Analysis of financial data using a Difference-Poisson Autoregressive Model

by
Hiba Baroud

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Actuarial Science

Waterloo, Ontario, Canada, 2011

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Box and Jenkins methodologies have massively contributed to the analysis of time series data. However, the assumptions used in these methods impose constraints on the type of the data. As a result, difficulties arise when we apply those tools to a more generalized type of data (e.g. count, categorical or integer-valued data) rather than the classical continuous or more specifically Gaussian type. Papers in the literature proposed alternate methods to model discrete-valued time series data, among these methods is Pegram's operator (1980).

We use this operator to build an AR(p) model for integer-valued time series (including both positive and negative integers). The innovations follow the differenced Poisson distribution, or Skellam distribution. While the model includes the usual AR(p) correlation structure, it can be made more general. In fact, the operator can be extended in a way where it is possible to have components which contribute to positive correlation, while at the same time have components which contribute to negative correlation. As an illustration, the process is used to model the change in a stocks price, where three variations are presented: Variation I, Variation II and Variation III. The first model disregards outliers; however, the second and third include large price changes associated with the effect of large volume trades and market openings.

Parameters of the model are estimated using Maximum Likelihood methods. We use several model selection criteria to select the best order for each variation of the model as well as to determine which is the best variation of the model. The most adequate order for all the variations of the model is $AR(3)$. While the best fit for the data is Variation II, residuals' diagnostic plots suggest that Variation III represents a better correlation structure for the model.

Acknowledgements

I would like to first thank God and the Virgin Mary for their blessings especially during challenging times.

I thank my parents, Sami and Nada and my brother, Charbel for their loving care and support.

I am thankful for Wassim, his unconditional love and encouragement constantly influence my work. I am grateful for all my friends in Lebanon and Canada, I am blessed to be always surrounded by them.

I thank my supervisor, Professor Keith Freeland for introducing me to the topic and assisting me to achieve a deep knowledge of the material. His understanding, guidance and insights helped me a lot in my thesis. I would also like to thank the members of the Statistics and Actuarial Science department, especially Mary Lou Dufton and Professors Yulia Gel and Brian Hartman.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General Overview

The classic tools used in modeling and analyzing time series data are the Box and Jenkins methodologies (Box & Jenkins, 1976). They presented a rich class of models for continuous distributed time series. Unfortunately those methods are not applicable to discrete-valued time series data because of the constraints imposed on the marginal distribution. As a result, a lot of research has been done to model discrete-valued time series.

Many discrete-valued time series are built with an operator. A number of these operators have specific assumptions leading to some restrictions while others can be used to build models with fewer restrictions. One of the main restrictions is the marginal distribution which depends on the type of discrete data, whether it is binary, count, integer or categorical. Another restriction focuses on the resemblance of the process to Box and Jenkins' which would affect the interpretation of the model as well as the expressions derived from it and the properties of the model. Developing new models and extending previous models and operators to fit integer-valued data has received great attention in the past two decades.

This research uses Pegram's mixing operator (Pegram, 1980) in order to develop an integer-valued stationary AR(p) process. This operator is known for its flexibility regarding the marginal distribution as well as its similitude to Box and Jenkins' methodologies. For the marginal distribution of the innovations we will use the Skellam distribution which is the difference between two Poisson random variables. An important feature of this distribution is its symmetry which has been missing in almost all the integer-valued time series models to date appearing in the literature. This particular feature has many advantages and allows us to construct processes covering both positive and negative correlations while maintaining the similarities with the traditional Box-Cox ARIMA models.

1.2 Thesis Layout

This thesis is composed of six chapters of which the first is the introduction. In chapter 2, we review the models developed so far in the literature and discuss their advantages and drawbacks. The model in its three variations is defined in detail in chapter 3 with a derivation of its properties. In chapter 4 we look at the estimation method and the model selection criteria. Score functions are derived to find Maximum Likelihood estimates for the parameters and several model selection criteria are discussed in this chapter. The application of the model is presented in chapter 5 and we conclude in chapter 6 with a summary of the findings and suggest some recommendations to improve and further extend the model.

Chapter 2

Literature Review

2.1 Introduction

This literature review is an overview of some of the models that were used previously to fit discrete-valued time series data. Due to the large number of models, we narrowed our focus in this review to models fitting similar forms of data and models from which our model has been expanded. Both advantages and disadvantages of each methodology are discussed next to the model's summary and outcomes.

As mentioned before, Box and Jenkins' models cover continuous-valued time series data but are not able to model discrete-valued time series. Therefore, the latter has been studied from different perspectives in the literature. We first look at several discrete-valued time series models built using different types of mixing schemes then we examine closely the binomial thinning operator along with the different models developed using this particular operator. A brief definition of each model/operator is included in the review along with a summary of its specification, advantages and drawbacks. We conclude the review by a brief definition of Pegram's operator on which our model is built and a discussion on models that appeared in the literature that are aimed at fitting a similar type of data.

2.2 Background

2.2.1 Discrete-Valued Time Series Models

Among several methods used to create discrete-valued time series models and more specifically, binary autoregressive processes, we mention "Addition mod 2" which is a stochastic operator proposed by Kanter (1975). Given a set of independent 0 - 1 random variables and k^{th} order Markov Chain with state space $\{0, 1\}$, the operator assigns different probabilities to different values of the process. Therefore if X_n is the set of independent binary random variables with $P(X_n = 1) = \beta$ and Y_n is the k^{th} Markov process after n periods, then

$$Y_n = \begin{cases} X_n & \text{with a probability } p_0 \\ X_n \oplus Y_{n-j} & \text{with a probability } p_j \text{ for } 1 \leq j \leq p \end{cases} \tag{2.1}$$

In this case $\oplus$ denotes "Addition mod 2". Kanter shows that this stochastic operator generates a stationary process. Morever, he looks into its covariance structure and derives several theorems that are deemed useful in the estimation and interpretation of the parameters under different values of beta ($\beta = 0.5$, $\beta < 0.5$, $\beta > 0.5$). Although it is well defined for the case of independent Bernoulli trials as well as for a more general case, this process cannot be applicable to categorical data as the operator used can only be manipulated arithmetically.

Jacobs & Lewis (1978a) consider a mixture as well to model discrete-valued time series. The special case of their model is binary too. However, the approach is different than Kanter's whereby they use a probabilistic linear model to build the process. Also, the set of independent random variables is spread over a wider discrete space and follows any discrete distribution. The model is derived given both a specified marginal distribution and a correlation structure and is called discrete mixed autoregressive moving average (DARMA). Therefore if X_n is the set of discrete independent random variables having each a certain distribution, the stationary process of dependent random variables is obtained using several other independent sequences. Let U_n and W_n be two independent sequences of independent 0 - 1 random variables such that $P(W_n = 1) = \alpha$ and $P(U_n = 1) = \beta$ for $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$. Also, let Z_n be a sequence of i.i.d. random variables taking values 0,1,2,...,N. Then Y_n, which is the stationary process of dependent random variables, will be constructed under the following model, for $n = 1, 2, ...$:

$$\begin{aligned} Y_n &= W_n X_{n-Z_n} + (1 - W_n) S_{n-(N+1)} \\ S_n &= U_n S_{n-1} + (1 - U_n) X_n \end{aligned} \tag{2.2}$$

This model is called $DARMA(1, N + 1)$. Properties and inferences of this model are drawn in Jacobs & Lewis (1978b). Later, Jacobs and Lewis extended the special case of this model which is $DAR(1)$. In a third paper (Jacobs & Lewis, 1978c), they construct an autoregressive process for representing a stationary sequence of discrete random variables with p-th order Markov dependence. The process is called mixed p-th order discrete autoregressive ($DAR(p)$). Similarly to $DARMA(1, N + 1)$, it is specified by a marginal distribution and a correlation structure and defined using several sequences of independent random variables. Let X_n be a sequence of i.i.d. random variables and Y_n a sequence of independent 0 - 1 random variables such that $P(Y_n = 1) = 1 - P(Y_n = 0) = \beta$ for $0 \leq \beta < 1$. In addition, let V_n be a sequence of independent random variables such that $P(V_n = i) = \alpha_i$ where $i = 1, 2, ..., p$, $\alpha_i \geq 0$, and $\sum_{i=1}^{p} \alpha_i = 1$. Then for $n = 1, 2, ...$:

$$Z_n = Y_n Z_{n-V_n} + (1 - Y_n) X_n \tag{2.3}$$

In the above definition, Z_n is called $DAR(p)$. Jacobs and Lewis extended the $DARMA(1, N+1)$

to develop $DARMA(p, N+1)$ and $NDARMA(p, N)$ and discussed some of their properties (Jacobs & Lewis, 1983). Note that, the structure of the models developed by Jacobs and Lewis allows for the time indices to be a sequence of independently and identically distributed random variables. This representation is not appropriate for modeling time series observed at deterministic time points. As a result, the model interpretation becomes problematic and the process differs from Box and Jenkins ARMA structure.

2.2.2 Thinning Operator

The thinning operation is one of the most popular means of building discrete-valued time series models. More specifically, the binomial thinning operator has been used and extended to construct several types of integer-valued processes. The thinning operator originates from the discussion on self-decomposability and stability of discrete-valued time series in Steutel and van Harn (1979) and is defined as follows:
Let X be a non-negative integer-valued random variable; then for any $\alpha \epsilon [0, 1]$ the operator is defined by

$$\alpha \circ X = \sum_{i=1}^{X} Y_i \tag{2.4}$$

where "$\circ$" is the thinning operator and Y_i is a sequence of i.i.d. Bernoulli random variables such that $P(Y_i = 1) = 1 - P(Y_i = 0) = \alpha$. Note that $0 \circ X = 0$, $1 \circ X = X$, and $E(\alpha \circ X) = \alpha E(X)$.

Poisson Integer-Valued processes

The thinning operator has been used in many discrete-valued time series models having different discrete marginal distributions. One example is the first-order integer-valued autoregressive (INAR(1)) process discussed in Al-Osh & Alzaid (1987). The model uses the binomial thinning operator to construct the discrete version of AR(1) process which is defined as follows

$$X_t = \alpha \circ X_{t-1} + \epsilon_t \tag{2.5}$$

where $\alpha \in [0, 1]$ and ϵ_t is a sequence of uncorrelated non-negative integer-valued random variables having mean μ and finite variance σ^2. In the traditional Box-Jenkin's AR(1) model, the current observation, X_t, is made up of the fraction of the previous observation, αX_{t-1}, plus a random innovation, ϵ_t. Since the sample space is made of real numbers, the lagged term is simply X_{t-1} multiplied by α, where α represents the correlation coefficient and is usually restricted to $-1 < \alpha < 1$ (Box & Jenkins, 1976). However, when the sample space is made of integers, the approach under the thinning operator is to replace the scalar multiplication, αX_{t-1}, with the operation defined in (2.4), $\alpha \circ X_{t-1}$, then add the random innovation ϵ_t. The purpose of the operator is to "thin out" X_{t-1}. For example, in the case of binomial thinning, the operator is defined as follows; given X_{t-1}, $\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} B_{it}$, where $B_{1t}, B_{2t}, ..., B_{X_{t-1}t}$ are i.i.d. Bernoulli random variables with $P(B_{it} = 1) = 1 - P(B_{it} = 0) = \alpha$. It follows that $\alpha \circ X_{t-1}$ has a binomial distribution with X_{t-1} trials and probability of "success" α since it is a sum of X_{t-1} Bernoulli random variables. The process is also referred to as a birth-death process in which the number of people at time t, i.e. X_t, is the number of survivals represented by $\alpha \circ X_{t-1}$ with a probability of survival α added to the number of newborn at time t given by the innovation term ϵ_t. Correlation structure and distributional properties are discussed in Al-Osh & Alzaid (1987) and shown similar to the Gaussian $AR(1)$. In addition, several methods of parameter estimation are presented such as Maximum Likelihood Estimation (MLE). An illustration of the different methods of parameter estimation was given when the innovation term followed a Poisson distribution. Inference on this particular process (Poisson INAR(1)) is developed in Freeland & McCabe (2004a) in which new expressions of the score function and the information matrix are derived which led to the construction of new types of residuals. As a result, several ways of assessing the model were derived in this paper including formal tests and graphical analyses. Freeland and McCabe used the conditional probability of X_t given X_{t-1} to derive the expressions mentioned earlier. The conditional probability is given by the convulotion of the two random components, $\alpha \circ X_{t-1}$ and ϵ_t, and is defined by

$$p(X_t|X_{t-1}) = \sum_{s=0}^{min(X_t,X_{t-1})} \binom{X_{t-1}}{s} \alpha^s (1-\alpha)^{X_{t-1}-s} \times \frac{e^{-\lambda}\lambda^{X_t - s}}{(X_t - s)!} \tag{2.6}$$

where () is the standard combinatorial symbol. Since the model is Markovian, the likelihood function would be formed by the product of $p(X_t|X_{t-1})$. The authors also studied forecasting methods under the Poisson $INAR(1)$ in Freeland & McCabe (2004b) using the median of the k-step-ahead conditional distribution. The methods were applicable to low count time series data such as wage loss claims. The $INAR(1)$ was later extended to higher orders (INAR(p)) in Alzaid & Al-Osh (1990) where they presented it in the following way

$$X_n = \sum_{i=1}^{p} \alpha_i \circ X_{n-i} + \epsilon_t \qquad \text{for } n = 0, \pm 1, \pm 2, ... \tag{2.7}$$

where the $\alpha_i' s$ for $i = 1, 2, ...p$ are non-negative constants such that $\sum_{i=1}^{p} \alpha_i < 1$. Some of the similarities with Gaussian $AR(p)$ process are not extended with higher orders such as the behaviour of the correlation which actually looks like a standard $ARMA(p, p-1)$. In addition, $INAR(p)$ does not have the Markovian property and also differs from the Gaussian $AR(p)$ in terms of the regression. The main similarity that holds for any order relies in the form of the process. In terms of parameter estimation, no illustration was presented but it is noted that using MLE to estimate the $INAR(p)'s$ could be complicated. Bu & McCabe (2008) developed a likelihood based Markov Chain approach for estimation, forecasting, and model selection for $INAR(p)$ suitable for time series count data. The work done was an extension of Freeland & McCabe (2004b) in which coherent forecasts were developed for higher order $INAR$ processes. Parameter estimation is done using Conditional Maximum Likelihood. In terms of forecasting, the authors develop an approach to produce k-step-ahead forecasts of the conditional probability distribution using the transition probability function of the $INAR(p)$ process. They assessed the parameter uncertainty by computing confidence intervals for probability forecasts and developed new model selection tools using residual processes. It is important to note that there are two main restrictions in the $INAR(p)$ process; one is on the innovation term, ϵ_t, and the other on the correlation coefficients, $\alpha_i' s$. Both terms must take positive values which limits the application of the process to only positively correlated integer-valued data. In addition the estimation, inference, and forecasting of the process were only applicable to count data, more specifically Poisson distributed time series data.

McKenzie (1988) developed a wider range of Poisson processes besides the $AR(1)$. The class of models developed in this paper includes the moving-average component as well as an extension to higher-order processes. Hence, models such as $MA(q)$ and $ARMA(1, q)$ were derived and their properties were discussed in detail such as time-reversibility and asymptotic behaviour. Note that McKenzie did not extent $AR(1)$ to a higher order.

Other Integer-Valued processes

The binomial thinning operator was also used to develop processes with other discrete marginal distributions. McKenzie (1986) developed a combination of AR, MA, and $ARMA$ processes to model sequences of dependent discrete random variables having negative binomial and geometric distributions. The models are found to be analogue to their continuous version, gamma and negative exponential. Those continuous models mirror the well-known standard Gaussian $ARMA$ processes. Inferential problems are not addressed as the paper focuses on the derivation of each process and studies in depth the similarities between the discrete and the continuous version of the models as well as the interval processes.

