(s)}(d),$$

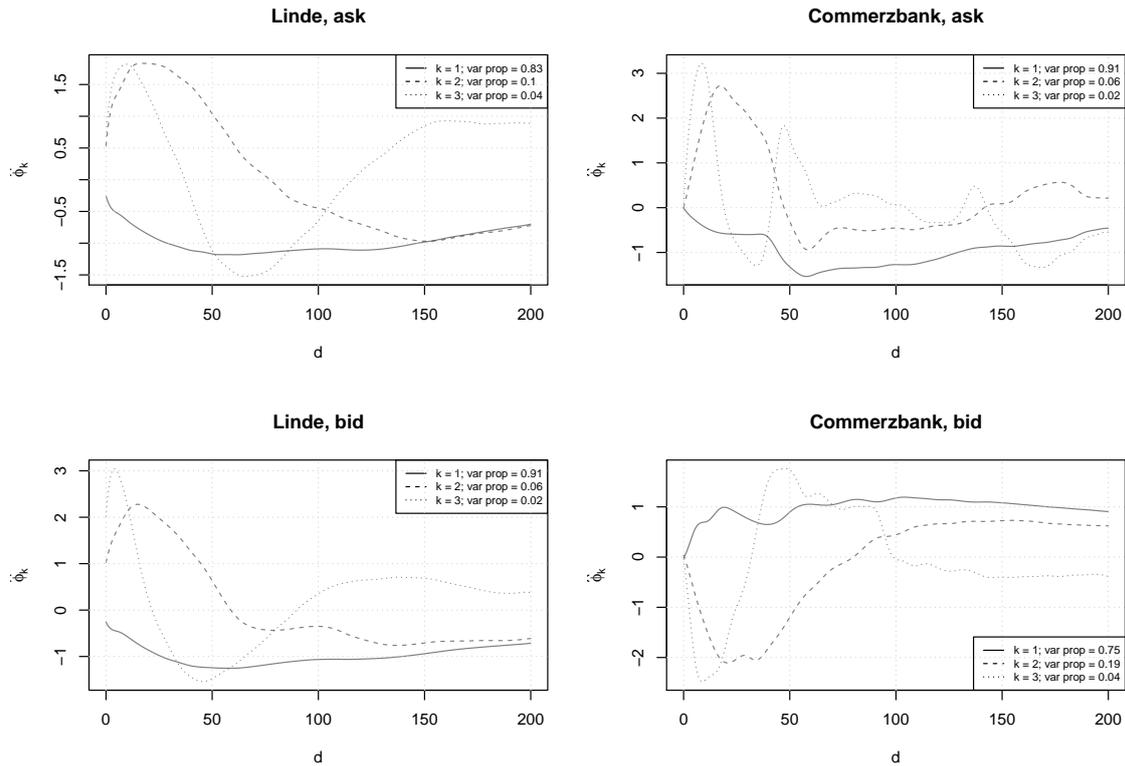
where  $v_t(d)$  is the deterministic diurnal liquidity surface and  $x_{\ell,t}^{(s)}(d)$  the stochastic liquidity component, which is of primary interest in our analysis.

In the situation considered in the previous Chapter, liquidity is observed regularly over time, as LOB snapshots are taken at constant frequency, and over  $d$ , as the relative price grid is equidistant and the same for all curves. While retaining the second property, snapshots in the present context are taken at random, irregularly spaced times which differ from day to day.

To estimate the pattern, however, we take *equidistant* snapshots (constant  $t_i - t_{i-1}$  for all  $i = 1, \dots, I$ ) at a frequency of one minute for all days in our sample of  $L$  trading days. A raw estimate of the pattern is then given by

$$\hat{v}_{t_i}^{(s)}(d_j) = L^{-1} \sum_{\ell=1}^L \tilde{x}_{\ell,t_i}^{(s)}(d_j),$$

where  $d_1, \dots, d_J$  is the observation grid along the price axis. The resulting raw estimate forms a  $I \times J$  matrix. We smooth this raw estimate using a tensor product

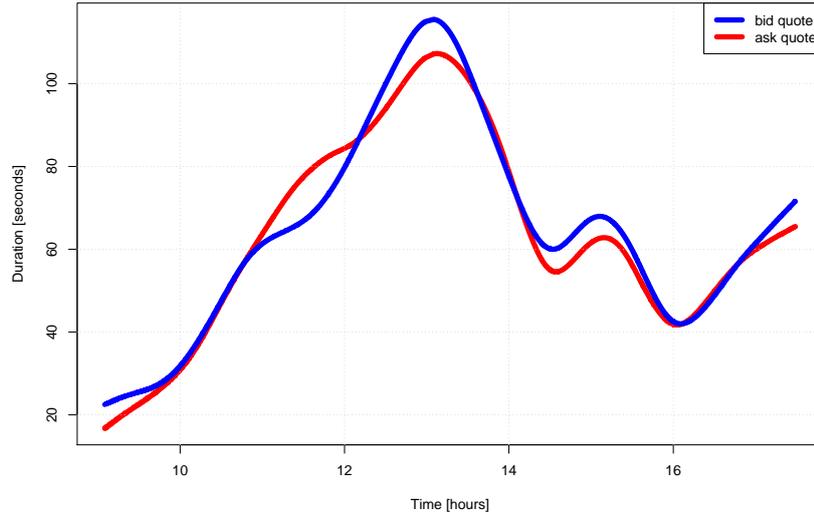


**Figure 4.4:** First three eigenfunction estimates, each for the full domain ( $D = 201$ ), based on  $\delta = 5$  tick price durations, Commerzbank and Linde, 10/2010. Note that eigenfunctions are identifiable only up to the sign, meaning that, e.g., all four first components capture the overall liquidity “level”, regardless if they are positive or negative in sign. The remaining components more or less capture different modes of variation near the quotes. Reported explained variance proportions slightly overestimate the true ones as only the first five components’ eigenvalues are taken into account. Typically, two components capture roughly 95 percent of overall liquidity variation. The eigenfunctions are quite similar to the medium-frequency case in the previous Chapter.

spline, which is basically the same method that is also used to smooth the covariance function of the liquidity curves (Xiao et al., 2013). The result is still a  $I \times J$  matrix. However, as events occur at arbitrary times during the trading day, we need to evaluate the diurnal surface not only at the gridpoints, but potentially for all  $t \in \mathcal{T}$ . As we have a regular grid for  $d$ , we interpolate the  $I$  estimates for each relative price tier  $d_j$  using a cubic spline. That is, our final surface estimate amounts to  $J$  cubic spline estimates for each market side.

### 4.3.5 Diurnal duration patterns

Finally, we also account for the diurnal pattern of price durations themselves. We fit a smoothing spline to the scatterplot of  $y_i$  against  $t_i$ , pooled for all trading days. The amount of smoothness is determined manually and is the same for all  $\delta$  considered. It is usually a bit higher than suggested by GCV when selected in a data-driven way. The pooling of all durations causes that volatile days are represented by more



**Figure 4.5:** Diurnal duration pattern for Commerzbank 5-tick price durations, computed by fitting a smoothing spline to the scatterplot of all durations vs. event times during October 2010.

observations than tranquil days, leading to a slight downward bias. However, we have compared this scheme with one that pools pre-averaged data for each day, finding that the difference is negligible. At the same time, by using all observations we decrease the variability of the estimation. All in all, we expect to be better off with our strategy as compared to the pre-averaging method.

Figure 4.5 shows the estimated diurnal pattern for Commerzbank in October 2010. Apart from the well-known U-shape of volatility, which here manifests as inverse U-shape of price durations, we see that the opening of US markets in the afternoon (at 3:30pm) reduces durations considerably. Moreover, the pattern is very similar for bid and ask prices.

## 4.4 Econometric model

### 4.4.1 The model

Let  $(\tilde{y}_i)_i$  be the sequence of (raw) price durations, i.e., durations between price changes of at least  $\delta$ , associated with a continuous-time process  $(P_t)_t$  of quoted prices. The event times from which durations are constructed are denoted by  $t_i$ ,  $t_i = t_{i-1} + \tilde{y}_i$ .  $(P_t)_t$  exhibits a deterministic diurnal volatility pattern which  $(\tilde{y}_i)_i$  inherits in form of a diurnal duration pattern,  $s_t$ .

Let us for the moment use a more explicit notation and write  $\tilde{y}_{\ell,t_i}$  for the  $i$ -th raw duration on trading day  $\ell$ . Durations are generated by

$$\tilde{y}_{\ell,t_i} = y_{\ell,t_i} s_{t_i}, \quad (4.4)$$

$$y_{\ell,t_i} = \psi_{\ell,t_i} \varepsilon_{\ell,t_i}, \quad \varepsilon_t \stackrel{iid}{\sim} \text{Exp}(1), \quad (4.5)$$

where  $\psi_{\ell,t_i}$  is the conditional mean of  $y_{\ell,t_i}$ , given a conditioning information set to be made explicit in a moment. This setup is the same as in most of the intraday volatility (Andersen and Bollerslev, 1997, 1998; Engle and Sokalska, 2012) and duration literature. In the remainder of the Chapter, we exclusively work with de-diurnalized durations  $y_{\ell,t_i}$ .

Note that the model generates only intraday durations, and assumes that  $\psi_{\ell,t_i}$  stays constant between the trading hours of subsequent trading days. The conditional duration follows an ACD specification which is augmented by exogeneous information in terms of the liquidity curves at the beginning of each intraday interval from both market sides, i.e.

$$\psi_{\ell,t_i} = f \left( y_{\ell,t_{i-1}}, \dots, x_{\ell,t_{i-1}}^{(ask)}, x_{\ell,t_{i-1}}^{(bid)} \right), \quad (4.6)$$

where  $y_{\ell,t_{i-1}}, \dots$  denotes the entire duration history, in the following denoted by  $\mathcal{F}_{\ell,t_{i-1}}$ . As explained before, the liquidity curves are, as the  $y_{\ell,t_i}$ , de-seasonalized quantities.

As we are primarily interested in de-seasonalized durations, we lighten notation in the following by dropping the  $\ell$ , and by writing  $i$  instead of  $t_i$ . Durations, even if observed over several days, are indexed by  $i = 1, \dots, I$ . Furthermore, we consider a log-EACD specification<sup>2</sup> (Bauwens and Giot, 2000) whose endogeneous part has order (1,1), and choose the functional liquidities to enter the model in a log-linear fashion as well. The conditional log-duration then becomes

$$\begin{aligned} \log \psi_i &= \omega + \alpha \log y_{i-1} + \beta \log \psi_{i-1} \\ &+ \int_0^1 \gamma^{(ask)}(m) x_{i-1}^{(ask)}(m) dm + \int_0^1 \gamma^{(bid)}(m) x_{i-1}^{(bid)}(m) dm. \end{aligned} \quad (4.7)$$

We denote this larger information set, consisting of past returns plus liquidities at  $i - 1$ , by  $\mathcal{F}_{i-1}^L$ .

Collecting all these and the ingredients from Section 4.3, the logarithmic ACD(1,1)-FunXL model for returns with conditional volatility influenced by  $x_{i-1}^{(s)}$ ,  $s \in \{\text{bid}, \text{ask}\}$ , can be defined as follows.

<sup>2</sup>More specifically, we consider what Fernandes and Grammig (2006) call a log-ACD (type I) specification. The authors report good overall results for this specification in contrast to type II, which resembles an EGARCH, and as compared to other models belonging to the AACD family in empirical applications similar to ours.

**Definition 4.1** (Logarithmic EACD(1,1)-FunXL process). Let  $x_i^{(ask)}, x_i^{(bid)}$  be drawn from curve-valued exogeneous liquidity processes as specified before. Then,  $y_i$  follows a logarithmic EACD(1,1)-FunXL process, if

$$y_i = \psi_i \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} \text{Exp}(1) \quad (4.8)$$

$$\begin{aligned} \log \psi_i &= \omega + \alpha \log y_{i-1} + \beta \log \psi_{i-1} \\ &+ \int_0^1 \gamma^{(ask)}(m) x_{i-1}^{(ask)}(m) dm + \int_0^1 \gamma^{(bid)}(m) x_{i-1}^{(bid)}(m) dm, \end{aligned} \quad (4.9)$$

where

$$x_i^{(s)} = \mu^{(s)} + \sum_{k=1}^{\infty} \zeta_{k;i}^{(s)} \phi_k^{(s)}. \quad (4.10)$$

Thus, the model is basically the same as the GARCH-FunXL model from Chapter 3, however, with two notable differences: Firstly, the ACD-FunXL model is (primarily) a model for the location parameter of the conditional distribution of the duration process as opposed to the conditional variance. Secondly, it is driven by an innovation term with positive support. For most candidate innovation distributions, the conditional mean and variance are intertwined, i.e., possible “overdispersion” is constant. For instance, in case of the exponential distribution, the variance is equal to the squared mean. As we are solely interested in the conditional mean specification, i.e., in the second model equation, we confine ourselves to exponential innovations. Other innovation specifications will certainly lead to more accurate conditional densities, but will not be considered here. This is justified from an inferential perspective by results on QML estimation as well as Monte Carlo evidence (Grammig and Maurer, 2000).

Further assumptions and properties are directly imported from the GARCH-FunXL specification: We assume that the terms  $\int_0^1 \gamma^{(s)}(m) x_{i-1}^{(s)}(m) dm$ ,  $s \in \{\text{bid}, \text{ask}\}$ , are non-degenerate in the sense that the coefficient functions are finite over  $[0, 1]$  and that both liquidity processes are stationary in the sense explained above. It also holds that log durations admit an ARMA(1,1)-(Fun)X representation.

Evidently, the innovation distribution of the ARMA-X representation is different from the GARCH case, but the basic structure is preserved. In particular,  $\log y_i$  is stationary if the roots of the AR polynomial lie outside the unit circle, i.e., for a model of order (1,1),  $-1 < \alpha + \beta < 1$ . The log-ACD specification is primarily attractive because it avoids non-negativity constraints for the functional exogeneous part of the model.

#### 4.4.2 Estimation

The liquidity part of the model is estimated based on the truncated FPC representation

$$x_i^{(s)}(d) \approx \mu(d)^{(s)} + \sum_{k=1}^K \phi_k^{(s)}(d) \zeta_{k,i}^{(s)}.$$

Empirically we find that the properties of the liquidity part are very much the same as in the GARCH-FunXL case as only the sampling scheme has changed. For low  $\delta$ , i.e., for short durations we get an even higher autocorrelation of liquidity curves while observing more curves, resulting in basically the same eigenfunctions. As shown already, few components explain most of liquidity variation. For our empirical model specification, we conjecture that these components also represent the features that potentially have an impact on the duration process.

### Assumptions

We impose the same assumptions as for the GARCH-FunXL model. In short, this means that (i) only a finite number,  $K$ , of liquidity components is relevant for duration dynamics, where  $K$  is the same for bid and ask liquidity, and (ii) liquidity dynamics are dominated by the diagonal of the multivariate score processes.

### Two-step estimation

We estimate the ACD-FunXL model in two steps.

1. Estimation of the liquidity curves using the orthonormal FPC expansion

$$\hat{x}_i^{(s)}(d) = \hat{\mu}^{(s)}(d) + \sum_{k=1}^K \hat{\phi}_k^{(s)}(d) \hat{\zeta}_{k,i}^{(s)},$$

where the true  $K$ , mean function  $\mu$ , and eigenfunctions  $\phi_k$  are unknown, and the  $\hat{\zeta}_{k,i}^{(s)} = \int_0^1 (x_i^{(s)}(m) - \hat{\mu}^{(s)}(m)) \hat{\phi}_k^{(s)}(m) dm$  are computed via numerical integration. This step has been outlined in detail before.

2. (Q)ML estimation of the EACD-FunXL parameters using the scores  $\hat{\zeta}_{k,i}^{(s)}$ ,  $k = 1, \dots, K$ ,  $i = 1, \dots, I$  from Step 1 and the return data. An alternative is given by inference via the ARMA-X representation.

The procedure can be termed QML as we do not claim that innovations are necessarily exponential. It has been shown, however, that the conditional duration dynamics can be consistently estimated using the EACD model regardless of the true innovation distribution. Taking the square root of durations, it is even possible to use GARCH software. However, in our empirical application, the conditional intensity and its implied instantaneous volatility, the quantity of primary interest, are computed based on the EACD specification.

The conditional distribution of the logarithmic EACD(1,1)-FunXL with exponential innovations is given by

$$y_i | \mathcal{F}_{i-1}^L \sim \text{Exp} \left( \exp \left( \omega + \alpha \log y_{i-1} + \beta \log \psi_{i-1} + \int_0^1 \gamma^{(ask)}(m) x_{i-1}^{(ask)}(m) dm + \int_0^1 \gamma^{(bid)}(m) x_{i-1}^{(bid)}(m) dm \right) \right).$$

The exponential quasi-log-likelihood is proportional to

$$l(\mathbf{y}, \mathbf{x}; \omega, \alpha, \beta, \gamma^{(bid)}, \gamma^{(ask)}) = - \sum_{i=2}^I \left( \log \psi_i + \frac{y_i}{\psi_i} \right),$$

where  $\mathbf{y}$  is the vector of de-seasonalized returns, and  $\mathbf{x}$  the "matrix" of de-seasonalized liquidity curves.

As both the  $x_i^{(s)}$  and the coefficients  $\gamma^{(s)}(\cdot)$  are infinite-dimensional objects, the term  $\int_0^1 \gamma^{(s)}(m) x_i^{(s)}(m) dm$  has to be approximated by some finite-dimensional representation. In our practical application, we use  $K = 1, \dots, 5$ . For both stocks considered,  $K = 4$  components explain at least 95 percent of the curves' variation.

Introducing a  $K$ -dimensional parameter vector  $\gamma^{(s)} = [\gamma_1^{(s)} \ \dots \ \gamma_K^{(s)}]$  for each market side, we expand the coefficient function using the same set of  $K$  eigenfunctions that is used to represent the curves themselves,

$$\gamma^{(s)}(d) = \sum_{k=1}^K \gamma_k^{(s)} \hat{\phi}_k^{(s)}(d),$$

so that, plugging in estimated mean, eigenfunctions and scores from the FPCA of the liquidity curves, the integral  $\int_0^1 \gamma^{(s)}(m) x_i^{(s)}(m) dm$  becomes

$$\int_0^1 \sum_{j=1}^K \sum_{k=1}^K \hat{\xi}_{j;i}^{(s)} \hat{\phi}_j^{(s)}(m) \gamma_k^{(s)} \hat{\phi}_k^{(s)}(m) dm = \sum_{k=1}^K \gamma_k^{(s)} \hat{\xi}_{k;i}^{(s)}$$

by orthonormality of the eigenfunctions. Defining

$$G_{i-1} := \alpha \log y_{i-1} + \beta \log \psi_{i-1}$$

for the endogeneous part of the model, we can now write the conditional duration in the following compact way

$$\log \psi_i = \omega' + G_{i-1} + \sum_s \sum_{k=1}^K \gamma_k^{(s)} \hat{\xi}_{k;i-1}^{(s)} \quad (4.11)$$

by the same arguments as in Chapter 3. The intercept  $\omega'$  includes the mean liquidity effect,  $\int_0^1 \gamma^{(s)}(m) \mu^{(s)}(m) dm$ .

The infinite-dimensional problem has thereby become much easier to handle as only  $2K$  additional scalar parameters compared to the original model,  $\omega, \alpha, \beta, \gamma_1^{(bid)}, \dots, \gamma_K^{(bid)}, \gamma_1^{(ask)}, \dots, \gamma_K^{(ask)}$ . We are nevertheless rather generous in the choice of  $K$  as low variation of liquidity features does not necessarily mean irrelevance regarding the duration process. Note that the properties of the estimators and potential problem of the non-parametric first step have already been discussed in Chapter 3.

### 4.4.3 Liquidity impact

The conditional duration of the EACD-FunXL model can be written as a product of the (endogeneous) ACD part and the exogeneous liquidity part, i.e.,

$$\begin{aligned} \psi_i = & \exp(\omega + \alpha \log y_{i-1} + \beta \log \psi_{i-1}) \\ & \times \exp\left(\int \gamma^{(ask)}(m) x_{i-1}^{(ask)}(m) dm + \int \gamma^{(bid)}(m) x_{i-1}^{(bid)}(m) dm\right). \end{aligned} \quad (4.12)$$

Defining

$$LI_i := \exp\left(\int \gamma^{(ask)}(m) x_{i-1}^{(ask)}(m) dm\right) \times \exp\left(\int \gamma^{(bid)}(m) x_{i-1}^{(bid)}(m) dm\right),$$

the second term is henceforth called the *liquidity impact*. Note that the sub-index is chosen according to the target variable, the conditional variance. Due to the multiplicative structure of the model, liquidity reduces volatility for  $LI_i < 1$  and increases it for  $LI_i > 1$ .  $LI_i$  can further be divided into the contributions of each market side,  $LI_i = LI_i^{(ask)} LI_i^{(bid)}$ .

The  $K$ -truncated, estimated version of the liquidity impact is given by

$$\widehat{LI}_i = \exp\left(\hat{\gamma}'^{(ask)} \hat{\xi}_{i-1}^{(ask)} + \hat{\gamma}'^{(bid)} \hat{\xi}_{i-1}^{(bid)}\right).$$

Confidence statements for the liquidity impact are directly linked to estimation uncertainty about the functional parameters discussed above.

## 4.5 In-sample evaluation

We fit log-EACD-FunXL models to price durations between quote changes of the Commerzbank and Linde stocks, considering the following constellations:

- Separate durations models for bid quotes and ask quotes.
- Durations between price changes of ( $\geq$ ) 1 (2, 3, 4, 5, 10, 20) ticks. Note that the tick size for Linde (EUR 0.01) is ten times as high as for Commerzbank.
- Liquidity curves up to  $D = 51$  (101, 151, 201) price levels (Cents) from the quotes, that is for  $J = D$  equidistant gridpoints, are used as FunX covariates. This

range, if related to the absolute price level, differs for Commerzbank and Linde. Liquidity is evaluated only centwise, which for Commerzbank is a rougher grid than would be possible in principle.

- We allow for impact of ask-side liquidity, bid-side liquidity, both simultaneously, and imbalance impact, respectively.
- We use up to  $K = 5$  FPCs to approximate the observed liquidity curves and to expand the functional parameter.

For each time series of durations, this amounts to 96 different models, allowing to give answers to, among others, the following questions:

- Does LOB-implied liquidity matter for the variation of quoted prices?
- Is liquidity at large distance from the quoted prices relevant for quote variation, or does it suffice to account for liquidity in the neighbourhood of the quotes?
- Does only the liquidity *level* matter, or the specific pattern of limit order schedules?
- Is liquidity impact symmetric?
- What role does the tick size play?

#### 4.5.1 Information criteria

To judge the in-sample fit of the models considered, we primarily use the usual AIC and BIC criteria.

$$AIC = -2 \text{ maximum log likelihood} + 2 \# \text{ parameters}$$

$$BIC = -2 \text{ maximum log likelihood} + \log(\# \text{ observations}) \# \text{ parameters,}$$

where the number of observations,  $\sum_{\ell=1}^L N_{\ell,T}$ , is stochastic and decreases with respect to  $\delta$ . To facilitate interpretation and comparisons, the criteria are normalized by the number of observations.

We emphasize, however, that these measures should be termed *naive* AIC/BIC, respectively, as the number of parameters in the respective second terms must not be taken literally. For a model of order (1,1) accounting for bid and ask liquidity, the number of parameters sums up to  $3 + 2K$ , where  $K$  is the number of FPCs selected. The FPC scores are constructed based on an estimate of the liquidity curves' contemporaneous covariance kernel, whose dimension is ignored in our informal comparison. Moreover, mean and eigenfunctions are estimated quantities subject to estimation error. The "number of parameters" in our model's penalty term ignores these facts and treats the FPCs like directly observable, scalar regressors.

However, this strategy can be justified as the construction of the functional regressors, i.e., the expansion in terms of their covariance operator's eigenfunctions, is not tailored with their explanatory power in the second estimation step in view. In other words: Dimension reduction for liquidity is done *independently* of the inference for the FunXL parameters  $\gamma^{(s)}(\cdot)$ . Thus, the problem of possible data snooping might be considered negligible in this case. Having this simplifying assumption in mind, we nevertheless judge our fits by these criteria. We return to this point when discussing the out-of-sample exercise below. A second issue, already raised in Chapter 3, is that QML theory for log-GARCH (and, thus, log-ACD) models is not fully understood, either. We conjecture it has the favorable properties of the linear version of the model based on our own numerical experiments and the Monte Carlo evidence in Sucarrat et al. (2013).

To address the question whether liquidity deep in the book is relevant, models with different domains  $[0, D]$  have to be compared. Note, however, that they are not nested as regressors are constructed using different eigen-expansions. Therefore, a comparison of in-sample results has to be drawn with some caution. We prefer to ultimately judge the models based on their out-of-sample performances.

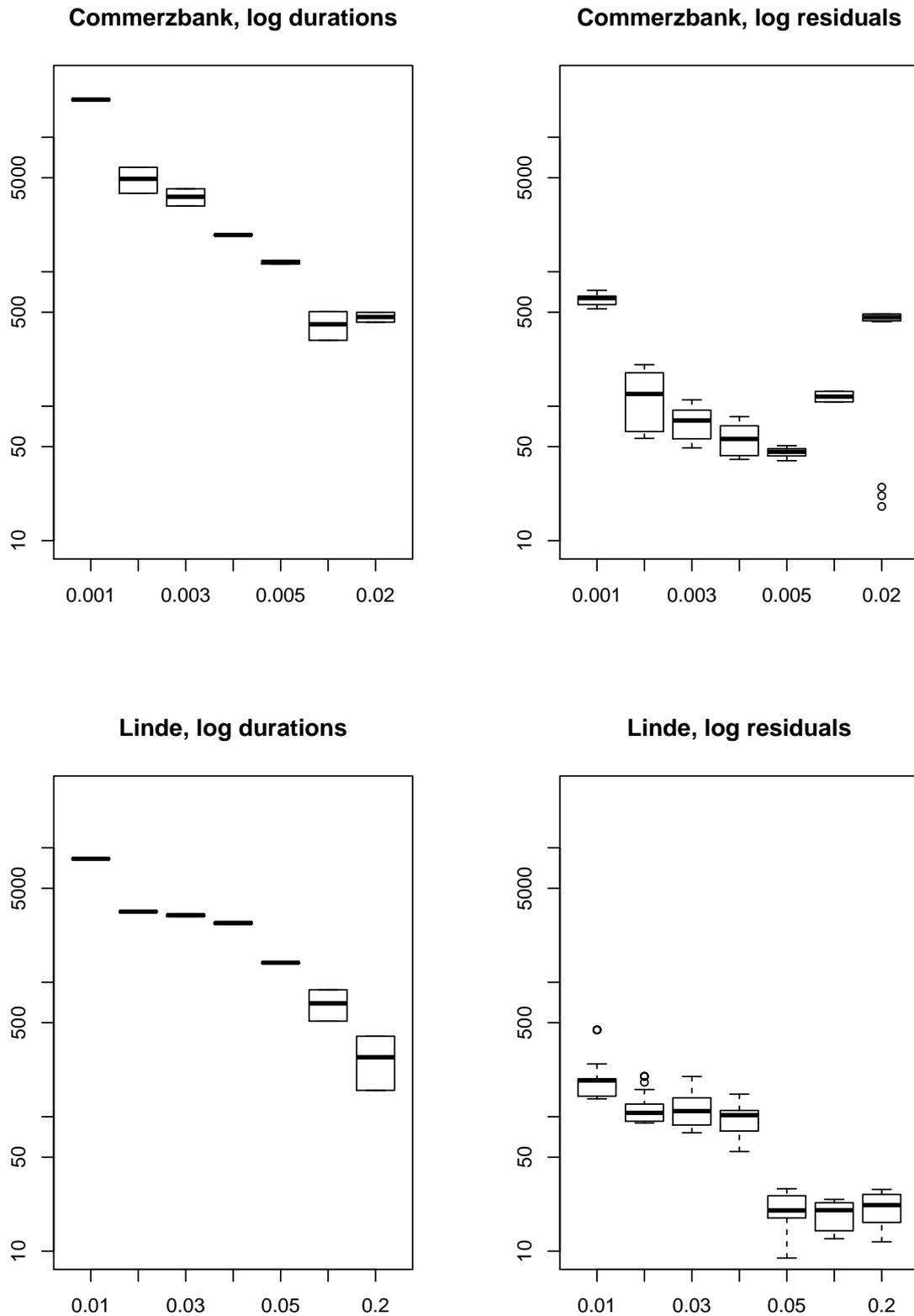
## 4.5.2 Results

To start with, we find that liquidity imbalance does not have any explanatory power, which is in line with the finding for GARCH-FunXL. This result indicates that the imbalance measure, by construction, eliminates information on the individual curves which is valuable for predicting jumps of quoted prices. This is quite plausible: As a gedankenexperiment, consider the cases of (i) a flat ask liquidity and an infinitely steep bid liquidity curve vs. (ii) curves with unit slope on both market sides. Both (i) and (ii) imply the same liquidity imbalance (the angle between the curves is 90 degrees in both cases) although representing completely different states of the market.

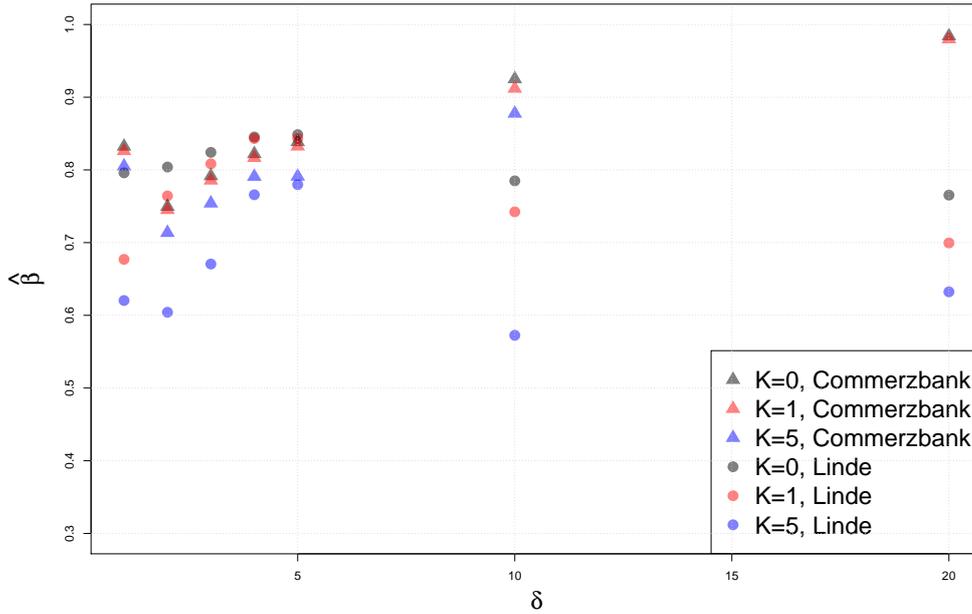
Secondly, we find that for all duration series considered, models employing bid *and* ask liquidity exhibit a superior goodness-of-fit compared to the one-sided models. In what follows, we therefore confine ourselves to models with bid and ask liquidity impact.

For reasons of clarity, we focus on the Commerzbank stock and  $\delta = 1$  (5) ticks, respectively. All remaining results for both Linde and Commerzbank, and for all  $\delta$  considered can be found in Appendix A.4.

A minimum requirement for a "good fit" of a time series model is its ability to decorrelate the data. All models considered achieve this primary goal — which, in the case of log-ACD models, consists in fitting the autocorrelation of log-durations—, reasonably well. Figure 4.6 shows Ljung-Box statistics (10 lags) for data and residuals. For each  $\delta$ , there are two time series (bid and ask durations), and for each time series, 6 different models are fitted ( $K = 0, \dots, 5$ ). We observe that in almost all cases, log-residuals have much lower LB statistics than the unfiltered log-durations, indicating that log-ACD models (also those without liquidity impact) are able to filter the largest part of serial correlation out of the data.



**Figure 4.6:** Ljung-Box statistics (10 lags) for log durations (left panel) and log residuals (right panel). Each boxplot in the right panel represents the statistics for all models ( $K=1, \dots, 5$ ), each fitted to one of the two quote duration series (bid and ask). Each boxplot in the left panel represents two durations series, the price durations computed based on bid and ask quotes, respectively.



**Figure 4.7:** Estimates of the persistence parameter  $\beta$  for models with  $K = 0$  (1, 5) FPCs,  $D = 201$ .

A further question of interest is whether liquidity provides an alternative explanation for what is usually called the “persistence” of durations. This would be the case if including liquidity resulted in much smaller autoregressive parameters, especially smaller  $\hat{\beta}$ . In a series of papers, Han and Kristensen (2014), Han and Park (2008), Han et al. (2015) show both theoretically and empirically that persistent covariates in a GARCH framework are able to explain almost all of volatility persistence. However, usually these persistent covariates are realized measures, i.e., observed proxies of the latent volatility component itself. This is quite different from our setup, where liquidity is supposed to be truly exogenous.

Figure 4.7 shows estimates for the “persistence” parameter  $\beta$  with respect to the number of liquidity components and  $\delta$ . In some cases, a fair amount of the persistence can alternatively be explained by liquidity impact. However, liquidity is not identified as the main driver of what otherwise would be termed persistence of volatility. The result is plausible in view of the LB statistics in Figure 4.6, where accounting for liquidity impact appeared not to affect the amount of residual autocorrelation in such a decisive manner, either.

### Goodness of fit: Information criteria

Let us now turn to the goodness-of-fit results in terms of the (naive) selection criteria discussed above. Overall, the results can be summarized as follows:

- Liquidity matters: Inclusion of liquidity curves improves the fit in all cases, i.e. for bid and ask quote processes, in terms of AIC and BIC.

- Both bid and ask liquidities have an impact on both quote processes.
- Considering liquidity for both market sides improves the fit as compared to only one liquidity curve.

Table 4.2 shows the AICs for the duration models fitted on bid and ask quotes' 2-tick and 5-tick price durations, respectively.  $K = 0$  corresponds to the pure log-ACD model without liquidity impact. We also consider different liquidity domains (parameter  $D$ ). We find in all cases that liquidity has a considerable impact on price durations, that is, on instantaneous volatility. Models with  $D = 51$  tend to slightly outperform models with larger  $D$  ceteris paribus, in particular for the same  $K$ , indicating that liquidity variation near the quotes is most important. Nevertheless, it appears that the (tuning) parameter  $D$  is, at least in the four settings considered here, of least importance. De-diurnalized liquidity varies most in the left part of its domain (i.e., near the quotes), and this phenomenon is captured almost equally well by eigen-components accounting for variation in a wider neighbourhood of up to 201 Cents from the quotes as for  $D = 51$ . Note that we found liquidity on both market sides to be influential, not only on the side of the respective quote (result not shown here). Detailed results for all models (bid and ask durations for all  $\delta$ ,  $D$ , and  $K$  considered) are shown in the Appendix.

Figure 4.8 shows the fit in terms of the log instantaneous variance over time, and allows for a comparison to the pure log-ACD model. We find that (on the log scale) implied variances are quite similar for both models, though (according to the ratio of the two) there seem to be periods where implied variances differ in a somewhat persistent manner.

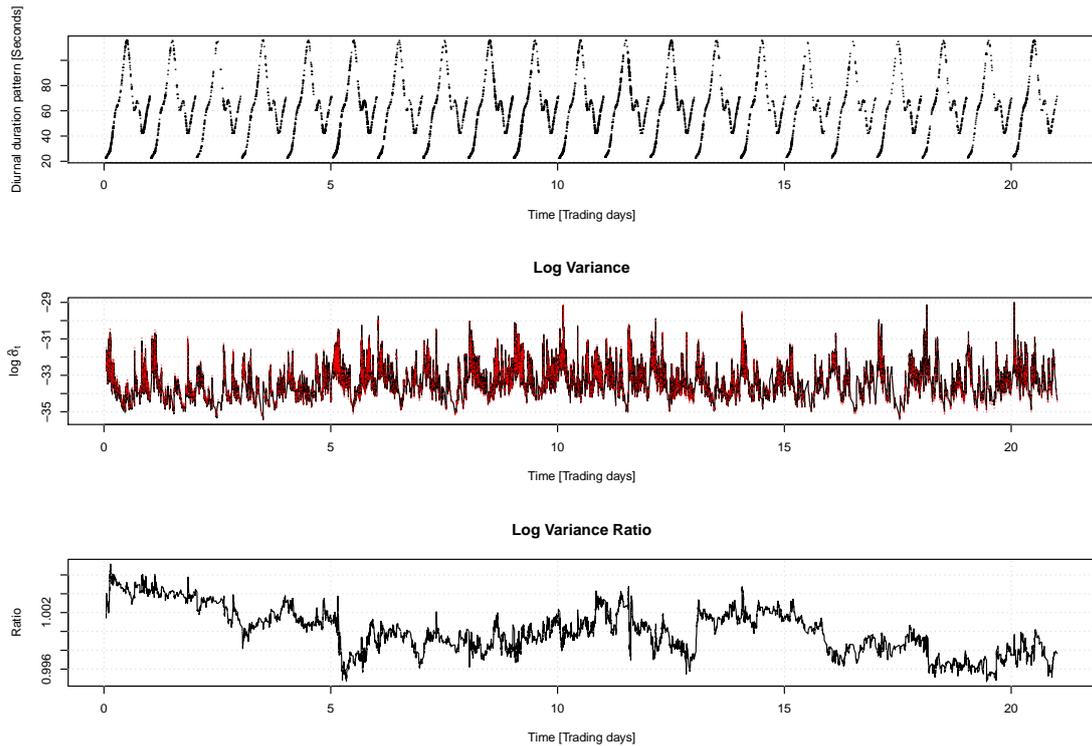
Accordingly, the fits do not vary much with respect to  $D$ . However, for Commerzbank the models with  $D = 201$  explain quote durations slightly better than those with lower  $D$ ; for Linde, the opposite is true. We conjecture that the choice of a less-than-maximal  $D$  only makes sense in finite samples. Asymptotically,  $K$  can be allowed to grow along with the number of observations at a suitable rate, so that the approximations of the liquidity curves and the functional parameter should converge to their true entities,

$$\sum_{j=1}^K \sum_{k=1}^K \hat{\xi}_{i,k} \hat{\phi}_k(d) \hat{\gamma}_j \hat{\phi}_j(d) \rightarrow \gamma(d) x_i(d) \quad \forall d \in [0, 1].$$

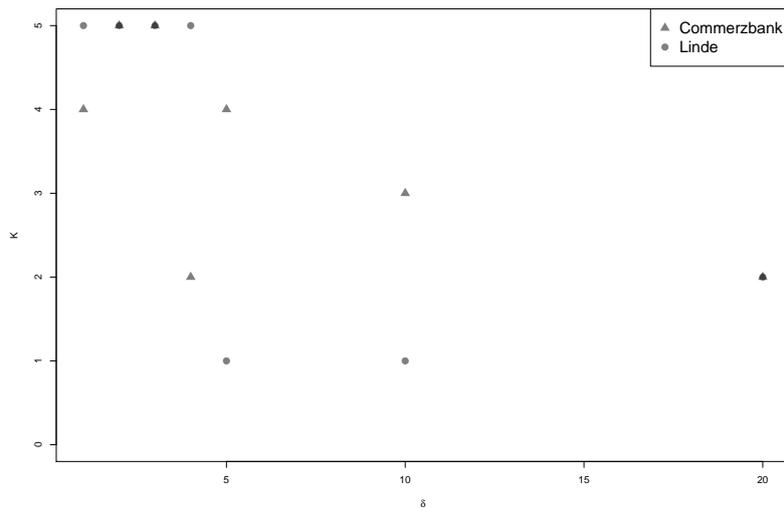
A thorough analysis of this conjectured convergence is left for future research.

