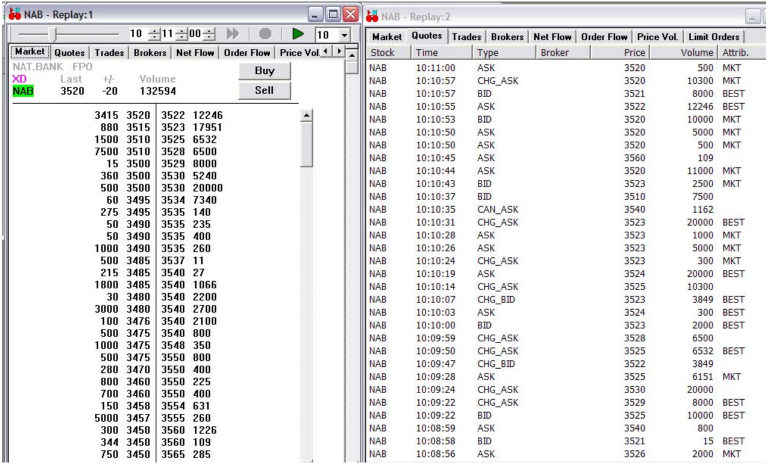
Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics

Wolfgang Härdle
Nikolaus Hautsch
Andrija Mihoci

Institute for Statistics and Econometrics
CASE - Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
http://ise.wiwi.hu-berlin.de

Snapshot of a Limit Order Book - ASX



3415	3520	3522	12246
880	3515	3523	17951
1500	3510	3525	6532
7500	3510	3528	6500
15	3500	3529	8000
360	3500	3530	5240
500	3500	3530	20000
60	3495	3534	7340
275	3495	3535	140
50	3490	3535	235
50	3490	3535	400
1000	3490	3535	260
500	3485	3537	11
215	3485	3540	27
1800	3485	3540	1066
30	3480	3540	2200
3000	3480	3540	2700
100	3476	3540	2100
500	3475	3540	800
1000	3475	3548	350
500	3475	3550	800
280	3470	3550	400
800	3460	3550	225
700	3460	3550	400
150	3458	3554	631
5000	3457	3555	260
300	3450	3560	1226
344	3450	3560	109
750	3450	3565	285

Stock	Time	Type	Broker	Price	Volume	Attrib.
NAB	10:11:00	ASK		3520	500	MKT
NAB	10:10:57	CHG_ASK		3520	10300	MKT
NAB	10:10:57	BID		3521	8000	BEST
NAB	10:10:55	ASK		3522	12246	BEST
NAB	10:10:53	BID		3520	10000	MKT
NAB	10:10:50	ASK		3520	5000	MKT
NAB	10:10:50	ASK		3520	500	MKT
NAB	10:10:45	ASK		3560	109	
NAB	10:10:44	ASK		3520	11000	MKT
NAB	10:10:43	BID		3523	2500	MKT
NAB	10:10:37	BID		3510	7500	
NAB	10:10:35	CAN_ASK		3540	1162	
NAB	10:10:31	CHG_ASK		3523	20000	BEST
NAB	10:10:28	ASK		3523	1000	MKT
NAB	10:10:26	ASK		3523	5000	MKT
NAB	10:10:24	CHG_ASK		3523	300	MKT
NAB	10:10:19	ASK		3524	20000	BEST
NAB	10:10:14	CHG_ASK		3525	10300	
NAB	10:10:07	CHG_BID		3523	3849	BEST
NAB	10:10:03	ASK		3524	300	BEST
NAB	10:10:00	BID		3523	2000	BEST
NAB	10:09:59	CHG_ASK		3528	6500	
NAB	10:09:50	CHG_ASK		3525	6532	BEST
NAB	10:09:47	CHG_BID		3522	3849	
NAB	10:09:28	ASK		3525	6151	MKT
NAB	10:09:24	CHG_ASK		3530	20000	
NAB	10:09:22	CHG_ASK		3529	8000	BEST
NAB	10:09:22	BID		3525	10000	BEST
NAB	10:08:59	ASK		3540	800	
NAB	10:08:58	BID		3521	15	BEST
NAB	10:08:56	ASK		3526	2000	MKT

