Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics

Wolfgang Karl Härdle
Nikolaus Hautsch
Andrija Mihoci

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt–Universität zu Berlin
University of Vienna
Brandenburg University of Technology
lvb.wiwi.hu-berlin.de
homepage.univie.ac.at
b-tu.de
Motivation

Snapshot of a Limit Order Book (LOB)

Figure 1: Snapshot of a LOB for National Australia Bank Ltd. (NAB)

Modelling and Forecasting Liquidity Supply
Figure 2: Limit order book for NAB on July 8, 2002
Motivation

Objectives

- Modelling the LOB spatial and time structure using a dynamic factor model
  - Estimating and predicting factors and factor loadings
  - Understanding the dynamics of factor loadings
  - Impact of explanatory variables capturing the state of the market

- Forecasting demand and supply curves → liquidity supply
  - Extensive rolling window out-of-sample forecasting exercise
  - Forecasting evaluation against naive benchmark
  - Financial and economic applications
Statistical Challenges

- Require flexible framework for modelling and forecasting high-dimensional time-varying phenomenon
- Dimension reduction: extraction of common factors
- No obvious parametric model for factors
- Modelling philosophy: smooth in space and parametric in time
- Capturing dynamics by parametric multivariate TS model for factor loadings
Economic Implications

- LOB reflects liquidity supply on both sides of the market
- Information content: LOB reflects market’s expectation
- Shape of order book curves drives instantaneous trading costs for given volumes
- Predicting transaction costs yields implications for splitting strategies: transaction costs vs. liquidity risks
Applications

Example: Trading Strategy

An investor decides to buy (sell) certain number of NAB shares (10,000 or 20,000) over the course of a trading day, starting from 10:30 until 15:55.

Which execution strategy should the investor follow:

(i) Splitting the buy (sell) order proportionally over the trading day (i.e. every 5 minutes)

(ii) Placing one buy (sell) order at a time where the predicted transaction costs using the DSFM approach are minimal?
Research Questions

- Does the DSFM successfully model liquidity supply?
- Do factor loadings and quote dynamics follow a vector error correction (VEC) specification?
- Does our method outperform a naive forecasting benchmark and improves order execution strategies?
Outline

1. Motivation ✓
2. Limit Order Book Data
3. The Dynamic Semiparametric Factor Model (DSFM)
4. Modelling LOB Dynamics
5. Forecasting LOB Dynamics
6. Conclusions
Limit Order Book Data

The Data

- Limit order data from the Australian Stock Exchange (ASX)
  - Allows for complete reconstruction of the LOB at any time
  - Accounting for all LOB activities outside continuous trading

- Analyzing 4 stocks
  - Broken Hill Proprietary Ltd. (BHP)
  - National Australia Bank Ltd. (NAB)
  - MIM
  - Woolworths (WOW)
The Data

- **Australian Stock Exchange (ASX)**
  - Period covered: July 8 - August 16, 2002 (30 trading days)
  - Daily trading period: 10:15 - 15:55
  - LOB sampling frequency: 5 minutes
## Descriptive Statistics

<table>
<thead>
<tr>
<th>Orders</th>
<th>BHP</th>
<th>NAB</th>
<th>MIM</th>
<th>WOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limit orders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(i) buy (bid side)</td>
<td>50012</td>
<td>28850</td>
<td>9551</td>
<td>13234</td>
</tr>
<tr>
<td>(ii) sell (ask side)</td>
<td>32053</td>
<td>25953</td>
<td>6474</td>
<td>11318</td>
</tr>
<tr>
<td>Market orders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(i) buy</td>
<td>28030</td>
<td>16304</td>
<td>4115</td>
<td>7260</td>
</tr>
<tr>
<td>(ii) sell</td>
<td>16755</td>
<td>15142</td>
<td>2789</td>
<td>6464</td>
</tr>
</tbody>
</table>

Table 1: Number of orders from July 8 to August 16, 2002
Notation and Data Preprocessing