As shown in Figure 4.9, durations between small price changes tend to be explained best by models accounting for many aspects of the liquidity curves' structure, whereas durations between large price changes are explained well by 1 or 2 FPCs already. For the vast majority of models, the first component, which may be regarded as a level component, provides the highest improvement of the fit as compared to pure ACD models.

We pick up the question of model choice again in the context of our forecast exercise.



**Figure 4.8:** Volatility of Commerzbank ask quotes over the full sample, 10/2010, based on 5-tick price movements. Top panel: Diurnal pattern, evaluated at the event times. Center panel: Log Variance for the pure log-ACD (black) and the log-ACD-FunXL with two principal liquidity components, i.e., the model with the best in-sample AIC. Bottom panel: Ratio of the variances of the pure log-ACD and the log-ACD-FunXL model.



**Figure 4.9:** Liquidity dimension  $K$  of the best model with respect to the duration series (number of ticks) considered. The liquidity domain is held fixed at  $D = 201$ .

$\delta = 0.002$  EUR (2 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.7788	0.7788	0.7788	0.7788	0.9279	0.9279	0.9279	0.9279
1	0.6919	0.7093	0.7186	0.7220	0.8226	0.8408	0.8486	0.8524
2	<b>0.6847</b>	0.6917	0.6995	0.7034	0.8222	0.8197	0.8231	0.8229
3	0.7078	0.7036	0.7029	0.6986	0.8262	0.8355	0.8339	0.8279
4	0.7070	0.7073	0.7086	0.7090	0.8182	0.8288	0.8288	0.8290
5	0.7129	0.7197	0.7110	0.7112	<b>0.8159</b>	0.8228	0.8223	0.8246

 $\delta = 0.005$  EUR (5 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.0164	1.0164	1.0164	1.0164	1.1071	1.1071	1.1071	1.1071
1	0.9777	0.9890	0.9937	0.9949	<b>1.0689</b>	1.0744	1.0768	1.0768
2	<b>0.9738</b>	0.9779	0.9839	0.9858	1.0800	1.0738	1.0772	1.0768
3	0.9986	0.9957	0.9887	0.9817	1.1103	1.0989	1.0949	1.0866
4	0.9853	1.0078	1.0072	1.0021	1.1033	1.1112	1.1056	1.0986
5	0.9563	1.0167	1.0093	1.0124	1.0981	1.1089	1.1113	1.1077

**Table 4.2:** Average AIC per observation for bid and ask price durations and the Commerzbank stock.  $\delta$  amounts to 1 tick (top panel) and 5 ticks (bottom panel), respectively.

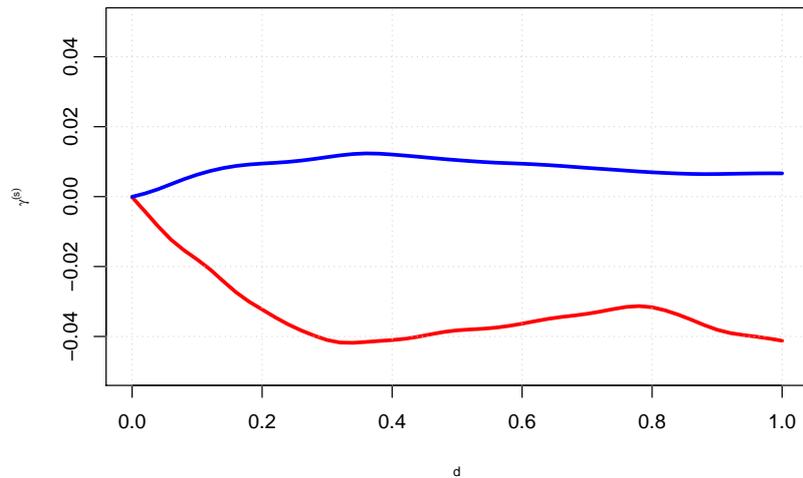
### 4.5.3 The functional parameter

Point estimates for the functional parameter of the best model according to the AIC for the Commerzbank data ( $\delta = 5$  ticks) are shown in Figure 4.10. Bid liquidity tends to increase, ask liquidity tends to decrease price durations over the entire domain of relative prices (50 Cents in the present case). As price durations are a reciprocal measure of volatility, the result is in line with the finding for the Commerzbank data in the GARCH-FunXL context, see Chapter 3, where bid (ask) liquidity has been found to decrease (increase) the conditional variance.

### 4.5.4 Liquidity impact

We recall that the EACD-FunXL model's conditional duration equation can be written as

$$\begin{aligned} \psi_i = & \exp(\omega + \alpha \log y_{i-1} + \beta \log \psi_{i-1}) \\ & \times \exp\left(\int \gamma^{(ask)}(m) x_{i-1}^{(ask)}(m) dm + \int \gamma^{(bid)}(m) x_{i-1}^{(bid)}(m) dm\right). \end{aligned}$$



**Figure 4.10:** Functional parameter estimates for bid and ask liquidity impact, Commerzbank data ( $\delta = 5$  ticks), for the model with  $K = 2$  and  $D = 51$  as chosen by the AIC.

Then, in analogy to the GARCH case, the *liquidity impact* on the conditional duration can be defined as

$$LI_i := \exp \left( \int \gamma^{(ask)}(m) x_{i-1}^{(ask)}(m) dm \right) \times \exp \left( \int \gamma^{(bid)}(m) x_{i-1}^{(bid)}(m) dm \right).$$

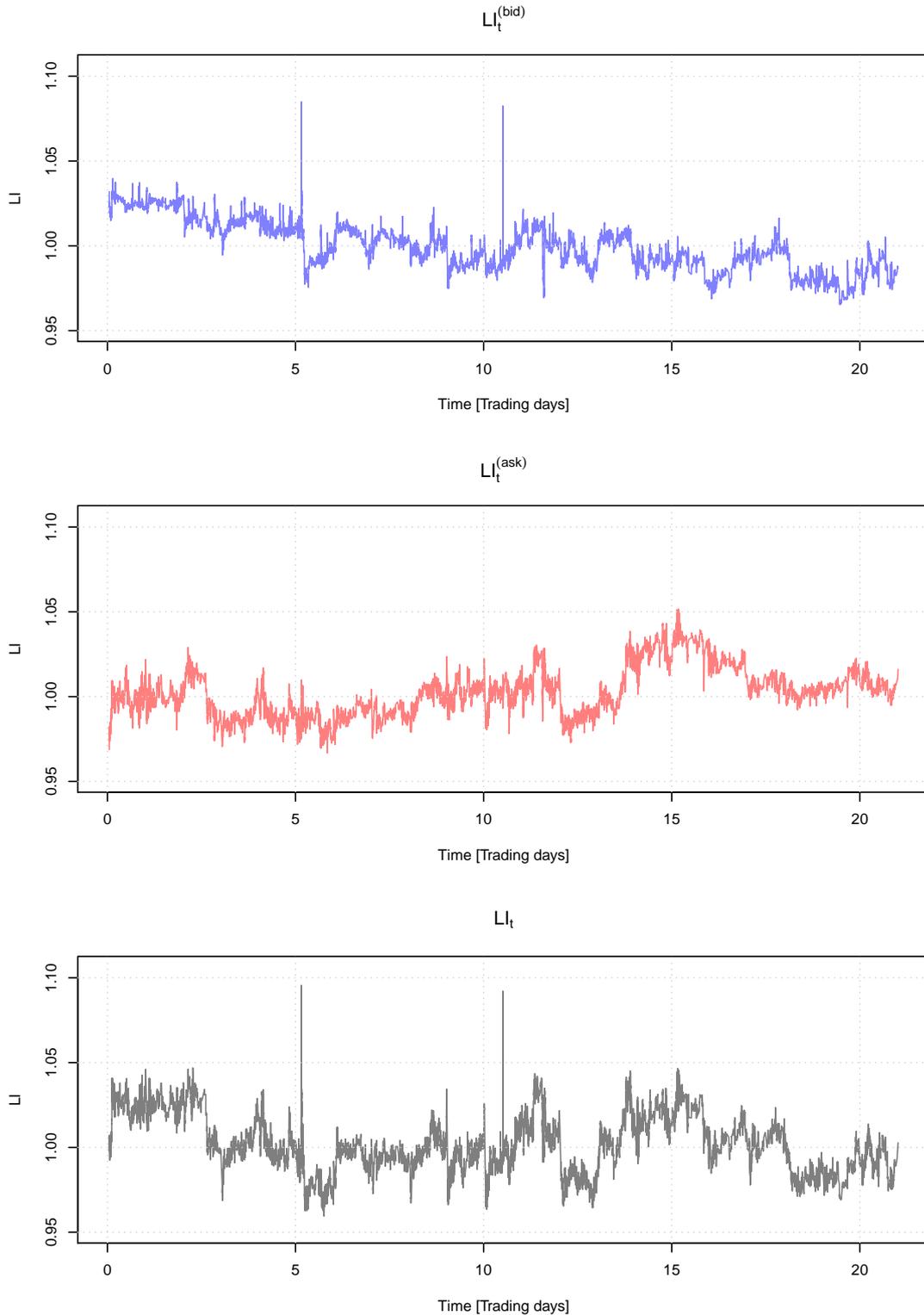
The liquidity impact for the Commerzbank data and the same “best” model as before is depicted in Figure 4.11. According to the total impact in the bottom panel, apart from two spikes the relative contribution of liquidity to conditional durations is relatively moderate, leveraging the endogeneous part by roughly  $\pm 5$  percent.

## 4.6 Forecasting price durations

We exemplarily comment on the out-of-sample results for the two data situations already discussed above, which are quite representative. All remaining results can be found in Appendix A.4.

Table 4.3 shows 1-step-ahead out-of-sample forecast results for the Commerzbank data. We find that liquidity-augmented models are able to outperform the pure ACD specification. Results are also significant in a Diebold-Mariano sense.<sup>3</sup> Models only accounting for liquidity near the quotes tend to perform better than those accounting for the full range of liquidity. This is in line with the in-sample results. Interestingly, for the present situation the fourth FPC, not that relevant for liquidity variation itself, provides a considerable improvement over the models with  $K = 3$ .

<sup>3</sup>We abstain from discussing the appropriateness of the DM assumptions here.



**Figure 4.11:** Liquidity impact ( $D = 51$ ) on the instantaneous volatility of Commerzbank's ask quotes over the full sample, 10/2010, based on 5-tick price movements and  $K = 2$  FPCs. The top panel (blue) shows the buy-side liquidity impact, the center panel (red) the impact from the sell-side. The total impact is depicted in the bottom panel (black).

$\delta = 0.002$  EUR (2 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	7.4409	7.4409	7.4409	7.4409	7.8037	7.8037	7.8037	7.8037
1	7.3842	7.4809	7.5794	7.6184	7.7112	7.8672	8.0188	8.0760
2	7.3566	7.5040	7.5307	7.5345	7.6573	7.8673	7.9076	7.9128
3	7.3542	7.4529	7.4672	7.4596	7.6427	7.8078	7.8306	7.8137
4	<b>7.2833</b>	7.4448	7.4569	7.4545	<b>7.5296</b>	7.7796	7.8093	7.7924
5	7.3634	7.4385	7.4348	7.4166	7.6059	7.6963	7.7354	7.7314

 $\delta = 0.005$  EUR (5 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	6.9429	6.9429	6.9429	6.9429	7.5744	7.5744	7.5744	7.5744
1	6.9091	6.9298	6.9629	6.9738	7.2801	7.4584	7.6317	7.6813
2	6.9416	6.8670	6.8937	6.9066	7.3498	7.4155	7.4599	7.4723
3	6.8634	6.8544	6.8546	6.8679	7.3081	7.4036	7.4450	7.4410
4	6.8883	<b>6.8293</b>	6.8400	6.8776	<b>7.2330</b>	7.4226	7.4725	7.4236
5	6.8775	6.8311	6.8674	6.8179	7.2633	7.3616	7.3868	7.3768

**Table 4.3:** Out-of-sample forecast performance for Commerzbank price durations. The model is fitted to the data from the first 10 (left panel) and 15 (right panel) days, respectively. Then, 1-step-ahead forecasts are computed for each observation during the days 16 to 21 in the sample (October 2010). The forecasts are evaluated using the RMSE.

## 4.7 Conclusion

In the present Chapter we have shown how functional liquidity can be included into a second class of multiplicative error models, ACD models, in much the same way as in the GARCH case. It has been shown empirically that (and in which respect) liquidity matters for the dynamics of quoted prices at ultra-high frequency. Moreover, we conjecture that the model could be of great use for other applications, for instance if one is interested in non-negative dynamic quantities like trading volume or realized variance.

Regarding the specific application, skepticism might come from the realized volatility corner of financial econometrics. Namely, one could argue that accounting for each quote change when measuring volatility is problematic. Pointedly, one might ask if we have just been modeling microstructure noise in the present Chapter. If one is solely interested (and believes) in an underlying *true, efficient* price process buried in noise, the answer to this question is certainly positive. However, when working with high-frequency data (to whatever purpose), the question really is whether there is such a thing as a *true* underlying price.

From a more practical, balanced perspective, we admit that it is certainly correct that

volatility measurement for the purpose of low-frequency decision making — such as portfolio allocation, derivative pricing, or risk management on a day-to-day or even 10-day basis —, should ignore the extra variance accumulated when sampling at ultra-high frequency. However, for instance from the perspective of a high-frequency trader, ignoring any movement of the quotes does not make any sense, simply because it can potentially be used to be traded on. The ACD model can even be used to estimate realized volatility via the ICV method which cumulates the instantaneous variance over time. Robustified by a reasonable choice of  $\delta$ , it can even be superior to other realized measures (Tse and Yang, 2012). For a recent contribution discussing the contrast between microstructural volatility (not being called *noise*) and fundamental price variation, see Hasbrouck (2015).



## **Part II**

# **Realized Volatility and Liquidity Impact**



## Realized Volatility and Liquidity Impact

In *Part II* of the thesis, we change the view by measuring temporal aggregates of liquidity and volatility. To this end, in Chapter 5, analogous to the well-known realized measures of price variation, some realized measures of liquidity are proposed. We motivate the realized measures by some of the stylized facts of ultra-high frequency LOB dynamics, and propose a compound nonhomogeneous Poisson process (NHPP) for intraday inventory revisions. We justify the use of the realized volatility estimator as a measure of liquidity risk under the NHPP model.

In Chapter 6, we introduce heterogeneous autoregressive (HAR-) Fun-XL models to model the connection between liquidity and price volatility dynamics. Finally, we show how Markovian models like HAR-FunXL, the Markovian version of ACD-FunXL, and further possible dynamic order book models can be viewed as (generalized, additive) linear regression models. This representation allows to import smoothness penalties and other regularization methods that have been developed for functional regression models in recent years. We apply one of these methods, penalized functional regression, to reassess the results obtained by using the FPC basis expansion.



## Realized measures of liquidity

### 5.1 Introduction

In Part I, we have considered models capturing the state of the LOB *inventories* in the form of snapshots taken at specific times during continuous trading, and modeled their impact on the price process. In the present and the following Chapter, LOB dynamics are viewed and modeled from a different, namely the *realized volatility* (RV) perspective.

Research on realized volatility aims at measuring uncertainty about asset prices *ex post*, based on high-frequency data. The basic claim made by this strand of research is that (in principle) volatility is no more a latent, unobservable quantity which can only be proxied by squared (day-to-day) returns, but directly observable using high-frequency observations of the data-generating, continuous time price process, (Andersen et al., 2001, 2003). For liquidity, however, to our knowledge no “realized liquidity” concept exists, with the only exception of Engle et al. (2012). Their notion of realized volatility is imported directly and rather on an ad-hoc basis from the RV literature in order to model liquidity uncertainty of treasury bonds. However, the authors report that the basic assumption of a martingale difference sequence appears to be fulfilled for their data, so that the application of the standard realized volatility measure is justified.

In this Chapter, we propose an alternative continuous-time, pure-jump model of liquidity, but for the same purpose: The ex-post measurement of important aspects of LOB inventories’ dynamics, *aggregated over intervals* during continuous trading, which is in close analogy to realized measures of price volatility. We call these *realized measures of liquidity* (RML). This model of liquidity allows for the construction of estimators of liquidity’s integrated drift and variation for every relative price  $d$ . In a second step, the measures can be viewed as functions of the relative price  $d$ . RML on the bid and ask side, respectively, can then be viewed as a stationary functional time series at a daily frequency, and its dynamics be captured by models introduced in Chapter 2.

The Chapter proceeds as follows. We first review the standard model of realized price variation, i.e., a diffusion process for logarithmic prices with stochastic volatility,

but without jumps. We then turn to liquidity. After shortly reviewing the approach by Engle et al. (2012) and discussing the implications of some other LOB models for liquidity variation, we put forth a pure-jump model of liquidity in the spirit of the compound Poisson model of Press (1967), originally constructed with prices in view. Our liquidity model is a difference of two compound nonhomogeneous Poisson processes (NHPPs) with independent increments or marks. In the realized (price) volatility literature, a similar model has been introduced in Oomen (2006), on which we heavily draw.

We then define three measures of *integrated* liquidity over some interval, typically the trading day, and show several ways of obtaining realized measures, i.e., estimators of integrated liquidity. In particular, it is shown analytically and by means of simulation that an estimator based on equidistant snapshots approximates integrated liquidity accurately. The simulation setup is based on an estimated version of the compound NHPP with independent exponential marks.

Finally, by considering limit order schedules at a wide range of relative price levels, functional RML are constructed. We illustrate the construction using XETRA LOB data and briefly look at RML dynamics by means of functional time series models. The interplay of RML and RV is then analyzed in the subsequent Chapter 6.

## 5.2 Realized volatility

### The price process

In the basic setting, one assumes that the logarithmic price in continuous time,  $(p_t)_{t \in [0, \infty)}$ , is generated by the diffusion process

$$p_t = p_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s,$$

implying

$$dp_t = \mu_t dt + \sigma_t dW_t.$$

Starting at some  $p_0$ , the process consists of a deterministic drift,  $\mu_t$ , and a stochastic part driven by standard Brownian motion  $(W_t)_t$  and a time-varying, random scale parameter, the stochastic volatility process  $\sigma_t$ .

Now consider the sum of squared returns in the interval  $[0, T]$ ,

$$\sum_{j=0}^{M-1} (p_{\tau_{j+1}} - p_{\tau_j})^2,$$

for partitions  $0 = \tau_0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_M = T$ . If we observed the process in each instant of time, i.e.  $M \rightarrow \infty$  and  $\tau_{j+1} - \tau_j \rightarrow 0$ , we would directly observe the

integrated variance,

$$[p]_T = \int_0^T \sigma_s^2 ds,$$

which is the quantity of primary interest in the following.

We turn to the task of *estimating* integrated variance (or volatility) over intervals of length 1 (day). If we observe the price equidistantly  $M$  times during day  $\ell$ , we can construct  $M - 1$  intraday returns,

$$r_{\ell,j} = p_{\ell-1+j/M} - p_{\ell-1+(j-1)/M}, \quad (5.1)$$

and define the *realized variance* for day  $\ell$ ,

$$RV_\ell = \sum_{j=1}^M r_{\ell,j}^2, \quad (5.2)$$

which should approach the integrated variance of the price process over day  $\ell$  for  $M \rightarrow \infty$ ,

$$RV_\ell \rightarrow \int_0^T \sigma_s^2 ds. \quad (5.3)$$

Using high-frequency data, the possible sample size  $M$  is very large, which gives rise to the notion of *realized* variance: Integrated variance is no more considered latent, but (almost) directly observable. Given that the consistency statement holds, there exists a worked-out inferential theory (Barndorff-Nielsen and Shephard, 2002; Barndorff-Nielsen, 2002).

We have stated that RV asymptotically *should approach* its population counterpart. However, it turns out in practice that the estimate of integrated variance is growing larger and larger rather than converging to what could be considered the true integrated variance as  $M$  increases. This phenomenon is usually explained by the presence of so-called *microstructure noise* which is attributed, among other things, to the discreteness of the price, which can be specified only as multiples of the tick size, to the bid-ask spread, etc.. In the statistical literature, such a phenomenon is typically termed *measurement error*.

In the simplest (still *equidistant*) setting, a model exhibiting microstructure noise may be given by

$$r_{\ell,j} = p_{\ell-1+j/M} - p_{\ell-1+(j-1)/M} + e_{\ell,j},$$

where  $e_{\ell,j}$  is *i.i.d.* with mean zero and constant variance.<sup>1</sup> This error is present at every measurement, thus, for  $M \rightarrow \infty$ ,

---

<sup>1</sup>There are also versions with autocorrelated errors, MA(1) being the most common.

$$RV_\ell \rightarrow \int_0^T \sigma_s^2 ds + \sum_{j=1}^{\infty} e_{\ell,j}^2 = \infty.$$

There is a huge body of literature on how to modify  $RV_t$  to account for and robustify against measurement error, while the “traditional” way to deal with the phenomenon is to sample at 5 minute frequency which is widely agreed to be the highest frequency not subject to noise while still delivering about 100 observations per day (XETRA). Recently, Liu et al. (2015) found that the improvement brought by more sophisticated methods like subsampling and pre-averaging is questionable. In our empirical application, we use 5-minute returns (in part to be consistent with our method of liquidity measurement) but also tried subsampling (Zhang et al., 2005) which did not change the essential findings in our study.

There are alternative models for the price, especially models including a discontinuous jump part, and corresponding realized measures that disentangle the continuous and the jump part. However, we stay within the basic setting explained above. Before turning to the measurement of liquidity, we stress the fact that realized volatility, as an ex-post measure of price variation, is computed conditional on a trading day, that is, conditional on that trading day’s infinite-dimensional parameter  $\sigma_{\ell,t}$ , the day- $\ell$  section of the stochastic spot volatility function. We adopt this conditional approach when measuring *liquidity* ex post.

## Realized volatility from a functional data perspective

When measuring realized volatility of the price, for each day a trajectory of  $M$  squared returns is constructed. If one is not only interested in the integrated variance  $\int \sigma_s^2 ds$ , but in the spot variance  $\sigma_t^2$  itself, squared intraday returns can be used to construct an estimate for this quantity. Müller et al. (2011) propose to view each such daily trajectory as realization of an underlying functional variance process,  $(\sigma_\ell^2(s))_{\ell \in \mathbb{Z}}$ .<sup>2</sup> The potential problem of measurement error is thereby considered negligible. Moreover, the realizations are assumed to be independently sampled.<sup>3</sup>

Whereas in case of the price process we are not interested in  $\sigma_\ell^2(s)$  but rather in  $\int \sigma_\ell^2(s) ds$ , a scalar quantity, we adopt this same basic idea for modeling *realized liquidity* in a functional manner.

## Towards realized liquidity

As demonstrated, there exists a well-elaborated theory for the volatility of asset *prices*, and the pure diffusion setting considered above is the standard case. Other studies

<sup>2</sup>Note that we move the continuous time argument by writing  $\sigma_\ell^2(s)$  instead of  $\sigma_s^2$  both to emphasize the change of perspective and to stay notationally consistent.

<sup>3</sup>It is well-known and will be demonstrated later on that realized variances (more precisely: their logarithm and square-root) are highly autocorrelated. Therefore, the independence assumption for the time series of spot volatility functions is disputable.

consider pure-jump environments as introduced in Press (1967), or combine both concepts to jump diffusions. Moreover, the presence of and correction for several kinds of measurement error (termed *microstructure noise*) has been a central topic of this literature in recent years. For an extensive treatment, see Aït-Sahalia and Jacod (2014).

For asset *liquidity* no such theory exists, and to our knowledge Engle et al. (2012) is the only reference so far using realized measures of liquidity uncertainty. However, Engle et al. (2012) do not provide a theory, but rather conjecture that standard realized volatility measures may be employed to measure liquidity variation as well. Using data for US treasury bonds, sampled at 1-second frequency and the first 5 tiers apart from the quotes on each market side, the authors compute both realized volatility of the price and realized liquidity volatility (volatility of *depth*).<sup>4</sup> The quantities constructed in that study are the realized volatility of price,  $RVP$ , and of depth,  $RVD$ , which can in (a slight abuse of) our notation be defined as

$$RVP_t = \sqrt{\sum_{k_t=1}^{300} (\Delta p_{k_t})^2},$$

$$RVD_t(d) = \sqrt{\sum_{k_t=1}^{300} (\Delta v_{k_t}(d) p_{k_t})^2}, \quad d = 0, 1, 2, 3, 4,$$

where  $\Delta p_{k_t}$  are second-to-second price changes, and  $\Delta v_{k_t}(d)$  are second-to-second volume changes at relative price  $d$  (ticks). The sampling frequency of high-frequency observations is constant at 1 second, the frequency of aggregates is 300 seconds (5 minutes).

First, we observe that  $RVD$  is not a pure liquidity measure as it depends on the price  $p_{k_t}$ . Moreover, a sensible use of the  $RVD$  measure requires assuming liquidity changes to form a martingale difference sequence, in close analogy to the price process. Finally, there may arise severe problems with microstructure noise at this sampling frequency. The authors report that for their data the martingale difference assumption is justified and claim to avoid microstructure noise by using a sampling frequency of 5 minutes.<sup>5</sup>

Another class of order book models, the most prominent ones being Smith et al. (2003) and Cont et al. (2010), make assumptions regarding the in- and outflow of liquidity which imply a specific kind of liquidity variation. In these so-called "zero-intelligence" models, market orders, limit order submissions, and limit order cancellations arrive at rates which are constant given the relative price, and all these inventory changes have unit size. In the case of Cont et al. (2010), both market sides evolve by the following rules.

<sup>4</sup>Price and depth measurement for treasury bonds is a bit different from prices and volumes of equities as considered here. For details see the data description in Engle et al. (2012).

<sup>5</sup>However, the sampling frequency relevant for the noise issue in their study is 1 second.

- Market orders arrive at constant rate  $\mu$ .
- Limit orders arrive at rate  $\lambda(d) = \frac{\beta}{d^\alpha}$ , where  $\beta$  and  $\alpha$  are parameters. This means that the rate decays with respect to  $d$  by a power law.
- Limit orders are cancelled at rate  $\theta(d)v_t(d)$  which is proportional to the number of outstanding orders at  $d$ ,  $v_t(d)$ .

The entire LOB is then a continuous-time Markov chain with finite state space. An implication is the fact that inventories' variation is constant over time.

Our liquidity model, which is introduced in the next Section, adopts some aspects of both approaches, but it also differs in many respects from them. Firstly, we adopt the idea of ex-post measurement of squared liquidity differences put forth in Engle et al. (2012). As outlined earlier, we use the number of offered/ requested shares rather than their monetary value as measure of pure liquidity. Instead of treating only a few price tiers and each of them separately, we construct *functional* realized measures of liquidity (FRML). Two of the proposed FRML do not only measure variation, but also the level of liquidity. Finally, in order to avoid potential effects of microstructure noise and for reasons of practical implementation, we advocate a much lower sampling frequency (1 minute and 5 minutes), constructing daily aggregates thereof.<sup>6</sup>

Secondly, we back our realized measures by sketching a "theory of liquidity variation", i.e., by constructing a plausible data-generating process for liquidity. Then, the properties of the realized measures will be analyzed for this process. The liquidity process is built upon the zero-intelligence model, but with two important generalizations: Order sizes are random instead of constant, and can be described by some parametric distribution. Moreover, the rates of liquidity arrivals and departures are time-varying.

## 5.3 A continuous-time pure-jump model of liquidity

With the final goal of constructing functional realized measures of liquidity, we first introduce our model and the theoretical quantities of interest within this model. Estimators of these quantities are proposed and their theoretical properties analyzed both analytically and by means of simulation.

### 5.3.1 The liquidity model

To develop a model for liquidity, let us first introduce some notation. In what follows,  $t$  denotes continuous intraday time. The subindex  $\ell = 1, \dots, L$  denotes the trading day and the superindex ( $s$ ),  $s \in \{\text{bid}, \text{ask}\}$ , denotes the market side. We consider a specific

---

<sup>6</sup>The acceptable amount of microstructure effects of course depends strongly on the goal of the study. In applications in earlier chapters, particularly in the case of ACD-FunXL, we have tacitly tolerated a large amount of what the RV literature would call "noise". One may even argue that the efficient price (and here, additionally: liquidity) process we are interested in does not exist at all and, thus, distinguishing its variation from noise is a pointless task. Of course, we do not argue that way.

trading day with trading hours  $t \in [0, T] =: \mathcal{T}$ , and only one market side. Thus, where not urgently needed,  $\ell$  and  $(s)$  are omitted in the following to avoid notational clutter. However, to emphasize the functional nature of liquidity, we entrain the argument  $d$ .

As an alternative to the diffusive stochastic volatility models discussed before and implicitly assumed by Engle et al. (2012) in a liquidity context, we consider a pure-jump model

$$x_t(d) = x_0(d) + \sum_{i=1}^{N_t^+(d)} u_i^+(d) - \sum_{j=1}^{N_t^-(d)} u_j^-(d),$$

where  $x_t(d)$  are the limit order inventories at relative price  $d$  and time  $t$ ,  $x_0(d)$  are the inventories when continuous trading starts,  $u_i^+(d)$  denotes increases of liquidity and  $u_j^-(d)$  denotes decreases of liquidity, and  $N_t^+(d)$  and  $N_t^-(d)$  are counting processes to be specified shortly. Note that  $x_t(d)$ , in contrast to *Part I*, is *not* the *cumulative volume*, but the volume at relative price  $d$ . Liquidity increases when a limit order of size  $u_i^+(d)$  is being submitted or a quote change leads  $x_t(d)$  to represent the inventory at a different absolute price than before, associated with higher liquidity. Liquidity decreases when a limit order of size  $u_j^-(d)$  is cancelled, a market order absorbs  $u_j^-(d)$  shares, or a quote change induces a decrease in liquidity at relative price  $d$ .

Inventory changes are thus viewed as marked point processes (MPPs) whose realizations consist of the tuples  $\{t_i(d), u_i^+(d)\}$  ( $\{t_j(d), u_j^-(d)\}$ ), i.e. of event *times* and sizes of changes in liquidity (the *marks*). Using the terminology of Daley and Vere-Jones (2008), the point process generating the event times is called the MPP's *ground process*, the stochastic law of the inventories is the *mark process*.

To complete the definition of the liquidity model, we assume the following:

1. Conditional on the trading day, liquidity arrives (departs) according to a *nonhomogeneous Poisson ground process* with rate  $\lambda_t^+(d)$  ( $\lambda_t^-(d)$ ).
2. Conditional on the trading day, liquidity increases and decreases are drawn from the same parametric distribution with positive support, finite mean  $\mu_u$ , and finite variance  $\sigma_u^2$ . The draws are independent of each other and independent of the ground process. The distribution is the same for all relative prices  $d$ .
3. For the cumulative intensities  $\Lambda_t^+(d) = \int_0^t \lambda_s^+(d) ds$  and  $\Lambda_t^-(d) = \int_0^t \lambda_s^-(d) ds$ ,

$$\Lambda_t^+(d) \lesseqgtr \Lambda_t^-(d)$$

holds for all  $t$  during continuous trading. That is, we allow for periods where liquidity extraction is higher than liquidity submission, and this may even hold for the entire trading day. However, we assume that the order book is never empty at any  $d$ , which requires  $x_0(d) \gg \mu_u (\Lambda_t^-(d) - \Lambda_t^+(d))$  for all  $t$ .

4. The point processes are assumed to be *simple*, i.e., there are no simultaneous jumps.

We briefly discuss the model assumptions. The intensities of liquidity arrivals and departures can be viewed as doubly-stochastic or *Cox* processes, i.e., as (nonhomogeneous) Poisson processes whose parameters are themselves generated by discrete-time stochastic processes,  $(\lambda_{\ell,t}^+(d))_{\ell \in \mathbb{Z}}$  and  $(\lambda_{\ell,t}^-(d))_{\ell \in \mathbb{Z}}$ , respectively. The intensities are independent of the sizes of order submissions. This allows to estimate the parameters of the size distribution and the intensities separately based on an observed trajectory, and possibly to build a time series model for the intensities. However, the day-to-day intensity dynamics are not the focus here. Instead, we are solely interested in the realized intensities,  $\lambda_t^+(d)$  and  $\lambda_t^-(d)$ , and in the parameters of the jump size distribution.<sup>7</sup>

An important implication of the NHPP structure is that liquidity events are independent of earlier liquidity events' timing and sizes. The number of liquidity jumps during some intraday interval depends solely on the intensity, i.e. for instance

$$\#\text{order submissions at } d \text{ during } [t, t+h] \sim \text{Pois} \left( \Lambda_{t+h}^+(d) - \Lambda_t^+(d) \right).$$

Thus, the latent NHPP intensities play the same role as the stochastic volatility process  $\sigma_t^2$  in realized volatility theory: From an ex-post perspective, it does not matter if volatility itself has an autoregressive or "self-exciting" structure as for example in the Heston model (Heston, 1993), where volatility follows a Cox-Ingersoll-Ross process, or is purely exogenous but time-varying. Similarly, liquidity dynamics may be well explained by self-exciting models like ACD or Hawkes processes, reflecting market participants' reaction to other market participants' actions ("endogeneous news"), but can *ex post* be viewed as NHPPs.

Regarding the jumps, as we will show later on, the constancy of the size distribution over  $d$  is realistic, although there seems to be a slight tendency of  $\mu_u$  to grow with  $d$ . The assumption of non-emptiness is not completely realistic empirically as the book is usually empty at some price levels. This is even possible near the quotes, when market participants do not use the full denseness of the price grid.

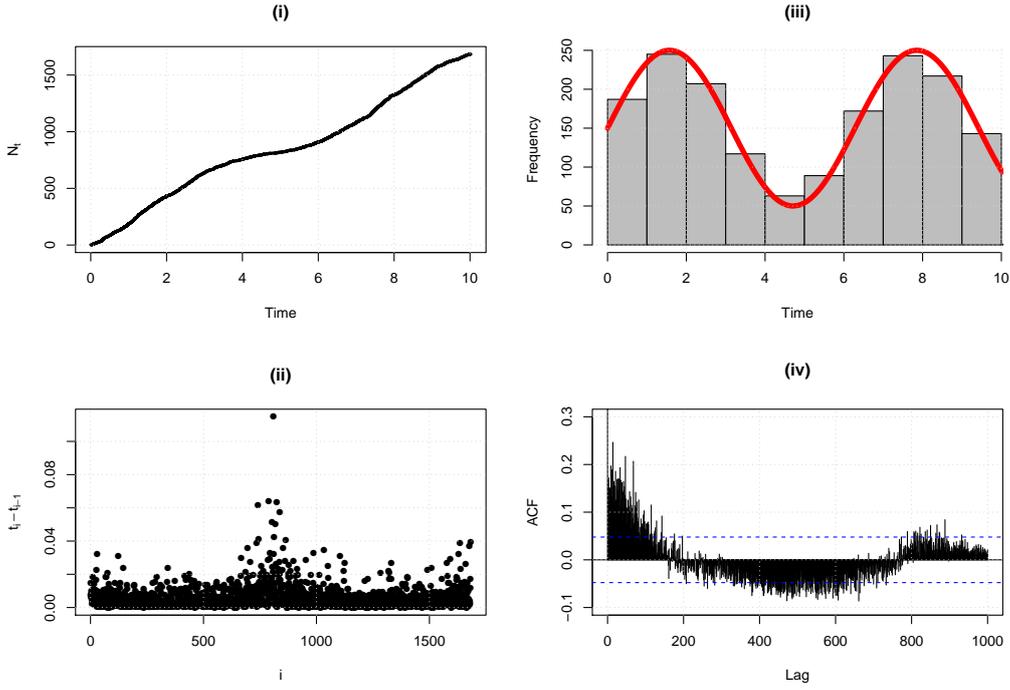
As already indicated, the "zero intelligence" models in Smith et al. (2003), Cont et al. (2010), Gatheral and Oomen (2010) are special cases of the present model, with homogeneous instead of nonhomogeneous Poisson liquidity arrival and departure processes and unit marks  $u_i^+(d) = u_i^-(d) = 1 \forall i, j$ .