Graphical Illustration

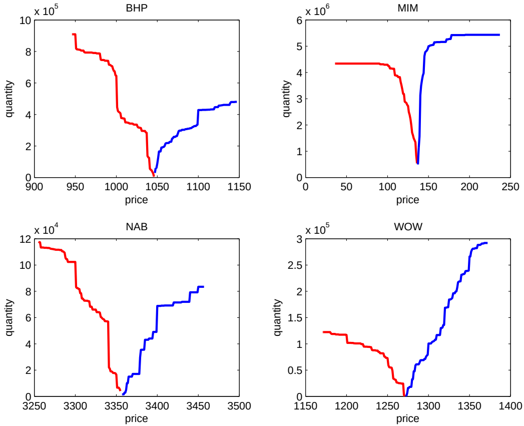


Figure 1: LOB for four stocks on the ASX, July 8, 2002, 10:15

Objectives

- Parsimonious statistical modelling of a limit order book
- Extracting common factors driving the book
- Understanding the dynamics of liquidity supply
- Analyzing the predictability of liquidity supply

Economic Motivation

- LOB displays instantaneous liquidity-induced transaction costs
- Shape of order book curves: marginal trading costs
- Optimal splitting strategies: transaction costs vs. liquidity risks
- Information content: LOB reflects market's expectation (Glosten, 1994, Bloomfield/OHara/Saar, 2002, Cao/Hansch/Wang, 2003)

Statistical Motivation

- Providing a flexible but unifying framework for orderbook modelling and forecasting
- Modelling approach: *smooth (non-parametrically) in space and parametrically in time*
- Dimension reduction: extraction of relevant common factors
- Time series properties of factors?

Outline

1. Motivation ✓
2. The Dynamic Semiparametric Factor Model (DSFM)
3. Data
4. In-Sample Fit
5. Out-of-Sample Forecasting
6. Conclusions

Notation

- t: time index,
- j: cross-sectional index, $j = 1, \ldots, J = 202$,
- $Y_{t,j}$: offered volume at time t at level j,
- $X_{t,j}$: limit price at time t at level j,
- L: number of underlying factors, $L \ll J$

The Dynamic Semiparametric Factor Model (DSFM)

- Orthogonal L-factor model of an observable J-dimensional random vector:

$$Y_{t,j} = m_{0,j} + Z_{t,1} m_{1,j} + \cdots Z_{t,L} m_{L,j} + \varepsilon_{t,j}$$

- $m(.) = (m_0, m_1, \ldots, m_L)^\top$ is a tuple of functions with $m_j : R^d \rightarrow R$ representing (time-invariant) factor loadings
- $Z_t = (1, Z_{t,1}, \ldots, Z_{t,L})^\top$ are the factors
- Including explanatory variables $X_{t,j}$:

$$Y_{t,j} = \sum_{l=0}^{L} Z_{t,l} m_l(X_{t,j}) + \varepsilon_{t,j} = Z_t^\top m(X_{t,j}) + \varepsilon_{t,j}$$

Principle of the DSF Model

$$Y_{t,j} = \sum_{l=0}^{L} Z_{t,l} m_l(X_{t,j}) + \varepsilon_{t,j} = Z_t^\top m(X_{t,j}) + \varepsilon_{t,j}$$

- Reducing the dimension of the process
- Nonparametric estimation of factor loadings
- Keeping the time structure
- Taking the structure of the high-dimensional object into account

Estimation: Series Estimator

$$Z_t^\top m(X) = \sum_{l=0}^{L} Z_{t,l} m_l(X_{t,j}) = \sum_{l=0}^{L} Z_{t,l} \sum_{k=1}^{K} a_{l,k} \psi_k(X) = Z_t^\top A \psi(X)$$

