# MODELLING HIGH FREQUENCY FINANCIAL COUNT DATA

Shahiduzzaman Quoreshi

## Abstract

This thesis comprises two papers concerning modelling of financial count data. The papers advance the integer-valued moving average model (INMA), a special case of integer-valued autoregressive moving average (INARMA) model class, and apply the models to the number of stock transactions in intra-day data.

Paper [1] advances the INMA model to model the number of transactions in stocks in intra-day data. The conditional mean and variance properties are discussed and model extensions to include, e.g., explanatory variables are offered. Least squares and generalized method of moment estimators are presented. In a small Monte Carlo study a feasible least squares estimator comes out as the best choice. Empirically we find support for the use of long-lag moving average models in a Swedish stock series. There is evidence of asymmetric effects of news about prices on the number of transactions.

Paper [2] introduces a bivariate integer-valued moving average model (BINMA) and applies the BINMA model to the number of stock transactions in intra-day data. The BINMA model allows for both positive and negative correlations between the count data series. The study shows that the correlation between series in the BINMA model is always smaller than 1 in an absolute sense. The conditional mean, variance and covariance are given. Model extensions to include explanatory variables are suggested. Using the BINMA model for AstraZeneca and Ericsson B it is found that there is positive correlation between the stock transactions series. Empirically, we find support for the use of long-lag bivariate moving average models for the two series.

**Key words:** Count data, Intra-day, High frequency, Time series, Estimation, Long memory, Finance.

*To my parents*

## Acknowledgement

It has been a long journey and I am grateful to many people for their cooperation during this journey. From the first day, Kurt has been very helpful by providing invaluable information. In the process of writing this thesis he has not only shown endless patience but also been very prompt in reading, correcting and improving the papers. He has always been accessible and generous with time. Moreover, he encouraged and recommended me to visit University of California, Berkeley. I am deeply indebted to him for his excellent supervision. Thank you Kurt!

I am grateful to Mikael Bask and Jörgen Hellström for reading the paper(s) and making important comments. I am also grateful to a number of people for their comments on the paper(s) when presented at a seminar or a conference. Thank you Magnus Wikström, Jonas Nordström, Thomas Aronsson, Karl-Gustaf Löfgren, Torben G. Andersen, Tim Bollerslev and Robert Engle.

A number of people has made this journey easy and enjoyable. I am grateful to Roger Axelsson, for nice conversations and information regarding housing and schools for my kids, Olle Westerlund and Curt Löfgren for late night conversations, Marie Hammarstedt and Eva Cederblad for practical help. I am also grateful to Tomas Sjögren for his positive remarks during the courses. I would like to thank Magdalena, Tarek and Torbjörn for sharing the office room, discussions and cooperation. I would also like to thank Ola, Torbjörn D. and Mattias for discussions and other cooperation. I am also thankful to Kjell-Göran, Carina, Linda H., Linda A. and many others at the department.

Many thanks to the Faculty of Social Science and the Department of Economics for giving me opportunity to study one semester at UC, Berkeley. Special thanks to Kalle for the nice trip and other entertainment in California. I would like to thank David R. Brillinger, for his excellent teaching and supervision in the project work and not least for discussions about Ice-Hockey, and Michael Jansson for suggestions and practical help at Berkeley. I would also like to thank Krisanto, Jonas, Krister, Joe, Roshan and others who made my stay in Berkeley enjoyable.

I would like to thank my parents, brother, sisters, maternal grand father, aunt Shahnaz and friends. Special thanks to my sons, Arvin and Nivan, and my wife Sakira. Arvin has always been understanding, caring to Nivan and sacrificing his deserved time from his parents for Nivan. Arvin has understood that Nivan needs more attention due to his health. Sakira has helped me in many ways, for example by taking major responsibilities for kids and home. She deserves very special thanks for taking care of two kids alone while I was visiting Berkeley. Finally, I want to remember and to express my gratitude to uncle Parvez who submitted himself to and was murdered by the Pakistani army due to saving our lives during the war between Bangladesh and Pakistan.

Umeå, May 31, 2005

Shahiduzzaman Quoreshi

***The following two papers and a summary are included in this thesis:***

[I] Brännäs, K. and Quoreshi, S. (2004). Integer-Valued Moving Average Modelling of the Number of Transactions in Stocks. *Umeå Economic Studies* **637**.

[II] Quoreshi, S. (2005). Bivariate Time Series Modelling of Financial Count Data. *Umeå Economic Studies* **655**.