### 5.3.2 Properties of the liquidity model

Conditional on the intensity  $\lambda_t^+(d)$ , the number of events  $N_{t+\Delta}^+(d) - N_t^+(d)$  during  $[t, t + \Delta]$  is Poisson distributed with parameter  $\Lambda_{t+\Delta}^+(d) - \Lambda_t^+(d)$ , thus

$$E[N_{t+\Delta}^+(d) - N_t^+(d) | \lambda_t^+(d)] = \text{Var}[N_{t+\Delta}^+(d) - N_t^+(d) | \lambda_t^+(d)] = \int_t^{t+\Delta} \lambda_s^+(d) ds, \Delta > 0$$

<sup>7</sup>In this context, *realized* means that the NHPP's intensity is deterministic ex post. However, the intensity is still a *latent* infinite-dimensional parameter of the observed point process of event times.



**Figure 5.1:** Properties of a NHPP with intensity  $\lambda_t = 100(1.5 + \sin(t))$ , simulated for  $t \in [0, 10]$ . (i) shows the counting process, (ii) shows the durations, i.e., differences between event times, (iii) shows a histogram of the simulated event times with the true  $\lambda_t$  overlaid. (iv) shows the SACF of the durations depicted in (ii). Evidently, the inhomogeneity of the intensity leads to strong autocorrelation of durations *unconditional* on  $\lambda_t$ , while jumps are independent *conditional* on  $\lambda_t$  (that is, on  $\ell$ ).

holds. The same property holds for  $N_t^-$ . All (conditional) means and variances in the following are to be understood as conditional on the day's intensity processes. We denote these conditional moments by  $E_\ell(\cdot)$ ,  $\text{Var}_\ell(\cdot)$  etc. The lack of autocorrelation or “self-excitement” of events conditional on  $\ell$ , however, still allows for (spuriously) autocorrelated durations between events through its time-inhomogeneity. Figure 5.1 illustrates this by a simulation example.<sup>8</sup>

The marks are independent of the ground process and of each other. Moreover, both ground processes are independent of each other as well. Thus, the integrated variance over some interval is given by the expected number of events, multiplied by the constant jump variance for each event,

$$\text{Var}(x_{t+\Delta}(d) - x_t(d)) =: \text{IVL}_{[t,t+\Delta]}(d) = \sigma_u^2 \left( \int_t^{t+\Delta} \lambda_s^+(d) ds + \int_t^{t+\Delta} \lambda_s^-(d) ds \right).$$

We assume that continuous trading starts after the opening auction at some initial state of the LOB,  $x_0(d)$ , which is random. The expected liquidity, conditional on

<sup>8</sup>The simulation algorithm is described below.

$x_0(d)$ , at some time during continuous trading is therefore given by

$$E_\ell[x_t(d)|x_0(d)] = x_0(d) + \mu_u \int_0^t \lambda_s^+(d) - \lambda_s^-(d) ds, \quad (5.4)$$

so that the *integrated liquidity* over the entire trading day can be defined as

$$IL_{[0,T]}(d) := \mu_u \int_0^T \lambda_s^+(d) - \lambda_s^-(d) ds. \quad (5.5)$$

Note that  $x_t(d)$  is in general *not* a martingale. To see this, consider the case where  $\lambda_t^+(d) = 2$  and  $\lambda_t^-(d) = 1$ . Then,  $x_t(d)$  is a supermartingale as  $E[x_{t+\Delta}(d) - x_t(d)|x_0(d)] > 0$  holds for all  $t$ . However, one may construct an appropriate compensator to make  $x_t(d)$  a martingale. Apart from that,  $x_t(d)$  is a martingale if  $\lambda_t^+(d) = \lambda_t^-(d)$  holds for all  $t$ .

As is shown below, we empirically find that indeed  $\lambda_t^+(d) \approx \lambda_t^-(d)$ , such that an equality assumption seems appropriate. Stated differently,  $\lambda_t^+(d) - \lambda_t^-(d)$  is very small relative to  $\lambda_t^+(d)$  and therefore contributes virtually nothing to liquidity variation. However, over the entire trading day there is typically some change in the level of liquidity. We account for this by replacing  $\int_0^T \lambda_s^+(d) - \lambda_s^-(d) ds$  by a deterministic drift  $\mu_t(d) < x_0(d) \forall t$ , yielding the model

$$x_t(d) = x_0(d) + \mu_t(d) + \sum_{i=1}^{N_t^+(d)} u_i^+(d) - \sum_{j=1}^{N_t^-(d)} u_j^-(d), \quad (5.6)$$

where  $\lambda_t^+(d) = \lambda_t^-(d)$ . Thus, conditional on  $x_0(d)$ ,  $x_t(d) - \mu_t(d)$  is a martingale.<sup>9</sup>

Due to this simplification, positive and negative jumps do not only have the same distribution, but also the same intensity. This gives rise to introducing the counting process  $N_t(d) = N_t^+(d) + N_t^-(d)$  with intensity  $\lambda_t(d) = \lambda_t^+(d) + \lambda_t^-(d)$ . Moreover, we construct a mean-zero symmetric random variable  $u_i$  as follows. Let  $f(u)$  denote the density of  $u_i^+(d)$  and  $u_i^-(d)$ , then the density of  $u_i(d)$  is given by

$$g(u) = \begin{cases} f(u)/2, & u \geq 0, \\ f(-u)/2, & u < 0. \end{cases}$$

---

<sup>9</sup>Note that if the mean of  $\lambda_t^+(d) - \lambda_t^-(d)$  was approximately zero unconditionally, but potentially large for a particular trading day, using a deterministic drift rather than an additional compound point process for the difference between bid and ask innovations would be problematic. However, we find empirically that the relative difference  $(\lambda_t^+(d) - \lambda_t^-(d))/\lambda_t^+(d)$  is usually less than 0.1 percent for all  $d$  and  $t$ . Therefore substituting the NHPP of differences by a drift is legitimate according to the data.

For instance, if liquidity jumps are exponentially distributed with mean  $\mu_u$  and variance  $\mu_u^2$ ,  $u_i(d)$  has a Laplace distribution with zero mean and variance  $2\mu_u^2$ . In the remainder of the chapter,  $\sigma_u^2$  denotes this variance which is twice the variance of  $u_i^+$  or  $u_i^-$ . Collecting all these ingredients, we finally define the liquidity process.

**Definition 5.1** (Nonhomogeneous Poisson liquidity process). *Conditional on the  $\ell$ th trading day, the outstanding number of shares at the relative price  $d$ ,  $d \in [0, D]$  for each market side  $s$ ,  $s \in \{\text{bid}, \text{ask}\}$  for times  $t$  during continuous trading,  $t \in \mathcal{T}$ , is given by*

$$x_t(d) = x_0(d) + \mu_t(d) + \sum_{i=1}^{N_t(d)} u_i(d), \quad (5.7)$$

where

- $x_0(d)$  and  $\mu_t(d)$  are fixed conditional on  $\ell$ ,
- $u_i(d)$  are i.i.d. symmetric random variables with zero mean and finite variance  $\sigma_u^2$ ,
- $N_t(d)$  is a nonhomogeneous Poisson counting process with intensity function  $\lambda_t(d)$  and cumulative intensity  $\Lambda_t(d) = \int_0^t \lambda_s(d) ds$  which is fixed conditional on  $\ell$ .

Unconditional on  $\ell$ ,

- the initial state  $x_0(d)$  has constant mean and variance for all  $d \in \mathcal{D}$ ,
- the drift is zero throughout the trading day at all relative prices,  $\mu_t(d) = 0 \forall t \in \mathcal{T}, d \in \mathcal{D}$ .

### 5.3.3 Realized measures

Having defined the liquidity process, we now turn to the task of estimation. For our purposes of liquidity measurement, it is not necessary to estimate all parameters of the process. Instead, we are interested in statistical inference on the following quantities:

- The initial state of the LOB,  $x_0(d)$ ,
- the integrated liquidity, i.e. the cumulative drift  $\int \mu_s(d) ds$ ,
- the integrated variance of liquidity,  $\sigma_u^2 \int_0^T \lambda_s(d) ds$ .

As both the initial state and the integrated drift reflect the average level of liquidity conditional on the trading day, we term them *level components* of liquidity. The integrated variance is called *variance component* of liquidity. The initial state of the LOB,  $x_0(d)$  can be observed directly. Integrated liquidity and integrated variance of liquidity as defined above are functions of the model's parameters, i.e. the variance  $\sigma_u^2$

of the marks (the liquidity jumps), the jump intensity  $\lambda_t(d)$ , and the integrated drift  $\int \mu_s(d)ds$ .

While the variance of the mark distribution is a scalar parameter, the NHPP's intensity is infinite-dimensional, and it can be estimated based on  $n_\ell$  observed jumps using kernel methods. However, it is not necessary to infer the entire intraday (spot) liquidity trajectory, but the following cumulative quantities for a given trading day only.

**Definition 5.2** (Integrated measures of liquidity). *The integrated liquidity over a trading day is given by*

$$IL_{[0,T]}(d) = \int \mu_s(d)ds. \quad (5.8)$$

*The integrated variance of liquidity is defined as*

$$IVL_{[0,T]}(d) = \sigma_u^2 \int_0^T \lambda_s(d)ds = \sigma_u^2 \Lambda_t(d) \quad (5.9)$$

Both measures depend solely on quantities that are deterministic conditional on the trading day. IVL can be splitted into the expected number of events, determined by the intensity of the ground process, and the mark variance, as times and marks are independent.

Turning to estimation, we observe that by definition of the data-generating process,

$$E_\ell(x_T(d) - x_0(d)) = \int_0^T \mu_s(d)ds = IL_{[0,T]}. \quad (5.10)$$

Now consider the case where  $M + 1$  (not necessarily equidistant) snapshots of the LOB are observed,  $x_{t_0}(d), x_{t_1}(d), \dots, x_{t_{M-1}}(d), x_{t_M}(d)$ ,  $t_0 = 0, t_M = T$ . A natural, unbiased estimator for  $IL$  is then given by

$$\hat{IL}_{[0,T]}(d) = \sum_{i=1}^M x_{t_i}(d) - x_{t_{i-1}}(d) = x_T(d) - x_0(d), \quad (5.11)$$

that is, the intraday observations do not provide any valuable information on the drift. This result is the same as for the drift in a (jump-) diffusion setting in the standard stochastic price volatility model in Section 5.2 (Andersen and Benzoni, 2009). Note that the variance of the estimator  $\hat{IL}_{[0,T]}(d)$  depends on the variance of liquidity.

A natural estimator of integrated liquidity variation has already been mentioned: The estimator of Engle et al. (2012), which for general, not necessarily equidistant LOB snapshots may be written as

$$I\hat{V}L_{[0,T]}(d) = \sum_{i=1}^M (x_{t_i}(d) - x_{t_{i-1}}(d))^2. \quad (5.12)$$

As

$$E[x_t(d) - x_0(d)] = E\left[\mu_t(d) + \sum_{i=1}^{N_t(d)} u_i(d)\right] = \mu_t(d)$$

holds, it is instantly clear that  $I\hat{V}L$ , conditional on the trading day, is biased due to the presence of a drift. A thorough analysis of the estimator's properties follows in the next Section. We call  $\hat{I}L$  and  $I\hat{V}L$  *realized measures of liquidity* (RML) in the remainder of the thesis, and collect them in Definition 5.3.

**Definition 5.3** (Realized measures of liquidity). *The realized liquidity at relative price  $d$  on trading day  $\ell$ ,  $RL_\ell(d)$ , is given by*

$$RL_\ell(d) = x_{T,\ell}(d) - x_{0,\ell}(d). \quad (5.13)$$

*Given LOB snapshots  $x_{t_0,\ell}(d), x_{t_1,\ell}(d), \dots, x_{t_{M-1},\ell}(d), x_{t_M,\ell}(d)$ ,  $t_0 = 0$ ,  $t_M = T$ , the realized variance of liquidity at relative price  $d$  on trading day  $\ell$ ,  $RVL_\ell(d)$ , is given by*

$$RVL_\ell(d) = \sum_{i=1}^M (x_{t_i,\ell}(d) - x_{t_{i-1},\ell}(d))^2. \quad (5.14)$$

### 5.3.4 Properties of realized measures of liquidity

We are interested in the error that is made when the realized measure is computed using LOB snapshots. To answer this question analytically, assume for the moment that the drift is negligible when measuring liquidity variation, i.e.

$$RVL_\ell(d) = \sum_{i=1}^M (x_{t_i,\ell}(d) - x_{t_{i-1},\ell}(d))^2 \approx \sum_{i=1}^M \left( x_{t_i,\ell}(d) - x_{t_{i-1},\ell}(d) - \int_{t_{i-1}}^{t_i} \mu_s(d) ds \right)^2,$$

which is quite realistic empirically. We therefore may write the model as

$$x_t(d) = x_0(d) + \sum_{i=1}^{N_t} u_i, \quad (5.15)$$

which, apart from its functional nature in the liquidity context, is the model of Oomen (2006) for the situation without microstructure noise. Let  $\Delta x_{[t,t+h],\ell}(d)$  denote the liquidity change in the interval  $[t, t+h]$  during day  $\ell$ , then

$$\Delta x_{[t,t+h],\ell}(d) = \sum_{i=N_t(d)+1}^{N_{t+h}(d)} u_i. \quad (5.16)$$

Now, for any sampling scheme  $TS_M = \{t_0, \dots, t_M\}$ , the sequence  $(\Delta x_{[t_m, t_{m+1}]})_{t_m \in TS_M}$  is a martingale difference sequence. To analyze the properties of RVL under alternative sampling schemes, we define the following sampling schemes, again following Oomen (2006).

**Definition 5.4** (Sampling schemes).

- *Calendar time sampling:* When liquidity  $x_t(d)$  is sampled equidistantly in calendar time,  $CTS_M$ , the sample is given by  $\{x_{t_i^c}(d)\}_{i=0}^M$ , where  $t_i^c = i/M$ .
- *Business time sampling:* When liquidity  $x_t(d)$  is sampled equidistantly in business time,  $BTS_M$ , the sample is given by  $\{x_{t_i^b}(d)\}_{i=0}^M$ , where  $t_i^b = \Lambda^{-1}(i\Lambda(T)/M)$ .
- *Event time sampling:* When liquidity  $x_t(d)$  is sampled equidistantly in event time,  $ETS_M$ , the sample is given by  $\{x_{t_i^e}(d)\}_{i=0}^M$ , where  $t_i^e = \inf N_{iN_T/M}^{-1}$ .

For  $CTS_M$ , the  $M$  observations are equidistant in the usual sense. For  $BTS_M$ , time is compressed and stretched according to the expected frequency of events which is completely determined by  $\Lambda_t(d)$ . Durations between events sampled in business time, measured in calendar time, are smaller than in calendar time for  $\lambda_t(d) > \Lambda_T(d)/T$ , and larger for  $\lambda_t(d) < \Lambda_T(d)/T$ . If liquidity innovations arrive according to a homogeneous Poisson process,  $BTS_M$  and  $CTS_M$  are equivalent. Finally, for  $ETS_M$ , time is stretched or compressed as in  $BTS_M$ , but according to the events' *actual* rather than their *expected* frequency.

We confine ourselves to  $CTS_M$  and  $ETS_M$  as we seek to gauge the potential loss of information by using  $M$  equidistant instead of all snapshots when, in principle, all LOB changes can be observed. Using the characteristic function of the compound non-homogeneous Poisson liquidity process, the mean squared error for the two schemes can be explicitly computed.

**Proposition 5.1** (MSE of the realized variance of liquidity under alternative sampling schemes, Oomen (2006)).

The MSE of the realized variance under calendar time sampling is given by

$$\begin{aligned} \text{MSE}^{\text{CTS}_M}(d) &= E_\ell \left[ \left( \text{RVL}_\ell^{\text{CTS}_M}(d) - \sigma_x^2(d) \right)^2 \right] \\ &= 3\sigma_u^2\sigma_x^2 - \sigma_x^4 + 2\sigma_u^4 \sum_{i=1}^M \lambda_i^2 + \sigma_u^4 \sum_{i=1}^M \sum_{j=1}^M \lambda_i\lambda_j, \end{aligned} \quad (5.17)$$

where

$$\lambda_i = \int_{t_{i-1}}^{t_i} \lambda_s(d) ds. \quad (5.18)$$

The MSE of the realized variance under event time sampling is given by

$$\text{MSE}^{\text{ETS}_M}(d) = E_\ell \left[ \left( \text{RVL}_\ell^{\text{ETS}_M}(d) - \sigma_x^2(d) \right)^2 \right] = 2\sigma_u^4/M, \quad (5.19)$$

where  $\sigma_x^2(d) = \sigma_u^2\Lambda_{T,\ell}(d)$  denotes the true integrated variance of liquidity at relative price  $d$  on day  $\ell$ .

The proof, i.e., the derivation of the characteristic function and moments can be found in Oomen (2006). The difference between the two MSEs is

$$\text{MSE}^{\text{CTS}_M}(d) - \text{MSE}^{\text{ETS}_M}(d) = 2\sigma_u^4(d) \sum_{i=1}^M (\lambda_i(d) - \bar{\lambda}(d))^2 + 3\sigma_u^4(d)\Lambda_T(d) > 0,$$

where  $\bar{\lambda}(d) = \Lambda_T(d)/T$ . That is, the first term is due to the temporal nonhomogeneity of liquidity increments. It would be zero for a homogeneous Poisson process. The second term is due to sampling variability, i.e., due to the difference between the expected and actual frequency of events.

The practical implications of these results for our desired application are investigated by simulation in Section 5.4.

## 5.4 The NHPP liquidity model: Estimation and simulation

We are primarily interested in the realized measures as defined above. To investigate the properties of the realized measures, we use a simulation in a realistic setting. To this end, we first show how the NHPP can be estimated, then use the estimated rate  $\hat{\lambda}_t$  to simulate.

### 5.4.1 Estimation

We illustrate the marked NHPP process of liquidity innovations using XETRA data, namely for the Deutsche Bank stock as traded on March 30, 2010, which is one of the most heavily traded stocks on XETRA. Some descriptive information on the number of limit order submissions and their average sizes are shown in Table 5.1.

	all	d=0	d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8	d=9	d=10
number buy	101555	25554	23259	17632	9353	7244	4480	2500	1854	1146	1074	846
number sell	111979	27892	24888	19034	10387	8203	5195	3007	2146	1352	1247	956
avg. size buy	2161	1684	2303	2363	2186	2057	2090	2147	2131	1876	2757	2158
avg. size sell	2159	2156	2200	2197	2133	2035	2158	2169	2091	2051	2155	2092

**Table 5.1:** Descriptive statistics for the number of orders and mean order sizes, computed for the XETRA data, Deutsche Bank stock, 2010-03-30. Most submissions are placed near the quotes, and the mean size is almost constant with respect to the relative price  $d$  (measured in Cents). The distributions of both quantities over  $d$  are very similar for buy and sell orders, respectively.

An estimate of the arrival (departure) intensity  $\lambda_t^+(d)$  ( $\lambda_t^-(d)$ ) is obtained using the nonparametric kernel estimator of Diggle and Marron (1988), which essentially is a rescaled *density* estimator. We use the quartic or biweight kernel,  $K(x) = .9375(1 - x^2)^2$  for  $x \in [-1, 1]$  and zero otherwise, and denote the estimator's bandwidth by  $h$ . Rescaling trading time to  $[0, 1]$ ,  $T = 1$ , the estimator is then given by

$$\hat{\lambda}_t(d) = \begin{cases} h^{-1} \sum_{i=0}^{N_T} K\left(\frac{t-t_i}{h}\right) + K\left(\frac{t+t_i}{h}\right), & t \in [0, h), \\ h^{-1} \sum_{i=0}^{N_T} K\left(\frac{t-t_i}{h}\right), & t \in [h, 1-h), \\ h^{-1} \sum_{i=0}^{N_T} K\left(\frac{t-t_i}{h}\right) + K\left(\frac{t+t_i-2}{h}\right), & t \in (1-h, 1], \end{cases}$$

where  $t_0, \dots, t_{N_T}$  are the event times of the nonhomogeneous Poisson ground process  $N_t$ . Estimates within a distance of  $h$  from the edges have a downward bias as there are no changes in liquidity outside  $[0, 1]$ . The additional terms for  $t \in [0, h) \cup (1-h, 1]$  correct for this effect.

We use Silverman's rule of thumb, Silverman (1986), to determine the bandwidth parameter  $h$ . The rule has originally been proposed for the Gaussian kernel. However, we find that theoretically more attractive strategies for bandwidth selection, namely cross validation or pilot estimation of derivatives, yield virtually the same results in our data-rich situation, while being computationally much more expensive. Figure 5.2 shows the estimation results. Submission activity has a first peak in the morning, then stays at roughly half of its peak level from about 11am to 2pm, then again increases before the opening of US markets at 3:30pm, peaking at about 4pm. We find that both market sides behave in a quite similar way (top panel). Moreover, limit order arrivals (i.e., submissions) and departures (cancellations and opposite-side market orders), conditional on  $d$ , behave in a very similar way as well, i.e.  $\lambda_t^+(d) \approx \lambda_t^-(d)$  holds throughout the trading day (center panel). The arrival frequency near the quotes exceeds the departure frequency slightly at the beginning of the day. At the end

of the day, the opposite is true. During most of the meantime, in- and outflow of liquidity balance each other.

We now turn to estimating the distribution of liquidity innovations. Mean and variance of limit order sizes as a function of  $d$  are quite flat, therefore we work with one innovation distribution for all  $d$ . Figure 5.3 shows a histogram of jumps, overlaid with an exponential and a log-normal density fitted to the data, respectively. Both parametric distributions underestimate the probability of medium-sized orders (size 3000-4000). Apart from that, we found that, according to raw autocorrelations, the assumption of serial independence (in event time) is realistic both unconditional on  $d$  and conditional on  $d$  at some distance from the quotes,  $d > 5$  say. However, liquidity innovations near the quotes exhibit positive autocorrelation which could be described by a low-order AR process. This means that our measure of liquidity uncertainty may be slightly biased near the quotes. Using thinned data (equidistant snapshots at lower frequency), no serial dependence remains. This finding is similar (though not identical) to the MA(1) serial dependence structure of high-frequency returns on security prices and may therefore also be seen as a "microstructure noise" phenomenon.

We choose to ignore serial dependence in the present analysis as we do not primarily aim at accurately characterizing liquidity innovations' behavior in every aspect, but to construct realized measures capturing the essential information.

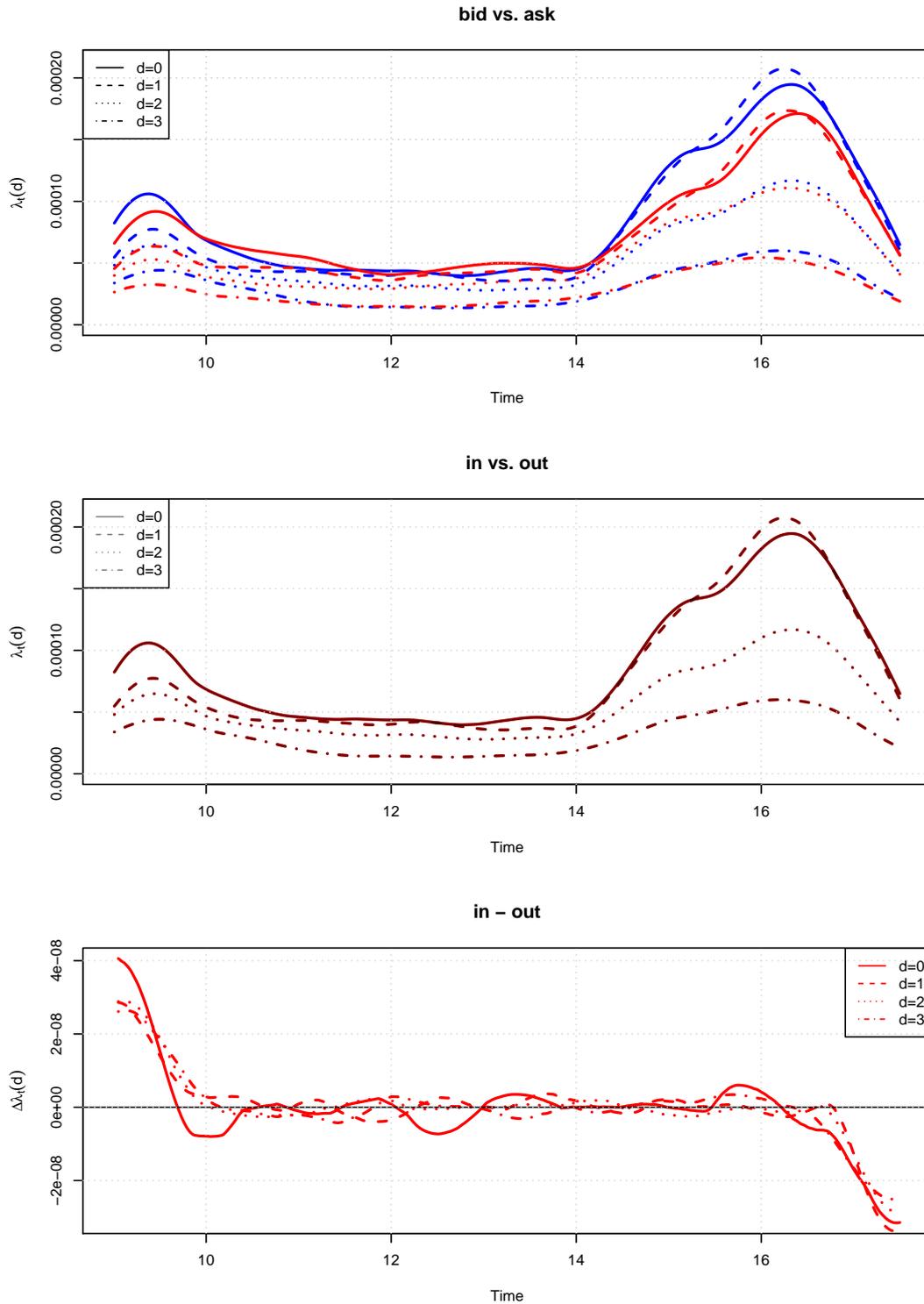
### 5.4.2 Simulation

The estimated intensities and innovation distribution can be used to investigate the finite-sample performance of realized liquidity estimators in a realistic setting, based on all observations and based on a smaller number of equidistant snapshots. As we have shown, estimation of the level components ( $x_0(d)$  and  $x_T(d) - x_0(d)$ ) does not at all benefit from using intraday data. We therefore confine ourselves to the analysis of realized variance of liquidity (RVL). Marks are simulated from an exponential distribution with a realistic parameter (1/2000, implying a mean of 2000). To simulate from the ground process, we use the above estimates, in particular  $\hat{\lambda}_t^+(d)$  and  $\hat{\lambda}_t^-(d)$  for  $d = 0, 5, 10$  and sell limit orders.

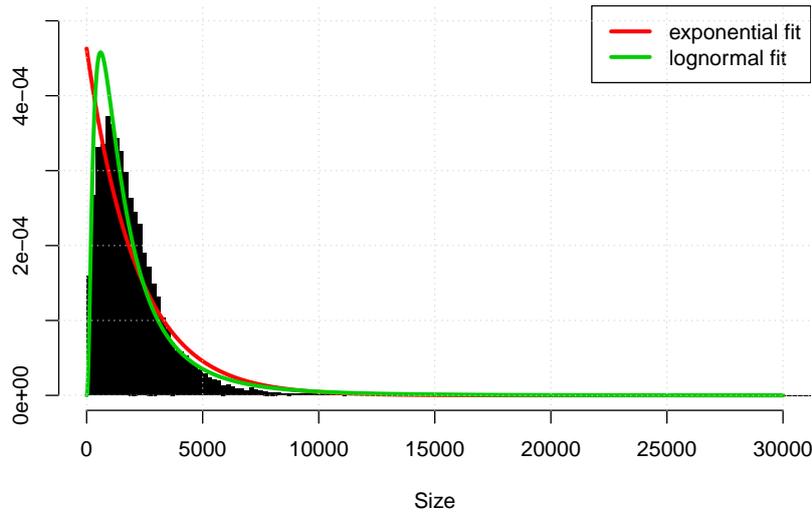
The simulation algorithm proceeds as follows.<sup>10</sup>

---

<sup>10</sup>For simulation of nonhomogeneous Poisson processes see Pasupathy (2011), moreover Charpentier (2012) for the implementation in R.



**Figure 5.2:** Estimated intensities for the Deutsche Bank stock. Top panel: Submission intensities  $\hat{\lambda}_t^+(d)$  of buy (blue) and sell (red) limit orders at a distance of  $d$  to the respective quote. Center panel: Submission intensities  $\hat{\lambda}_t^+(d)$  vs. extraction intensities  $\hat{\lambda}_t^-(d)$  for sell limit orders and several  $d$ . Bottom panel: Difference between liquidity inflow and outflow intensities,  $\hat{\lambda}_t^\Delta(d) = \hat{\lambda}_t^+(d) - \hat{\lambda}_t^-(d)$ . The rate is estimated based on time measured in  $10^{-4}$  seconds. For instance, this means that a rate of 0.0001 corresponds to one event per second.



**Figure 5.3:** Empirical distribution of liquidity innovations vs. two parametric fits. An exponential fit, with only one parameter for location and scale, is insufficient, particularly as the size distribution has a mode apart from zero. A location-scale distribution like the lognormal is more, but still not entirely appropriate.

**Algorithm 5.1** (Simulation from a marked nonhomogeneous Poisson process).

1. Specify some  $T$ , the time until which the process is simulated,
2. start in  $t = 0$ ,
3. draw  $y$ , the duration until the next jump, from its conditional distribution, which depends on  $t$  only and is exponential with cdf

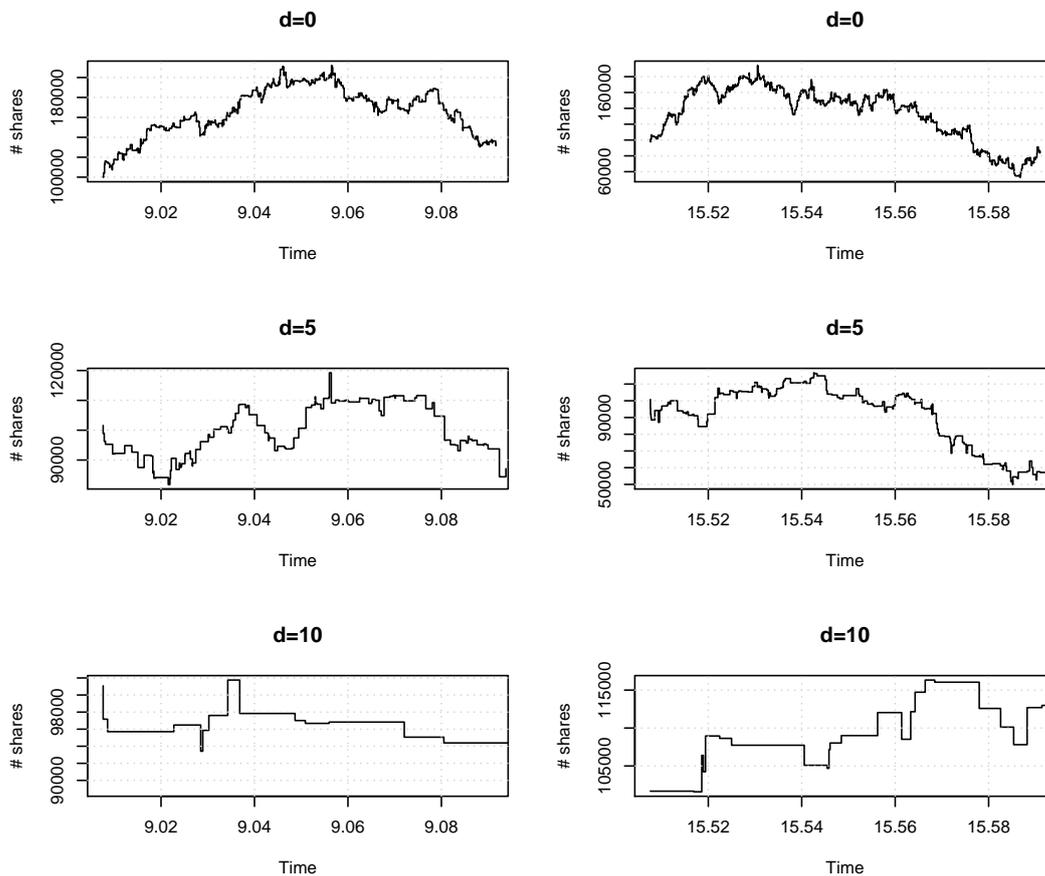
$$F_t(y) = 1 - \exp(\Lambda(y) - \Lambda(y + t)) = 1 - \exp\left(-\int_t^{t+y} \lambda(s) ds\right),$$

4. set  $t$  to  $t + y$ ,
5. store  $t$ ,
6. proceed with step 3 unless the stopping criterion  $t \geq T$  is met.
7. Simulate  $n$  i.i.d. realizations from the distribution of marks,  $u_i$ , yielding the simulated path of the MPP,

$$(t_1, u_1), (t_2, u_2), \dots, (t_n, u_n).$$

The compound NHPP is strictly nondecreasing in  $t$ . Liquidity is given by the difference of two compound NHPPs. For illustration, Figure 5.4 depicts simulated trajectories of the model at different  $d$  and times during the trading day.

For each of the intensity processes, we simulate  $B = 100$  trajectories of  $x_t(d)$  with



**Figure 5.4:** Simulated trajectories of ask volume generated by the nonhomogeneous Poisson liquidity process with submission and extraction intensities estimated from sell order submissions for one exemplary trading day. From top to bottom: Volume evolution at a distance of  $d = 0$  (5, 10) ticks from the ask quote. The left panel shows trajectories directly after the start of continuous trading, the right panel after opening of US markets. Apparently, submission and revision activity is higher in the afternoon, leading to higher fluctuations and, thus, higher liquidity uncertainty. Moreover, activity is much higher near the quote than at greater distance.

exponential jumps whose mean size is 2000, starting at  $x_0(d) = 10^6$ . For each simulation run, we compute RVL using the following strategies:

1.  $(x_T(d) - x_0(d))^2$ , i.e., the squared open-to-close difference,
2.  $\sum_{i=1}^M (x_{iT/M}(d) - x_{(i-1)T/M}(d))^2$ , i.e., the sum of squared differences over  $M$  equidistant intraday snapshots of the LOB,
3.  $\sum_{i=1}^{N_T} D_i^2$ , the sum over all squared intraday changes,
4.  $\underbrace{2\mu_D^2}_{=\text{Var}(D_i)} \underbrace{2 \int_0^T \lambda_s(d) ds}_{=\Lambda_T^+ + \Lambda_T^-}$ , i.e., the truth based on the model's parameters.

