Online Detection of Mean Reversion in Algorithmic Pairs Trading

Seung Jin Han

School of Mathematics and Statistics

University of Sheffield

A thesis submitted for the degree of

Doctor of Philosophy

November 2013

Acknowledgements

It is great pleasure to acknowledge people who have given me guidance, help, and encouragement.

First and foremost, I would like to offer my sincere gratitude to Dr. Kostas Triantafyllopoulos. As a supervisor, he has encouraged and helped me in many ways during the last four years of my studies.

I extend my gratitude to my family members. My late father (Keung-woo Han), my mother (Eunkyoung Oh), my sister (Jiyeon Han) and brother-in-law (Taekyoung Ko), their daughter and my nephew (Jaehee Ko), my wife (Sun Young Park), parents-in-laws (Byoungil Park and Yunsun Lee), and a sister-in-law(Hyunyoung Park), they are and always has been my supporters.

My deepest appreciation goes to my wife, Sun Young Park. She is and always has been a good friend and colleague of mine.

Abstract

This thesis is concerned with online detection of mean-reversion in algorithmic pairs trading where a pair of assets is chosen when their prices are expected to show similar movements. In pairs trading, mean-reversion of the spread, defined as the price difference of a pair of financial instruments, is assumed, and we propose that the mean-reverted patterns can be detected online. For this, a new algorithm for variable forgetting factor using the conjugacy of distributions and an inference for multicategorical time series in dynamic models are developed. Two algorithms for variable forgetting factor are also introduced using the steepest descent method and the Gauss-Newton method each from the field of signal processing and control engineering. Performances of the three variable forgetting factor algorithms are evaluated by the mean square errors and the detection rate. However, the mean-reversion is not related to the stationarity in time series, in particular when it is locally detected. Thus, the behaviour of the spread or its mean-reversion needs to be carefully monitored as well. Considering that the detection of mean-reversion relies on a parameter estimate of the state in dynamic linear model, the estimate is located to a category specified by the modeller. This behaviour of the estimate is monitored in dynamic generalised linear model for multicategorical time series where sequential Monte Carlo methods are applied for an inference as a simulation-based approach. As an illustration, algorithmic pairs trading is implemented and shown to be successful even with simple trading rules, given the daily stock prices from the stock exchange.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces algorithmic pairs trading, gives the aims and objectives of the thesis, and provides information on the organisation, the terminology and the notation of the thesis. An introduction of algorithmic pairs trading is also given as an introductory part of Triantafyllopoulos and Han (2013).

1.1 Algorithmic Pairs Trading

Pairs trading is a market neutral trading strategy so that a trader's risk exposure is indifferent to the market trend by holding a long and a short position together. A long position is held by buying an asset and a short position by short-selling the other. According to Vidyamurthy (2004), the outset of pairs trading is marked by a group of quants at Morgan Stanley in the mid 1980s. Algorithmic trading deploys trading strategies and related decisions that can be implemented on a computer system and executed without human intervention. Thus, algorithmic pairs trading can be an algorithmic trading dealing with a pair of financial instruments. Recently, there has been a growing interest in pairs trading and related market neutral trading approaches which can be found in Elliott et al. (2005), Gatev et al. (2006), Montana and Parrella (2008), Montana et al. (2009), and Zhang and Zhang (2008). For a book-length discussion, the reader is referred to Pole (2007), Vidyamurthy (2004), Whistler (2004), Ehrman (2005), Chan (2008), and Chan (2013).

Defining the spread of two assets A and B as the difference of the prices of A and B, pairs trading assumes that the spread attains an equilibrium or that the spread in the long run reverts to its historical mean. The main idea behind pairs trading is to propose trades based upon the relative temporary mispricings between the two assets. For example, suppose that the equilibrium of the spread is \$10 (in USD) and today the two assets trade at \$40 and \$10 respectively, or with spread of \$30 (=\$40-\$10). Then, pairs trading suggests to go short (or short-sell) asset A (as this is likely to be overpriced at \$40) and to go long (or buy) asset B (as this is likely to be underpriced at \$10). If the spread reverts to its historical mean, the price of asset A will decrease and/or the price of asset B will increase.

This approach is heavily dependent on the assumption of mean-reversion of the spread. If this assumption is violated, the trader may buy an overpriced asset, which is losing its value, or may short-sell an undervalued asset, which commit the trader to high buying costs in the future; both of these actions result in significant loss. Mean reversion implies that the spread fluctuates around the equilibrium level and thus if today the price of an asset goes up, it will go down in the near future and vice versa. Conversely, a breakdown of mean-reversion implies that any shock in the spread may be permanent and hence there is no guarantee that if today the price of an asset goes up, it will go down in the future. This is what happened at a Wall Street operating hedge fund of Long Term Capital Management, which had to be bailed out in 1998 by the Federal Reserve Bank of New York over a \$3.625 billion loss including \$286 million in equity pairs according to Lowenstein (2002). This story reveals that spread speculation, in particular regarding to short-selling assets, may lead to significant loss if mean-reversion is not monitored systematically and if the uncertainty of spread prediction is not studied carefully. In practice, assets may exhibit local mean-reversion. For example, there may be periods of mean-reversion followed by periods of a breakdown of mean-reversion according to Pole (2007) and Triantafyllopoulos and Montana (2011). As a result, it is proposed that by detecting periods of mean-reversion, the trader can find opportunities for trading.

1.2 Aims and Objectives of The Thesis

This thesis is concerned with online detection of mean-reversion in algorithmic pairs trading. Considering a dynamic linear model for the spread time series, we propose that mean-reverted patterns can be detected in real time. Adopting a dynamic generalised linear model for multi-categorical time series, the mean-reversion of the spread can be monitored online.

In this thesis, a newly developed variable forgetting factor algorithm in dynamic linear model and the dynamic generalised linear model proposed for multi-categorical time series are applied for online detection of mean-reversion of the spread time series and algorithmic pairs trading. As an illustration, an opportunity of algorithmic pairs trading is proposed with simple trading rules.

This thesis introduces the variability of forgetting factor, also known as discount factor, in the class of dynamic linear model. This is motivated by variable forgetting factor algorithms using the steepest descent method and the Gauss-Newton method from the field of signal processing and control engineering. New algorithm for variable forgetting factor is developed and proposed using the conjugacy of distributions where the prior and the posterior distributions belong to the same family of distributions.

For multi-categorical time series, the class of dynamic generalised linear model is employed with multinomial distribution. The recursions of parameter estimation in the model are proposed when the posterior distribution of the states is approximated by two different approaches. In the first approach, the mean vector and the covariance matrix of the posterior distribution of the states are approximated using Bayes linear methods while the particle filter as a simulation-based method is applied in the second approach.

1.3 Layout of the Thesis

Chapter 2 provides literature review on Bayesian time series and forecasting, pairs trading, the class of the dynamic linear model and the Kalman filter as a method of recursive inference, variable forgetting factor, the class of the dynamic generalised linear model, and the linear Bayesian method and the particle filters as for the approximate inference on the posterior distribution. This literature review is focused on the relevant methodologies covered in this thesis. It aims to help a reader understand the key ideas from the selection of references for each subject.

Chapter 3 discusses the algorithms for variable forgetting factor from the fields of signal processing and control engineering. In the chapter, two most widely used algorithms, each of which employs the steepest-descent method and the Gauss-Newton method respectively, are introduced and implemented in dynamic linear model. A new algorithm for variable forgetting factor, which relies on the conjugacy of distributions in Bayesian statistics for sequential updating, is developed and proposed as an improvement on previous work by Triantafyllopoulos and Montana (2011) for online detection of mean-reversion.

Chapter 4 employs dynamic generalised linear model with multinomial distribution for multi-categorical time series. As no closed form of the posterior distribution is available, the chapter considers the linear Bayesian method and the particle filter for approximate inference. While applying the particle filter for the approximation, multivariate normal distribution is assumed for an importance density of the states. In Chapter 5, the results of this chapter are used for real-time monitoring of mean-reversion.

Chapter 5 proposes an opportunity of algorithmic pairs trading and illustrates successful implementation of the newly developed algorithm for variable forgetting factor from Chapter 3 and the dynamic generalised linear model for multi-categorical time series as an online monitoring process of mean-reversion from Chapter 4. It is shown that algorithmic pairs trading can make a profit in the market even with a simple trading rule.

Chapter 6 gives final remarks as conclusion, and discusses further research opportunities.

1.4 Terminology & Notation

Greek letters are used for the parameters, and Roman letters for the observed or the observable scalars, vectors and matrices. The vectors are in bold, but the scalars and the matrices are in plain. For example, Y_t represents the univariate series while $\mathbf{Y}_t$ does for the vector of $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{i,t})'$ for $i = 1, 2, \ldots, N$ at t.

As this thesis covers several topics from different fields of academics, some terminologies need to be clearly explained what they represent in this thesis. For example, a forgetting factor is a widely used common terminology among the engineers for a discount factor in Bayesian forecasting. In this thesis, a forgetting factor as a terminology is adopted rather than a discount factor because the variability of it is rigorously discussed in the fields of engineering. In addition, the state space model, the hidden Markov model, and the dynamic models are meant to be the same in literature by many authors, although they have different roots and usages in theory. The state space model has its roots in the space program where the states are the actual locations of a target. In the hidden Markov model, the states are unknown, but finite, taking one of the possible and expected outcomes in discrete while they are continuous and unknown in the state space model. The dynamic linear model assumes the Gaussianity and the linearity although the other two does not. The dynamic generalised linear model assumes the linearity as well. Authors even in the field of time series and econometrics are found not to take much care of these differences, using mostly 'the state space model' for the dynamic model and the hidden Markov model in literature. Thus, in this thesis, the state space model are meant to be the dynamic linear model and the dynamic generalised linear model.

Chapter 2

Pairs Trading, Time Series, and Dynamic Models

2.1 Introduction

This chapter reviews literature on time series and Bayesian forecasting, pairs trading, the class of dynamic linear model, variable forgetting factors, the class of dynamic generalised linear model for non-normal multivariate time series, and the particle filters. Some topics have vast amounts of literature, but only a few related to this thesis are reviewed in this chapter.

2.2 Time Series and Bayesian Forecasting

A time series is a collection of data recorded at equally spaced time intervals over a period of time and indexed by time t. The aim of the classical time series analysis is to understand the data and extract patterns such as trend, seasonality, and irregular variation from given observations. Assuming that the findings and/or the patterns continue, future values of the time series can be predicted.

Forecasting in time series dates back to the 1950s. The exponential smoothing of Holt (1957) is followed by a one-parameter polynomial forecasting model of Brown (1959), the adaptive smoothing techniques of Box and Jenkins (1962) and Brown

(1963). While smoothing depends on all available observations, filtering aims to update the system as each observation becomes available. In the fields of system and control engineering, the Kalman filter, developed by Kalman (1960) and Kalman and Bucy (1961), provides the recursions for the parameters in the state space model.

Harrison and Stevens (1971) introduce the application of the Kalman filter for short-term forecasting based on Bayesian principles, and then Harrison and Stevens (1976) define the dynamic linear model, which assumes the Gaussianity and the linearity in the state space model. When the Gaussianity and the linearity of a model are assumed, the Kalman filter can be the natural choice for online inference in Bayesian formulation, giving the analytic solutions for sequential updating.

When an assumption on the Gaussianity is lost, the dynamic linear model can be extended to the dynamic generalised linear model discussed in West et al. (1985). Non-normal distributions can be presented in the exponential family form of distributions, but the analytic solution may not be available. In that case, approximate inference for the moments of the posterior distribution can be done using the linear Bayesian method or using the simulation-based inference such as the sequential Monte Carlo methods.

2.3 Pairs Trading

Statistical arbitrage deploys statistical methods in order to construct trading strategies in the financial market, especially popular among hedge funds and investment banks. It takes advantage of relative mispricings between two or more financial instruments.

Pairs trading is a particular methodology of statistical arbitrage. According to Vidyamurthy (2004), the outset of pairs trading is marked by a group of quants at Morgan Stanley in the mid 1980s. It assumes that there is, for example, a pair of stocks showing a pattern for similar movements of their prices. The pattern may be valid for most of times, but temporarily be disrupted. It is known that a series of disruptions from and restorations to the pattern makes pairs trading profitable.

A pair trader opens her position to buy the underpriced and short-sell the overpriced asset. By holding a long and a short position at the same time, pairs trading is regarded as a market neutral strategy. A convergence trader bets her luck on the convergence by opening her position when a disruption occurs, and closing it with the convergence. On the other hand, there is a divergence trader betting her luck against the convergence. A divergence trader opens her position when the share prices of a pair converge, and closes it when they diverge. A day trader opens and closes her position day after day while a swing trader keeps her position from two days to several weeks.

Book-length references on statistical arbitrage, pairs trading, and algorithmic trading can be found in Pole (2007), Vidyamurthy (2004), Whistler (2004), Ehrman (2005), Chan (2008), and Chan (2013).

2.3.1 The Correlation Approach

Ehrman (2005) suggests an approach using a correlation from a pair of shares. The correlation coefficient in his analysis is interpreted as the strength of a relationship between the two. When it is greater than or equal to 0.7, the pair is identified as relevant to trading. To see if the relationship changes, the correlations are measured and monitored over some periods of time such as 30-, 90-, and 180-calendar-day for short-term traders and 90-, 180-, and 365-day for long-term traders.

2.3.2 The Distance Approach

Gatev et al. (2006) suggest the distance approach. It assumes that the standardised price differences of a pair follow a standardised normal distribution. Among hundreds of pairs from the market, the pairs, having the smallest sum of squared deviations, are selected for trading. An advantage of this approach is that computation is relatively cheap. However, the assumption is often breached because the share prices are known to have a log-normal distribution.

2.3.3 The Co-integration Approach

Co-integration is introduced by Engle and Granger (1987) and Engle and Yoo (1987), where price time series of two shares, for example, need not to be differenced for stationarity. Two non-stationary time series are said to be cointegrated if their linear combination is a stationary process.

Vidyamurthy (2004) illustrates the co-integration approach for pairs trading. Suppose that there are two non-stationary price time series of x_t and y_t. x_t and y_t are said to be cointegrated if there is a relationship of $z_t = y_t - a - b \cdot x_t$ so that z_t is stationary. The stationarity of z_t is evaluated by the Engle-Granger two-step method. This co-integrated pair is assumed to have an equilibrium. When there is a deviation from the long-term equilibrium, either one or both shares prices are believed to adjust themselves towards the equilibrium.

This approach has a couple of drawbacks in practice. For example, we need a full set of data to make a trading decision for pairs trading. The estimation of a and b and the stationarity test of z_t has to be done at each time t. This requires expensive computation costs and prevents the fast and efficient application of pairs trading in real time.

2.4 Mean Reversion and The Spread Model

There has been much interest on the long-term property of equity prices among researchers, though mostly in the fields of economics and finance. In particular, research is focused on whether time series of share prices follows either a random walk or a mean reverting process.

Kendall and Hill (1953) propose that share prices randomly move, and Fama (1965) and Malkiel (2004) make the random walk hypothesis as one of the most influential in finance under the assumption of the efficient market hypothesis developed by Fama (1970). If share prices evolve according to a random walk, any shock is permanent and a share price is not predictable using historical prices because the

volatility of the process would grow with no bound in the long run.

DeBondt and Thaler (1985) document the evidence of mean reversion in the market, which is supported by Fama and French (1988) as well as by Poterba and Summers (1988). If share prices revert to a mean over a period of time, then they can be forecasted using past observations. However, even when the share prices are supposed to follow the mean reversion process, questions arise such as whether the mean is constant, and if so, how long the constant mean would be valid.

Empirical studies by Gatev et al. (2006) show that there are trading opportunities for a pair of shares from the same industry. However, they do not assume if the pairs follow a mean-reversion process.

Elliott et al. (2005) assume a portfolio with two shares from the same industry, looking for an opportunity within pairs trading. They define the spread as the difference of prices from two shares at time t, and model the spread time series as a mean-reverting process. Assuming that the spread time series from a pair of shares follow a mean-reverting process, Elliott et al. (2005) propose the arithmetic Ornstein-Uhlenbeck model, proposed by Uhlenbeck and Ornstein (1930), in a Gaussian linear state space model, where the observed process is seen as a noisy realisation of the true spread. The expectation-maximisation (EM) algorithm is employed to estimate the model parameters.

While the arithmetic Ornstein-Uhlenbeck model is the most basic form of mean reversion model, the autoregressive model of order 1, AR(1), is a discrete time version of the arithmetic Ornstein-Uhlenbeck model. In an AR(1) model, shocks are transitory while any shock is permanent in a random walk.

Triantafyllopoulos and Montana (2011) propose time-dependency of parameters in the model with an on-line estimation procedure. The model by Triantafyllopoulos and Montana (2011) has couple of advantages over the model by Elliott et al. (2005). One advantage is that their proposed model is more flexible with time varying parameters. Another advantage is that it needs even less costs at computing for the

parameter estimation, using an adaptive and recursive algorithm. A third advantage is that the estimation procedure by Triantafyllopoulos and Montana (2011), unlike the expectation-maximisation algorithm employed by Elliott et al. (2005), produces the uncertainty measures without any additional computational cost.

2.4.1 The Spread Model

Elliott et al. (2005) propose a "mean-reverting Gaussian Markov chain model", or a Gaussian linear state-space model with time-invariant parameters A, B, C and D, for the observations in Gaussian noise and call it as the spread model. They assume that there are two shares showing similar price movements over time and the spread between the two have an equilibrium. Thus, the observations are the spread time series, where the spread at each time t is defined by the price difference between the two shares.

In Elliott et al. (2005), the observed spread series $\{Y_t\}$ is a noisy realisation of the state process $\{X_t\}$, a true but unobserved spread series. It is represented as follows.

$$\begin{aligned} Y_t &= X_t + D \cdot \omega_t, \quad \omega_t \sim N(0,1) \\ X_{t+1} &= A + B \cdot X_t + C \cdot \epsilon_{t+1}, \quad \epsilon_{t+1} \sim N(0,1) \end{aligned}$$

where ω_t and ϵ_t are mutually independent and assumed to be uncorrelated with X_t for $t = 1, 2, \ldots$, and $N(0,1)$ denotes the standard normal distribution. With B inside the unit circle, the spread series $\{Y_t\}$ is said to follow a mean-reverting process, as it can be easily verified in Elliott et al. (2005) that both $\mathrm{E}(Y_t)$ and $\mathrm{Var}(Y_t)$ converge to constant values.

Triantafyllopoulos and Montana (2011) consider the spread time series $\{Y_t\}$ following a state space model with time-varying parameters driven by an autoregressive model of order 1, which is specified by

$$Y_t = A_t + B_t \cdot Y_{t-1} + \nu_t, \qquad \nu_t \sim N(0, V_t) \tag{2.1}$$

$$A_t = \phi_1 \cdot A_{t-1} + \omega_{1,t}, \quad B_t = \phi_2 \cdot B_{t-1} + \omega_{2,t} \tag{2.2}$$

where ϕ_1 and ϕ_2 are the AR coefficients.

Triantafyllopoulos and Montana (2011) provide the conditions of this model to be a mean reverted process applying the Kalman filter for the recursive and sequential estimation of parameters in the model. A forgetting factor, also known as a discount factor, is adopted to control the local durability of the model. They illustrate a possible application of the model to pairs trading as a statistical arbitrage.

The model by Elliott et al. (2005) is extended by Triantafyllopoulos and Montana (2011) at least in three different ways. First of all, the parameters of the model in Triantafyllopoulos and Montana (2011) are time-varying, which makes a model adaptive to changes in the data stream. Secondly, by applying the Kalman filter to the dynamic linear model as shown in Harrison and Stevens (1971) and Harrison and Stevens (1976), the parameters can be estimated in real time when new observation is available. Thirdly, distributional assumptions on the normality allow the computation of posterior quantiles as well as the more traditional point estimates.

In the next section, a time-varying autoregressive model of order 1 is described in the state space model for the spread time series $\{Y_t\}$, and it is shown how the Kalman filter, proposed by Kalman (1960) and Kalman and Bucy (1961), is applied and understood in a Bayesian formulation. More details on the dynamic linear model and on-line estimation of parameters by the Kalman filter can be found in Harrison and Stevens (1971), Harrison and Stevens (1976), West and Harrison (1997) and Triantafyllopoulos and Montana (2011) as well. A book-length exposition on Bayesian forecasting with time series data can be found in West and Harrison (1997), Petris et al. (2009), Prado and West (2010), and Durbin and Koopman (2012).

2.4.2 The Spread Time Series in Dynamic Linear Model

For a univariate spread time series, the dynamic linear model is defined by

$$Y_t = \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim N(0, V_t) \tag{2.3}$$

$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_2(\mathbf{0}, W_t) \tag{2.4}$$

where $\mathbf{F}_t$ is the design vector, G_t is the evolution, system, transfer or state matrix, V_t is the observational variance, and W_t is the evolution covariance matrix. (2.3) is called as the observation equation, and (2.4) as the evolution, state or system equation. The observation error, ν_t, and the evolution or system error vector, $\boldsymbol{\omega}_t$, are assumed to be individually and mutually uncorrelated. The initial state $(\boldsymbol{\theta}_0 \mid D_0) \sim N(\mathbf{m}_0, C_0)$ is determined by a modeller, and the state vector $\boldsymbol{\theta}_t$ is assumed not to be correlated with ν_t and $\boldsymbol{\omega}_t$.

Equations (2.1) and (2.2) by Triantafyllopoulos and Montana (2011) can be represented as dynamic linear model by setting $\mathbf{F}_t' = (1, Y_{t-1})$, $\boldsymbol{\theta}_t = (A_t, B_t)'$, $G_t = \text{diag}(\phi_1, \phi_2)$, and $\boldsymbol{\omega}_t = (\omega_{1,t}, \omega_{2,t})'$. Therein, A_t and B_t are considered to evolve via AR models over time. Accordingly, ϕ_1 and ϕ_2 make the AR coefficients as $A_t = \phi_1 A_{t-1} + \omega_{1,t}$ and $B_t = \phi_2 B_{t-1} + \omega_{2,t}$, and G_t becomes G. These coefficients of ϕ_1 and ϕ_2 are assumed to lie inside the unit circle so that A_t and B_t be weakly stationary.

$\mathbf{F}_t$ and G_t are usually determined by the modeller. However, the observational variance V_t is often not known. When it is unknown over time, a coherent Bayesian learning is available for the observational variance V_t. For example, when V_t is unknown but constant as V, the precision τ is defined as $\frac{1}{V}$ and the posterior distribution of τ at time $t-1$ can be specified as a gamma distribution with parameters of $\frac{n_{t-1}}{2}$ and $\frac{d_{t-1}}{2}$ so that $(\tau \mid D_{t-1}) \sim Gamma\left(\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right)$. As new observation arrives in at t, the posterior distribution of τ is updated as $(\tau \mid D_t) \sim Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$ at t. The details of this updating are shown in Section 2.5.3.

W_t represents any stochastic change of the states over period of time, and measures the durability of the model. With no evolution error, say $W_t = 0$ at any time t, the system equation reduces to $\theta_t = G\theta_{t-1}$ and the model is considered to be globally true. On the other hand, when $W_t = \infty$, the model is totally useless. Considering the local durability of the model, an optimal value of W_t, if any, may vary over time.

2.4.3 Conditions for Mean Reversion

Given a set of data, we will be able to estimate $\boldsymbol{\theta}_t = (A_t, B_t)'$, and Triantafyllopoulos and Montana (2011) provide a theorem for the state space model of (2.3)-(2.4), giving the sufficient conditions of the series $\{Y_t\}$ to be mean-reverting.

Theorem 1. *If $\{Y_t\}$ is generated from the model of (2.3)-(2.4), then, conditionally on a realised sequence $B_1, \ldots, B_t$, $\{Y_t\}$ is mean revering if one of the two conditions apply.*

(a) *$\phi_1 = \phi_2 = 1, W_t = 0$ and $|B_1| < 1$;*

(b) *ϕ_1 and ϕ_2 lie inside the unit circle, W_t is bounded and $|B_t| < 1$, for all $t \geq t_0$ and for some integer $t_0 > 0$.*

By setting $\phi_1 = \phi_2 = 1$ and the covariance matrix $W_t = 0$ for all t, the first condition of **(a)** implies $A_t = A_{t-1} = \cdots = A_1 = A$ and $B_t = B_{t-1} = \cdots = B_1 = B$, resulting in a static AR model. In this case, $|B_1| = |B| < 1$ gives the known condition for mean-reversion in static AR models. Considering that ϕ_1 and ϕ_2 can be set to lie inside the unit circle initially by the modeller and W_t to be bounded, the second condition of **(b)** implies that $|B_t| < 1$.

Corollary 1. *If $\{Y_t\}$ is generated from the model of (2.3)-(2.4) with $\phi_1 = 1$, $|\phi_2| < 1$, $V_{11,t} = V_{22,t} = 0$, then $\{Y_t\}$ is mean revering if $|B_1| < 1$ for all $t \geq t_0$ for some $t_0 > 0$ where $V_t = (V_{ij,t})$ for $i, j = 1, 2$.*

This corollary allows that $A_t = A$ for all t as in Elliott et al. (2005), but B_t evolves as a weakly stationary AR model.

Given a set of data, $Y_1, \ldots, Y_t$, we can estimate B_t as $\hat{B}_t$ in the dynamic linear model discussed in this section so that the spread Y_t is detected as mean-reversion at time t when $|\hat{B}_t| < 1$. Since $\hat{B}_t$ is uncertain, the true value of B_t might well be greater than or equal to 1 at time t even if $|\hat{B}_t| < 1$ due to the associated uncertainty of $\hat{B}_t$. To deal with this issue, Triantafyllopoulos and Montana (2011) suggest to check the 95% credible bounds of B_t and see whether they lie inside the unit circle. However,

from the results of the above reference as well as from our own experimentation with several data sets, this approach results in conservative detection of mean-reversion: if B_t is much less than one, then the algorithm detects mean-reversion well, but if B_t is close to one, the algorithm results in credible bounds that are outside the unit circle.

2.5 Recursions in Dynamic Linear Model

For sequential updating of the recursive estimation, we employ the Kalman filter. To make the model be invariant to the measurement scale of observations, all the variances in the model are multiplied by V, and the variances are independent of the measurement scales. In this thesis, the state space model for the univariate spread time series $\{Y_t\}$ is specified with V as a scaling factor.

The derivation details for recursive estimation in dynamic linear model can be found in Appendix B.

2.5.1 Conditional on V

When V_t is known as V, the specification of a model is obtained as

$$
\begin{aligned}
Y_t &= \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim N(0, V) && (2.5)\\
\boldsymbol{\theta}_t &= G\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_2(\mathbf{0}, VW_t^*) && (2.6)\\
&\quad (\boldsymbol{\theta}_0 \mid V, D_0) \sim N_2(\mathbf{m}_0, VC_0^*) && (2.7)
\end{aligned}
$$

where N_2 denotes the bivariate normal distribution, and the starred variance matrices of C_0^* and W_t^* represent the scale-free variance-covariance matrices. The quantities of $\mathbf{m}_0$ and C_0^* are specified by the modeller initially.

The recursive estimation procedure with updating equations are given by

(a1) Posterior at $t-1$
$(\boldsymbol{\theta}_{t-1} \mid V, D_{t-1}) \sim N_2(\mathbf{m}_{t-1}, VC_{t-1}^*)$

(a2) Prior at t

$(\boldsymbol{\theta}_t \mid V, D_{t-1}) \sim N_2(\mathbf{a}_t, VR_t^*)$

where $\mathbf{a}_t = G\mathbf{m}_{t-1}$ and $R_t^* = GC_{t-1}^*G' + W_t^*$

(a3) One-step forecast

$(Y_t \mid V, D_{t-1}) \sim N_2(f_t, VQ_t^*)$

where $f_t = F_t'G\mathbf{m}_{t-1}$ and $Q_t^* = F_t'R_t^*F_t + 1$

(a4) Posterior at t

$(\boldsymbol{\theta}_t \mid V, D_t) \sim N_2(\mathbf{m}_t, VC_t^*)$

where $\mathbf{m}_t = \mathbf{a}_t + K_te_t$ and $C_t^* = R_t^* - K_tQ_t^*K_t'$

with $K_t = R_t^*F_t/Q_t^*$ and $e_t = Y_t - f_t$

2.5.2 Unconditional on V

When V_t is unknown but constant as V, the specification of a model is obtained as

$$
\begin{aligned}
Y_t &= \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim N(0, V) && (2.8)\\
\boldsymbol{\theta}_t &= G\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_2(\mathbf{0}, VW_t^*) && (2.9)\\
&\quad (\boldsymbol{\theta}_0 \mid D_0) \sim N_2(\mathbf{m}_0, VC_0^*) && (2.10)\\
&\quad (\tau \mid D_0) \sim Gamma(\frac{n_0}{2}, \frac{d_0}{2}) && (2.11)
\end{aligned}
$$

where N_2 denotes the bivariate normal distribution, and the starred variance matrices of C_0^* and W_t^* represent the scale-free variance-covariance matrices. $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$ and U_0 is a prior point estimate of V. The unknown V or $\tau = \frac{1}{V}$ is sequentially updated as new observation is obtained at each time t. The posterior mean of τ is $\mathrm{E}(\tau \mid D_t) = \frac{n_t}{d_t} = \frac{1}{U_t}$ where U_t is a posterior point estimate of V at t. The quantities of $\mathbf{m}_0$, C_0^*, n_0, and d_0 are specified by the modeller initially.

The recursive estimation procedure with updating equations are given by

(b1) Posterior at $t-1$

$(\boldsymbol{\theta}_{t-1} \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{m}_{t-1}, C_{t-1})$

where $C_t = U_{t-1}C_{t-1}^*$

(b2) Prior at t

$(\boldsymbol{\theta}_t \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{a}_t, R_t)$

where $\mathbf{a}_t = G\mathbf{m}_{t-1}$ and $R_t = U_{t-1}R_t^*$

(b3) One-step forecast

$(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}(f_t, Q_t)$

where $f_t = F_t'G\mathbf{m}_{t-1}$ and $Q_t = U_{t-1}Q_t^*$

(b4) Posterior at t

$(\boldsymbol{\theta}_t \mid D_t) \sim T_{n_t}(\mathbf{m}_t, C_t)$

where $\mathbf{m}_t = \mathbf{a}_t + K_t e_t$ and $C_t = U_t C_t^*$

with $K_t = R_t^* F_t / Q_t^*$ and $e_t = Y_t - f_t$

where $T_{n_{t-1}}$ and T_{n_t} denotes the student t distribution with n_{t-1} and n_t degrees of freedom respectively.