## 1. INTRODUCTION

In recent years, the scale of the activities in the financial arena, especially in stock and currency markets, have grown enormously. More people are engaged in trading, especially in electronic trade, in financial markets than ever before. The financial markets have also become a source of high frequency data. For obvious reasons, interest in understanding and forecasting future market developments have long been of interest. With high frequency data there are new ways of answering some of the old questions with respect to, e.g., volatility in stocks but also to posing new questions. Both new and old count data models have been employed to study, e.g., the number of traded stocks. For example, Rydberg and Shephard (1999) propose a time series model for the number of transactions made in intervals of a fixed length of time, while Engle and Russell (1998) propose a model for durations between irregularly spaced data.

Until now there is no study of pure time series models for count data in this area and this thesis contributes to filling this gap. A time series of count data is an integer-valued non-negative sequence of count observations observed at equidistant instants of time. There is a growing literature of various aspect of how to model, estimate and use such data. Jacobs and Lewis (1978ab, 1983) develop discrete ARMA (DARMA) models that introduce time dependence through a mixture process. McKenzie (1986) and Al-Osh and Alzaid (1987) introduce independently the integer-valued autoregressive moving average (INARMA) model for pure time series data, while Brännäs (1995) extends the model to incorporate explanatory variables. The regression analysis of count data is relatively new, though the statistical analysis of count data has a long and rich history. The increased availability of count data in recent years has stimulated the development of models for both panel and time series count data. For reviews of these and other models, see, e.g., Cameron and Trivedi (1998, ch. 7) and McKenzie (2003). In INARMA, the parameters are interpreted as probabilities and hence restricted to unit intervals. Some empirical applications of INARMA are due to Blundell, Griffith and Windmeijer (2002), who studied the number of patents in firms, Rudholm (2001), who studied competition in the generic pharmaceuticals market, and Brännäs, Hellström and Nordström (2002), who estimated a nonlinear INMA(1) model for tourism demand.

In this thesis, we focus on advancing and employing an integer-valued moving average model of order $q$ [INMA($q$)], i.e. a special case of the INARMA model class, for analyzing high frequency financial data in the form of stock transactions data aggregated over one or five minute intervals of time. Each transaction refers to a trade between a buyer and a seller in a volume of stocks for a given price. Besides volume and price, a transaction is impounded with other information like, e.g., spread, i.e., the difference between bid and ask price. Later, we propose a bivariate integer-valued moving average (BINMA) model. The BINMA is employed for the same type of data. A description of high frequency data, the INMA model and the BINMA model is given below.

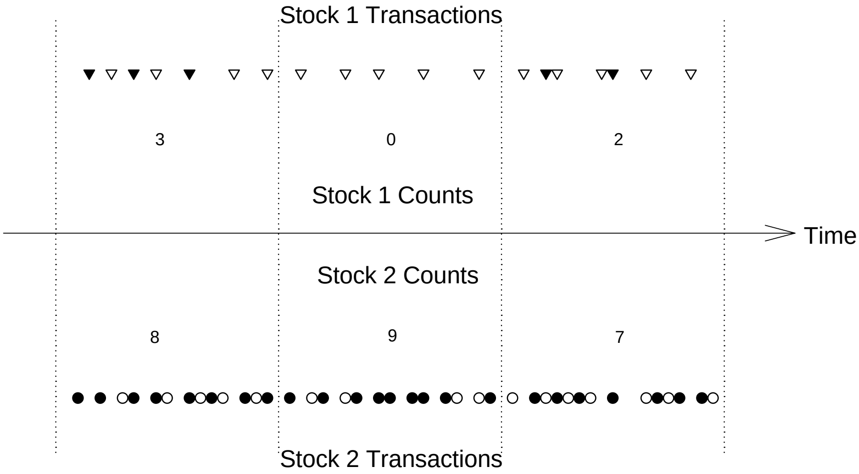

Figure 1: An illustration of how transactions data are generated. The black triangles and circles represent transactions for stock 1 and stock 2, respectively, while the white triangles and circles represent all other activities in an order book. The stock counts record the number of black triangles/circles falling into a time interval, i.e. falling between vertical lines.

