Non-linear logit models for high frequency currency exchange data

N. Sazuka¹ & T. Ohira²

¹Department of Physics, Tokyo Institute of Technology, Japan
²Sony Computer Science Laboratories, Japan

Abstract

High frequency market data has become available with recent developments in computer technology. These data have some unique characteristics that do not appear in low frequency data. They are important in understanding financial markets. We present evidence of a unique property of high frequency data by proposing a new model. In this paper, we analyze tick-by-tick data, the most high frequency data available, of yen-dollar currency exchange rates. Focusing on the direction of up or down price movement, we show that a non-trivial structure exists in conditional probabilities of binarized data, which is apparently invisible from the price change itself. The probabilistic structure has a strong bias not only in the first order conditional probabilities but also in the higher order ones. Logit models are often applied to the analysis of the direction of price change. However, we found that a conventional logit model was not sufficient for our data due to its non-linear behavior. This motivated us to develop a new extended non-linear logit model to reproduce the binary probabilistic structure. This new model has overcome some of the shortcomings of the conventional analysis such as AR models and logit models. It can successfully show that the structure is such that it refers up to the previous few minutes by a model selection based on Akaike Information Criterion. The empirical result is consistent with dealers' perceptions that their strategies change slowly in the time scale of a few minutes. Our analysis here opens a possibility that this new non-linear logit model can be applied to a wide range of binary time series to extract their non-trivial probabilistic structures. Finally, in order to investigate the generality of our model, we are now analyzing the tick-by-tick GE data on NYSE, which is also one of the most active stocks.

Keywords: non-linear logit models, high frequency data, tick-by-tick data, the direction of price movement, conditional probabilities of binary data, probabilistic structure, Akaike Information Criterion.
1 Introduction

High frequency market data has become available with recent developments of computer technology [1]. High frequency data mirrors the dealers' actions with high resolution. It is now possible to follow the price change in real-time. The novelty of high frequency data has revealed distinct properties [1], which could not be detected by analysis of low frequency data. We present evidence of a unique property of high frequency data by proposing a new model.

2 Tick-by-tick data

We use two data sets of tick-by-tick "trade" data provided by Bloomberg (Ohira et al [2]) for the yen-dollar exchange rate for the period of 10/26/1998 to 11/30/1998 (data set A) and 1/4/1999 to 3/12/1999 (data set B) (fig. 1). The time series data sets are composed of values $Y(t)$ of yen value at "tick step" $t$. Note that $t$ is not real time, but rather discrete steps with variable time intervals at which the exchange took place. The data set A and B contains 267,398 and 578,509 data points respectively. On average two ticks are separated by 7 seconds.

![Figure 1: Time series plots of original data sets for A in (A) and B in (B).](image-url)
3 Binary data

We here binarized the data to extract the direction of up down movement of prices in the following way,
\[ X(t) = +1, \quad (Y(t) - Y(t-1) > 0), \]
\[ X(t) = -1, \quad (Y(t) - Y(t-1) < 0). \]

(1)

\[ X(t) = +1 \] and \( X(t) = -1 \) represent an up or down movement of prices, respectively. To focus on up down movement, we will disregard the cases, which the prices stay the same value between consecutive two ticks (\( Y(t+1) - Y(t) = 0 \)). With this reduction, the data sets A and B contain 145,542 and 344,791 data points, respectively, and the average time between ticks is about 10-13 seconds (Ohira et al [2]).

The analysis of the two sets of binary data shows surprisingly similar characteristics (Sazuka et al [3]), although no common features are apparently seen from the price change itself. First of all, the correlation functions for \( X(t) \) are almost the same (fig. 2). Secondly, the estimated parameters of a linear autoregression type stochastic model (AR model) for \( X(t) \) are very close to each other (fig. 3) (Sazuka et al [3]). Before making data binary, however, both properties are completely different for two data sets (fig. 4).

![Figure 2: Correlation functions C(k)=<X(t), X(t+k)> for binary data A (triangle) and binary data B (circle). They are almost overlapped.](image)
same direction, there is a greater tendency for the steps to continue to be in the same direction. This may be a reflection of dealers' behaviours which are trying to follow the market trend. We note the striking similarity of the two data sets in table 1, indicating that there is a probabilistic structure with non-linear trend in high frequency up down movement. Zhang looked at similar higher order statistics every half hour, and observed notable deviations for commodities like silver but not for yen-dollar trading (Zhang [4]). The result of other studies [5] mentioned the bias of the price movement between two consecutive ticks but not among multiple relations.

Table 1: The values of conditional probabilities computed from data sets A and B and from a simulation of the non-linear logit model. The number of training data for a simulation is 57,891 including the cases that the prices stay the same between consecutive ticks, which corresponds with the amount of data for a week. We compute the maximum likelihood estimate of the parameters using the first week data of Data B then simulate 20,000 points from the maximum likelihood model. The output of this model is estimated to be accurate to around 0.01.