The results, shown in Figure 5.5, confirm the unbiasedness of the estimators. Not surprisingly, the estimation based on  $M$  equidistant sampling points (as compared to all  $N_T(d)$  observed changes) is the better, the lower  $N_T(d) - M$  is. In our empirical application, we assume that there are no microstructure effects and use  $M = 500$  sampling points.<sup>11</sup>

## 5.5 Empirical results

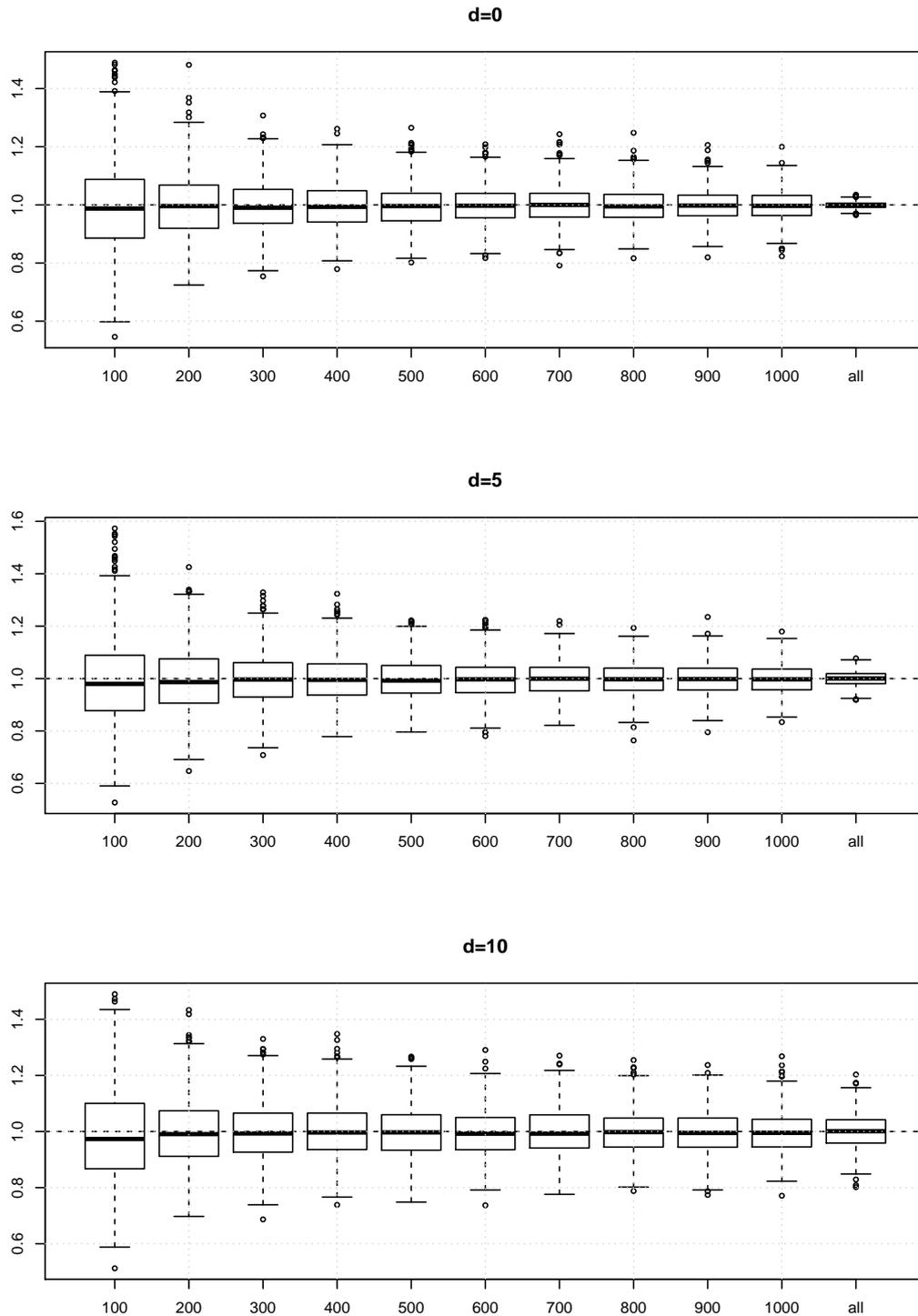
From a sample of  $L$  trading days, we obtain the following time series of realized measures,

- $(x_{0,\ell}^{(s)}(d))_{\ell \in \{1, \dots, L\}}$
- $(RL_{[0,T],\ell}^{(s)}(d))_{\ell \in \{1, \dots, L\}}$
- $(RVL_{[0,T],\ell}^{(s)}(d))_{\ell \in \{1, \dots, L\}}$

where  $d$  is observed along  $J = 201$  equidistant grid points (EUR 0 to EUR 2 by 1-Cent steps), and  $s \in \{\text{bid}, \text{ask}\}$ . To exclude auction effects, we set " $t = 0$ " to 9:05am, exclude the interval from 12:59pm to 1:05pm, and set  $T$  to 5:29pm. We use equidistant LOB snapshots, taken every minute, providing  $M = 511 - 15 = 496$  observations per day.

$x_{0,\ell}^{(s)}$  is already a cumulative volume curve. From  $RL$  and  $RVL$ , estimated individually at every  $d$ , we construct *cumulative* realized liquidity curves in the following way.

<sup>11</sup>A one-minute sampling scheme provides 511 snapshots. Excluding the first 5 minutes, 5 minutes near the midday auction at 1pm, and the last one at 5:30pm leaves us with exactly  $M = 500$ .



**Figure 5.5:** Results of the simulation study. The realized liquidity variance at distances of  $d = 0, 5, 10$  from the ask quote is computed for each of the  $B = 1000$  simulation runs, based on  $M = 100, 200, \dots, 1000$  equidistant LOB snapshots, and based on all (approximately  $\hat{\Lambda}_T^+(0) + \hat{\Lambda}_T^-(0) \approx 10^5$ ) liquidity jumps. We have normalized all RVLs by  $\hat{\Lambda}_T^+(0) + \hat{\Lambda}_T^-(0)$ , so that the truth (dashed line) is 1, and errors can be interpreted as relative deviations from the truth.

**Definition 5.1** (Cumulative realized measures). *The cumulative realized liquidity (CRL) at side  $s$ , tick  $d$  and day  $\ell$  is defined by*

$$x_{\ell}^{(s),RL}(d) = \sum_{k=0}^d \left( x_{T,\ell}^{(s)}(k) - x_{0,\ell}^{(s)}(k) \right),$$

*and the cumulative realized variance of liquidity (CRVL) by*

$$x_{\ell}^{(s),RVL}(d) = \sum_{k=0}^d \left( \sum_{i=1}^M (x_{t_i,\ell}(k) - x_{t_{i-1},\ell}(k))^2 \right).$$

For the sake of notational consistency, we rewrite the third realized measure,  $x_{0,\ell}^{(s)}$ , as  $x_{\ell}^{(s),0}$  in the remainder of the Chapter. Moreover, in analogy to the variation of the price, we employ logarithmic liquidity variation,  $\log \left( x_{\ell}^{(s),RVL}(d) \right)$ , instead of the original series.

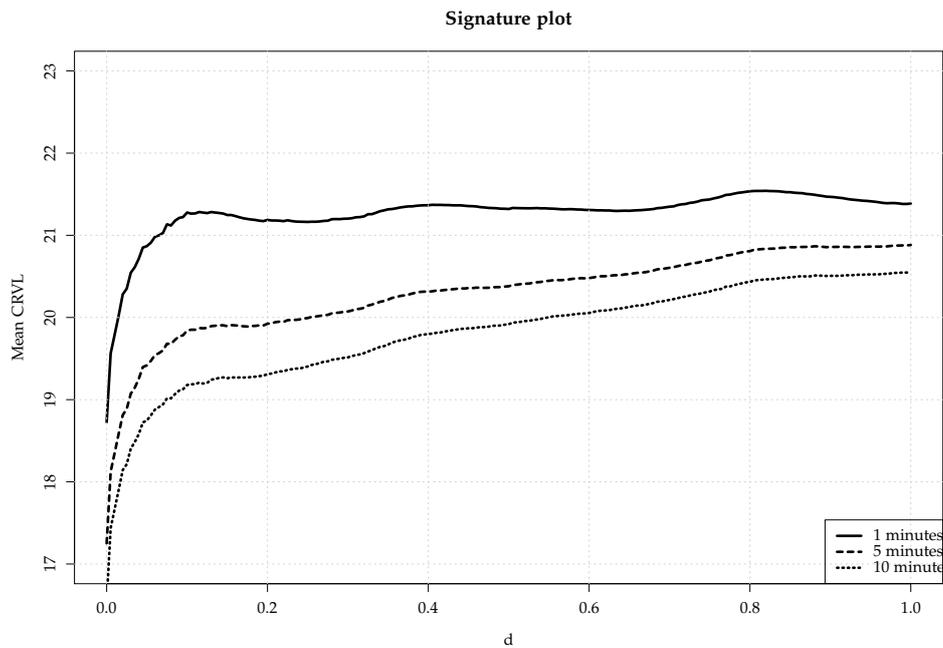
Whereas the continuous-time model describing liquidity's intraday behavior is essentially a random walk with time-varying innovation arrivals and drift, i.e., non-stationary, we assume the time series of realized measures to stem from a weakly dependent functional stochastic process as defined in Chapter 2. Each of the six time series (three measures for two market sides) can then be represented in terms of a constant mean function and time-varying FPCs,

$$x_{\ell}^{(s),\cdot}(d) = \mu^{(s),\cdot}(d) + \sum_{k=1}^{\infty} \tilde{\zeta}_{k,\ell}^{(s),\cdot} \phi_k^{(s),\cdot}(d).$$

The mean function  $\mu^{(s),\cdot}(d)$  and eigenfunctions  $\phi_k^{(s),\cdot}(d)$  can then be estimated, and the empirical scores  $\hat{\zeta}_{k,\ell}^{(s),\cdot}(d)$  be computed as usual. Figure 5.7 shows the realized measures exemplarily for the Linde stock. For CRVL, a sampling frequency of 5 minutes has been used. We found that there is a signature pattern present, i.e. a monotonic connection between sampling frequency and realized variance, as it is already known in case of price volatility, see Figure 5.6. In the empirical application, we use 5-minute sampling for both prices (RV) and liquidity (RML).

To assess the structure of realized measures, we consider Figure 5.7, which shows the three functional realized measures for the three stocks, Figure 5.8, which depicts the corresponding eigenfunctions for the first three FPCs, and Figure 5.10, which shows the cumulative normalized eigenvalues of the first 20 components for all three stocks. The corresponding Figures for MunichRe and Commerzbank, A.1, A.4 (realized measures), and A.2, A.5 (eigenfunctions), can be found in Appendix A.5. All these Figures reveal information on all three realized measures and both market sides.

Regarding the structure, we find that all three realized measures vary mainly in their level and slope, CRVL even seems to consist of parallel shifts of just one concave factor. CRVL's major mode of variation, reflected by the first eigenfunction, explains at least 85 percent of the variation in all cases. In case of RVL, the first component even explains more than 95 percent. For the two realized measures of the liquidity *level*,

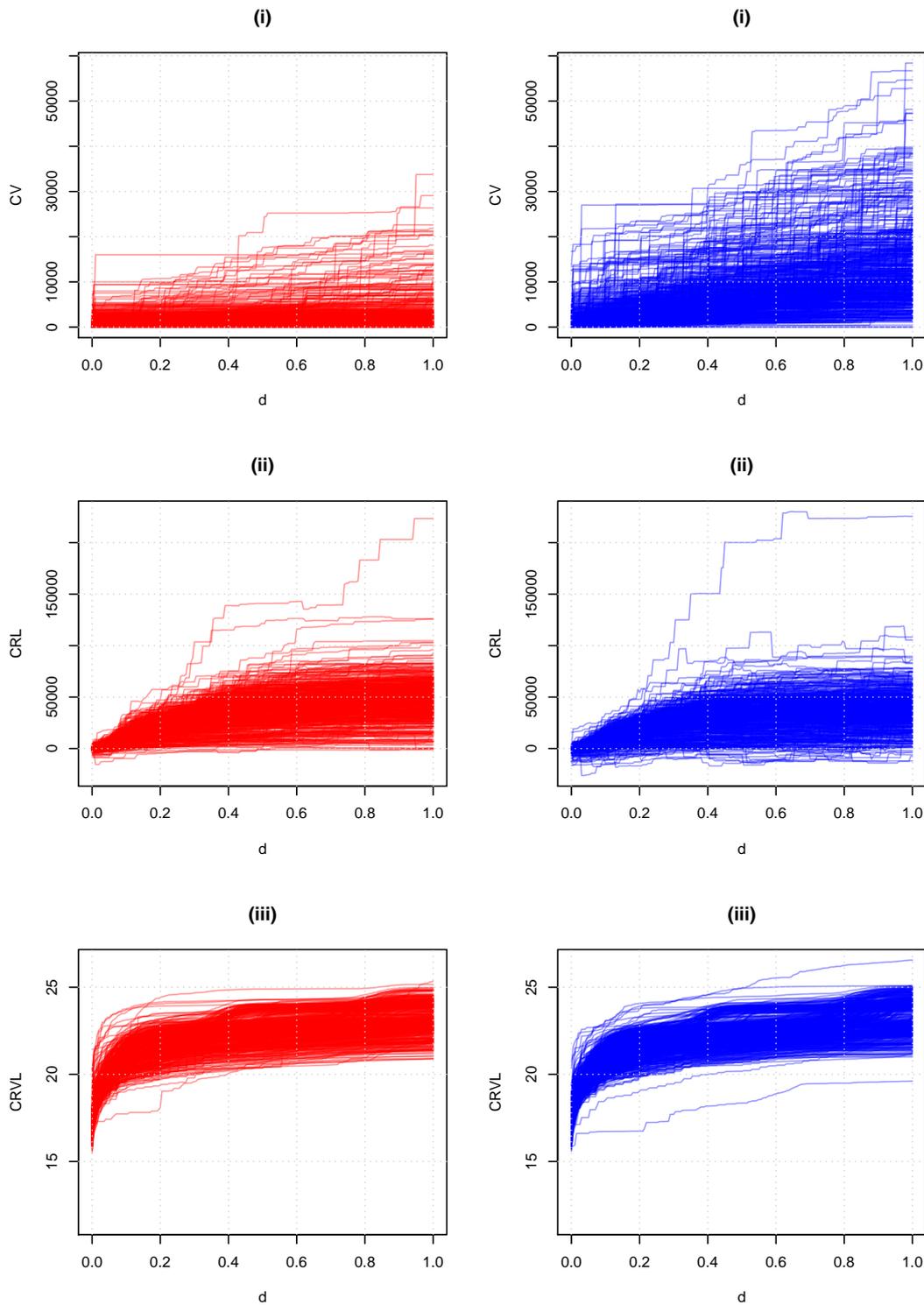


**Figure 5.6:** Mean cumulative realized variance curves for Linde’s ask curves, for sampling frequencies of 60, 12, and 6 snapshots per hour (corresponding to 1, 5, 10 minutes between snapshots). The three values for given  $d$  can be interpreted as signature plot, in analogy to the price volatility. In analogy to price volatility, there is a positive relationship between sampling frequency and CRVL.

namely initial liquidity and liquidity drift, the corresponding eigenfunction is almost linear with respect to  $d$ .<sup>12</sup> That is, the first component reflects the slope of initial liquidity and the drift. For the third realized measure, the log of cumulative liquidity variance which is supposed to capture liquidity uncertainty, the first eigenfunction is almost flat and parallel to the relative price axis. This means that variation is dominated by the variation near the quotes.

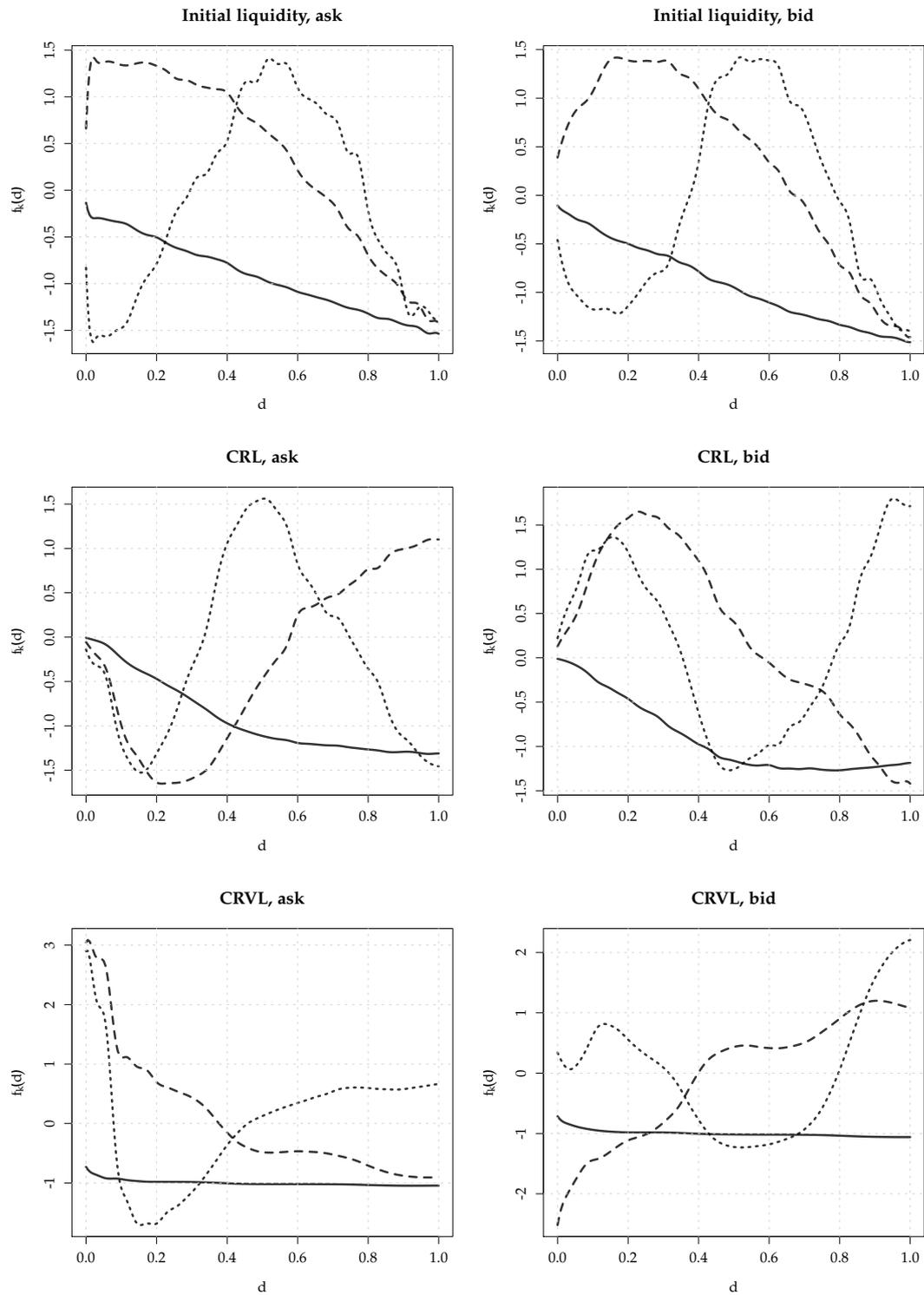
For the third measure and the ask side, we additionally look at the dynamic properties by means of the first three FPCs’ weights over time, which are depicted in Figure 5.9 for Linde, along with their SACFs. Figures A.3 and A.6 for MunichRe and Commerzbank are in Appendix A.5. We find that especially the first component exhibits a quite persistent serial dependence for all stocks considered, in roughly the same way as it is the case for realized variation of prices.

<sup>12</sup>Note that the eigenfunctions are unique only up to their sign, it therefore does not matter whether the slope is positive or negative.

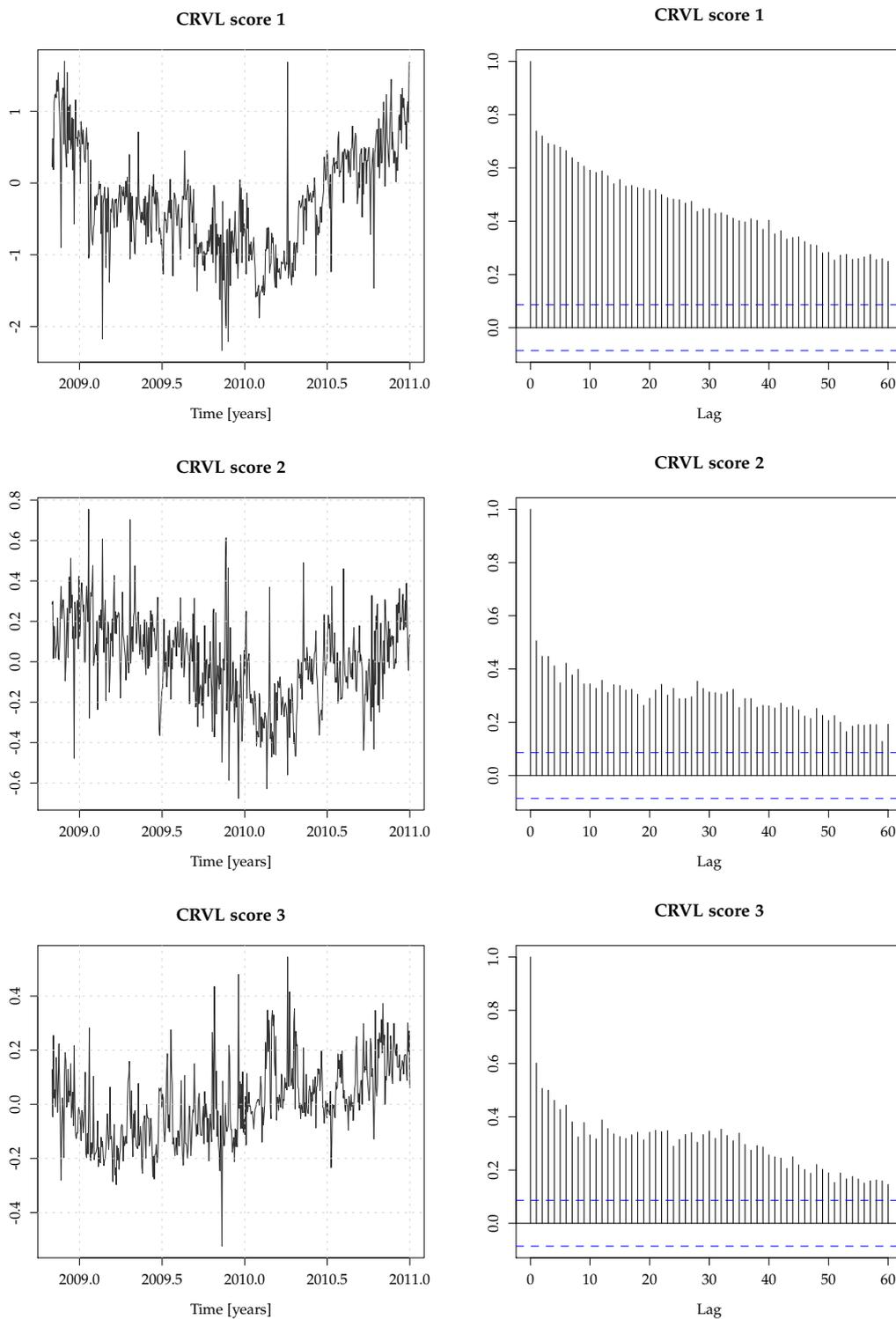


**Figure 5.7:** Top to bottom: (i) Initial liquidity, (ii) cumulative realized liquidity, and (iii) cumulative realized log variance of liquidity for the Linde stock. Each curve represents one day between 3 November 2008 to 30 December 2010 (incomplete trading days removed). Initial liquidity is strictly positive and strictly non-decreasing by construction. The drift is positive for most of the days, which is in line with the findings for the diurnal pattern in Chapter 3. The liquidity variance, computed based on 5-minute snapshots, almost exclusively varies in its level, i.e., CRVL curves appear to be almost parallel, which is also reflected by their eigenstructure.

## 5 Realized measures of liquidity

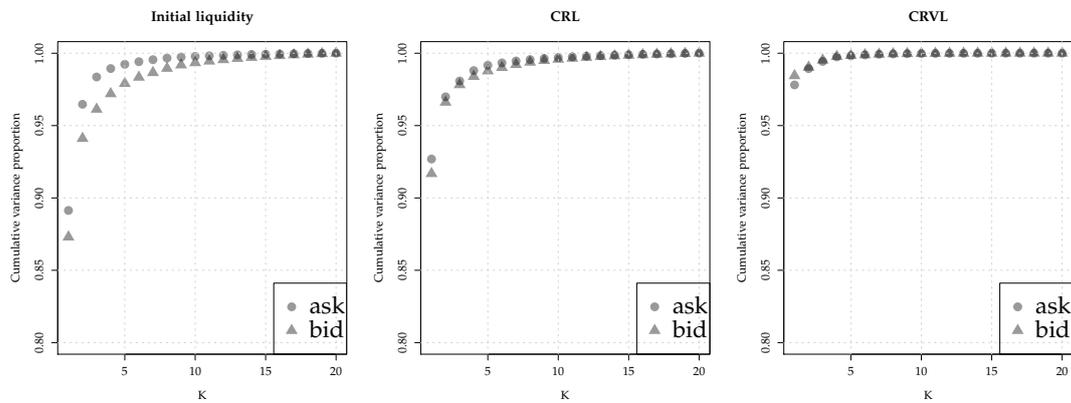


**Figure 5.8:** First three estimated eigenfunctions for all six realized measures, Linde.

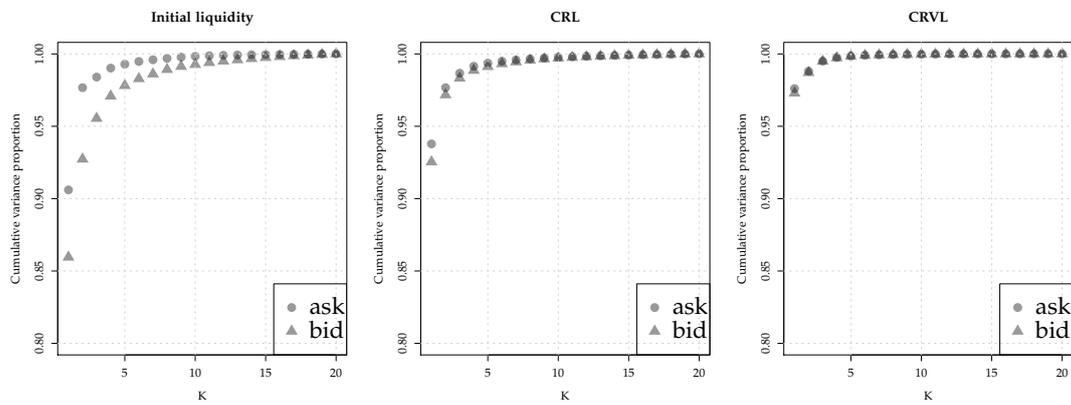


**Figure 5.9:** First three FPC score series and their SACF for Linde's ask curves. The first series, which carries most of the information according to the eigenvalues analysis, is strongly serially dependent and quite persistent, indicating that liquidity variation itself has these properties as well.

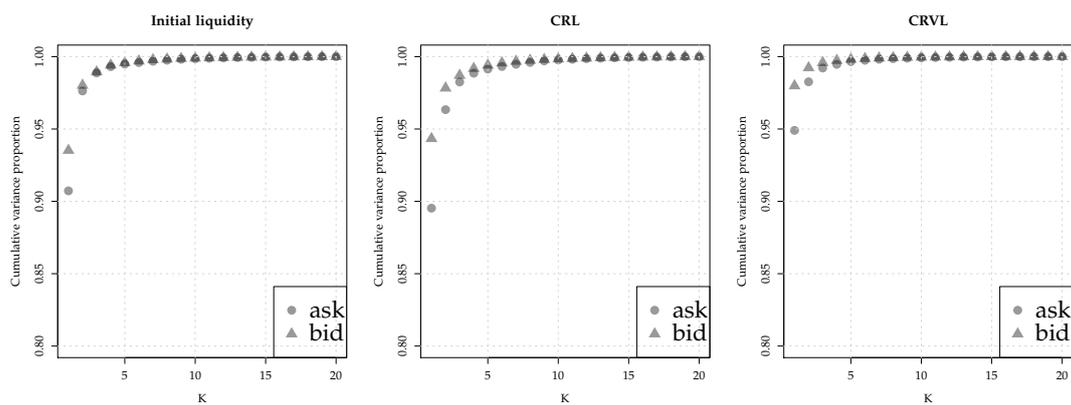
(i)



(ii)



(iii)



**Figure 5.10:** Cumulative relative eigenvalues as a function of  $K$ , i.e., the variance proportion explained by the first  $K$  FPCs, for (i) Linde, (ii) MunichRe, and (iii) Commerzbank.

## 5.6 Conclusion

In the present Chapter, we introduce realized measures of liquidity (RML) based on equidistant snapshots of the LOB, in close analogy to the realized volatility literature. Cumulating these measures along LOB locations  $d$ , *functional* RML arise. By constructing a semiparametric model with ingredients from Press (1967) and Oomen (2006), capturing many of the LOB's typical properties, we provide theoretical justifications for the ad-hoc adoption of realized volatility in the liquidity context. By doing this, the model supplies some supporting arguments for the RML proposed in Engle et al. (2012) en passant.

We again note that the assumptions regarding the data-generating process of liquidity, as far as the intraday dynamics are concerned, are quite different from those in *Part I*: In *Part I*, we assume a stationary process throughout, while in *Part II* we, basically, assume a stationary sequence of "parameters" governing the nonstationary intraday dynamics. Our empirical analysis reveals that liquidity increments, at least at very high frequency, are serially dependent, leading to a non-flat signature plot. As far as subsequent analyses in this thesis are concerned, we alleviate this problem by using a rather low sampling frequency. However, we may develop realized measures accounting for this "noise" phenomenon in future research.

In Chapter 6, functional RML are employed to investigate the liquidity impact on the dynamics of price variation from a *realized* perspective.



## Realized volatility and realized liquidity: The HAR-FunXL model

### 6.1 Introduction

In *Part I*, we have considered models capturing the state of the LOB *inventories* in the form of snapshots taken at specific times during continuous trading. We have modeled their impact on the price process in the framework of multiplicative error models (MEMs). In the present Chapter, we introduce the HAR-FunXL process to model realized volatility dynamics with liquidity impact. Liquidity is measured by means of realized measures introduced in Chapter 5.

The MEMs in Part I are non-Markovian models<sup>1</sup> as the log of squared returns (the log of durations) has an AR( $\infty$ ) representation. In contrast, the model to be introduced in the present Chapter is Markovian of finite order  $m$ . As we will show, in a Markovian framework the absence of a latent component ( $\sigma_t$  for GARCH,  $\psi_i$  for ACD), that is, the absence of unobservable regressors, facilitates the use of additional estimation and roughness penalization techniques originally proposed within the framework of non- and semiparametric regression.

The Chapter proceeds as follows. In Section 6.2, the HAR-FunXL model is introduced. In Section 6.3 its estimation is discussed. The realized liquidity curves are expanded in terms of their functional principal components as in earlier Chapters. However, in case of the functional parameter, some alternatives to the FPC expansion are introduced and discussed. In particular, we exploit the Markovianity of the model to import methods based on penalized splines from the FDA literature. These expansions include explicit roughness penalties and are estimated in a mixed model framework. In an empirical application using XETRA data for Linde, Commerzbank, and MunichRe (in Section 6.4), we demonstrate that realized measures of liquidity contribute substantially to the explanation of realized volatility dynamics, while not being particularly useful to forecast RV in an out-of-sample exercise.

Finally, in Section 6.5, we show that the HAR-FunXL model can be viewed as a dynamic functional regression model. From an inferential point of view, it is a spe-

<sup>1</sup>To be more precise, they are Markovian of infinite order.

cial case of the more general framework of (functional) generalized additive models. We discuss some possible extensions, namely the case of modeling all features of the conditional distribution, and functional autoregressive models. Some concluding remarks close the Chapter.

## 6.2 The HAR-FunXL model

Let  $\exp(y_\ell)$  denote the daily realized variance of the price of a financial asset, i.e., an estimate of its integrated variance on day  $\ell$ . The realized variance can either be computed based on returns on the quoted mid-price or by using returns on the transaction price. We define the  $n$ -period logarithmic realized variance by

$$y_{\ell:(\ell+1-n)} = \frac{1}{n} \sum_{i=1}^n y_{\ell-i}$$

We capture serial dependence of  $y_\ell$  by employing the HAR model of Corsi (2009), which is a linear autoregressive model of the form

$$y_\ell = \alpha_0 + \alpha_d y_{\ell-1} + \alpha_w y_{(\ell-1):(\ell-5)} + \alpha_m y_{(\ell-1):(\ell-22)} + \varepsilon_\ell, \quad (6.1)$$

which can easily be shown to be a restricted AR(22) model with only 3 instead of 22 free autoregressive parameters,

$$y_\ell = \alpha_0 + \sum_{i=1}^{22} a_i y_{\ell-i} + \varepsilon_\ell, \quad (6.2)$$

where

$$\begin{aligned} a_1 &= \alpha_d + \alpha_w/5 + \alpha_m/22 \\ a_i &= \alpha_w/5 + \alpha_m/22, \quad i = 2, \dots, 5 \\ a_i &= \alpha_m/22, \quad i = 6, \dots, 22. \end{aligned} \quad (6.3)$$

In the standard case, which is the situation we confine to,  $\varepsilon_\ell$  is assumed to be Gaussian white noise, that is, homoskedastic and inducing a Gaussian marginal distribution to  $y_\ell$ . For possible extensions see for example Corsi et al. (2008) and Corsi et al. (2012). Though formally having short memory, the model has become the workhorse in the empirical RV literature as it is able to imitate the autocorrelation structure of fractional ARIMA-type long memory models quite well. In what follows, we use the notation

$$\begin{aligned} \mathbf{y}_\ell &= (1 \quad y_\ell \quad y_{\ell:(\ell-4)} \quad y_{\ell:(\ell-21)})', \\ \boldsymbol{\alpha} &= (\alpha_0 \quad \alpha_d \quad \alpha_w \quad \alpha_m)'. \end{aligned} \quad (6.4)$$

Now consider a  $C$ -dimensional “vector” of functional regressors,

$$\mathbf{x}_\ell(d) = (x_{1,\ell}(d) \cdots x_{C,\ell}(d))',$$

where the argument  $d$  is used to emphasize the functional nature of its entries. The subindex  $\ell$  does not need to be taken too literally as the object may contain different variables at the same lag, the same variable at different lags, or even a mixture of such ingredients. In our empirical application, we intend to use  $C = 6$  functional regressors, namely the three realized measures defined above for both market sides each, observed contemporaneously at  $\ell$  or lagged by one day,  $\ell - 1$ . Observations lie on an equidistant, rescaled grid of relative prices,  $d = 0, 1/(J - 1), 2/(J - 1), \dots, (J - 2)/(J - 1), 1$ , thus  $\mathbf{x}_\ell$  is a  $C \times J$  matrix in practice.

Furthermore, we introduce the “vector” of functional parameters,

$$\boldsymbol{\gamma}(d) = (\gamma_1(d) \cdots \gamma_C(d))',$$

which are coupled with the predictors in a functional linear fashion using the function  $g$  which maps the measures to the real line,

$$g(\boldsymbol{\gamma}(d), \mathbf{x}_\ell(d)) = \sum_{i=1}^C \int_0^1 \gamma_i(m) x_{i,\ell}(m) dm.$$

Using all these ingredients, we can finally define our model.

**Definition 6.1** (HAR-FunXL model). *The HAR model with functional exogeneous liquidity is given by*

$$y_t = \mathbf{y}'_{\ell-1} \boldsymbol{\alpha} + g(\boldsymbol{\gamma}(d), \mathbf{x}_\ell(d)) + \varepsilon_\ell, \quad (6.5)$$

where  $\varepsilon_\ell$  is Gaussian white noise.

## 6.3 Estimation

Exploiting the Gaussianity of innovations, and ignoring the unconditional part, the log-likelihood of the model can be written as

$$\begin{aligned} l &= \sum_{\ell=23}^L \log f(y_\ell | \mu_\ell = \mathbf{y}'_{\ell-1} \boldsymbol{\alpha} + g(\mathbf{x}_\ell(d), \boldsymbol{\gamma}(d)), \sigma_\ell^2 = \sigma_\varepsilon^2) \\ &= (23 - L) \left[ \log \sqrt{2\pi} + \log \sigma_\varepsilon^2 \right] - \frac{1}{2\sigma_\varepsilon^2} \sum_{\ell=23}^L y_\ell - \mathbf{y}'_{\ell-1} \boldsymbol{\alpha} - g(\boldsymbol{\gamma}(d), \mathbf{x}_\ell(d)), \end{aligned}$$

where  $f$  denotes the normal density. Apart from the exogeneous part, this is still the HAR model of Corsi (2009). In recent years, several aspects of HAR models have been generalized. Innovations may be modeled as being non-Gaussian or even conditionally heteroskedastic (Corsi et al., 2008), the standard choice for  $d, w, m$  may be sub-optimal, and the approximate long memory generated by using  $p = 22$  lags may alternatively

be created by a short-memory model with time-varying parameter (Chen et al., 2010). However, we are primarily interested in an alternative explanation of the dynamics due to the presence of liquidity effects, which are taken as exogeneous. As shown in Corsi et al. (2008), homoskedastic NIG or heteroskedastic NIG-GARCH innovations usually provide more realistic conditional densities than a Gaussian model. Being solely interested in the conditional mean, augmented by functional liquidity impact, we nevertheless stick to the homoskedastic Gaussian case, which may therefore be considered a QML approach.<sup>2</sup>

### 6.3.1 The functional part

Not for the first time, we mention that the main contribution of this thesis is the inclusion of functional liquidity information. In the present setting, we assume linearity,

$$g(\gamma(d), \mathbf{x}_\ell(d)) = \sum_{i=1}^C \int_0^1 \gamma_i(m) x_{i,\ell}(m) dm.$$

So far, we have always pursued a purely FPC-based estimation strategy, where the functional parameters are expanded in terms of basis functions estimated from the data. As a consequence, the truncation parameter  $K$  tunes both the complexity of the regressors and the complexity of the functional parameters. One might argue that these are different aspects longing for separation. To accomplish this, we introduce an alternative estimation strategy. The first strategy is still based on a functional principal component expansion of both liquidity and functional parameters. We term this strategy FPCR henceforth. The second strategy, being called PFR (for penalized functional regression) in the following, exploits the fact that the HAR model is Markovian. For estimation this means that all variables on the right-hand side can be directly observed, which facilitates the use of penalized splines to expand the functional coefficients.