## 2. HIGH FREQUENCY DATA

Financial market data are tick-by-tick data. Each tick represents a change in, e.g., a quote or corresponds to a transaction. For a liquid stock or a currency, these tick-by-tick data generate high frequency data. Such financial data are also characterized by lack of synchronization, in the sense that only exceptionally there is more than one transaction at a given instant of time. For reviews of high frequency data and their characteristics, see, e.g., Tsay (2002, ch. 5), Dacorogna et al. (2001) and Gourieroux and Jasiak (2001, ch. 14). The access to high frequency data is getting less and less of a problem for individual researchers and costs are low. As a consequence, many issues related to the trading process and the market microstructure are under study.

Transactions data are collected from an electronic limited order book for each stock. Incoming orders are ranked according to price and time of entry and are continuously updated. Hence, new incoming buy and sell orders and the automatic match of the buy and sell orders are recorded. The automatic match of a buy and a sell order generates a transaction. In Figure 1, we see that the transactions in the two stocks are not synchronized, i.e. the transactions appear at different points of time. The counts in the intervals are the number of transactions for corresponding intervals. In paper [1] a one minute time

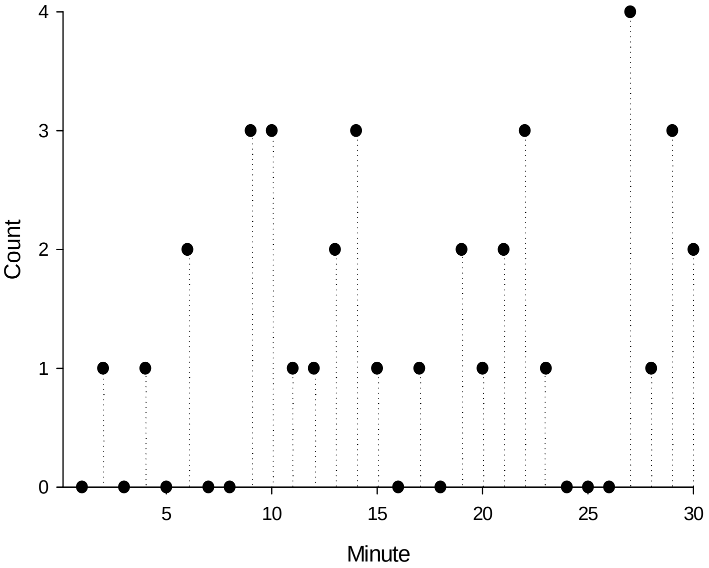

Figure 2: The number of transactions data over minute intervals for 30 minutes of trading in AstraZeneca.

scale is employed and for paper [2] a five minute scale. The collection of the number of transactions over a time period makes up a time series of count data. The time series of transactions or count data are synchronized between stocks in the sense that all the numbers of transactions are aggregated transactions over the same time interval. An example of real transactions data over a 30 minute period for the stock AstraZeneca is exhibited in Figure 2. Each observation number corresponds to one minute of time. This type of data series comprises frequent zero frequencies and motivates a count data model.

For this thesis the Ecovision system is utilized. Daily downloads are stored to files and count data are calculated from the tick-by-tick data using Matlab programs.

## 3. THE INMA AND BINMA MODELS

The INMA model is a special case of the INARMA model. The INMA model of order $q$, INMA($q$), is introduced by Al-Osh and Alzaid (1988) and in a slightly different form by McKenzie (1988). The single thing that most visibly makes the INMA model different from its continuous variable MA counterpart is that multiplication of variables with real valued parameters is no longer a viable operation, when the result is to be integer-valued. Multiplication is therefore replaced by the binomial thinning operator

$$\alpha \circ u = \sum_{i=1}^{u} v_i,$$

where $\{v_i\}_{i=1}^{u}$ is an iid sequence of $0-1$ random variables, such that $\Pr(v_i=1)=\alpha=1-\Pr(v_i=0)$. Conditionally on the integer-valued $u$, $\alpha \circ u$ is binomially distributed with $E(\alpha \circ u|u)=\alpha u$ and $V(\alpha \circ u|u)=\alpha(1-\alpha)u$. Unconditionally it holds that $E(\alpha \circ u)=\alpha\lambda$, where $E(u)=\lambda$, and $V(\alpha \circ u)=\alpha^2\sigma^2+\alpha(1-\alpha)\lambda$, where $V(u)=\sigma^2$. Obviously, $\alpha \circ u$ take an integer-value in the interval $[0,u]$.