- Seasonally adjusted bid/ask side volume
  \[ Y_{b,t}^j = \tilde{Y}_{b,t}^j/s_t^b \in \mathbb{R}^{101} \text{ and } Y_{a,t}^j = \tilde{Y}_{a,t}^j/s_t^a \in \mathbb{R}^{101} \]

- Best bid/ask price: \( \tilde{S}_{t,101}^b, \tilde{S}_{t,1}^a \)

- Relative price deviations from best bid/ask quotes: \( S_{b,t}^j, S_{a,t}^j \)

- Capturing intraday seasonality using FFF approximation

  \[ s_t = \delta \cdot \bar{t} + \sum_{m=1}^{M} \left\{ \delta_{c,m} \cos (\bar{t} \cdot 2\pi m) + \delta_{s,m} \sin (\bar{t} \cdot 2\pi m) \right\}, \]

  where \( \bar{t} \in (0, 1] \).
Intraday Seasonalities in Liquidity Supply

Figure 3: Seasonal factors for quantities at $\tilde{S}_{t,101}^b$ (red) and $\tilde{S}_{t,1}^a$ (blue)
The Dynamic Semiparametric Factor Model (DSFM)

Orthogonal $L$-factor model of an observable $J$-dimensional random vector - Park et al. (2009), Fengler et al. (2007)

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

$m(\cdot) = (m_0, m_1, \ldots, m_L)^\top$ - tuple of functions

$m_l : \mathbb{R}^d \rightarrow \mathbb{R}$ - time-invariant factors

$Z_t = (1_T, Z_{t,1}, \ldots, Z_{t,L})^\top$ - factor loadings

Including explanatory variables $X_{t,j}$

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

Estimation

Efficient nonparametric method

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

\[
\psi(\cdot) = (\psi_1, \ldots, \psi_K)^\top - \text{basis functions (tensor B-spline basis)}
\]

\[
A = (a_{l,k}) \in \mathbb{R}^{(L+1) \times K} - \text{coefficient matrix}
\]

\[
\left( \hat{Z}_t, \hat{A} \right) = \arg \min_{Z_t, A} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\{ Y_{t,j} - Z_t^\top A \psi(X_{t,j}) \right\}^2
\]

Minimization by Newton-Raphson algorithm

Modelling and Forecasting Liquidity Supply
Implementation

Selection of $L$ and $K$

- Explained variance

\[ EV(L) = 1 - \frac{\sum_{t=1}^{T} \sum_{j=1}^{J} \{ Y_{t,j} - \sum_{l=0}^{L} \hat{Z}_{t,l} \hat{m}_{l} (X_{t,j}) \}^2}{\sum_{t=1}^{T} \sum_{j=1}^{J} \{ Y_{t,j} - \bar{Y} \}^2} \]

Statistical Inference

- Difference between $\hat{Z}_{t}$ and $Z_{t}$ can be asymptotically neglected
- TS models can be used for modelling of $\hat{Z}_{t}$
Modelling Liquidity Supply

- DSFM approaches
  - "Separated" approach - demand and supply separately, i.e. \( Y_{t,j}^b \in \mathbb{R}^{101} \) and \( Y_{t,j}^a \in \mathbb{R}^{101} \)
  - "Combined" approach - whole LOB, \( (-Y_{t,j}^b, Y_{t,j}^a) \in \mathbb{R}^{202} \)

- Explanatory variables, \( X_{t,j} \)
  - Relative price levels, \( S_{t,j}^b \) and \( S_{t,j}^a \)
  - Deseasonalized lagged 5 min buy/sell volume, \( Q_{t}^b \) and \( Q_{t}^s \)
  - Lagged 5 min log return and realized volatility, \( r_t \) and \( V_t = r_t^2 \)

Modelling and Forecasting Liquidity Supply
LOB Based on Relative Price Levels - Explained Variance