Extension of the thinning operator

Joe (1996) extended the binomial thinning operator to include all distributions in the family of convolution-closed infinitely divisible class. This class includes but is not limited to gamma, inverse Gaussian, Poisson, negative binomial, and generalized Poisson margins. Note that the Gaussian is considered to be a special case in these models. To define an infinitely divisible distribution we consider a random variable X having a probability distribution F, $X \sim F$. Then for every positive integer n, $n \in \mathbb{N}$, there exist n i.i.d. random variables, $X_1, X_2, ..., X_n$, such that $Y = \sum_{i=1}^{n} X_i$ is equal in distribution to X, i.e. the sum follows the same distribution as X, $Y \sim F$ (Steutel , 1979). This means that if the innovations of the processes were added together they would remain in the same distribution. Joe developed AR, MA and $ARMA$ models of arbitrary orders and discussed their special cases. He also mentioned the possible extension to non-stationary AR processes. The marginal distributions of the processes covered in this paper are used either for count data (Poisson, negative binomial, or generalized Poisson margin) or for positive response variable (gamma or inverse Gaussian). Joe constructed a generalized form of $AR(1)$ with marginal distribution F_θ and autocorrelation $0 < \alpha < 1$ as follows:

$$Y_t = A_t(Y_{t-1}) + \epsilon_t \tag{2.8}$$

where the innovations ϵ_t are i.i.d. following an $F_{(1-\alpha)\theta}$ distribution for all t and $\{A_t : t \geq 1\}$ are independent replications of the operator A. The author discussed the special cases of this model under the different marginal distributions mentioned above. It is noted that when F_θ is Poisson(θ) the resulting process is the one that appeared in McKenzie (1988) and Al-Osh & Alzaid (1987). However, when F_θ is negative binomial, the process obtained differs from the one developed in McKenzie (1986). Joe also extended the model to higher order AR processes by defining $AR(2)$ as follows:

$$Y_t = A_t(Y_{t-1}, Y_{t-2}) + \epsilon_t \tag{2.9}$$

The extension is based on an operator applied to a vector of p random variables. This approach differs from the extension suggested by Alzaid & Al-Osh (1990) which was based on the sum of operators applied to different random variables. Joe studied in detail the properties of $AR(2)$ noting that the generalization to $AR(p)$, $p > 2$, is straightforward using the multivariate extension of the convolution-closed infinitely divisible univariate family but the notation cannot be simplified. Note that the Binomial distribution is not part of this family of distributions; hence this case was excluded from Joe's extension. In addition to that restriction, the models studied in Joe's paper are only suitable for positively correlated data; either count data or positive response variable can be modeled under this extension of the thinning operator.

True Integer Value Time Series model

Freeland (2010) extended the autoregressive process developed in Joe (1996) and constructed the True Integer Value Time Series of order one ($TINAR(1)$) which is a discrete valued stationary symmetric $AR(1)$ model having either negative or positive autocorrelation. The innovations are distributed on both the positive and negative integers. In Freeland's model the innovations follow a symmetric distribution, more specifically their distribution is built by taking the difference between two i.i.d. Poisson random variables which is called the Skellam distribution. This is also the distribution for the innovations of our model which will be defined and discussed in detail in chapter 3. Freeland studied the model's basic properties as well as the parameter estimation by conditional least squares. In order to construct $TINAR(1)$, the author defined a new operator "$\star$". Let Z_t be the difference of two latent i.i.d. Poisson random variables, X_t and Y_t. Then the operator would be defined as

$$\alpha \star Z_t | Z_t = \alpha \circ X_t - \alpha \circ Y_t | X_t - Y_t \tag{2.10}$$

where $0 < \alpha < 1$ and "$\circ$" is the binomial thinning operator. $TINAR(1)$ process is then defined by

$$Z_t = \alpha \star Z_{t-1} + \epsilon_t \tag{2.11}$$

where Z_t is an integer valued stochastic process and "$\star$" is the operator defined in (2.10) by replacing t by $t-1$. For the negative correlation the process would be defined as

$$Z_t = \alpha \star -Z_{t-1} + \epsilon_t \tag{2.12}$$

given the same specifications of α and Z_t. $TINAR(1)$ is also seen as the difference between two latent independent Poisson $INAR(1)$. It follows that the process preserved some important properties from $INAR(1)$ such as being ergodic and strongly stationary and could be extended to other processes having a Skellam marginal distribution such as $MA(q)$, $ARMA(1,q)$, and ARMA(p, p-1). One main limitation of this model is the fact that it is defined in terms of a latent process which makes higher order $TINAR$ models difficult to use.

2.2.3 Pegram's Operator

A more flexible operator which allows the construction of Box and Jenkins' type $ARMA$ models is Pegram's operator, (Pegram, 1980). The operator will be defined and discussed in detail in Chapter 3, prior to defining our model. At this point, we will discuss briefly the models developed by Biswas & Song (2009) using this operator to model binomial and categorical data that were excluded from Joe's extension of the thinning operator since they weren't infinitely divisible, (Joe, 1996). Pegram's operator's flexibility lies in the marginal distribution; the operator does not impose any restrictions on the margins. It was originally proposed to develop AR processes. Biswas and Song extended the operator and developed discrete-valued $ARMA$ processes with arbitrary discrete marginal distributions. More specifically, the authors developed $AR(p)$, $MA(q)$, and $ARMA(p,q)$ processes as well as discussed different methods for parameter estimation including maximum likelihood. Biswas and Song conducted a simulation study to compare $AR(1)$ processes developed by the thinning operator and by Pegram's operator. They concluded that neither of the operators outperforms the other unless the data generation mechanism is similar to Pegram's mechanism (for e.g. $X_t = X_{t-1}$ for a large portion of the data points); then in that case $AR(1)$ under Pegram's operator is better. However, there are two main advantages in using Pegram's operator; one is the absence of restrictions for the type of discrete-valued time series data and the other the flexibility in building higher order processes. For these two reasons, we build our model using Pegram's operator.

2.2.4 Models for stock price change data

Modeling stock price change data has received a lot of attention recently. We are looking specifically at a discrete type of data in which the stock price is recorded after each transaction during the day. Engle & Russel (1998) developed a complex model, Autoregressive Conditional Duration (ACD) process, which aims at fitting positive and negative integer-valued time series and applied their model to a financial data which is the price change in each transaction of IBM stock. The ACD is an approach to model irregularly spaced transactions and the focus is therefore on the expected duration between the events. More specifically, the model derives the density of the duration at time t conditional on past durations. The main variable in the ACD model is then the arrival time and other variables could be associated to the model such as price and volume. Our model is focused on the price change and we look at the conditional probabiliy of the price change at time t given past observations; volume and duration are two covariates added to the variance term of our innovations. Russel & Engle (1998) developed another model applicable to financial data in which the price of the stock is dependent on both previous events and the historic distribution, they called their model the Autoregressive Conditional Multinomial (ACM) process. Both the ACD and ACM were later combined in Russel & Engle (2005) to develop one model, Autoregressive Conditional Multinomial-Autoregressive Conditional Duration model (ACM-ACD), in which the price change distribution is conditional on past information as well as on the duration between transactions. The price change covered is limited to a maximum of two-ticks jump (5 states) which means that large values of price change are not taken into consideration. The authors included the duration between transactions in their model but excluded large price changes. Moreover, the models developed involve a large number of parameters to estimate. For instance, in Russel & Engle (1998), the number of parameters estimated for ACM(2,2,2) is 46 after applying a restriction and consequently reducing the number of estimated parameters.

In Freeland (2010), the model is applied to a discrete stock price change data for one day. Freeland excluded both duration and large price changes from his model as they had insignificant impact on the price change for that day. However, he presented a much simpler model with only two parameters to be estimated.

Chapter 3

The Model

In this chapter we define and discuss the properties of our AR model; we present the model in three variations. We start by defining Pegram's operator which we use to build our model and we compare this operator with the binomial thinning operator by studying two Poisson $AR(1)$ models, one constructed using Pegram's operator and the other using the binomial thinning operator. We then define the distribution of the innovations, the Skellam distribution, and we derive the first variation of our model which is its simplest form, excluding outliers in the data. We discuss some of the model's properties and derive some useful expressions and present briefly an extension to this model. The other two variations are defined at the end of this chapter, one includes a covariate that relates the large price changes to the volume of the trade and the other adds a covariate to account for the duration between each transaction and also related to large price changes.

3.1 Pegram's Operator

3.1.1 Definition

The Pegram operator or mixing operator will be denoted by "$*$" and is defined as follows. For any two independent discrete random variables U and V over the same sample space Ω and any coefficient $\alpha \in (0, 1)$, the Pegram operator mixes the two random variables to produce a third random variable Z. The mixing operation is denoted as

$$Z = (U, \alpha) * (V, 1 - \alpha) \tag{3.1}$$

and the resulting marginal probability function is defined as

$$P(Z \in A) = \alpha P(U \in A) + (1 - \alpha)P(V \in A) \quad \text{for all } A \in \Omega \tag{3.2}$$

Note that if $\alpha = 0$ then $Z = V$ with probability 1. The operator is mixing two discrete distributions with respective weights of α and $1 - \alpha$. There is no restriction over the type of discrete distributions used for V and U, only that they must have the same sample space. Furthermore, the sample space does not have to be integers or a subset of integers and can include categorical sample spaces.

In this thesis we will use the integers as our sample space. Further, since our marginal distribution will be symmetric around zero, we will extend the Pegram operator to allow both positive and negative values of our correlation parameter α. To do that, we assume three cases of possible probability function. Let U and V be any two independent integer valued random variables and let $0 < \alpha < 1$.
Then for $Z = (U, \alpha) * (V, 1 - \alpha)$ the probability function is

$$P(Z = j) = \alpha P(U = j) + (1 - \alpha) P(V = j), \quad j = 0, \pm 1, \pm 2, \ldots \tag{3.3}$$

and for $Z = (U, -\alpha) * (V, 1 - \alpha)$ the probability function is

$$P(Z = j) = \alpha P(U = -j) + (1 - \alpha) P(V = j), \quad j = 0, \pm 1, \pm 2, \ldots \tag{3.4}$$

as for $Z = (-U, \alpha) * (V, 1 - \alpha)$ the probability function is

$$P(Z = j) = \alpha P(-U = j) + (1 - \alpha) P(V = j), \quad j = 0, \pm 1, \pm 2, \ldots \tag{3.5}$$

Thus $Z = (U, -\alpha) * (V, 1 - \alpha)$ is the same as $Z = (-U, \alpha) * (V, 1 - \alpha)$. As noted above, $0 < \alpha < 1$, which means that the parameter's value will always be positive. Therefore, in order to account for the negative correlation, the structure of the mixing operation would differ and we would use (3.5) to construct the model.

The conditional probability of the new random variable is a useful tool in the parameter estimation that will be discussed in Chapter 4. Conditional on U and using the same case as (3.3), the probability function the new random variable would be

$$P(Z = j | U) = \alpha I[U = j] + (1 - \alpha) P(V = j), \quad j = 0, \pm 1, \pm 2, \ldots \tag{3.6}$$

where $I[B]$ is an indicator function taking value 1 if event B is present and 0 otherwise. This means that the conditional probability could be constructed as follows

$$P(Z = j | U) = \begin{cases} \alpha + (1 - \alpha) P(V = j) & \text{if } U = j \\ (1 - \alpha) P(V = j) & \text{if } U \neq j \end{cases} \tag{3.7}$$

In order to express negative correlation, the conditional probability remains the same but the indicator events would be $U = -j$ or $U \neq -j$.
One of the nice features of Pegram's operator is that it can be easily extended to handle more than two discrete random variables which will later be useful in developing higher order AR models. For instance, if we start with three discrete random variables U, V, and W, then the

new random variable Z is defined under the mixing operator as follows:

$$Z = (U, \alpha_1) * (V, \alpha_2) * (W, 1 - \alpha_1 - \alpha_2) \tag{3.8}$$

and its conditional probability function is

$$P(Z = j|U, V) = \alpha_1 I[U = j] + \alpha_2 I[V = j] + (1 - \alpha_1 - \alpha_2)P(W = j), \quad j = 0, \pm 1, \pm 2, \ldots \tag{3.9}$$

3.1.2 Comparison

We consider in this section two Poisson integer-valued $AR(1)$ processes each constructed using a different operator: one is the Pegram's operator, Y_t, and the other is the binomial thinning operator, X_t. This comparison study is useful to observe how each model behaves under different operators.

Table 3.1: Comparison between the Binomial Thinning and Pegram's operator

	Binomial Thinning	**Pegram's Operator**
Model	$X_t = \alpha \circ X_{t-1} + \omega_t$	$Y_t = (Y_{t-1}, \alpha) \star (\epsilon_t, 1 - \alpha)$
Innovations' distribution	$\omega_t \sim Poisson((1 - \alpha)\lambda)$	$\epsilon_t \sim Poisson(\lambda)$
Marginal Distribution	$X_t \sim Poisson(\lambda)$	$X_t \sim Poisson(\lambda)$
Conditional Expectation	$E[X_t|X_{t-1}] = \alpha X_{t-1} + (1 - \alpha)\lambda$	$E[Y_t|Y_{t-1}] = \alpha Y_{t-1} + (1 - \alpha)\lambda$

We notice that the main similarities lie in the marginal distribution, both processes follow marginally the Poisson distribution. In addition to that, the expressions of the conditional expectations are the same. However, the innovations' distributional specifications are different. Although under both models the innovations are Poisson distributed, the parameters are not the same. The most significant difference is in terms of conditional distributions, and we will examine these graphically in a numerical example.

Suppose we have $\alpha = 0.5$ and $\lambda = 2$. Then our processes marginal distribution is $Poisson(2)$. The innovations' distributions are $Poisson(1)$ for X_t and $Poisson(2)$ for Y_t. In figures 3.1, 3.2, and 3.3 we have the plots of the marginal probability functions of X_t and Y_t as well as the conditional probability functions given three different values of X_{t-1} and Y_{t-1}; the values are 1, 3 and 5. We notice, as expected, that the marginal distribution is the same for both processes, it is Poisson distribution. As for the conditional distribution, the model constructed under Pegram's operator represents more consistency among the different values given for Y_{t-1}. The main pattern is that Y_t will most likely take a value equal to the value of Y_{t-1}. As for the rest of the values, the probability that Y_t might take them is very small. However, for the model built under the binomial thinning operator, this consistency does not hold; the model behaves differently with larger values. For small values such as in figure 3.1, X_t would most likely have a value equal to X_{t-1} but note that the probability is not as large as for the case of Y_t and that the probability of getting other values is not very small. As the value of X_{t-1} gets larger, we notice that the probability of X_t taking the value of X_{t-1} becomes smaller and smaller and the probability for other values to occur is higher than before to the point where the highest probability is no longer assigned to X_t taking the value of X_{t-1}, see figures 3.2 and 3.3.

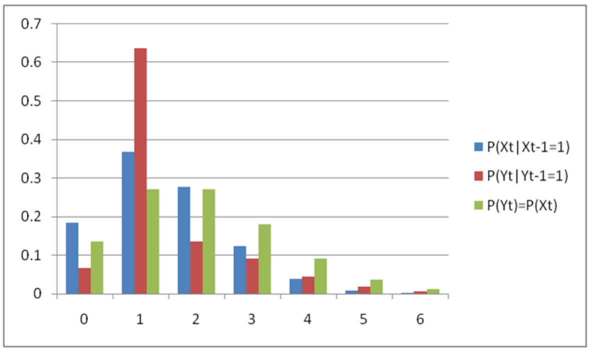

Figure 3.1: Marginal and Conditional Probability function of X_t and Y_t given $X_{t-1} = Y_{t-1} = 1$

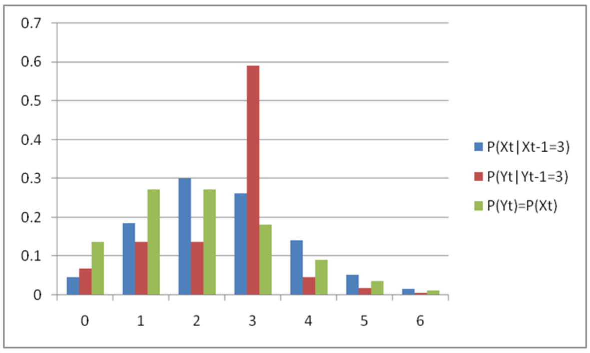

Figure 3.2: Marginal and Conditional Probability function of X_t and Y_t given $X_{t-1} = Y_{t-1} = 3$

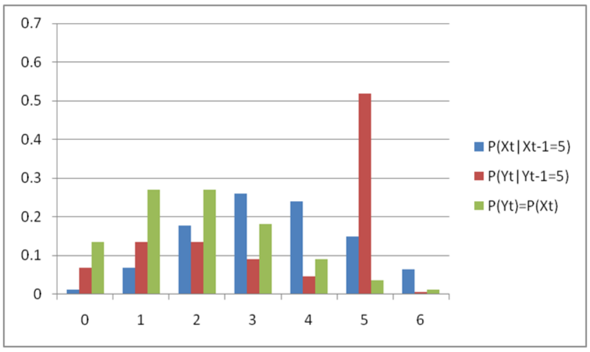

Figure 3.3: Marginal and Conditional Probability function of X_t and Y_t given $X_{t-1} = Y_{t-1} = 5$

3.2 Skellam Distribution

For the marginal distribution of our time series model we use the difference between two i.i.d. Poisson random variables. This symmetric distribution is studied in Irwin (1937) and further generalized by Skellam (1946). The general case would be having two independent Poisson random variables $Z_1 \sim P_o(\lambda_1)$ and $Z_2 \sim P_o(\lambda_2)$ and taking their difference. Then $D = Z_1 - Z_2$ follows a Skellar distribution with parameters λ_1 and λ_2. The mean and variance would respectively be $\lambda_1 - \lambda_2$ and $\lambda_1 + \lambda_2$. Note that in the special case where we have two i.i.d Poisson(λ) random variables, the Skellam distribution would have a mean equal to zero and a variance of 2λ. As stated previously, we will be working with this special case of Skellam distribution.