#### Functional principal component regression (FPCR)

In the context of MEMs, we have expanded both the curves and the parameters in terms of the curves' functional principal components, that is in terms of data-driven basis functions. We use the same strategy here. For the  $i$ th predictor, truncating at  $K$  components, and otherwise using the usual notation, we write

$$x_{i,\ell}(d) = \mu_i(d) + \sum_{k=1}^K \tilde{\zeta}_{k,i,\ell} \phi_{k,i}(d) + v_{i,\ell}(d), \quad (6.6)$$

---

<sup>2</sup>Note that Corsi et al. (2008) by means of simulations find that such a misspecification leads to an inefficient statistical inference (provided the true data-generating process is HAR-NIG-GARCH), a possible drawback of Gaussian QML that we have to accept.

where  $v_{i,\ell}$  is the truncation error. The corresponding parameter expansion reads

$$\gamma_i(d) = \sum_{j=1}^K \gamma_{j,i} \phi_{j,i}(d) \quad (6.7)$$

where the  $\gamma_{j,i}$  (without argument  $d$ ),  $j = 1, \dots, K$ , on the right hand side are scalar parameters. This yields the  $K$ -FPC expansion

$$\begin{aligned} \int_0^1 \gamma_i(m) x_{i,\ell}(m) dm &\approx \sum_{j=1}^K \sum_{k=1}^K \tilde{\zeta}_{k,i,\ell} \gamma_{j,i} \int_0^1 \phi_{k,i}(m) \phi_{j,i}(m) dm \\ &= \sum_{k=1}^K \gamma_{k,i} \tilde{\zeta}_{k,i,\ell} \\ &= \tilde{\boldsymbol{\zeta}}_{i,\ell}' \boldsymbol{\gamma}_i \end{aligned} \quad (6.8)$$

where

$$\begin{aligned} \tilde{\boldsymbol{\zeta}}_{i,\ell} &= (\tilde{\zeta}_{1,i,\ell} \cdots \tilde{\zeta}_{K,i,\ell})' \\ \boldsymbol{\gamma}_i &= (\gamma_{1,i} \cdots \gamma_{K,i})' \end{aligned} \quad (6.9)$$

A crucial difference between the HAR-FunXL and the MEM-FunXL models is the Markovianity of the first. Its endogeneous part has an  $AR(22)$  structure, so that each conditional density in the log-likelihood can directly be evaluated at the regressors. As shown above, in contrast MEMs have an ARMA representation due to the latent component ( $\sigma_t$  for GARCH,  $\psi_i$  for ACD) and are therefore non-Markovian. In a MEM of order  $p, q$ , the first  $q$  pseudo-observations of the latent component needed for estimation have to be initialized. This is usually done by plugging in an estimate of the unconditional return variance (GARCH) or mean duration (ACD). However, the trajectory of pseudo-observations, the "regressor", depends on all parameters in the conditional variance (mean) equation which leads to both theoretical and practical problems, especially for nonparametric estimation. For a review of the respective literature see Linton and Yan (2011).

In contrast, apart from some aspects of the estimator's asymptotic properties, the estimation of Markovian models is not different from the estimation of regression models for *i.i.d.* data. For such scalar-on-function regression models, a powerful machinery of nonparametric techniques already exists, see Reiss et al. (2015) for a comprehensive survey, and Goldsmith and Scheipl (2014) for an extensive comparison of empirical performances in cross-sectional settings. In case of the purely FPC-based approach employed so far, only the number of components  $K$  is used to calibrate the model's complexity.<sup>3</sup> Alternative techniques use two leverage points: Firstly, the regressors may be expanded by other basis functions or, if observations are regular and

<sup>3</sup>Additionally, the covariance surface from which the eigenfunctions are estimated is smoothed to account for its functional nature. However, the effect of smoothing the covariance surface can be expected to be negligible compared to the sensitivity of results to the choice of  $K$ .

dense, even not be expanded at all. Secondly, and certainly more importantly in our setting, the functional parameter  $\gamma(d)$  may be forced to be smooth. Such an estimator is typically biased, but has lower variance than its unpenalized counterpart.

### Penalized functional regression (PFR)

To gauge the possible benefits of roughness penalization, we empirically investigate one such technique, penalized functional regression (PFR), which is due to Goldsmith et al. (2011). For ease of exposition, we consider the case of only one FunX component, dropping the subindex  $i$ . In PFR, the functional predictor is expanded in terms of its FPCs as before, but the number of FPCs,  $K$ , is not a crucial tuning parameter here because regularization concentrates on the expansion of  $\gamma(d)$ , which is achieved by using a different set of basis functions  $\psi_j(d)$ ,  $j = 1, \dots, K_\gamma$ ,

$$\gamma(d) = \sum_{j=1}^{K_\gamma} \gamma_j \psi_j(d),$$

whose choice does not depend on the data. To underline the difference, let us denote the truncation parameter of the FPC by  $K_x := K$  in the following. The functional linear predictor then becomes

$$\int_0^1 \gamma(m) x_\ell(m) dm \approx \sum_{j=1}^{K_\gamma} \sum_{k=1}^{K_x} \xi_{k,\ell} \gamma_j \int_0^1 \phi_k(m) \psi_j(m) dm.$$

However,  $\int_0^1 \phi_k(m) \psi_j(m) dm$  does no longer collapse to either zero or 1 as before, but will be nonzero in general. We collect all these integrals in the  $K_x \times K_\gamma$  matrix  $\mathbf{J}_{\phi\psi}$  whose  $(k, j)$ th element is given by  $\int_0^1 \phi_k(m) \psi_j(m) dm$ .

Goldsmith et al. (2011) use a truncated power spline basis for the functional parameter,

$$\gamma(d) = \gamma_1 + \gamma_2 d + \sum_{j=3}^{K_\gamma} \gamma_j (d - \kappa_j)_+,$$

with knots  $\{\kappa_j\}_{j=3}^{K_\gamma}$ . The roughness penalty to be imposed on  $\gamma(d)$  uses the idea of representing the basis coefficients as mixed effects in a linear regression setup (Ruppert et al., 2003). To this end, the coefficients of the nonlinear part of the basis expansion are assumed to come from a joint normal distribution with zero mean and homoskedastic as well as diagonal covariance structure,  $\{\gamma_j\}_{j=3}^{K_\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I})$ . The motivation for this assumption will become obvious shortly.

### 6.3.2 Inference for the HAR-FunXL model

We write the model for the pure FPC-based and the penalized case in a unified way. Denote the  $(L - 22) \times K$  matrix of empirical FPC scores by  $\hat{\Xi}$ , where the *hat* indi-

cates that this is not an observable regressor but one which is constructed in a first estimation step. The  $\ell$ th row of  $\hat{\Xi}$  is denoted by  $\hat{\xi}_\ell$ . The unknown parameter in the first estimation step is the contemporaneous liquidity covariance surface. We further denote the  $(L - 22) \times 3$  matrix of lagged HAR components by  $Y$ , and the  $(L - 22) \times 1$  vector of dependent log variances by  $y$ .

The components which PFR and FPCR have in common are then explicitly given by

$$\begin{aligned} y &= (y_{23} \cdots y_L)' \\ Y &= [y_{22} \cdots y_{L-1}]' \\ \hat{\Xi} &= [\hat{\xi}_{23} \cdots \hat{\xi}_L]' \\ \alpha &= (\alpha_0 \alpha_d \alpha_w \alpha_m)' \end{aligned} \quad (6.10)$$

To make the difference between the two approaches explicit, we observe that the PFR penalty matrix is given by

$$J_{\phi\psi} = \begin{bmatrix} \int_0^1 \phi_1(m)\psi_1(m)dm & \int_0^1 \phi_1(m)\psi_2(m)dm & \cdots & \int_0^1 \phi_1(m)\psi_{K_\gamma}(m)dm \\ \int_0^1 \phi_2(m)\psi_1(m)dm & \int_0^1 \phi_2(m)\psi_2(m)dm & \cdots & \int_0^1 \phi_2(m)\psi_{K_\gamma}(m)dm \\ \vdots & \vdots & \ddots & \vdots \\ \int_0^1 \phi_K(m)\psi_1(m)dm & \int_0^1 \phi_K(m)\psi_2(m)dm & \cdots & \int_0^1 \phi_K(m)\psi_{K_\gamma}(m)dm \end{bmatrix}. \quad (6.11)$$

Analogously to  $J_{\phi\psi}$ , we define the  $K_x \times K_\gamma$  matrix  $J_{\phi\phi}$  (which is quadratic as  $K_x = K_\gamma$ ) for the pure FPC case. Due to the orthonormality of eigenfunctions, we have

$$J_{\phi\phi} = \begin{bmatrix} \int_0^1 \phi_1(m)\phi_1(m)dm & \int_0^1 \phi_1(m)\phi_2(m)dm & \cdots & \int_0^1 \phi_1(m)\phi_{K_\gamma}(m)dm \\ \int_0^1 \phi_2(m)\phi_1(m)dm & \int_0^1 \phi_2(m)\phi_2(m)dm & \cdots & \int_0^1 \phi_2(m)\phi_{K_\gamma}(m)dm \\ \vdots & \vdots & \ddots & \vdots \\ \int_0^1 \phi_K(m)\phi_1(m)dm & \int_0^1 \phi_K(m)\phi_2(m)dm & \cdots & \int_0^1 \phi_K(m)\phi_{K_\gamma}(m)dm \end{bmatrix} = \mathbf{I}_{K_x \times K_x}. \quad (6.12)$$

The weight vector of the basis functions,

$$\gamma = (\gamma_1 \cdots \gamma_{K_\gamma})',$$

is typically chosen to be larger for PFR than for FPCR,  $K_\gamma > K_x$ , as in the PFR approach regularization is not accomplished through the choice of  $K_\gamma$  but by shrinking the elements of  $\gamma$ . In contrast, in FPCR the tuning parameter  $K = K_\gamma = K_x$  is crucial as it determines *both* the amount of liquidity features represented in the model and the shape of the functional parameter.

Using the parameter expansion, the HAR-FunXL model with one FunXL component can be rewritten as

$$\mathbf{y} = \mathbf{Y}\boldsymbol{\alpha} + \hat{\boldsymbol{\Xi}}\mathbf{J}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}),$$

where  $\mathbf{J}$  equals either  $\mathbf{J}_{\phi\phi}$  (pure FPCR) or  $\mathbf{J}_{\phi\psi}$  (PFR).

While in FPCR,  $\mathbf{y}$  is directly regressed on  $\hat{\boldsymbol{\Xi}}$ , and  $\boldsymbol{\gamma}$  is considered as an unknown but fixed parameter vector, in PFR  $\boldsymbol{\gamma}$  is assumed to come from a multivariate normal distribution with mean zero and diagonal covariance matrix,  $\boldsymbol{\gamma}_- = (\gamma_3 \cdots \gamma_{K_\gamma})' \sim N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I})$ . Assuming independence of innovations  $\boldsymbol{\varepsilon}$  and random coefficients  $\boldsymbol{\gamma}_-$ , and stacking them in one vector, their joint mean and covariance may be written (Ruppert et al., 2003),

$$E\left(\begin{bmatrix} \boldsymbol{\gamma}_- \\ \boldsymbol{\varepsilon} \end{bmatrix}\right) = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \text{Cov}\left(\begin{bmatrix} \boldsymbol{\gamma}_- \\ \boldsymbol{\varepsilon} \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{\Sigma}_\gamma & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\varepsilon \end{bmatrix} \quad (6.13)$$

The joint GLS estimator of fixed and random coefficients accounting for this covariance structure is therefore given by

$$\begin{bmatrix} \hat{\boldsymbol{\gamma}}_- \\ \hat{\boldsymbol{\varepsilon}} \end{bmatrix} = \left( \begin{bmatrix} \mathbf{Y} \\ \hat{\boldsymbol{\Xi}}\mathbf{J}_{\hat{\phi}\psi} \end{bmatrix} \boldsymbol{\Sigma}_\varepsilon^{-1} \begin{bmatrix} \mathbf{Y} \\ \hat{\boldsymbol{\Xi}}\mathbf{J}_{\hat{\phi}\psi} \end{bmatrix}' + \mathbf{B} \right)^{-1} \begin{bmatrix} \mathbf{Y} \\ \hat{\boldsymbol{\Xi}}\mathbf{J}_{\hat{\phi}\psi} \end{bmatrix} \boldsymbol{\Sigma}_\varepsilon^{-1} \mathbf{y}, \quad (6.14)$$

where  $\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\gamma^{-1} \end{bmatrix}$

Note that we have emphasized the fact that the eigenfunctions have been estimated in a first step by writing  $\mathbf{J}_{\hat{\phi}\psi}$ . If we assume homoskedastic and serially uncorrelated innovations, the estimator is of OLS type except for  $\mathbf{B}$  which contains the (unknown) parameters in  $\boldsymbol{\Sigma}_\gamma$ . This can be interpreted as a ridge penalty on  $\hat{\boldsymbol{\gamma}}_-$  which is the larger, the smaller  $\sigma_\gamma^2$  is. Unpenalized estimation therefore corresponds to  $\sigma_\gamma^2 = \infty$ . The variance parameters appearing in (6.14) are estimated via restricted maximum likelihood (REML), see Ruppert et al. (2003).

## 6.4 Empirical results

### 6.4.1 Explaining realized volatility

In our empirical analysis, we seek to answer the following questions based on daily realized measures:

- Are liquidity and volatility associated contemporaneously?
- Does past liquidity provide any information on future volatility?
- Which features of liquidity matter the most?
- Are liquidity effects time-varying?

In all cases, the term *liquidity* stands for all aspects of realized liquidity: its level, changes in level, and its intraday variation. In the previous Chapter, we have introduced three functional realized measures of liquidity (RML), namely:

1. Initial liquidity, that is, liquidity at the start of continuous trading on trading day  $\ell$ ,

$$x_{\ell}^{(s),IL}(d) = x_{0,\ell}^{(s)}(d).$$

2. Realized liquidity, that is, an estimate of the liquidity drift on trading day  $\ell$ ,

$$x_{\ell}^{(s),RL}(d) = \sum_{k=0}^d \left( x_{T,\ell}^{(s)}(k) - x_{0,\ell}^{(s)}(k) \right).$$

3. Realized variance of liquidity, that is, the sum of squared changes over trading day  $\ell$ ,

$$x_{\ell}^{(s),RVL}(d) = \sum_{k=0}^d \left( \sum_{i=1}^M \left( x_{t_i,\ell}(k) - x_{t_{i-1},\ell}(k) \right)^2 \right).$$

In our empirical analysis, we additionally consider

4. a second measure of liquidity variation,

$$x_{\ell}^{(s),RVLII}(d) = \sum_{i=1}^M \left( \sum_{k=0}^d x_{t_i,\ell}(k) - \sum_{k=0}^d x_{t_{i-1},\ell}(k) \right)^2,$$

while re-termining the first realized variation measure  $x_{\ell}^{(s),RVL_I}(d)$ . Note that we have developed our inferential theory for the original proposal,  $RVL_I$ , where variation is first aggregated over time for each observed relative price level individually, and then cumulated along the relative price. As  $RVL_I$  is positive for each  $d$ , this leads to strictly non-decreasing curves. The alternative measure,  $RVL_{II}$ , is directly computed from cumulative volume curves as modeled in Part I, i.e., cumulation along  $d$  comes first, then the squared differences of liquidity curves are computed. These curves are not necessarily non-decreasing with respect to  $d$ , and we do not provide an inferential theory. However, we conjecture that they behave similar to  $RVL_I$ .

We collect the RML for the two market sides in the “vector”

$$\mathbf{x}_{\ell} = \left( x_{\ell}^{(ask),IL} \ x_{\ell}^{(bid),IL} \ x_{\ell}^{(ask),RL} \ x_{\ell}^{(bid),RL} \ x_{\ell}^{(ask),RVL} \ x_{\ell}^{(bid),RVL} \right)'$$

Lightening notation a bit, the elements of  $\mathbf{x}_{\ell}$  are denoted simply by  $x_{i,\ell}(d)$ ,  $i = 1, \dots, 6$ . We construct the following two models, firstly the model with contemporaneous liq-

liquidity impact,

$$y_t = \mathbf{y}'_{\ell-1} \boldsymbol{\alpha} + \sum_{i=1}^C \int_0^1 \gamma_i(m) x_{i,\ell}(m) dm + \varepsilon_\ell, \quad (6.15)$$

and, secondly, the model with lagged liquidity impact,

$$y_t = \mathbf{y}'_{\ell-1} \boldsymbol{\alpha} + \sum_{i=1}^C \int_0^1 \gamma_i(m) x_{i,\ell-1}(m) dm + \varepsilon_\ell, \quad (6.16)$$

where  $\varepsilon_\ell$  is Gaussian white noise. Of course, one could easily think of many other specifications, especially including further lagged liquidity curves, but we leave such analyses for future investigation.

We split the potential predictors into three sets, (i) the level components ( $i \in \{1, 2, 3, 4\}$ ), (ii) the variance components ( $i \in \{5, 6\}$ ), and (iii) all components ( $i \in \{1, \dots, 6\}$ ). For all HAR models considered, the “minimum requirement” of a good fit, i.e. the decorrelation of residuals, was met according to Ljung-Box tests.<sup>4</sup> Tables 6.1 and 6.2 present the estimation results for the three stocks with contemporaneous and lagged liquidity impact, respectively. Both the level and realized variation of liquidity explain realized price variation to a considerable extent. However, we again emphasize that the AIC’s and BIC’s should be interpreted with some caution as they are naive in the sense that the ascribed dimension of the model used in the penalty term treats the constructed regressors as observed ones.

When fitting, we use either  $RVL_I$  or  $RVL_{II}$  (not both in one model) to measure liquidity variation, finding that the latter better explains price variation. The importance of liquidity is somewhat higher if the price is measured in terms of bid and ask quotes than if transaction prices are used. This may be attributed to the fact that bid and ask quotes are quantities implied by the limit order schedules (though not measured by the curves themselves). The quotes tend to exhibit several spikes which, at least partly, may be explained by liquidity variation. In contrast, transaction prices do not exhibit such spikes, and their association with liquidity is weaker, but still present. It appears that both level and variance components explain RV to some extent. However, quote variation is clearly best explained by liquidity *variation* rather than by its level, while the results for the variation of the transaction prices are rather mixed. Not surprisingly, we find that lagged liquidity impact is not as strongly associated with price variation as contemporaneous liquidity impact. Figures 6.1 and 6.2 reveal to what extent liquidity is able to alternatively explain RV’s autocorrelation. Although the results for the three stocks analyzed, especially the composition of short-to-long-term dependencies as reflected by the HAR parameters  $\hat{\alpha}_d, \hat{\alpha}_w, \hat{\alpha}_m$ , are quite mixed, there appears a slight negative relationship between autocorrelation and  $K$  which may be interpreted as a liquidity explanation of serial dependence.

<sup>4</sup>However, when looking at the SACF for many lags, the short memory character of the HAR model becomes finally apparent, as sample autocorrelations exceed those generated by the model.

The results commented on so far are based on FPCR. In comparison, we found that the overall performance of PFR was quite poor. The comparison has been drawn on the grounds of information criteria and residuals, respectively. As PFR is basically an additive regression model fitted with mixed model software, we can compute AIC and BIC which account for the dimension of  $J_{\phi\psi}$  and  $\gamma$ , and reward the presence of a roughness penalty.<sup>5</sup> Still, the dimension of the FPC-based part is intrinsically infinite. As in FPCR, this intrinsic dimension of the regressors, the FPC scores, is neglected.

However, it is possible to compare the fits provided by FPCR and PFR by means of their residuals or based on  $R^2$ , which relates the residual sum of squares to overall RV variation. For both criteria we find that PFR performs rather poorly as it is even outperformed by the pure HAR model (which it nests!) in all cases. However, such poor performance may not come as a total surprise as it is also documented in Goldsmith and Scheipl (2014), where it is applied in several cross-sectional regression settings. The same results indicate that other regularization methods may be superior. We leave this question for further research.

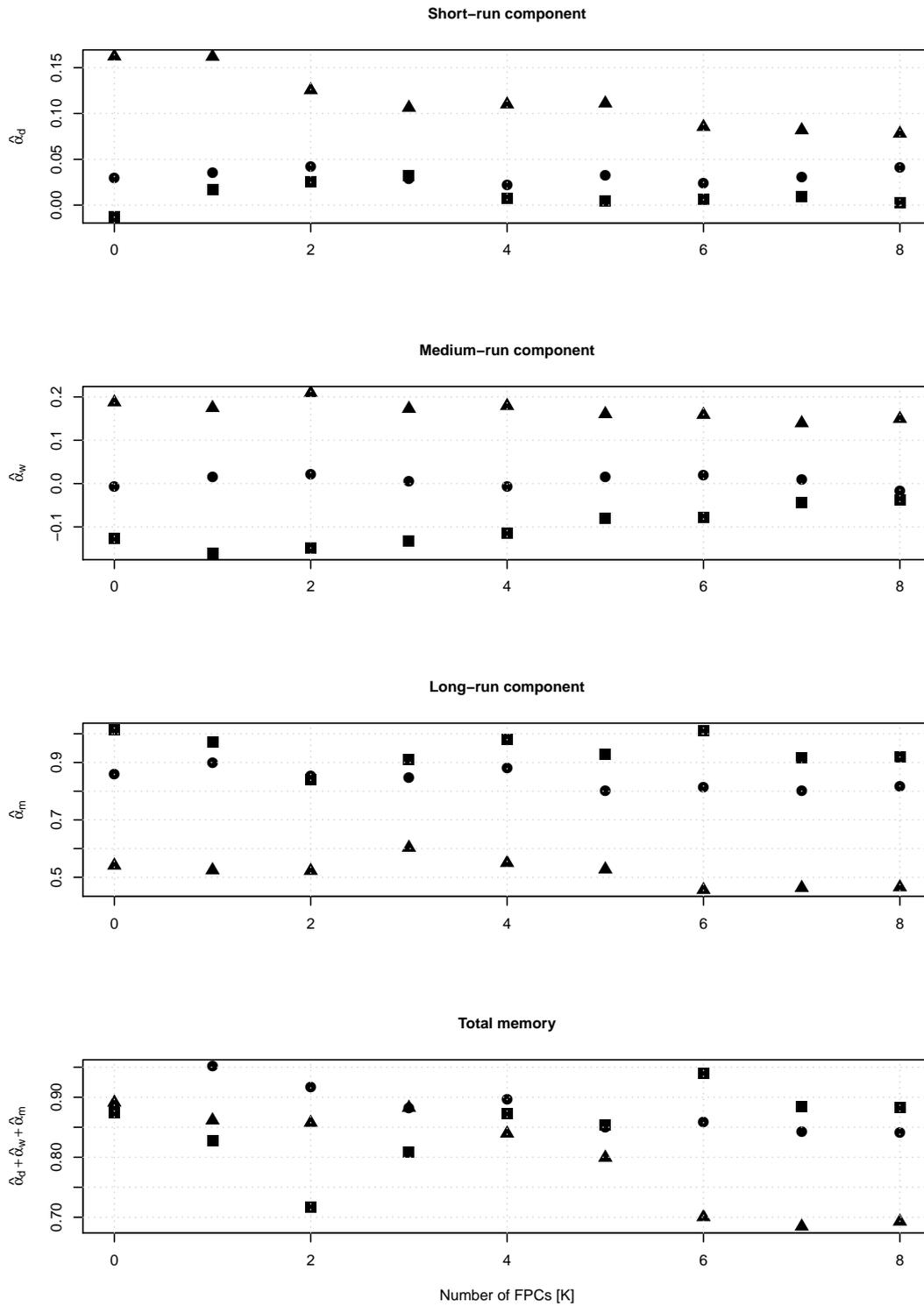
### Liquidity impact

The transitory effects of liquidity can be measured in terms of liquidity impact. In close analogy to the MEM models introduced in Part I, we write the model as

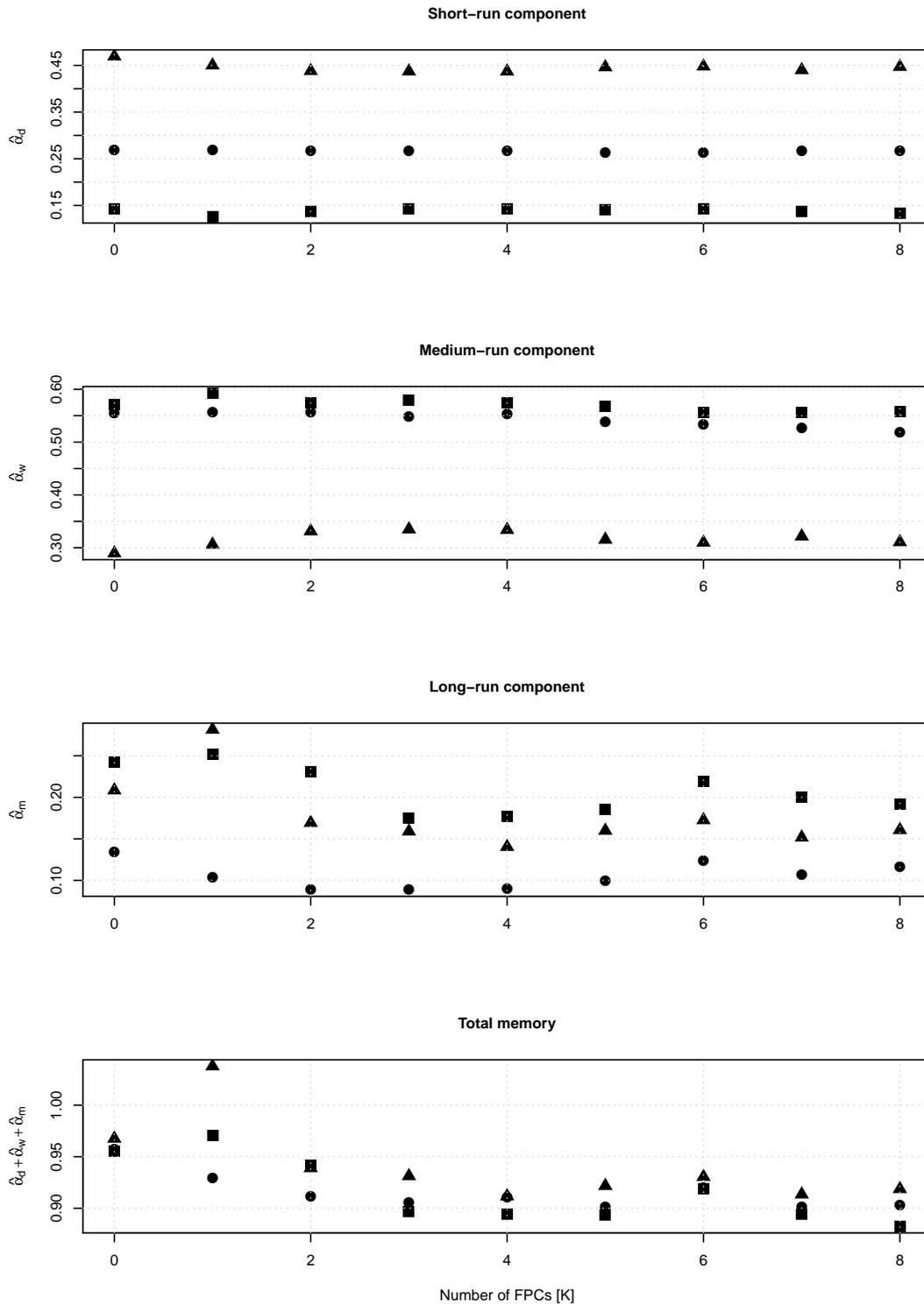
$$y_\ell = \underbrace{\alpha_0 + \alpha_d \mathbf{y}_{(\ell-1):(\ell-d)} + \alpha_w \mathbf{y}_{(\ell-1):(\ell-w)} + \alpha_m \mathbf{y}_{(\ell-1):(\ell-m)}}_{=:G_\ell} + \underbrace{\sum_{i=1}^C \int_0^1 \gamma_c(m) x_{i,\ell}(m) dm}_{=:LI_\ell} + \varepsilon_\ell, \quad (6.17)$$

where  $G_\ell$  is the *endogeneous* part of the model and  $LI_\ell$  is the exogeneous *liquidity impact*, consisting of several measures  $x_{i,\ell}(d)$  which are functions of the relative price. In terms of notation, the exogeneous part is treated as contemporaneous, but it may also consist of lagged quantities. The setting (i) *level impact* means that the opening liquidity  $x_{0,\ell}(d)$  and realized drift  $x_{T,\ell}(d) - x_{0,\ell}(d)$ , for each market side, are used as predictors. Setting (ii) *variance impact* means that  $RVL_\ell$  for the bid and ask side is used. In the following, we consider setting (iii), where both effects are present, which makes  $C = 6$  (three measures, each for bid and ask). Plugging in the FPC-based estimation results, each of the functional predictors is approximated using  $K$  FPC scores by the representation

<sup>5</sup>We used the `pfr` function in the R package `refund` (Crainiceanu, Reiss, Goldsmith, Huang, Huo, and Scheipl, Crainiceanu et al.) for this task, which wraps `gam` (authored by Simon Wood) and `lme4` (Douglas Bates), the latter two providing the state-of-the-art methods in semiparametric regression and mixed models, respectively.



**Figure 6.1:** HAR parameters for Linde (■), MunichRe (●), Commerzbank (▲), for models with  $K = 0, \dots, 8$  realized liquidity components (realized level and variance). Realized price variance is computed in terms of the mid-price. The results are quite heterogeneous, but there is a slight downward trend with respect to  $K$ , i.e., liquidity explains some of realized volatility’s “memory”.



**Figure 6.2:** HAR parameters for Linde (■), MunichRe (●), Commerzbank (▲), for models with  $K = 0, \dots, 8$  realized liquidity components (realized level and variance). Realized price variance is computed in terms of transaction prices. The results are quite heterogeneous, but there is a slight downward trend with respect to  $K$ , i.e., liquidity explains some of realized volatility’s “memory”.

$$\int_0^1 \hat{\gamma}_i(m) x_{i,\ell}(m) dm = \sum_{k=1}^K \hat{\gamma}_{i,k} \hat{\xi}_{i,k,\ell}, \quad (6.18)$$

where  $\hat{\xi}_{i,k,\ell}$  are FPC scores obtained via numerical integration as explained in Chapter 2. Thus, the estimated liquidity impact can be written

$$\hat{LI}_\ell = \sum_{i=1}^C \sum_{k=1}^K \hat{\gamma}_{i,k} \hat{\xi}_{i,k,\ell}, \quad (6.19)$$

and, exploiting the fact that we have an additive structure, we additionally define the *relative liquidity impact*,

$$RLI_\ell = \frac{LI_t}{|G_\ell| + |LI_\ell|}, \quad (6.20)$$

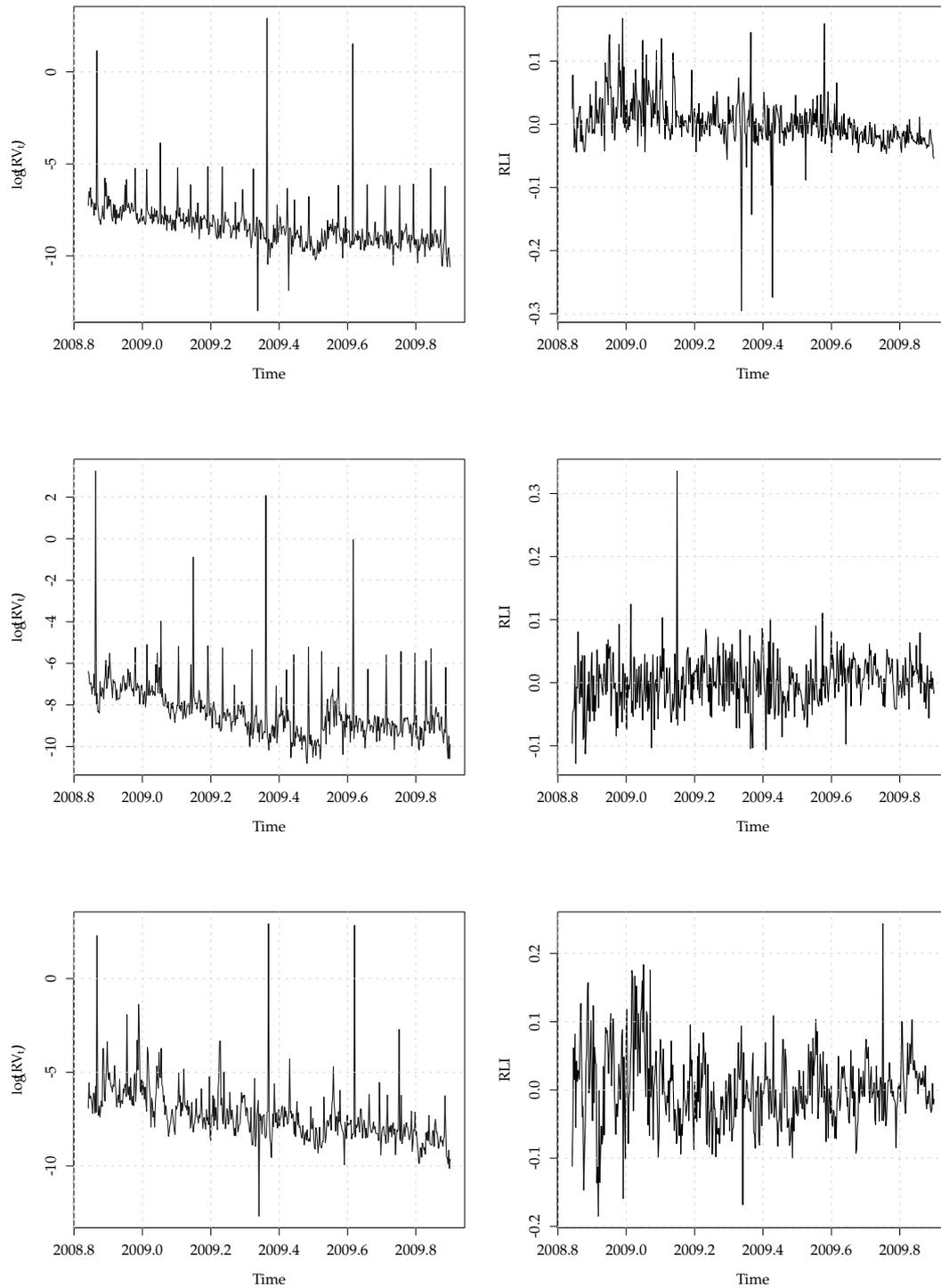
where  $-1 < RLI_\ell < 1$ . When  $RLI_\ell \approx 0$ , serial dependence dominates the dynamics of realized variance. However, when  $RLI_\ell$  is distinctly positive (negative), liquidity increases (decreases) the volatility of the price. Moreover, as the components of  $RLI_\ell$  have an additive structure, it can be further decomposed into its bid/ask and level/variance parts, respectively.

Figures 6.3 and 6.4 present the results for the three stocks. Despite their less-than-optimal fit according to the naive AIC and BIC criteria reported in Tables 6.1 and 6.2, we use models with only  $K = 2$  FPCs in all cases as they already provide the largest log-likelihood improvement as compared to the pure HAR model. We find a substantial liquidity impact for mid-price variation, which exhibits many spikes. The liquidity impact on transaction price variation is much lower, but still considerable. Moreover, we found that LI heavily depends on the model choice, that is on  $K$ , so that model choice is a crucial task all the more.

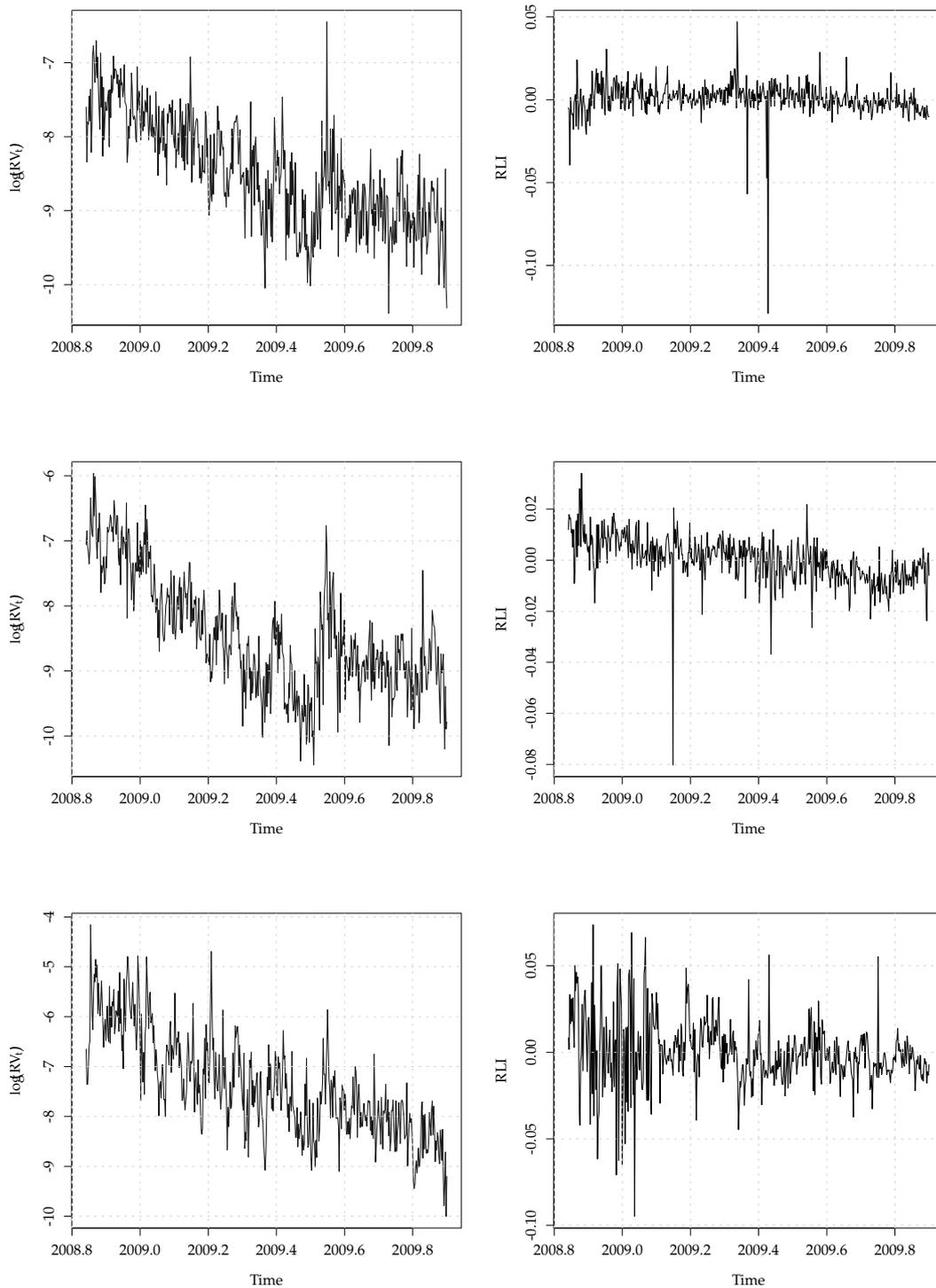
Including further lagged FunX components might further improve the fit and is straightforward to implement. Moreover, as contemporaneous liquidity explains RV better than lagged liquidity does, a one-step-ahead forecast obtained by one of the methods from Chapter 2 may be used to augment the model if forecasting is of primary interest.