2.5.3 Recursion of $\tau = \frac{1}{V}$

The prior distribution of τ at t is also $(\tau \mid D_{t-1}) \sim Gamma\left(\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right)$ where the posterior distribution at $t-1$ is the prior distribution at t. After an observation is made at t, the posterior distribution of τ at t becomes $(\tau \mid D_t) \sim Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$ where $n_t = n_{t-1} + 1$ and $d_t = d_{t-1} + \frac{e_t^2}{Q_t^*}$.

The initial information on the precision is specified as $(\tau \mid D_0) \sim Gamma(\frac{n_0}{2}, \frac{d_0}{2})$ by the modeller where both n_0 and d_0 are positive integers. The updating equations are summarised as

(c1) $(\tau \mid D_{t-1}) \sim Gamma\left(\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right)$

(c2) $(\tau \mid D_t) \sim Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$

where $n_t = n_{t-1} + 1$ and $d_t = d_{t-1} + e_t^2/Q_t^*$.

The posterior mean of $\mathrm{E}(\tau \mid D_t) = \frac{n_t}{d_t} = \frac{1}{U_t}$, and U_t is a posterior point estimate of V at t.

2.6 Variable Forgetting Factor and Least Squares

Since the notion of adaptive control emerged in 1950s by Drenick and Shahbender (1957), it took couple of decades to see several adaptive algorithms in practice. In adaptive control, parameters of a model are estimated by adaptive algorithms. Adaptive algorithms are classified as either the stochastic gradient algorithm or the recursive least squares algorithm according to the method of least squares applied in the algorithm. The stochastic gradient algorithm is also known as the least-mean-square algorithm.

An adaptive algorithm is known as a self-tuning regulator in the field of control theory. Åström and Wittenmark (1973) design the original self-tuning regulator for a process with parameters which are constant but unknown. Adaptive algorithms, or self-tuning regulators, help time-varying parameters of a model to recursively adjust to the input. However, if the constant forgetting factor is not carefully chosen, the covariance matrix may grow exponentially, and a system becomes extremely sensitive to the changes in the process.

Variable forgetting factor first appears in Åström et al. (1977), and its recursion in Fortescue et al. (1981). According to Åström et al. (1977), when a process contains time-varying and nonlinear dynamics, a forgetting factor makes the recursive estimator to adjust to the changes, preventing it from the converging. In Fortescue et al. (1981), the forgetting factor is determined by the recursion at each time t, and the parameter estimates are obtained by a recursive least squares with variable weighting on the past data.

Haykin (1996) suggests that the tracking performance of a model is enhanced with an adaptive scheme of a forgetting factor, and introduces the cost function J_t to be minimised as

$$J_t = \frac{1}{2}E(|e_t|^2) \tag{2.12}$$

where $e_t = Y_t - f_t$, defined as the difference between the observation and the estimation or prediction of Y_t. e_t indicates the estimation or prediction error. In this case, the aim of the adaptive scheme is to find the particular value of λ minimising the

cost function, J_t. The forgetting factor λ gives exponentially less weight to the older errors. The tracking performance of a time-varying system is usually measured by the mean-square error, defined as $\mathrm{E}(|e_t|^2)$.

Chun et al. (1998) develop a generalized recursive least squares algorithm, presenting possible applications of variable forgetting factor recursive least squares to the general state space model. Song et al. (2000) suggest that the speed of tracking is achieved by incorporating the second derivatives of the cost function J_t in (2.12). They propose the Gauss-Newton method from Ljung and Soderstrom (1983) for variable forgetting factor.

2.7 Dynamic Generalised Linear Model

Since the mid-1980s, there is a growing literature on fully Bayesian analysis with models for categorical data. West et al. (1985) show early work with non-normal univariate time series, proposing conjugate analysis and the linear Bayes method as an approximate inference for the dynamic generalised linear model (DGLM). Fahrmeir and Kaufmann (1991) and Fahrmeir (1992) use direct analytic approximations for an analysis with multinomial time series. Grunwald et al. (1993) model multivariate series of proportions on conditionally Dirichlet distributed vectors of multinomial probabilities, developing time evolution for the probabilities as well. Cargnoni et al. (1997) propose the class of conditionally Gaussian dynamic models for non-normal multivariate time series. In that paper, logistically transformed probabilities are formulated in dynamic linear model, and a posterior is simulated based on appropriately modified Markov chain Monte Carlo algorithms including Metropolis-Hastings components. All of these can be seen as an extension of the work by West et al. (1985).

For time series analysis, a number of models such as autoregressive moving average models, structural time series, and dynamic regression models can be described and dealt with in a flexible and unifying way of the state space form. An exponential family state space model or a DGLM has been developed by West et al. (1985). As in the class of dynamic linear model, a DGLM consists of an observation model,

which distribution belongs to the exponential family, and an evolution model for the states or the unknown parameters.

The generalised state space model implies the non-linear non-Gaussian state space model. The simplest example of the generalised state space model would be the dynamic linear model assuming both the Gaussianity and the linearity. The dynamic linear model is extended and generalized to the DGLM with no distributional assumption of Gaussianity. The exponential family class of distributions forms the large class of the DGLM which consists of the observation model and the evolution model as in the dynamic linear model. However, the posterior distribution of the states is not analytically available, hence inference is based on approximations, such as in West et al. (1985), or in simulation-based procedures, such as Markov chain Monte Carlo by Gamerman (1998) and particle filters.

Triantafyllopoulos (2009) provides a discussion on online estimation of DGLM for several response distributions. A book-length treatment of DGLM for multivariate and multicategorical responses can be found in Fahrmeir and Tutz (2010).

A DGLM for a univariate time series Y_t is defined in West and Harrison (1997) by

$$\begin{aligned} p(Y_t \mid \eta_t) &= \exp\left\{[Y_t'\eta_t - a(\eta_t)] + b(Y_t)\right\} && (2.13)\\ g(\eta_t) &= \mu_t = \mathbf{F_t'}\boldsymbol{\theta}_t && (2.14)\\ \boldsymbol{\theta}_t &= G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \quad \text{with} \quad \boldsymbol{\epsilon}_t \sim [\mathbf{0}, W_t] && (2.15) \end{aligned}$$

In (2.15), $[\mathbf{0}, W_t]$ represents a distribution with mean vector of $\mathbf{0}$ and variance matrix of W_t and indicates that there is no specific form of distributional assumption for $\boldsymbol{\epsilon}_t$. (2.13), (2.14), and (2.15) are an observation model, a link function and an evolution equation respectively, and the details of them are as follows.

- $p(\cdot)$ is the joint probability function;
- η_t is the natural parameter;
- $g(\cdot)$ is a link function of a known, continuous and monotonic function mapping η_t to the real line;

- $\mu_t = \mathbf{F}_t'\boldsymbol{\theta}_t$ is a linear function of the state vector parameters;
- $\mathbf{F}_t$ is a known n-dimensional regression vector;
- $\boldsymbol{\theta}_t$ is an n-dimensional state vector at time t;
- G_t is a known $n \times n$ evolution matrix;
- $\boldsymbol{\epsilon}_t$ is an n-vector of evolution errors having zero mean and known variance matrix W_t;

The model definition is completed via

$$(\boldsymbol{\theta}_0|D_0) \sim [\mathbf{m}_0, C_0] \tag{2.16}$$

where the quantities of $\mathbf{m}_0$ and C_0 are specified initially by the modeller.

2.7.1 The Linear Bayesian Method

Hartigan (1969) develops the linear Bayesian methods where the inference is linearly approximated using only the first two moments of the distribution of the prior and likelihood. A similar approach for nonlinear regression problems is found in Goldstein (1976).

Given no full distributional assumption for the evolution error $\boldsymbol{\epsilon}_t$, and the state vector $\boldsymbol{\theta}_t$, their first- and second- order moments can be approximated using Bayes linear methods as detailed in West et al. (1985). The recursive updating proceeds as follows.

Assuming the posterior for the state vector at $t-1$ as

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim [\mathbf{m}_{t-1}, C_{t-1}]$$

the joint prior distribution of μ_t and $\boldsymbol{\theta}_t$ at time t is partially specified by the first

two moments only, and it follows that

$$\begin{pmatrix} \mu_t \\ \boldsymbol{\theta}_t \end{pmatrix} \Bigg| \; D_{t-1} \Bigg) \sim \left[\begin{pmatrix} f_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} Q_t & \mathbf{F}_t' R_t \\ R_t \mathbf{F}_t & R_t \end{pmatrix} \right] \tag{2.17}$$

where $f_t = \mathbf{F}_t' \mathbf{a}_t$, $Q_t = \mathbf{F}_t' R_t \mathbf{F}_t$, $\mathbf{a}_t = G_t \mathbf{m}_{t-1}$, and $R_t = G_t C_{t-1} G_t' + W_t$ are approximately the mean and variance of $(\boldsymbol{\theta}_t \mid D_{t-1})$.

The posterior mean and variance of η_t are approximated by the linear Bayesian method and then by using the tower property of expectations as

$$\mathrm{E}(\mu_t \mid D_t) \;=\; \mathrm{E}(g(\eta_t) \mid D_t) = f_t^* \quad \text{and} \quad \mathrm{Var}(\mu_t \mid D_t) \;=\; \mathrm{Var}(g(\eta_t) \mid D_t) = Q_t^* \tag{2.18}$$

where f_t^* and Q_t^* represents the posterior mean and variance of η_t respectively. The expressions for f_t^* and Q_t^* would differ for each distribution.

The mean and variance of $(\boldsymbol{\theta}_t \mid D_t)$ are approximated as

$$(\boldsymbol{\theta}_t \mid D_t) \sim [\mathbf{m}_t, C_t] \tag{2.19}$$

where $\mathbf{m}_t = \mathbf{a}_t + R_t \mathbf{F}_t (f_t^* - f_t)/Q_t$ and $C_t = R_t - R_t \mathbf{F}_t \mathbf{F}_t' R_t (1 - Q_t^*/Q_t)/Q_t$. The details of the approximation of (2.19) are given in the below.

The posterior of $(\boldsymbol{\theta}_t \mid D_t)$ can be derived from the joint posterior for μ_t and $\boldsymbol{\theta}_t$ which is obtained by Bayes' theorem as follows.

$$\begin{aligned} p(\mu_t, \boldsymbol{\theta}_t \mid D_t) &\propto p(\mu_t, \boldsymbol{\theta}_t \mid D_{t-1}) p(Y_t \mid \mu_t) \\ &\propto \{p(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) p(\mu_t \mid D_{t-1})\} p(Y_t \mid \mu_t) \\ &\propto p(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) \{p(\mu_t \mid D_{t-1}) p(Y_t \mid \mu_t)\} \\ &\propto p(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) p(\mu_t \mid D_t) \end{aligned} \tag{2.20}$$

From (2.20), we can see that $\boldsymbol{\theta}_t$ is conditionally independent of Y_t given μ_t and D_{t-1},

and it follows that

$$p(\boldsymbol{\theta}_t \mid D_t) = \int p(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) p(\mu_t \mid D_t) d\mu_t \tag{2.21}$$

$p(\mu_t \mid D_t)$, the second component of the integrand in (2.21), can be obtained directly from (2.18) in the conjugate form posterior for η_t. However, due to the incomplete of the joint prior distribution in (2.17), conditional moments of $p(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1})$ are unknown and non-linear functions of μ_t. Given only the partial moments, the posterior mean and variance matrix of $\boldsymbol{\theta}_t$ can be estimated using the linear Bayesian method by Goldstein and Wooff (2007).

Conditional moments of $\mathrm{E}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1})$ and $\mathrm{Var}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1})$ are obtained as the optimal estimate by the linear Bayesian method. For all μ_t, they are

$$\begin{aligned} \hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) &= \mathbf{a}_t + R_t \mathbf{F}_t (\mu_t - f_t)/Q_t & (2.22) \\ \hat{\mathrm{Var}}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) &= R_t - R_t \mathbf{F}_t \mathbf{F}_t R_t / Q_t & (2.23) \end{aligned}$$

From (2.21), $\mathrm{E}(\boldsymbol{\theta}_t \mid D_t) = \mathrm{E}\{\mathrm{E}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) \mid D_t\}$ and $\mathrm{Var}(\boldsymbol{\theta}_t \mid D_t) = \mathrm{Var}\{\mathrm{E}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) \mid D_t\} + \mathrm{E}\{\mathrm{Var}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) \mid D_t\}$. Thus, the posterior moments of $\boldsymbol{\theta}_t$ may be estimated based on the optimal estimates of (2.22) and (2.23).

$$\begin{aligned} \mathbf{m}_t &= \mathrm{E}\{\hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1} \mid D_t)\} \\ &= \mathrm{E}\{\mathbf{a}_t + R_t \mathbf{F}_t (\mu_t - f_t)/Q_t \mid D_t\} \\ &= \mathbf{a}_t + R_t \mathbf{F}_t \{\mathrm{E}(\mu_t \mid D_t) - f_t\}/Q_t \\ &= \mathbf{a}_t + R_t \mathbf{F}_t (f_t^* - f_t)/Q_t & (2.24) \\ C_t &= \mathrm{Var}\{\hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) \mid D_t\} + \mathrm{E}\{\hat{\mathrm{Var}}(\boldsymbol{\theta}_t \mid \mu_t, D_{t-1}) \mid D_t\} \\ &= \mathrm{Var}\{\mathbf{a}_t + R_t \mathbf{F}_t (\mu_t - f_t)/Q_t \mid D_t\} + \mathrm{E}(R_t - R_t \mathbf{F}_t \mathbf{F}_t R_t / Q_t \mid D_t) \\ &= R_t \mathbf{F}_t \mathbf{F}_t R_t \mathrm{Var}(\mu_t \mid D_t)/Q_t^2 + R_t - R_t \mathbf{F}_t \mathbf{F}_t R_t / Q_t \\ &= R_t - R_t \mathbf{F}_t \mathbf{F}_t R_t (1 - Q_t^*/Q_t)/Q_t & (2.25) \end{aligned}$$

2.7.2 Particle Filters

Sequential Monte Carlo methods, also referred to as particle filters, are regarded as sequential application of importance sampling. When importance sampling as one of Monte Carlo integration methods is sequentially applied, it is called as the sequential importance sampling and resampling. Unfortunately, sequential importance sampling may fail as t increases. Thus, it is followed by the selection step of the generated particles, known as the resampling step.

Since Gordon et al. (1993) first demonstrate the utility of sequential Monte Carlo approaches to nonlinear/non-Gaussian Bayesian state estimation, several related algorithms are developed and actively used in the names of the bootstrap filter, the condensation, the auxiliary particle filter, the Monte Carlo filter, the interacting particle approximations and the survival of the fittest according to Cappé et al. (2007) and Doucet and Johansen (2009).

There is a large literature on particle filters. Among them, Cappé et al. (2007), Doucet and Johansen (2009), and Doucet et al. (2010a) review and summarise the sequential Monte Carlo methods while Liu (2001) reviews the general Monte Carlo methods. Douc et al. (2005) compare several resampling schemes in the particle filters. Petris et al. (2009) suggest the basic approach of the particle filters with codings in R.

2.8 Conclusion

The variable forgetting factor (VFF) would improve the flexibility of the model, making it more adaptive to the changes of the observations in time series, and enhance the performance in prediction by the model when new data point is observed. Thus, if the variability of forgetting factor is introduced to the spread model by Triantafyllopoulos and Montana (2011), more trading opportunities can be expected in the pairs trading. The VFF algorithms using the steepest descent and the Gauss-Newton methods are applied to the class of DLM from the field of signal processing and control engineering. However, the steepest descent and the Gauss-

Newton methods are originally from the optimisation theory, looking for the optimal value over a period of time. Now that both of them are applied to find the value of VFF sequentially at each time t, the application may not be a reasonable choice in some cases, and they are complicated to follow. Considering the role of VFF is to help prediction of the model, the VFF algorithm needs to be rather simple but sequential using the error analysis and the conjugacy of distributions.

According to the conditions of mean-reversion by Triantafyllopoulos and Montana (2011), the spread of a pair of financial instruments is said to be in mean-reversion when $|\hat{B}_t| < 1$. The detection of mean-reversion of the spread at time t solely relies on the value of $\hat{B}_t$ at that time, and whatever happens before and after the time t, the algorithm detects mean-reversion as long as $|\hat{B}_t| < 1$. A trader or an investor may think that this is too dangerous to take the risks of algorithmic pairs trading. In particular, when the spread shows volatile movements, there would be no way to avoid the extreme, and algorithmic pairs trading may end up with huge loss. Thus, we find the need to monitor the behaviour of $|\hat{B}_t|$, slicing the range which $|\hat{B}_t|$ can be located into categories. Now that the behavior of $\hat{B}_t$ is regarded as a process over the period of time, it is closely monitored online at each time t. As a process, the dynamic variation in $|\hat{B}_t|$ such as the trend, seasonality, and cycle, if any, can be monitored and analyzed more in depth with the evolution model of DGLM. For this, in Chapter 4, the DGLM for multi-categorical time series and its recursions are derived where the particle filter as inference are proposed. This is novel extending the works by West et al. (1985) and Triantafyllopoulos (2009) to multi-categorical time series in the DGLM using multinomial distribution for the observation model. The particle filter, developed in the chapter for categorical time series data, is not necessarily restricted to the application of mean reversion considered in the thesis. Categorical time series appear frequently and the contribution of the proposed particle filter is general.

Chapter 3

Dynamic Linear Model with Variable Forgetting

3.1 Introduction

The tracking performance of the dynamic linear model, defined as (2.3) and (2.4) in Section 2.4.2, is known to be better with small forgetting factor when there is a sudden variation or change in the data. On the other hand, it is better to have forgetting factor close to unity when the data stream is stable, giving the longer memory. When a forgetting factor is small, the memory is low, using few past data to predict future values of the data. Hence, the model can adapt quickly to changes, as changes are supposed to be local and influenced from immediate past observations, and results in non-smooth or noisy predictions; when the memory is high, we use a large number of past data to predict future values of the data, and then predictions are smooth. Smooth predictions mean low variance while noisy predictions mean high variance. We want a system that is adaptive when there is a change and smooth with low variance when there is not a change. This observation naturally leads us to variable forgetting.

From the updating equations **(a1)**-**(a4)** and **(b1)**-**(b4)** in Section 2.5, the covariance matrix C^*_{t-1} of the posterior for the state $(\theta_{t-1} \mid V, D_{t-1})$ at time $t-1$ proceeds to $R^*_t = GC^*_{t-1}G' + W^*_t$ which is the covariance matrix of the prior of θ_t at time

t. When the model is globally true, $W_t^* = 0$ and $R_t^* = GC_{t-1}^*G'$. However, the model is not supposed to be globally true, but locally true. Thus, if the actual precision of the posterior for the state $(\theta_t \mid V, D_{t-1})$ is denoted by $(R_t^*)^{-1}$, then proportionally it is reduced to $\lambda(GC_{t-1}^*G')^{-1}$, or $R_t^* = \frac{1}{\lambda}GC_{t-1}^*G'$. With the relation of $R_t^* = GC_{t-1}^*G' + W_t^*$, this implies a specification of W_t^* as $W_t^* = \frac{1-\lambda}{\lambda}(GC_{t-1}^*G')$ where λ is a forgetting factor, taking a value between 0 and 1, say $0 < \lambda \leq 1$.

It is an important question how appropriate values for the variances V_t and W_t are chosen in the DLM. For the observational variance V_t, variance learning procedure is applied using the precision and the Gamma distribution when it is not known but constant. For the evolution variance W_t, it is very difficult to directly quantify the elements, which are often grossly misspecified. If known, it would hold only temporarily, or be hypothetical since no "true" evolution process of the states can be exactly represented by mathematical model. Considering that W_t controls the uncertainty of the states between $t-1$ and t and determines the stability in the evolution equation of the DLM over time, there will be no optimal value of W_t suitable for all times. The forgetting factor is an aid to choose the evolution variance W_t.

The forgetting factor λ controls the local durability, ensuring that the data in the distant past are forgotten, and measures the memory of an algorithm as $(1-\lambda)^{-1}$. For example, when $\lambda = 1$, the evolution error W_t in the dynamic model equals to 0 and the whole system becomes globally true with the memory of ∞. If $\lambda \to 0$, $W_t \to \infty$ and the system can be said to be totally unreliable or useless.

First of all, this chapter aims to introduce the existing algorithms using the steepest descent and the Gauss-Newton methods for the variable forgetting factor (VFF) from the field of signal processing and control engineering, deriving the recursions in the class of dynamic linear model (DLM). The role of the VFF is to improve the adaptability to the changes of time series and enhance the predictability of the model. Thus, secondly, based on the sequential analysis of the prediction error at each time t, new algorithm for the VFF is proposed.

Widely applied methods of the steepest descent method and the Gauss-Newton

method for variable forgetting factor (VFF) algorithms are taken into account of their use to dynamic linear model (DLM). It may be difficult to say that the application of VFF algorithm to DLM is new. However, in DLM, forgetting factor has been a constant with no consideration on its variability yet. Thus, no derivation of the recursions has been proposed and shown with VFF in the class of DLM. In that sense, new names are given to VFF algorithms in DLM of the steepest descent method and the Gauss-Newton method each as the steepest descent VFF (SDvFF) algorithm and the Gauss-Newton VFF (GNvFF) algorithm respectively.

An algorithm for VFF needs to catch up the variation when there is a sudden change in the data, and adjust itself to the system when there is a smooth data stream. To achieve these two goals in one algorithm, new algorithm of the beta-Bernoulli VFF (BBvFF) is devised and proposed as an alternative to the existing VFF algorithms. The rationale of the BBvFF is as follows. When the prediction error e_t, defined by $e_t = Y_t - f_t$, is small, high value of forgetting factor close to unity should be chosen to retain as much information as possible from the new observation. On the other hand, when e_t is large, low value of forgetting factor should be chosen to increase the sensitivity of parameter estimates.

Derivation of the recursions for the algorithms of the SDvFF and the GNvFF is shown, and the tracking performances of three algorithms, the SDvFF, the GNvFF, and the BBvFF, are compared and evaluated by mean squared error. Comparisons are made with generated time series.

3.2 Dynamic Linear Model with Variable Forgetting Factor

Assuming the unknown but constant observational variance of $V_t = V$, a univariate DLM is specified as (2.8)-(2.11), which are

$$
\begin{aligned}
Y_t &= \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim N(0, V) && (3.1)\\
\boldsymbol{\theta}_t &= G\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim T_{n_{t-1}}(\mathbf{0}, VW_t^*) && (3.2)\\
&(\boldsymbol{\theta}_0 \mid D_0) \sim T_0(\mathbf{m}_0, VC_0^*) && (3.3)\\
&(\tau \mid D_0) \sim Gamma(\frac{n_0}{2}, \frac{d_0}{2}) && (3.4)
\end{aligned}
$$

where $\mathbf{F}_t' = (1, Y_{t-1})'$, $\boldsymbol{\theta}_t = (A_t, B_t)$, $G = \text{diag}(\phi_1, \phi_2)$, and $\boldsymbol{\omega}_t = (\omega_{1,t}, \omega_{2,t})$. $T_{n_{t-1}}$ denotes the t-distribution with degrees of freedom n_{t-1}, and the starred variance matrices of C_0^* and W_t^* represent the scale-free variance-covariance matrices. $\text{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$ and U_0 is a prior point estimate of V. The quantities of $\mathbf{m}_0$, C_0^*, n_0, and d_0 are also specified by the modeller initially.

3.2.1 Relationship Between K_t and C_t^*

Before looking into the derivation of the recursions in the VFF algorithms, it is worth looking into a relation between K_t and C_t^*.

First of all, from the updating equation **(a4)** in Section 2.5.1, C_t^* and K_t can be represented as

$$
\begin{aligned}
C_t^* &= R_t^* - K_t Q_t^* K_t' \\
&= R_t^* - K_t \mathbf{F}_t'(R_t^*)' \quad = \quad (I - K_t\mathbf{F}_t')R_t^* \qquad (3.5)\\
\text{or} \quad &= R_t^* - R_t^*\mathbf{F}_t K_t' \quad = \quad R_t^*(I - \mathbf{F}_t K_t')
\end{aligned}
$$

where $K_t = \frac{R_t^*\mathbf{F}_t}{Q_t^*}$ and $R_t^* = (R_t^*)'$ because of the assumption that the observation error ν_t and the evolution error ω_t are individually and mutually uncorrelated. Actually, considering that $R_t = U_{t-1}R_t^*$ and $Q_t = U_{t-1}Q_t^*$, $K_t = \frac{R_t\mathbf{F}_t}{Q_t}$ and $R_t = (R_t)'$.

As seen from the previous chapter, $Q_t^* = 1 + \mathbf{F}_t' R_t^* \mathbf{F}_t$ in the univariate case, and $K_t = \frac{R_t^* \mathbf{F}_t}{Q_t^*} = \frac{R_t^* \mathbf{F}_t}{1+\mathbf{F}_t' R_t^* \mathbf{F}_t}$, from which $K_t + K_t \mathbf{F}_t' R_t^* \mathbf{F}_t = R_t^* \mathbf{F}_t$. Thus, K_t can be rewritten as

$$\begin{aligned} K_t &= R_t^* \mathbf{F}_t - K_t \mathbf{F}_t' R_t^* \mathbf{F}_t \\ &= (I - K_t \mathbf{F}_t') R_t^* \mathbf{F}_t \\ &= C_t^* \mathbf{F}_t \end{aligned} \tag{3.6}$$

where $(I - K_t \mathbf{F}_t') R_t^* = C_t^*$ from (3.5).

In general, $Q_t = U_{t-1} + \mathbf{F}_t' R_t \mathbf{F}_t$ and $K_t = \frac{R_t \mathbf{F}_t}{Q_t} = \frac{R_t \mathbf{F}_t}{U_{t-1}+\mathbf{F}_t' R_t \mathbf{F}_t}$, from which $K_t U_{t-1} + K_t \mathbf{F}_t' R_t \mathbf{F}_t = R_t \mathbf{F}_t$. Thus, K_t still can be rewritten as

$$\begin{aligned} K_t &= \frac{R_t \mathbf{F}_t - K_t \mathbf{F}_t' R_t \mathbf{F}_t}{U_{t-1}} \\ &= R_t^* \mathbf{F}_t - K_t \mathbf{F}_t' R_t^* \mathbf{F}_t \\ &= (I - K_t \mathbf{F}_t') R_t^* \mathbf{F}_t \\ &= C_t^* \mathbf{F}_t \end{aligned} \tag{3.7}$$

where $R_t^* = \frac{R_t}{U_{t-1}}$.

Additionally, because of the assumption that the observation error ν_t and the evolution error ω_t are individually and mutually uncorrelated again, $(C_t^*)' = C_t^*$, and therefore, $K_t' = \mathbf{F}_t' C_t^*$.

3.2.2 Recursions of Parameter Estimates

Now that forgetting factor λ is variable, it is decided at each time t as λ_t. Following the result of (3.6) and (b1)-(b4) in Section 2.5, the recursive estimation procedure with updating equations can be achieved by

(d1) Posterior at $t-1$
$(\theta_{t-1} \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{m}_{t-1}, C_{t-1})$
where $C_t = U_{t-1} C_{t-1}^*$ and $C_{t-1}^* = \frac{G C_{t-2} G'}{\lambda_{t-2} + F_{t-1}' G C_{t-2} G' F_{t-1}}$

(d2) Prior at t

$(\theta_t \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{a}_t, R_t)$

where $\mathbf{a}_t = G\mathbf{m}_{t-1}$, $R_t = U_{t-1}R_t^*$ and $R_t^* = \frac{1}{\lambda_{t-1}} GC_{t-1}^* G'$

(d3) One-step forecast

$(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}(f_t, Q_t)$

where $f_t = \mathbf{F}_t' G\mathbf{m}_{t-1}$, $Q_t = U_{t-1}Q_t^*$ and $Q_t^* = \frac{1}{\lambda_{t-1}}(\mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t) + 1$

(d4) Posterior at t

$(\theta_t \mid D_t) \sim T_{n_t}(\mathbf{m}_t, C_t)$

where $\mathbf{m}_t = \mathbf{a}_t + K_t e_t$, $C_t = U_t C_t^*$ and $C_t^* = \frac{GC_{t-1}^* G'}{\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t}$

with $K_t = R_t^*\mathbf{F}_t / Q_t^*$ and $e_t = Y_t - f_t$

3.2.3 Recursion of $\tau = \frac{1}{V}$

The unknown V or $\tau = \frac{1}{V}$ is sequentially updated as new observation is obtained at each time t. The posterior mean of τ is $\mathrm{E}(\tau \mid D_t) = \frac{n_t}{d_t} = \frac{1}{U_t}$ where U_t is a posterior point estimate of V at t. As for (c1) and (c2) in Section 2.5, the updating equations are summarised as

(e1) $(\tau \mid D_{t-1}) \sim Gamma\left(\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right)$

(e2) $(\tau \mid D_t) \sim Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$

where $n_t = n_{t-1} + 1$, $d_t = d_{t-1} + e_t^2/Q_t^*$ and $Q_t^* = \frac{1}{\lambda_{t-1}}(\mathbf{F}_t' GC_{t-1} G'\mathbf{F}_t) + 1$ as in **(d3)**.

The posterior mean of $\mathrm{E}(\tau \mid D_t) = \frac{n_t}{d_t} = \frac{1}{U_t}$, and U_t is a posterior point estimate of V at t. As $t \to \infty$, the posterior of τ eventually converges about the mode.

When $(\tau \mid D_t) \sim Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$, the density function of a Gamma distribution is

$$\begin{aligned} p(\tau \mid D_t) &= \frac{d_t^{n_t}}{\Gamma(n_t)} \tau^{n_t - 1} \mathrm{e}(-\tau d_t) \\ &\propto \tau^{n_t - 1} \mathrm{e}(-\tau d_t) \end{aligned}$$

where both n_t and d_t are positive integers, and d_t follows a chi-squared distribution with degrees of freedom n_t, $d_t \sim \chi^2_{n_t}$.

In the following sections for the variable forgetting factor algorithms, both the upper limit λ_+ and the lower limit λ_- are specified by the modeller. According to Haykin (2001), λ_+ is normally set close to the unity while the user determines λ_- by experiment. The bracket followed by λ_- and λ_+, $[]_{\lambda_-}^{\lambda_+}$, indicates "truncation", restricting the forgetting factor to the interval $[\lambda_-, \lambda_+]$.