Suppose that our time series is denoted by Y_t, which is symmetrically distributed over the integers and has a marginal probability function which is time-dependent. We will denote the probability function as $p_j(\lambda) = P(Y_t = j) = P(Y_t = -j)$ for $j = 0, 1, \dots,$ which has the following form

$$\begin{aligned} p_j(\lambda) &= \textstyle\sum_{x=0}^{\infty} \frac{e^{-2\lambda}\lambda^{2x+j}}{x!(x+j)!} \\ &= e^{-2\lambda} I_j(2\lambda), \end{aligned} \tag{3.10}$$

where x is the difference between the two i.i.d. Poisson random variables, λ is the sole parameter of the distribution and $I_j(2\lambda)$ is the modified Bessel function of the first kind of order j with argument 2λ. We will call this the Skellam distribution with parameter λ. Note that the moment generating function of this distribution is $M_D(s) = e^{\lambda(e^s + e^{-s} - 2)}$.

The probablity distribution for different values of λ is presented in the histogram in figure 3.4.

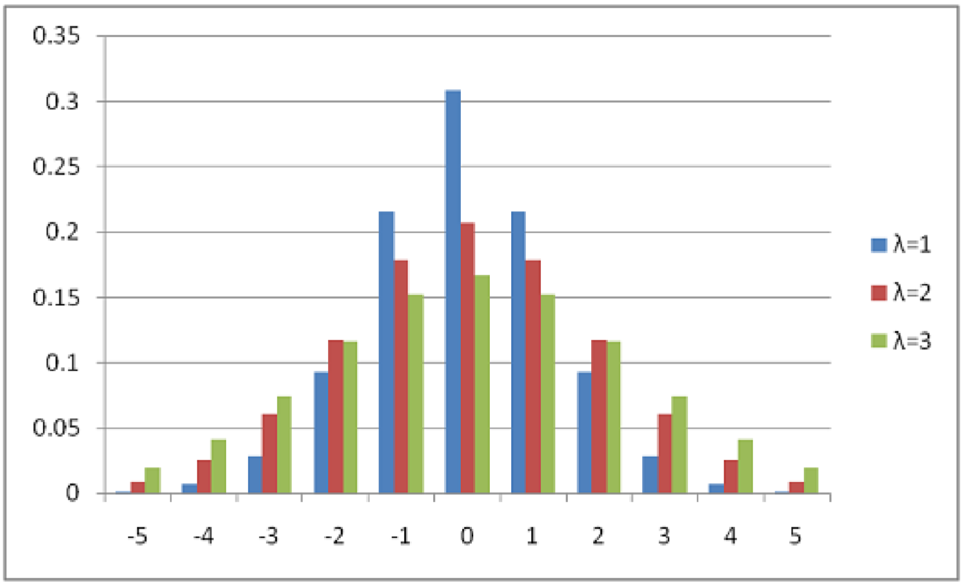

Figure 3.4: Probability function for Skellam distribution

3.3 Variation I

We now define our integer-valued AR(p) process, which is essentially a special case of the discrete-valued AR(p) process of Biswas & Song (2009). As noted before this is the simplest form of our model in which the variance of the innovations, 2λ, is constant and does not depend on t. This means that we are assuming the variation in the stock price change is not related to either the volume of the trade or the duration between trades. We will discuss some of the model's properties and conclude this section by an important extension of our model.

3.3.1 Definition

Suppose that Y_t is an integer valued time series and that ϵ_t's are i.i.d. random variables following the Skellam distribution with parameter λ. Respectively, denote μ and σ^2 the mean and variance of the Skellam distribution. It follows that $\mu = 0$ and $\sigma^2 = 2\lambda$. The probability function of the marginal distribution is denoted by p_{y_t}.

Definition 1 *Let Y_t be an integer-valued discrete-time stochastic process such that*

$$Y_t = (Y_{t-1}, \alpha_1) * (Y_{t-2}, \alpha_2) * \cdots * (Y_{t-p}, \alpha_p) * (\epsilon_t, 1 - \alpha_1 - \cdots \alpha_p) \tag{3.11}$$

which is a mixture of $p+1$ discrete distributions, where conditionally $Y_{t-1}, \ldots, Y_{t-p}$ are p point masses and ϵ_t has a Skellam distribution with parameter λ, with respective weights $\alpha_1, \ldots, \alpha_p$ and $1 - \alpha_1 - \cdots - \alpha_p$, $\alpha_k \in (0, 1)$, $k = 1, \ldots, p$ and $\sum_{k=1}^{p} \alpha_k \in (0, 1)$. This implies that for every $t \in 0, \pm 1, \pm 2, \ldots$, the conditional probability function is

$$P(Y_t = y_t | Y_{t-1}, \ldots, Y_{t-p}) = \alpha_1 I[Y_{t-1} = y_t] + \cdots + \alpha_p I[Y_{t-p} = y_t] + (1 - \alpha_1 - \cdots - \alpha_p) p_{y_t} \tag{3.12}$$

where α_k, $k = 1, \ldots, p$ are such that $1 - \alpha_1 z - \cdots - \alpha_p z^p = 0$ has roots outside of the unit disc.

3.3.2 Properties

Marginal Stationarity

Biswas & Song (2009) note that the marginal distribution of Y_t is the same as ϵ_t, which in our case is a Skellam distribution with parameter λ. Note that if $P(Y_{t-h} = j) = p_j$ for $h = 1, ..., p$, then $P(Y_t = j) = p_j$, which implies marginal stationarity. This means that, marginally, $Y_t \sim$ Skellam(λ).

The conditional mean

Denoting p_{y_t} as the marginal distribution defined in (3.10), the conditional mean is derived as follows

$$\begin{aligned}
E[Y_t|Y_{t-1},\ldots,Y_{t-p}] &= \sum_{y_t} y_t P(Y_t = y_t|Y_{t-1},...,Y_{t-p}) \\
&= \sum_{y_t} y_t\{\alpha_1 I[Y_{t-1}=y_t] + ... + \alpha_p I[Y_{t-p}=y_t] + (1-\alpha_1 - ... - \alpha_p)p_{y_t}\} \\
&= \alpha_1 Y_{t-1} + ... + \alpha_p Y_{t-p} + (1-\alpha_1 - ... - \alpha_p)\sum_{y_t} y_t p_{y_t} \\
&= \alpha_1 Y_{t-1} + ... + \alpha_p Y_{t-p} + (1-\alpha_1 - ... - \alpha_p)E(Y_t) \\
&= \alpha_1 Y_{t-1} + ... + \alpha_p Y_{t-p} \\
&= \sum_{k=1}^{p} \alpha_k Y_{t-k}
\end{aligned} \tag{3.13}$$

The autocorrelation function

Given that Y_t and Y_{t-h} are uncorrelated when $Y_{t-1},...,Y_{t-p}$ are conditionally given, the autocovariance function (ACVF) is

$$\begin{aligned}
\gamma(h) &= E[cov(Y_t, Y_{t-h}|Y_{t-1},...,Y_{t-p})] + cov[E(Y_t|Y_{t-1},...,Y_{t-p}), E(Y_{t-h}|Y_{t-1},...,Y_{t-p})] \\
&= cov(\textstyle\sum_{k=1}^{p} \alpha_k Y_{t-k}, Y_{t-h}) \\
&= \alpha_1 cov(Y_{t-1}, Y_{t-h}) + ... + \alpha_p cov(Y_{t-p}, Y_{t-h}) \\
&= \alpha_1 \gamma(h-1) + ... + \alpha_p \gamma(h-p)
\end{aligned} \tag{3.14}$$

Dividing both sides of (3.14) by $\gamma(0)$, we obtain a recursive relation for the autocorrelation function (ACF)

$$\rho(h) = \alpha_1 \rho(h-1) + ... + \alpha_p \rho(h-p) \tag{3.15}$$

Note that the ACF we obtain is similar to Box and Jenkins' $AR(p)$ ACF. For instance, the ACVF of an $AR(1)$ process is

$$\gamma(h) = cov(Y_t, Y_{t-h}) = \alpha^{|h|}\sigma^2,$$

consequently the ACF is

$$\rho(h) = corr(Y_t, Y_{t-h}) = \alpha^{|h|}$$

which is the same as the ACF of a Box and Jenkins' $AR(1)$ process.

The moment generating function

$$\begin{aligned}
M_{Y_t|Y_{t-1}\ldots Y_{t-p}}(s) &= \sum e^{sy_t} P(Y_t = y_t|Y_{t-1}, ..., Y_{t-p}) \\
&= \sum e^{sy_t}\{\alpha_1 I[Y_{t-1} = y_t] + ... + \alpha_p I[Y_{t-p} = y_t] + (1 - \alpha_1 - ... - \alpha_p)p_{y_t}\} \\
&= \alpha_1 \sum e^{sy_t} I[Y_{t-1} = y_t] + ... + \alpha_p \sum_s e^{sy_t} I[Y_{t-p} = y_t] + (1 - \alpha_1 - ... - \alpha_p) \sum e^{sy_t} p_{y_t} \\
&= \alpha_1 e^{sY_{t-1}} + \cdots + \alpha_p e^{sY_{t-p}} + (1 - \alpha_1 - \cdots - \alpha_p)M_\epsilon(s)
\end{aligned} \tag{3.16}$$

where $M_\epsilon(s)$ is the moment generating function of the Skellam distribution and has been defined in section 3.2. Since the Skellam distribution is symmetric around zero, we have the added feature that the lags may also have a negative correlation. To accomplish this we continue to let the $\alpha_k' s$ denote the mixing weights and simply change the sign of the lagged value of Y_t in our model. This has been show in section 3.1 and will be illustrated in a simple numerical example.

Example

We will first consider two $AR(1)$ processes, one with a positive correlation and the other with a negative correlation. We will define them and look at each model's properties. Then we will consider a numerical example in which we use an $AR(2)$ process having one positive correlation and one negative correlation.

Starting with $AR(1)$ where we have a lag one positive correlation. The process is defined as follows:

$$Y_t = (Y_{t-1}, \alpha) * (\epsilon_t, 1 - \alpha), \tag{3.17}$$

with the following conditional probability function,

$$P(Y_t = y_t|Y_{t-1}) = \alpha I[Y_{t-1} = y_t] + (1 - \alpha)p_{y_t} \tag{3.18}$$

The properties of the model are:

(i) Conditional mean: $E[Y_t|Y_{t-1}] = \alpha Y_{t-1}$
(ii) Autocorrelation function: $\rho(h) = (\alpha)^{|h|}$
(iii) Moment generating function:$M_{Y_t|Y_{t-1}}(s) = \alpha e^{sY_{t-1}} + (1 - \alpha)M_\epsilon(s)$

If $AR(1)$ had a lag one negative correlation, it would be defined by

$$Y_t = (-Y_{t-1}, \alpha) * (\epsilon_t, 1 - \alpha), \tag{3.19}$$

with the following conditional probability function,

$$P(Y_t = y_t | Y_{t-1}) = \alpha I[Y_{t-1} = -y_t] + (1 - \alpha) p_{y_t} \tag{3.20}$$

As for the properties of this model, they are:

(i) Conditional mean: $E[Y_t|Y_{t-1}] = -\alpha Y_{t-1}$
(ii) Autocorrelation function: $\rho(h) = (-\alpha)^{|h|}$
(iii) Moment generating function:$M_{Y_t|Y_{t-1}}(s) = \alpha e^{-sY_{t-1}} + (1 - \alpha) M_\epsilon(s)$

As a numerical example, consider an AR(2) model where the second lag has a negative correlation and the first lag has a positive correlations. Suppose the correlations are respectively 0.5 and -0.3. The mixing weights for this model are then 0.5, 0.3 and (1 - 0.5 - 0.3)=0.2 and the model is written as

$$Y_t = (Y_{t-1}, 0.5) * (-Y_{t-2}, 0.3) * (\epsilon_t, 0.2) \tag{3.21}$$

The conditional probability function is

$$P(Y_t = y_t) = 0.5I[Y_{t-1} = y_t] + 0.3I[Y_{t-2} = -y_t] + 0.2p_{y_t}, \tag{3.22}$$

and the model's properties are:

(i) Conditional mean: $E[Y_t|Y_{t-1}] = 0.5Y_{t-1} - 0.3Y_{t-2}$
(ii) Autocorrelation function: $\rho(h) = 0.5\rho(h-1) - 0.3\rho(h-2)$
(iii) Moment generating function:$M_{Y_t|Y_{t-1}}(s) = 0.5e^{sY_{t-1}} + 0.3e^{-sY_{t-1}} + 0.2M_\epsilon(s)$

Both the conditional and the marginal probability functions of Y_t are illustrated in figure 3.5. It has been noted earlier that the marginal distribution of Y_t is the Skellam distribution which explains the symmetrical shape of the graphical representation of its probability function. In terms of the conditional probability, Y_t is more likely to take values of 1 or -1 given that $Y_{t-1} = Y_{t-2} = 1$; this is due to having both the negative and positive correlations. Also note that the first lag has a larger weight which explains the higher conditional probability of obtaining a value equal to 1 at time t as opposed to obtaining -1.

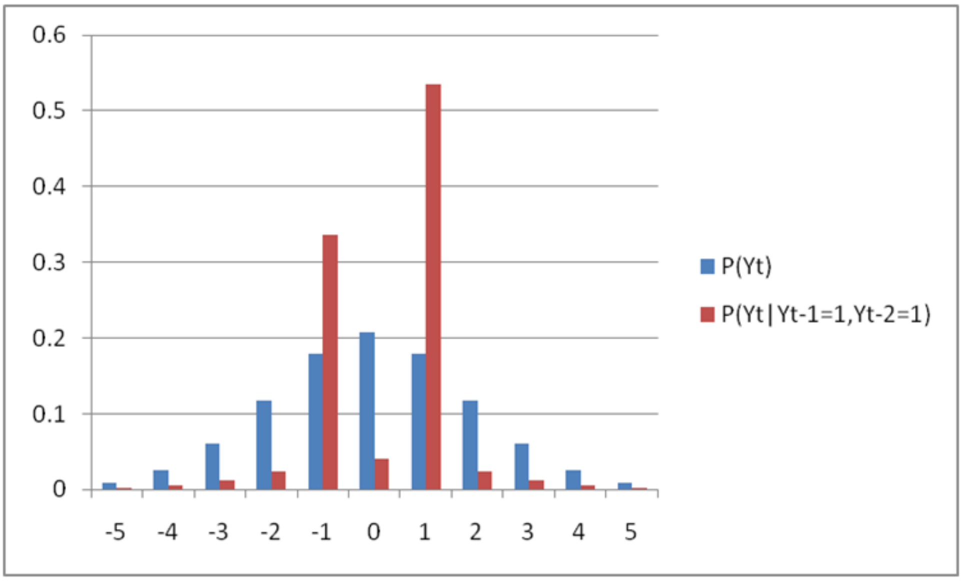

Figure 3.5: Conditional and Marginal Probability function of Y_t

3.3.3 Extension: Mixed Lag Correlation

Definition

As mentioned previously, the model is a special case of the $AR(p)$ process proposed by Biswas & Song (2009). This special case is quite rich because the correlation structure can be broadened. In particular it may contain both positive and negative dependence within the same lag. This allows more flexibility in terms of expressing the correlations properly at different lags. Models of this form can quickly become unmanageable for higher order lags. To illustrate we will only consider the AR(1) case.