### 6.4.2 Forecasting realized volatility

We fitted the models with lagged liquidity impact to the first 300 observations (roughly 15 months) to investigate the out-of-sample forecast performance, measured in terms of the forecast RMSE. This means that we are interested in forecasts of the conditional mean, not of the complete conditional density. As has been shown previously in the literature, alternative innovation specifications will certainly provide much more realistic conditional densities, but forecasts of the conditional mean will typically not



**Figure 6.3:** Log realized mid-price variance (left panel) and relative liquidity impact (right panel) for Linde (top), MunichRe (center), and Commerzbank (bottom) and  $K = 2$ . Realized variance trajectories exhibit several spikes which are usually due to few extreme price changes during a trading day. However, experimenting with the sampling grid we find that extreme quotes occur quite often during a trading day, indicating that the phenomenon is not spurious. Liquidity impact, apart from some spikes, typically ranges between  $\pm 0.2$ . Some variance spikes can be directly linked to liquidity impact by pure visual inspection already.



**Figure 6.4:** Log realized transaction price variance (left panel) and relative liquidity impact (right panel) for Linde (top), MunichRe (center), and Commerzbank (bottom) and  $K = 2$ . In contrast to the mid-price, realized variance trajectories for transaction prices do not exhibit many spikes. Consequently, the relative liquidity impact on the transaction prices' variation is much smaller than the impact on the variation of the mid-price, but still amounts to several percent.

differ much (Corsi et al., 2008). The results are shown in Tables 6.7, 6.8, and 6.9. Apart from few exceptions, the pure HAR model provides better forecasts than its liquidity-augmented competitors. All HAR models outperform the naive no-change random walk forecast. In terms of the Diebold-Mariano test, all HAR models appear to be equally good, and superior to the random walk. The PFR version of the model has not been investigated as, due to its poor in-sample performance, we expect it to be inferior to FPCR.

An alternative to the model with lagged liquidity would be one that explicitly models liquidity dynamics as in Section 2.5, either based on functional autoregression or based on the FPCs' dynamics. One could also think of a full joint model with volatility and liquidity interacting. Some comments on this prospect follow in Section 6.5. However, multivariate models are typically outperformed by univariate ones when it comes to forecasts. Therefore, it is questionable whether the construction of such a full model will be beneficial in this respect.

## 6.5 Generalizations

We briefly discuss some possible extensions of the functional time series approach with regularizing penalties. Some comments on the possibility of constructing multivariate models follow.

### 6.5.1 Families of conditional distributions

For (quasi-) likelihood inference on the parameters of the HAR-FunXL model, we have used a specific conditional distribution of  $y_\ell$ ,

$$y_\ell | \mathbf{y}_{\ell-1}, \gamma(d), \mathbf{x}_\ell(d) \sim N \left( \mathbf{y}'_{\ell-1} \boldsymbol{\alpha} + g(\gamma(d), \mathbf{x}_\ell(d)), \sigma_\varepsilon^2 \right),$$

in which the conditioning set includes only a finite number of lagged, observable quantities. However, the strategies for estimating the functional parameters in  $g(\cdot)$  explained before are not limited to the Gaussian case. More generally, from an inferential point of view the model may be viewed as a generalized regression model (GRM). (Univariate) GRMs consist of three ingredients.

1. A **distributional assumption**. (Univariate) GRMs assume that the conditional distribution is a univariate exponential family distribution and focus on the *mean* of this distribution.
2. A real-valued **predictor**, i.e., a (possibly nonlinear) combination of the predictor variables of interest. In our case, this is a mixture of an endogeneous (the lagged  $y_\ell$ ) and an exogeneous part ( $g(\gamma(d), \mathbf{x}_\ell(d))$ ). If the predictor is linear, the model is a generalized *linear* model (GLM). If the predictor variables are transformed in a data-driven way, and this transformation is estimated nonparametrically, but still additive, the model is a generalized *additive* model (GAM).

3. A **transformation function** or link function  $h(\cdot)$ , i.e., a continuous function which maps the predictor to the mean of the conditional distribution.

This framework encompasses, among others, the following models as special cases.

- The HAR-FunXL model. The conditional distribution is normal, the predictor maps scalar lagged realized measures and functional realized measures to  $\mathbb{R}$ , and the link function is the identity link.
- The ACD-FunXL model with  $\beta = 0$ . The conditional distribution is exponential (or Gamma), the predictor is as before, and the link function is the logarithmic link.

However, the framework does not encompass non-Markovian models like GARCH or (unrestricted) ACD because the predictor variables must be observable, and there must be a finite number of them. Even the ARCH-FunXL model, being Markovian in contrast to GARCH, is not included for a second reason: It does not feature the mean, but the dispersion of the conditional distribution as response variable.

### 6.5.2 Conditional heteroskedasticity and higher-order moments

Building on the GAMLSS model of Rigby and Stasinopoulos (2005), in Brockhaus et al. (2015) we propose a class of functional regression models in which more features of the conditional distribution than only the mean are allowed to depend on predictor variables. While leaving the framework of GRMs, these models still do not allow for latent components on the right hand side the way they appear in GARCH and ACD models. For instance, in terms of the HAR model this means that the Brockhaus et al. (2015) framework encompasses the HAR-FunXL with ARCH-FunXL innovations, but not with GARCH-FunXL innovations.

The framework is also able to capture time-varying conditional skewness and kurtosis. The latter can for instance be introduced using the degrees-of-freedom parameter of a  $t$ -ARCH model. However, while being very flexible, a major drawback of this approach is that estimation is only accomplished by means of boosting techniques. The statistical properties of boosting are not well understood in many respects. On the other hand, the approach can easily be extended to functional autoregressive settings like those discussed in Chapter 2, which is quite attractive.

### 6.5.3 Multivariate models

The term *multivariate* can mean several things in the HAR-FunXL context, namely

- A model for several,  $p$  say, realized price volatility series with exogeneous liquidity.
- A joint model for one or  $p$  price volatility series along with the corresponding liquidity measures.

The first extension is straightforward to implement by means of a VHAR model, even if there are cross-effects between, say, asset  $i$ 's liquidity on asset  $j$ 's price (but not vice versa). The second extension is much more involved. If we approach the problem by means of semiparametric factors as in Section 2.5, we could simply stack FPC scores of curves and scalar realized measures in a large vector and, again, use standard V(H)AR methodology. This strategy might even enjoy some success in terms of forecasting. However, describing dynamic dependencies between finite-dimensional representations of curves and scalars, such a model would be hardly interpretable.

## 6.6 Conclusion

In this Chapter, we have constructed a model to investigate the impact of realized measures of liquidity on realized price variation, extending the HAR model of Corsi (2009) by including FunXL terms. While the FunXL part of the model is very similar to the FunXL part in MEM-FunXL models, the construction of liquidity measures is fundamentally different as they are *aggregates* over intervals of time. Empirically, we find that there is a considerable contemporaneous liquidity impact on price variation, while the impact of lagged liquidity on present volatility is rather moderate. Therefore, at least in the simple cases considered here, realized liquidity does not help to predict realized price variation at a daily frequency.

FunXL measure	number of FPCs								
	0	1	2	3	4	5	6	7	8
<b>Linde</b>									
AIC									
(i) level	1551	1513	1505	1497	1482	1475	1471	1468	1460
(ii) variance	1551	1515	1514	1510	1506	1503	1501	1495	1495
(iii) all	1551	1499	1487	1468	1449	1437	1427	1419	1405
BIC									
(i) level	1572	1523	1515	1508	1492	1485	1481	1478	1471
(ii) variance	1572	1526	1525	1520	1517	1514	1511	1506	1506
(iii) all	1572	1509	1497	1478	1460	1447	1438	1429	1416
<b>MunichRe</b>									
AIC									
(i) level	1602	1580	1569	1552	1549	1543	1541	1535	1531
(ii) variance	1602	1572	1570	1569	1567	1556	1550	1548	1545
(iii) all	1602	1562	1544	1525	1521	1500	1495	1490	1486
BIC									
(i) level	1623	1591	1579	1562	1560	1554	1552	1546	1542
(ii) variance	1623	1583	1580	1580	1578	1566	1561	1558	1555
(iii) all	1623	1573	1554	1535	1531	1510	1505	1500	1497
<b>Commerzbank</b>									
AIC									
(i) level	1642	1623	1616	1606	1602	1600	1592	1589	1587
(ii) variance	1642	1633	1604	1600	1599	1573	1561	1561	1559
(iii) all	1642	1618	1574	1551	1545	1531	1512	1509	1506
1142									
BIC									
(i) level	1663	1633	1627	1616	1612	1611	1602	1599	1598
(ii) variance	1663	1644	1614	1610	1610	1584	1572	1571	1569
(iii) all	1663	1629	1584	1561	1555	1542	1523	1520	1517

**Table 6.1:** AIC and BIC of the in-sample fits for the HAR-FunXL model with contemporaneous realized liquidity impact on RV computed based on the mid-price. The model with  $K = 0$  FPCs is the pure HAR model. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRV_{L,t}$ , and (iii) level and variance are used as functional predictors. Both level and variance components tend to improve the fit to a considerable extent in terms of both criteria.

FunXL measure	number of FPCs								
	0	1	2	3	4	5	6	7	8
<b>Linde</b>									
AIC									
(i) level	1549	1533	1517	1514	1511	1503	1502	1487	1478
(ii) variance	1549	1541	1539	1537	1534	1530	1529	1520	1518
(iii) all	1549	1533	1516	1510	1507	1495	1491	1470	1461
BIC									
(i) level	1570	1544	1527	1524	1522	1514	1512	1497	1489
(ii) variance	1570	1551	1549	1548	1544	1540	1539	1530	1529
(iii) all	1570	1544	1526	1521	1517	1505	1501	1480	1471
<b>MunichRe</b>									
AIC									
(i) level	1600	1590	1587	1583	1579	1578	1562	1550	1544
(ii) variance	1600	1588	1586	1584	1584	1582	1581	1578	1569
(iii) all	1600	1585	1580	1574	1568	1567	1552	1536	1522
BIC									
(i) level	1621	1601	1597	1593	1589	1589	1573	1561	1554
(ii) variance	1621	1598	1596	1595	1594	1593	1592	1589	1580
(iii) all	1621	1595	1590	1585	1579	1578	1562	1546	1532
<b>Commerzbank</b>									
AIC									
(i) level	1639	1623	1623	1617	1616	1610	1609	1607	1602
(ii) variance	1639	1622	1617	1617	1616	1613	1612	1608	1607
(iii) all	1639	1617	1610	1606	1605	1597	1596	1590	1584
BIC									
(i) level	1660	1634	1633	1627	1626	1620	1619	1617	1613
(ii) variance	1660	1632	1627	1627	1626	1623	1622	1619	1618
(iii) all	1660	1628	1620	1617	1616	1608	1606	1600	1595

**Table 6.2:** AIC and BIC of the in-sample fits for the HAR-FunXL model with lagged realized liquidity impact on RV computed based on the mid-price. The model with  $K = 0$  FPCs is the pure HAR model. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_I$ , and (iii) level and variance are used as functional predictors. Both level and variance components tend to improve the fit considerably in terms of both criteria. However, the improvement is smaller than for contemporaneous liquidity.

FunXL measure	number of FPCs								
	0	1	2	3	4	5	6	7	8
<b>Linde</b>									
AIC									
(i) level	556	545	543	538	535	532	526	525	515
(ii) variance	556	532	532	530	528	528	527	518	516
(iii) all	556	530	525	519	517	513	503	496	481
BIC									
(i) level	577	556	553	548	545	543	536	535	526
(ii) variance	577	542	542	540	539	539	537	529	527
(iii) all	577	541	535	530	527	523	513	506	492
<b>MunichRe</b>									
AIC									
(i) level	583	573	568	568	564	554	551	547	546
(ii) variance	583	572	570	570	570	569	566	561	560
(iii) all	583	569	563	562	557	544	539	531	529
BIC									
(i) level	604	583	579	578	574	564	561	558	556
(ii) variance	604	582	581	581	581	579	576	572	571
(iii) all	604	579	574	573	568	555	550	541	539
<b>Commerzbank</b>									
AIC									
(i) level	787	770	767	762	758	752	748	746	745
(ii) variance	787	779	766	764	763	763	761	755	754
(iii) all	787	766	747	744	740	732	726	721	720
BIC									
(i) level	808	781	778	773	768	762	758	756	756
(ii) variance	808	790	776	775	774	773	771	765	765
(iii) all	808	777	757	754	750	742	736	732	731

**Table 6.3:** AIC and BIC of the in-sample fits for the HAR-FunXL model with contemporaneous realized liquidity impact on RV computed based on the mid-price. The model with  $K = 0$  FPCs is the pure HAR model. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_t$ , and (iii) level and variance are used as functional predictors. Both level and variance components tend to improve the fit to a considerable extent in terms of both criteria.

FunXL measure	number of FPCs									
	0	1	2	3	4	5	6	7	8	
<b>Linde</b>										
AIC										
(i) level	556	544	538	538	533	528	518	517	515	
(ii) variance	556	548	547	546	545	545	539	538	535	
(iii) all	556	544	537	536	531	524	511	509	504	
BIC										
(i) level	577	555	549	548	544	538	528	528	526	
(ii) variance	577	558	558	557	555	555	549	549	545	
(iii) all	577	554	548	547	541	535	521	520	514	
71										
<b>MunichRe</b>										
AIC										
(i) level	583	574	570	565	563	562	560	554	550	
(ii) variance	583	572	568	568	565	564	561	558	557	
(iii) all	583	571	562	557	554	551	546	537	530	
BIC										
(i) level	604	584	580	576	574	572	571	564	560	
(ii) variance	604	582	579	578	576	575	572	569	567	
(iii) all	604	581	573	568	564	562	556	547	540	
<b>Commerzbank</b>										
AIC										
(i) level	787	772	766	763	761	755	747	746	744	
(ii) variance	787	777	773	770	766	763	764	762	761	
(iii) all	787	768	759	755	747	735	731	728	726	
BIC										
(i) level	808	782	777	774	771	765	757	756	755	
(ii) variance	808	787	784	781	777	774	774	773	771	
(iii) all	808	779	770	765	758	745	742	739	737	

**Table 6.4:** AIC and BIC of the in-sample fits for the HAR-FunXL model with lagged realized liquidity impact on RV computed based on the mid-price. The model with  $K = 0$  FPCs is the pure HAR model. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_I$ , and (iii) level and variance are used as functional predictors. Both level and variance components tend to improve the fit considerably in terms of both criteria. However, the improvement is smaller than for contemporaneous liquidity.

FunXL measure	number of FPCs								
	0	1	2	3	4	5	6	7	8
<b>Linde</b>									
AIC									
(i) level	1427	1401	1400	1382	1379	1374	1372	1368	1364
(ii) variance	1427	1354	1336	1333	1328	1328	1327	1322	1321
(iii) all	1427	1350	1326	1306	1293	1290	1287	1280	1273
BIC									
(i) level	1448	1411	1410	1393	1389	1385	1382	1378	1374
(ii) variance	1448	1364	1346	1343	1339	1338	1338	1333	1332
(iii) all	1448	1360	1337	1316	1304	1300	1297	1290	1284
<b>MunichRe</b>									
AIC									
(i) level	1520	1493	1478	1461	1459	1454	1453	1447	1443
(ii) variance	1520	1467	1431	1428	1425	1415	1414	1411	1404
(iii) all	1520	1458	1420	1399	1394	1382	1381	1377	1367
BIC									
(i) level	1541	1504	1489	1472	1469	1464	1463	1458	1453
(ii) variance	1541	1477	1441	1439	1435	1425	1424	1422	1415
(iii) all	1541	1468	1430	1409	1404	1393	1392	1388	1378
<b>Commerzbank</b>									
AIC									
(i) level	1360	1343	1340	1329	1320	1319	1311	1304	1302
(ii) variance	1360	1311	1274	1228	1220	1210	1209	1208	1206
(iii) all	1360	1304	1266	1196	1183	1167	1154	1148	1142
BIC									
(i) level	1381	1353	1351	1339	1331	1329	1322	1315	1313
(ii) variance	1381	1322	1285	1238	1230	1221	1219	1219	1217
(iii) all	1381	1314	1277	1206	1193	1178	1164	1159	1152

**Table 6.5:** AIC and BIC of the in-sample fits for the HAR-FunXL model with contemporaneous realized liquidity impact. The model with  $K = 0$  FPCs is the pure HAR model. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_{II}$ , and (iii) level and variance are used as functional predictors. Both level and variance components tend to improve the fit to a considerable extent in terms of both criteria.

FunXL measure	number of FPCs								
	0	1	2	3	4	5	6	7	8
<b>Linde</b>									
AIC									
(i) level	1425	1414	1394	1391	1391	1390	1389	1387	1381
(ii) variance	1425	1417	1411	1409	1406	1406	1403	1401	1399
(iii) all	1425	1414	1390	1386	1382	1381	1376	1372	1366
BIC									
(i) level	1446	1425	1405	1402	1402	1401	1399	1398	1392
(ii) variance	1446	1427	1421	1420	1416	1416	1413	1411	1409
(iii) all	1446	1425	1400	1396	1392	1392	1386	1382	1376
<b>MunichRe</b>									
AIC									
(i) level	1518	1507	1501	1497	1494	1494	1482	1470	1466
(ii) variance	1518	1508	1507	1505	1504	1504	1503	1500	1499
(iii) all	1518	1502	1495	1491	1488	1488	1475	1458	1452
BIC									
(i) level	1539	1518	1511	1508	1505	1504	1492	1481	1476
(ii) variance	1539	1518	1517	1515	1515	1514	1513	1510	1510
(iii) all	1539	1513	1505	1502	1499	1498	1485	1468	1463
<b>Commerzbank</b>									
AIC									
(i) level	1357	1341	1341	1336	1330	1327	1325	1322	1321
(ii) variance	1357	1335	1335	1332	1331	1330	1329	1327	1326
(iii) all	1357	1327	1319	1313	1307	1303	1300	1295	1293
BIC									
(i) level	1378	1352	1352	1346	1341	1337	1336	1333	1331
(ii) variance	1378	1346	1345	1342	1341	1341	1339	1337	1336
(iii) all	1378	1338	1329	1323	1317	1314	1310	1306	1304

**Table 6.6:** AIC and BIC of the in-sample fits for the HAR-FunXL model with lagged realized liquidity impact. The model with  $K = 0$  FPCs is the pure HAR model. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_{LL}$ , and (iii) level and variance are used as functional predictors. Both level and variance components tend to improve the fit considerably in terms of both criteria. However, the improvement is smaller than for contemporaneous liquidity.

FunXL measure	number of FPCs									naive
	0	1	2	3	4	5	6	7	8	
<b>Linde</b>										
(i) level	0.6269	0.6643	0.7180	0.7197	0.7181	0.7219	0.7160	0.7065	0.7009	0.7514
(ii) variance	0.6269	0.6325	0.6371	0.6727	0.7067	0.7307	0.7549	0.7681	0.7855	0.7514
(iii) all	0.6269	0.6697	0.7416	0.7785	0.7912	0.8120	0.8220	0.8348	0.8441	0.7514
<b>MunichRe</b>										
(i) level	0.6940	0.7144	0.7228	0.7155	0.7078	0.7066	0.6956	0.7001	0.7125	0.8586
(ii) variance	0.6940	0.6780	0.6837	0.6849	0.6874	0.6836	0.6747	0.6930	0.6950	0.8586
(iii) all	0.6940	0.7019	0.7171	0.7128	0.7074	0.6990	0.6836	0.7195	0.7414	0.8586
<b>Commerzbank</b>										
(i) level	0.6646	0.7314	0.7344	0.7289	0.7415	0.7583	0.7650	0.7630	0.7724	0.7470
(ii) variance	0.6646	0.7259	0.6709	0.6726	0.7036	0.7082	0.7316	0.7161	0.7260	0.7470
(iii) all	0.6646	0.7719	0.7742	0.7913	0.8137	0.8453	0.8663	0.8302	0.8478	0.7470

**Table 6.7:** RMSE of one-step-ahead out-of-sample forecasts for the HAR-FunXL model with lagged realized liquidity impact (expanding window scheme). Price RV is computed based on the mid-price. The model with  $K = 0$  FPCs is the pure HAR model. The naive forecast, the random walk forecast, is always outperformed by the pure HAR specification. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_I$ , and (iii) level and variance are used as functional predictors. In contrast to  $CRVL_{II}$ ,  $CRVL_I$  does not improve the RV forecast, nor do the other RML. As compared to the analysis represented in Table 6.9, one outlier (2009-09-15) has been excluded.

FunXL measure	number of FPCs									naive
	0	1	2	3	4	5	6	7	8	
	<b>Linde</b>									
(i) level	0.3527	0.3519	0.3551	0.3551	0.3556	0.3567	0.3558	0.3564	0.3566	0.4487
(ii) variance	0.3527	0.3530	0.3554	0.3575	0.3591	0.3670	0.3719	0.3719	0.3716	0.4487
(iii) all	0.3527	0.3520	0.3591	0.3612	0.3613	0.3748	0.3803	0.3821	0.3811	0.4487
	<b>MunichRe</b>									
(i) level	0.3726	0.3733	0.3729	0.3726	0.3726	0.3729	0.3729	0.3733	0.3731	0.4404
(ii) variance	0.3726	0.3727	0.3740	0.3742	0.3735	0.3744	0.3753	0.3732	0.3748	0.4404
(iii) all	0.3726	0.3734	0.3732	0.3727	0.3731	0.3746	0.3760	0.3736	0.3744	0.4404
	<b>Commerzbank</b>									
(i) level	0.3851	0.3861	0.3867	0.3886	0.3869	0.3880	0.3866	0.3873	0.3877	0.4405
(ii) variance	0.3851	0.3873	0.3858	0.3854	0.3880	0.3887	0.3882	0.3894	0.3908	0.4405
(iii) all	0.3851	0.3873	0.3916	0.3908	0.3894	0.3940	0.3908	0.3941	0.3978	0.4405

**Table 6.8:** RMSE of one-step-ahead out-of-sample forecasts for the HAR-FunXL model with lagged realized liquidity impact (expanding window scheme). Price RV is computed based on transaction prices. The model with  $K = 0$  FPCs is the pure HAR model. The naive forecast, the random walk forecast, is always outperformed by the pure HAR specification. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_I$ , and (iii) level and variance are used as functional predictors. Lagged RML do not improve the forecast of price volatility.

FunXL measure	number of FPCs									naive
	0	1	2	3	4	5	6	7	8	
<b>Linde</b>										
(i) level	0.6111	0.6280	0.6615	0.6683	0.6677	0.6686	0.6636	0.6682	0.6708	0.7514
(ii) variance	0.6111	0.6131	0.6252	0.6277	0.6577	0.6712	0.6774	0.6826	0.6959	0.7514
(iii) all	0.6111	0.6294	0.6712	0.6750	0.7151	0.7290	0.7352	0.7449	0.7543	0.7514
<b>MunichRe</b>										
(i) level	0.6791	0.7114	0.7196	0.7153	0.7078	0.7061	0.6978	0.7029	0.7097	0.8617
(ii) variance	0.6791	0.6711	0.6730	0.6692	0.6746	0.6749	0.6745	0.6755	0.6757	0.8617
(iii) all	0.6791	0.7121	0.7275	0.7182	0.7139	0.7111	0.7039	0.7293	0.7395	0.8617
<b>Commerzbank</b>										
(i) level	0.5787	0.6132	0.6170	0.6196	0.6421	0.6478	0.6571	0.6578	0.6611	0.6734
(ii) variance	0.5787	0.5852	0.5798	0.5734	0.5718	0.5713	0.5713	0.5752	0.5730	0.6734
(iii) all	0.5787	0.6118	0.6074	0.6016	0.6288	0.6320	0.6574	0.6563	0.6519	0.6734

**Table 6.9:** RMSE of one-step-ahead out-of-sample forecasts for the HAR-FunXL model with lagged realized liquidity impact (expanding window scheme). Price RV is computed based on the mid-price. The model with  $K = 0$  FPCs is the pure HAR model. The naive forecast, the random walk forecast, is always outperformed by the pure HAR specification. For each RV series, (i) liquidity's level components (snapshot at the beginning of trading plus drift estimate), (ii) its log realized variance,  $CRVL_{II}$ , and (iii) level and variance are used as functional predictors. Including realized liquidity variance,  $CRVL_{II}$ , as predictor improves the RV forecast slightly for MunichRe and Commerzbank, but not for Linde. The liquidity level does not help to predict RV in any of the cases.

## Conclusions and perspectives

In this thesis, the concept of *functional exogeneous liquidity* (FunXL) is introduced which makes use of the full information provided by limit order book data in order to understand and forecast financial risk.

In Part I, logarithmic versions of the GARCH (Bollerslev, 1986) and ACD (Engle and Russell, 1998) models are augmented by FunXL components. In Part II, in analogy to realized volatility, *realized measures of liquidity* are designed, and their use for understanding realized variation of prices is investigated in the celebrated yet simple framework of heterogeneous autoregressive models (Corsi, 2009). The FunXL part, capturing the additional information provided by the limit order book, is in all cases viewed as a weakly stationary functional time series, allowing for (i) the adoption of dimension reduction techniques for functional time series (Hörmann and Kokoszka, 2010), and (ii) the adaptation of ideas from the field of functional regression analysis (Ramsay and Silverman, 2005), which in particular provide means of mapping random functional observations to scalar parameters of the conditional distribution of financial returns or durations.

The usefulness of the newly introduced models is investigated in extensive applications to limit order book data from the XETRA system of Deutsche Börse. Namely, the GARCH-FunXL model is used to model financial returns for three instruments over 26 months, sampled at a frequency of 20 minutes, the ACD-FunXL model is applied to one month of price durations data, and the realized measures fed into the HAR-FunXL model, computed at daily frequency, are based on 26 months of high-frequency trading data. While the FunXL parts of multiplicative error models are exclusively expanded in terms of functional principal components, the Markovian HAR-FunXL model allows to import alternative techniques originally developed for functional regression models in cross-sectional or longitudinal contexts.<sup>1</sup>

Apart from the different models introduced, the major difference between Parts I and II is that in Part I, LOB *snapshots* are modeled, while in Part II, snapshots are *aggregated* over time. Summarizing, we find empirically that the impact of LOB-implied liquidity on the price process is very pronounced at high frequency, whereas its suc-

---

<sup>1</sup>Note that the meaning of the word *longitudinal* differs somewhat between time series econometrics and (bio-) statistics.

cess in explaining realized price variation is rather modest. However, we are (with the notable exception of Engle et al. (2012)) to the best of our knowledge the first to measure realized liquidity variation, and the first to measure *functional* realized liquidity variation. We expect to achieve an improvement regarding this measurement issue in future work. For instance, we observe phenomena similar to those that are termed *microstructure noise* in the RV literature. Such phenomena will certainly be investigated in future research, and, possibly, sophisticatedly be accounted for.

Recalling the liquidity definition of Kyle (1985), provided at the very beginning of Chapter 2, a good liquidity measure should, besides other things, inform about *resiliency*, i.e., the speed at and extent to which the LOB recovers to its former shape after large trades. This dynamic aspect of liquidity is hard to capture, and our liquidity measure of choice, cumulative volume, does not capture it — at least not directly. A possible operationalization, to be addressed in future research, might approach the problem via the *persistence* of liquidity curves. From an economic perspective, the ideas put forward in this thesis may eventually also be of use when it comes to better judging the pros and cons of high-frequency trading (HFT), which has recently been criticized from both academic (Budish et al., 2013) and non-academic (Lewis, 2014) directions.<sup>2</sup> However, as LOB data usually do not reveal any information on individual behavior — our data do not include IDs of market participants —, such inferences would be highly speculative. The heuristic identification of algo strategies from anonymous data might nevertheless be a feasible alternative (Hasbrouck and Saar, 2013).

Given the success of the FunXL approach in modeling high-frequency conditional variance dynamics, an obvious next step for future research is the construction of multivariate models. In the case of MEMs, depending on the intended application, such models could be either vector MEMs or multivariate (marked) point process models with direct specification of the conditional intensity function (Hawkes, 1971; Bowsher, 2007; Bauwens and Hautsch, 2009). Alternative specifications of the conditional variance (or duration) equation allowing for general forms of liquidity impact, for instance as in Amado and Teräsvirta (2013), could also be of interest. A further desirable goal is to endogenize liquidity curves, that is, to construct models for the joint dynamics of liquidity curves of several instruments, or for prices and liquidity. However, even in the Markovian case discussed in Part II, there are some major obstacles to overcome. One possible, yet unexplored way to handle such complex models is to extend the boosting approach we employ in Brockhaus et al. (2015). This may also open the door to conveniently including additional features like ARCH effects in a functional setting (Hörmann et al., 2013). Another step towards multivariate models could be the use of predetermined factors as demonstrated in Diebold and Li (2006) and Diebold et al. (2008) in the context of term structure dynamics. Such a strategy comes at the cost of flexibility and statistical efficiency, but keeps the models *sophisticatedly simple*.

---

<sup>2</sup>Still, most studies, for example Hendershott et al. (2011), find that HFT improves liquidity. The criticism in Budish et al. (2013) mainly concerns the waste of resources attributed to the technological “arms race” of HF traders.

## Bibliography

- Aït-Sahalia, Y. and J. Jacod (2014). *High-Frequency Financial Econometrics*. Princeton University Press.
- Alvarez, A., F. Panloup, M. Pontier, and N. Savy (2012). Estimation of the instantaneous volatility. *Statistical Inference for Stochastic Processes* 15(1), 27–59.
- Amado, C. and T. Teräsvirta (2013). Modelling volatility by variance decomposition. *Journal of Econometrics* 175(2), 142–153.
- Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets* 5(1), 31–56.
- Amihud, Y., H. Mendelson, and L. H. Pedersen (2013). *Market Liquidity: Asset Pricing, Risk, and Crises*. Cambridge: Cambridge University Press.
- Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2000). Great realisations. *Risk* 13, 105–108.
- Andersen, T. G. and L. Benzoni (2009). Realized volatility. In *Handbook of Financial Time Series*, pp. 555–575. Springer.
- Andersen, T. G. and T. Bollerslev (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance* 4(2), 115–158.
- Andersen, T. G. and T. Bollerslev (1998). Deutsche mark–dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *The Journal of Finance* 53(1), 219–265.
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The distribution of realized stock return volatility. *Journal of Financial Economics* 61(1), 43–76.
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96(453), 42–55.

## Bibliography

- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica* 71(2), 579–625.
- Andersen, T. G. and O. Bondarenko (2014). Vpin and the flash crash. *Journal of Financial Markets* 17, 1–46.
- Andersen, T. G., D. Dobrev, and E. Schaumburg (2008). Duration-based volatility estimation. *Manuscript, Northwestern University*.
- Ash, R. B. and M. F. Gardner (1975). *Topics in Stochastic Processes*. New York: Academic Press.
- Aue, A., D. D. Norinho, and S. Hörmann (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* 110(509), 378–392.
- Bachelier, L. (1900). *Théorie de la spéculation*. Paris: Gauthier-Villars.
- Bacry, E., S. Delattre, M. Hoffmann, and J.-F. Muzy (2013). Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance* 13(1), 65–77.
- Barndorff-Nielsen, O. E. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(2), 253–280.
- Barndorff-Nielsen, O. E. and N. Shephard (2002). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics* 17(5), 457–477.
- Bauwens, L. and P. Giot (2000). The logarithmic acid model: An application to the bid-ask quote process of three nyse stocks. *Annales d'Economie et de Statistique* (60), 117–149.
- Bauwens, L. and N. Hautsch (2006). Stochastic conditional intensity processes. *Journal of Financial Econometrics* 4(3), 450–493.
- Bauwens, L. and N. Hautsch (2009). *Modelling Financial High Frequency Data Using Point Processes*, pp. 953–981. *Handbook of Financial Time Series*. Springer.
- Bauwens, L., W. Pohlmeier, and D. Veredas (2008). *High Frequency Financial Econometrics: Recent Developments*. Springer.
- Bauwens, L. and D. Veredas (2004). The stochastic conditional duration model: a latent variable model for the analysis of financial durations. *Journal of Econometrics* 119(2), 381–412.
- Benko, M., W. Härdle, and A. Kneip (2009). Common functional principal components. *The Annals of Statistics* 37(1), 1–34.
- Besse, P. C., H. Cardot, and D. B. Stephenson (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* 27(4), 673–687.

- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31(3), 307–327.
- Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Springer.
- Bowsher, C. G. (2004). Modelling the dynamics of cross-sectional price functions: an econometric analysis of the bid and ask curves of an automated exchange. *Working Paper, Nuffield College, University of Oxford*.
- Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics* 141, 876–912.
- Bowsher, C. G. and R. Meeks (2008). The dynamics of economic functions: Modeling and forecasting the yield curve. *Journal of the American Statistical Association* 103(484), 1413–1437.
- Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. Siam.
- Brockhaus, S., A. Fuest, A. Mayr, and S. Greven (2015). Functional regression models for location, scale and shape applied to stock returns. In *IWSM 2015: 30th International Workshop on Statistical Modelling*. Statistical Modelling Society.
- Budish, E. B., P. Cramton, and J. J. Shim (2013). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Working Paper*.
- Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & Probability Letters* 45(1), 11–22.
- Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton: CRC press.
- Charpentier, A. (2012). Generating a non-homogeneous poisson process. URL: <http://freakonometrics.hypotheses.org/724>.
- Chen, Y., W. K. Härdle, and U. Pigorsch (2010). Localized realized volatility modeling. *Journal of the American Statistical Association* 105(492), 1376–1393.
- Conrad, J., S. Wahal, and J. Xiang (2015). High-frequency quoting, trading, and the efficiency of prices. *Journal of Financial Economics* 116(2), 271–291.
- Cont, R. and A. De Larrard (2013). Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics* 4(1), 1–25.
- Cont, R., A. Kukanov, and S. Stoikov (2014). The price impact of order book events. *Journal of Financial Econometrics* 12(1), 47–88.
- Cont, R., S. Stoikov, and R. Talreja (2010). A stochastic model for order book dynamics. *Operations Research* 58, 549–563.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7(2), 174–196.

- Corsi, F., F. Audrino, and R. Reno (2012). Har modeling for realized volatility forecasting. In: *Handbook of Volatility Models and Their Applications*. New York: Wiley.
- Corsi, F., S. Mittnik, C. Pigorsch, and U. Pigorsch (2008). The volatility of realized volatility. *Econometric Reviews* 27(1), 46–78.
- Crainiceanu, C., P. Reiss, J. Goldsmith, L. Huang, L. Huo, and F. Scheipl. refund: Regression with functional data, 2012. URL: <http://CRAN.R-project.org/package=refund>, r package version 0.1-6.
- Daley, D. J. and D. Vere-Jones (2003). *An Introduction to the Theory of Point Processes* (2nd ed.), Volume I. New York: Springer.
- Daley, D. J. and D. Vere-Jones (2008). *An Introduction to the Theory of Point Processes* (2nd ed.), Volume II. New York: Springer.
- Di, C.-Z., C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics* 3(1), 458.
- Didericksen, D., P. Kokoszka, and X. Zhang (2012). Empirical properties of forecasts with the functional autoregressive model. *Computational Statistics* 27(2), 285–298.
- Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. *Journal of Business & Economic Statistics* 33.
- Diebold, F. X. and C. Li (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics* 130(2), 337–364.
- Diebold, F. X., C. Li, and V. Z. Yue (2008). Global yield curve dynamics and interactions: A dynamic nelson–siegel approach. *Journal of Econometrics* 146(2), 351–363.
- Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13, 253–265.
- Easley, D., M. López de Prado, and M. O’Hara (2011). The microstructure of the flash crash: Flow toxicity, liquidity crashes and the probability of informed trading. *Journal of Portfolio Management* 37(2), 118–128.
- Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 89–102.
- Elezović, S. (2009). Functional modelling of volatility in the swedish limit order book. *Computational Statistics & Data Analysis* 53(6), 2107–2118.
- Engle, R. (2000). The econometrics of ultra-high-frequency data. *Econometrica* 68(1), 1–22.
- Engle, R., M. Fleming, E. Ghysels, and G. Nguyen (2012). Liquidity, volatility, and flights to safety in the us treasury market: evidence from a new class of dynamic order book models. *FRB of New York Staff Report* (590).