3.3 The Steepest Descent Variable Forgetting Factor (SDvFF) Algorithm

In Haykin (2001), the steepest descent method is described as

$$\lambda_t = [\lambda_{t-1} - \alpha \cdot \nabla_\lambda(t)]_{\lambda_-}^{\lambda_+} \tag{3.8}$$

where α, the convergence rate *or* a learning-rate parameter, is set to be 0.0005 in Malik (2006) and λ_+ and λ_- are the upper and lower limits of λ respectively. In the steepest descent method, α can take any real value, and it is fixed, normally as 0.5, by the modeller. When $\alpha = 0$, $\lambda_t = \lambda_{t-1} = \cdots = \lambda_1$, indicating that a forgetting factor is constant. In the field of neural network, α itself is sought after using an algorithm such as the steepest descent method again. Haykin (2001) and his co-authors have developed the steepest descent method above for recursive least squares models. In this section, we adopt (3.8) and we extend the steepest descent for variable forgetting for state space models.

$\nabla_\lambda(t)(\equiv \frac{\partial J_t}{\partial \lambda})$ is recursively updated as in the followings, the details of which are given in the subsections.

$$\nabla_\lambda(t) \approx -e_t \mathbf{F}_t' G \psi_{t-1} \tag{3.9}$$

where $\psi_t \equiv \frac{\partial m_t}{\partial \lambda}$ and m_t is the first moment of the posterior density for $(\theta_t \mid D_t)$ at t. Considering the adaptive scheme of a forgetting factor λ, the recursions of ψ_t and

$S_t(\equiv \frac{\partial C_t^*}{\partial \lambda})$ are obtained by

$$\begin{aligned} \psi_t &= (I - C_t^* \mathbf{F}_t \mathbf{F}_t') G \psi_{t-1} + S_t \mathbf{F}_t e_t & (3.10) \\ S_t &= \frac{G S_{t-1} G' (\lambda_{t-1} + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t) - G C_{t-1}^* G' (1 + \mathbf{F}_t' G S_{t-1} G' \mathbf{F}_t)}{(\lambda_{t-1} + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t)^2} & (3.11) \end{aligned}$$

where also the updating equations for C_t^* and K_t, respectively, reduce to

$$\begin{aligned} C_t^* &= \lambda_{t-1}^{-1} (I - K_t \mathbf{F}_t') G C_{t-1}^* G' & (3.12) \\ &= \frac{\lambda_{t-1}^{-1} G C_{t-1}^* G'}{1 + \lambda_{t-1}^{-1} \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t} & (3.13) \\ K_t &= \frac{\lambda_{t-1}^{-1} G C_{t-1}^* G' \mathbf{F}_t}{1 + \lambda_{t-1}^{-1} \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t} & (3.14) \end{aligned}$$

3.3.1 Recursion of $\nabla_\lambda(t)$

By defining $\psi_t \equiv \frac{\partial m_t}{\partial \lambda}$, $\nabla_\lambda(t)$ is given by

$$\begin{aligned} \nabla_\lambda(t) &= \frac{\partial J_t}{\partial \lambda} = \frac{\partial}{\partial \lambda} \{ \frac{1}{2} E(\mid e_t \mid^2) \} \\ &= \frac{1}{2} 2 E(e_t \frac{\partial e_t}{\partial \lambda}) \\ &= -E(e_t \mathbf{F}_t' G \psi_{t-1}) \\ &\approx -e_t \mathbf{F}_t' G \psi_{t-1} & (3.15) \end{aligned}$$

$$\begin{aligned} \text{where} \quad e_t &= Y_t - f_t = \mathbf{F}_t' \theta_t + \epsilon_t - \mathbf{F}_t' G m_{t-1} \quad \text{and} \\ \frac{\partial e_t}{\partial \lambda} &= -\mathbf{F}_t' G \frac{\partial m_{t-1}}{\partial \lambda} = -\mathbf{F}_t' G \psi_{t-1} \end{aligned}$$

3.3.1.1 Updating Equation for ψ_t

Using $K_t = C_t^* \mathbf{F}_t$ from (3.6) and $m_t = a_t + K_t e_t$ from the updating equation (d4), the recursion for $\psi_t \equiv \frac{\partial m_t}{\partial \lambda}$ is easily obtained by defining $S_t \equiv \frac{\partial C_t^*}{\partial \lambda}$ as follows.

$$\begin{aligned}
\psi_t &= \frac{\partial m_t}{\partial \lambda} = \frac{\partial (a_t + K_t e_t)}{\partial \lambda} = \frac{\partial (G m_{t-1} + C_t^* \mathbf{F}_t e_t)}{\partial \lambda} \\
&= G \frac{\partial m_{t-1}}{\partial \lambda} + \frac{\partial C_t^*}{\partial \lambda} \mathbf{F}_t e_t + C_t^* \mathbf{F}_t \frac{\partial e_t}{\partial \lambda} \\
&= G\psi_{t-1} + S_t \mathbf{F}_t e_t + C_t^* \mathbf{F}_t (-\mathbf{F}_t' G \psi_{t-1}) \\
&= G\psi_{t-1} + S_t \mathbf{F}_t e_t - C_t^* \mathbf{F}_t \mathbf{F}_t' G \psi_{t-1} \\
&= (I - C_t^* \mathbf{F}_t \mathbf{F}_t') G\psi_{t-1} + S_t \mathbf{F}_t e_t \\
&= (I - K_t \mathbf{F}_t') G\psi_{t-1} + S_t \mathbf{F}_t e_t
\end{aligned} \tag{3.16}$$

where $\frac{\partial K_t}{\partial \lambda} = S_t \mathbf{F}_t$ and $S_t \equiv \frac{\partial C_t^*}{\partial \lambda}$.

3.3.1.2 Updating Equation for S_t

From the recursion,

$$\begin{aligned}
K_t &= \frac{R_t^* \mathbf{F}_t}{Q_t^*} = \frac{R_t^* \mathbf{F}_t}{1 + \mathbf{F}_t' R_t^* \mathbf{F}_t} \\
&= \frac{\lambda^{-1} G C_{t-1}^* G' \mathbf{F}_t}{1 + \lambda^{-1} \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t} \qquad \text{from (3.14)} \\
&= \frac{G C_{t-1}^* G'}{\lambda + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t} \mathbf{F}_t
\end{aligned} \tag{3.17}$$

From the relation of C_t^* and K_t in (3.6), we obtain

$$C_t^* = \frac{G C_{t-1}^* G'}{\lambda + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t} \tag{3.18}$$

A partial differentiation of (3.18) with regard to λ gives

$$\begin{aligned}
S_t &\equiv \frac{\partial C_t^*}{\partial \lambda} = \frac{\partial \left(\frac{G C_{t-1}^* G'}{\lambda + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t} \right)}{\partial \lambda} \\
&= \frac{G S_{t-1} G' (\lambda + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t) - G C_{t-1}^* G' (1 + \mathbf{F}_t' G S_{t-1} G' \mathbf{F}_t)}{(\lambda + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t)^2}
\end{aligned} \tag{3.19}$$

Considering the adaptive scheme of a forgetting factor λ, a recursion of S_t at time t can be obtained by

$$S_t = \frac{GS_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)}{(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2} \tag{3.20}$$

3.4 The Gauss-Newton Variable Forgetting Factor (GNvFF) Algorithm

Song et al. (2000) suggest that the speed of tracking is improved by incorporating the second derivatives of the cost function, J_t, as in Ljung and Soderstrom (1983), proposing the recursion by

$$\lambda_t = \left[\lambda_{t-1} - \alpha \cdot \frac{\nabla_\lambda(t)}{\nabla_\lambda^2(t)}\right]_{\lambda_-}^{\lambda_+} \tag{3.21}$$

where $\nabla_\lambda(t) = \frac{\partial J_t}{\partial \lambda}$, $\nabla_\lambda^2(t) = \frac{\partial^2 J_t}{\partial \lambda^2}$, and α is the convergence rate, *or* a learning-rate parameter, is set to be 0.1 in Song et al. (2000). λ_+ and λ_+ are the upper and lower limit of λ respectively. Herein, the Gauss-Newton method is extended for the use to a DLM.

$\nabla_\lambda(t)$, followed by ψ_t and S_t, is recursively updated as seen in the SDvFF.

$$\begin{aligned}
\nabla_\lambda(t) &\approx -e_t\mathbf{F}_t'G\psi_{t-1} \\
\psi_t &= (I - C_t^*\mathbf{F}_t\mathbf{F}_t')G\psi_{t-1} + S_t\mathbf{F}_te_t \\
S_t &= \frac{GS_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)}{(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2}
\end{aligned}$$

where $\psi_t \equiv \frac{\partial m_t}{\partial \lambda}$ and $S_t \equiv \frac{\partial C_t^*}{\partial \lambda}$.

The recursions of $\nabla_\lambda^2(t)$, $\eta_t \equiv \frac{\partial \psi_t}{\partial \lambda}$, and $L_t \equiv \frac{\partial S_t}{\partial \lambda}$ are

$$\nabla_\lambda^2(t) \approx (\mathbf{F}_t'G\psi_{t-1})^2 - e_t\mathbf{F}_t'G\frac{\partial \psi_{t-1}}{\partial \lambda} \tag{3.22}$$

$$\eta_t = (I - C_t^*\mathbf{F}_t\mathbf{F}_t')G\eta_{t-1} + L_t\mathbf{F}_te_t - 2S_t\mathbf{F}_t\mathbf{F}_t'G\psi_{t-1} \tag{3.23}$$

where L_t is also recursively updated with a recursion of $L_t = \frac{A}{B}$ and

$$\begin{aligned} A &= \left\{GL_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(\mathbf{F}_t'GL_{t-1}G'\mathbf{F}_t)\right\} \\ &\quad \cdot \left\{(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2\right\} \\ &\quad - \left\{GS_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\} \\ &\quad \cdot \left\{2(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\} \qquad (3.24) \\ B &= (\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^4 \qquad (3.25) \end{aligned}$$

The detailed derivation of recursions follows as subsections.

3.4.1 Recursion of $\nabla_\lambda^2(t)$

$\nabla_\lambda^2(t)$ is defined and can be approximated as follows.

$$\begin{aligned} \nabla_\lambda^2(t) &= \frac{\partial^2 J_t}{\partial \lambda^2} = \frac{\partial}{\partial \lambda}\left(-E(e_t\mathbf{F}_t'G\psi_{t-1})\right) \\ &= -E(\frac{\partial e_t}{\partial \lambda}\mathbf{F}_t'G\psi_{t-1} + e_t\mathbf{F}_t'G\frac{\partial \psi_{t-1}}{\partial \lambda}) \\ &\approx (\mathbf{F}_t'G\psi_{t-1})^2 - e_t\mathbf{F}_t'G\frac{\partial \psi_{t-1}}{\partial \lambda} \qquad (3.26) \\ \text{where } \frac{\partial e_t}{\partial \lambda} &= -\mathbf{F}_t'G\frac{\partial m_{t-1}}{\partial \lambda} = -\mathbf{F}_t'G\psi_{t-1} \end{aligned}$$

3.4.1.1 Updating Equation for η_t

By defining $\eta_t \equiv \frac{\partial \psi_t}{\partial \lambda}$ and $L_t \equiv \frac{\partial S_t}{\partial \lambda}$, a recursion of $\eta_t \equiv \frac{\partial \psi_t}{\partial \lambda}$ can be derived as follows.

$$\begin{aligned} \eta_t &\equiv \frac{\partial \psi_t}{\partial \lambda} = \frac{\partial}{\partial \lambda}\{(I - C_t^*\mathbf{F}_t\mathbf{F}_t')G\psi_{t-1} + S_t\mathbf{F}_te_t\} \\ &= -\frac{\partial C_t^*}{\partial \lambda}\mathbf{F}_t\mathbf{F}_t'G\psi_{t-1} + (I - C_t^*\mathbf{F}_t\mathbf{F}_t')G\frac{\partial \psi_{t-1}}{\partial \lambda} + \frac{\partial S_t}{\partial \lambda}\mathbf{F}_te_t + S_t\mathbf{F}_t\frac{\partial e_t}{\partial \lambda} \\ &= -S_t\mathbf{F}_t\mathbf{F}_t'G\psi_{t-1} + (I - C_t^*\mathbf{F}_t\mathbf{F}_t')G\eta_{t-1} + L_t\mathbf{F}_te_t + S_t\mathbf{F}_t(-\mathbf{F}_t'G\psi_{t-1}) \\ &= (I - C_t^*\mathbf{F}_t\mathbf{F}_t')G\eta_{t-1} + L_t\mathbf{F}_te_t - 2S_t\mathbf{F}_t\mathbf{F}_t'G\psi_{t-1} \qquad (3.27) \\ &= (I - K_t\mathbf{F}_t')G\eta_{t-1} + L_t\mathbf{F}_te_t - 2S_t\mathbf{F}_t\mathbf{F}_t'G\psi_{t-1} \end{aligned}$$

3.4.1.2 Updating Equation for L_t

A recursion for $L_t(\equiv \frac{\partial S_t}{\partial \lambda})$ also can be derived as follows.

$$L_t \equiv \frac{\partial}{\partial \lambda}\left\{\frac{GS_{t-1}G'(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1+\mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)}{(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2}\right\} \tag{3.28}$$

A partial differentiation of a numerator in (3.28) gives

$$\begin{aligned}
&\frac{\partial}{\partial \lambda}\left\{GS_{t-1}G'(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1+\mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\} \\
&\quad = GL_{t-1}G'(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) + GS_{t-1}G'(1+\mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t) \\
&\qquad -GS_{t-1}G'(1+\mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t) - GC_{t-1}^*G'(\mathbf{F}_t'GL_{t-1}G'\mathbf{F}_t) \\
&\quad = GL_{t-1}G'(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(\mathbf{F}_t'GL_{t-1}G'\mathbf{F}_t)
\end{aligned} \tag{3.29}$$

A partial differentiation of a denominator in (3.28) gives

$$\frac{\partial(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2}{\partial \lambda} = 2(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)(1+\mathbf{F}_t'GS_{t-1}^*G'\mathbf{F}_t) \tag{3.30}$$

Therefore, combining terms in (3.29) and (3.29) by the principle of differentiation gives $L_t = \frac{A}{B}$ where

$$\begin{aligned}
A &= \left\{GL_{t-1}G'(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(\mathbf{F}_t'GL_{t-1}G'\mathbf{F}_t)\right\} \\
&\quad \cdot \left\{(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2\right\} \\
&\quad - \left\{GS_{t-1}G'(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1+\mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\} \\
&\quad \cdot \left\{2(\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)(1+\mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\} \qquad (3.31) \\
B &= (\lambda + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^4 \qquad (3.32)
\end{aligned}$$

Considering the adaptive scheme of a forgetting factor, a recursion to compute a

numerator and a denominator of L_t finally reduces to $L_t = \frac{A}{B}$ where

$$\begin{aligned} A &= \left\{GL_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(\mathbf{F}_t'GL_{t-1}G'\mathbf{F}_t)\right\} \\ &\quad \cdot \left\{(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2\right\} \\ &\quad - \left\{GS_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\} \\ &\quad \cdot \left\{2(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\} \quad (3.33) \\ B &= (\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^4 \quad (3.34) \end{aligned}$$

3.5 The Beta-Bernoulli Variable Forgetting Factor (BBvFF) Algorithm

A new algorithm referred to as the BBvFF(d, k) is proposed to achieve an adaptive forgetting to track down the changes of the data stream, where $d(> 0)$ is a threshold for a Bernoulli process x_t and k $(0 < k \leq 1)$ is a discount factor in the steady forecasting models by Smith (1979). The idea is that λ_t will take the upper limit λ_+ with probability π_t and the lower limit λ_- with probability $1 - \pi_t$. Hence, to determine λ_t, we define

$$\lambda_t = \pi_t\lambda_+ + (1 - \pi_t)\lambda_- \quad (3.35)$$

where π_t is estimated as $\hat{\pi}_t = \text{mode}(\pi_t \mid x_t) = \frac{\alpha_{1,t}-1}{\alpha_{1,t}+\alpha_{2,t}-2}$ and $\pi_t \sim \text{Beta}(\alpha_{1,t}, \alpha_{2,t})$. Also, λ_+ and λ_- are upper and lower limits of λ_t where $0 < \lambda_t \leq 1$ as before.

Steady forecasting models by Smith (1979) may propose an evolution for π_t using the conjugacy of the beta-Bernoulli as follows.

$$p(\pi_t \mid D_{t-1}) \propto p(\pi_{t-1} \mid D_{t-1})^k \quad (3.36)$$

where k is a discount factor and $0 < k \leq 1$. Especially when $k = 1$, $p(\pi_t \mid D_{t-1}) = p(\pi_{t-1} \mid D_{t-1})$ and $\pi_t = \pi$. For $0 < k < 1$, π_t changes over time.

Suppose that $(\pi_{t-1} \mid D_{t-1}) \sim \text{Beta}(\alpha_{1,t-1}, \alpha_{2,t-1})$. If we write $p_1(\pi \mid D_{t-1}) = p(\pi_t \mid$

$D_{t-1})$ and $p_2(\pi \mid D_{t-1}) = p(\pi_{t-1} \mid D_{t-1})$, then

$$\begin{aligned} p_1(\pi \mid D_{t-1}) &\propto p_2(\pi \mid D_{t-1})^k && (3.37)\\ &\propto \pi^{(\alpha_{1,t-1}-1)k}(1-\pi)^{(\alpha_{2,t-1}-1)k} && (3.38)\\ &= \pi^{\alpha_{1,t-1}k-k+1-1}(1-\pi)^{\alpha_{2,t-1}k-k+1-1} && (3.39) \end{aligned}$$

so that $(\pi_t \mid D_{t-1}) \sim \text{Beta}(\alpha_{1,t-1}k - k + 1, \alpha_{2,t-1}k - k + 1)$.

This approach keeps the mode unchanged from $t-1$ to t such that

$$\begin{aligned} \text{mode}(\pi_{t-1} \mid D_{t-1}) &= \frac{\alpha_{1,t-1} - 1}{\alpha_{1,t-1} + \alpha_{2,t-1} - 2}\\ \text{mode}(\pi_t \mid D_{t-1}) &= \frac{\alpha_{1,t-1}k - k + 1 - 1}{\alpha_{1,t-1}k - k + 1 + \alpha_{2,t-1}k - k + 1 - 2}\\ &= \frac{\alpha_{1,t-1}k - k}{\alpha_{1,t-1}k - k + \alpha_{2,t-1}k - k}\\ &= \frac{\alpha_{1,t-1} - 1}{\alpha_{1,t-1} + \alpha_{2,t-1} - 2} \end{aligned}$$

Suppose that x_t is a binary series, taking a value of either 1 or 0 at each time t according to

$$x_t = \begin{cases} 1, & \text{if } \frac{|e_t|}{\sqrt{Q_t}} \le d, \quad \text{with probability } \pi \\ 0, & \text{if } \frac{|e_t|}{\sqrt{Q_t}} > d, \quad \text{with probability } 1-\pi \end{cases}$$

where $d(>0)$ is a threshold specified by the modeller.

We define x_t to be a Bernoulli process at any time t, and herein, in a broad sense, π_t can be a probability of "success" as a measure of uncertainty. Considering a sequence of Bernoulli trials, having a result as "success" or "failure" at t, π_t can be regarded as the proportion of "successes" in the population up to time t, or the probability in a trial at t. Still, in the above framework of (3.35) in the BBvFF(d, k), π_t is not known, but to be estimated. Using the conjugacy of the beta-Bernoulli distributions, it is easily obtained as follows.

Given π, the likelihood function at t from the observed data x_t is a Bernoulli distribution, and written as

$$p(x_t \mid \pi) = \text{Bernoulli}(x_t \mid \pi) = \pi^{x_t}(1-\pi)^{1-x_t}$$

where $\pi \in [0, 1]$.

A prior distribution for π is specified to be a beta distribution with parameters $(\alpha_{1,t-1}k-k+1)$ and $(\alpha_{2,t-1}k-k+1)$. Thus, $\pi \sim \text{Beta}(\alpha_{1,t-1}k-k+1, \alpha_{2,t-1}k-k+1)$ and its density is

$$\begin{aligned} p(\pi) &= \text{Beta}(\pi \mid \alpha_{1,t-1}k-k+1, \alpha_{2,t-1}k-k+1) \\ &= \frac{\Gamma(\alpha_{1,t-1}k-k+1, \alpha_{2,t-1}k-k+1)}{\Gamma(\alpha_{1,t-1}k-k+1)\Gamma(\alpha_{2,t-1}k-k+1)}\pi^{\alpha_{1,t-1}k-k+1-1}(1-\pi)^{\alpha_{2,t-1}k-k+1-1} \end{aligned}$$

where both $(\alpha_{1,t-1}k-k+1)$ and $(\alpha_{2,t-1}k-k+1) > 0$ and $\Gamma(\cdot)$ denotes the gamma function. It is noted that the distribution of π is implicitly conditional on data up to time $t-1$.

By applying the Bayes' theorem, the posterior distribution for π is

$$\begin{aligned} p(\pi \mid x_t) &\propto p(x_t \mid \pi)p(\pi) \\ &\propto \pi^{x_t}(1-\pi)^{1-x_t}\pi^{\alpha_{1,t-1}k-k+1-1}(1-\pi)^{\alpha_{2,t-1}k-k+1-1} \\ &= \pi^{\alpha_{1,t-1}k-k+1+x_t-1}(1-\pi)^{\alpha_{2,t-1}k-k+1+1-x_t-1} \\ &= \text{Beta}(\alpha_{1,t-1}k-k+1+x_t, \alpha_{2,t-1}k-k+2-x_t) \end{aligned} \tag{3.40}$$

Applying the above sequentially, we obtain that $(\pi \mid x_1, \ldots, x_t) \equiv (\pi \mid D_t) \sim \text{Beta}(\alpha_{1,t}, \alpha_{2,t})$ where $\alpha_{1,t} = \alpha_{1,t-1}k-k+1+x_t$ and $\alpha_{2,t} = \alpha_{2,t-1}k-k+2-x_t$ so that $(\pi_t \mid D_t) \sim \text{Beta}(\alpha_{1,t}, \alpha_{2,t})$. Bearing in mind the above results, π at t can be estimated as follows.

$$\hat{\pi}_t = \text{mode}(\pi_t \mid D_t) = \frac{\alpha_{1,t}-1}{\alpha_{1,t}+\alpha_{2,t}-2} = \frac{\alpha_{1,t-1}k-k+x_t}{\alpha_{1,t-1}k+\alpha_{2,t-1}k-2k+1} \tag{3.41}$$

Now that we write $p_1(\pi \mid D_t) = p(\pi_{t+1} \mid D_t)$ and $p_2(\pi \mid D_t) = p(\pi_t \mid D_t)$, then

$$\begin{aligned} p_1(\pi \mid D_t) &\propto p_2(\pi \mid D_t)^k && (3.42)\\ &\propto \pi^{(\alpha_{1,t}-1)k}(1-\pi)^{(\alpha_{2,t}-1)k} && (3.43)\\ &= \pi^{\alpha_{1,t}k-k+1-1}(1-\pi)^{\alpha_{2,t}k-k+1-1} && (3.44) \end{aligned}$$

so that $(\pi_{t+1} \mid D_t) \sim \text{Beta}(\alpha_{1,t}k - k + 1, \alpha_{2,t}k - k + 1)$.

This approach keeps the mode unchanged from t to $t+1$ such that

$$\begin{aligned} \text{mode}(\pi_t \mid D_t) &= \frac{\alpha_{1,t} - 1}{\alpha_{1,t} + \alpha_{2,t} - 2}\\ \text{mode}(\pi_{t+1} \mid D_t) &= \frac{\alpha_{1,t}k - k + 1 - 1}{\alpha_{1,t}k - k + 1 + \alpha_{2,t}k - k + 1 - 2}\\ &= \frac{\alpha_{1,t}k - k}{\alpha_{1,t}k - k + \alpha_{2,t}k - k}\\ &= \frac{\alpha_{1,t} - 1}{\alpha_{1,t} + \alpha_{2,t} - 2} \end{aligned}$$

Given initial values of $\alpha_{1,0}$ and $\alpha_{2,0}$, the above development suggests a sequential algorithm, which basically runs the Kalman filter conditional on the forgetting factor λ_{t-1} and then updates the forgetting factor according to the beta-Bernoulli procedure. For $\pi_0 \sim \text{Beta}(\alpha_{1,0}, \alpha_{2,0})$, the initial values can be proposed to be set as $\alpha_{1,0} = \alpha_{2,0} = 2$. Noting that $\hat{\pi}_0 = (\alpha_{1,0} - 1) \cdot (\alpha_{1,0} + \alpha_{2,0} - 2)^{-1}$, $\hat{\pi}_1$ becomes $1/2$. This is motivated by the reasoning that since we have no data observed, the probability π is assumed to be 0.5 at the beginning.

3.5.1 Advantages of the BBvFF over the other algorithms

The BBvFF(d, k) has a couple of advantages of its own. First of all, at each time t, λ_t of the BBvFF(d, k) from (3.35) is guaranteed to lie between λ_+ and λ_-, which is

easily proven in the following:

$$\begin{aligned}
\lambda_t &= \pi_t\lambda_+ + (1-\pi_t)\lambda_- \\
&= \pi_t(\lambda_+ - \lambda_-) + \lambda_- \\
&\geq \lambda_- \\
&\text{and} \\
\lambda_t &= (\pi_t - 1 + 1)\lambda_+ + (1-\pi_t)\lambda_- \\
&= -(1-\pi_t)\lambda_+ + \lambda_+ + (1-\pi_t)\lambda_- \\
&= (1-\pi_t)(\lambda_- - \lambda_+) + \lambda_+ \\
&\leq \lambda_+
\end{aligned}$$

as $\lambda_+ > \lambda_-$ and $1 - \pi_t \geq 0$.

Secondly, in the BBvFF(d, k), the forgetting factor can be regarded as a random variable as it is a function of π from (3.35). Rearranging (3.35) gives $\pi_t = \frac{\lambda_t - \lambda_-}{\lambda_+ - \lambda_-}$. It is known that $(\pi_t \mid D_t) \sim \text{Beta}(\alpha_{1,t}, \alpha_{2,t})$, and

$$p(\pi_t \mid D_t) = \frac{\Gamma(\alpha_{1,t}, \alpha_{2,t})}{\Gamma(\alpha_{1,t})\Gamma(\alpha_{2,t})}\pi_t^{\alpha_{1,t}-1}(1-\pi_t)^{\alpha_{2,t}-1}$$

where both $\alpha_{1,t}$ and $\alpha_{2,t} > 0$ and $\Gamma(\cdot)$ denotes the gamma function. Hence, the distribution of $(\lambda_t \mid D_t)$ can be obtained by

$$\begin{aligned}
p(\lambda_t \mid D_t) &= \frac{\Gamma(\alpha_{1,t}, \alpha_{2,t})}{\Gamma(\alpha_{1,t})\Gamma(\alpha_{2,t})}\left(\frac{\lambda_t - \lambda_-}{\lambda_+ - \lambda_-}\right)^{\alpha_{1,t}-1}\left(1 - \frac{\lambda_t - \lambda_-}{\lambda_+ - \lambda_-}\right)^{\alpha_{2,t}-1} \\
&= \frac{\Gamma(\alpha_{1,t} + \alpha_{2,t})}{\Gamma(\alpha_{1,t})\Gamma(\alpha_{2,t})}\frac{(\lambda_t - \lambda_-)^{\alpha_{1,t}-1}(\lambda_+ - \lambda_t)^{\alpha_{2,t}-1}}{(\lambda_+ - \lambda_-)^{\alpha_{1,t}+\alpha_{2,t}-2}} \qquad (3.45)
\end{aligned}$$

This notion of the distribution of λ helps to analyze the uncertainty associated with λ_t. From (3.35), λ_t can be rearranged as $\lambda_t = \pi_t(\lambda_+ - \lambda_-) + \lambda_-$, and therefore, noting that $(\pi_t \mid D_t) \sim \text{Beta}(\alpha_{1,t}, \alpha_{2,t})$, the variance of $(\lambda_t \mid D_t)$ is

$$\text{Var}(\lambda_t \mid D_t) = (\lambda_+ - \lambda_-)^2\frac{\alpha_{1,t}\alpha_{2,t}}{(\alpha_{1,t} + \alpha_{2,t})^2(\alpha_{1,t} + \alpha_{2,t} + 1)} \qquad (3.46)$$

where

$$\mathrm{Var}(\pi_t \mid D_t) = \frac{\alpha_{1,t}\alpha_{2,t}}{(\alpha_{1,t}+\alpha_{2,t})^2(\alpha_{1,t}+\alpha_{2,t}+1)}$$

For $\alpha_{1,t} > 1$ and $\alpha_{2,t} > 1$, even the mode of $(\lambda_t \mid D_t)$ can be found. Firstly, by taking the log of $p(\lambda_t \mid D_t)$, we have $\log(\lambda_t \mid D_t) = \log K + (\alpha_{1,t}-1)\cdot\log(\lambda_t - \lambda_-) + (\alpha_{2,t}-1)\cdot\log(\lambda_+ - \lambda_t)$ where $K = \frac{\Gamma(\alpha_{1,t}+\alpha_{2,t})}{\Gamma(\alpha_{1,t})\Gamma(\alpha_{2,t})}\frac{1}{(\lambda_+-\lambda_-)^{\alpha_{1,t}+\alpha_{2,t}-2}}$. Secondly, partial differentiation of $\log(\lambda_t \mid D_t)$ with regard to λ_t gives

$$\frac{\partial \log(\lambda_t \mid D_t)}{\partial \lambda_t} = \left(\frac{\alpha_{1,t}-1}{\lambda_t - \lambda_-}\right) - \left(\frac{\alpha_{2,t}-1}{\lambda_+ - \lambda_t}\right)$$
$$\frac{\partial^2 \log(\lambda_t \mid D_t)}{\partial \lambda_t^2} = -\frac{\alpha_{1,t}-1}{(\lambda_t-\lambda_-)^2} - \frac{\alpha_{2,t}-1}{(\lambda_+-\lambda_t)^2}$$

By making $\frac{\partial \log(\lambda_t|D_t)}{\partial \lambda_t} = 0$ and the fact that $\frac{\partial^2 \log(\lambda_t|D_t)}{\partial^2 \lambda_t} < 0$, we find the mode of $(\lambda_t \mid D_t)$ is

$$\mathrm{mode}(\lambda_t \mid D_t) = \frac{(\alpha_{1,t}-1)\lambda_+ + (\alpha_{2,t}-1)\lambda_-}{\alpha_{1,t}+\alpha_{2,t}-2} \tag{3.47}$$

3.6 Pseudo-code Implementations of The VFF Algorithms

3.6.1 The SDvFF

Table 3.1 shows the pseudo-code implementation of the SDvFF. An example of the initialisation at $t = 0$ could be that $\mathbf{m}_0 = (1,1)'$, $C_0 = I_2$, $n_0 = d_0 = 1$ where $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$, $\alpha = 0.5$, $\nabla_\lambda(0) = 0$, $\boldsymbol{\psi}_0 = (1,1)'$, and $S_0 = I_2$ with $G = diag(0.95, 2)$, an 2×2 evolution matrix. As for the variable forgetting factor λ, $\lambda_0 = 0.8$ as $0 < \lambda_t \leq 1$, $\lambda_+ = 1$ and $\lambda_- = 0.01$ as $0 < \lambda_t \leq 1$.