Definition 2 *Let Y_t be an integer-valued discrete-time stochastic process such that*

$$Y_t = (Y_{t-1}, \alpha_+) * (-Y_{t-1}, \alpha_-) * (\epsilon_t, 1 - (\alpha_+ + \alpha_-)) \tag{3.23}$$

which is a mixture of 3 discrete distributions, where conditionally Y_{t-1} and $-Y_{t-1}$ are two point masses and ϵ_t has a Skellam distribution with parameter λ, with respective weights α_+, α_- and $1 - \alpha_+ - \alpha_-$, with $0 < \alpha_+ < 1$, $0 < \alpha_- < 1$, and $0 < \alpha_+ + \alpha_- < 1$. This implies that for every $t \in 0, \pm 1, \pm 2, \ldots,$ the conditional probability function is

$$P(Y_t = j | Y_{t-1}) = \alpha_+ I[Y_{t-1} = j] + \alpha_- I[Y_{t-1} = -j] + (1 - \alpha_+ - \alpha_-) p_j \tag{3.24}$$

Model properties

Moment Generating Function

$$M_{Y_t|Y_{t-1}}(s) = \alpha_+ e^{sY_{t-1}} + \alpha_- e^{-sY_{t-1}} + (1 - (\alpha_+ + \alpha_-)) M_\epsilon(s) \tag{3.25}$$

where $M_\epsilon(s)$ is the moment generating function for the Skellam distribution with parameter λ.

Conditional Expectation

$$E[Y_t|Y_{t-1}] = \alpha_+ Y_{t-1} - \alpha_- Y_{t-1} \tag{3.26}$$

Autocorrelation Function

$$\rho(1) = \alpha_+ - \alpha_- \tag{3.27}$$

Note that if $\alpha_+ = \alpha_-$ then $\rho(1) = 0$ and in that case we don't have correlation but we would still have dependence.

3.4 Other variations of our model

The other variations of our model lie in the specification of the marginal distribution which is the Skellam distribution. More specifically, the variance of the marginal distribution will have a different expression in model variations II and III.

Since we are interested in the behaviour of the price change and want to specifically model the large values of price changes, then we are looking at the variability of our random variable. This is represented by the variance of the marginal distribution. As mentioned in Section 3.2, the variance of our Skellam distribution is 2λ since the distribution is constructed using two i.i.d. Poisson random variables, i.e. having the same parameter λ.

In the first variation of the model, λ was a constant parameter and did not depend on time or any other factor. This means that at any given time (including opening or closing of the market) and for any transaction (including large or small volumes), λ had the same value. In order to capture the effect of these factors on the price change we need to let λ vary with duration and volume.

3.4.1 Variation II

In this variation of the model we develop an expression in which λ_t is a function of the volume as we believe that the extreme decreases or increases in the stock price are due to large trades. Hence, λ_t is modeled as follows:

$$\lambda_t = \gamma + \beta \left(\frac{|V_t - V_{t-1}|}{max(Vol)} \right)^p \tag{3.28}$$

where V_t is the volume of the trade at time t and $max(Vol)$ is a constant to be determined based on the data set. There are two components in the new expression of λ_t; the first is a constant parameter, γ and the second component which contains the additional parameter β

Table 3.2: Stock Price Change Data

Time	Volume(V_t)	$\|V_t - V_{t-1}\|$	Price	Price Change
1	2681	NA	9.21	NA
2	6835	4154	9.21	0
3	150,000	143,165	9.09	-0.12
4	4352	145,648	9.20	0.11
5	3080	1272	9.20	0

depends on the size of the volume of the trade. The reason behind using the difference of the trade volumes instead of relating the parameter directly to the volume is that a large portfolio trade at a certain time will not only affect the price at this specific time but also the price during the next transaction. To illustrate this we look at a small portion of a stock price change data, table 3.2. We notice that the transaction at time $t = 3$ has a very large volume, $V_3 = 150,000$ which explains the big drop of twelve ticks in the price of the stock. However, looking at the transaction at time 4, the price change is large but it is not associated with the corresponding volume of the transaction, V_4. In fact, the increase in the price at time 4 occurred to restore the original level of the price of the stock. Comparing the price change and the absolute value of the difference in the volumes of the trades, we can see that when the price change is very large, it is associated with a large value of the difference in the volumes such as at times $t = 3$ and $t = 4$.

Another parameter appears in the second component; it is the power of the fraction, p. We will discuss in detail how to estimate this parameter in Chapter 4. The values of p can range from 1, where the variance and the volume would be linearly related, to 2, where the relation is quadratic.

The way λ_t is defined in this model suggests that when the difference between the volume of the trade at time t and the one at time $t - 1$ is large enough (greater than the pre-defined $max(Vol)$), the fraction will consequently be large and β is therefore significant and has an impact on λ_t. However, when the difference in the volumes is small, the fraction will be very small and the term with β will not be significant. Hence, the value of λ_t will be a constant equal to the first component of the expression, γ.

3.4.2 Variation III

In the third variation of the model, the variance is not only related to the volume of the trade but also to the duration between two trades. It is known that the effect of closing and opening of the market has a great impact on the price of the stock. We add a covariate to capture the duration between the transactions and relate it to the variance. The new expression for λ_t is

$$\lambda_t = \gamma + \beta \left(\frac{|V_t - V_{t-1}|}{max(Vol)} \right)^p + \delta \left(\frac{D_t - D_{t-1}}{max(D_t)} \right)^q \tag{3.29}$$

where D_t is the time of the transaction; it is expressed in the number of seconds in a given day. The duration is then expressed in terms of the difference in time between the current transaction and the previous one. The value of $max(D_t)$ is determined in the same way as $max(Vol)$ is determined; it is a constant chosen according the data. The power of the fraction q will be shown to have a value of 2 according to the data we have which makes the relation between the variance and the duration quadratic. This value could change in case we had a different set of data. More details on this topic are discussed in chapter 4.

The idea behind the additional covariate is the same as the covariate of the volume; if the duration is large enough, δ will be significant and will have more effect on λ_t. However, if the duration is small to the point where $\left(\frac{D_t - D_{t-1}}{max(D_t)} \right)^q$ is very close to zero then there is no additional time effect on λ_t. For instance, the duration between the first transaction of today and the last transaction of yesterday is significant enough to make a difference in the value of λ_t which actually explains the effect of market opening on the price of the stock resulting in a large price change.

Chapter 4

Parameter Estimation and Model Selection

4.1 Maximum Likelihood Estimation

In order to get estimates for the parameters of our model we use Maximum Likelihood Estimation (MLE). Freeland (2010) did not use this method to estimate the parameteres of his model due to the complication of the likelihood function; he used instead Yule-Walker and conditional least squares. In our model, the likelihood function can be easily derived along with the log-likelihood and the score functions. MLE method in this case is not computationally complicated and it gives a rich set of tools useful in the model selection procedure.

To construct the likelihood function, we will denote $P_t(p) = P(Y_t = y_t|Y_{t-1}, \ldots, Y_{t-p})$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$. In our case we are basing the MLE on the conditional likelihood function, which is

$$\begin{aligned} L(\boldsymbol{\alpha}, \lambda) &= \Pi_{t=p+1}^{n} P_t(p) \\ &= \Pi_{t=p+1}^{n} P(Y_t = y_t|Y_{t-1}, \ldots, Y_{t-p}) \end{aligned}$$

then the log-likelihood for our AR(p) model is

$$\begin{aligned} l(\boldsymbol{\alpha}, \lambda) &= ln\{\Pi_{t=p+1}^{n} P_t(p)\} \\ &= \sum_{t=p+1}^{n} lnP_t(p) \\ &= \sum_{t=p+1}^{n} ln\{(1 - \alpha_1 - \cdots - \alpha_p)p_{y_t} + \alpha_1 I(y_{t-1} = y_t) + \cdots + \alpha_p I(y_{t-p} = y_t)\} \end{aligned}$$

where p_{y_t} is the marginal probability function and can be taken from any discrete valued distribution. In our case, $p_{y_t} = p_{y_t}(\lambda) = \sum_{x=0}^{\infty} \frac{e^{-2\lambda}\lambda^{2x+y_t}}{x!(x+y_t)!}$ which is the Skellam probability function.

In order to estimate the parameters, we need to find the score functions by computing the partial derivative of the log-likelihood with respect to each parameter. Each variation of the model has a certain number of parameters to be estimated, and since the variance of Skellam distribution has different specifications for each variation of the model we will consider the three variations separately.

Note that $\boldsymbol{\alpha}$ is a common parameter among all models. Thus to estimate any α_k under any model, the score function would be

$$\begin{aligned}
\frac{\partial l}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k} \sum_{t=p+1}^{n} lnP_t(p) \\
&= \sum_{t=p+1}^{n} \frac{\frac{\partial}{\partial \alpha_k} P_t(p)}{P_t(p)} \\
&= \sum_{t=p+1}^{n} \frac{\frac{\partial}{\partial \alpha_k}\{(1-\alpha_1-\cdots-\alpha_p)p_{y_t}+\alpha_1 I(y_{t-1}=y_t)+\cdots+\alpha_p I(y_{t-p}=y_t)\}}{P_t(p)} \\
&= \sum_{t=p+1}^{n} \frac{-p_{y_t}+I(y_{t-k}=y_t)}{P_t(p)}
\end{aligned} \tag{4.1}$$

The same procedure is followed to derive the rest of the score functions.

4.1.1 Variation I: λ

Recall that the first variation is the simplest version of the model: the variance of the marginal distribution is constant. Thus, beside $\boldsymbol{\alpha}$, the parameter to estimate in this model is λ. The corresponding score function is

$$\begin{aligned}
\frac{\partial l}{\partial \lambda} &= \sum_{t=p+1}^{n} \frac{\frac{\partial}{\partial \lambda}\{(1-\alpha_1-\cdots-\alpha_p)p_{y_t}+\alpha_1 I(y_{t-1}=y_t)+\cdots+\alpha_p I(y_{t-p}=y_t)\}}{P_t(p)} \\
&= \sum_{t=p+1}^{n} \frac{(1-\alpha_1-\cdots-\alpha_p)\frac{\partial p_{y_t}}{\partial \lambda}}{P_t(p)}
\end{aligned} \tag{4.2}$$

where $\frac{\partial p_{y_t}}{\partial \lambda}$ is the partial derivative of the Skellam probability function with respect to its parameter λ.

$$\begin{aligned}
\frac{\partial p_{y_t}}{\partial \lambda} &= \frac{\partial}{\partial \lambda} \sum_{x=0}^{\infty} \frac{e^{-2\lambda}\lambda^{2x+y_t}}{x!(x+y_t)!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-2\lambda}\lambda^{2x+y_t}\left(\frac{2x+y_t}{\lambda}-2\right)}{x!+(x+y_t)!}
\end{aligned} \tag{4.3}$$

4.1.2 Variation II: γ and β

In the second variation of the model, the parameter λ of the Skellam distribution (i.e. the marginal distribution) is no longer a constant and is expressed as a function of two other parameters γ and β; $\lambda_t = \gamma + \beta \left(\frac{|V_t - V_{t-1}|}{max(Vol)}\right)^p$.

Therefore the parameters we need to estimate beside $\boldsymbol{\alpha}$ are γ and β. Let us denote $V = \left(\frac{|V_t - V_{t-1}|}{max(Vol)}\right)^p$, then the score functions are

$$\frac{\partial l}{\partial \gamma} = \sum_{t=p+1}^{n} \frac{(1 - \alpha_1 - \cdots - \alpha_p)\frac{\partial p_{y_t}}{\partial \gamma}}{P_t(p)} \tag{4.4}$$

$$\frac{\partial l}{\partial \beta} = \sum_{t=p+1}^{n} \frac{(1 - \alpha_1 - \cdots - \alpha_p)\frac{\partial p_{y_t}}{\partial \beta}}{P_t(p)} \tag{4.5}$$

The partial derivatives are derived in the same way as $\frac{\partial p_{y_t}}{\partial \lambda}$ but in this case we replace λ with $\gamma + \beta V$. Then the partial derivatives of the Skellam probability function with respect to γ and β are

$$\begin{aligned} \frac{\partial p_{y_t}}{\partial \gamma} &= \frac{\partial}{\partial \gamma} \sum_{x=0}^{\infty} \frac{e^{-2(\gamma+\beta V)}(\gamma + \beta V)^{2x+y_t}}{x!(x+y_t)!} \\ &= \sum_{x=0}^{\infty} \frac{e^{-2(\gamma+\beta V)}(\gamma + \beta V)^{2x+y_t}\left(\frac{2x+y_t}{(\gamma+\beta V)} - 2\right)}{x! + (x+y_t)!} \end{aligned} \tag{4.6}$$

and

$$\begin{aligned} \frac{\partial p_{y_t}}{\partial \beta} &= \frac{\partial}{\partial \beta} \sum_{x=0}^{\infty} \frac{e^{-2(\gamma+\beta V)}(\gamma + \beta V)^{2x+y_t}}{x!(x+y_t)!} \\ &= \sum_{x=0}^{\infty} \frac{Ve^{-2(\gamma+\beta V)}(\gamma + \beta V)^{2x+y_t}\left(\frac{2x+y_t}{(\gamma+\beta V)} - 2\right)}{x! + (x+y_t)!} \end{aligned} \tag{4.7}$$

4.1.3 Variation III: γ, β, and δ

As for the third variation, $\lambda_t = \gamma + \beta \left(\frac{|V_t - V_{t-1}|}{max(Vol)}\right)^p + \delta \left(\frac{D_t - D_{t-1}}{max(D_t)}\right)^q$, thus we have one additional score function for the additional parameter, δ. Note that the score functions of γ and β are the same. We will denote $D = \left(\frac{D_t - D_{t-1}}{max(D_t)}\right)^q$. Then the additional score function needed is

$$\frac{\partial l}{\partial \delta} = \sum_{t=p+1}^{n} \frac{(1 - \alpha_1 - \cdots - \alpha_p)\frac{\partial p_{y_t}}{\partial \delta}}{P_t(p)} \tag{4.8}$$

The partial derivatives of the Skellam probability function with respect to γ and β under the third variation of the model are slightly different than the ones under the second variation. To find the partial derivatives of p_{y_t} with respect to the three parameters, γ, β, and δ, we replace

λ with $\gamma + \beta V + \delta D$.

$$\begin{aligned} \frac{\partial p_{y_t}}{\partial \gamma} &= \frac{\partial}{\partial \gamma} \sum_{x=0}^{\infty} \frac{e^{-2(\gamma+\beta V+\delta D)}(\gamma + \beta V + \delta D)^{2x+y_t}}{x!(x+y_t)!} \\ &= \sum_{x=0}^{\infty} \frac{e^{-2(\gamma+\beta V+\delta D)}(\gamma + \beta V + \delta D)^{2x+y_t}\left(\frac{2x+y_t}{(\gamma+\beta V+\delta D)} - 2\right)}{x! + (x+y_t)!} \end{aligned} \tag{4.9}$$

$$\begin{aligned} \frac{\partial p_{y_t}}{\partial \beta} &= \frac{\partial}{\partial \beta} \sum_{x=0}^{\infty} \frac{e^{-2(\gamma+\beta V+\delta D)}(\gamma + \beta V + \delta D)^{2x+y_t}}{x!(x+y_t)!} \\ &= \sum_{x=0}^{\infty} \frac{Ve^{-2(\gamma+\beta V+\delta D)}(\gamma + \beta V + \delta D)^{2x+y_t}\left(\frac{2x+y_t}{(\gamma+\beta V+\delta D)} - 2\right)}{x! + (x+y_t)!} \end{aligned} \tag{4.10}$$

$$\begin{aligned} \frac{\partial p_{y_t}}{\partial \delta} &= \frac{\partial}{\partial \delta} \sum_{x=0}^{\infty} \frac{e^{-2(\gamma+\beta V+\delta D)}(\gamma + \beta V + \delta D)^{2x+y_t}}{x!(x+y_t)!} \\ &= \sum_{x=0}^{\infty} \frac{De^{-2(\gamma+\beta V+\delta D)}(\gamma + \beta V + \delta D)^{2x+y_t}\left(\frac{2x+y_t}{(\gamma+\beta V+\delta D)} - 2\right)}{x! + (x+y_t)!} \end{aligned} \tag{4.11}$$

4.2 Model Selection Criteria

We look at model selection criteria under two perspectives. We examine first each variation of the model individually to determine the best order for our $AR(p)$ model. Then we use the models in their most adequate order and compare the three variations to determine which one fits best the data.