- Engle, R. F. and V. K. Ng (1993). Measuring and testing the impact of news on volatility. *The Journal of Finance* 48(5), 1749–1778.
- Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* 66(5), 1127–1162.
- Engle, R. F. and M. E. Sokalska (2012). Forecasting intraday volatility in the us equity market. multiplicative component garch. *Journal of Financial Econometrics* 10(1), 54–83.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* 25(2), 383–417.
- Fengler, M. R., W. K. Härdle, and E. Mammen (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics* 5(2), 189–218.
- Fernandes, M. and J. Grammig (2006). A family of autoregressive conditional duration models. *Journal of Econometrics* 130(1), 1–23.
- Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics* 82(4), 540–554.
- Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005). The generalized dynamic factor model. *Journal of the American Statistical Association* 100(471), 830–840.
- Francq, C., O. Wintenberger, and J.-M. Zakoïan (2013). Garch models without positivity constraints: Exponential or log garch? *Journal of Econometrics* 177(1), 34–46.
- Frey, S. and J. Grammig (2008). Liquidity supply and adverse selection in a pure limit order book market. In L. Bauwens, W. Pohlmeier, and D. Veredas (Eds.), *High Frequency Financial Econometrics: Recent Developments*. Springer.
- Fuest, A. and S. Mittnik (2015). Modeling liquidity impact on volatility: A garch-funxl approach. *CEQURA Working Paper*.
- Gabrys, R., L. Horváth, and P. Kokoszka (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association* 105(491), 1113–1125.
- Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form: the fourier flexible form. *Journal of Econometrics* 15(2), 211–245.
- Gallant, A. R., P. E. Rossi, and G. Tauchen (1992). Stock prices and volume. *Review of Financial Studies* 5(2), 199–242.
- Gatheral, J. and R. Oomen (2010). Zero-intelligence realized variance estimation. *Finance and Stochastics* 14(2), 249–283.

- Giot, P. and J. Grammig (2008). How large is liquidity risk in an automated auction market? In L. Bauwens, W. Pohlmeier, and D. Veredas (Eds.), *High Frequency Financial Econometrics: Recent Developments*. Springer.
- Giot, P., S. Laurent, and M. Petitjean (2010). Trading activity, realized volatility and jumps. *Journal of Empirical Finance* 17(1), 168–175.
- Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics* 20(4), 830–851.
- Goldsmith, J. and F. Scheipl (2014). Estimator selection and combination in scalar-on-function regression. *Computational Statistics & Data Analysis* 70, 362–372.
- Gomber, P. and U. Schweickert (2002). The market impact liquidity measure in electronic securities trading. *Die Bank* 7.
- Gomber, P., U. Schweickert, and E. Theissen (2015). Liquidity dynamics in an electronic open limit order book: an event study approach. *European Financial Management* 21(1), 52–78.
- Gould, M. D., M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison (2013). Limit order books. *Quantitative Finance* 13(11), 1709 – 1742.
- Gouriéroux, C., G. Le Fol, and B. Meyer (1998). Étude du carnet d’ordres. *Banque et Marchés* 36, 5–20.
- Grammig, J. and K.-O. Maurer (2000). Non-monotonic hazard functions and the autoregressive conditional duration model. *Econometrics Journal* 3(1), 16–38.
- Grammig, J. and M. Wellner (2002). Modeling the interdependence of volatility and inter-transaction duration processes. *Journal of Econometrics* 106(2), 369–400.
- Hall, A. D. and N. Hautsch (2007). Modelling the buy and sell intensity in a limit order book market. *Journal of Financial Markets* 10, 249–286.
- Hall, A. D. and N. Hautsch (2008). Order aggressiveness and order book dynamics. In L. Bauwens, W. Pohlmeier, and D. Veredas (Eds.), *High Frequency Financial Econometrics: Recent Developments*. Springer.
- Han, H. (2013). Asymptotic properties of garch-x processes. *Journal of Financial Econometrics* (forthcoming).
- Han, H. and D. Kristensen (2014). Asymptotic theory for the qmle in garch-x models with stationary and non-stationary covariates. *Journal of Business & Economic Statistics* (forthcoming).
- Han, H. and D. Kristensen (2015). Multiplicative garch-x model: Adopting economic variables to explain volatility of financial time series. *Working Paper*.

- Han, H. and J. Y. Park (2008). Time series properties of arch processes with persistent covariates. *Journal of Econometrics* 146(2), 275–292.
- Han, H., M. D. Park, and S. Zhang (2015). A multiplicative error model with heterogeneous components for forecasting realized volatility. *Journal of Forecasting* 34(3), 209–219.
- Hansen, P. R., Z. Huang, and H. H. Shek (2012). Realized garch: a joint model for returns and realized measures of volatility. *Journal of Applied Econometrics* 27(6), 877–906.
- Härdle, W. K., N. Hautsch, and A. Mihoci (2012a). Local adaptive multiplicative error models for high-frequency forecasts. *Working Paper, HU Berlin*.
- Härdle, W. K., N. Hautsch, and A. Mihoci (2012b). Modelling and forecasting liquidity supply using semiparametric factor dynamics. *Journal of Empirical Finance* 19(4), 610–625.
- Hasbrouck, J. (2015). High frequency quoting: Short-term volatility in bids and offers. *Working Paper available at SSRN*.
- Hasbrouck, J. and G. Saar (2013). Low-latency trading. *Journal of Financial Markets* 16(4), 646–679.
- Hautsch, N. (2012). *Econometrics of financial high-frequency data*. Berlin: Springer.
- Hautsch, N. and R. Huang (2012). The market impact of a limit order. *Journal of Economic Dynamics and Control* 36(4), 501–522.
- Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1), 83–90.
- Hays, S., H. Shen, and J. Z. Huang (2012). Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics* 6(3), 870–894.
- Hendershott, T., C. M. Jones, and A. J. Menkveld (2011). Does algorithmic trading improve liquidity? *The Journal of Finance* 66(1), 1–33.
- Hendershott, T. and R. Riordan (2013). Algorithmic trading and the market for liquidity. *Journal of Financial and Quantitative Analysis* 48(04), 1001–1024.
- Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6(2), 327–343.
- Hollifield, B., R. A. Miller, and P. Sandås (2004). Empirical analysis of limit order markets. *The Review of Economic Studies* 71, 1027–1063.
- Hörmann, S., L. Horváth, and R. Reeder (2013). A functional version of the arch model. *Econometric theory* 29(02), 267–288.

## Bibliography

- Hörmann, S., Ł. Kidziński, and M. Hallin (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society (Series B)* (forthcoming).
- Hörmann, S. and P. Kokoszka (2010). Weakly dependent functional data. *The Annals of Statistics* 38(3), 1845–1884.
- Hörmann, S. and P. Kokoszka (2012). Functional time series. *Handbook of Statistics: Time Series Analysis: Methods and Applications* 30, 157.
- Horváth, L. and P. Kokoszka (2012). *Inference for functional data with applications*. Berlin: Springer.
- Horváth, L., P. Kokoszka, and G. Rice (2014). Testing stationarity of functional time series. *Journal of Econometrics* 179(1), 66–82.
- Horváth, L. and G. Rice (2015). Testing for independence between functional time series. *Journal of Econometrics* (forthcoming).
- Hu, Y.-P. and R. S. Tsay (2014). Principal volatility component analysis. *Journal of Business & Economic Statistics* 32(2), 153–164.
- Hyndman, R. J. and M. Shahid Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis* 51(10), 4942–4956.
- Hyndman, R. J. and H. L. Shang (2009). Forecasting functional time series. *Journal of the Korean Statistical Society* 38(3), 199–211.
- Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 31(3), 300–303.
- Jones, C., G. Kaul, and M. Lipson (1994). Transactions, volume, and volatility. *Review of Financial Studies* 7(4), 631–651.
- Kargin, V. and A. Onatski (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis* 99(10), 2508–2526.
- Kirilenko, A. A., A. S. Kyle, M. Samadi, and T. Tuzun (2014). The flash crash: The impact of high frequency trading on an electronic market. *Available at SSRN*.
- Kokoszka, P., H. Miao, and X. Zhang (2014). Functional dynamic factor model for intraday price curves. *Journal of Financial Econometrics* (forthcoming).
- Kowal, D. R., D. S. Matteson, and D. Ruppert (2014). A Bayesian Multivariate Functional Dynamic Linear Model. *ArXiv e-prints*.
- Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica* 53(6), 1315–1336.
- Large, J. (2007). Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets* 10(1), 1–25.

- Lewis, M. (2014). *Flash Boys: A Wall Street Revolt*. New York: WW Norton & Company.
- Li, Y., N. Wang, and R. J. Carroll (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association* 108(504), 1284–1294.
- Linton, O. B. and Y. Yan (2011). Semi- and nonparametric arch processes. *Journal of Probability and Statistics* 2011.
- Liu, L. Y., A. J. Patton, and K. Sheppard (2015). Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. *Journal of Econometrics* 187(1), 293–311.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.
- Mikosch, T. and C. Stărică (2004). Nonstationarities in financial time series, the long-range dependence, and the igarch effects. *Review of Economics and Statistics* 86(1), 378–390.
- Müller, H.-G., R. Sen, and U. Stadtmüller (2011). Functional data analysis for volatility. *Journal of Econometrics* 165(2), 233–245.
- Müller, H.-G., U. Stadtmüller, and F. Yao (2006). Functional variance processes. *Journal of the American Statistical Association* 101(475), 1007–1018.
- Munk, A., J. Schmidt-Hieber, et al. (2010). Nonparametric estimation of the volatility function in a high-frequency model corrupted by noise. *Electronic Journal of Statistics* 4, 781–821.
- Oomen, R. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business and Economic Statistics* 24(2), 219–237.
- Panaretos, V. M. and S. Tavakoli (2013). Cramér–karhunen–loève representation and harmonic principal component analysis of functional time series. *Stochastic Processes and their Applications* 123(7), 2779–2807.
- Park, B. U., E. Mammen, W. Härdle, and S. Borak (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association* 104(485), 284–298.
- Pasupathy, R. (2011). Generating nonhomogeneous poisson processes. *Wiley Encyclopedia of Operations Research and Management Science*.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160(1), 246–256.
- Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association* 89(428), 1303–1313.
- Preis, T. (2010). *Quantifying and modeling financial market fluctuations*. Ph. D. thesis, JGU Mainz.

## Bibliography

- Press, S. J. (1967). A compound events model for security prices. *J. Business* 40(3), 317–335.
- Ramsay, J. and B. Silverman (2005). *Functional data analysis*. New York: Springer.
- Reiss, P. T., J. Goldsmith, H. L. Shang, and R. T. Ogden (2015). Methods for scalar-on-function regression. *Working Paper*.
- Reiss, P. T. and R. T. Ogden (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association* 102(479), 984–996.
- Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(3), 507–554.
- Roşu, I. (2009). A dynamic model of the limit order book. *Review of Financial Studies* 22(11), 4601–4641.
- Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.
- Sen, R. and C. Klüppelberg (2010). Time series of functional data. *Working paper, TU Munich*.
- Smith, E., J. D. Farmer, L. s. Gillemot, and S. Krishnamurthy (2003). Statistical theory of the continuous double auction. *Quantitative Finance* 3(6), 481–514.
- Staniswalis, J. G. and J. J. Lee (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 93(444), 1403–1418.
- Straumann, D., T. Mikosch, et al. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *The Annals of Statistics* 34(5), 2449–2495.
- Sucarrat, G. (2014). lgarch: Simulation and estimation of log-garch models. *R package version 0.2*.
- Sucarrat, G., S. Grønneberg, and A. Escibano (2013). Estimation and inference in univariate and multivariate log-garch-x models when the conditional density is unknown. *Working Paper*.
- Tse, Y.-k. and T. T. Yang (2012). Estimation of high-frequency volatility: An autoregressive conditional duration approach. *Journal of Business & Economic Statistics* 30(4), 533–545.
- van Kervel, V. (2015). Competition for order flow with fast and slow traders. *Review of Financial Studies* (forthcoming).

- Wolak, F. A. (1991). The local nature of hypothesis tests involving inequality constraints in nonlinear models. *Econometrica* 59(4), 981–95.
- Xiao, L., Y. Li, and D. Ruppert (2013). Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(3), 577–599.
- Xiao, L., D. Ruppert, V. Zipunnikov, and C. Crainiceanu (2013). Fast covariance estimation for high-dimensional functional data. *arXiv preprint arXiv:1306.5718*.
- Yao, F., H.-G. Müller, and J.-L. Wang (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100(470), 577–590.
- Yao, F., H.-G. Müller, and J.-L. Wang (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics* 33(6), 2873–2903.
- Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005). A tale of two time scales. *Journal of the American Statistical Association* 100(472), 1394–1411.



## Additional empirical results

### A.4 Chapter 4: Price durations and the ACD-FunXL model

#### A.4.1 Goodness-of-Fit

In the following, results of estimation for models for bid and ask durations are shown, for several  $\delta$  (1,2,3,4,5,10,20 Ticks), each with bid and ask liquidity impact. The goodness-of-fit criteria are (naive) AIC and BIC (mean per observation) for the EACD log-likelihood.

#### Commerzbank

Table A.1: AIC for EACD-FunXL, Commerzbank.

$\delta = 0.001$  EUR (1 Tick)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.4554	0.4554	0.4554	0.4554	0.5514	0.5514	0.5514	0.5514
1	0.3612	0.3610	0.3619	0.3645	0.4157	0.4212	0.4245	0.4314
2	0.3568	0.3585	0.3610	0.3638	0.4145	0.4180	0.4208	0.4227
3	0.3432	0.3537	0.3572	0.3609	0.4096	0.4163	0.4189	0.4248
4	0.3430	0.3488	0.3511	0.3499	0.4076	0.4156	0.4170	0.4177
5	0.3434	0.3497	0.3477	0.3483	0.4075	0.4139	0.4166	0.4183

$\delta = 0.002$  EUR (2 Ticks)

*Continued on next page*

Table A.1 – Continued from previous page

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.7788	0.7788	0.7788	0.7788	0.9279	0.9279	0.9279	0.9279
1	0.6919	0.7093	0.7186	0.7220	0.8226	0.8408	0.8486	0.8524
2	0.6847	0.6917	0.6995	0.7034	0.8222	0.8197	0.8231	0.8229
3	0.7078	0.7036	0.7029	0.6986	0.8262	0.8355	0.8339	0.8279
4	0.7070	0.7073	0.7086	0.7090	0.8182	0.8288	0.8288	0.8290
5	0.7129	0.7197	0.7110	0.7112	0.8159	0.8228	0.8223	0.8246

$\delta = 0.003$  EUR (3 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.8941	0.8941	0.8941	0.8941	1.0360	1.0360	1.0360	1.0360
1	0.8205	0.8387	0.8484	0.8510	0.9635	0.9777	0.9841	0.9866
2	0.8198	0.8173	0.8233	0.8281	0.9659	0.9635	0.9663	0.9665
3	0.8322	0.8308	0.8269	0.8206	0.9615	0.9771	0.9764	0.9711
4	0.8198	0.8333	0.8352	0.8345	0.9309	0.9647	0.9655	0.9649
5	0.8347	0.8435	0.8358	0.8377	0.9253	0.9423	0.9519	0.9551

$\delta = 0.004$  EUR (4 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.2607	1.2607	1.2607	1.2607	1.3779	1.3779	1.3779	1.3779
1	1.2463	1.2561	1.2627	1.2635	1.3549	1.3633	1.3666	1.3666
2	1.2542	1.2427	1.2453	1.2498	1.3605	1.3587	1.3637	1.3639
3	1.2291	1.2568	1.2534	1.2485	1.3546	1.3767	1.3764	1.3703
4	1.1680	1.2429	1.2530	1.2494	1.3175	1.3604	1.3624	1.3563
5	1.1660	1.2527	1.2450	1.2497	1.3090	1.3401	1.3552	1.3520

$\delta = 0.005$  EUR (5 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.0164	1.0164	1.0164	1.0164	1.1071	1.1071	1.1071	1.1071
1	0.9777	0.9890	0.9937	0.9949	1.0689	1.0744	1.0768	1.0768
2	0.9738	0.9779	0.9839	0.9858	1.0800	1.0738	1.0772	1.0768

Continued on next page

Table A.1 – Continued from previous page

3	0.9986	0.9957	0.9887	0.9817	1.1103	1.0989	1.0949	1.0866
4	0.9853	1.0078	1.0072	1.0021	1.1033	1.1112	1.1056	1.0986
5	0.9563	1.0167	1.0093	1.0124	1.0981	1.1089	1.1113	1.1077

 $\delta = 0.010$  EUR (10 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.1015	1.1015	1.1015	1.1015	1.0517	1.0517	1.0517	1.0517
1	1.1411	1.1357	1.1336	1.1316	0.9702	0.9726	0.9753	0.9763
2	1.1453	1.1391	1.1434	1.1408	0.9994	0.9851	0.9882	0.9831
3	1.1913	1.1481	1.1454	1.1409	1.0564	1.0515	1.0351	1.0228
4	1.2311	1.1349	1.1403	1.1455	1.0617	1.0497	1.0444	1.0498
5	1.3564	1.1915	1.1460	1.1235	1.0812	1.0727	1.0615	1.0661

 $\delta = 0.020$  EUR (20 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.8024	0.8024	0.8024	0.8024	1.1946	1.1946	1.1946	1.1946
1	0.7782	0.7813	0.7796	0.7790	1.0814	1.0883	1.0804	1.0761
2	0.8108	0.8049	0.7837	0.7800	1.1667	1.1186	1.0857	1.0739
3	0.8256	0.8243	0.8392	0.8191	1.1886	1.1431	1.1499	1.1676
4	0.8332	0.8280	0.8442	0.8510	1.2514	1.1457	1.1319	1.1652
5	0.8407	0.8944	0.8569	0.8491	1.3504	1.2049	1.2007	1.2165

Table A.2: BIC for EACD-FunXL, Commerzbank.

 $\delta = 0.001$  EUR (1 Tick)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.4562	0.4562	0.4562	0.4562	0.5523	0.5523	0.5523	0.5523
1	0.3626	0.3624	0.3633	0.3659	0.4173	0.4227	0.4260	0.4329
2	0.3587	0.3605	0.3630	0.3658	0.4166	0.4201	0.4229	0.4249
3	0.3458	0.3562	0.3597	0.3635	0.4123	0.4190	0.4216	0.4275

*Continued on next page*

Table A.2 – Continued from previous page

4	0.3461	0.3519	0.3542	0.3529	0.4109	0.4189	0.4203	0.4210
5	0.3470	0.3534	0.3514	0.3520	0.4114	0.4178	0.4205	0.4222

 $\delta = 0.002$  EUR (2 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.7804	0.7804	0.7804	0.7804	0.9296	0.9296	0.9296	0.9296
1	0.6946	0.7120	0.7212	0.7247	0.8255	0.8437	0.8514	0.8553
2	0.6884	0.6954	0.7032	0.7071	0.8262	0.8237	0.8271	0.8269
3	0.7126	0.7084	0.7077	0.7033	0.8313	0.8407	0.8390	0.8331
4	0.7128	0.7131	0.7144	0.7148	0.8245	0.8351	0.8351	0.8352
5	0.7198	0.7265	0.7179	0.7181	0.8234	0.8302	0.8297	0.8320

 $\delta = 0.003$  EUR (3 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.8959	0.8959	0.8959	0.8959	1.0379	1.0379	1.0379	1.0379
1	0.8235	0.8418	0.8515	0.8541	0.9668	0.9810	0.9874	0.9899
2	0.8241	0.8215	0.8276	0.8324	0.9705	0.9681	0.9709	0.9710
3	0.8377	0.8363	0.8325	0.8261	0.9674	0.9830	0.9823	0.9770
4	0.8266	0.8400	0.8420	0.8413	0.9381	0.9719	0.9727	0.9721
5	0.8426	0.8515	0.8438	0.8456	0.9338	0.9508	0.9604	0.9636

 $\delta = 0.004$  EUR (4 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.2636	1.2636	1.2636	1.2636	1.3809	1.3809	1.3809	1.3809
1	1.2511	1.2609	1.2675	1.2684	1.3600	1.3684	1.3717	1.3717
2	1.2610	1.2495	1.2521	1.2566	1.3676	1.3658	1.3709	1.3710
3	1.2378	1.2654	1.2621	1.2572	1.3638	1.3859	1.3855	1.3794
4	1.1786	1.2535	1.2636	1.2600	1.3287	1.3716	1.3736	1.3675
5	1.1785	1.2652	1.2575	1.2622	1.3222	1.3533	1.3684	1.3652

 $\delta = 0.005$  EUR (5 Ticks)*Continued on next page*

Table A.2 – Continued from previous page

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.0205	1.0205	1.0205	1.0205	1.1114	1.1114	1.1114	1.1114
1	0.9845	0.9958	1.0005	1.0017	1.0759	1.0814	1.0838	1.0839
2	0.9833	0.9874	0.9934	0.9953	1.0898	1.0836	1.0871	1.0866
3	1.0109	1.0079	1.0009	0.9939	1.1229	1.1115	1.1076	1.0992
4	1.0002	1.0228	1.0222	1.0170	1.1187	1.1266	1.1210	1.1141
5	0.9740	1.0343	1.0270	1.0300	1.1163	1.1271	1.1295	1.1259

 $\delta = 0.010$  EUR (10 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.1101	1.1101	1.1101	1.1101	1.0609	1.0609	1.0609	1.0609
1	1.1554	1.1500	1.1480	1.1460	0.9855	0.9879	0.9906	0.9916
2	1.1654	1.1593	1.1635	1.1609	1.0208	1.0065	1.0096	1.0046
3	1.2171	1.1739	1.1712	1.1668	1.0840	1.0791	1.0627	1.0504
4	1.2627	1.1665	1.1719	1.1771	1.0954	1.0833	1.0781	1.0835
5	1.3937	1.2288	1.1833	1.1609	1.1210	1.1125	1.1013	1.1059

 $\delta = 0.020$  EUR (20 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.8187	0.8187	0.8187	0.8187	1.2136	1.2136	1.2136	1.2136
1	0.8052	0.8083	0.8066	0.8061	1.1132	1.1201	1.1122	1.1078
2	0.8487	0.8428	0.8216	0.8178	1.2112	1.1630	1.1302	1.1183
3	0.8743	0.8730	0.8878	0.8678	1.2458	1.2003	1.2070	1.2247
4	0.8927	0.8875	0.9037	0.9106	1.3212	1.2156	1.2017	1.2350
5	0.9111	0.9647	0.9272	0.9194	1.4329	1.2874	1.2832	1.2990

Linde

**Table A.3:** AIC for EACD-FunXL, Linde.

$\delta = 0.01$  EUR (1 Tick)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.6190	0.6190	0.6190	0.6190	0.7493	0.7493	0.7493	0.7493
1	0.4909	0.4906	0.4919	0.4954	0.5650	0.5724	0.5768	0.5863
2	0.4849	0.4873	0.4907	0.4945	0.5633	0.5680	0.5718	0.5745
3	0.4665	0.4808	0.4855	0.4906	0.5566	0.5657	0.5693	0.5773
4	0.4662	0.4741	0.4772	0.4755	0.5538	0.5647	0.5667	0.5676
5	0.4667	0.4753	0.4726	0.4734	0.5537	0.5625	0.5661	0.5684

$\delta = 0.02$  EUR (2 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.9377	0.9377	0.9377	0.9377	1.0945	1.0945	1.0945	1.0945
1	0.8331	0.8541	0.8652	0.8694	0.9703	0.9917	1.0009	1.0054
2	0.8244	0.8329	0.8422	0.8469	0.9697	0.9668	0.9708	0.9706
3	0.8523	0.8472	0.8464	0.8411	0.9745	0.9855	0.9836	0.9765
4	0.8513	0.8517	0.8532	0.8536	0.9651	0.9776	0.9776	0.9778
5	0.8584	0.8665	0.8561	0.8563	0.9624	0.9705	0.9699	0.9726

$\delta = 0.03$  EUR (3 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.0280	1.0280	1.0280	1.0280	1.1760	1.1760	1.1760	1.1760
1	0.9433	0.9643	0.9754	0.9785	1.0937	1.1099	1.1172	1.1199
2	0.9426	0.9396	0.9466	0.9521	1.0964	1.0938	1.0969	1.0971
3	0.9568	0.9552	0.9508	0.9435	1.0914	1.1092	1.1084	1.1024
4	0.9426	0.9581	0.9603	0.9595	1.0567	1.0951	1.0960	1.0953
5	0.9597	0.9699	0.9610	0.9631	1.0504	1.0696	1.0806	1.0842

$\delta = 0.04$  EUR (4 Ticks)

K	ask quote	bid quote
---	-----------	-----------

*Continued on next page*

Table A.3 – Continued from previous page

	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.1860	1.1860	1.1860	1.1860	1.2884	1.2884	1.2884	1.2884
1	1.1725	1.1817	1.1879	1.1887	1.2669	1.2747	1.2779	1.2778
2	1.1799	1.1691	1.1716	1.1758	1.2721	1.2705	1.2752	1.2753
3	1.1563	1.1823	1.1792	1.1746	1.2666	1.2873	1.2870	1.2812
4	1.0988	1.1693	1.1788	1.1754	1.2319	1.2720	1.2739	1.2682
5	1.0969	1.1785	1.1713	1.1756	1.2240	1.2531	1.2672	1.2642

 $\delta = 0.05$  EUR (5 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.2455	1.2455	1.2455	1.2455	1.3690	1.3690	1.3690	1.3690
1	1.1980	1.2118	1.2176	1.2191	1.3217	1.3285	1.3314	1.3315
2	1.1933	1.1982	1.2056	1.2079	1.3354	1.3277	1.3320	1.3315
3	1.2236	1.2200	1.2115	1.2028	1.3729	1.3588	1.3539	1.3436
4	1.2073	1.2349	1.2342	1.2279	1.3642	1.3740	1.3671	1.3585
5	1.1718	1.2458	1.2367	1.2405	1.3578	1.3711	1.3741	1.3696

 $\delta = 0.10$  EUR (10 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.5709	1.5709	1.5709	1.5709	1.4454	1.4454	1.4454	1.4454
1	1.6273	1.6196	1.6167	1.6139	1.3334	1.3368	1.3404	1.3418
2	1.6334	1.6246	1.6306	1.6270	1.3735	1.3539	1.3582	1.3512
3	1.6989	1.6373	1.6335	1.6272	1.4519	1.4452	1.4227	1.4058
4	1.7557	1.6185	1.6263	1.6337	1.4592	1.4426	1.4354	1.4428
5	1.9344	1.6993	1.6343	1.6023	1.4859	1.4743	1.4589	1.4652

 $\delta = 0.20$  EUR (20 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.5984	1.5984	1.5984	1.5984	2.0033	2.0033	2.0033	2.0033
1	1.5501	1.5562	1.5528	1.5517	1.8135	1.8251	1.8119	1.8045
2	1.6150	1.6034	1.5611	1.5536	1.9566	1.8758	1.8207	1.8009
3	1.6445	1.6420	1.6715	1.6316	1.9933	1.9170	1.9283	1.9580

*Continued on next page*

Table A.3 – Continued from previous page

4	1.6598	1.6493	1.6816	1.6952	2.0985	1.9213	1.8981	1.9540
5	1.6747	1.7816	1.7068	1.6913	2.2645	2.0206	2.0135	2.0400

Table A.4: BIC for EACD-FunXL, Linde.

$\delta = 0.01$  EUR (1 Tick)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.6201	0.6201	0.6201	0.6201	0.7506	0.7506	0.7506	0.7506
1	0.4928	0.4925	0.4938	0.4973	0.5670	0.5745	0.5788	0.5883
2	0.4876	0.4900	0.4934	0.4972	0.5661	0.5709	0.5747	0.5773
3	0.4700	0.4842	0.4889	0.4940	0.5603	0.5694	0.5729	0.5809
4	0.4704	0.4783	0.4814	0.4797	0.5583	0.5692	0.5712	0.5721
5	0.4717	0.4803	0.4776	0.4784	0.5591	0.5678	0.5714	0.5737

$\delta = 0.02$  EUR (2 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	0.9396	0.9396	0.9396	0.9396	1.0965	1.0965	1.0965	1.0965
1	0.8363	0.8573	0.8684	0.8725	0.9736	0.9951	1.0043	1.0088
2	0.8288	0.8373	0.8467	0.8514	0.9744	0.9715	0.9755	0.9753
3	0.8580	0.8529	0.8521	0.8468	0.9806	0.9916	0.9896	0.9826
4	0.8583	0.8586	0.8602	0.8606	0.9725	0.9850	0.9850	0.9852
5	0.8667	0.8748	0.8644	0.8646	0.9712	0.9793	0.9787	0.9814

$\delta = 0.03$  EUR (3 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.0301	1.0301	1.0301	1.0301	1.1782	1.1782	1.1782	1.1782
1	0.9469	0.9678	0.9790	0.9820	1.0974	1.1136	1.1209	1.1237
2	0.9475	0.9446	0.9515	0.9570	1.1016	1.0990	1.1021	1.1023
3	0.9632	0.9615	0.9571	0.9498	1.0981	1.1159	1.1151	1.1091
4	0.9504	0.9658	0.9680	0.9672	1.0649	1.1033	1.1042	1.1035

Continued on next page

Table A.4 – Continued from previous page

5	0.9688	0.9790	0.9702	0.9722	1.0600	1.0793	1.0902	1.0939
---	--------	--------	--------	--------	--------	--------	--------	--------

$\delta = 0.04$  EUR (4 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.1887	1.1887	1.1887	1.1887	1.2912	1.2912	1.2912	1.2912
1	1.1770	1.1862	1.1925	1.1932	1.2716	1.2795	1.2826	1.2826
2	1.1863	1.1755	1.1779	1.1821	1.2788	1.2771	1.2818	1.2819
3	1.1644	1.1905	1.1873	1.1827	1.2752	1.2958	1.2955	1.2898
4	1.1088	1.1792	1.1888	1.1854	1.2424	1.2825	1.2844	1.2786
5	1.1087	1.1902	1.1830	1.1874	1.2363	1.2654	1.2795	1.2765

$\delta = 0.05$  EUR (5 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.2504	1.2504	1.2504	1.2504	1.3742	1.3742	1.3742	1.3742
1	1.2063	1.2202	1.2260	1.2274	1.3303	1.3372	1.3401	1.3402
2	1.2049	1.2098	1.2173	1.2195	1.3476	1.3399	1.3442	1.3436
3	1.2386	1.2349	1.2264	1.2178	1.3885	1.3744	1.3695	1.3592
4	1.2255	1.2532	1.2525	1.2462	1.3833	1.3931	1.3862	1.3776
5	1.1934	1.2674	1.2583	1.2621	1.3803	1.3937	1.3967	1.3922

$\delta = 0.10$  EUR (10 Ticks)

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.5832	1.5832	1.5832	1.5832	1.4580	1.4580	1.4580	1.4580
1	1.6478	1.6401	1.6372	1.6344	1.3544	1.3578	1.3615	1.3628
2	1.6621	1.6533	1.6593	1.6557	1.4030	1.3833	1.3876	1.3807
3	1.7358	1.6742	1.6704	1.6641	1.4898	1.4831	1.4605	1.4436
4	1.8008	1.6636	1.6714	1.6788	1.5055	1.4889	1.4817	1.4891
5	1.9877	1.7525	1.6876	1.6556	1.5407	1.5290	1.5137	1.5199

$\delta = 0.20$  EUR (20 Ticks)

*Continued on next page*

Table A.4 – Continued from previous page

K	ask quote				bid quote			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	1.6307	1.6307	1.6307	1.6307	2.0352	2.0352	2.0352	2.0352
1	1.6039	1.6101	1.6067	1.6056	1.8667	1.8783	1.8651	1.8578
2	1.6905	1.6788	1.6366	1.6291	2.0311	1.9504	1.8952	1.8754
3	1.7415	1.7390	1.7685	1.7286	2.0891	2.0128	2.0242	2.0538
4	1.7783	1.7678	1.8001	1.8138	2.2157	2.0384	2.0153	2.0711
5	1.8148	1.9217	1.8469	1.8314	2.4029	2.1590	2.1519	2.1784

## A.4.2 Forecast results

### Commerzbank

Table A.5: 1-step forecast MSEs, EACD-FunXL, bid-quote durations, Commerzbank, for days 16 to 21 in the sample.

$\delta = 0.001$  EUR (1 Tick)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	11.7733	11.7733	11.7733	11.7733	12.0061	12.0061	12.0061	12.0061
1	11.6583	11.7720	11.9552	12.0172	11.8365	12.0122	12.2384	12.3067
2	11.7566	11.8544	11.9515	11.9830	11.8086	12.0847	12.1704	12.1877
3	11.5565	11.7722	11.7873	11.7950	11.7377	12.0069	12.0325	12.0157
4	11.5718	11.8199	11.8680	11.8592	11.5628	12.0012	11.9907	11.9633
5	11.6905	11.8026	11.8127	11.7939	11.6854	11.8228	11.8539	11.8475

$\delta = 0.002$  EUR (2 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	7.4409	7.4409	7.4409	7.4409	7.8037	7.8037	7.8037	7.8037
1	7.3842	7.4809	7.5794	7.6184	7.7112	7.8672	8.0188	8.0760
2	7.3566	7.5040	7.5307	7.5345	7.6573	7.8673	7.9076	7.9128
3	7.3542	7.4529	7.4672	7.4596	7.6427	7.8078	7.8306	7.8137
4	7.2833	7.4448	7.4569	7.4545	7.5296	7.7796	7.8093	7.7924
5	7.3634	7.4385	7.4348	7.4166	7.6059	7.6963	7.7354	7.7314

Continued on next page

Table A.5 – Continued from previous page

$\delta = 0.003$  EUR (3 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	7.2367	7.2367	7.2367	7.2367	7.5309	7.5309	7.5309	7.5309
1	7.2040	7.2923	7.3717	7.3999	7.4669	7.6296	7.7674	7.8138
2	7.2065	7.2965	7.3265	7.3343	7.4822	7.6625	7.7015	7.7088
3	7.1971	7.2634	7.2738	7.2695	7.4641	7.6214	7.6407	7.6267
4	7.1326	7.2629	7.2711	7.2701	7.3684	7.6057	7.6334	7.6132
5	7.1872	7.2444	7.2679	7.2452	7.4346	7.5304	7.5739	7.5643

$\delta = 0.004$  EUR (4 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	7.3139	7.3139	7.3139	7.3139	7.8651	7.8651	7.8651	7.8651
1	7.2771	7.3403	7.4070	7.4290	7.7267	7.8998	8.0547	8.1017
2	7.2904	7.2968	7.3452	7.3623	7.7999	7.9311	7.9900	8.0076
3	7.2500	7.2625	7.2749	7.2775	7.7596	7.8983	7.9271	7.9245
4	7.1999	7.2268	7.2408	7.2637	7.6514	7.8903	7.9458	7.9320
5	7.1923	7.2358	7.2570	7.2097	7.7089	7.8233	7.8774	7.8502

$\delta = 0.005$  EUR (5 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	6.9429	6.9429	6.9429	6.9429	7.5744	7.5744	7.5744	7.5744
1	6.9091	6.9298	6.9629	6.9738	7.2801	7.4584	7.6317	7.6813
2	6.9416	6.8670	6.8937	6.9066	7.3498	7.4155	7.4599	7.4723
3	6.8634	6.8544	6.8546	6.8679	7.3081	7.4036	7.4450	7.4410
4	6.8883	6.8293	6.8400	6.8776	7.2330	7.4226	7.4725	7.4236
5	6.8775	6.8311	6.8674	6.8179	7.2633	7.3616	7.3868	7.3768

$\delta = 0.010$  EUR (10 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	5.8732	5.8732	5.8732	5.8732	7.3693	7.3693	7.3693	7.3693
1	5.9407	5.9232	5.8736	5.8538	6.4405	6.5086	6.6184	6.5913
2	5.9317	5.8119	5.7944	5.7934	6.5003	6.0153	6.0008	6.0124

Continued on next page

Table A.5 – Continued from previous page

3	5.9419	5.8087	5.7854	5.7774	6.1178	6.0315	5.9226	5.8983
4	5.9916	5.8028	5.7783	5.7750	6.1607	5.7621	5.7950	5.8302
5	5.9290	5.8223	5.8020	5.7978	6.0241	5.7718	5.7979	5.7847

$\delta = 0.020$  EUR (20 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	6.7894	6.7894	6.7894	6.7894	11.3246	11.3246	11.3246	11.3246
1	6.7644	6.2842	5.8911	5.7117	8.4256	14.1300	12.7087	11.0349
2	28.0357	77.5566	5.4520	5.4428	80.0214	26.4208	10.0981	8.3073
3	23.7000	29.8716	5.9803	6.9722	99.7300	96.0933	9.1658	7.8474
4	7.1163	6.6939	8.3304	8.8277	5.9705	32.2277	10.2098	30.4344
5	8.2596	489.8598	340.6562	445.5293	87.3178	106.5640	60.9214	94.8828

Table A.6: 1-step forecast negative log-likelihoods, EACD-FunXL, bid-quote durations, Commerzbank.