- $\psi(.) = (\psi_1, \ldots, \psi_K)^\top$ vector of basis functions, e.g. a tensor B-spline basis
- $A = (a_{l,k}) \in R^{(L+1)\times K}$ is a coefficient matrix
- K bandwidth parameter

Least Squares Estimation

$$
\begin{aligned}
(\widehat{Z}_t, \widehat{A}) &= argmin_{Z_t,A} S(A, Z) \\
&= argmin_{Z_t,A} \sum_{t=1}^{T} \sum_{j=1}^{J} \{Y_{t,j} - Z_t^\top A\psi(X_{t,j})\}^2 \\
\widehat{Z}_t &= (1, \widehat{Z}_{t,1}, \ldots, \widehat{Z}_{t,L})^\top \\
\widehat{A} &= (\widehat{a}_{l,k})_{l=0,\ldots,L;k=1,\ldots,K}
\end{aligned}
$$

- Minimization by Newton-Raphson algorithm

Identification Issues

The minimization problem has no unique solution. If $(\hat{Z}_t, \hat{A}_t)$ is a minimizer then also

$$(\tilde{B}^\top \hat{Z}_t, \tilde{B}^\top \hat{A})$$

is a minimizer. Here $\tilde{B}$ is an arbitrary matrix of the form

$$\tilde{B} = \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix}$$

for an invertible matrix B.

Inference

The differences in the inference based on $\hat{Z}_t$ instead of the (true unobservable) Z_t are asymptotically negligible (Borak et al, 2007).

This asymptotic equivalence carries over to estimation and testing procedures in the framework of fitting a VAR or VEC model.

Therefore it is justified to fit vector autoregressive model and proceed as if $\hat{Z}_t$ were observed.

The Data

- Four traded stocks at the Australian Stock Exchange (ASX) in 2002:
  - Broken Hill Proprietary Ltd. (BHP)
  - MIM
  - National Australia Bank Ltd. (NAB)
  - Woolworths Ltd.(WOW)
- Period covered: July 8, 2002 until August 23, 2002 (7 weeks, 35 trading days)
- 202 dimensional vector of price-volume pairs each minute for each stock

Trading Frequencies

	Recorded		Analyzed	
Stock	Total	Per day	Total	Per day
BHP	123281	3522	11258	322
MIM	27394	783	8339	238
NAB	86106	2460	10811	309
WOW	39127	1118	9272	265

Table 1: Number of transactions for selected stocks in the period under review

Data Preprocessing

- 330 vectors of quantities per day for selected stocks starting from 10:15 until 15:45
- 'Relative' prices were computed as deviations from the best bid and best ask price (respectively)
- Bid and ask side are modelled separately to obtain better fit around the inside quotes

Selection of K and L

Explained variance:

$$1 - RV(L) = 1 - \frac{\sum_t^T \sum_j^{J_t} \left\{ Y_{t,j} - \sum_{l=0}^{L} \hat{Z}_{t,l} \hat{m}_l \left(X_{t,j} \right) \right\}^2}{\sum_t^T \sum_j^{J_t} \left(Y_{t,j} - \bar{Y} \right)}$$

	BHP, BID		BHP, ASK	
L	$K = 15$	$K = 25$	$K = 15$	$K = 25$
1	0.937	0.939	0.953	0.955
2	0.976	0.978	0.977	0.979
3	0.985	0.987	0.983	0.986
4	0.988	0.990	0.986	0.989
5	0.989	0.992	0.988	0.990

In-Sample Parameterisation

- Identical parameterisation for bid and ask side
  - 2 dynamic factors ($L = 2$)
  - B-splines of order 2 (linear)
  - 15 knots ($K = 15$)

In-Sample Fit

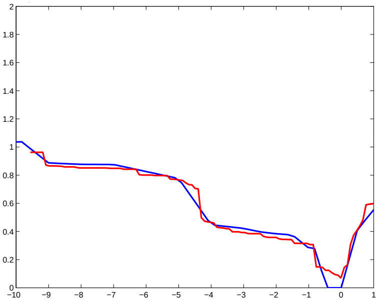


Figure 2: In-sample fit for BHP on July 12, 2002 (13:15)