<table>
<thead>
<tr>
<th>Approach</th>
<th>BHP</th>
<th>NAB</th>
<th>MIM</th>
<th>WOW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bid side</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(i) Separated</td>
<td>0.964</td>
<td>0.965</td>
<td>0.996</td>
<td>0.975</td>
</tr>
<tr>
<td>(ii) Combined</td>
<td>0.921</td>
<td>0.936</td>
<td>0.975</td>
<td>0.914</td>
</tr>
<tr>
<td><strong>Ask side</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(i) Separated</td>
<td>0.941</td>
<td>0.948</td>
<td>0.953</td>
<td>0.959</td>
</tr>
<tr>
<td>(ii) Combined</td>
<td>0.930</td>
<td>0.912</td>
<td>0.951</td>
<td>0.948</td>
</tr>
</tbody>
</table>

Table 2: EV of the estimated LOB data from July 8 to August 16, 2002
Figure 4: True (solid) and estimated (dashed) LOB using the separated approach with two factors (EV ≈ 95%) on July 8, 2002
Modelling LOB Dynamics

Estimated LOB Factors

Figure 5: Estimated factors vs. relative price levels
Figure 6: Estimated factor loadings vs. relative price levels

The diagrams show the estimated factor loadings for BHP, NAB, MIM, and WOW over trading days. The loadings are plotted against relative price levels, with the y-axis representing the factor loadings and the x-axis representing the trading days.
Vector Error Correction (VEC) Specification

Engle and Patton (2004), Hautsch and Huang (2011)

\[ z_t = \left( \hat{Z}^{b}_{1,t}, \hat{Z}^{b}_{2,t}, \hat{Z}^{a}_{1,t}, \hat{Z}^{a}_{2,t}, \Delta \log \tilde{S}^{b}_{t,101}, \Delta \log \tilde{S}^{a}_{t,1} \right)^{\top} \]

\( \Delta \log \tilde{S}^{b}_{t,101} \) and \( \Delta \log \tilde{S}^{a}_{t,1} \) - best bid and ask price return

\[ z_t = c + \Gamma_1 z_{t-1} + \ldots + \Gamma_q z_{t-q} + \gamma \left( \log \tilde{S}^{b}_{t-1,101} - \log \tilde{S}^{a}_{t-1,1} \right) + \epsilon_t \]

Findings: BHP and WOW \( (q = 3) \), NAB \( (q = 2) \), MIM \( (q = 4) \)

- Strong own-process dynamics, weak cross-dependencies
- Quote changes are short-run predictable (up to 10-15 minutes)
Drivers of the Order Book Shape

<table>
<thead>
<tr>
<th>Variable</th>
<th>BHP</th>
<th>NAB</th>
<th>MIM</th>
<th>WOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bid side</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q^s_t$</td>
<td>10.37</td>
<td>8.17</td>
<td>5.41</td>
<td>6.31</td>
</tr>
<tr>
<td>$Q^b_t$</td>
<td>10.42</td>
<td>8.41</td>
<td>4.37</td>
<td>6.29</td>
</tr>
<tr>
<td>$r_t$</td>
<td>21.93</td>
<td>23.09</td>
<td>39.47</td>
<td>175.40</td>
</tr>
<tr>
<td>$V_t$</td>
<td>95.74</td>
<td>87.12</td>
<td>258.37</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: RMSE of the estimated LOB data from July 8, 2002 to August 16, 2002
Drivers of the Order Book Shape

<table>
<thead>
<tr>
<th>Variable</th>
<th>BHP</th>
<th>NAB</th>
<th>MIM</th>
<th>WOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask side</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q^s_t$</td>
<td>7.38</td>
<td>8.30</td>
<td>5.72</td>
<td>9.18</td>
</tr>
<tr>
<td>$Q^b_t$</td>
<td>7.30</td>
<td>8.42</td>
<td>7.22</td>
<td>8.88</td>
</tr>
<tr>
<td>$r_t$</td>
<td>18.00</td>
<td>22.13</td>
<td>45.54</td>
<td>236.08</td>
</tr>
<tr>
<td>$V_t$</td>
<td>78.62</td>
<td>63.63</td>
<td>192.87</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: RMSE of the estimated LOB data from July 8, 2002 to August 16, 2002
Drivers of the Order Book Shape