$\delta = 0.001$  EUR (1 Tick)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	132.7807	132.7807	132.7807	132.7807	29.8599	29.8599	29.8599	29.8599
1	139.6271	128.9054	115.5066	110.4986	30.9869	29.6028	27.6126	26.9261
2	137.7622	88.9822	93.5263	96.5894	30.3827	27.2940	26.5760	26.4274
3	120.9574	87.6850	88.2370	89.1542	31.4384	28.2724	28.1910	28.2311
4	118.7791	87.4673	89.1304	89.5400	31.7164	27.7267	27.3395	27.4409
5	107.6624	90.6671	90.7568	90.5395	29.7584	28.0682	27.9855	28.1182

$\delta = 0.002$  EUR (2 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	59.1568	59.1568	59.1568	59.1568	18.5709	18.5709	18.5709	18.5709
1	60.4795	58.1004	53.4885	51.2249	19.4968	17.9053	16.3049	15.6767
2	61.9666	39.9391	40.0715	40.7493	19.5863	14.8124	14.7150	14.8107
3	59.6486	40.4612	39.8595	40.4464	19.0041	15.0816	14.9093	15.0125
4	58.4316	40.1380	39.9959	41.0751	19.9714	15.0566	14.9531	15.0568

Continued on next page

Table A.6 – Continued from previous page

5	57.2028	44.6556	41.0928	39.8062	18.1517	15.4141	15.2985	15.3290
---	---------	---------	---------	---------	---------	---------	---------	---------

 $\delta = 0.003$  EUR (3 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	28.9968	28.9968	28.9968	28.9968	16.4561	16.4561	16.4561	16.4561
1	30.1558	27.3000	24.5889	23.5189	17.0893	15.4850	14.0861	13.5586
2	27.7395	20.1435	20.1755	20.3685	16.0805	12.7744	12.5325	12.5230
3	27.0654	20.0158	19.8301	19.9122	16.2439	12.9336	12.8074	12.7896
4	26.8899	19.8413	19.7670	19.9004	16.5630	12.8761	12.8197	12.7532
5	24.2970	20.1681	19.3991	19.9258	15.0496	12.8898	12.5835	13.1228

 $\delta = 0.004$  EUR (4 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	25.6233	25.6233	25.6233	25.6233	14.7380	14.7380	14.7380	14.7380
1	27.4991	25.0252	22.3705	21.3776	15.7498	14.4858	13.1938	12.7296
2	24.9505	18.0113	18.1738	18.3635	14.4643	11.3446	11.3368	11.4021
3	23.5617	17.9353	17.7206	17.8414	14.3869	11.3835	11.3096	11.2997
4	23.8119	17.6775	17.6173	17.7938	14.5608	11.4965	11.4672	11.3300
5	21.6686	17.5247	17.0196	17.1341	13.2487	11.5093	11.3121	11.7351

 $\delta = 0.005$  EUR (5 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	29.3623	29.3623	29.3623	29.3623	12.5518	12.5518	12.5518	12.5518
1	32.4616	31.2226	28.0930	26.7449	13.4456	13.2925	12.4849	12.1237
2	29.7700	19.9115	19.8425	20.0242	12.1602	10.0174	9.9393	9.9642
3	28.3197	19.8475	19.2470	19.4021	12.5336	10.0228	9.8871	9.7718
4	27.8488	19.3438	19.1319	19.1642	12.0738	9.9555	9.8157	9.6330
5	27.4472	19.6357	18.8044	19.1162	11.8335	9.9507	9.5093	9.9011

 $\delta = 0.010$  EUR (10 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	20.4311	20.4311	20.4311	20.4311	4.0387	4.0387	4.0387	4.0387

*Continued on next page*

Table A.6 – Continued from previous page

1	23.5379	22.5895	18.7491	17.2437	4.6775	4.5829	4.4516	4.4279
2	22.8695	14.7489	13.6697	13.5037	4.5980	4.3234	4.3318	4.3469
3	21.5865	14.6012	13.1728	12.7361	4.6215	4.3395	4.3311	4.2950
4	21.7560	12.9870	12.2638	12.2254	4.7306	4.2985	4.2101	4.1712
5	18.8513	12.9576	12.4847	12.2813	4.7083	4.3297	4.2306	4.2119

$\delta = 0.020$  EUR (20 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	15.6184	15.6184	15.6184	15.6184	3.8491	3.8491	3.8491	3.8491
1	15.2867	9.2762	6.4388	5.4752	3.7055	3.9904	3.8985	3.7861
2	4.5803	6.1659	3.9645	3.9427	6.2466	4.6576	3.6892	3.5632
3	4.4822	4.6041	3.4377	3.4302	6.6005	6.5996	3.5919	3.4818
4	28.7744	12.6216	3.5602	193475.3932	3.5513	4.8991	3.6763	4.8113
5	1131.3715	15.4259	23.1420	23.4375	6.2110	7.4356	17.2510	15.0367

**Linde**

Table A.7: 1-step forecast MSEs, EACD-FunXL, bid-quote durations, Linde.

$\delta = 0.01$  EUR (1 Tick)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	43.7432	43.7432	43.7432	43.7432	44.1697	44.1697	44.1697	44.1697
1	42.4187	42.5101	42.5422	42.6163	42.7396	42.7265	42.7393	42.8271
2	42.3925	42.4078	42.4395	42.4485	42.7268	42.7798	42.8170	42.8275
3	42.3141	42.3878	42.4180	42.4831	42.6266	42.7503	42.7900	42.8529
4	42.2957	42.3173	42.3529	42.3435	42.6000	42.7064	42.7303	42.7028
5	42.3209	42.3099	42.3272	42.3770	42.6153	42.6903	42.7263	42.7229

$\delta = 0.02$  EUR (2 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	11.1626	11.1626	11.1626	11.1626	11.0121	11.0121	11.0121	11.0121
1	11.2425	11.1748	11.1925	11.1956	10.9147	10.9439	10.9675	10.9753

*Continued on next page*

Table A.7 – Continued from previous page

2	11.1629	11.3918	11.2923	11.2583	10.9152	10.9400	10.8999	10.8960
3	10.9165	10.9222	10.9928	11.1154	10.7432	10.7471	10.7856	10.8241
4	10.9013	10.8980	10.8914	10.9049	10.6816	10.7257	10.7215	10.7255
5	10.9705	10.8716	10.8949	10.9096	10.6472	10.6881	10.7129	10.7265

$\delta = 0.03$  EUR (3 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	9.0828	9.0828	9.0828	9.0828	8.6437	8.6437	8.6437	8.6437
1	9.1184	9.0960	9.1203	9.1286	8.6229	8.6413	8.6567	8.6577
2	8.9988	9.0636	8.9535	8.9473	8.6005	8.5779	8.5208	8.5214
3	8.5816	8.5703	8.6714	8.8251	8.4136	8.3575	8.3906	8.4293
4	8.5646	8.5434	8.5368	8.5497	8.3733	8.3637	8.3496	8.3400
5	8.6213	8.5131	8.5542	8.5777	8.3646	8.3308	8.3567	8.3546

$\delta = 0.04$  EUR (4 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	8.3959	8.3959	8.3959	8.3959	8.0286	8.0286	8.0286	8.0286
1	8.5698	8.4868	8.4863	8.5035	8.2990	8.2762	8.2753	8.2894
2	7.8831	8.7728	8.5267	8.4903	7.8842	8.1064	7.7435	7.7653
3	6.1145	6.5340	7.1367	7.7253	6.2265	6.3679	6.7208	7.1904
4	5.7990	6.1987	6.2714	6.2261	5.7867	6.3117	6.3666	6.2951
5	5.8167	5.9706	6.1959	6.2071	5.6962	6.0074	6.2683	6.2836

$\delta = 0.05$  EUR (5 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	7.7961	7.7961	7.7961	7.7961	6.0869	6.0869	6.0869	6.0869
1	8.0578	7.9821	8.0071	8.0067	6.0395	6.0654	6.0787	6.0841
2	7.9026	8.1152	7.8426	7.7804	6.0658	6.0220	5.9901	5.9966
3	6.8263	6.9161	7.2284	7.3896	5.8611	5.8716	5.9257	5.9608
4	6.6332	6.9352	6.9526	6.8585	5.8373	5.8746	5.8637	5.8598
5	6.6215	6.6776	6.9507	6.8738	5.8305	5.8591	5.8806	5.8716

$\delta = 0.10$  EUR (10 Ticks)

Continued on next page

Table A.7 – Continued from previous page

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	6.4713	6.4713	6.4713	6.4713	5.8833	5.8833	5.8833	5.8833
1	6.2527	6.0697	6.1576	6.1797	5.9565	5.8708	5.9489	5.9893
2	6.1950	6.7156	6.3372	6.0978	6.1246	6.2156	5.7962	5.7352
3	4.9532	5.6174	5.9126	6.3516	4.9544	5.5827	5.7998	5.9545
4	4.7693	5.1821	5.4136	5.4959	4.8873	5.1658	5.3862	5.4246
5	4.8137	5.0285	5.0974	5.2098	4.8836	5.0309	5.1343	5.1819

$\delta = 0.20$  EUR (20 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	37.3599	37.3599	37.3599	37.3599	24.5085	24.5085	24.5085	24.5085
1	36.2510	30.3680	29.7287	29.4673	26.8256	31.4838	31.8711	31.8474
2	49.4079	44.8076	37.5068	32.1943	60.4037	56.0472	39.3286	33.8787
3	30.1036	43.1636	50.5767	55.2667	30.1453	55.0503	65.4426	69.8417
4	29.0450	24.2792	27.4535	34.8537	29.4290	27.2701	31.3452	39.3851
5	25.6843	24.3878	26.3501	28.1792	22.7121	25.8964	28.5219	30.0226

**Table A.8:** 1-step forecast negative log-likelihoods, EACD-FunXL, bid-quote durations, Linde.

$\delta = 0.01$  EUR (1 Tick)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	525.2768	525.2768	525.2768	525.2768	1413.6026	1413.6026	1413.6026	1413.6026
1	195.7372	207.2154	207.4348	215.0429	305.1424	351.0625	370.9741	401.2723
2	200.8595	210.0136	215.1555	216.2921	309.8952	313.6336	317.4923	314.4418
3	192.8245	208.1792	213.1704	224.5930	302.6874	313.4489	316.7142	324.1369
4	197.6982	204.2048	210.6150	214.2571	299.9014	305.0893	314.0129	308.2225
5	196.7881	205.8797	208.1212	215.7185	300.6993	298.9430	306.1110	314.7011

$\delta = 0.02$  EUR (2 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201

*Continued on next page*

Table A.8 – Continued from previous page

0	28.9570	28.9570	28.9570	28.9570	101.4032	101.4032	101.4032	101.4032
1	20.7777	21.5261	21.6877	21.8041	49.8157	59.8620	63.8068	66.0162
2	21.5460	22.0787	23.7480	23.7285	52.2479	44.0402	45.2661	45.6860
3	25.8711	26.1188	26.1721	24.2874	47.4167	46.9846	48.6573	46.2368
4	26.9474	26.3800	26.6571	26.8229	42.4920	44.8713	45.3428	45.1752
5	27.4611	28.4975	27.8333	27.2799	40.5651	41.2852	44.2171	44.8035

$\delta = 0.03$  EUR (3 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	24.6267	24.6267	24.6267	24.6267	77.7280	77.7280	77.7280	77.7280
1	21.0543	22.3655	22.4739	22.5149	52.7818	58.3250	59.9376	60.5456
2	22.6873	19.6835	20.4350	20.6815	55.6740	51.2873	52.7984	53.6609
3	26.3786	22.2762	22.4808	20.9670	50.6536	52.8633	55.5913	52.0816
4	26.7085	24.2579	23.8261	23.3063	41.7905	49.6224	49.1432	49.0689
5	25.7048	24.8729	25.4544	24.6324	38.9985	44.4366	46.7154	47.1737

$\delta = 0.04$  EUR (4 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	28.5780	28.5780	28.5780	28.5780	22.0163	22.0163	22.0163	22.0163
1	24.6614	26.3713	26.8384	26.8225	19.6127	19.5300	19.8755	19.8536
2	25.4609	25.6625	26.5752	27.2370	20.8221	18.5638	18.8311	19.0306
3	22.4884	23.7031	25.4002	24.0095	22.1201	21.2403	21.6892	21.4554
4	17.5658	23.2528	22.7994	22.6002	16.0429	19.6350	19.4242	19.0693
5	15.7546	24.0781	23.7958	23.4362	17.0737	17.2200	19.8164	18.8923

$\delta = 0.05$  EUR (5 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	15.3983	15.3983	15.3983	15.3983	21.9687	21.9687	21.9687	21.9687
1	13.7136	14.0220	14.0684	14.0334	17.6103	18.8468	19.2209	19.2657
2	13.9487	13.6764	14.0844	14.1567	18.0874	17.2030	17.7543	18.1752
3	16.5435	15.1184	14.8876	14.1080	20.3910	18.8871	18.2998	17.1210
4	16.2340	15.9240	15.2942	15.7887	18.6770	21.0448	19.9973	19.6900
5	16.8840	16.4470	15.7257	15.9943	19.7793	19.9884	20.9584	20.5544

Continued on next page

Table A.8 – Continued from previous page

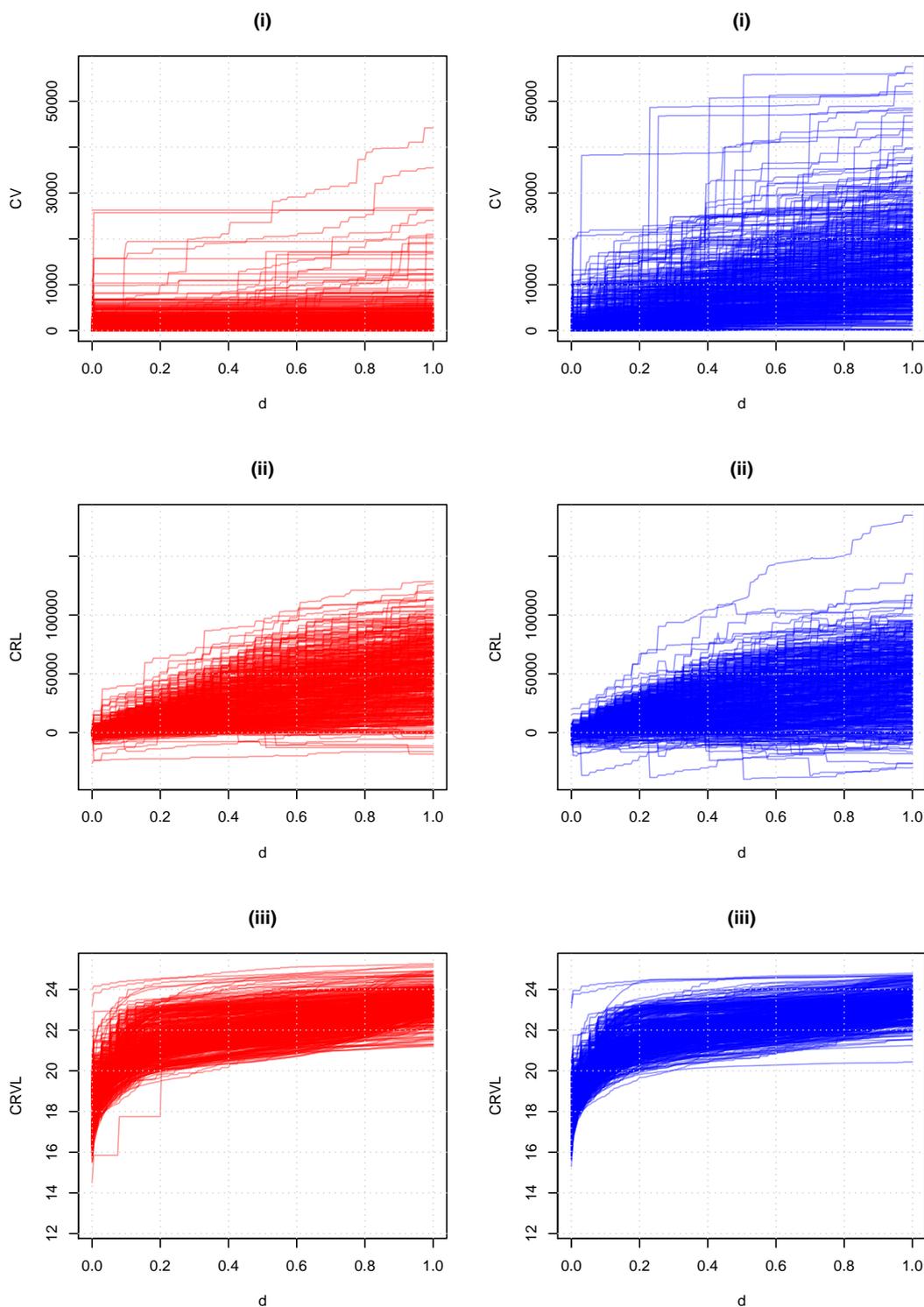
$\delta = 0.10$  EUR (10 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	22.8848	22.8848	22.8848	22.8848	56.4386	56.4386	56.4386	56.4386
1	14.5633	14.5801	14.6390	14.6539	43.3427	41.9081	41.5853	41.4188
2	15.8618	17.4962	19.4047	18.0145	48.5597	53.4011	60.2529	54.6936
3	14.6560	16.5633	17.7124	17.1940	39.0345	56.2466	55.1553	53.5709
4	13.1235	12.0970	11.6287	13.3530	39.7532	36.8111	36.4276	41.6271
5	11.3723	18.4836	12.5065	12.7178	33.1281	56.4665	35.0704	35.7900

$\delta = 0.20$  EUR (20 Ticks)

K	trained on days 1 to 10				trained on days 1 to 15			
	D=51	D=101	D=151	D=201	D=51	D=101	D=151	D=201
0	72.7945	72.7945	72.7945	72.7945	17.1619	17.1619	17.1619	17.1619
1	19.9910	19.2941	18.7423	18.1642	9.8586	5.9428	5.6036	5.5831
2	16.4248	15.3489	15.3027	16.0595	6.0241	6.7923	7.4523	7.1161
3	25.9401	14.7324	15.7002	16.3236	6.2674	7.3947	6.1170	6.0627
4	19.2920	25.2561	16.0749	14.9718	5.4159	6.9554	6.6788	7.3295
5	12.5037	42.9097	33.2648	33.9488	4.5610	6.3519	7.0543	8.3908

## A.5 Chapter 5: Realized measures of liquidity



**Figure A.1:** Top to bottom: (i) Initial liquidity, (ii) cumulative realized liquidity, and (iii) cumulative realized log variance of liquidity for the MunichRe stock. Each curve represents one day during 3 November 2008 to 30 December 2010 (incomplete trading days removed). Initial liquidity is strictly positive and strictly non-decreasing by construction. The drift is positive for most of the days, which is in line with the findings for the diurnal pattern in Chapter 3. The liquidity variance, computed based on 5-minute snapshots, almost exclusively varies in its level, i.e., CRVL curves appear to be almost parallel, which is also reflected by their eigenstructure.

A Additional empirical results

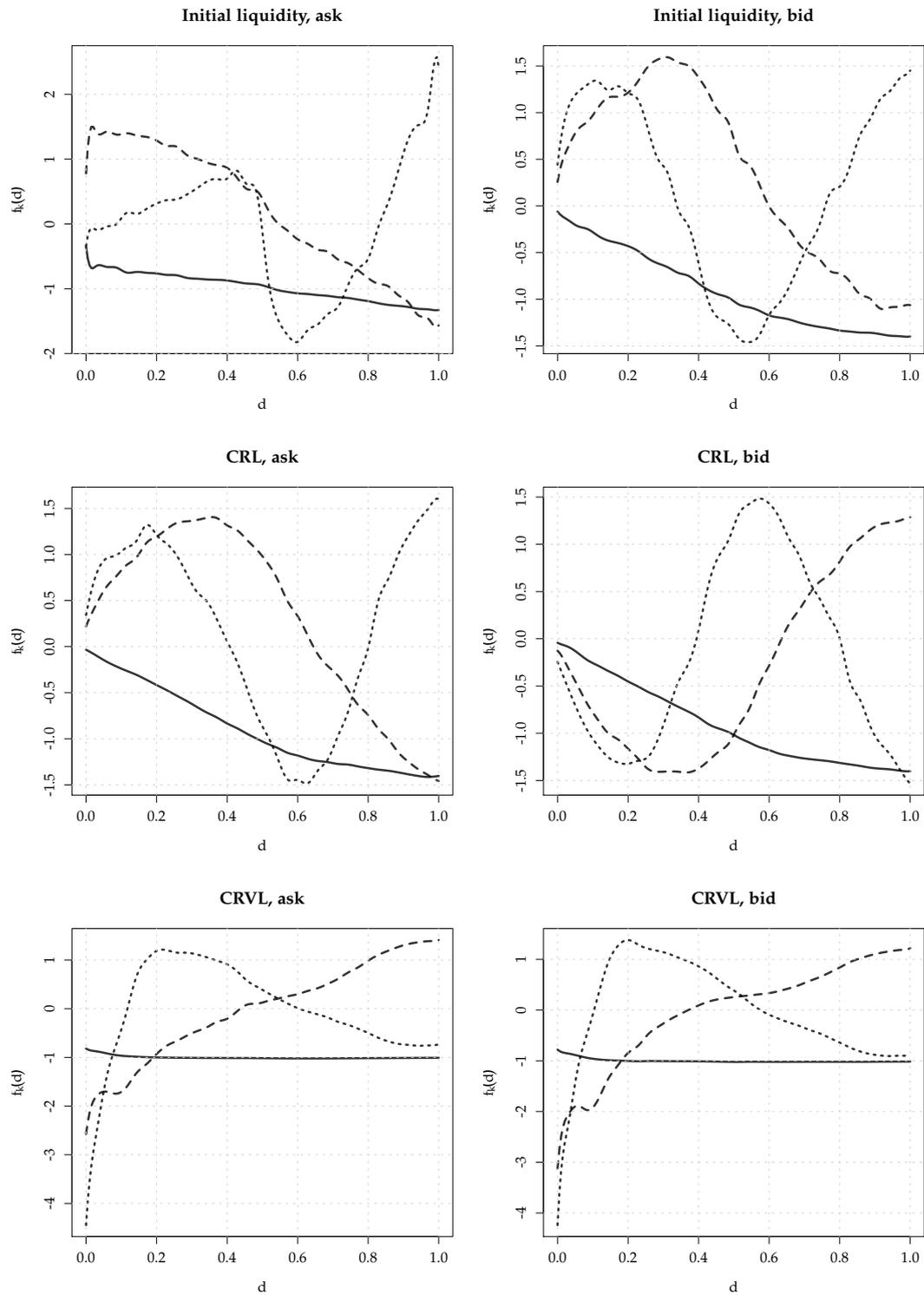
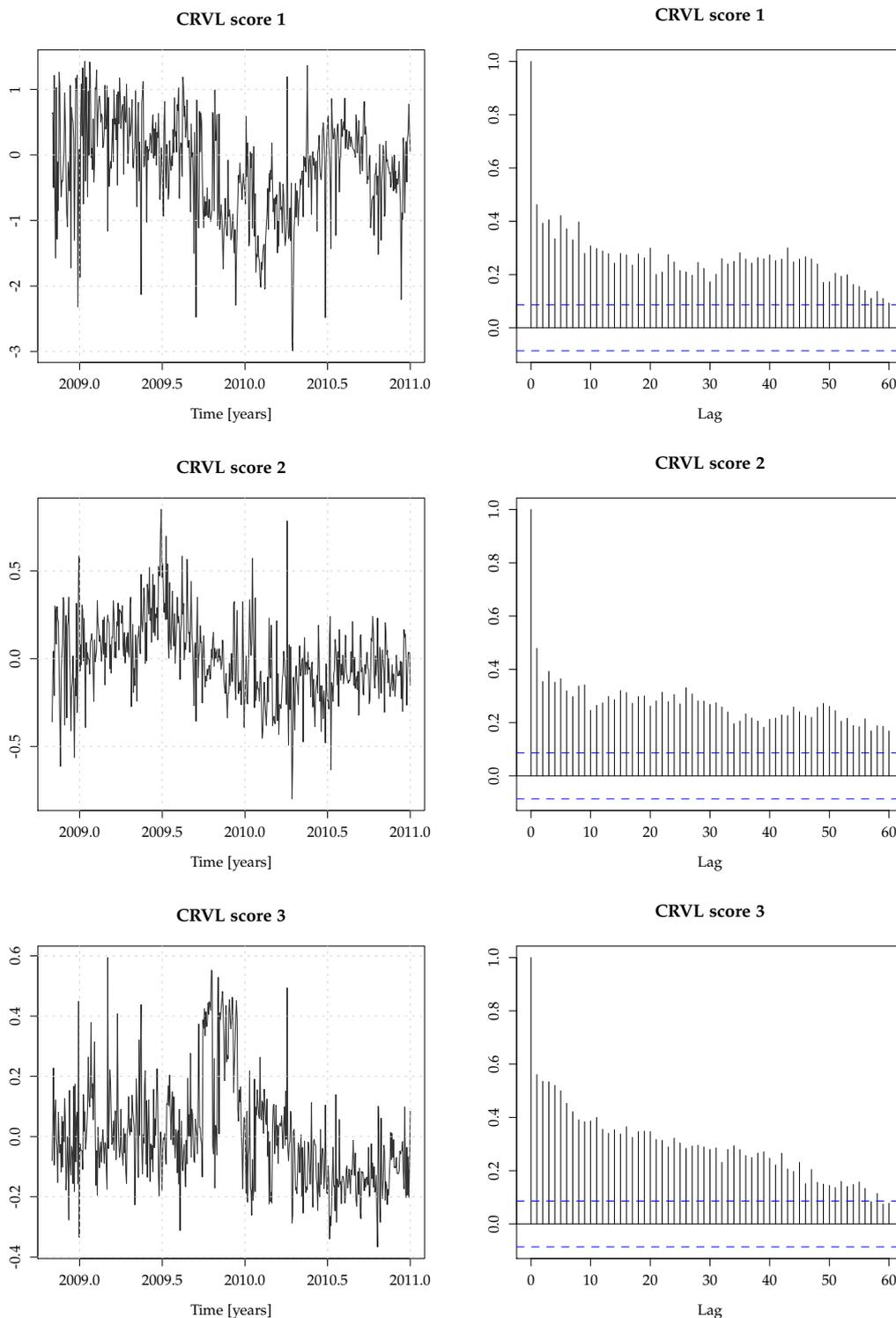
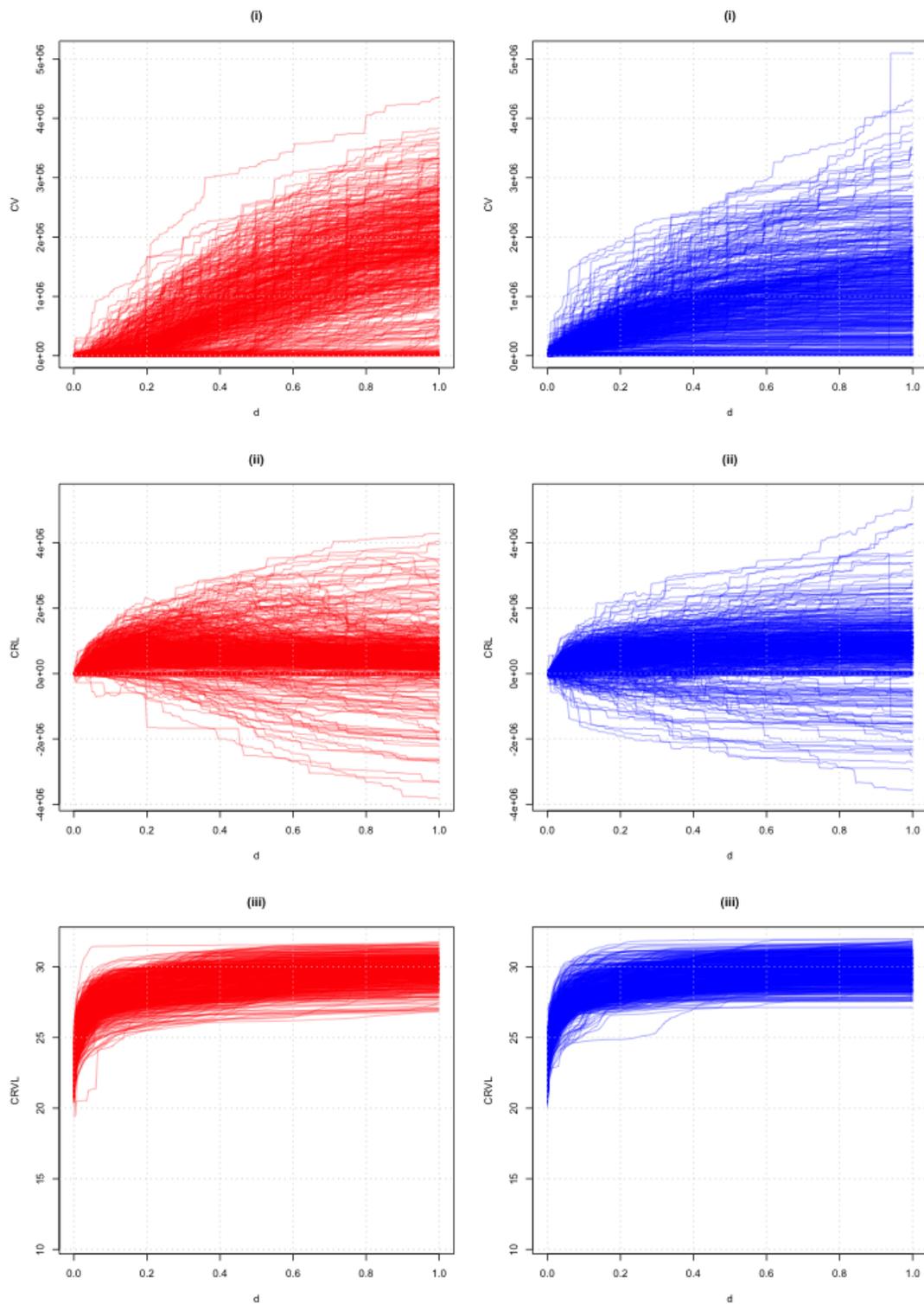


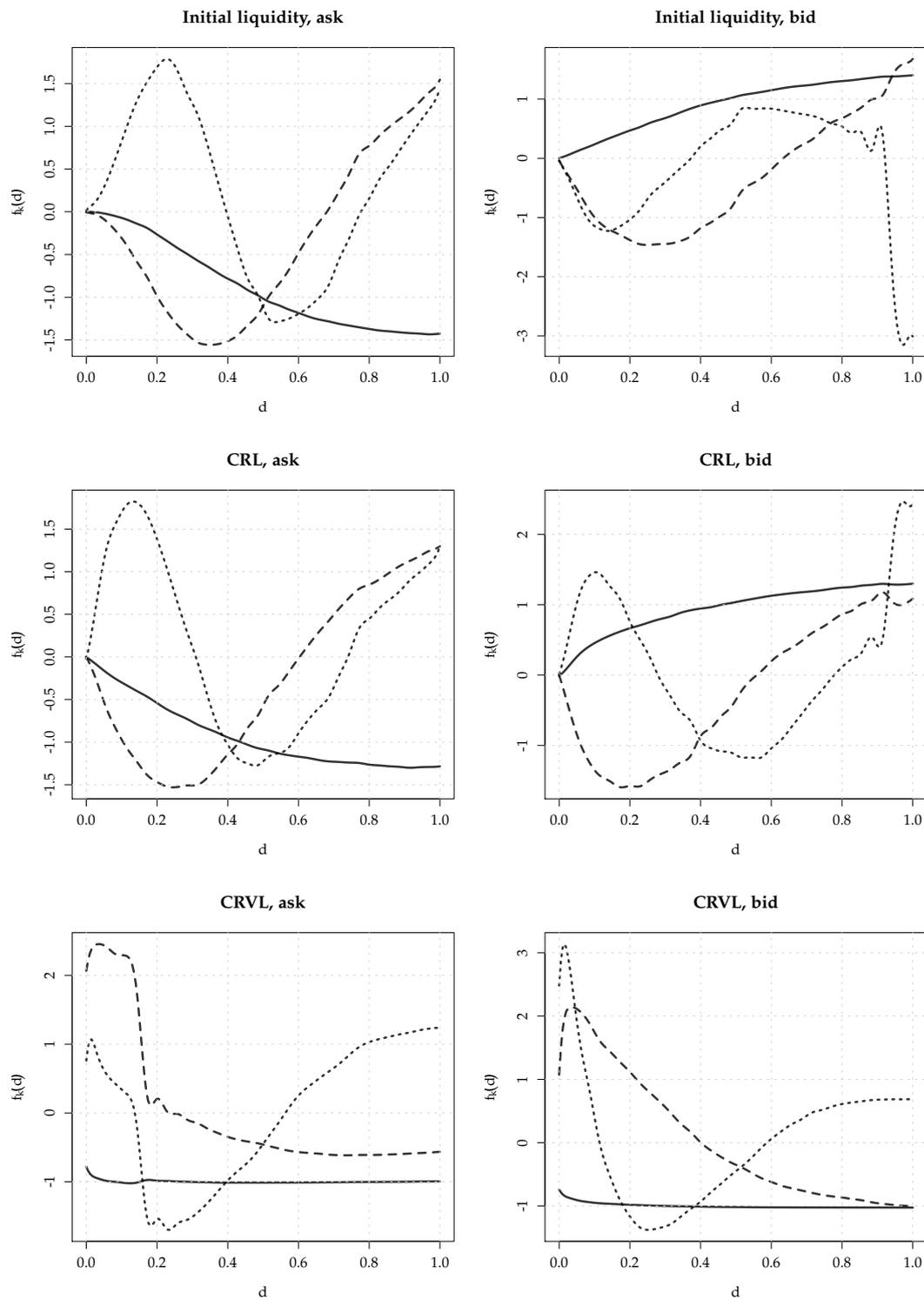
Figure A.2: First three estimated eigenfunctions for all six realized measures, MunichRe.



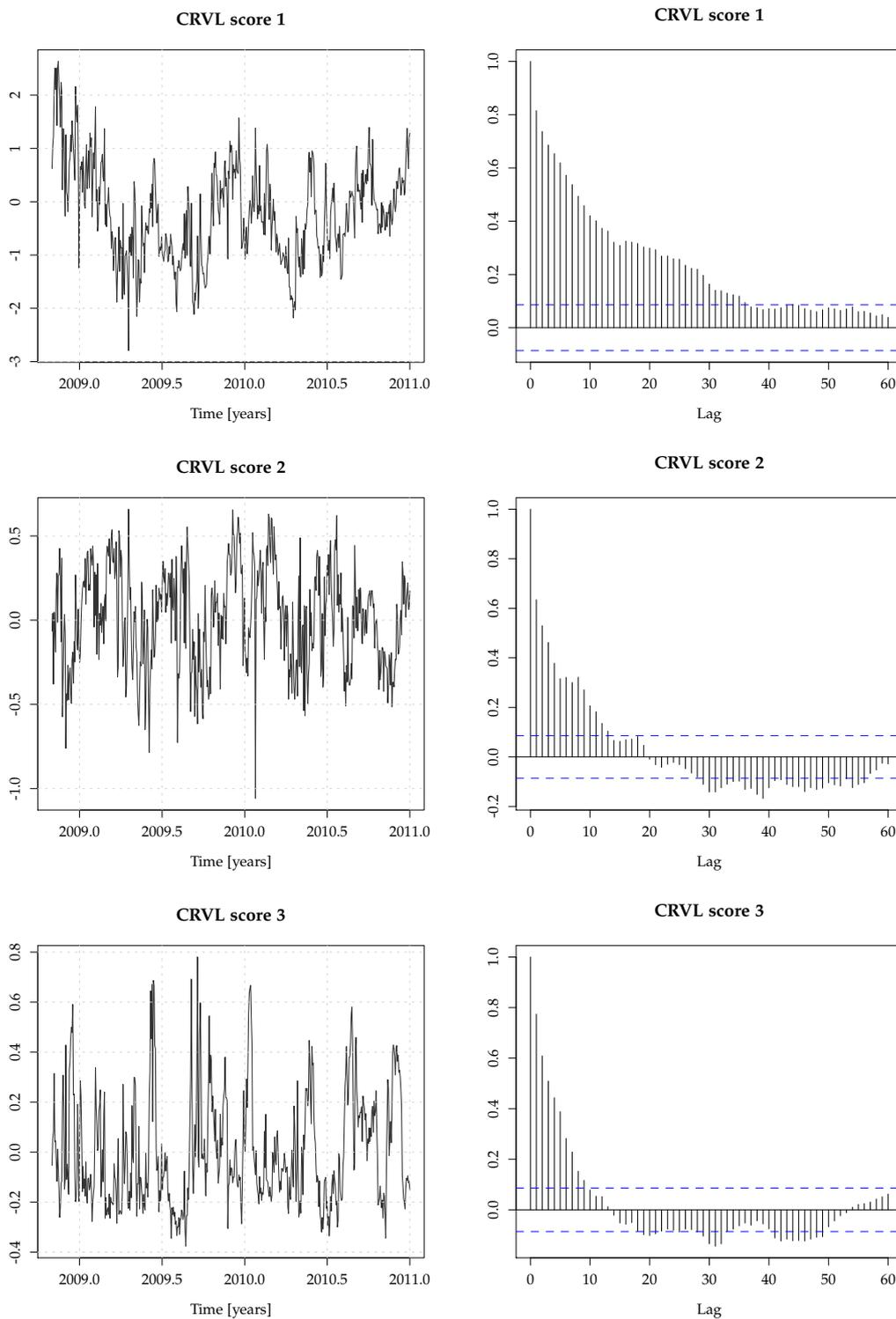
**Figure A.3:** First three FPC score series and their SACF for MunichRe's ask curves. The first series, which carries most of the information according to the eigenvalues analysis, is strongly serially dependent and quite persistent, indicating that liquidity variation itself has these properties as well.



**Figure A.4:** Top to bottom: (i) Initial liquidity, (ii) cumulative realized liquidity, and (iii) cumulative realized log variance of liquidity for the Commerzbank stock. Each curve represents one day between 3 November 2008 to 30 December 2010 (incomplete trading days removed). Initial liquidity is strictly positive and strictly non-decreasing by construction. The drift is positive for most of the days, which is in line with the findings for the diurnal pattern in Chapter 3. The liquidity variance, computed based on 5-minute snapshots, almost exclusively varies in its level, i.e., CRVL curves appear to be almost parallel, which is also reflected by their eigenstructure.



**Figure A.5:** First three estimated eigenfunctions for all six realized measures, Commerzbank.



**Figure A.6:** First three FPC score series and their SACF for Commerzbank's ask curves. The first series, which carries most of the information according to the eigenvalues analysis, is strongly serially dependent and quite persistent, indicating that liquidity variation itself has these properties as well.

# Eidesstattliche Versicherung

gemäß der Promotionsordnung vom 12. Juli 2011, §8 Abs. 2 Pkt. 5

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, 18. Juni 2015

---

Andreas Fuest