Modelling the BID side: 1st Factor Loadings

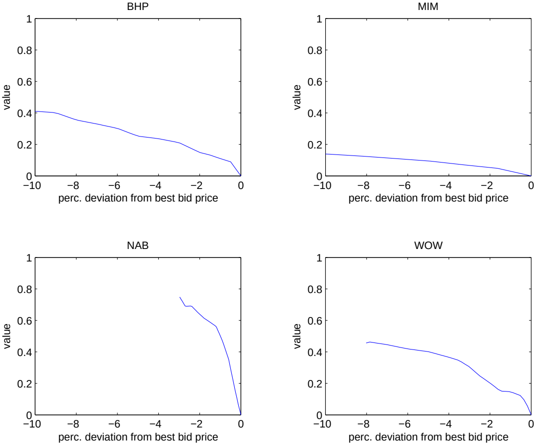


Figure 3: 1st factor loadings for the BID side

Modelling the BID side: 2nd Factor Loadings

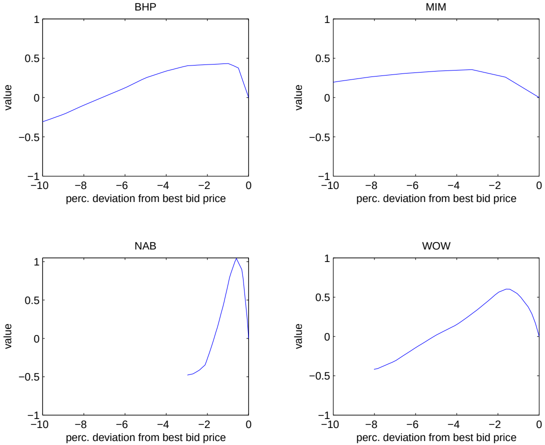


Figure 4: 2nd factor loading for the bid side

Modelling the BID side: 1st Factor

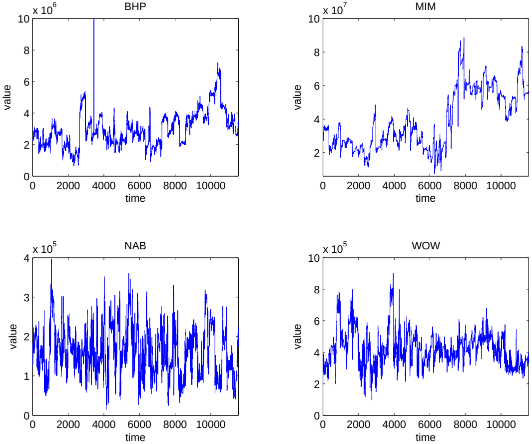


Figure 5: 1st factor for the bid side

Modelling the BID side: 2nd Factor

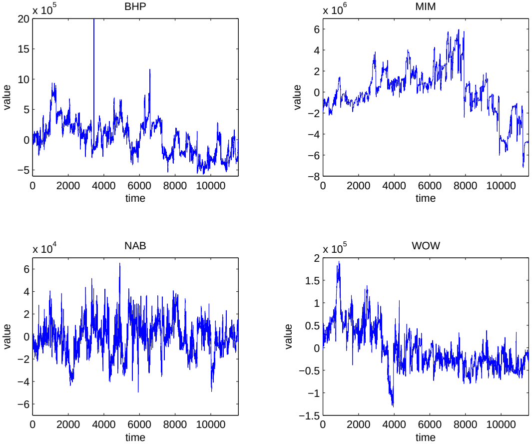


Figure 6: 2nd factor for the bid side

Modelling the ASK side: 1st Factor Loadings

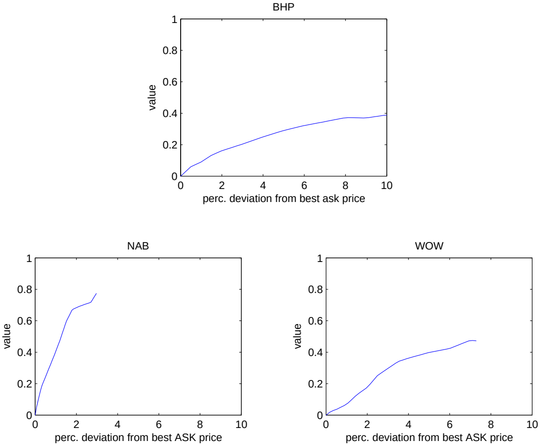