3.6.2 The GNvFF

Table 3.2 shows the pseudo-code implementation of the GNvFF. An example of the initialisation at $t = 0$ could be that $\mathbf{m}_0 = (1,1)'$, $C_0 = I_2$, $n_0 = d_0 = 1$

Table 3.1: Pseudo-code implementations for the SDvFF

1.	Initialisation at $t = 0$
·	Set $\mathbf{m}_0$, C_0, U_0 and n_0 where $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$ for $(\boldsymbol{\theta}_0 \mid D_0) \sim T_{n_0}[\mathbf{m}_0, C_0]$ and $(\tau \mid D_0) \sim Gamma\left(\frac{n_0}{2}, \frac{d_0}{2}\right)$
·	Set λ_0 as $0 < \lambda_t \leq 1$, α, $\nabla_\lambda(0)$, $\boldsymbol{\psi}_0$, and S_0
·	Set G an 2×2 evolution matrix
2.	Recursions for $t = 1, 2, \ldots, T$
·	$(\boldsymbol{\theta}_t \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{a}_t, R_t)$ where $\mathbf{a}_t = G\mathbf{m}_{t-1}$, $R_t = U_{t-1}R_t^*$ and $R_t^* = \frac{1}{\lambda_{t-1}}GC_{t-1}^*G'$
·	$(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}(f_t, Q_t)$ where $f_t = \mathbf{F}_t'G\mathbf{m}_{t-1}$, $Q_t = U_{t-1}Q_t^*$ and $Q_t^* = \frac{1}{\lambda_{t-1}}(\mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) + 1$
·	$(\theta_t \mid D_t) \sim T_{n_t}(\mathbf{m}_t, C_t)$ where $\mathbf{m}_t = \mathbf{a}_t + K_te_t$, $C_t = U_tC_t^*$ and $C_t^* = \frac{GC_{t-1}^*G'}{\lambda_{t-1}+\mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t}$ with $K_t = R_t^*\mathbf{F}_t/Q_t^*$ and $e_t = Y_t - f_t$
·	$\lambda_t = [\lambda_{t-1} - \alpha \cdot \nabla_\lambda(t)]_{\lambda_-}^{\lambda_+}$ where $\nabla_\lambda(t) \approx -e_t\mathbf{F}_t'G\psi_{t-1}$, $\psi_t = (I - C_t^*\mathbf{F}_t\mathbf{F}_t')G\psi_{t-1} + S_t\mathbf{F}_te_t$, and $S_t = \frac{GS_{t-1}G'(\lambda_{t-1}+\mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1+\mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)}{(\lambda_{t-1}+\mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2}$

where $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$, $\alpha = 0.5$, $\nabla_\lambda(0) = 0$, $\boldsymbol{\psi}_0 = (1, 1)'$, and $S_0 = I_2$ with $G = diag(0.95, 2)$, an 2×2 evolution matrix, which are the same as for the SDvFF. In addition, $\boldsymbol{\eta}_0 = (1, 1)'$, and $L_0 = I_2$ for the GNvFF. As for the variable forgetting factor λ, $\lambda_0 = 0.8$ as $0 < \lambda_t \leq 1$, $\lambda_+ = 1$ and $\lambda_- = 0.01$ as $0 < \lambda_t \leq 1$.

3.6.3 The BBvFF(d, k)

Table 3.3 shows the pseudo-code implementation of the BBvFF(d, k). An example of the initialisation at $t = 0$ could be that $\mathbf{m}_0 = (1, 1)'$, $C_0 = I_2$, and $n_0 = d_0 = 1$ where $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$ with $G = diag(0.95, 2)$, an 2×2 evolution matrix. As for the variable forgetting factor λ, $\alpha_{1,0} = \alpha_{2,0} = 2$ for $(\pi_0 \mid D_0) \sim \mathrm{Beta}(\alpha_{1,0}, \alpha_{2,0})$ as $\lambda_0 = \mathrm{mode}(\pi_0) = \hat{\pi}_0 \cdot \lambda_+ + (1 - \hat{\pi}_0) \cdot \lambda_-$ where $\hat{\pi}_0 = \frac{\alpha_{1,0}-1}{\alpha_{1,0}+\alpha_{2,0}-2}$, $\lambda_+ = 1$ and $\lambda_- = 0.01$ as $0 < \lambda_t \leq 1$. Additionally, a threshold d=0.1 or 1.96 as $d > 0$ and a discount factor k as $0 < k \leq 1$ as seen in Smith (1979).

Table 3.2: Pseudo-code implementations for the GNvFF

1. Initialisation at $t = 0$

- Set $\mathbf{m}_0$, C_0, U_0 and n_0 where $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$
  for $(\boldsymbol{\theta}_0 \mid D_0) \sim T_{n_0}[\mathbf{m}_0, C_0]$ and $(\tau \mid D_0) \sim Gamma\left(\frac{n_0}{2}, \frac{d_0}{2}\right)$
- Set λ_0 as $0 < \lambda_t \leq 1$, α, $\nabla_\lambda(0)$, $\boldsymbol{\psi}_0$, and S_0
- Set $\boldsymbol{\eta}_0$, and L_0
- Set G an 2×2 evolution matrix

2. Recursions for $t = 1, 2, \ldots, T$

- $(\boldsymbol{\theta}_t \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{a}_t, R_t)$
  where $\mathbf{a}_t = G\mathbf{m}_{t-1}$, $R_t = U_{t-1}R_t^*$ and $R_t^* = \frac{1}{\lambda_{t-1}} GC_{t-1}^* G'$
- $(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}(f_t, Q_t)$
  where $f_t = \mathbf{F}_t' G\mathbf{m}_{t-1}$, $Q_t = U_{t-1}Q_t^*$ and $Q_t^* = \frac{1}{\lambda_{t-1}}(\mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t) + 1$
- $(\theta_t \mid D_t) \sim T_{n_t}(\mathbf{m}_t, C_t)$
  where $\mathbf{m}_t = \mathbf{a}_t + K_t e_t$, $C_t = U_t C_t^*$ and $C_t^* = \frac{GC_{t-1}^* G'}{\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t}$
  with $K_t = R_t^* \mathbf{F}_t / Q_t^*$ and $e_t = Y_t - f_t$
- $\lambda_t = \left[\lambda_{t-1} - \alpha \cdot \frac{\nabla_\lambda(t)}{\nabla_\lambda^2(t)}\right]_{\lambda_-}^{\lambda_+}$
  where $\nabla_\lambda(t) \approx -e_t \mathbf{F}_t' G\psi_{t-1}$, $\psi_t = (I - C_t^* \mathbf{F}_t \mathbf{F}_t')G\psi_{t-1} + S_t \mathbf{F}_t e_t$,
  $S_t = \frac{GS_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t) - GC_{t-1}^* G'(1 + \mathbf{F}_t' GS_{t-1}G'\mathbf{F}_t)}{(\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t)^2}$,
  $\nabla_\lambda^2(t) \approx (\mathbf{F}_t' G\psi_{t-1})^2 - e_t \mathbf{F}_t' G\frac{\partial \psi_{t-1}}{\partial \lambda}$,
  $\eta_t = (I - C_t^* \mathbf{F}_t \mathbf{F}_t')G\eta_{t-1} + L_t \mathbf{F}_t e_t - 2S_t \mathbf{F}_t \mathbf{F}_t' G\psi_{t-1}$, and $L_t = \frac{A}{B}$
  with $A = \left\{GL_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t) - GC_{t-1}^* G'(\mathbf{F}_t' GL_{t-1}G'\mathbf{F}_t)\right\}$
  $\cdot \left\{(\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t)^2\right\}$
  $- \left\{GS_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t) - GC_{t-1}^* G'(1 + \mathbf{F}_t' GS_{t-1}G'\mathbf{F}_t)\right\}$
  $\cdot \left\{2(\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t)(1 + \mathbf{F}_t' GS_{t-1}G'\mathbf{F}_t)\right\}$
  and $B = (\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G'\mathbf{F}_t)^4$

3.7 Comparisons with Simulated Time Series

In algorithmic pairs trading, accurate forecast of the series is critical. In this section, the performance of three algorithms of the SDvFF, the GNvFF, and the BBvFF(d, k) are compared and assessed.

For simulation study, a time series is generated by $Y_t = a \cdot Y_{t-1} + \epsilon_t, \epsilon_t \sim \mathrm{N}(0, V)$ where $(a, V) = \{(0.1, 1), (0.1, 100), (0.5, 1), (0.5, 100), (0.9, 1), (0.9, 100)\}$. Each combination of a and V is iterated for 1,000 times, generating 1,000 times series of 30 data points each. For example, with $a = 0.1$ and $V = 1$, $Y_t = 0.1 \cdot Y_{t-1} + \epsilon_t, \epsilon_t \sim \mathrm{N}(0, 1)$

Table 3.3: Pseudo-code implementations for the BBvFF(d, k)

1. Initialisation at $t = 0$
· Set $\mathbf{m}_0$, C_0, U_0 and n_0 where $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$ for $(\boldsymbol{\theta}_0 \mid D_0) \sim T_{n_0}[\mathbf{m}_0, C_0]$ and $(\tau \mid D_0) \sim Gamma\left(\frac{n_0}{2}, \frac{d_0}{2}\right)$
· Set $\alpha_{1,0}$ and $\alpha_{2,0}$ for $(\pi_0 \mid D_0) \sim \mathrm{Beta}(\alpha_{1,0}, \alpha_{2,0})$
· Set λ_+ and λ_- as $0 < \lambda_+, \lambda_- \leq 1$ for $\lambda_0 = \mathrm{mode}(\pi_0) = \hat{\pi}_0 \cdot \lambda_+ + (1 - \hat{\pi}_0) \cdot \lambda_-$ where $\hat{\pi}_0 = \frac{\alpha_{1,0}-1}{\alpha_{1,0}+\alpha_{2,0}-2}$
· Set d as $d > 0$ and k as $0 < k \leq 1$
· Set G an 2×2 evolution matrix
2. Recursions for $t = 1, 2, \ldots, T$
· $(\boldsymbol{\theta}_t \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{a}_t, R_t)$ where $\mathbf{a}_t = G\mathbf{m}_{t-1}$, $R_t = U_{t-1}R_t^*$ and $R_t^* = \frac{1}{\lambda_{t-1}} G C_{t-1}^* G'$
· $(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}(f_t, Q_t)$ where $f_t = \mathbf{F}_t' G \mathbf{m}_{t-1}$, $Q_t = U_{t-1}Q_t^*$ and $Q_t^* = \frac{1}{\lambda_{t-1}}(\mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t) + 1$
· $(\theta_t \mid D_t) \sim T_{n_t}(\mathbf{m}_t, C_t)$ where $\mathbf{m}_t = \mathbf{a}_t + K_t e_t$, $C_t = U_t C_t^*$ and $C_t^* = \frac{G C_{t-1}^* G'}{\lambda_{t-1} + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t}$ with $K_t = R_t^* \mathbf{F}_t / Q_t^*$ and $e_t = Y_t - f_t$
· $x_t = \begin{cases} 1, & \text{if } \frac{\lvert e_t \rvert}{\sqrt{Q_t}} \leq d \\ 0, & \text{if } \frac{\lvert e_t \rvert}{\sqrt{Q_t}} > d \end{cases}$ where $d(> 0)$ is a threshold specified by the modeller
· $\lambda_t = \hat{\pi}_t \cdot \lambda_+ + (1 - \hat{\pi}_t) \cdot \lambda_-$ from $(\pi_t \mid D_t) \sim \mathrm{Beta}(\alpha_{1,t}, \alpha_{2,t})$ where $\alpha_{1,t} = \alpha_{1,t-1}k - k + 1 + x_t$, $\alpha_{2,t} = \alpha_{2,t-1}k - k + 2 - x_t$ and $\hat{\pi}_t = \mathrm{mode}(\pi_t \lvert D_t) = \frac{\alpha_{1,t}-1}{\alpha_{1,t}+\alpha_{2,t}-2} = \frac{\alpha_{1,t-1}k-k+x_t}{\alpha_{1,t-1}k+\alpha_{2,t-1}k-2k+1}$

and 1,000 time series, each of which contains 30 data points, are generated. To each time series of 30 data points, the DLM with the VFF algorithm is applied to see its performance in prediction and assessed by mean squared errors (MSE).

Table 3.4, 3.5, 3.6, and 3.7 summarise the mean and the standard errors (s.e.) of the MSE by the DLM with each of the VFF algorithms such as the SDvFF, the GNvFF, and the BBvFF(d,k) where d=0.1, 1.96 and k=1, 0.99, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5 for the purpose of comparisons. A difference between the tables of Table 3.4 and 3.5 and of Table 3.6 and 3.7 lies in the different d for the BBvFF(d,k). Thus, no additional information is necessary from the SDvFF and the GNvFF for the tables

of Table 3.6 and 3.7.

From the simulation study, it is found that the performance of the BBvFF(d,k) may differ with different choices of d and k as seen in the summary tables. From the mean and the s.e. of the MSE in Table 3.4 and 3.5, it can be seen that the BBvFF(0.1,k) outperforms the SDvFF and the GNvFF except for the case of $(a, V) = (0.9, 1)$ where the BBvFF(0.1,k) outperforms the SDvFF only. Among the BBvFF(0.1,k) with different k, performances in forecasting get better as k decreases except for the cases of $(a, V) = \{(0.9, 1), (0.9, 100)\}$ where the performance of the BBvFF(0.1,k) rather deteriorates with decreasing k. From the mean and the s.e. of the MSE in Table 3.6 and 3.7, it is seen that the BBvFF(0.1,k) outperforms the SDvFF only for all the cases of (a, V). Among the BBvFF(1.96,k) with different k, performances in forecasting get worse as k decreases.

Table 3.4: The mean and the standard errors (s.e.) in the bracket of the MSE by the SDvFF, the GNvFF, and the BBvFF(d,k) with d=0.1 and k=1,0.99,0.95

(a, V)	SDvFF	GNvFF	BBvFF(0.1,1)	BBvFF(0.1,0.99)	BBvFF(0.1,0.95)
(0.1,1)	1.924(0.778)	1.580(0.535)	1.430(0.382)	1.430(0.381)	1.428(0.380)
(0.1,100)	3.766(6.148)	2.966(4.389)	2.226(1.548)	2.225(1.546)	2.220(1.539)
(0.5,1)	1.879(0.679)	1.541(0.463)	1.459(0.377)	1.459(0.378)	1.459(0.377)
(0.5,100)	3.362(3.582)	2.584(2.211)	2.199(1.513)	2.199(1.511)	2.196(1.501)
(0.9,1)	1.819(0.632)	1.537(0.417)	1.601(0.400)	1.603(0.401)	1.609(0.404)
(0.9,100)	3.068(4.745)	2.501(2.406)	2.320(1.317)	2.321(1.315)	2.327(1.306)

Table 3.5: The mean and the standard errors (s.e.) in the bracket of the MSE by the BBvFF(d,k) with d=0.1 and k=0.9,0.8,0.7,0.6,0.5

(a, V)	BBvFF(0.1,0.9)	BBvFF(0.1,0.8)	BBvFF(0.1,0.7)	BBvFF(0.1,0.6)	BBvFF(0.1,0.5)
(0.1,1)	1.426(0.379)	1.423(0.377)	1.420(0.374)	1.417(0.372)	1.415(0.370)
(0.1,100)	2.214(1.532)	2.205(1.519)	2.197(1.510)	2.191(1.506)	2.188(1.511)
(0.5,1)	1.459(0.377)	1.459(0.378)	1.458(0.377)	1.457(0.375)	1.455(0.374)
(0.5,100)	2.193(1.489)	2.187(1.465)	2.181(1.440)	2.174(1.413)	2.168(1.384)
(0.9,1)	1.616(0.407)	1.625(0.412)	1.629(0.416)	1.630(0.418)	1.629(0.419)
(0.9,100)	2.334(1.296)	2.342(1.280)	2.345(1.269)	2.344(1.264)	2.341(1.264)

Table 3.6: The mean and the standard errors (s.e.) in the bracket of the MSE by the BBvFF(d,k) with d=1.96 and k=1,0.99,0.95,0.9

(a, V)	BBvFF(1.96,1)	BBvFF(1.96,0.99)	BBvFF(1.96,0.95)	BBvFF(1.96,0.9)
(0.1,1)	1.852(0.614)	1.855(0.616)	1.867(0.623)	1.879(0.633)
(0.1,100)	3.704(5.123)	3.717(5.187)	3.735(4.898)	3.781(5.027)
(0.5,1)	1.755(0.548)	1.757(0.550)	1.765(0.555)	1.774(0.560)
(0.5,100)	3.249(3.091)	3.257(3.105)	3.288(3.158)	3.314(3.173)
(0.9,1)	1.598(0.487)	1.599(0.488)	1.604(0.492)	1.608(0.497)
(0.9,100)	3.016(4.255)	3.020(4.269)	3.045(4.334)	3.068(4.390)

Table 3.7: The mean and the standard errors (s.e.) in the bracket of the MSE by the BBvFF(d,k) with d=1.96 and k=0.8,0.7,0.6,0.5

(a, V)	BBvFF(1.96,0.8)	BBvFF(1.96,0.7)	BBvFF(1.96,0.6)	BBvFF(1.96,0.5)
(0.1,1)	1.901(0.649)	1.921(0.666)	1.933(0.674)	1.948(0.689)
(0.1,100)	3.858(5.207)	3.940(5.653)	4.003(5.933)	4.090(6.311)
(0.5,1)	1.791(0.574)	1.807(0.589)	1.821(0.607)	1.834(0.622)
(0.5,100)	3.366(3.264)	3.401(3.315)	3.431(3.380)	3.480(3.527)
(0.9,1)	1.617(0.508)	1.623(0.516)	1.631(0.524)	1.638(0.530)
(0.9,100)	3.099(4.515)	3.100(4.322)	3.129(4.487)	3.163(4.758)

3.8 Conclusion

The BBvFF(d, k) is different with the other two existing VFF algorithms in that the BBvFF(d, k) considers the error analysis. The steepest descent and the Gauss-Newton methods are originally from the optimisation theory, looking for the optimal value over a period of time. Now that both of them are applied to find the value of VFF sequentially at each time t, the application may not be a reasonable choice in some cases. Considering that the VFF algorithm is to improve the flexibility of the model, making it adaptive to the changes of the observations in time series, and enhance the performance in prediction of the model when new data point is observed, the BBvFF(d, k) may be more reasonable choice for sequential application of the VFF algorithm.

From the simulation study, the BBvFF(0.1,k) is found to be the best or at least a competitive choice for better forecasting rather than the other two existing VFF

algorithms of the SDvFF and the GNvFF and the BBvFF(1.96,k). In pairs trading, the spread is assumed to follow an AR(1) model. Thus, the BBvFF(0.1,k) is expected to improve the performance in forecasting of the model.

Chapter 4

Inference for Multi-categorical Time Series

4.1 Introduction

In this chapter, an online monitoring process is newly defined and proposed to monitor a process in real time. In the field of process control, there are two different approaches to monitoring **(a)** the statistical process control and **(b)** the automatic process control according to Joe Qin (2003) and Box and Kramer (1992). The former is originated from the parts industry while the latter is from the process industry. The statistical process control aims to achieve the highest possible mean or a fixed target value with the smaller possible variation for the measurement of the targets. For example, there are lower and upper control limits and warning lines from a normal distribution or a normal approximation to the reference distribution. The automatic process control focuses on the feedback control trying to adjust the process accordingly to the external and uncontrollable variables. Newly proposed online monitoring process is different from the automatic process control in that it does not try to adjust itself to the external and uncontrollable variables. The online monitoring process is also different from the monitoring process of the statistical process control such as the Shewart chart, the Page-Barnard CUSUM chart, and the Roberts EWMA chart. In the online monitoring process, categories are set up by thresholds, and the location of a target is monitored and confirmed by the poste-

rior probabilities for each category. A logistic transformation of the probabilities for the categories is represented with the evolution model. Thus, the posterior probabilities are not just the counted proportions of the responses in the corresponding categories. The probability vector itself is a vector of random variables.

This chapter derives the recursions and proposes the particle filter as inference for multi-categorical time series in the DGLM. This is novel extending the works by West et al. (1985) and Triantafyllopoulos (2009) to multi-categorical time series in the DGLM using multinomial distribution for the observation model. The particle filter, developed in this chapter for categorical time series data, is not necessarily restricted to the application of mean reversion considered in the thesis. Categorical time series appear frequently and the contribution of the proposed particle filter is general.

In the classical time series analysis, the stationarity is a tool to find the periods when the moments such as the mean, the variance, and the auto-covariance of the process do not change. This is achieved often by differencing the original time series. A common but usual question on finding the stationary of time series is how to difference and how many more times to difference the original time series. There also exist two types stationarity: strong stationarity and weak stationarity. The strong stationary process share the same joint probability distribution while the weak stationary process does mainly the same first two moments. However, it is hardly possible to find the strong stationarity of the process. "Perfect" stationarity of kind can be obtained from the only one data point, clearly having the constant moments. Thus, to avoid claiming that each data point is strongly stationary, there are several formal tests to find the stationarity of the process over a period of time such as Augmented Dickey-Fuller test. The time series analyst roughly decides the periods where this much closeness is enough to be declared as stationary so that the stationarity is agreed and accepted by test results. For all this, a question always remains in time series analysis on how closeness is enough to be stationary. A concept of mean-reversion by Triantafyllopoulos and Montana (2011) is nothing but the stationarity in the class of dynamic models. A reader may think that we can calculate the probabilities of B_t when it is estimated by a normal or t posterior distribution

as in chapter 3. However, this is based on the assumption of normality on the data. Such an assumption may be unrealistic, especially given that the assets and their spreads normally have fat tails in real. Thus, considering the assumption on the distribution of the assets and their spreads with fat tails, the idea of calculating the probabilities of B_t is not recommendable, at least as far as the decision of mean reversion is concerned. Still, of course, we use the normal model to estimate B_t, but, in chapter 4, we suppose that this may not be perfect. In fact, B_t does not depend much on normality, in the sense that the moments of $\hat{B}_t$ is obtained by the Kalman filter, which can be obtained assuming no distribution and having the Bayes liner optimality. There is another advantage of treating the data as categorical. We are now able not only to estimate the probabilities, but also to say something about the uncertainty around these estimates with the posterior variances of the probabilities.

As a general formulation, a multi-categorical response is considered, which can be represented as a multinomial distribution at each time t. An example of the online monitoring process is to track the movements of $|\hat{B}_t|$, which is discretized within several categories. For this, a multinomial distribution, a member of the exponential family of distributions, is adopted within the class of dynamic generalised linear model (DGLM). Thus, sequential estimation of parameters can be approximated only by the moments. In this chapter, we propose two approaches for sequential estimation: **(a)** adopting Bayes linear methods together with conjugate prior distributions we approximate the first two moments of the states, and **(b)** adopting sequential Monte Carlo methods which we provide sequential simulation of the states for. West et al. (1985) show how the approximation using the linear Bayesian method can be done in the class of DGLM, which is discussed earlier in Section 2.7.1. Triantafyllopoulos (2009) provides a critical discussion on the topic of online estimation. For **(a)**, we extend the approximate inference of DGLM, due to West et al. (1985), to responses of the multivariate exponential family of distributions. Using the ideas of the above paper, developed for binomial responses, we extend their approach to the multinomial distribution. This produces an approximation of the first two moments of the state distribution, based on Bayes linear methods. For **(b)**, we use the posterior of **(a)** in order to approximate the importance function in particle filters.

From Chapter 2, we know that the value of $|\hat{B}_t|$ determines whether the spread of a pair is in the mean-reversion or not at time t. The one-step ahead forecast of a spread is recursively obtained, but it is believed to be valid as long as the spread at t is detected in the mean-reversion. When $|\hat{B}_t| < 1$, the pair is said to be in mean-reversion. Otherwise, $|\hat{B}_t| \geq 1$ and the pair is in non-mean-reversion according to the conditions of mean-reversion in Section 2.4.3; see Triantafyllopoulos and Montana (2011). However, at any time t when $|\hat{B}_t| < 1$, how confident a trader can be is an issue. Considering that the detection is online, it also would be better if a trader can have more information on the pair. Thus, the behavior of $|\hat{B}_t|$ over the period of time is regarded as a process, and it is closely monitored in the online monitoring process at each time t. As a process, the dynamic variation in $|\hat{B}_t|$ such as the trend, seasonality, and cycle, if any, can be monitored and analyzed more in depth with the evolution model of DGLM.

Suppose that the probability of π_t is defined as the counted proportion of the data points in the mean-reversion from the spread time series available up to time t. At each time t, this probability would provide numerical information, revealing how many data points are in mean-reversion state up to t. In the binomial case where there are only two possible outcomes of 'Mean Reversion' or 'Non-Mean Reversion' at each time t, the less than half proportion for the mean-reversion may represent that the pair is not reliable for pairs trading. On the other hand, the increased proportion for the mean-reversion from $t-1$ to t would add more support of the pair for trading. The binomial case of either 'in mean-reversion' or 'not in mean-reversion' can be extended to the multiple cases by closely monitoring the behavior of $|\hat{B}_t|$. For example, with thresholds of 0.9 and 1.0 for $|\hat{B}_t|$, the three different categories can be assumed according to the value of $|\hat{B}_t|$. They can be set as follows: 'mean-reversion' for $0 \leq |\hat{B}_t| < 0.9$, 'semi-mean-reversion' for $0.9 \leq |\hat{B}_t| < 1$, and 'non-mean-reversion' for $1 \leq |\hat{B}_t|$. Over the period of time, each value of $|\hat{B}_t|$ belongs to one of three categories, and at each time t, a vector of the observation is made such that $\mathbf{X}_t = (x_{1,t}, x_{2,t}, x_{3,t})'$.

A statistical model for time series of multi-categorical, or polychotomous responses

can be categorized into a class of multivariate dynamic models. Suppose that there is a $(n+1)$-categorical time series, and a response vector of $\mathbf{X}_t = (x_{1,t}, \ldots, x_{n+1,t})'$ with $\mathbf{\Pi}_t = (\pi_{1,t}, \ldots, \pi_{n+1,t})'$ for each category. Assuming that $N_t = \sum_{i=1}^{n+1} x_{i,t}$ and $\sum_{i=1}^{n+1} \pi_{i,t} = 1$ at t, the $(n+1)^{th}$ category can be understood to have $x_{n+1,t} = N_t - \sum_{i=1}^{n} x_{i,t}$ with $\pi_{n+1,t} = 1 - \sum_{i=1}^{n} \pi_{i,t}$. Thus, a vector of responses at t can be described by $\mathbf{x}_t = (x_{1,t}, \ldots, x_{n,t})'$ with probabilities of $\boldsymbol{\pi}_t = (\pi_{1,t}, \ldots, \pi_{n,t})'$ for each category, having the joint probability function as

$$p(\mathbf{x}_t \mid \boldsymbol{\pi}_t) = \frac{N_t!}{x_{1,t}! \cdots x_{n,t}! (N_t - \sum_{i=1}^{n} x_{i,t})!} \pi_{1,t}^{x_{1,t}} \cdots \pi_{n,t}^{x_{n,t}} (1 - \sum_{i=1}^{n} \pi_{i,t})^{N_t - \sum_{i=1}^{n} x_{i,t}} \quad (4.1)$$

where $N_t = \sum_{i=1}^{n+1} x_{i,t}$ and $\sum_{i=1}^{n+1} \pi_{i,t} = 1$ satisfying $0 \leq \pi_{i,t} \leq 1$.

If only one response is observed for multiple categories at each time t, $N_t = 1$ and components of an observation vector $\mathbf{X}_t$ is determined by

$$x_{i,t} = \begin{cases} 1, & \text{if a category } i \text{ is observed} \\ 0, & \text{otherwise} \end{cases}$$

where $i = 1, 2, \ldots, n+1$. For example, when there are three categories and only one observation is available at t, $N_t = \sum_{i=1}^{3} x_{i,t} = 1$ and $\mathbf{X}_t$ becomes either one of the three possible vector such as $\mathbf{X}_t \in \{(1,0,0)', (0,1,0)', (0,0,1)'\}$, making $\mathbf{x}_t$ to be one of $\{(1,0)', (0,1)', (0,0)'\}$.

In this chapter, as the response vector $\mathbf{X}_t$ is observed from the values of $|\hat{B}_t|$ in the dynamic linear model for the spread time series Y_t, the DGLM with a multinomial distribution works together with the dynamic linear model for pairs trading. However, each of the two models is mathematically separate and applicable independently to any appropriate time series.

4.2 Model Specification

A DGLM is comprised of the observation model with a link function and the evolution model. For multi-categorical time series with $(n+1)$ categories, the online

monitoring process can be set up by the DGLM with a multinomial distribution where the response or observation vector is $\mathbf{X}_t = (x_{1,t}, \ldots, x_{n+1,t})'$ at time t. Thus, a continuous and monotonic link function $g(\cdot)$ becomes a logistic transformation of the vector of probabilities $\mathbf{\Pi}_t = (\pi_{1,t}, \ldots, \pi_{n+1,t})'$, mapping $\mathbf{\Pi}_t$ to the real line as $g(\mathbf{\Pi}_t) = \boldsymbol{\eta}_t$. The corresponding response probabilities $\mathbf{\Pi}_t = (\pi_{1,t}, \ldots, \pi_{n+1,t})'$ are specified by a dynamic multivariate logistic model so that $\pi_{i,t} = \frac{\mathrm{e}^{\eta_{i,t}}}{1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}}$ for $i = 1, \ldots, n$, leaving the $(n+1)^{th}$ component of $\mathbf{\Pi}_t$ as $\pi_{n+1,t} = \frac{1}{1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}}$. In the evolution model, the moments of the state vector $\boldsymbol{\theta}_t$ are shown to be approximated first using the conjugate analysis and the linear Bayesian method, and then using the particle filters as a simulation-based approach.