Employing this binomial thinning operator, an INMA($q$) model can be written

$$y_t = u_t + \beta_1 \circ u_{t-1} + \ldots + \beta_q \circ u_{t-q}$$

with $\beta_i \in [0,1]$, $i=1,\ldots,q-1$, and $\beta_q \in (0,1]$. Brännäs and Hall (2001) discuss model generalizations and interpretations resulting from different thinning operator structures, and an empirical study and approaches to estimation are reported by Brännäs et al. (2002). McKenzie (1988), Joe (1996), Jørgensen and Song (1998) and others stress exact distributional results for $y_t$, while we emphasize in paper [1] only the first two conditional and unconditional moments of the model. Moreover, we discuss and introduce more flexible conditional mean and heteroskedasticity specifications for $y_t$ than implied by the above equation. There is an obvious connection between the introduced count data model and the conditional duration model of, e.g., Engle and Russell (1998) in the sense that long durations in a time interval correspond to a small count and vice versa. Hence, a main use of the count data models discussed here is also one of measuring reaction times to shocks or news.

In paper [2], we focus on the modelling of bivariate time series of count data that are generated from stock transactions. The used data are aggregates over five minutes intervals and computed from tick-by-tick data. One obvious advantage of the introduced model over the conditional duration model is that there is no synchronization problem between the time series.[1] Hence, the spread of shocks and news is more easily studied in the present framework. Moreover, the bivariate count data models can easily be extended to multivariate models without much complication. The introduced bivariate time series count data model allows for negative correlation between the counts and the integer-value property of counts is taken into account. With this model we can model the stock series that also may have a long memory property. Moreover, this model is capable of capturing the conditional heteroskedasticity.

A large number of studies have considered the modelling of bivariate or multivariate count data assuming an underlying Poisson distribution (e.g., Gourieroux, Monfort and Trognon, 1984). Heinen and Rengifo (2003) introduce multivariate time series count data models based on the Poisson and the double Poisson distribution. Other extensions to traditional count data regression models are considered by, e.g., Brännäs and Brännäs (2004) and Rydberg and Shephard (1999).

[1] For a bivariate duration model the durations for transactions typically start at different times and as a consequence measuring the covariance between the series becomes intricate.

## 4. SUMMARY OF THE PAPERS

### *Paper [1]: Integer-valued Moving Average Modelling of the Number of Transactions in Stocks*

The integer-valued moving average model is advanced to model the number of transactions in intra-day data of stocks. The conditional mean and variance properties are discussed and model extensions to include, e.g., explanatory variables are offered. Least squares and generalized method of moment estimators are presented. In a small Monte Carlo experiment we study the bias and MSE properties of the CLS, FGLS and GMM estimators for finite-lag specifications, when data is generated according to an infinite-lag INMA model. In addition, we study the serial correlation properties of estimated models by the Ljung-Box statistic as well as the properties of forecasts one and two steps ahead. In this Monte Carlo study, the feasible least squares estimator comes out as the best choice. However, the CLS estimator which is the simplest to use of the three considered estimators is not far behind. The GMM performance is weaker than that of the CLS estimator. It is also clear that the lag length should be chosen large and that both under and overparameterization give rise to detectable serial correlation.

In its practical implementation for the time series of the number of transactions in Ericsson B, we found both promising and less advantageous features of the model. There is evidence of asymmetric effects of news about prices on the number of transactions. With the CLS estimator it was relatively easy to model the conditional mean in a satisfactory way in terms of both interpretation and residual properties. It was more difficult to obtain satisfactory squared residual properties for the conditional variance specifications that were tried. The FGLS estimator reversed this picture and we suggest that more empirical research is needed on the interplay between the conditional mean and heteroskedasticity specifications for count data. Depending on research interest the conditional variance parameters are or are not of particular interest. For studying reaction times to shocks or news it is the conditional mean that matters, in much the same way as for conditional duration models. In addition, the conditional variance has no direct ties to, e.g., risk measures included in, e.g., option values or portfolios.

### *Paper [2]: Modelling A Bivariate Time Series of Financial Count Data*

This study introduces a bivariate integer-valued moving average model (BINMA) and applies the BINMA model to the number of stock transactions in intra-day data. The BINMA model allows for both positive and negative correlations between the count data series. The conditional mean, variance and covariance are given. The study shows that the correlation between series in the BINMA model is always smaller than 1 in an absolute sense. Applying the BINMA model for the number of transactions in Ericsson B and AstraZeneca, we find promising and less promising features of the model. The conditional mean, variance and covariance have successfully been estimated. The standardized residuals based on FGLS are serially uncorrelated. But the model could not eliminate the serial correlation in the squared standardized residual series that is not of particular interest in this study. Further study is required to eliminate such serial correlation. One way of eliminating serial correlation may be to use extended model by letting, e.g., $\lambda_j$ or $\sigma_j$ be time-varying. Alternatively, by introducing non-diagonal $\mathbf{A}$ matrices as in (8), we could allow for an asymmetric flow of news from say Ericsson B to AstraZeneca but not the other way.