Figure 7: 1st factor loadings for the ask side

Modelling the ASK side: 2nd Factor Loadings

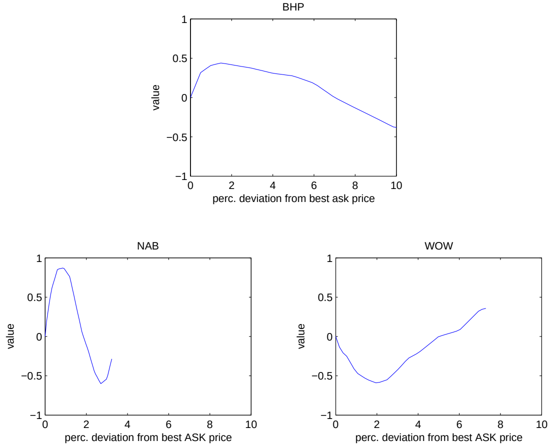


Figure 8: 2nd factor loadings for the ask side

Modelling the ASK side: 1st Factor

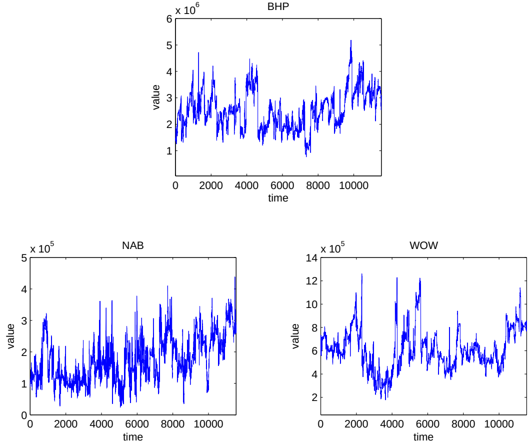


Figure 9: 1st factor for the ask side

Modelling the ASK side: 2nd Factor

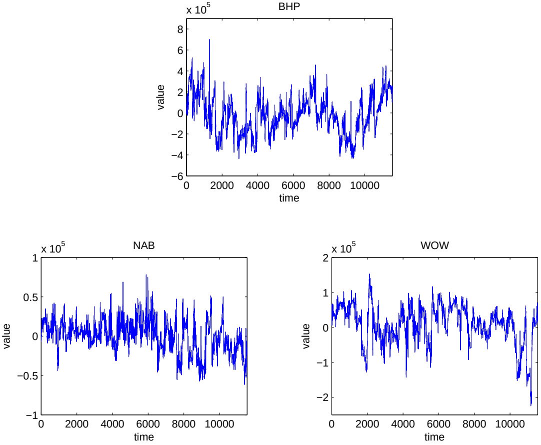


Figure 10: 2nd factor for the ask side

Time Series Properties of Factors

- First factor integrated
- Second factor mostly integrated
- For most periods first and second factors are cointegrated
- Evidence for GARCH effects