4.2.1 The Observation Model

For the multi-categorical time series $\mathbf{X}_t$ with $(n+1)$ categories, the observation model of an exponential family state space model or a DGLM can be described in the exponential family form as

$$
\begin{aligned}
p(\mathbf{x}_t \mid \boldsymbol{\eta}_t) &= \exp\left\{[\mathbf{x}_t'\boldsymbol{\eta}_t - a(\boldsymbol{\eta}_t)] + b(\mathbf{x}_t)\right\} \qquad (4.2)\\
\mathbf{x}_t &= (x_{1,t}, \ldots, x_{n,t})'\\
\boldsymbol{\eta}_t &= (\eta_{1,t}, \ldots, \eta_{n,t})'\\
&= \left(\log\left(\frac{\pi_{1,t}}{1-\sum_{i=1}^{n}\pi_{i,t}}\right), \cdots, \log\left(\frac{\pi_{n,t}}{1-\sum_{i=1}^{n}\pi_{i,t}}\right)\right)'\\
a(\boldsymbol{\eta}_t) &= N_t \cdot \log\left(1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}\right)\\
b(\mathbf{x}_t) &= \log\left(\frac{N_t!}{x_{1,t}!\cdots x_{n,t}!(N_t-\sum_{i=1}^{n}x_{i,t})!}\right)
\end{aligned}
$$

where $\boldsymbol{\eta}_t$ is the natural parameter.

$a(\boldsymbol{\eta}_t)$ is assumed to be twice differentiable in $\boldsymbol{\eta}_t$ such that

$$\begin{aligned}
\mathrm{E}(x_{i,t} \mid \eta_{i,t}) &= \frac{da(\boldsymbol{\eta}_t)}{d\eta_{i,t}} = \dot{a}(\eta_{i,t}) \\
&= N_t \cdot \frac{\mathrm{e}^{\eta_{i,t}}}{1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}} = N_t \cdot \pi_{i,t} \\
\mathrm{Var}(x_{i,t} \mid \eta_{i,t}) &= \frac{d\dot{a}(\eta_{i,t})}{d\eta_{i,t}} = \ddot{a}(\eta_{i,t}) \\
&= N_t \cdot \frac{d}{d\eta_{i,t}}\left(\frac{\mathrm{e}^{\eta_{i,t}}}{1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}}\right) \\
&= N_t \cdot \frac{\mathrm{e}^{\eta_{i,t}}\left(1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}\right) - \mathrm{e}^{\eta_{i,t}}\left(\mathrm{e}^{\eta_{i,t}}\right)}{\left(1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}\right)^2} \\
&= N_t \cdot \pi_{i,t} \cdot (1-\pi_{i,t})
\end{aligned}$$

where $\mathrm{e}^{\eta_{i,t}} = \frac{\pi_{i,t}}{1-\sum_{i=1}^{n}\pi_{i,t}}$ and $1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}} = \frac{1}{1-\sum_{i=1}^{n}\pi_{i,t}}$. Each of $\dot{a}(\eta_{i,t})$ and $\ddot{a}(\eta_{i,t})$ is termed as the mean function and the variance function of the distribution respectively.

The link function for the observation model is given by

$$g(\boldsymbol{\Pi}_t) = \boldsymbol{\eta}_t \simeq F_t'\boldsymbol{\theta}_t \tag{4.3}$$

so that $g(\cdot)$ maps $\boldsymbol{\Pi}_t$ to the linear predictor $\boldsymbol{\eta}_t$. The link function of the logit model is given by $g_{j,t}(\pi_{i,t},\ldots,\pi_{n+1,t}) = \log\left(\frac{\pi_{j,t}}{1-\sum_{i=1}^{n}\pi_{i,t}}\right)$ and the response function $h(\cdot)$, defined as the inverse function of $g(\cdot)$, becomes $h_{j,t}(\eta_{i,t},\ldots,\eta_{n,t}) = \frac{\mathrm{e}^{\eta_{j,t}}}{1+\sum_{i=1}^{n}\mathrm{e}^{\eta_{i,t}}}$ for $j = 1,2,\ldots,n$. F_t is a known $(n \times n)$ design matrix, and $\boldsymbol{\theta}_t$ is an n-dimensional state vector of $\boldsymbol{\theta}_t = (\theta_{1,t},\ldots,\theta_{n,t})'$ at t.

In (4.3), the relationship between $\boldsymbol{\eta}_t$ and $\boldsymbol{\theta}_t$ is represented as $\boldsymbol{\eta}_t \simeq F_t'\boldsymbol{\theta}_t$ as in West et al. (1985), indicating that there is no actual but a "guide" relationship by $\simeq$. This guide relationship is brought into to explain the recursions between $\boldsymbol{\eta}_t$ and $\boldsymbol{\theta}_t$ for any F_t' in general, justifying the application of the linear Bayesian method to obtain the optimal estimate for conditional moments of $\mathrm{E}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1})$ and $\mathrm{Var}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1})$.

4.2.2 The Evolution Model

The evolution model of the state vector $\boldsymbol{\theta}_t$ is

$$\boldsymbol{\theta}_t \;=\; G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathbf{MVN}[0, W_t] \tag{4.4}$$

where G_t is a known $(n \times n)$ transition matrix, $\boldsymbol{\epsilon}_t$ is an n-dimensional vector of evolution errors, and W_t is a known $(n \times n)$ covariance matrix. For the evolution errors of $\boldsymbol{\epsilon}_t$, distributional assumption is not necessary, and the zero mean assumption may be relaxed. However, in this chapter, the evolution errors are assumed to be a white noise sequence with multivariate normal distribution and uncorrelated over time. Conditional on $\boldsymbol{\eta}_t$, $\mathbf{x}_t$ is assumed to be independent of $\boldsymbol{\epsilon}_t$. When $\boldsymbol{\theta}_t$ is time-invariant, $G_t = I$ and $W_t = 0$, the model is reduced to a generalized linear model.

(4.2), (4.3), and (4.4) define the DGLM for the time series $\mathbf{X}_t$ with polychotomous responses at t. In this chapter, the simplest choice for $\boldsymbol{\theta}_t$ would be the first order random walk model of $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \sim \mathbf{MVN}[0, W_t]$, having $G_t = I$.

4.2.3 Recursions of Parameter Estimates

Although F_t and G_t are assumed to be known as an identity matrix for an application to pairs trading in this chapter, the derivation of the recursions is made for the general formulation of a model with F_t and G_t.

With the model definitions of a DGLM, $\mathbf{r}_0^*$, $\mathbf{m}_0$ and C_0 are decided initially by the modeller when $(\boldsymbol{\Pi}_0 \mid D_0) \sim \text{Dirichlet}(\mathbf{r}_0^*)$ and $(\boldsymbol{\theta}_0 \mid D_0) \sim [\mathbf{m}_0, C_0]$. In the dynamic linear model, the prior and posterior distributions for the state vector $\boldsymbol{\theta}_t$ is assumed to be the normal distribution, but in the DGLM, they are meant to be any exponential family form of distributions. When the recursions are not available to get the posterior distribution for the state vector $\boldsymbol{\theta}_t$, the moments are approximated. In this chapter, as the first approach, the moments of the posterior distribution for the state vector $\boldsymbol{\theta}_t$ are approximated using the conjugate analysis and the linear Bayes method. As the second approach, a simulation-based methodology of the particle

filters is employed to adequately summarise such distributions with the moments.

While applying the particle filters to a DGLM with a multinomial distribution, the required distributions are partially achieved in terms of the first and the second moments. When choosing the importance transition density in the sequential Monte Carlo methods, the easiest choice at t would be the prior distribution of the states $\boldsymbol{\theta}_t$ and it is called as the bootstrap filter. However, with consideration on information from the newly observed, the online estimation can be improved. Thus, the optimal importance kernel is chosen to be the importance transition density in the particle filter, where its moments are obtained also by the linear Bayesian method. Both the bootstrap filter and the particle filter are the sequential Monte Carlo methods, also know as the particle filters, but the only difference lies in the choice of the importance transition density from which the particles are simulated.

4.2.3.1 Prior Distributions

The prior distributions are approximated by the moments only. Given that $(\boldsymbol{\theta}_{t-1} \mid D_{t-1}) \sim [\mathbf{m}_{t-1}, C_{t-1}]$, the recursions for the prior distributions at t are achieved by

(f1) $(\boldsymbol{\theta}_t \mid D_{t-1}) \sim [\mathbf{a}_t, R_t]$
where $\mathbf{a}_t = G_t \mathbf{m}_{t-1}$ and $R_t = G_t C_{t-1} G_t' + W_t$

(f2) $(\boldsymbol{\eta}_t \mid D_{t-1}) \sim [\mathbf{f}_t, Q_t]$
where $\mathbf{f}_t = F_t' \mathbf{a}_t$ and $Q_t = F_t' R_t F_t$

(f3) $(\boldsymbol{\Pi}_t \mid D_{t-1}) \sim \text{Dirichlet}(\mathbf{r}_t)$
where $r_{i,t} = \frac{1+e^{f_{i,t}}}{Q_{ii,t}}$ for $i = 1, 2, \ldots, n$ and $r_{n+1,t} = \frac{\sum_{i=1}^{n} r_{i,t}}{\sum_{i=1}^{n} e^{f_{i,t}}}$

The derivation of updating equations in **(f1)** and **(f2)** parallel the normal theory which can be found in Appendix B as shown for the dynamic linear model in Chapter 2 and 3, but with no full distributional assumptions, $(\boldsymbol{\theta}_t \mid D_{t-1})$ and $(\boldsymbol{\eta}_t \mid D_{t-1})$ are

partially specified in terms of the moments. **(f1)**, for example, is achieved by

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\theta}_t \mid D_{t-1}) &= \mathrm{E}(G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \mid D_{t-1}) \;=\; G_t\mathrm{E}(\boldsymbol{\theta}_{t-1} \mid D_{t-1}) \\
&= G_t\mathbf{m}_{t-1} \;=\; \mathbf{a}_t \\
\mathrm{Var}(\boldsymbol{\theta}_t \mid D_{t-1}) &= \mathrm{Var}(G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \mid D_{t-1}) \\
&= G_t\mathrm{Var}(\boldsymbol{\theta}_{t-1} \mid D_{t-1})G_t' + \mathrm{Var}(\boldsymbol{\epsilon}_t \mid D_{t-1}) \\
&= G_tC_{t-1}G_t' + W_t \;=\; (G_tC_{t-1}G_t' + W_t) \\
&= R_t
\end{aligned}$$

The derivation of **(f2)** can be seen in Section 4.3.1. A vector of parameters $\mathbf{r}_t$ in **(f3)** for $(\boldsymbol{\Pi}_t \mid D_{t-1})$ are achieved by the moments of $(\boldsymbol{\eta}_t \mid D_{t-1})$ where $\boldsymbol{\Pi}_t$ and $\boldsymbol{\eta}_t$ are linked via a multivariate logistic transformation. The details are shown in (4.14) and (4.15) of Section 4.3.2.5 on how a vector of parameters $\mathbf{r}_t$ for $(\boldsymbol{\Pi}_t \mid D_{t-1})$ are achieved by the moments of $(\boldsymbol{\eta}_t \mid D_{t-1})$.

4.2.3.2 Posterior Distributions

When an observation $\mathbf{X}_t$ is made at time t, the parameters of the Dirichlet distribution are updated from $(\boldsymbol{\Pi}_t \mid D_{t-1}) \sim \mathrm{Dirichlet}(\mathbf{r}_t)$ to $(\boldsymbol{\Pi}_t \mid D_t) \sim \mathrm{Dirichlet}(\mathbf{r}_t^*)$ where $\mathbf{r}_t^* = (r_{1,t}^*, \ldots, r_{n+1,t}^*)$ and $r_{i,t}^* = \mathbf{r}_t + \mathbf{x}_t$ for $i = 1, 2, \ldots, n$ and $r_{n+1,t}^* = r_{n+1,t} + N_t - \sum_{i=1}^n x_{i,t}$.

(g1) $(\boldsymbol{\Pi}_t \mid D_t) \sim \mathrm{Dirichlet}(\mathbf{r}_t^*)$
$r_{i,t}^* = r_{i,t} + x_{i,t}$ and $r_{n+1,t}^* = r_{n+1,t} + N_t - \sum_{i=1}^n x_{i,t}$
where $\mathbf{X}_t = (x_{1,t}, x_{2,t}, \ldots, x_{n+1,t})'$ and $\mathbf{r}_t^* = (r_{1,t}, r_{2,t}, \ldots, r_{n+1,t})$

(g2) $(\boldsymbol{\eta}_t \mid D_t) \sim [\mathbf{f}_t^*, Q_t^*]$
$f_{i,t}^* \approx \log(r_{i,t} + x_{i,t}) - \log(r_{n+1,t} + N_t - \sum_{i=1}^n x_{i,t})$ where $\mathbf{f}_t^* = (f_{1,t}^*, \ldots, f_{n,t}^*)'$,
$Q_{ii,t}^* \approx \frac{1}{r_{i,t}+x_{i,t}} + \frac{1}{r_{n+1,t}+N_t-\sum_{i=1}^n x_{i,t}}$ and $Q_{ij,t}^* \approx \left(\frac{1}{r_{n+1,t}+N_t-\sum_{i=1}^n x_{i,t}}\right)$ for $i \neq j$
where $Q_{i,j}$ represents (i,j) entries of an $(n \times n)$ matrix Q_t

(g3) $(\boldsymbol{\theta}_t \mid D_t) \sim [\mathbf{m}_t, C_t]$

As new observation $\mathbf{X}_t$ is made, the parameters of the Dirichlet distribution for $\boldsymbol{\Pi}_t$ are updated from $\mathbf{r}_t$ for $(\boldsymbol{\Pi}_t \mid D_{t-1})$ to $\mathbf{r}_t^* = \mathbf{r}_t + \mathbf{X}_t$ for $(\boldsymbol{\Pi}_t \mid D_t)$ as given in **(g1)**.

By these updated parameters of $\mathbf{r}_t^*$, the moments of the posterior distribution for $(\boldsymbol{\eta}_t \mid D_t)$ are approximately obtained as given in **(g2)**, the approximation details of which can be found in Section 4.3.3. The distribution of $(\boldsymbol{\theta}_t \mid D_t)$ in **(g3)** is analytically intractable so that the first and the second moments are approximated using the linear Bayesian method in Section 4.4 and using the particle filters in Section 4.5. The estimation methods make use of conjugate prior distributions and of Bayes linear methods, which are combined with particle filters. In Section 4.3, we give all the necessary details of the conjugate prior distributions together with derivations of moments of the linear predictor. In Section 4.4, we develop an inference that uses Bayes linear methods as the main approximation or as a guide to approximate the optimal importance kernel in particle filtering.

4.3 Recursive Updating for $\boldsymbol{\Pi}_t$ and $\boldsymbol{\eta}_t$

Parameters of $(\boldsymbol{\Pi}_t \mid D_{t-1})$ are determined by the moments of $(\boldsymbol{\eta}_t \mid D_{t-1})$. $(\boldsymbol{\Pi}_t \mid D_{t-1})$ are updated to $(\boldsymbol{\Pi}_t \mid D_t)$ with new vector of observations $\mathbf{X}_t$ at time t. The moments of $(\boldsymbol{\eta}_t \mid D_t)$ are approximately obtained from the parameters of $(\boldsymbol{\Pi}_t \mid D_t)$. These recursive updating is explored and shown in this section.

4.3.1 Moments of $(\boldsymbol{\eta}_t \mid D_{t-1})$

(f2) from Section 4.2.3.1 can be derived as follows.

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{\eta}_t \mid D_{t-1}) &= \mathrm{E}(F_t'\boldsymbol{\theta}_t \mid D_{t-1}) = \mathrm{E}(F_t'G_t\boldsymbol{\theta}_{t-1} + F_t'\boldsymbol{\epsilon}_t \mid D_{t-1}) \\
&= F_t'G_t\mathrm{E}(\boldsymbol{\theta}_{t-1} \mid D_{t-1}) + F_t'\mathrm{E}(\boldsymbol{\epsilon}_t \mid D_{t-1}) \\
&= F_t'G_t\mathbf{m}_{t-1} = F_t'\mathbf{a}_t = \mathbf{f}_t \\
\mathrm{Var}(\boldsymbol{\eta}_t \mid D_{t-1}) &= \mathrm{Var}(F_t'\boldsymbol{\theta}_t \mid D_{t-1}) = \mathrm{Var}(F_t'G_t\boldsymbol{\theta}_{t-1} + F_t'\boldsymbol{\epsilon}_t \mid D_{t-1}) \\
&= F_t'G_t\mathrm{Var}(\boldsymbol{\theta}_{t-1} \mid D_{t-1})G_t'F_t + F_t'\mathrm{Var}(\boldsymbol{\epsilon}_t \mid D_{t-1})F_t \\
&= F_t'G_tC_{t-1}G_t'F_t + F_t'W_tF_t \\
&= F_t'(G_tC_{t-1}G_t' + W_t)F_t \\
&= F_t'R_tF_t \\
&= Q_t
\end{aligned}
$$

where $R_t = G_t C_{t-1} G_t' + W_t$.

4.3.2 Relationship between $(\boldsymbol{\eta}_t \mid D_{t-1})$ and $(\boldsymbol{\Pi}_t \mid D_{t-1})$

While $\boldsymbol{\Pi}_t$ follows a Dirichlet distribution, $\boldsymbol{\eta}_t$ is related to $\boldsymbol{\Pi}_t$ via a link function $g(\cdot)$ and by a logistic variable transformation of

$$\boldsymbol{\eta}_t = (\eta_{1,t}, \ldots, \eta_{n,t})' = \left(\log\left(\frac{\pi_{1,t}}{1 - \sum_{i=1}^{n} \pi_{i,t}} \right), \cdots, \log\left(\frac{\pi_{n,t}}{1 - \sum_{i=1}^{n} \pi_{i,t}} \right) \right)'$$

By investigating the relationship between $(\boldsymbol{\Pi}_t \mid D_{t-1})$ and $(\boldsymbol{\eta}_t \mid D_{t-1})$, the details of **(f3)** are shown on how the parameters of $(\boldsymbol{\Pi}_t \mid D_{t-1})$ are determined by the first and second moments of $(\boldsymbol{\eta}_t \mid D_{t-1})$. For a random vector $\boldsymbol{\Pi}_t = (\pi_{1,t}, \ldots, \pi_{n+1,t})'$, $(\boldsymbol{\Pi}_t \mid D_{t-1}) \sim \text{Dirichlet}(\mathbf{r}_t)$ is assumed where $\mathbf{r}_t = (r_{1,t}, \ldots, r_{n+1,t})'$. However, $\mathbf{r}_t = (r_{1,t}, \ldots, r_{n+1,t})'$ is unknown. This unknown parameter vector of $\mathbf{r}_t$ can be obtained from the relationship between $(\boldsymbol{\Pi}_t \mid D_{t-1})$ and $(\boldsymbol{\eta}_t \mid D_{t-1})$.

4.3.2.1 The Density Function of $(\boldsymbol{\eta}_t \mid D_{t-1})$

Suppose that $\boldsymbol{\Pi}_t = (\pi_{1,t}, \ldots, \pi_{n+1,t})'$ is a random vector which follows the Dirichlet distribution such as $(\boldsymbol{\Pi}_t | D_{t-1}) \sim \text{Dirichlet}(\mathbf{r}_t)$ where $\mathbf{r}_t = (r_{1,t}, \ldots, r_{n+1,t})'$ is a parameter vector.

The joint density function of $(\boldsymbol{\Pi}_t \mid D_{t-1})$ is

$$p(\boldsymbol{\Pi}_t \mid D_{t-1}) = \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1} \Gamma(r_{i,t})} \pi_{1,t}^{r_{1,t}-1} \cdots \pi_{n,t}^{r_{n,t}-1} \left(1 - \sum_{i=1}^{n} \pi_{i,t} \right)^{r_{n+1,t}-1}$$

where $0 \leq \pi_{1,t}, \ldots, \pi_{n+1,t} \leq 1$ and $\pi_{n+1,t} = 1 - \sum_{i=1}^{n} \pi_{i,t}$.

By a logistic variable transformation,
$\boldsymbol{\eta}_t = (\eta_{1,t}, \ldots, \eta_{n,t})' = \left(\log\left(\frac{\pi_{1,t}}{1-\sum_{i=1}^{n} \pi_{i,t}} \right), \ldots, \log\left(\frac{\pi_{n,t}}{1-\sum_{i=1}^{n} \pi_{i,t}} \right) \right)'$, which is equally likely to be $\boldsymbol{\pi}_t = (\pi_{1,t}, \ldots, \pi_{n,t})' = \left(\frac{e^{\eta_{1,t}}}{1+\sum_{i=1}^{n} e^{\eta_{i,t}}}, \frac{e^{\eta_{2,t}}}{1+\sum_{i=1}^{n} e^{\eta_{i,t}}}, \ldots, \frac{e^{\eta_{n,t}}}{1+\sum_{i=1}^{n} e^{\eta_{i,t}}} \right)'$.

Thus, by the variable transformation, the prior distribution at t, or the density

function of $(\boldsymbol{\eta}_t \mid D_{t-1})$ is

$$\begin{aligned} p(\boldsymbol{\eta}_t \mid D_{t-1}) &= \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1}\Gamma(r_{i,t})} \frac{e^{\sum_{i=1}^{n}\eta_{i,t}(r_{i,t}-1)}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{\sum_{i=1}^{n+1}(r_{i,t}-1)}} \cdot |J| \\ &= \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1}\Gamma(r_{i,t})} \frac{e^{\sum_{i=1}^{n} r_{i,t}\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{\sum_{i=1}^{n+1} r_{i,t}}} \end{aligned} \tag{4.5}$$

where

$$|J| = \frac{\prod_{i=1}^{n} e^{\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{n+1}} = \frac{e^{\sum_{i=1}^{n}\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{n+1}} \tag{4.6}$$

The computation of the Jacobian $|J|$ in (4.6) is shown in Appendix C.

4.3.2.2 Generating Functions for $\boldsymbol{\eta}_t$

For a real valued vector $\mathbf{z} = (z_1, \ldots, z_n)'$, the moment generating function of $\boldsymbol{\eta}_t$ is

$$\begin{aligned} M_{\boldsymbol{\eta}_t}(\mathbf{z}) &= E(e^{\mathbf{z}'\boldsymbol{\eta}_t}) = \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1}\Gamma(r_{i,t})} \int_{\mathbb{R}^n} \cdots \int \frac{e^{\sum_{i=1}^{n}(r_{i,t}+z_i)\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{\sum_{i=1}^{n+1} r_{i,t}}} d\eta_{1,t}\cdots d\eta_{n,t} \\ &= \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1}\Gamma(r_{i,t})} \int_{\mathbb{R}^n} \cdots \int \frac{e^{\sum_{i=1}^{n}(r_{i,t}+z_i)\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{\sum_{i=1}^{n}(r_{i,t}+z_i)+(r_{n+1,t}-\sum_{i=1}^{n} z_i)}} d\eta_{1,t}\cdots d\eta_{n,t} \\ &= \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1}\Gamma(r_{i,t})} \frac{\prod_{i=1}^{n}\Gamma(r_{i,t}+z_i)\Gamma(r_{n+1,t}-\sum_{i=1}^{n} z_i)}{\Gamma(\sum_{i=1}^{n}(r_{i,t}+z_i)+(r_{n+1,t}-\sum_{i=1}^{n} z_i))} \\ &= \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1}\Gamma(r_{i,t})} \frac{\prod_{i=1}^{n}\Gamma(h_{i,t})\Gamma(h_{n+1,t})}{\Gamma(\sum_{i=1}^{n+1} h_{i,t})} \end{aligned} \tag{4.7}$$

where $h_{i,t} = r_{i,t} + z_i$ for $i = 1, \ldots, n$ and $h_{n+1,t} = r_{n+1,t} - \sum_{i=1}^{n} z_i$ and (4.7) is justified by noting the followings.

Firstly, by taking $h_{i,t} = r_{i,t} + z_i$ for $i = 1, \ldots, n$ and $h_{n+1,t} = r_{n+1,t} - \sum_{i=1}^{n} z_i$,

$$\sum_{i=1}^{n+1} r_{i,t} = \sum_{i=1}^{n+1} h_{i,t} = \sum_{i=1}^{n} h_{i,t} + h_{n+1,t} = \sum_{i=1}^{n}(r_{i,t}+z_i) + (r_{n+1,t} - \sum_{i=1}^{n} z_i)$$

Secondly, from the fact that

$$\int p(\boldsymbol{\eta}_t)d\boldsymbol{\eta}_t = \int_{\mathbb{R}^n} \cdots \int \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1}\Gamma(r_{i,t})} \frac{e^{\sum_{i=1}^{n} r_{i,t}\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{\sum_{i=1}^{n+1} r_{i,t}}} d\eta_{1,t}\cdots d\eta_{n,t} = 1$$

we obtain

$$\int_{\mathbb{R}^n} \cdots \int \frac{e^{\sum_{i=1}^{n} r_{i,t}\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{\sum_{i=1}^{n+1} r_{i,t}}} d\eta_{1,t}\cdots d\eta_{n,t} = \frac{\prod_{i=1}^{n+1}\Gamma(r_{i,t})}{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}$$

By definition, the cumulant generating function for $\boldsymbol{\eta}_t$ is equal to

$$\begin{aligned} K_{\boldsymbol{\eta}_t}(\mathbf{z}) &= \log\left\{M_{\boldsymbol{\eta}_t}(\mathbf{z})\right\} \\ &= \log\left\{\Gamma(\sum_{i=1}^{n+1} r_{i,t})\right\} - \log\left\{\prod_{i=1}^{n+1}\Gamma(r_{i,t})\right\} \\ &\quad + \log\left\{\prod_{i=1}^{n}\Gamma(r_{i,t}+z_i)\Gamma(r_{n+1,t}-\sum_{i=1}^{n} z_i)\right\} \\ &\quad -\log\left\{\Gamma(\sum_{i=1}^{n}(r_{i,t}+z_i)+(r_{n+1,t}-\sum_{i=1}^{n} z_i))\right\} \end{aligned}$$

The partial derivatives of the cumulant generating function in the above are

$$\begin{aligned} \frac{\partial K_{\boldsymbol{\eta}_t}(\mathbf{z})}{\partial z_i} &= \Psi(r_{i,t}+z_i) - \Psi(r_{n+1,t}-\sum_{i=1}^{n} z_i) \\ &= \Psi(h_{i,t}) - \Psi(h_{n+1,t}), \quad i=1,\ldots,n \qquad (4.8) \\ \frac{\partial^2 K_{\boldsymbol{\eta}_t}(\mathbf{z})}{\partial z_i^2} &= \Psi^{(1)}(r_{i,t}+z_i) + \Psi^{(1)}(r_{n+1,t}-\sum_{i=1}^{n} z_i) \\ &= \Psi^{(1)}(h_{i,t}) + \Psi^{(1)}(h_{n+1,t}), \quad i=1,\ldots,n \qquad (4.9) \\ \frac{\partial^2 K_{\boldsymbol{\eta}_t}(\mathbf{z})}{\partial z_j \partial z_i} &= \frac{\partial\{\Psi(r_{i,t}+z_i) - \Psi(r_{n+1,t}-\sum_{i=1}^{n} z_i)\}}{\partial z_j} = \Psi^{(1)}(r_{n+1,t}-\sum_{i=1}^{n} z_i) \\ &= \Psi^{(1)}(h_{n+1,t}) \qquad (4.10) \end{aligned}$$

where $\Psi(\cdot)$ and $\Psi^{(1)}(\cdot)$ denote the digamma function and the trigamma function respectively as in Abramowitz and Stegun (1965). The details on generating functions,

digamma and trigamma functions are found in Appendix A.

4.3.3 Parameters of $(\boldsymbol{\Pi}_t \mid D_{t-1})$

With $\Psi(z) \approx \log(z)$ and $\Psi^{(1)}(z) \approx \frac{1}{z}$, the moments of $(\boldsymbol{\eta}_t \mid D_{t-1})$ can be represented as

$$\begin{aligned}
\mathrm{E}(\eta_{i,t} \mid D_{t-1}) = f_{i,t} &= \Psi(r_{i,t}) - \Psi(r_{n+1,t}) \\
&\approx \log(r_{i,t}) - \log(r_{n+1,t}) \\
&= \log\left(\frac{r_{i,t}}{r_{n+1,t}}\right) \qquad (4.11) \\
\mathrm{Var}(\eta_{i,t} \mid D_{t-1}) = Q_{ii,t} &= \Psi^{(1)}(r_{i,t}) + \Psi^{(1)}(r_{n+1,t}) \\
&\approx \frac{1}{r_{i,t}} + \frac{1}{r_{n+1,t}} \qquad (4.12) \\
\mathrm{Cov}(\eta_{i,t}, \eta_{j,t} \mid D_{t-1}) = Q_{ij,t} &= \Psi^{(1)}(r_{n+1,t}) \\
&\approx \frac{1}{r_{n+1,t}} \qquad \text{for} \quad i \neq j \qquad (4.13)
\end{aligned}$$

where $\mathbf{f}_t = (f_{1,t}, \ldots, f_{n,t})'$, $Q_{ii,t}$ are diagonal elements and $Q_{ij,t}$ for $i \neq j$ are non-diagonal elements of the covariance matrix, Q_t.

From (4.11), we know that $e^{f_{i,t}} = \frac{r_{i,t}}{r_{n+1,t}}$. By taking $\sum$ for both sides, (4.11) reduces to $\sum_{i=1}^{n} e^{f_{i,t}} = \frac{\sum_{i=1}^{n} r_{i,t}}{r_{n+1,t}}$. Thus, $r_{n+1,t}$ can be written as

$$r_{n+1,t} = \frac{\sum_{i=1}^{n} r_{i,t}}{\sum_{i=1}^{n} e^{f_{i,t}}} \qquad (4.14)$$

On the other hand, rewriting (4.11) with regard to $r_{n+1,t}$ gives $r_{n+1,t} = \frac{r_{i,t}}{e^{f_{i,t}}}$, and by substituting it into (4.12), $r_{i,t}$ is obtained as

$$r_{i,t} = \frac{1 + e^{f_{i,t}}}{Q_{ii,t}}, \qquad \text{for } i = 1, 2, \ldots, n \qquad (4.15)$$

4.3.4 Moments of $(\boldsymbol{\eta}_t \mid D_t)$

The posterior distribution of $\mathbf{\Pi}_t$ given D_t follows the Dirichlet$(\mathbf{r}_t + \mathbf{X}_t)$ with an observation $\mathbf{X}_t$ at time t. When we derive formula for $\mathbf{r}_t$ in Section 4.3.3, $\Psi(z)$ and $\Psi^{(1)}(z)$ are approximated by $\log(z)$ and $\frac{1}{z}$ for computational simplicity. While computing the moments of $(\boldsymbol{\eta}_t \mid D_t)$, we use more terms up to the second term for approximations, which are $\log(z) - \frac{1}{2z}$ and $\frac{1}{z}\left(1 + \frac{1}{2z}\right)$ for $\Psi(z)$ and $\Psi^{(1)}(z)$. More details on a digamma and a trigamma functions can be found in Abramowitz and Stegun (1965) and in Appendix A.2 of this thesis.