## REFERENCES

Al-Osh, M.A. and Alzaid, A.A. (1987). First Order Integer-Valued Autoregressive (INAR(1)) Process. *Journal of Time Series Analysis* **8**, 261-275.

Al-Osh M. and Alzaid A. (1988). Integer-Valued Moving Average (INMA) Process. *Statistical Papers* **29**, 281-300.

Blundell, R., Griffith, R. and Windmeijer, F. (2002). Individual Effects and Dynamics in Count Data Models. *Journal of Econometrics* **108**, 113-131.

Brännäs, K. (1995). Explanatory Variables in the AR(1) Model. *Umeå Economic Studies* **381**.

Brännäs, K. and Brännäs, E. (2004). Conditional Variance in Count Data Regression. *Communication in Statistics; Theory and Methods* **33**, 2745-2758.

Brännäs, K. and Hall, A. (2001). Estimation in Integer-Valued Moving Average Models. *Applies Stochastic Models in Business and Industry* **17**, 277-291.

Brännäs, K., Hellström, J. and Nordström, J. (2002). A New Approach to Modelling and Forecasting Monthly Guest Nights in Hotels. *International Journal of Forecasting* **18**, 19-30.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

Dacorogna, M.M., Gencay, R., Muller, U., Olsen, R.B. and Pictet, O.V. (2001). *An Introduction to High-Frequency Finance.* San Diego: Elsevier.

Engle, R.F. and Russell, J.R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica* **66**, 1127-1162.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Application to Poisson Models. *Econometrica* **52**, 701-720.

Gourieroux, C and Jasiak, J. (2001). *Financial Econometrics.* Princeton: Princeton University Press.

Heinen, A. and Rengifo, E. (2003). Multivariate Modelling of Time Series Count Data: An Autoregressive Conditional Poisson Model. CORE Discussion Paper 2003/25, Université Catholique de Louvain.

Jacobs, P.A. and Lewis, P.A.W. (1978a). Discrete Time Series Generalized by Mixtures I: Correlational and Runs Properties. *Journal of the Royal Statistical Society* **B40**, 94-105.

Jacobs, P.A. and Lewis, P.A.W. (1978b). Discrete Time Series Generalized by Mixtures II: Asymptotic Properties. *Journal of the Royal Statistical Society* **B40**, 222-228.

Jacobs, P.A. and Lewis, P.A.W. (1983). Stationary Discrete Autoregressive Moving Average Time Series Generated by Mixtures. *Journal of Time Series Analysis* **4**, 19-36.

Joe, H. (1996). Time Series Models with Univariate Margins in the Convolution-Closed Infinitely Divisible Class. *Journal of Applied Probability* **33**, 664-677.

Jørgensen, B. and Song, P.X-K. (1998). Stationary Time Series Models with Exponential

Dispersion Model Margins. *Journal of Applied Probability* **35**, 78-92.

McKenzie, E. (1986). Autoregressive Moving-Average Processes with Negative Binomial and Geometric Marginal Distributions. *Advances in Applied Probability* **18**, 679-705.

McKenzie, E. (1988). Some ARMA models for Dependent Sequences of Poisson Counts. *Advances in Applied Probability* **20**, 822-835.

McKenzie, E. (2003). Discrete Variate Time Series. In Shanbhag, D.N. and Rao, C.R. (eds.) *Handbook of Statistics*, Volume 21, pp. 573-606. Amsterdam: Elsevier Sciences.

Rudholm, N. (2001). Entry and the Number of Firms in the Swedish Pharmaceuticals Market. *Review of Industrial Organization* **19**, 351-364.

Rydberg, T.H. and Shephard, N. (1999). BIN Models for Trade-by-trade Data. Modelling the Number of Trades in a Fixed Interval of Time. Working Paper Series W23. Nuffield College, Oxford.

Tsay, R.S. (2002). *Analysis of Financial Time Series.* New York: John Wiley & Sons, Inc.