4.2.1 Order selection of AR(p)

An important issue in fitting our model to a set of data is to determine the appropriate order p using the most effective and least biased model selection criterion. Numerous model selection criteria for the order of $AR(p)$ can be suggested. Some are classical and some were developed in the literature when the classical approach performs poorly.

Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) are two classical model selection criteria widely used as order selection criteria for $AR(p)$ models. However, both AIC and BIC perform poorly in the context of Box and Jenkins' $ARMA$ models due to overfitting issues and severe bias caused by autocorrelation. Alternative approaches to select the best order for time series models were discussed in the literature.

For instance, Hurvich & Tsai (1989) proposed a bias correction to AIC, AIC_c, defined as the sum of AIC and a nonstochastic penalty term. The authors developed a generalized form of AIC_c in Hurvich & Tsai (1993) for vector AR models containing the special case of univariate AR models. AIC_c is applicable to Box Jenkins' $AR(p)$ models and was found to be a better approach then the classical AIC. In an unpublished work, Song, Freeland and Biswas extended the AIC_c to discrete-valued time series models, more specifically to determine the order of $AR(p)$ constructed under Pegram's operator. By means of simulation studies, the authors concluded that this new approach outperforms AIC and BIC in selecting the most adequate order for the model which is in agreement with Hurvich and Tsai's results. Note that the biasness of AIC discussed in these papers is only in the case where we have a small sample size or when the number of parameters is a moderate to large fraction of the sample size.

Azzam (2007) compares three different order selection criteria for $AR(p)$ based on the idea that all model selection criteria perform poorly in small samples. The author performs simulations to determine how each order selection criterion behaves relatively to the other. The three approaches are Schwarz's criterion (SC), Shao's criterion and AIC. Azzam concluded that for small samples SC and Shao's are similar and they both outperform the AIC. This result confirms the biasness of AIC in small samples and is in compliance with the results found in the papers mentioned above. Another interesting conclusion is that for larger samples, AIC and SC are better than Shao's.

It is important to note that in the simulations performed in the papers discussed earlier the largest sample size was $n = 1,000$. Since our data set contains a total of more than $20,000$ transactions, this puts it in the category of large samples. Thus it is best for us to use the classical AIC to determine the order of our AR(p) model. The formula for this criterion is

$$AIC = 2k - 2ln(L) \tag{4.12}$$

where k is the number of parameters, n is the number of observations or the sample size and L is the value of the likelihood function at its minimum. Normally, this value is obtained by plugging the parameters' estimates in the log-likelihood function.

As mentioned previously, Azzam (2007) concluded that for large samples AIC and SC are better than the third criterion, Shao's. Looking at the formula of SC,

$$SC = \frac{kln(n)}{n} - \frac{2ln(L)}{n} \tag{4.13}$$

we notice that it is nothing but the BIC divided by the sample size. In fact,

$$BIC = kln(n) - 2ln(L) \tag{4.14}$$

Therefore we will also use BIC as an order selection criterion.

The procedure of selecting the most adequate order for our $AR(p)$ model is to start by a preliminary analysis of the data using the Partial Autocorrelation Function (PACF) plot to obtain an idea about the interval of the order. Given several candidates for the order of the model, we use AIC and BIC to choose the best model among them. The order to be chosen is the one providing the minimal value for AIC and BIC.

4.2.2 Model Selection among the variations of the model

After determining the order of $AR(p)$ separately under each variation, we try to determine which variation of the model is the best fit. In order to do so, we compute the standardized sum of squared errors (SSSE) according to this formula

$$\begin{aligned} SSSE &= \sum_{t=0}^{t=n} \left(\frac{E[Y_t|Y_{t-1}, ...Y_{t-p}] - Y_t}{\sqrt{2\lambda_t}} \right)^2 \\ &= \sum_{t=0}^{t=n} \left(\frac{\alpha_1 Y_{t-1} + ... + \alpha_p Y_{t-p} - Y_t}{\sqrt{2\lambda_t}} \right)^2 \end{aligned} \tag{4.15}$$

The conditional expectation of Y_t given $(Y_{t-1}, ..., Y_{t-p})$ was derived in Chapter 3. We took the standardized error because the variance of our marginal distribution is not constant and varies with time. Therefore, dividing by the standard deviation would give a more accurate error term.

The variation of the model with the smallest SSSE would be the best fit for the data.

In a second stage, we examine the residuals' diagnostic plots. We focus on the time series plots, the ACF (Autocorrelation Function) and the PACF plots of the standardized residuals. We determine how well each model behaves based on the residuals.

Chapter 5

Application

5.1 Data analysis

The data to which we fit our model is composed of the stock price change during the financial transactions of an Australian firm, Broken Hill Proprietary (BHP) Limited. The price of the stock is observed and recorded each time a stock is traded and this could happen hundreds of times in a day. Our Y_t is then the change in stock price for the t^{th} trade of the month.
BHP was incorporated in 1885 and operated in the mining of silver and lead at Broken Hill, which is in western New South Wales, Australia. In 2001, BHP merged with the British-Dutch company Billiton to form BHP Billiton, which is the world's largest mining company. In August 2010, the company offered an all cash-bid for the Canadian fertilizer producer, PotashCorp located in Saskatchewan, worth about $40 million. According to the Conference Board in Canada, if the takeover occurs, Saskatchewan government revenues will be cut by at least $2 billion. After a series of negotiations, the offer is rejected and BHP Billiton withdrew the bid in November 2010.

Freeland (2010) examined the same data for the day of October 1, 2001. He noted that the transactions occurring in that day did not present any particular relationship with the duration or the volume of the transactions. The duration ranged from 1 second to 8 minutes and most of the trade volumes are less than 10,000 shares.

We will look at the entire month of October, there are 20,524 trades recorded in total. The average stock price for this month is $9.16 and the standard deviation is $0.24. The histogram of the stock price changes is presented in figure 5.1. The histogram shows a relatively symmetric data with most of the price changes being zero. In fact, 69.79% of the transactions occurred with an unchanged stock price. 14.35% moved up one tick while 14.48% moved down one tick. Also, 0.59% moved up by two ticks while and 0.51% moved down by two ticks. Note that having the frequency of up and down movements almost equal makes the data close enough to a symmetrical distribution. As for movements of higher values, 0.29% of the changes are greater than or equal to three ticks. In terms of duration the range is from 1 second to 2.8 days and the difference in the volumes of the trades between time t and time $t-1$ ranged from 0 to 4,997,003 shares. This preliminary analysis of the data suggests that there should be a certain relationship between the price change and the duration and difference in the volume between two trades.

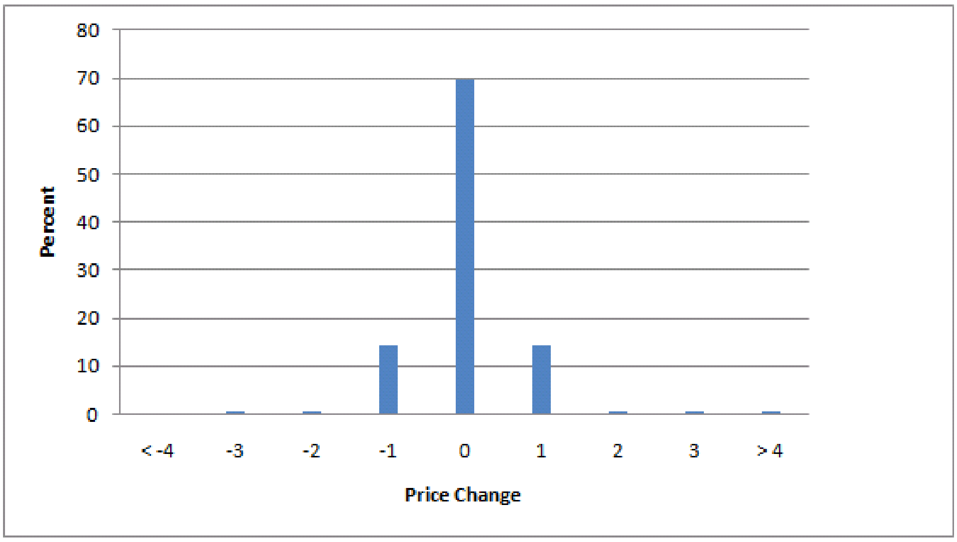

Figure 5.1: Histogram of Stock Price Changes

5.2 Parameter Estimation and Order Selection

5.2.1 Variation I

In this variation of the model, the outliers are not modeled and therefore for fitting this particular model, outliers have been excluded from the data. By outliers we mean all price changes that are greater than 3 ticks, i.e. $Y_t > 3$ and $Y_t < -3$. Although the number of outliers represents a very small portion of the data, their impact is still significant. We confirm this by comparing the behaviour of the data set with and without the outliers.

We then estimate the parameters and explain the procedure followed as well by the assumptions taken. Finally, using PACF, AIC and BIC we decide which order is the best for this model.

Excluding Outliers

Out of the 20,524 trades, 41 trades result in a stock price change greater than ±3. This number of outliers seems to be small since it barely constitutes 0.2% of the complete data. However, if we examine some factors we find that the presence of the outliers in the data makes a significant difference. We will refer to the two data sets as: complete data (which includes the outliers) and modified data (which excludes the outliers).

We first look at the estimates of the parameters. Note that the method of estimation is simplified here and merely serves the purpose of comparison between the two sets of data. A more specific method of estimation will be outlined later in this section and will be used to estimate the parameters and select the best model. We compare the parameter estimates for three models: AR(1), AR(2) and AR(3) for both the complete and the modified data in table 5.1. The estimation here treats the data as one group and does not take into account each day separately, more details on this topic will follow in this section. In addition, the starting values used to find MLE's are the estimates of the parameter using Yule-Walker approach. We notice that values of $\alpha_p' s$ are identical for both sets of data with a possible difference in the third digit. However, in the estimates of λ we have a significant difference. This difference is due to the reduced variability in the data with the absence of outliers. As noted previously, the variance of the skellam distribution is 2λ. Since this variance is decreased in the modified data so does the estimate of λ.

The absence of these outliers has also an impact on PACF. Figures 5.2 and 5.3 show the respective plots of the PACF of the complete and modified data. Similarly to the parameter estimation, those plots are only used for comparison reasons not for model selection purposes. The PACF plot will be later constructed differently to obtain some candidates for the order of our model. The plots shown below represent the partial autocorrelation computed when the data is considered as one entity.

For the complete data, the partial autocorrelation vanishes after the third lag. As for the data without the 41 oultiers, the partial autocorrelation is significant for a larger number of lags and does not vanish until the sixth lag. This affects the procedure of model selection since the order of the AR(p) will be different for each set of data.

One last note is that after excluding the outliers we remain with 20,483 data points out of which 69.92% correspond to an unchanged price. 14.37% correspond to a one tick up move and 14.51% to a one tick down move and 0.59% to a two ticks up move and 0.50% to a two ticks down move.

Table 5.1: Comparison of the estimates of the complete and modified data

	Complete Data				Modified Data			
	λ	α_1	α_2	α_3	λ	α_1	α_2	α_3
$AR(1)$	0.2021	0.2263			0.1787	0.2187		
$AR(2)$	0.2020	0.2299	0.0638		0.1756	0.2222	0.0617	
$AR(3)$	0.2032	0.2299	0.0664	0.0299	0.1746	0.2219	0.0642	0.0288

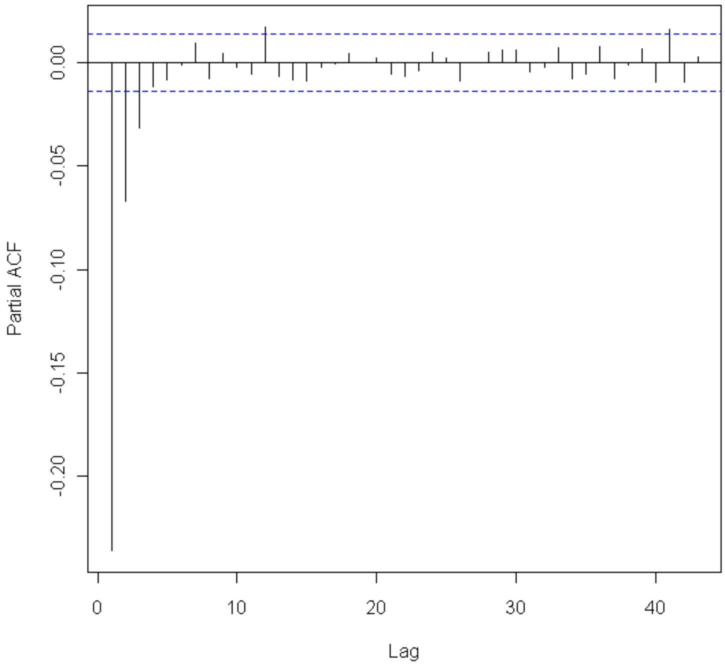

Figure 5.2: Partial Autocorrelation Function of the complete data

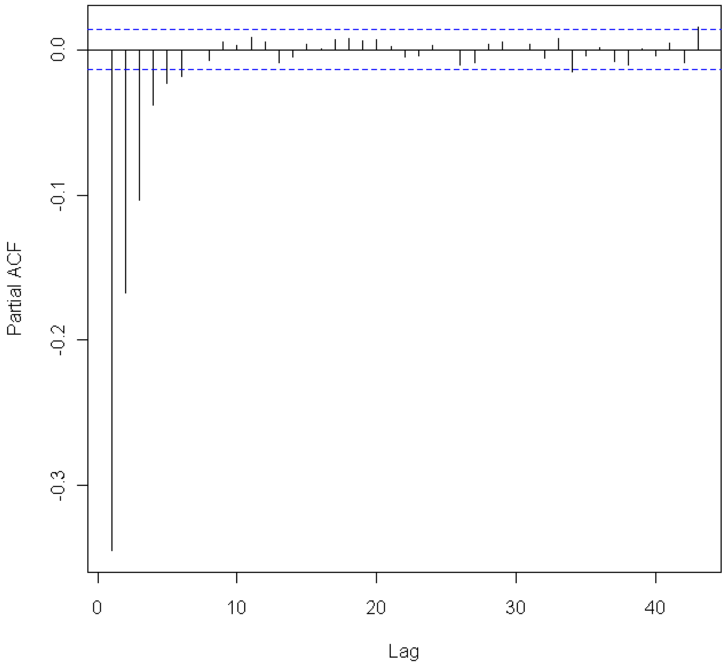

Figure 5.3: Partial Autocorrelation Function of the modified data

Parameter Estimation and Order Selection

A first step in the procedure of parameter estimation and order selection is to compute and plot the PACF, this will give us an idea of the maximum order the model might take, p_{max}. Based on the analysis of the plot, we fit several $AR(p)$ for all orders p such that $p = 1, ..., p_{max}$ and compute AIC and BIC for each model. The order selection takes place at this stage, where we examine the values of AIC and BIC and choose the order that gives the smallest value for these model selection criteria.

Before showing the results of the estimation and the order selection, we explain how the data was managed to produce the most accurate results. We assume that there is no correlation between the days in the month of October. This means that the value of Y_t depends on previous values of the price change, $Y_{t-1}, Y_{t-2}, ...$, for a given day and could not depend on previous values of the price change from previous days. Therefore, after removing all outliers, we divide the data into smaller portions according to the number of days in which transactions are taking place. Hence, excluding the week-ends during which the company is closed and no transactions are taking place, we have 23 days of activity.

We plot the PACF for each day separately, figures 5.4 and 5.5, and we also plot the weighted average PACF, figure 5.6, which should give us an idea about the range of the possible orders of the model. Note that the first three lags are significant in almost all days. PACF's for some of the days suggest a higher order such as October 5, 15 and 25, 2001. Other PACF's suggest a smaller model such as October 1 and 4, 2001.