While $\mathbf{f}_t$ and $\mathbf{a}_t$ are obtained from the recursions of $\mathbf{f}_t = F_t'\mathbf{a}_t$ and $\mathbf{a}_t = G_t\mathbf{m}_{t-1}$, a vector of $\mathbf{r}_t$ for the prior distribution of $(\mathbf{\Pi}_t \mid D_{t-1})$ has to be computed from $\mathbf{f}_t$ and Q_t. A vector of $\mathbf{r}_t^*$ for the posterior distribution of $(\mathbf{\Pi}_t \mid D_t)$ is obtained by conjugacy of a Dirichlet distribution when new observation $\mathbf{X}_t$ is made at t.

The posterior moments of $(\boldsymbol{\eta}_t \mid D_t)$ can be found by the approximations as follows.

$$
\begin{aligned}
\mathrm{E}(\eta_{i,t} \mid D_t) = f_{i,t}^* &\approx \Psi(r_{i,t} + x_{i,t}) - \Psi(r_{n+1,t} + x_{n+1,t}) \quad \text{for } i = 1, 2, \ldots, n \\
&= \left\{ \log(r_{i,t} + x_{i,t}) - \frac{1}{2(r_{i,t} + x_{i,t})} \right\} \\
&\qquad - \left\{ \log\left(r_{n+1,t} + x_{n+1,t}\right) - \frac{1}{2(r_{n+1,t} + x_{n+1,t})} \right\} \\
&= \log\left(\frac{r_{i,t} + x_{i,t}}{r_{n+1,t} + x_{n+1,t}} \right) \\
&\qquad - \left\{ \frac{(r_{i,t} + x_{i,t}) + (r_{n+1,t} + x_{n+1,t})}{2(r_{i,t} + x_{i,t})(r_{n+1,t} + x_{n+1,t})} \right\}
\end{aligned} \tag{4.16}
$$

$$
\begin{aligned}
\mathrm{Var}(\eta_{i,t} \mid D_t) = Q^*_{ii,t} &\approx \Psi^{(1)}(r_{i,t} + x_{i,t}) + \Psi^{(1)}(r_{n+1,t} + x_{n+1,t}) \\
&= \left(\frac{1}{r_{i,t} + x_{i,t}}\right)\left(1 + \frac{1}{2(r_{i,t} + x_{i,t})}\right) \\
&\quad + \left(\frac{1}{r_{n+1,t} + x_{n+1,t}}\right)\left(1 + \frac{1}{2(r_{n+1,t} + x_{n+1,t})}\right) \\
&= \left\{\frac{(r_{i,t} + x_{i,t}) + (r_{n+1,t} + x_{n+1,t})}{(r_{i,t} + x_{i,t})(r_{n+1,t} + x_{n+1,t})}\right\} \\
&\quad + \frac{1}{2}\left\{\frac{(r_{i,t} + x_{i,t})^2 + (r_{n+1,t} + x_{n+1,t})^2}{(r_{i,t} + x_{i,t})^2(r_{n+1,t} + x_{n+1,t})^2}\right\} \qquad (4.17)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(\eta_{i,t}, \eta_{j,t} \mid D_t) = Q^*_{ij,t} &\approx \Psi^{(1)}(r_{n+1,t} + x_{n+1,t}) \quad \text{for } i \neq j \\
&= \left(\frac{1}{r_{n+1,t} + x_{n+1,t}}\right)\left(1 + \frac{1}{2(r_{n+1,t} + x_{n+1,t})}\right) \\
&= \frac{1}{2}\left\{\frac{2(r_{n+1,t} + x_{n+1,t}) + 1}{(r_{n+1,t} + x_{n+1,t})^2}\right\} \qquad (4.18)
\end{aligned}
$$

4.4 Inference for The Posterior of $(\boldsymbol{\theta}_t \mid D_t)$

In this section, two approaches are introduced for inference of $(\boldsymbol{\theta}_t \mid D_t)$. Following the approach by Hartigan (1969), Goldstein (1976), West et al. (1985), and Triantafyllopoulos (2009), the Bayes linear methods are applied to DGLM for multi-categorical time series for the approximate inference. As a simulation-based approach, sequential Monte Carlo methods are also applied for the approximations.

4.4.1 Approximate Inference by the Bayes Linear Methods

With the model specification of DGLM in Section 4.2, the first- and second- moments of the posterior of $(\boldsymbol{\theta}_t \mid D_t)$ can be approximated using the moments of the joint posterior of $\boldsymbol{\eta}_t$ and $\boldsymbol{\theta}_t$. To the following, we generalise the utility of Bayes linear methods for univariate DGLMs by West et al. (1985) to the multivariate case. In West and Harrison (1997), the linear Bayes estimate $\mathbf{m}_t$ is known to minimise the associated risk matrix C_t.

Assuming the posterior for the state vector at $t-1$ as

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim [\mathbf{m}_{t-1}, C_{t-1}]$$

the joint prior distribution of $\boldsymbol{\eta}_t$ and $\boldsymbol{\theta}_t$ at time t is partially specified by the first two moments only, and it follows that

$$\begin{pmatrix} \boldsymbol{\eta}_t \\ \boldsymbol{\theta}_t \end{pmatrix} \Bigg| \; D_{t-1} \Bigg) \sim \left[\begin{pmatrix} \mathbf{f}_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} Q_t & F_t' R_t \\ R_t F_t & R_t \end{pmatrix} \right] \tag{4.19}$$

where $\mathbf{f}_t = F_t' \mathbf{a}_t$, $Q_t = F_t' R_t F_t$, $\mathbf{a}_t = G_t \mathbf{m}_{t-1}$, and $R_t = G_t C_{t-1} G_t' + W_t$ are approximately the mean and variance of $(\boldsymbol{\theta}_t \mid D_{t-1})$.

The posterior mean vector and covariance matrix of $\boldsymbol{\eta}_t$ are approximated by linear Bayesian methods and by using the tower property of expectations as

$$\mathrm{E}(\boldsymbol{\eta}_t \mid D_t) = \mathrm{E}(g(\boldsymbol{\Pi}_t) \mid D_t) = \mathbf{f}_t^* \text{ and } \mathrm{Var}(\boldsymbol{\eta}_t \mid D_t) = \mathrm{Var}(g(\boldsymbol{\Pi}_t) \mid D_t) = Q_t^* \tag{4.20}$$

The mean vector and covariance matrix of $(\boldsymbol{\theta}_t \mid D_t)$ are approximated as

$$(\boldsymbol{\theta}_t \mid D_t) \sim [\mathbf{m}_t, C_t] \tag{4.21}$$

where $\mathbf{m}_t = \mathbf{a}_t + R_t F_t Q_t^{-1}(\mathbf{f}_t^* - \mathbf{f}_t)$ and $C_t = R_t - R_t F_t Q_t^{-1}(I - Q_t^* Q_t^{-1}) F_t' R_t$. The details of the approximation in (4.21) are given below,

$$\begin{aligned} p(\boldsymbol{\eta}_t, \boldsymbol{\theta}_t \mid D_t) &\propto p(\boldsymbol{\eta}_t, \boldsymbol{\theta}_t \mid D_{t-1}) p(\mathbf{Y}_t \mid \boldsymbol{\eta}_t) \\ &\propto \{p(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) p(\boldsymbol{\eta}_t \mid D_{t-1})\} p(\mathbf{Y}_t \mid \boldsymbol{\eta}_t) \\ &\propto p(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) \{p(\boldsymbol{\eta}_t \mid D_{t-1}) p(\mathbf{Y}_t \mid \boldsymbol{\eta}_t)\} \\ &\propto p(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) p(\boldsymbol{\eta}_t \mid D_t) \end{aligned} \tag{4.22}$$

From (4.22), we can see that $\boldsymbol{\theta}_t$ is conditionally independent of $\mathbf{Y}_t$ given $\boldsymbol{\eta}_t$ and D_{t-1}, and it follows that

$$p(\boldsymbol{\theta}_t \mid D_t) = \int p(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) p(\boldsymbol{\eta}_t \mid D_t) d\eta_t \tag{4.23}$$

$p(\boldsymbol{\eta}_t \mid D_t)$, the second component of the integrand in (4.23), can be obtained directly from (4.20) in the conjugate form posterior for $\boldsymbol{\eta}_t$. However, due to the incomplete of the joint prior distribution in (4.19), conditional moments of $p(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1})$ are unknown and non-linear functions of $\boldsymbol{\eta}_t$. Given only the partial moments, the posterior mean and variance matrix of $\boldsymbol{\theta}_t$ can be estimated using the linear Bayesian method by Goldstein and Wooff (2007).

Conditional moments of $\mathrm{E}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1})$ and $\mathrm{Var}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1})$ are obtained as the optimal estimate by the linear Bayesian method. For all $\boldsymbol{\eta}_t$, they are

$$\begin{aligned} \hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) &= \mathbf{a}_t + R_t F_t Q_t^{-1}(\boldsymbol{\eta}_t - f_t) && (4.24) \\ \hat{\mathrm{Var}}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) &= R_t - R_t F_t Q_t^{-1} F_t R_t && (4.25) \end{aligned}$$

From (4.23), $\mathrm{E}(\boldsymbol{\theta}_t \mid D_t) = \mathrm{E}\{\mathrm{E}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) \mid D_t\}$ and $\mathrm{Var}(\boldsymbol{\theta}_t \mid D_t) = \mathrm{Var}\{\mathrm{E}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) \mid D_t\} + \mathrm{E}\{\mathrm{Var}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) \mid D_t\}$. Thus, the posterior moments of $\boldsymbol{\theta}_t$ may be estimated based on the optimal estimates of (4.24) and (4.25).

$$\begin{aligned} \mathbf{m}_t &= \mathrm{E}\{\hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1} \mid D_t)\} \\ &= \mathrm{E}\{\mathbf{a}_t + R_t F_t Q_t^{-1}(\boldsymbol{\eta}_t - \mathbf{f}_t) \mid D_t\} \\ &= \mathbf{a}_t + R_t F_t Q_t^{-1}\{\mathrm{E}(\boldsymbol{\eta}_t \mid D_t) - \mathbf{f}_t\} \\ &= \mathbf{a}_t + R_t F_t Q_t^{-1}(\mathbf{f}_t^* - \mathbf{f}_t) && (4.26) \\ C_t &= \mathrm{Var}\{\hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) \mid D_t\} + \mathrm{E}\{\hat{\mathrm{Var}}(\boldsymbol{\theta}_t \mid \boldsymbol{\eta}_t, D_{t-1}) \mid D_t\} \\ &= \mathrm{Var}\{\mathbf{a}_t + R_t F_t Q_t^{-1}(\boldsymbol{\eta}_t - \mathbf{f}_t) \mid D_t\} + \mathrm{E}(R_t - R_t F_t Q_t^{-1} F_t' R_t \mid D_t) \\ &= R_t F_t Q_t^{-1} \mathrm{Var}(\boldsymbol{\eta}_t \mid D_t) Q_t^{-1} F_t' R_t + R_t - R_t F_t Q_t^{-1} F_t' R_t \\ &= R_t F_t Q_t^{-1} Q_t^* Q_t^{-1} F_t' R_t + R_t - R_t F_t Q_t^{-1} F_t' R_t \\ &= R_t - R_t F_t Q_t^{-1}(I - Q_t^* Q_t^{-1}) F_t' R_t && (4.27) \end{aligned}$$

4.4.2 Particle Filters

Sequential Monte Carlo (SMC) methods, also known as the particle filters, are a set of simulation approaches to infer the posterior distributions in case analytically they are not possible to get. The particle filters apply importance sampling meth-

ods sequentially: at each time t, particles are generated from a suitable importance density and then the importance weights, which compensate to account for the fact we do not sample from the true posterior distribution, are updated recursively over time. Unfortunately, the weights are known to degenerate over time, meaning that very few particles have significant weights and all the rest are virtually zero. This effect fails the Monte Carlo approximation and hence a resampling step is applied at each time t when the particles are thought to degenerate.

The particle filter successfully works in online filtering application when the posterior distribution is not easy to obtain analytically. The posterior distribution is sequentially approximated by the Monte Carlo method for integration with the samples from a proposal density $q_t(\cdot)$, also known as an importance function or an importance density. When the posterior distribution $p(\boldsymbol{\theta}_t \mid D_t)$ is known, $q_t(\cdot) \equiv p(\boldsymbol{\theta}_t \mid D_t)$. When it is unknown, we write $q_t(\cdot)$ explicitly as $q_t(\boldsymbol{\theta}_t \mid \mathbf{X}_t)$, which can be represented as

$$q_t(\boldsymbol{\theta}_t \mid \mathbf{X}_t) = q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) \cdot q_{t-1}(\boldsymbol{\theta}_{t-1} \mid \mathbf{X}_{t-1}) \tag{4.28}$$

where $q_{t|t-1}(\cdot)$ is the importance transition density.

While applying the particle filter to the online monitoring process with a multinomial distribution for the response vector $\mathbf{X}_t$, N vectors of $\Theta_t = (\boldsymbol{\theta}_t^{(1)}, \cdots, \boldsymbol{\theta}_t^{(N)})'$ are sampled at each time t from the importance transition density $q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$. Each sample vector of $\boldsymbol{\theta}_t^{(i)}$ for $i = 1, \ldots, N$ consists of $(\theta_{1,t}^{(i)}, \ldots, \theta_{n,t}^{(i)})$, thus having $\Theta_t = \left((\theta_{1,t}^{(1)}, \ldots, \theta_{n,t}^{(1)}), \cdots, (\theta_{1,t}^{(N)}, \ldots, \theta_{n,t}^{(N)})\right)'$ as the sample matrix with the weights of $\mathbf{w}_t = (w_t^{(1)}, \cdots, w_t^{(N)})$ for each sample vector at t. For example, a sample vector of $\boldsymbol{\theta}_t^{(1)} = (\theta_{1,t}^{(1)}, \ldots, \theta_{n,t}^{(1)})$ has a weight of $w_t^{(1)}$, and a discrete approximation of $\hat{p}(\boldsymbol{\theta}_t \mid D_t)$ is done by the weighted vectors $(\boldsymbol{\theta}_t^{(i)}, w_t^{(i)})$, $i = 1, \ldots, N$.

4.4.2.1 The Importance Density

The selection of the importance transition density is one of the most common, but important issues considered in the particle filter. One of the most commonly used importance transition densities is such that $q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) = p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$. In this case, particles are drawn from the prior distribution irrespective of newly

made observation at t, making the calculation of the incremental weights straightforward; this is known as the bootstrap filter in Cappé et al. (2007) and Doucet and Johansen (2009). However, with no additional information from the observation $\mathbf{X}_t$ made at t, approximated moments of the posterior density from the generated particles easily become obsolete. To get over this problem, particles are generated from a conditional distribution of $\boldsymbol{\theta}_t$ given both $\boldsymbol{\theta}_{t-1}$ and $\mathbf{X}_t$, say $q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$.

According to Petris et al. (2009), the bootstrap filter is the suboptimal choice for the importance density while the optimal choice is $q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) = p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$. This is optimal in that the optimal density behaves in a similar way as the posterior of $(\boldsymbol{\theta}_t \mid D_t)$; see **Proposition 2** of Doucet et al. (2010b) and Doucet et al. (2010a). We propose that we approximate the optimal choice by simulating from a multivariate normal distribution with the mean vector and covariance matrix approximated by Bayes linear methods.

The moments of $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$ and $p(\boldsymbol{\eta}_t \mid \boldsymbol{\theta}_{t-1})$ are obtained as follows.

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) &= \mathrm{E}(G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \mid \boldsymbol{\theta}_{t-1}) = G_t\boldsymbol{\theta}_{t-1} \\
\mathrm{Var}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) &= \mathrm{Var}(G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \mid \boldsymbol{\theta}_{t-1}) = W_t \\
\mathrm{E}(\boldsymbol{\eta}_t \mid \boldsymbol{\theta}_{t-1}) &= \mathrm{E}(F_t'\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) = F_t'\mathrm{E}(G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \mid \boldsymbol{\theta}_{t-1}) = F_t'G_t\boldsymbol{\theta}_{t-1} \\
\mathrm{Var}(\boldsymbol{\eta}_t \mid \boldsymbol{\theta}_{t-1}) &= \mathrm{Var}(F_t'\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) = F_t'\mathrm{Var}(G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \mid \boldsymbol{\theta}_{t-1})F_t \\
&= F_t'W_tF_t \\
\mathrm{Cov}(\boldsymbol{\eta}_t, \boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) &= \mathrm{Cov}(F_t'\boldsymbol{\theta}_t, \boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) = F_t'\mathrm{Var}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) \\
&= F_t'\mathrm{Var}(G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t \mid \boldsymbol{\theta}_{t-1}) = F_t'W_t
\end{aligned}$$

Thus, the joint density of $\boldsymbol{\eta}_t$ and $\boldsymbol{\theta}_t$ given $\boldsymbol{\theta}_{t-1}$ is

$$\begin{pmatrix} \boldsymbol{\eta}_t \\ \boldsymbol{\theta}_t \end{pmatrix} \Bigg| \boldsymbol{\theta}_{t-1} \sim \left[\begin{pmatrix} F_t'G_t\boldsymbol{\theta}_{t-1} \\ G_t\boldsymbol{\theta}_{t-1} \end{pmatrix}, \begin{pmatrix} F_t'W_tF_t & F_t'W_t \\ W_tF_t & W_t \end{pmatrix} \right]$$

Using the linear Bayes' estimation, we have the mean vector and covariance matrix

of $(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \boldsymbol{\eta}_t)$ as

$$(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \boldsymbol{\eta}_t) \sim [G_t\boldsymbol{\theta}_{t-1} + W_tF_t(F_t'W_tF_t)^{-1}(\boldsymbol{\eta}_t - F_t'G_t\boldsymbol{\theta}_{t-1}), W_t - W_tF_t(F_t'W_tF_t)^{-1}F_t'W_t] \quad (4.29)$$

Using the tower property of conditional expectation and the variance decomposition formula with the results from (3) in Appendix B, the moments of the optimal importance kernel or the importance transition density $q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$ can be approximated as

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) &= \mathrm{E}(\hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \boldsymbol{\eta}_t) \mid \mathbf{X}_t) \\
&= \mathrm{E}(G_t\boldsymbol{\theta}_{t-1} + W_tF_t(F_t'W_tF_t)^{-1}(\boldsymbol{\eta}_t - F_t'G_t\boldsymbol{\theta}_{t-1}) \mid \mathbf{X}_t) \\
&= G_t\boldsymbol{\theta}_{t-1} + W_tF_t(F_t'W_tF_t)^{-1}(\mathbf{f}_t^* - F_t'G_t\boldsymbol{\theta}_{t-1}) \quad (4.30) \\
\mathrm{Var}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t) &= \mathrm{Var}(\hat{\mathrm{E}}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \boldsymbol{\eta}_t) \mid \mathbf{X}_t) + \mathrm{E}(\hat{\mathrm{Var}}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \boldsymbol{\eta}_t) \mid \mathbf{X}_t) \\
&= \mathrm{Var}(G_t\boldsymbol{\theta}_{t-1} + W_tF_t(F_t'W_tF_t)^{-1}(\boldsymbol{\eta}_t - F_t'G_t\boldsymbol{\theta}_{t-1}) \mid \mathbf{X}_t) \\
&\quad + \mathrm{E}(W_t - W_tF_t(F_t'W_tF_t)^{-1}F_t'W_t \mid \mathbf{X}_t) \\
&= W_tF_t(F_t'W_tF_t)^{-1}Q_t^*(F_t'W_tF_t)^{-1}F_t'W_t \\
&\quad + W_t - W_tF_t(F_t'W_tF_t)^{-1}F_t'W_t \\
&= W_t + W_tF_t(F_t'W_tF_t)^{-1}(Q_t^*(F_t'W_tF_t)^{-1} - I)F_t'W_t \quad (4.31)
\end{aligned}$$

where $\mathbf{f}_t^* = \mathrm{E}(\boldsymbol{\eta}_t \mid D_t)$ and $Q_t^* = \mathrm{Var}(\boldsymbol{\eta}_t \mid D_t)$.

Thus, N particles for the particle filter are sampled from the multivariate normal distribution with moments of $(\mathbf{m}_t^{OIK}, C_t^{OIK})$ where $\mathbf{m}_t^{OIK} = G_t\boldsymbol{\theta}_{t-1} + W_tF_t(F_t'W_tF_t)^{-1}(\mathbf{f}_t^* - F_t'G_t\boldsymbol{\theta}_{t-1})$ from (4.30) and $C_t^{OIK} = W_t + W_tF_t(F_t'W_tF_t)^{-1}(Q_t^*(F_t'W_tF_t)^{-1} - I)F_t'W_t$ from (4.31).

4.4.2.2 The Incremental Weights

From a link function for the observation model in (4.3), we can see that conditioning upon $\boldsymbol{\theta}_t$ implies conditioning upon $\boldsymbol{\eta}_t$; i.e. $p(\mathbf{X}_t \mid \boldsymbol{\eta}_t) \equiv p(\mathbf{X}_t \mid \boldsymbol{\theta}_t)$.

By Bayes' theorem, the weights $\mathbf{w}_t$ are updated as

$$\begin{aligned}
\mathbf{w}_t &\propto \frac{p(\boldsymbol{\theta}_t \mid \mathbf{X}_t)}{q_t(\boldsymbol{\theta}_t \mid \mathbf{X}_t)} \propto \frac{p(\boldsymbol{\theta}_t, \mathbf{X}_t \mid \mathbf{X}_{t-1})}{q_t(\boldsymbol{\theta}_t \mid \mathbf{X}_t)} \\
&\propto \frac{p(\boldsymbol{\theta}_t, \mathbf{X}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_{t-1})}{q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)} \cdot \frac{p(\boldsymbol{\theta}_{t-1} \mid \mathbf{X}_{t-1})}{q_{t-1}(\boldsymbol{\theta}_{t-1} \mid \mathbf{X}_{t-1})} \\
&\propto \frac{p(\mathbf{X}_t \mid \boldsymbol{\theta}_t) \cdot p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})}{q_{t|t-1}(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{X}_t)} \cdot \mathbf{w}_{t-1}
\end{aligned} \tag{4.32}$$

For each particle vector of $\boldsymbol{\theta}_t^{(i)}$, the weight of $w_t^{(i)}$ is calculated from

$$w_t^{(i)} = \frac{p(\mathbf{X}_t \mid \boldsymbol{\theta}_t^{(i)}) \cdot p(\boldsymbol{\theta}_t^{(i)} \mid \boldsymbol{\theta}_{t-1}^{(i)})}{q_{t|t-1}(\boldsymbol{\theta}_t^{(i)} \mid \boldsymbol{\theta}_{t-1}^{(i)}, \mathbf{X}_t)} \cdot w_{t-1}^{norm(i)} \tag{4.33}$$

where $\frac{p(\mathbf{X}_t|\boldsymbol{\theta}_t^{(i)}) \cdot p(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)})}{q_{t|t-1}(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, \mathbf{X}_t)}$ is called as the incremental weight. The optimal importance kernel $q_{t|t-1}(\cdot)$ is Markovian, and the incremental weight depends only on $\boldsymbol{\theta}_t^{(i)}$ and $\boldsymbol{\theta}_{t-1}^{(i)}$.

At each time t when the particles are sampled, each incremental weight $w_t^{(i)}$ is normalized as

$$w_t^{norm(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}} \tag{4.34}$$

4.4.2.3 Resampling Methods

At each time t, the incremental weights $w_t^{(i)}$ are computed by $\frac{p(\mathbf{X}_t|\boldsymbol{\theta}_t^{(i)}) \cdot p(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)})}{q_{t|t-1}(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, \mathbf{X}_t)}$ as in (4.33). Accordingly, the weights are updated with new weights at t from the old weights at $t-1$ and normalized to be $\sum_{i=1}^{N} w_t^{norm(i)} = 1$ as in (4.34). With these normalized weights, the effective sample size N_{eff} is computed as $1/\sum_{i=1}^{N}(w_t^{norm(i)})^2$, and compared with a threshold N_{thr} to see whether to employ the resampling step or not. The threshold is decided by the modeller, but it is normally $N/2$. When $N_{eff} < N_{thr}$, the resampling step is employed and the weights are set to be $1/N$.

Many different resampling algorithms are developed to keep its Monte Carlo variance as small as possible and to reduce the computational complexity. Among them,

multinomial, stratified, residual, and systematic resamplings are the most frequently employed in literature. A theoretical framework and the differences among different methods for resampling can be found in Hol et al. (2006).

We employ the simplest of multinomial resampling. When we have N particles from the importance sampling and need to do the resampling by $N_{eff} < N_{thr}$, a sample of size N are drawn with replacement from the discrete distribution of $p(\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^{(i)}) = w_t^{(i)}$. As mentioned, the weights become $1/N$ after the resampling.

4.5 Comparison of The Bootstrap Filter and The Particle Filter

When the prior distribution is chosen as the importance density to sample from, the particle filter is known as the bootstrap filter. To see the benefit of using the optimal importance kernel in the particle filter, a comparison is made between the particle filter and the bootstrap filter. To help a reader to understand both the bootstrap filter and the particle filter, a summary is provided as a table for each in the following section.

For an illustration, the states $\boldsymbol{\theta}_t = (\theta_{1,t}, \theta_{2,t})$ are generated for categorical time series with 3 categories by a random walk process, where the probabilities $\boldsymbol{\Pi}_t = (\pi_{1,t}, \pi_{2,t}, \pi_{3,t})$ are determined via the link function and the logistic transformation assuming that $F_t = I$. Observations $\mathbf{X}_t$ are simulated from Multinomial$(1, \boldsymbol{\Pi}_t)$ for $t = 1, 2, \ldots$, indicating the sum of the count at each time t is 1. Comparisons between the particle filter and the bootstrap filter are made with firstly 100 and secondly 1,000 generated time series. The number of particles N are set as 1,000 to generate at time t for both filters. Each comparison provides a plot of the original probabilities $\boldsymbol{\Pi}_t = (\pi_{1,t}, \pi_{2,t}, \pi_{3,t})$ used to generate multi-categorical time series at each time t and the estimated probabilities $\hat{\boldsymbol{\Pi}}_t = (\hat{\pi}_{1,t}, \hat{\pi}_{2,t}, \hat{\pi}_{3,t})$ for each category.

In addition to Multinomial$(1, \boldsymbol{\Pi}_t)$, observations $\mathbf{X}_t$ are simulated from Multinomial$(100, \boldsymbol{\Pi}_t)$ for $t = 1, 2, \ldots$, indicating the sum of the counts are 100 at each time t, when the

results from both filters are compared. This is to see if the only one count at each time t may cause some poor performance.

4.5.1 The Bootstrap Filter

Table 4.1 describes the bootstrap filter in detail. In this chapter, the bootstrap filter does not store the paths of the particles from time 0 to time t when the interest lies only in estimating $p(\boldsymbol{\theta}_t|D_t)$ or the moments of the distribution. At time t, the filter starts with $\{\boldsymbol{\theta}_t^{(i)}, \frac{1}{N}\}$, and then updates the importance weights, using the information given at time t, to $\{\boldsymbol{\theta}_t^{(i)}, w_t^{(i)}\}$. In the resampling step, only the fittest are selected to obtain the unweighted measure $\{\boldsymbol{\theta}_t^{(i)}, \frac{1}{N}\}$. These are used to approximate $p(\boldsymbol{\theta}_t|D_t)$. The bootstrap filter adopts the prior distribution $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)})$ as an importance distribution to sample from. As a distributional assumption, the multivariate normal distribution is considered for an importance density.

Table 4.1: Pseudo-code implementations for the bootstrap filter

1. Initialisation at $t = 0$
· Set N as the number of particles generated for the filtering
· Set $\mathbf{m}_0$ and C_0 for MVN($\mathbf{m}_0$, C_0)
· Sample $\boldsymbol{\theta}_0^{(i)} = \{\theta_{1,0}^{(i)}, \theta_{2,0}^{(i)}, \ldots, \theta_{n,0}^{(i)}\}$ for $i = 1, 2, \ldots, N$ from MVN($\mathbf{m}_0$, C_0) where $(n+1)$ is the number of categories
· Set $w_0^{(i)} = \frac{1}{N}$ for $i = 1, 2, \ldots, N$
2. Importance sampling step for $t = 1, 2, \ldots, T$
· Sample $\boldsymbol{\theta}_t^{(i)}$ for $i = 1, 2, \ldots, N$ from $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}^{(i)})$ or MVN$(G_t\boldsymbol{\theta}_{t-1}^{(i)}, W_t)$
· Update the importance weights $w_t^{(i)}$ for $i = 1, 2, \ldots, N$ $w_t^{(i)} = p(\mathbf{X}_t|\boldsymbol{\theta}_t^{(i)})$
· Normalise the importance weights as $w_t^{norm(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}}$
3. Resampling step for $t = 1, 2, \ldots, T$
· Resample N particles with replacement according to the importance weights

4.5.2 The Particle Filter

Table 4.2 describes the particle filter. Key differences between the bootstrap filter and the particle filter are on the choice of the importance density to sample the

particles from, the weights updating, and the decision criterion at the resampling step. The particle filter uses the optimal importance kernel $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{X}_t)$ as an importance desnsity, and updates the weights at t using the incremental weights and the weights at $t-1$. To decide whether to resample the particles, the modeller specifies a threshold N_0 which is compared with the effective sample size N_{eff}.