Estimated Parameters in the VECM

- Equation with estimated parameters:

$$\begin{bmatrix} \Delta Z_{1,t} \\ \Delta Z_{2,t} \end{bmatrix} = \begin{bmatrix} -0.022^{*} \\ 0.005^{*} \end{bmatrix} \cdot$$

$$\left\{ \begin{bmatrix} 1.000^{*} & 0.826 \end{bmatrix} \begin{bmatrix} Z_{1,t-1} \\ Z_{2,t-1} \end{bmatrix} + \begin{bmatrix} 2164110.119^{*} \end{bmatrix} \begin{bmatrix} const \end{bmatrix} \right\} +$$

$$+ \begin{bmatrix} -0.060 & -0.450 \\ 0.004 & -0.006 \end{bmatrix} \begin{bmatrix} \Delta Z_{1,t-1} \\ \Delta Z_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}$$

- Significant estimates are denoted by *

Out-of-Sample Forecasting Setup

- Model estimation based on past 1320 trading minutes (4-day period)
- Re-estimation and model selection for factor loadings every 15 minutes
- Forecast for every minute during the 15-minute interval
- Model selection:
  - ADF and KPSS test
  - Johansen trace test
  - BIC

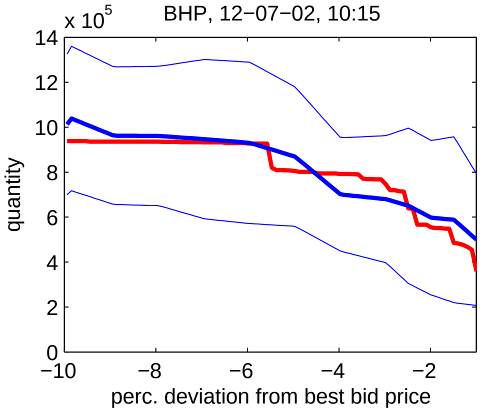


Figure 11: Forecasted LOB (blue) and observed LOB (red)

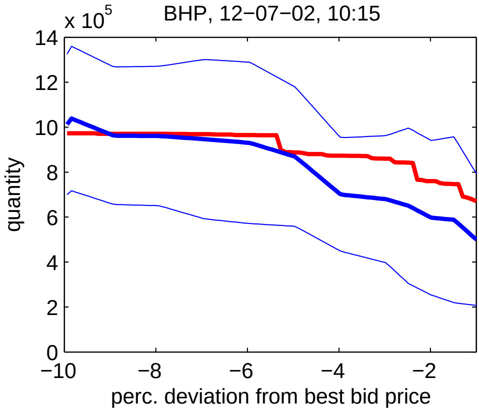


Figure 12: Forecasted LOB (blue) and naive forecast (red)

RMSPEs for DSFM and Naive Forecasts

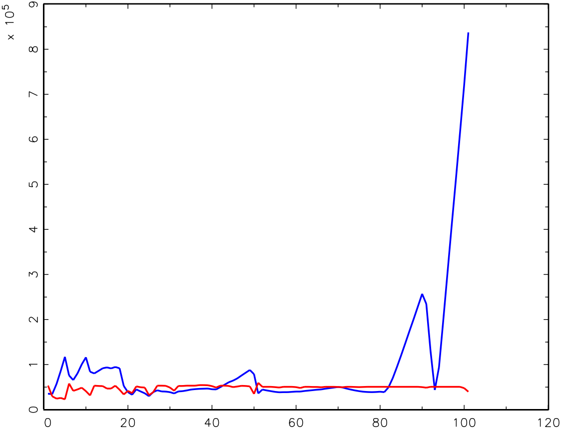


Figure 13: RMSPEs for DSFM forecasts (blue) and naive forecasts (red)

Conclusions

- 2 factors sufficient to model order book dynamics
- 1st factor captures slope
- 2nd factor captures curvature
- Order book factors are (co-)integrated
- DSFM-VEC based factors (partly) superior to naive forecast
- Confidence intervals for predicted liquidity are provided

Further Steps

- Linking liquidity factors to other variables (e.g. liquidity demand, volatility, PIN)
- Studying market elasticities
- Linking factors and loadings to execution risks and execution probabilities
- Studying liquidity risks, GARCH, 'default' risks
- Studying liquidity interdependencies between both sides of the market