Figure 7: Estimated first factor of the bid (left) and the ask (right) side with respect to relative price levels and the past log traded sell (left) and buy (right) volume using the DSFM-Separated approach with two factors from 20020708 to 20020816
Figure 8: Responses of the best bid quote return to a one standard deviation shock in the estimated first bid factor loadings from July 8 to August 16, 2002. We employ the DSFM-separated approach with two factors. The response variable always enters the VEC specification in the first position. 95% confidence intervals are shown with dashed lines.
Figure 9: Responses of the best ask quote return to a one standard deviation shock in the estimated first ask factor loadings from July 8 to August 16, 2002. We employ the DSFM-separated approach with two factors. The response variable always enters the VEC specification in the first position. 95% confidence intervals are shown with dashed lines.
Forecasting Liquidity Supply

Setup
- 4 stocks, forecasting period: 20020722 - 20020816 (20 days)
- Forecasts for all 5 minute intervals until the end of a day

Strategies
- "DSFM-Separated" approach - estimated factors and factor loadings every 5 minutes (10 trading days)
- "Naive" approach - last observed LOB curve
Figure 10: True (solid) and forecasted LOB using the "DSFM-Separated" (dashed) and the "Naive" approach (black) on July 22, 2002
Figure 11: RMSPEs using DSFM (red and blue) and naive forecast (black) for all intervals during the day.
Applications

Example: Trading Strategy

An investor decides to buy (sell) certain number of shares (5 or 10 times the average best bid/ask volume) over the course of a trading day, starting from 10:30 until 15:55.

Which execution strategy should the investor follow:

(i) Splitting the buy (sell) order proportionally over the trading day (i.e. every 5 minutes)

(ii) Placing one buy (sell) order at a time where the predicted transaction costs using the DSFM approach are minimal?
Figure 12: Average percentage DSFM gains by reduced transaction costs compared to an equal-splitting strategy when buying and selling shares. Daily volumes: BHP (175,000), NAB (25,000), MIM (1,860,000) and WOW (50,000). Period covered: 20020722-20020816.
Figure 13: Average percentage DSFM gains by reduced transaction costs compared to an equal-splitting strategy when buying and selling shares. Daily volumes: BHP (350,000), NAB (50,000), MIM (4,650,000) and WOW (100,000). Period covered: 20020722-20020816.
Conclusions

(i) Modelling LOB Dynamics

- Two factors are sufficient to model LOB dynamics (slope, curvature) - explained variance approximately 95%
- Estimated factor loadings and quote dynamics follow a VEC specification
- Strong own-process dynamics, weak cross-dependencies
- The shape of the order book curves depends stronger on past trading volume than on past price movements or past volatility
(ii) Forecasting LOB Dynamics

- The DSFM approach outperforms a naive benchmark and a proportional trading strategy.
- Quote changes are short-run predictable (up to 10-15 minutes).
- Applications: improved order execution strategies.
- Demand and supply curves are modelled and forecasted successfully.
Modelling and Forecasting Liquidity Supply Using Semiparametric Factor Dynamics

Wolfgang Karl Härdle
Nikolaus Hautsch
Andrija Mihoci

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
University of Vienna
Brandenburg University of Technology
lvb.wiwi.hu-berlin.de
homepage.univie.ac.at
b-tu.de
References

Engle, R.F. and Patton, A.J.
*Impacts of trades in an error-correction model of quote prices*

Fengler, M. R., Härdle, W. and Mammen, E.
*A Dynamic Semiparametric Factor Model for Implied Volatility String Dynamics*

Hautsch, N. and Huang, R.
*The market impact of a limit order*
References

Park, B., Mammen, E., Härdle, W. and Borak, S.
*Time Series Modelling With Semiparametric Factor Dynamics*