Table 4.2: Pseudo-code implementations for the particle filter

1. Initialisation at $t = 0$
· Set N as the number of particles generated for the filtering
· Set $\mathbf{m}_0$ and C_0 for MVN($\mathbf{m}_0$, C_0)
· Sample $\boldsymbol{\theta}_0^{(i)} = \{\theta_{1,0}^{(i)}, \theta_{2,0}^{(i)}, \ldots, \theta_{n,0}^{(i)}\}$ for $i = 1, 2, \ldots, N$ from MVN($\mathbf{m}_0$, C_0) where $(n+1)$ is the number of categories
· Set $w_0^{(i)} = \frac{1}{N}$ for $i = 1, 2, \ldots, N$
2. Importance sampling step for $t = 1, 2, \ldots, T$
· Sample $\boldsymbol{\theta}_t^{(i)}$ for $i = 1, 2, \ldots, N$ from MVN($\mathbf{m}_t^{oik}, C_t^{oik}$) where $\mathbf{m}_t^{oik} = G_t\boldsymbol{\theta}_{t-1}^{(i)} + W_tF_t(F_t'W_tF_t)^{-1}(\mathbf{f}_t^* - F_t'G_t\boldsymbol{\theta}_{t-1}^{(i)})$ and $C_t^{oik} = W_t + W_tF_t(F_t'W_tF_t)^{-1}(Q_t^*(F_t'W_tF_t)^{-1} - I)F_t'W_t$ with $\mathbf{f}_t^* = \mathrm{E}(\boldsymbol{\eta}_t \mid D_t)$ and $Q_t^* = \mathrm{Var}(\boldsymbol{\eta}_t \mid D_t)$
· Update the importance weights $w_t^{(i)}$ for $i = 1, 2, \ldots, N$ $w_t^{(i)} = \frac{p(\mathbf{X}_t\|\boldsymbol{\theta}_t^{(i)})\cdot p(\boldsymbol{\theta}_t^{(i)}\|\boldsymbol{\theta}_{t-1}^{(i)})}{q_{t\|t-1}(\boldsymbol{\theta}_t^{(i)}\|\boldsymbol{\theta}_{t-1}^{(i)}, \mathbf{X}_t)} \cdot w_{t-1}^{(i)}$
· Normalise the importance weights as $w_t^{norm(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}$
3. Resampling step for $t = 1, 2, \ldots, T$
· Set the threshold N_0
· Compute the effective sample size N_{eff} defined as $1/\sum_{i=1}^N (w_t^{norm(i)})^2$
· If $N_{eff} < N_0$, resample N particles with replacement and set the weights $w_t^{(i)} = \frac{1}{N}$ for $i = 1, 2, \ldots, N$

4.5.3 Comparison Results

To see the performance of both the particle filter and the bootstrap filter in the class of DGLM for a multi-categorical time series, both filters are applied to the generated time series for comparisons. Multi-categorical time series of 100 data points with 3 categories, having the sum of the counts as 1, 100, or 1,000 at each time t, are generated from a random walk process. For Monte Carlo simulation, a

multi-categorical time series generation of 100 data points is iterated for 100 times, where 1,000 particles are generated at each time t in each iteration.

As an example from 100 iterations, Figure 4.1, 4.2, and 4.3 show the probabilities of each category by both the particle filter and the bootstrap filter when the sum of the count at each time t is 1, 100, and 1,000 respectively. The upper plots in the figure are the results by the particle filter while the lower plots are by the bootstrap filter. In each of the upper and the lower plots of the figure, the original probabilities are drawn in black to generate a multi-categorical time series of 100 data points with 3 categories while the probabilities estimated by the filters are in red. From the figures, it is found that the filters work better with multi-categorical time series having more sums of counts at each time t, but not clearly seen which filter performs better than the other.

To see the details of the comparisons, the distance ($D_{i,t}$) between the original probabilities to generate the time series and the estimated probabilities by the filters is defined and measured by the differences between the two at each time t, say $D_{i,t} = \pi_{i,t} - \hat{\pi}_{i,t}$ for $i = 1, 2, 3$ in this section. Absolute deviation of the estimated probabilities by each of the filters from the original probabilities is measured at each time t and averaged in each iteration. Thus, the smaller mean and the smaller standard error for each category would mean the better estimation of probabilities by the filter. As an example from 100 iterations, Figure 4.4, 4.5, and 4.6 show the absolute deviation $|D_{i,t}|$ of the estimated probabilities by each of the filters from the original probabilities of each category by both the particle filter and the bootstrap filter when the sum of the count at each time t is 1, 100, and 1,000 respectively. In the plots of the figure, the absolute deviation $|D_{i,t}|$ is drawn in black by the particle filter and in red by the bootstrap filter. Table 4.5.3 shows the mean and the standard error of the absolute deviation measured by $|D_{i,t}|$ for $i = 1, 2, 3$ from each of the particle filter and the bootstrap filter according to the sum of counts at each time t. From the means and the standard errors in the table, it can be said that the particle filter works better than the bootstrap filter, although slightly even for the case with sum of count of 1 at each time t. As sum of counts at t increases, the performance by the particle filter is enhanced relatively to the bootstrap filter.

Table 4.3: The mean and the standard error (s.e.) of the absolute deviation measured by $|D_{i,t}|$ for $i = 1, 2, 3$ from each of the particle filter (PF) and the bootstrap filter (BF) with the sum of counts (Nt = 1, 100, and 1,000) at each time t in the bracket

Category 1	PF(1)	BF(1)	PF(100)	BF(100)	PF(1,000)	BF(1,000)
mean	0.0755	0.0788	0.0134	0.0167	0.0041	0.0072
s.e.	0.0423	0.0464	0.0071	0.0092	0.0023	0.0042
Category 2	PF(1)	BF(1)	PF(100)	BF(100)	PF(1,000)	BF(1,000)
mean	0.0816	0.0848	0.0134	0.0169	0.0042	0.0076
s.e.	0.0400	0.0425	0.0076	0.0097	0.0027	0.0053
Category 3	PF(1)	BF(1)	PF(100)	BF(100)	PF(1,000)	BF(1,000)
mean	0.0810	0.0855	0.0137	0.0175	0.0052	0.0094
s.e.	0.0427	0.0473	0.0074	0.0093	0.0027	0.0054

4.6 Conclusion

Both filters are sequentially applied to estimate the moments and therefore the probabilities of each category as a result. According to the simulation study, the particle filter works better than the bootstrap filter in the class of DGLM for multi-categorical time series. Better performance of the particle filter may be caused by using the optimal importance kernel as an importance density.

Since the particles are simulated from the prior distribution of the states in the bootstrap filter and the prior is just a normal distribution with the mean of the state at previous time, then the generated particles at time t may be similar to those at time $t-1$. In the proposed particle filter, the particles are simulated by taking into account the prior and the observed data $\mathbf{X}_t$. Hence, the particles may well adapt to new information. The performance of the particle filter should not be influenced by whether $F_t = I$ or not. When $F_t = I$ and $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t$, we can infer the estimates of $\boldsymbol{\theta}_t$ from the posterior distribution of $\boldsymbol{\Pi}_t$, which is known by the conjugate method. However, it should be clear that the particle filter is needed as the posterior distribution of $\boldsymbol{\Pi}_t$ depends on the $\mathbf{r}_t$, which are only

calculated approximately using the Bayes linear methods. The particle filter helps doing this estimation more accurately. If $F_t = I$, the distribution of $\boldsymbol{\theta}_t$ could be inferred by the conjugate posterior distribution of $\boldsymbol{\Pi}_t$. However, as noted earlier, this distribution depends on either unknown or approximated parameters, namely the $\mathbf{r}_t$. The advantage of the proposed particle filter is that it utilizes that approximation in order to achieve a much more accurate estimation.

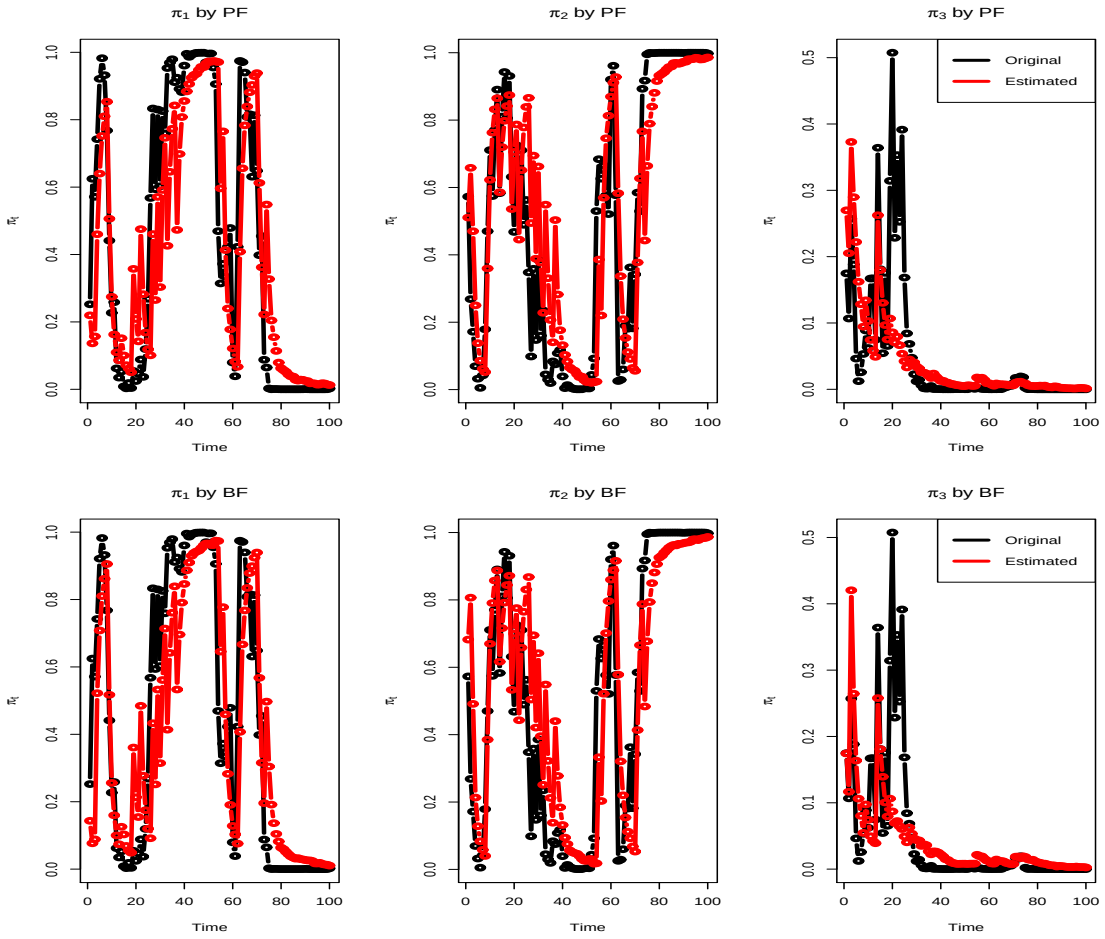

Figure 4.1: The probabilities used to generate a multi-categorical time series of 100 data points with the counts of 1 at each time t (in black for the upper and the lower plots) with the estimated posterior probabilities by the particle filter (PF) (in red for the upper plots) and by the bootstrap filter (BF) (in red for the lower plots)

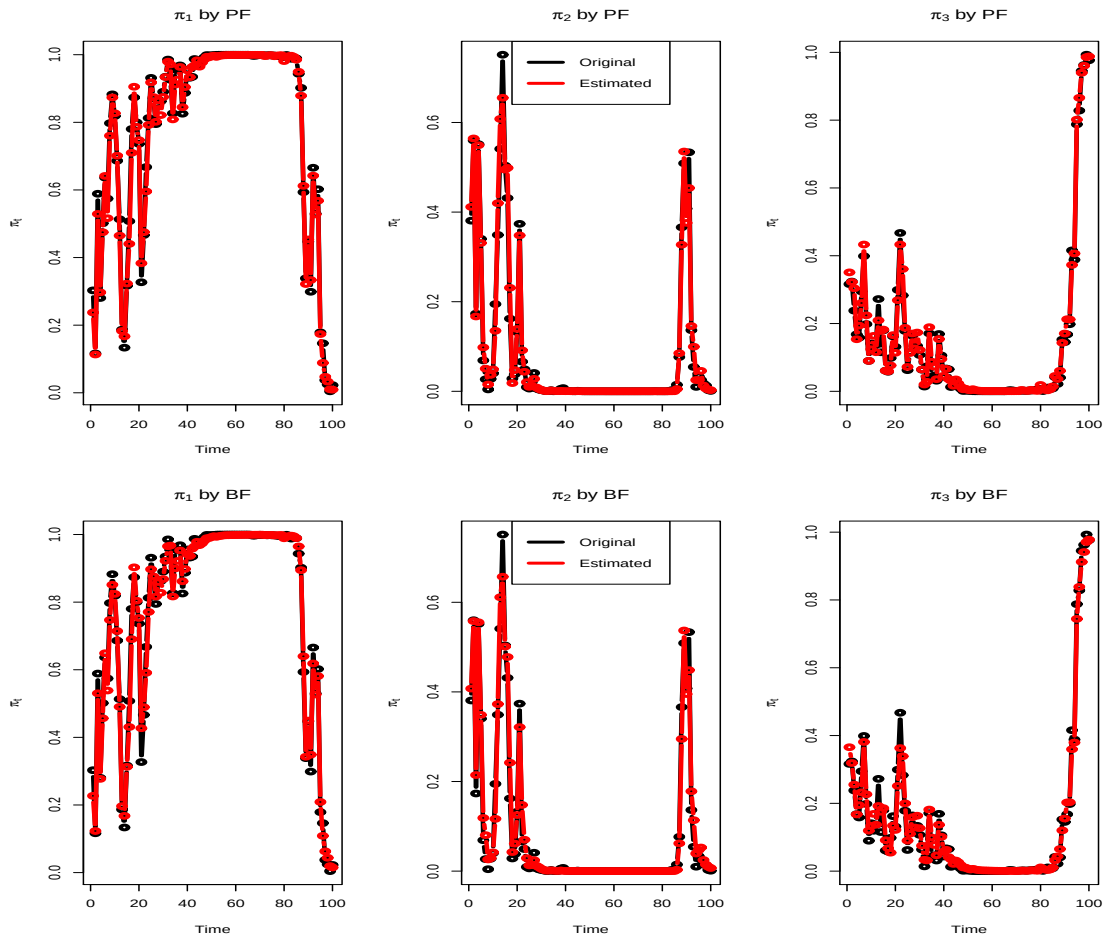

Figure 4.2: The probabilities used to generate a multi-categorical time series of 100 data points with the counts of 100 at each time t (in black for the upper and the lower plots) with the estimated posterior probabilities by the particle filter (PF) (in red for the upper plots) and by the bootstrap filter (BF) (in red for the lower plots)

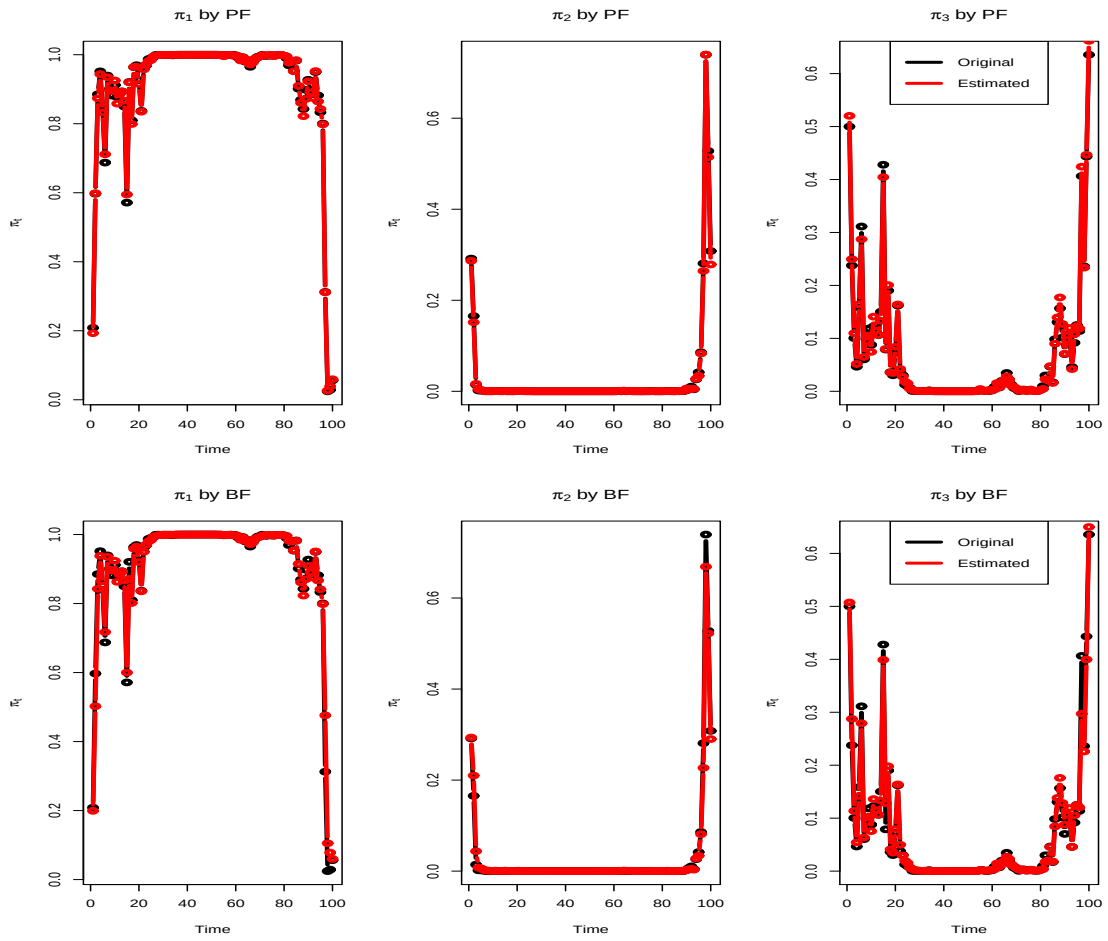

Figure 4.3: The probabilities used to generate a multi-categorical time series of 100 data points with the counts of 1,000 at each time t (in black for the upper and the lower plots) with the estimated posterior probabilities by the particle filter (PF) (in red for the upper plots) and by the bootstrap filter (BF) (in red for the lower plots)

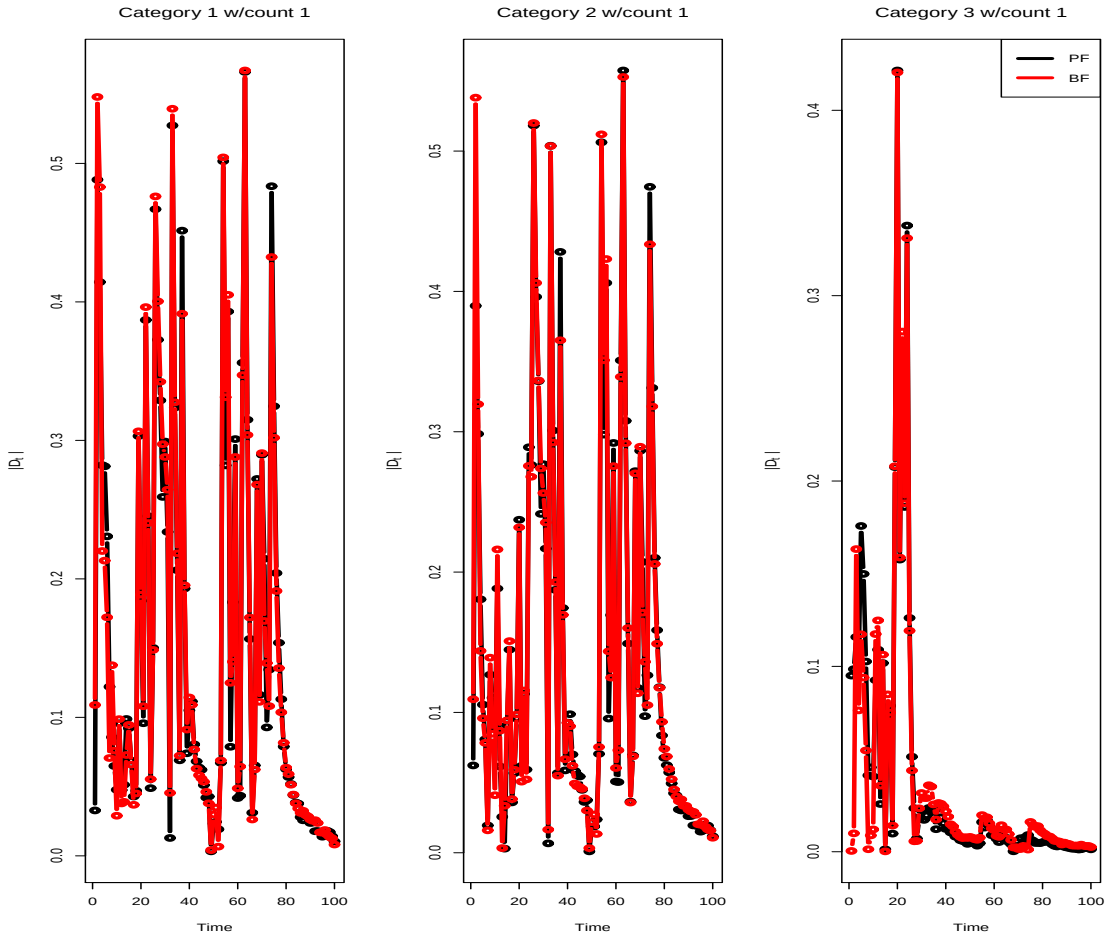

Figure 4.4: The probabilities used to generate a multi-categorical time series of 100 data points with the counts of 1 at each time t (in black for the upper and the lower plots) with the estimated posterior probabilities by the particle filter (PF) (in red for the upper plots) and by the bootstrap filter (BF) (in red for the lower plots)

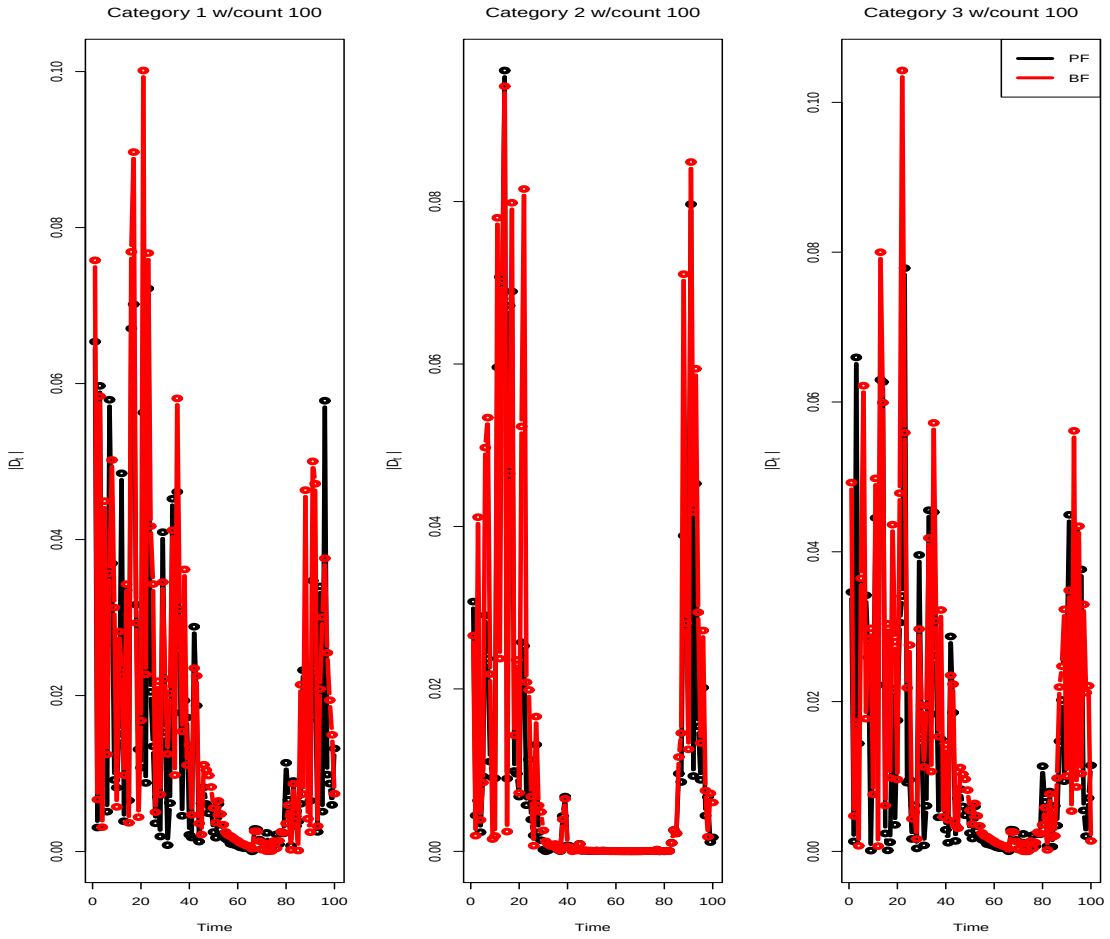

Figure 4.5: The probabilities used to generate a multi-categorical time series of 100 data points with the counts of 100 at each time t (in black for the upper and the lower plots) with the estimated posterior probabilities by the particle filter (PF) (in red for the upper plots) and by the bootstrap filter (BF) (in red for the lower plots)

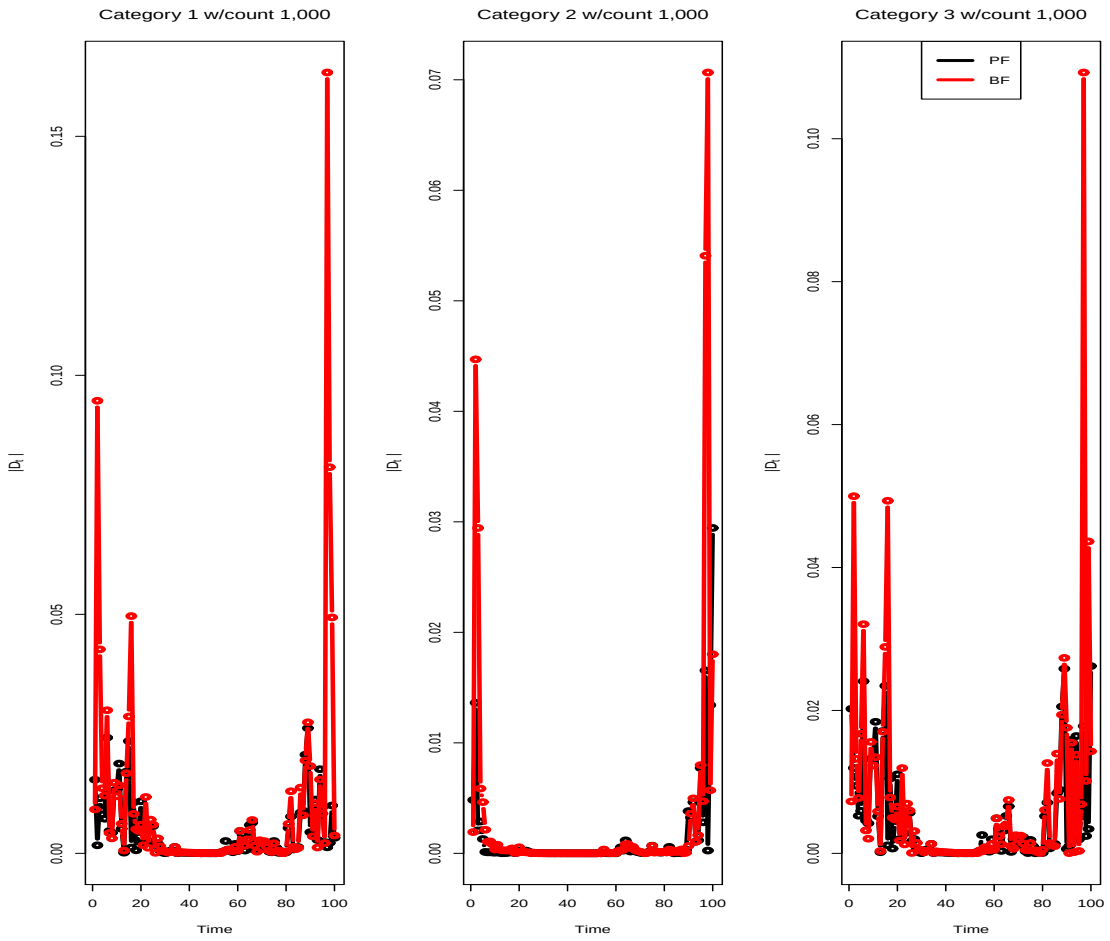

Figure 4.6: The probabilities used to generate a multi-categorical time series of 100 data points with the counts of 1,000 at each time t (in black for the upper and the lower plots) with the estimated posterior probabilities by the particle filter (PF) (in red for the upper plots) and by the bootstrap filter (BF) (in red for the lower plots)

Chapter 5

Algorithmic Pairs Trading

5.1 Introduction

Algorithmic trading implements a trading strategy and related decisions as an algorithm on a computer system where trading orders are entered without human intervention. Algorithmic pairs trading is regarded as an algorithmic trading to deal with a pair of financial instruments. In this chapter, simple implementation of algorithmic pairs trading is proposed, employing the dynamic linear model with variable forgetting factor to the spread time series from Chapter 3, and the dynamic generalised linear model developed for multi-categorical time series in Chapter 4. The dynamic linear model with variable forgetting factor detects the mean-reversion of the spread based on the value of $|\hat{B}_t|$, and the dynamic generalised linear model developed in Chapter 4 monitors the behaviour of $|\hat{B}_t|$.

Algorithmic pairs trading needs an algorithm to decide when and how many shares to trade, and probably a facility to feed the share prices to the system when the bid and ask prices are used for trading. Figure 5.1 illustrates algorithmic pairs trading as a flow chart. In the flow chart, the stages of START and END would mean that you power on and off a computer system for algorithmic pairs trading. At the Pattern Recognition & Modeling step, a pattern or a status of the spread is recognised, which can be done by applying the dynamic linear model with variable forgetting factor to the spread time series. As discussed earlier in Chapter 3, the

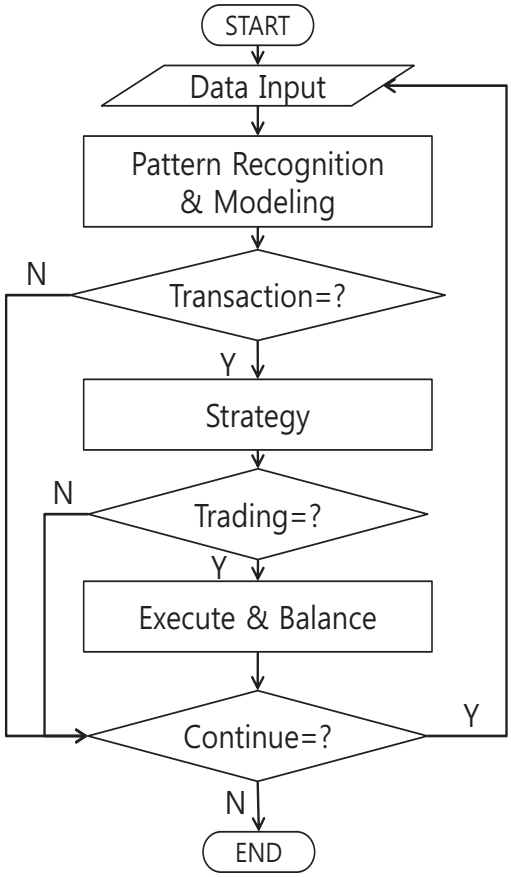

Figure 5.1: A flow chart of algorithmic pairs trading

value of $|\hat{B}_t|$ as a state of the model determines whether it is in mean-reversion or not. When in mean-reversion, a decision is made to open a position, and trading rules are taken into account such as which asset and how many to buy and short-sell at the Strategy step. The dynamic generalised linear model developed for multi-categorical time series in Chapter 4 may be involved in this Strategy step as an online monitoring process to consider if a trading should be executed at t. At the step of the Execute & Balance, trading orders are entered when a position is opened and any outstanding positions from the previous trading are cleared and closed by counter-orders. An income statement is drawn up every day to see the daily profit

and loss (P/L) which is accrued for the cumulative P/L.