We can see that there is a well diversified correlation structure among the days in the month of October. The best way to group these results into one PACF plot is by computing their weighted average according to the number of transactions happening in each day. We obtain the weighted PACF in figure 5.6. This plot suggests a maximum order for $AR(p)$ equal to 6. This means that we will be fitting models for orders ranging from 1 to 6.

Table 5.2 is a summary of the MLE estimates of the parameters and table 5.3 shows the values for AIC and BIC for each order. The model selected is of order 3.

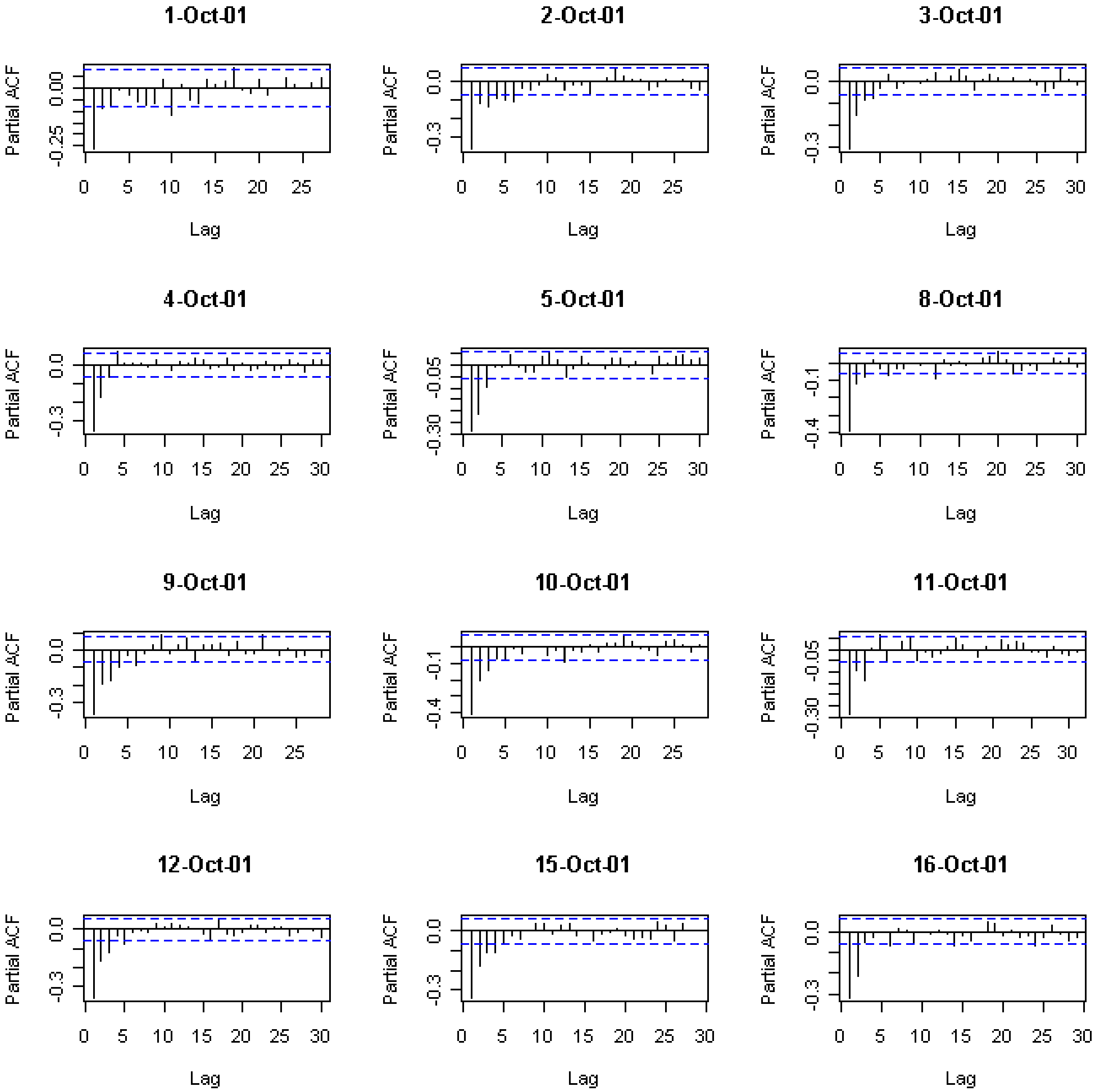

Figure 5.4: Partial Autocorrelation Function for each day, excluding outliers (a)

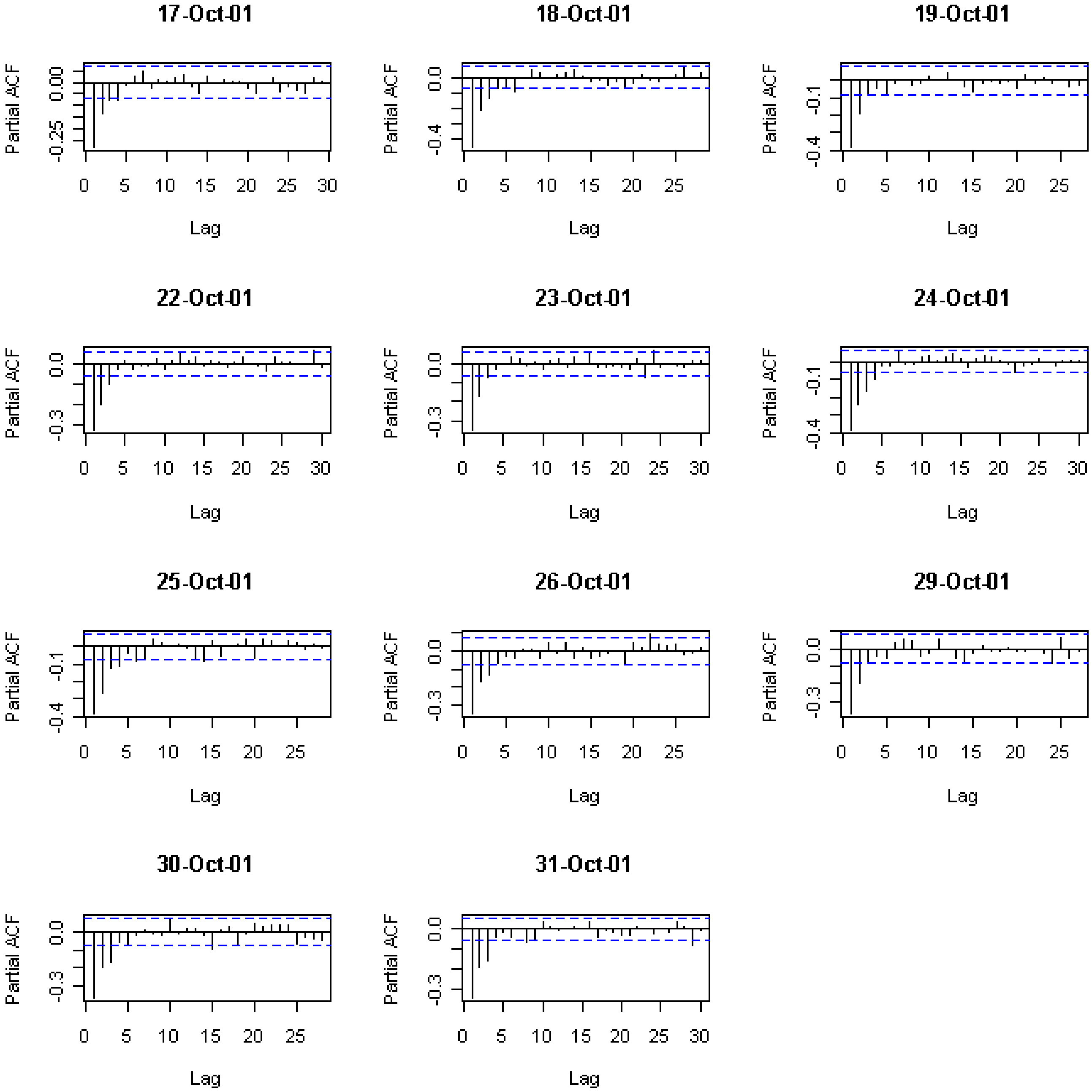

Figure 5.5: Partial Autocorrelation Function for each day, excluding outliers (b)

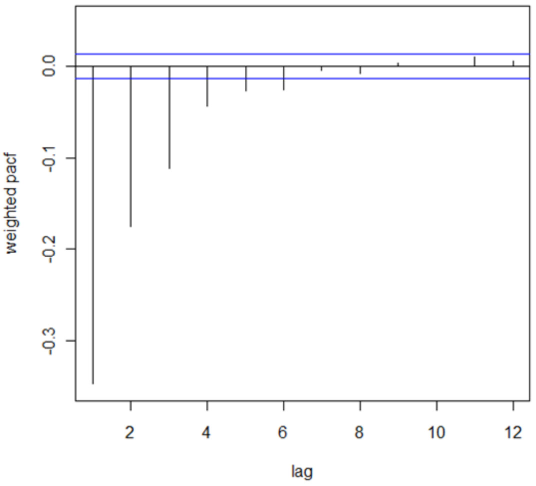

Figure 5.6: Weighted Partial Autocorrelation Function (excluding outliers)

Table 5.2: MLE estimates for Variation I

	λ	α_1	α_2	α_3	α_4	α_5	α_6
$AR(1)$	0.1782	0.2197					
$AR(2)$	0.1751	0.2234	0.0592				
$AR(3)$	0.1744	0.2232	0.0614	0.0234			
$AR(4)$	0.1743	0.2232	0.0615	0.0236	0.0025		
$AR(5)$	0.1742	0.2232	0.0614	0.0236	0.0026	0.0019	
$AR(6)$	0.1741	0.2232	0.0614	0.0235	0.0026	0.0020	0.0016

Table 5.3: Order Selection for Variation I

	AIC	BIC
$AR(1)$	33516.57	33532.41
$AR(2)$	33288.06	33311.82
$AR(3)$	33254.43	33286.11
$AR(4)$	33255.79	33295.39
$AR(5)$	33255.79	33295.39
$AR(6)$	33255.79	33295.39

5.2.2 Variation II

Here we let the variance λ vary with time which allows us to include outliers. This does not actually mean that all values of Y_t are considered. In fact, some large values of Y_t are not the result of a large trade but rather the effect of the market opening. Those data points correspond to the first transaction of the day and they are still excluded from the model at this point.

We examine the individual PACF's for each day, figures 5.7 and 5.8, as well as the weighted PACF, figure 5.9. Comparing figures 5.4 and 5.5 with figures 5.7 and 5.8, the differences are obviously insignificant for the days that originally did not include any large price changes. However, we can see that the PACF's for the days having large-volume transactions show some differences. For instance, PACF for October 9, 11, 15 and 23 showed a more significant autocorrelation for higher lags. This result is in compliance with the comparative analysis between the complete and modified data discussed in section 5.2.1.

If we examine the weighted PACF plot in figure 5.9, the maximum order suggested is 6. Therefore we will fit six models and choose among them the one with the smallest AIC and BIC.

Let us recall the expression of λ_t under this model:

$$\lambda_t = \gamma + \beta \left(\frac{|V_t - V_{t-1}|}{max(Vol)} \right)^p$$

We already noted that we must specify a constant value for $max(Vol)$ and for the power p. A first suggestion is to choose $max(Vol)$=10,000 and a value for p and fit the six models and choose the most adequate order. Once we determine the order of our model, we fix all parameters and treat the power p as a variable ranging between 1 and 2. We fit the model with all possible values of p and choose the one having the smallest minimum of the log-likelihood function. We start by using $p = 1.5$ as a starting value for p, we estimate the parameters and compute the AIC and BIC for each model; AR(3) is chosen, see tables 5.4 and 5.5.

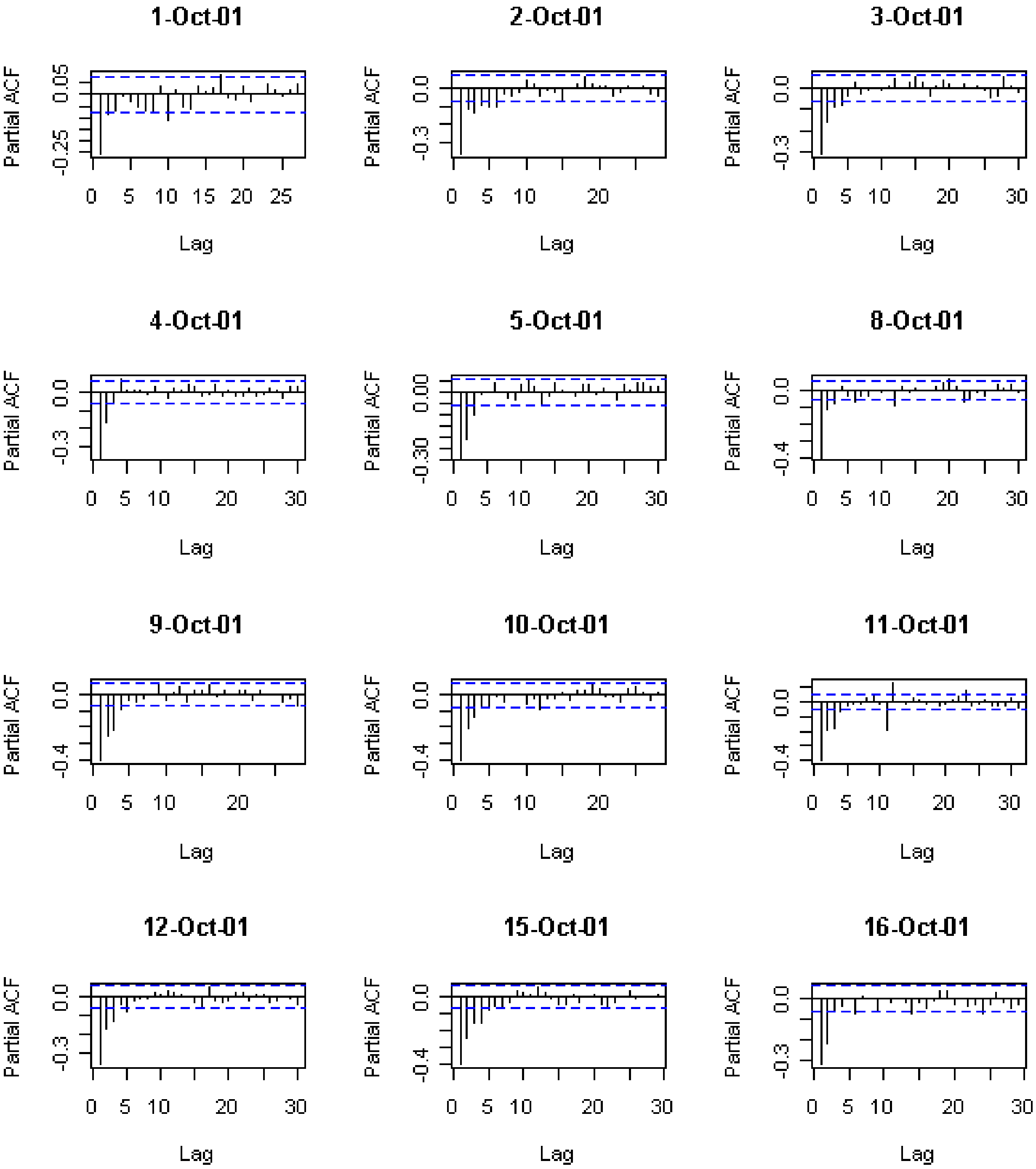

Figure 5.7: Partial Autocorrelation Function for each day, including outliers (a)