<table>
<thead>
<tr>
<th></th>
<th>DataA</th>
<th>DataB</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of points</td>
<td>145,542</td>
<td>344,791</td>
<td>20,000</td>
</tr>
<tr>
<td>(P(+))</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>(P(+, +))</td>
<td>0.32</td>
<td>0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>(P(+, +, +))</td>
<td>0.27</td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td>(P(+, -))</td>
<td>0.66</td>
<td>0.66</td>
<td>0.67</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.28</td>
<td>0.23</td>
<td>0.27</td>
</tr>
<tr>
<td>(P(+, -))</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>(P(+, +))</td>
<td>0.33</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>(P(+, +, +))</td>
<td>0.26</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.36</td>
<td>0.31</td>
<td>0.35</td>
</tr>
<tr>
<td>(P(+, -))</td>
<td>0.63</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.32</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.26</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.25</td>
<td>0.21</td>
<td>0.24</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.72</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.67</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>(P(+, +, +, +))</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
</tbody>
</table>

At the tick level, a very notable probabilistic rule may exist in the up down movement of prices for yen dollar exchange rate as found in these data sets. If there is no time correlation in the data, no similarity should be observed even in the binary data. Therefore, the dynamical structure which we have found could indicate that dealers' decision making is based on a binary strategy, even if they
are unconscious of this fact. It could be useful to know this property for actual trading.

Figure 3: Estimated parameters fitted to an AR(4) process for binary data A (triangle) and binary data B (circle).

Figure 4: Before making data binary, both properties are completely different for two data sets. (Up) Correlation functions C(k)=[[X(t), X(t+k)] for differenced data A (triangle) and differenced data B (circle). (Down) Estimated parameters fitted to an AR(10) process for differenced data A (triangle) and differenced data B (circle).
4 Model

The logit model is known to be suitable for binary analysis \([6, 7]\). However, we have found that a conventional logit model was not sufficient for our data due to its non-linear behaviour. This motivated us to develop a new extended non-linear logit model to reproduce the binary probabilistic structure. The model of order \(k\) is defined by

\[
\ln \left[ \frac{P}{1-P} \right] = q_0 + \sum_{i_1=1}^{k} q_{i_1} X(t-i_1) + \sum_{i_1, i_2=1}^{k} q_{i_1 i_2} X(t-i_1) X(t-i_2) \\
+ \ldots + \sum_{i_1, i_2, \ldots, i_k=1}^{k} q_{i_1 i_2 \ldots i_k} X(t-i_1) X(t-i_2) \ldots X(t-i_k),
\]

(2)

where \(P=P(X(t)=+1|X(t-1),\ldots,X(t-k))\) is the probability to be +1 at time \(t\) given the previous \(k\) states in the past. The model of order \(k\) takes into account factors up to \(k\) times product of \(X\). We note that the probability \(P\) in the above eqn (2) is appropriately in the range of \([0,1]\).

This model overcomes the shortcomings of the AR model applied for binary data where corresponding variables are computed as non-binary. Our model also

![Figure 5: AIC values of the non-linear logit model for data A (triangle) in (A) and data B (circle) in (B). AIC values of the model have minimum value at the 5th order for both data sets.](image)
avoids another problem of the scale discrepancy between AR coefficients and noise amplitude in which the same models are chosen when both of them are scaled.

Once you decide on the model, the order selection has to be done. The model itself does not have a procedure to choose an appropriate order. However, there is an objective criterion for order selection called information criterion. We here customize one of the common criteria, AIC (Akaike Information Criterion) which gives the smallest possible order by minimizing a penalty function [8, 9]. The penalty function is defined by -2(Maximum log likelihood)+2(the number of the model parameters).

![AIC values of the AR model for data A (triangle) in (A) and data B (circle) in (B). It keeps decreasing as the order of the model grows, which means it cannot decide on which order is appropriate.](image)

AIC values of the model have minimum value at the 5th order for both data sets and the function shapes are similar (fig. 5). Therefore the non-linear logit model with reference to the previous 5 states is appropriate according to AIC. The 5th order model can also capture the probabilistic structure very well as you can see on the table 1. A period of the 5 states is roughly equivalent to few minutes in real time. This result is consistent with the dealers' perceptions, which is said that their strategy is slowly changing in the time scale of few minutes (Takayasu et al [10]).
On the other hand, AIC values of the AR model do not have a minimum value and keep decreasing as the order of the model grows, which means that many orders are needed to be modelled properly and it cannot decide on which order is appropriate (fig. 6). For the normal logit model instead, AIC values have a minimum value but not at the same order of the model for both of data set (fig. 7). It also does not capture the probabilistic structure, especially a tendency to continue to move in the same direction (i.e. \( P(+|+,+,+,+) > P(+|+,+,+) = P(+|+,+) \)). Our analysis with non-linear logit models has overcome this difficulty as well.

![Figure 7: AIC values of the normal logit model for data A (triangle) in (A) and data B (circle) in (B). AIC values have a minimum value but not at the same order of the model for both of data set.](image)

5 Conclusion

In this paper, we have analyzed tick-by-tick data, the most high frequency data available, of yen-dollar currency exchange market using a non-linear logit model. This new model has been able to capture the non-trivial probability structure in the binary currency exchange data, which was impossible using the conventional methods such as AR models or linear logit models. This is an indication that this new non-linear logit model is a useful tool for analyzing a wide range of binary time series with non-trivial dynamical structures. Applications to other time series to prove the model's potential are currently being explored.
Finally, in order to investigate the generality of our model, we are now analyzing the tick-by-tick GE data on NYSE, which is also one of the most active stocks. Interestingly, we have found a much strong bias in conditional probabilities of binary GE stock data. So we expect that the model would be useful in this case as well.

Acknowledgments

N.S. thanks Yoshiyuki Kabashima of Tokyo Institute of Technology (TIT) for his fruitful comments and suggestions, and Jun-ichi Inoue of Hokkaido University and A.C.C. Coolen, Giulia Iori, Peter Sollich of King's College, London, Toshiyuki Tanaka of Tokyo Metropolitan University and Hidetoshi Nishimori of TIT for stimulating discussions. Grant-in-Aid for JSPS Fellows (NS) is acknowledged.

References