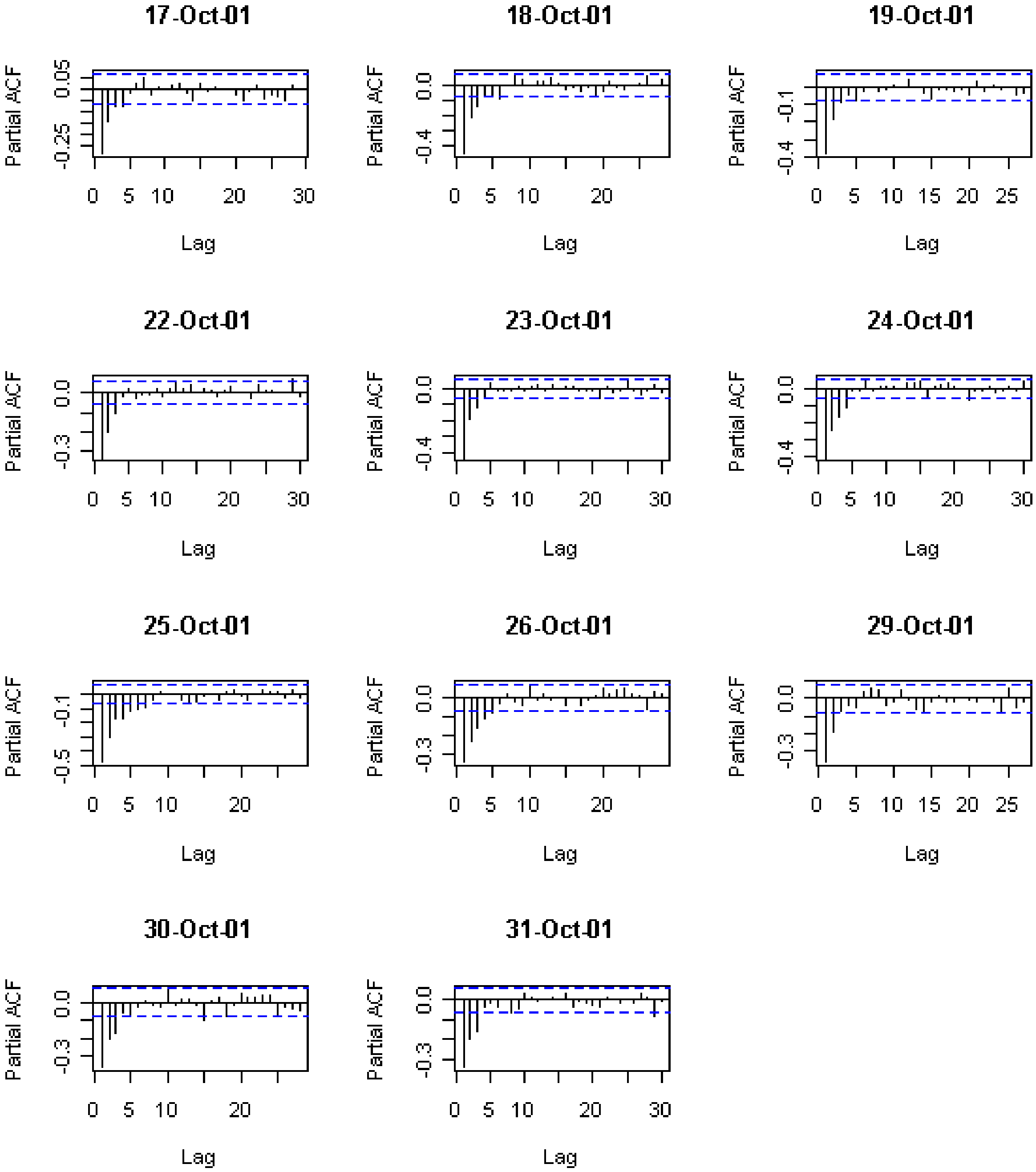

Figure 5.8: Partial Autocorrelation Function for each day, including outliers (b)

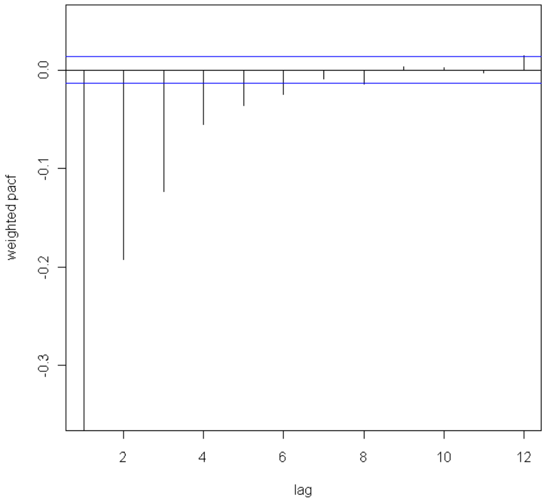

Figure 5.9: Weighted Partial Autocorrelation Function (including outliers)

Table 5.4: MLE estimates for Variation II, $max(Vol) = 10,000$

	γ	β	α_1	α_2	α_3	α_4	α_5	α_6
$AR(1)$	0.1702	0.0109	0.2219					
$AR(2)$	0.1666	0.0120	0.2256	0.0604				
$AR(3)$	0.1655	0.0127	0.2254	0.0627	0.0246			
$AR(4)$	0.1653	0.0128	0.2255	0.0628	0.0249	0.0036		
$AR(5)$	0.1651	0.0129	0.2254	0.0627	0.0248	0.0038	0.0024	
$AR(6)$	0.1650	0.0130	0.2254	0.0627	0.0248	0.0038	0.0024	0.0021

Table 5.5: Order Selection for Variation II, $max(Vol) = 10,000$

	AIC	BIC
$AR(1)$	34187.69	34211.47
$AR(2)$	33948.49	33980.21
$AR(3)$	33911.27	33950.91
$AR(4)$	33911.94	33959.51
$AR(5)$	33911.29	33966.79
$AR(6)$	33912.81	33976.24

Now we fix all the parameters and let p vary from 1 to 2. We fit the models with several iterations and compare the minima of the log-likelihood function. Figure 5.10 shows a plot of the minimum of the log-likelihood function as a function of the power p. We conclude that the most adequate power to use is 1.7.

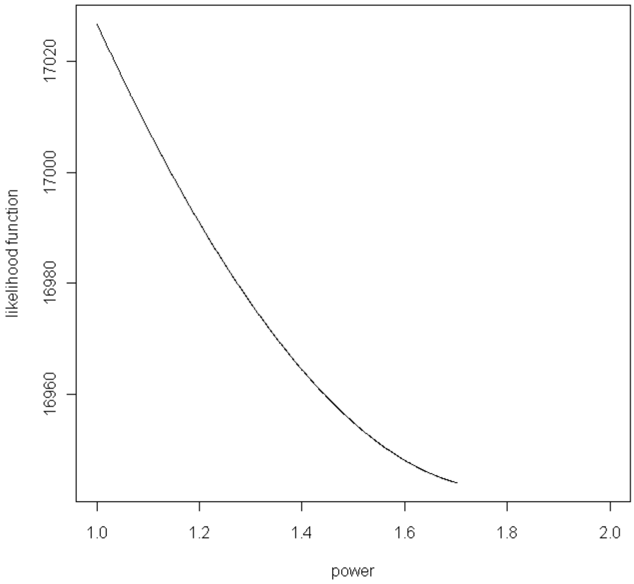

Figure 5.10: Likelihood function as a function of the power p, $(max(Vol) = 10,000$ shares)

Notice that there is a trend followed in the relationship between the likelihood function and the power; the higher the power, the lower is the minimum of the log-likelihood. Another thing to note is that 1.7 is the highest power for which the process is converging. All powers $p > 1.7$ are not applicable when we have $max(Vol) = 10,000$ shares.

In a final stage we compute the parameters' estimates given $max(Vol) = 10,000$ and $p = 1.7$ and find:

Parameter	Estimate
γ	0.1667
β	0.0090
α_1	0.2254
α_2	0.0628
α_3	0.0248

Another suggestion is to let $max(Vol) = 100,000$ and $p = 2$ and proceed directly to find the ML estimates for the rest of the parameters. We will see that when we choose this value for $max(Vol)$, we will no longer have the issue with the convergence for the values of $p > 1.7$. To confirm this, we look at figure 5.11 and see the same trend as in figure 5.10. The only difference is that the plot continues up to $p = 2$ with the latter being the best power value.

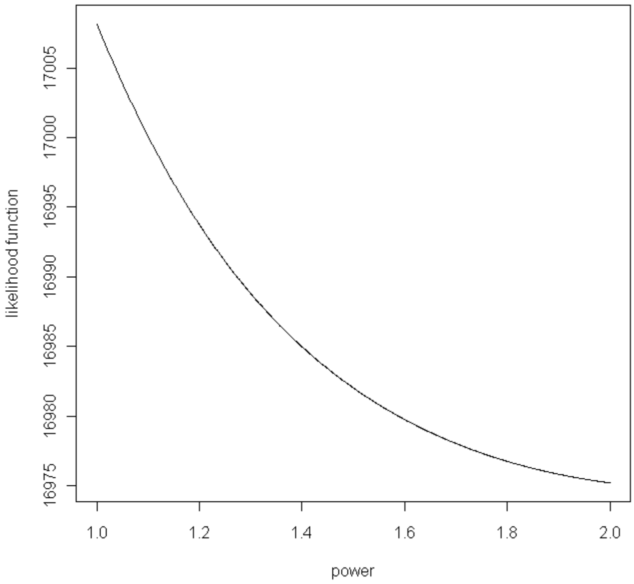

Figure 5.11: Likelihood function as a function of the power p, ($max(Vol) = 100,000$ shares)

Since we are still working with the same data, it means that we refer to the same PACF and consequently fit models up to order 6 before deciding on the order according to AIC and BIC values. The estimates of the parameters are in table 5.6 and according to table 5.7, we choose once again AR(3).

Table 5.6: MLE estimates for Variation II, $max(Vol) = 100,000$

	γ	β	α_1	α_2	α_3	α_4	α_5	α_6
$AR(1)$	0.1783	0.1419	0.2258					
$AR(2)$	0.1742	0.1419	0.2113	0.0604				
$AR(3)$	0.1723	0.1419	0.2105	0.0628	0.0243			
$AR(4)$	0.1725	0.1419	0.2113	0.0647	0.0267	0.0034		
$AR(5)$	0.1709	0.1419	0.2105	0.0659	0.0280	0.0038	0.0017	
$AR(6)$	0.1713	0.1419	0.2084	0.0669	0.0291	0.0030	0.0034	0.0024

Table 5.7: Order Selection for Variation II, $max(Vol) = 100,000$

	AIC	BIC
$AR(1)$	34218.36	34211.47
$AR(2)$	33992.10	34023.81
$AR(3)$	33960.39	34000.03
$AR(4)$	33962.26	34009.83
$AR(5)$	33963.92	34019.42
$AR(6)$	33968.89	34032.31

It would be interesting to graphically look at how λ_t behaves as a function of the difference in the volumes of the trades. If we look at figure 5.12 we notice that the value of λ_t is increasing as the difference in the volume increases, meaning the variability of the price change is more significant when we have a transactions with large portfolio.

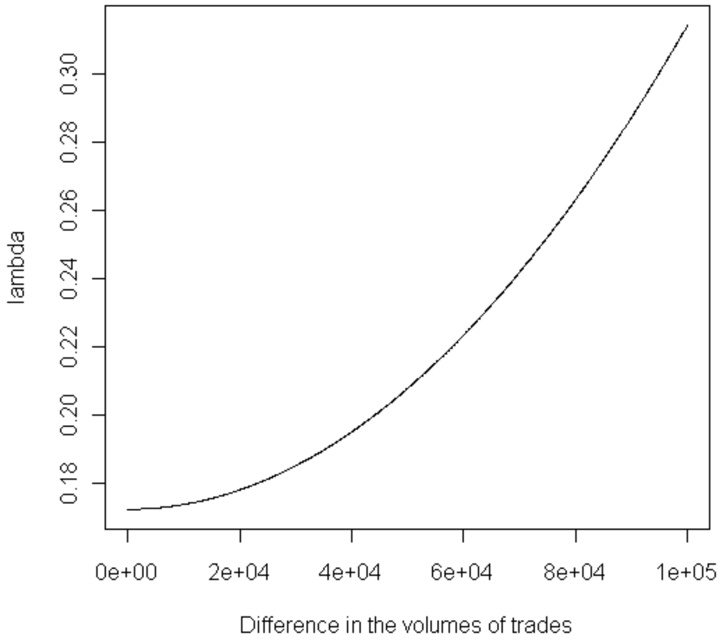

Figure 5.12: λ_t as a function of $|V_t - V_{t-1}|$

5.2.3 Variation III

In this model, λ_t is expressed as a function of both the volume and the duration. In terms of the relation of the price change to the volume we will use the second suggestion in Variation II of the model which sets $max(Vol) = 100,000$ shares and the power $p = 2$. As for the part that relates the duration to the price change which is $\left(\frac{D_t - D_{t-1}}{max(D_t)}\right)^q$, we choose $max(D_t) = 10,000$ seconds and $q = 2$.

We are now looking at the data as one entity since we related the duration to the price change and therefore defined a correlation structure that links all the transactions together. This means that the first transaction of the day is no longer excluded and it is related to the last transaction of the previous day.

In order to determine a starting value for the order of our model, we look at the PACF of the complete data, since it is now considered as one group, figure 5.2; the maximum order suggested is 3. We fit our three models, see table 5.8. AIC and BIC values suggest $AR(3)$ to be the best model, table 5.9.

Table 5.8: MLE estimates for Variation III

	γ	β	δ	α_1	α_2	α_3
$AR(1)$	0.1779	0.1419	0.3053	0.2170		
$AR(2)$	0.2816	0.1419	0.1715	0.2407	0.0709	
$AR(3)$	0.1719	0.1419	0.3012	0.2109	0.0593	0.0249

Table 5.9: Order Selection for Variation III

	AIC	BIC
$AR(1)$	34786.20	34817.92
$AR(2)$	35737.97	35777.61
$AR(3)$	34495.57	34543.15

5.3 Model Selection

In this final section of the application we look at the three variations of the model we fitted and try to determine which one fits best the data. As mentioned in chapter 4, we will be using both SSSE and the residuals' diagnostic plots to determine the best variation of the model.

5.3.1 Standardized Sum of Squared Errors

Looking at table 5.10, we notice that the model with the lowest SSSE corresponds to the second variation with a $max(Vol) = 100,000$ shares.

The worst model is the third variation, with the duration covariate. The reason behind this could be related to having a small portion of the data with large duration and it could also be that the relation between the duration and the variability of the price of the stock is not quadratic.

Table 5.10: Model Selection

Model	SSSE
Variation I	17340.96
Variation II, $max(Vol) = 10,000$	21396.64
Variation II, $max(Vol) = 100,000$	15942.12
Variation III	38934.76

5.3.2 Residuals' diagnostic plots

We examine the time series plots as well as the PACF and ACF plots of the standardized residuals to assess each variation of the model.

Time Series Plots

We notice the absence of any particular pattern in all the residuals plots. However, as the model gets more complicated and include more outliers in the data, we witness the presence of more outliers in the times series plots of the residuals. For instance, in Variation I of the model, the residuals' values range approximately between -4 and 4, figure 5.13. This range becomes larger in the subsequent plots.

One important thing to note is that, although Variation II of the model with $max(Vol) = 100,000$ shares includes outliers, it has a more concised interval for the residuals. The plot is similar to the one for Variation I and it shows a more constant standardized variance except for the sole outlier around $t = 12,000$, figure 5.15. We notice further that the residuals' plot of Variation III, figure 5.16, does not show a constant standardized variance and represents the largest amount of outliers among all the plots. This suggests that the model does not fit very well the data. This result is in compliance with the outcome of the SSSE.

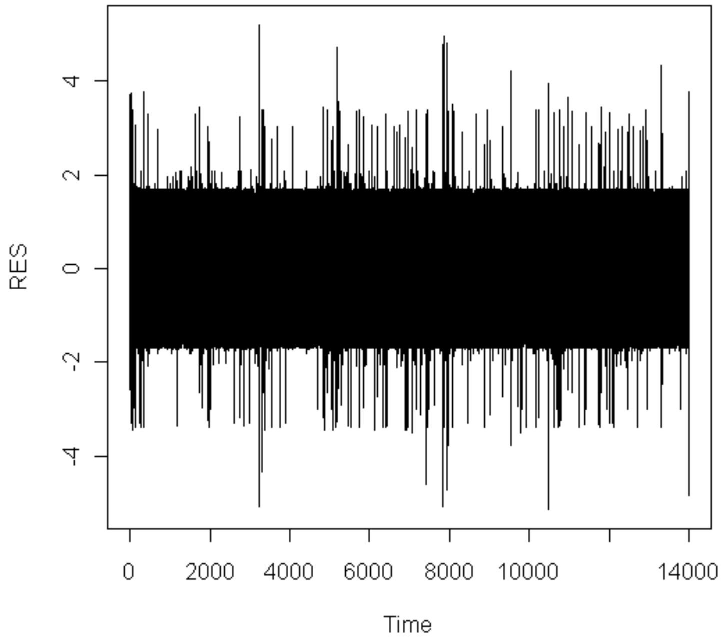

Figure 5.13: Residuals' Time Series plot, Variation I

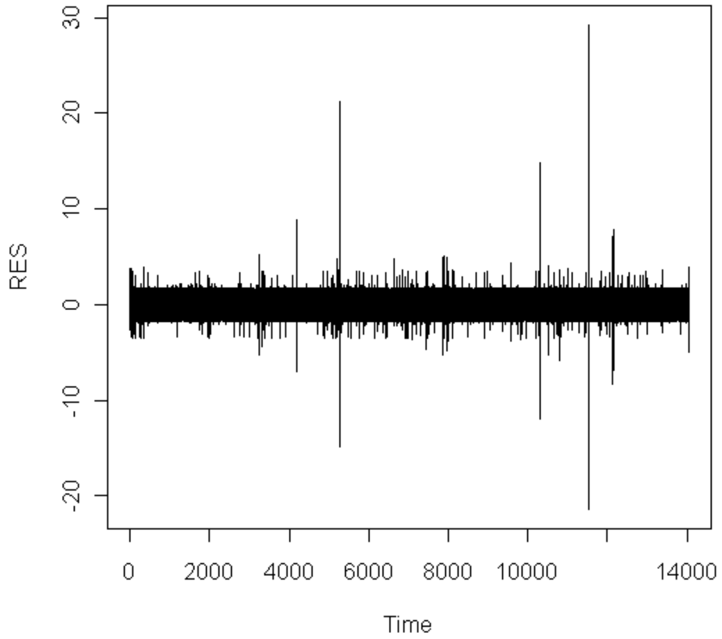

Figure 5.14: Residuals' Time Series plot, Variation II with $max(Vol) = 10,000$ shares

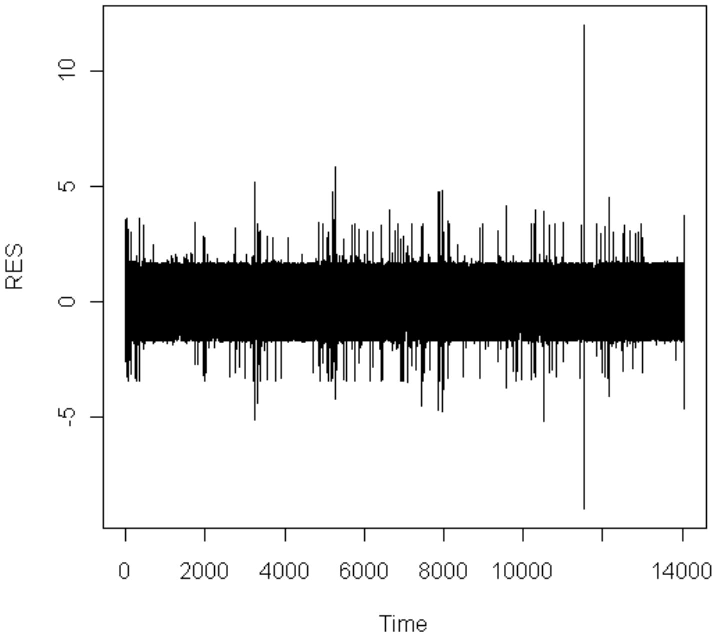

Figure 5.15: Residuals' Time Series plot, Variation II with $max(Vol) = 100,000$ shares

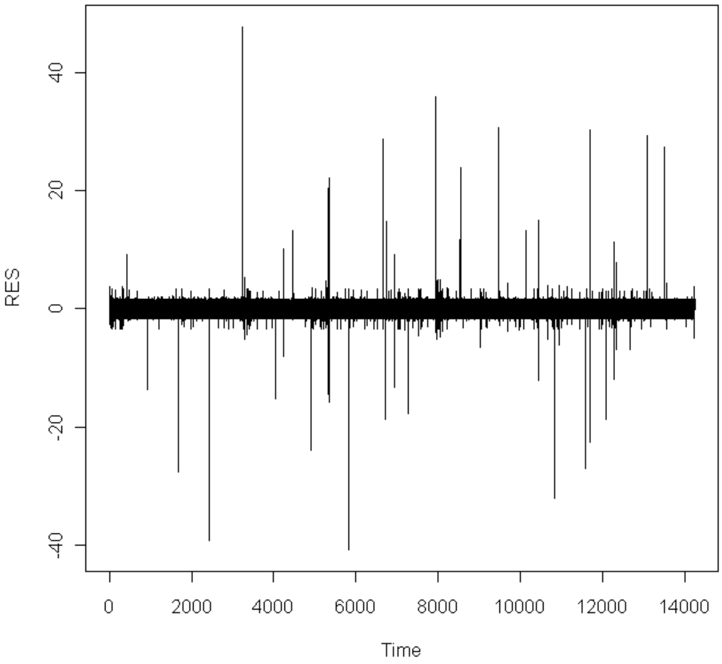

Figure 5.16: Residuals' Time Series plot, Variation III

PACF and ACF plots

Both PACF and ACF plots behave similarly among the models and consequently lead to the same conclusion, therefore we examine only PACF plots in this section. Furthermore, the plots show comparable results for Variation I and Variation II models so we only compare the plots of Variation II with $max(Vol) = 100,000$ shares, figure 5.17, with Variation III, figure 5.18.

We notice in that case that Variation III behaves much better than Variation I and II. More specifically, the correlation is better captured as it is noticeable that in figure 5.18 the standardized residuals are not correlated as opposed to the significant correlation present among the residuals in figure 5.17. This leads us to the conclusion that the covariate of time which is only included in Variation III, plays a major role in capturing the correlation in the data. However, according to the results of the SSSE and the time series plots, the way the duration was modeled in that case should be investigated as it does not provide the best fit for the data.

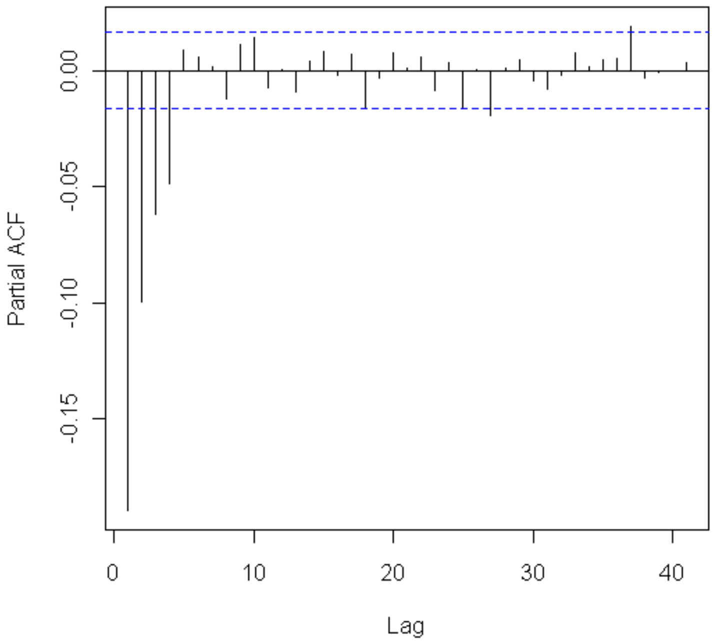

Figure 5.17: Residuals' PACF plot, Variation II with $max(Vol) = 100,000$ shares

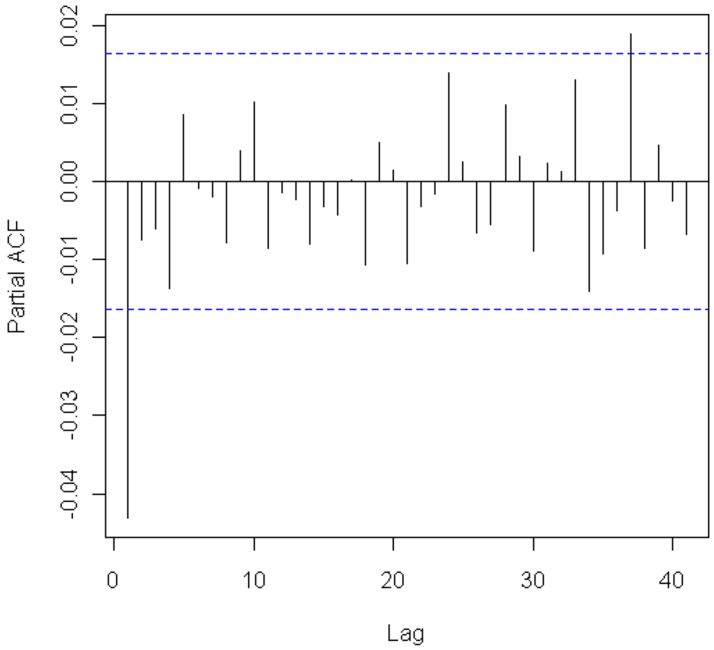

Figure 5.18: Residuals' PACF plot, Variation III

Chapter 6

Conclusions

In this chapter we present a summary of the study done in the thesis as well as an outline of the results of the application. We also make some recommendations for future studies that could further extend this research.

6.1 Summary

In this thesis we developed a discrete-valued time series model, which is a type of model that cannot be covered by Box and Jenkins' methodologies. We use Pegram's operator to build our $AR(p)$ model which is an extension as well as a special case of some of the previous models developed in the literature. Some of these models did not cover all types of discrete data (e.g. excluding negative integers) while others presented some restrictions in the interpretation and extension of the model. For instance, some had issues in their interpretation due to the structure of the model and others could not be extended to arbitrary orders. Pegram's operator served as a very flexible operator taking any type of discrete distribution and making it easier to develop higher order models. In addition, models constructed using this operator resembled Box and Jenkins' models in terms of form, correlation structure and other properties.

The marginal distribution of our $AR(p)$ is the Skellam distribution which is the difference between two independent Poisson random variables. This is a symmetric distribution which means that both negative and positive integers are in the sample space. As a result, we can have both negative and positive correlations. Pegram's operator was extended in a way to accomodate for this type of correlation and this was shown in simple numerical examples. Properties of our model were also discussed such as the conditional expectation, the autocorrelation function and the moment generating function. The extension for Pegram's operator can also be used to develop models where we have negative and positive correlation at the same lag, we discussed an $AR(1)$ model under this specification.

We presented three variations of our model. The first one is the simplest with a Skellam marginal distribution with parameter λ. The other two variations let λ vary with time and included more covariates inside this parameter to model extreme values in financial data. We looked at the parameter estimation under the ML method and developed score functions for the three variations of the model.

We also discussed order selection criteria for each model and used AIC and BIC to determine the best order. Then we studied the selection of the best variation of the model using the Standardized Sum of Squared Errors (SSSE) and the standardized residuals' diagnostic plots.

6.2 Results

As an illustration, we fit our models to a stock price change data during the financial transactions of BHP Billiton Proprietary Limited. We analyze the data and fit the three variations of the model. In the second variation, we try two different scenarios: the first takes $max(Vol) = 10,000$ shares and $p = 1.7$, and the second scenario takes $max(Vol) = 100,000$ shares and $p = 2$.

In the order selection procedure, we start by examining PACF plots to find starting values for the order and, after finding the ML estimates of the parameteres, we compute AIC and BIC for each possible order of the model and choose the order for which AIC and BIC have the smallest values. $AR(3)$ was selected for all variations of the models.

We last look at finding the best fit among the three variations we had. According to SSSE and the time series plots of the standardized residuals, the best model for our data is the second variation of $AR(3)$ with $max(Vol) = 100,000$ shares and $p = 2$. However, by examining the PACF and ACF plots of the standardized residuals, we notice that Variation III is the best model to capture the correlation in the data which suggests that the duration must be included in the model to assure the residuals are not correlated.

6.3 Recommendations

Two main extensions could follow this research. One is related to the third variation of the model and the other extension looks at the mixed lag correlation.

As noted previously, SSSE picked the second variation. This variation is not the most complicated version of the model, more specifically it does not relate the duration to the price change and does not model well the correlation. Therefore, it is believed that a relationship exists between the price change and the duration between transactions. This is due to the price jump at the opening of the market and the outcome of the PACF and ACF plots of the standardized residuals for Variation III. Therefore, it would be interesting to investigate the true relation between λ_t and D_t and try to improve this portion of the model. This means that duration could be re-modeled or it can be fitted to another type of data.

Another recommendation would be to study further the idea of mixed lag correlation. It has been noted that higher orders in such model can become cumbersome and that is why only the first order was developed. It would be interesting to investigate other possibilities to model mixed lag correlation such as trying a different operator.

Bibliography

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

Alzaid, A.A., Al-Osh, M.A. (1990). An Integer-Valued pth-Order Autoregressive Structure (INAR(p)) Process. Journal of Applied Probability 27, 314-324.

Azzam, I. (2007) The Effect of Model-Selection Uncertainty on Autoregressive Model Estimates International Research Journal of Finance and Economics 11, 79-92

Biswas, A., Song, P. X.-K. (2009). Discrete-Valued ARMA Processes. Statistics and Probability Letters 79, 1884-1889.

Box, G.E.P., Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, revised edition. Holden-Day, San Francisco.

Bu, R., McCabe, B.P.M. (2008) Model Selection, Estimation and Forecasting in INAR(p) models: a likelihood based Markov Chain approach International Journal of Forecasting, 24, 151-162.

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

Freeland, R.K. (2010) True Integer Value Time Series AStA Advances in Statistical Analysis 94, 217-229

Freeland, R.K., McCabe, B.P.M. (2004a). Analysis of Low Count Time Series Data by Poisson Autoregression. Journal of Time Series Analysis 25, 701-722.

Freeland, R.K., McCabe, B.P.M. (2004b). Forecasting discrete valued low count time sereis. International Journal of Forecasting 20, 427-434.

Hurvich, M.C., Tsai, C.-L. (1989). Regression and time series model selection in small samples. Biometrika 76, 297-307.

Hurvich, M.C., Tsai, C.-L. (1993). A Corrected Akaike Information Criterion for Vector Autoregressive Model Selection Journal of Time Series Analysis, 14: 271279.

Irwin, J.O. (1937). The Frequency Distribution of the Difference between two Independent Variates following the same Poisson Distribution. Journal of the Royal Statistical Society: Series A 100(3), 415-416.

Jacobs, P.A., Lewis, P.A.W. (1978a) Discrete Time Series generated by Mixtures. I: Conditional and runs properties. Journal of the Royal Statistical Society B 40, 94-105.

Jacobs, P.A., Lewis, P.A.W. (1978b) Discrete Time Series generated by Mixtures. II: Asymptotic Properties. Journal of the Royal Statistical Society B 40, 222-228.

Jacobs, P.A., Lewis, P.A.W. (1978c) Discrete Time Series generated by Mixtures. III: Autoregressive Processes (DAR(p)). Naval Postgraduate School Technical Report NPS55L-78-022.

Jacobs, P.A., Lewis, P.A.W. (1983) Stationary discrete autoregressive-moving average time series generated by mixtures. Journal of Time Series Analysis 4, 9-36

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

Kanter, M., (1975). Autoregression for Discrete Processes mod 2. Journal of Applied Probability 12, 371-375.

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

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

Pegram, G.G.S. (1980). An Autoregressive Model for Multilag Markov Chains Journal of Applied Probability 17, 350-362.

Russel, J. R., Engle, R. F. (1998). Econometric Analysis of Discrete-Valued Irregularly-Spaced Financial Transactions Data using a New Autoregressived Conditional Multinomial Model CRSP Working Paper No. 470; University of California at San Diego Working Paper No. 98-10

Russel, J. R., Engle, R. F. (2005). A Discrete-State Continuous-Time Model of Financial Transaction Prices and Times: The Autoregressive Conditional Multinomial-Autoregressive Duration Model Journal of Business and Economic Statistics 23, 166-180

Skellam, J.G. (1946). The frequency distribution of the difference between two variates belonging to different populations. J. of the Royal Statistical Society: Series A, 109 (3), 296.

Steutel, F.W. (1979). Infinite Divisibility in Theory and Practice (with discussion) Scandinavian Journal of Statistics. 6, 57-64.

Steutel, F.W., van Harn, K. (1979). Discrete Analogues of Self-Decomposability and Stability. The Annals of Probability 7, 893-899.