5.2 Trading Rules

An investor or a trader can open her long position by buying the shares, and open her short position by short-selling them. "Short-selling" is an investment strategy to sell the shares which one does not own. When an investor anticipates that the price of any share will fall, she borrows the shares from other investors and pay back to the lender later on, usually within a few days according to the rules of the exchange. In practice, a fund manager borrows the shares via the prime broker of her own so that she does not need to find a lender and eventually does not know whom the shares are borrowed from.

The basic rule for trading is to buy an asset which value is expected to go up and short-sell an asset which value is expected to go down. Assuming mean-reversion of the spread, any asset of the two, which is expected to lose its value, is short-sold and will be bought back to pay the borrowing. The other asset is bought to be sold when its price goes up. A pair trader buys number of shares of Stock A, for example, with the money earned from short-selling the borrowed shares of Stock B at the same time, and vice versa. Thus, she does not need large capital for initial investment.

As discussed earlier, pairs trading is based on the relative mis-pricing of the pair. From the dynamic linear model to the spread time series discussed in Chapter 3, the one-step ahead forecast of the spread can be obtained. However, the prices of the individual assets, forming the spread, are not known. Thus, there is no absolute rule of thumb for the guaranteed profits as long as the betting is not on the spread or the range of the future spread. As an online monitoring process of the behaviour of $|\hat{B}_t|$, the dynamic generalised linear model, discussed earlier in Chapter 4, may allow for further restrictions to open a position at t. In addition, although a decision is made to open a position at t, questions remain such as which asset of the two to buy and/or short-sell and how many shares to trade.

In this chapter, two simple trading rules, named as Trading Rule 1 and Trading Rule 2, are proposed to decide which asset of the two to buy and/or short-sell. The trading rules follow the very basics of the investment, 'buy an asset at the low price and sell another at the high price'. When a position is opened at t, the position is closed at $t+1$ as a rule. For a question of how many to trade, the ratio approach found in Whistler (2004) is adopted. Two stocks are chosen from the New York Stock Exchange, and the number of shares determined by the ratio approach is adjusted to avoid possible loss incurred by opening a position in the illustration.

The spread Y_t is observed at t and defined by $Y_t = P_{A,t} - P_{B,t}$ where $P_{A,t}$ and $P_{B,t}$ represent the prices of the two stocks, say Stock A and Stock B, at time t. For the purpose of comparison, the price differences of a stock between t and $t+1$ are defined as $\Delta P_{A,t+1} = P_{A,t+1} - P_{A,t}$ and $\Delta P_{B,t+1} = P_{B,t+1} - P_{B,t}$. The prices of the stocks in the market are non-negative, but the spread can be negative. From the spread model discussed in Chapter 3, $\mu_t = \mathrm{E}(Y_t|\boldsymbol{\theta}_t) = \mathbf{F}_t'\boldsymbol{\theta}_t$, μ_t is regarded as the mean response of the spread at t, and f_{t+1} is the forecast of the spread obtained at t.

5.2.1 Trading Rule 1

In Trading Rule 1, Y_t and μ_t are compared to make a decision on which asset to buy and short-sell. When $\mu_t \geq Y_t$, the spread Y_t is expected to move up towards the mean response μ_t at $t+1$. Thus, it is believed that one of the following will happen at $t+1$, widening the price gap or the spread between Stock A and Stock B: (1) $\Delta P_{A,t+1} > 0$ but $\Delta P_{B,t+1} < 0$, (2) $\Delta P_{A,t+1} > \Delta P_{B,t+1} \geq 0$, (3) $0 \geq \Delta P_{A,t+1} > \Delta P_{B,t+1}$. In case of (1), an investor makes double profits from the long position of Stock A and the short position of Stock B. While an investor loses her money from the short position of Stock B, she earns more money from the long position of Stock A in (2). Even in (3), while an investor loses her money from the long position of Stock A, she can make profits from the short position of Stock B. Therefore, when $\mu_t \geq Y_t$, it seems reasonable to buy Stock A and short-sell Stock B. When $\mu_t < Y_t$, the spread Y_t is expected to get smaller towards the mean response μ_t at $t+1$, narrowing down the price gap or the spread between Stock A and Stock B. Thus, it is believed to be one

Table 5.1: Trading Rule 1: which stock to buy and short-sell at t

$\mu_t \geq Y_t$		Stock		$\mu_t < Y_t$		Stock	
μ_t	Y_t	A	B	μ_t	Y_t	A	B
+	+	Buy	SS	+	+	SS	Buy
+	-	Buy	SS	+	-	-	-
-	+	-	-	-	+	SS	Buy
-	-	Buy	SS	-	-	SS	Buy

of the followings: (1) $\Delta P_{A,t+1} < 0$ but $\Delta P_{B,t+1} > 0$, (2) $\Delta P_{B,t+1} > \Delta P_{A,t+1} \geq 0$, (3) $0 \geq \Delta P_{B,t+1} > \Delta P_{A,t+1}$. In case of (1), an investor makes double profits from the long position of Stock B and the short position of Stock A. While an investor loses her money from the short position of Stock A, she earns money from the long position of Stock B in (2). Even in (3), while an investor loses her money from the long position of Stock B, she can make profits from the short position of Stock A. Therefore, when $\mu_t < Y_t$, it seems reasonable to buy Stock B and short-sell Stock A.

Regardless of the signs of μ_t and Y_t each, Stock A is bought and Stock B is borrowed to short-sell when $\mu_t \geq Y_t$ at t. On the other hand, when $\mu_t < Y_t$, Stock B is bought and Stock A is short-sold. With this trading rule, an investor anticipates that the price of a relatively cheaper stock would bounce back and that of a relatively expensive one would fall. This Trading Rule 1 is summarised in Table 5.1.

5.2.2 Trading Rule 2

If Y_{t+1} were known at t, we would know which to buy and short-sell with ease. However, Y_{t+1} is not known, but forecasted as f_{t+1} at time t. Thus, in Trading Rule 2, f_{t+1}, the one-step ahead forecast of the spread, is compared with Y_t at t. This trading rule is introduced in Triantafyllopoulos and Han (2013) and illustrated for pairs trading of Walmart Stores Inc. and Target Corporation.

Before making the direct comparisons between f_{t+1} and Y_t at t, a prediction margin h $(0 < h < 1)$ is introduced to allow uncertainty of the prediction by the model. This prediction margin h is assumed to make sure that Y_{t+1}, unknown

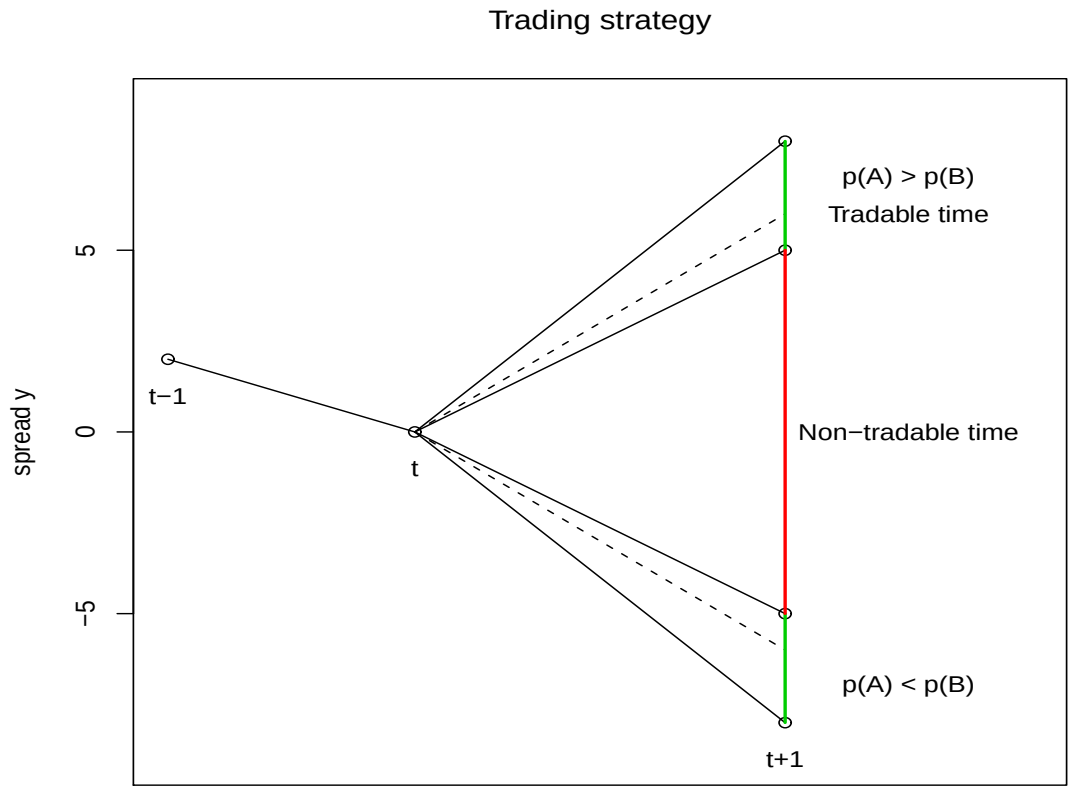

Figure 5.2: Trading Rule 2

at t, would fall in the range of $(f_{t+1} - h \cdot |f_{t+1}|, f_{t+1} + h \cdot |f_{t+1}|)$. Thus, when $f_{t+1} - h \cdot |f_{t+1}| < Y_t < f_{t+1} + h \cdot |f_{t+1}|$, a decision is made not to open a position even when $|\hat{B}_t| < 0.9$ at time t. Figure 5.2 illustrates the proposed strategy.

Now that f_{t+1} with a prediction margin h is in comparison with Y_t, Stock B is bought and Stock A is short-sold when $f_{t+1} + h \cdot |f_{t+1}| \leq Y_t$ and $|\hat{B}_t| < 1$. When $f_{t+1} - h \cdot |f_{t+1}| \geq Y_t$ and $|\hat{B}_t| < 1$, a decision is made to buy Stock A and to short-sell Stock B. This Trading Rule 2 is summarised in Table 5.2.

Table 5.2: Trading Rule 2: which stock to buy and short-sell at t

Condition	Buy	Short-sell (SS)
When $f_{t+1} - h \cdot \lvert f_{t+1} \rvert \geq Y_t$	Stock A	Stock B
When $f_{t+1} + h \cdot \lvert f_{t+1} \rvert \leq Y_t$	Stock B	Stock A

5.2.3 Number of Shares To Buy and Short-sell

Even when one is sure which stock to buy and short-sell at t, another question remains on how many shares to buy and short-sell. As in Whistler (2004), we use the ratio r_t which is defined as $P_{A,t}/P_{B.t}$.

If $Y_t \geq 0$, it indicates $P_{A,t} \geq P_{B,t}$, leading to $r_t \geq 1$. Thus, when $Y_t \geq 0$ or $r_t \geq 1$, 100 shares of Stock A and $100 \cdot r_t$ shares of Stock B is traded. For example, when $Y_t \geq 0$ (or $r_t \geq 1$) and $\mu_t \geq Y_t$ in Trading Rule 1, 100 shares of Stock A is bought and $100 \cdot r_t$ shares of Stock B is short-sold. In Trading Rule 2, when $Y_t \geq 0$ (or $r_t \geq 1$) and $f_{t+1} - h \cdot |f_{t+1}| \geq Y_t$, 100 shares of Stock A is bought and $100 \cdot r_t$ shares of Stock B is short-sold. On the other hand, when $Y_t \geq 0$ (or $r_t \geq 1$) and $\mu_t < Y_t$ in Trading Rule 1, 100 shares of Stock A is short-sold and $100 \cdot r_t$ shares of Stock B is bought. In Trading Rule 2, when $Y_t \geq 0$ (or $r_t \geq 1$) and $f_{t+1} + h \cdot |f_{t+1}| \leq Y_t$, 100 shares of Stock A is short-sold and $100 \cdot r_t$ shares of Stock B is bought.

If $Y_t < 0$, it indicates $P_{A,t} < P_{B,t}$, leading to $r_t < 1$. Thus, when $Y_t < 0$ (or $r_t < 1$), 100 shares of Stock B and $100 \cdot r_t$ shares of Stock A is traded. For example, when $Y_t < 0$ (or $r_t < 1$) and $\mu_t \geq Y_t$ in Trading Rule 1, 100 shares of Stock B is short-sold while $100 \cdot r_t$ shares of Stock A is bought. In Trading Rule 2, when $Y_t < 0$ (or $r_t < 1$) and $f_{t+1} - h \cdot |f_{t+1}| \geq Y_t$, 100 shares of Stock B is short-sold while $100 \cdot r_t$ shares of Stock A is bought. On the other hand, when $Y_t \geq 0$ and $\mu_t < Y_t$ in Trading Rule 1, 100 shares of Stock B is bought while $100 \cdot r_t$ shares of Stock A is short-sold. In Trading Rule 2, when $Y_t \geq 0$ and $f_{t+1} + h \cdot |f_{t+1}| \leq Y_t$, 100 shares of Stock B is bought while $100 \cdot r_t$ shares of Stock A is short-sold.

The earlier discussion on how many shares to trade is summarised in Table 5.3.

Table 5.3: Number of Shares to Trade

r_t	Y_t	Stock A	Stock B
≥ 1	≥ 0	min(100,100 $*$ r_t)	max(100,100 $*$ r_t)
< 1	< 0	max(100,100 $*$ r_t)	min(100,100 $*$ r_t)

5.3 Illustration: AEM-NEM

We consider the pair consisting of Agnico-Eagle Mines Limited (abbreviated as AEM) and Newmont Mininig Corporation (abbreviated as NEM) in the exchange. The former has its headquarter in Toronto, Canada while the latter is based in Colorado, U.S.A. However, both of them are listed and classified as in the same sector of Basic Materials and also the same industry of Gold on the New York Stock Exchange. More details on the companies can be found on the corporate web sites of their own, which are `http://www.agnicoeagle.com/` for Agnico-Eagle Mines Limited (AEM) and `http://www.newmont.com/` for Newmont Mining Corporation (NEM) respectively.

The daily prices of AEM and NEM are considered over a period from 3rd Jan. 2012 to 18th Oct. 2013, and their historical share prices are available from Yahoo!Finance (`http://finance.yahoo.com/`). Figure 5.3 shows the daily adjusted closing share prices of the two stocks and their spread, defined by the difference of the two, as an inset. For example, the spread Y_t is obtained by $Y_t = P_{AEM,t} - P_{NEM,t}$ where $P_{AEM,t}$ and $P_{NEM,t}$ are the share prices of AEM and NEM each at time t. A historical mean of the spread from the two is indicated as -4.172301 over the period of 452 trading days from 3rd Jan. 2012 to 18th Oct. 2013.

5.3.1 The Spread Models

For the detection of mean-reversion, a time-varying autoregressive model of order 1, which is represented in dynamic linear model, is applied to the spread of the pair of AEM and NEM with variable forgetting factor. Assuming the unknown but

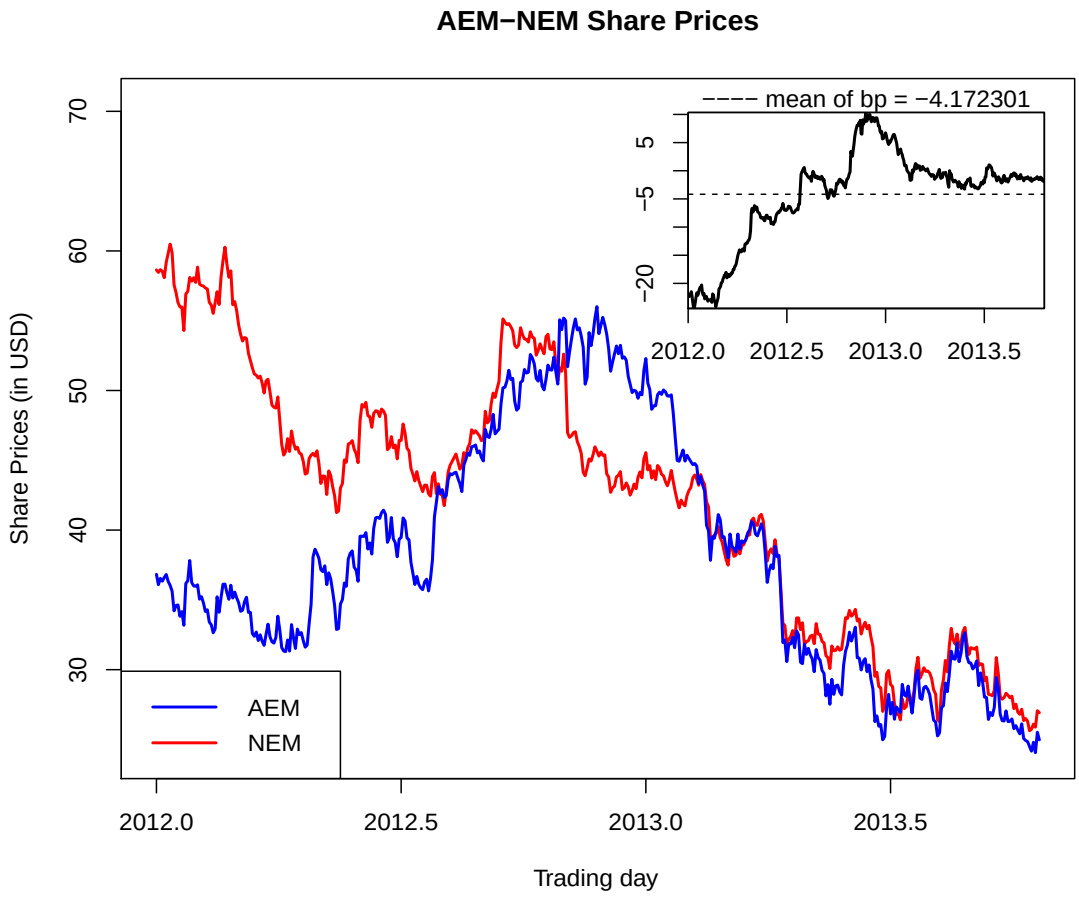

Figure 5.3: Share prices of Agnico-Eagle Mines Limited (AEM) and Newmont Mining Corporation (NEM) with their spread time series as an inset

constant observational variance of $V_t = V$, a univariate DLM is specified as follows.

$$
\begin{aligned}
Y_t &= \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t, \qquad \nu_t \sim N(0, V) && (5.1)\\
\boldsymbol{\theta}_t &= G\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim T_{n_{t-1}}(\mathbf{0}, VW_t^*) && (5.2)\\
&\quad (\boldsymbol{\theta}_0 \mid D_0) \sim T_0(\mathbf{m}_0, VC_0^*) && (5.3)\\
&\quad (\tau \mid D_0) \sim Gamma(\frac{n_0}{2}, \frac{d_0}{2}) && (5.4)
\end{aligned}
$$

where $\mathbf{F}_t' = (1, Y_{t-1})'$, $\boldsymbol{\theta}_t = (A_t, B_t)$, $G = \text{diag}(\phi_1, \phi_2)$, and $\boldsymbol{\omega}_t = (\omega_{1,t}, \omega_{2,t})$. $T_{n_{t-1}}$ denotes the t-distribution with degrees of freedom n_{t-1}, and the starred variance

matrices of C_0^* and W_t^* represent the scale-free variance-covariance matrices.

As in Triantafyllopoulos and Montana (2011), both A_t and B_t are considered to evolve via AR models over time, making ϕ_1 and ϕ_2 the AR coefficients as $A_t = \phi_1 A_{t-1} + \omega_{1,t}$ and $B_t = \phi_2 B_{t-1} + \omega_{2,t}$. Both coefficients of ϕ_1 and ϕ_2 are set as 0.95 so that A_t and B_t be weakly stationary.

At $t = 0$, the parameters are initialised such that $\mathbf{m}_0 = (1, 1)'$, $C_0 = I_2$, and $n_0 = U_0 = 1$ for $(\boldsymbol{\theta}_0 \mid D_0) \sim T_{n_0}[\mathbf{m}_0, C_0]$ and $(\tau \mid D_0) \sim Gamma\left(\frac{n_0}{2}, \frac{d_0}{2}\right)$ where $\mathrm{E}(\tau \mid D_0) = \frac{n_0}{d_0} = \frac{1}{U_0}$ and U_0 is a posterior point estimate of V at 0.

5.3.1.1 Recursions of Parameter Estimates

Now that the forgetting factor λ is variable, it is decided at each time t as λ_t. The recursive estimation procedure with updating equations can be achieved by

(d1) Posterior at $t-1$
$(\theta_{t-1} \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{m}_{t-1}, C_{t-1})$
where $C_t = U_{t-1} C_{t-1}^*$ and $C_{t-1}^* = \frac{GC_{t-2}G'}{\lambda_{t-2} + F_{t-1}' GC_{t-2} G' F_{t-1}}$

(d2) Prior at t
$(\theta_t \mid D_{t-1}) \sim T_{n_{t-1}}(\mathbf{a}_t, R_t)$
where $\mathbf{a}_t = G\mathbf{m}_{t-1}$, $R_t = U_{t-1} R_t^*$ and $R_t^* = \frac{1}{\lambda_{t-1}} GC_{t-1}^* G'$

(d3) One-step forecast
$(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}(f_t, Q_t)$
where $f_t = \mathbf{F}_t' G\mathbf{m}_{t-1}$, $Q_t = U_{t-1} Q_t^*$ and $Q_t^* = \frac{1}{\lambda_{t-1}}(\mathbf{F}_t' GC_{t-1}^* G' \mathbf{F}_t) + 1$

(d4) Posterior at t
$(\theta_t \mid D_t) \sim T_{n_t}(\mathbf{m}_t, C_t)$
where $\mathbf{m}_t = \mathbf{a}_t + K_t e_t$, $C_t = U_t C_t^*$ and $C_t^* = \frac{GC_{t-1}^* G'}{\lambda_{t-1} + \mathbf{F}_t' GC_{t-1}^* G' \mathbf{F}_t}$
with $K_t = R_t^* \mathbf{F}_t / Q_t^*$ and $e_t = Y_t - f_t$

5.3.1.2 Recursion of $\tau = \frac{1}{V}$

The unknown V or $\tau = \frac{1}{V}$ is sequentially updated as new observation is obtained at each time t. The posterior mean of τ is $\mathrm{E}(\tau \mid D_t) = \frac{n_t}{d_t} = \frac{1}{U_t}$ where U_t is a posterior

point estimate of V at t. The updating equations are summarised as

(e1) $(\tau \mid D_{t-1}) \sim Gamma\left(\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right)$

(e2) $(\tau \mid D_t) \sim Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$

where $n_t = n_{t-1} + 1$, $d_t = d_{t-1} + e_t^2/Q_t^*$ and $Q_t^* = \frac{1}{\lambda_{t-1}}(\mathbf{F}_t' G C_{t-1} G' \mathbf{F}_t) + 1$ as in **(d3)**.

5.3.2 The Variable Forgetting Factor Algorithms

According to Haykin (2001), the cost function J_t is defined as $J_t = \frac{1}{2}E(|e_t|^2)$ where $e_t = Y_t - f_t$. At $t = 0$, the upper and lower limits of the variable forgetting factor λ are set as $\lambda_+ = 1$ and $\lambda_- = 0.01$ respectively for $0 < \lambda_t \leq 1$.

5.3.2.1 The SDvFF

In the SDvFF, the variable forgetting factor λ_t is recursively updated as follows.

$$\lambda_t = [\lambda_{t-1} - \alpha \cdot \nabla_\lambda(t)]_{\lambda_-}^{\lambda_+} \tag{5.5}$$

where $\nabla_\lambda(t) \approx -e_t \mathbf{F}_t' G \psi_{t-1}$, $\psi_t = (I - C_t^* \mathbf{F}_t \mathbf{F}_t') G \psi_{t-1} + S_t \mathbf{F}_t e_t$, and
$S_t = \frac{G S_{t-1} G' (\lambda_{t-1} + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t) - G C_{t-1}^* G' (1 + \mathbf{F}_t' G S_{t-1} G' \mathbf{F}_t)}{(\lambda_{t-1} + \mathbf{F}_t' G C_{t-1}^* G' \mathbf{F}_t)^2}$.

The cost function J_t is $J_t = \frac{1}{2}E(|e_t|^2)$ when $\nabla_\lambda(t) \equiv \frac{\partial J_t}{\partial \lambda}$, $\psi_t \equiv \frac{\partial m_t}{\partial \lambda}$, and $S_t \equiv \frac{\partial C_t^*}{\partial \lambda}$.

At $t = 0$, λ_0 is set as 0.8 for $0 < \lambda_t \leq 1$, $\alpha = 0.5$, $\nabla_\lambda(0) = 0$, $\boldsymbol{\psi}_0 = (1, 1)'$, and $S_0 = I_2$.

5.3.2.2 The GNvFF

In the GNvFF, the variable forgetting factor λ_t is recursively updated as follows.

$$\lambda_t = \left[\lambda_{t-1} - \alpha \cdot \frac{\nabla_\lambda(t)}{\nabla_\lambda^2(t)}\right]_{\lambda_-}^{\lambda_+} \tag{5.6}$$

where $\nabla_\lambda^2(t) \approx (\mathbf{F}_t' G \psi_{t-1})^2 - e_t \mathbf{F}_t' G \frac{\partial \psi_{t-1}}{\partial \lambda}$,
$\eta_t = (I - C_t^* \mathbf{F}_t \mathbf{F}_t') G \eta_{t-1} + L_t \mathbf{F}_t e_t - 2 S_t \mathbf{F}_t \mathbf{F}_t' G \psi_{t-1}$, and $L_t = \frac{A}{B}$

with $A = \left\{GL_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(\mathbf{F}_t'GL_{t-1}G'\mathbf{F}_t)\right\}$
$\cdot \left\{(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^2\right\}$
$- \left\{GS_{t-1}G'(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t) - GC_{t-1}^*G'(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\}$
$\cdot \left\{2(\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)(1 + \mathbf{F}_t'GS_{t-1}G'\mathbf{F}_t)\right\}$
and $B = (\lambda_{t-1} + \mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t)^4$.

The cost function J_t and $\nabla_\lambda(t)$ are the same as those in the SDvFF while $\eta_t \equiv \frac{\partial \psi_t}{\partial \lambda}$ and $L_t \equiv \frac{\partial S_t}{\partial \lambda}$.

At $t = 0$, λ_0 is set as 0.8 for $0 < \lambda_t \leq 1$, $\alpha = 0.5$, $\nabla_\lambda(0) = 0$, $\boldsymbol{\psi}_0 = (1, 1)'$, and $S_0 = I_2$, which are the same as for the SDvFF. In addition, $\boldsymbol{\eta}_0 = (1, 1)'$, and $L_0 = I_2$ for the GNvFF.

5.3.2.3 The BBvFF(d, k)

In the BBvFF(d, k), the variable forgetting factor λ_t is recursively updated as follows.

$$\lambda_t = \hat{\pi}_t \cdot \lambda_+ + (1 - \hat{\pi}_t) \cdot \lambda_- \tag{5.7}$$

where $\hat{\pi}_t = \text{mode}(\pi_t) = \frac{\alpha_{1,t}-1}{\alpha_{1,t}+\alpha_{2,t}-2}$ from $(\pi_t \mid x_t) \sim \text{Beta}(\alpha_{1,t}, \alpha_{2,t})$.

Suppose that $(\pi_{t-1} \mid D_{t-1}) \sim \text{Beta}(\alpha_{1,t-1}, \alpha_{2,t-1})$ and $p(\pi_t \mid D_{t-1}) \propto p(\pi_{t-1} \mid D_{t-1})^k$ as in Smith (1979) where k is a discount factor. When x_t is defined as a binary series, taking a value of either 1 or 0 at each time t according to

$$x_t = \begin{cases} 1, & \text{if } \frac{|e_t|}{\sqrt{Q_t}} \leq d, \quad \text{with probability } \pi \\ 0, & \text{if } \frac{|e_t|}{\sqrt{Q_t}} > d, \quad \text{with probability } 1 - \pi \end{cases}$$

where $d(> 0)$ is a threshold specified by the modeller, $(\pi_t \mid D_t) \sim \text{Beta}(\alpha_{1,t}, \alpha_{2,t})$ with $\alpha_{1,t} = \alpha_{1,t-1}k - k + 1 + x_t$ and $\alpha_{2,t} = \alpha_{2,t-1}k - k + 2 - x_t$.

At $t = 0$, the parameters $\alpha_{1,0}$ and $\alpha_{2,0}$ are set as 2 for $(\pi_0 \mid D_0) \sim \text{Beta}(\alpha_{1,0}, \alpha_{2,0})$, and $d = 0.1$ $(d > 0)$ for a threshold of the binary series x_t while a discount factor k is $0 < k \leq 1$ and chosen to be either 0.95 or 0.5 for the illustration of this

chapter. In the BBvFF(d, k), λ_0 is not necessary because it is determined from $\lambda_0 = \hat{\pi}_0 \cdot \lambda_+ + (1 - \hat{\pi}_0) \cdot \lambda_-$ where $\hat{\pi}_0 = \text{mode}(\pi_0) = \frac{\alpha_{1,0}-1}{\alpha_{1,0}+\alpha_{2,0}-2}$.

5.4 Comparisons By The VFF Algorithms

A daily balance represents the profits and losses of the day. For example, the position from time $t-1$, if any, is closed at time t, and new position is opened at t when needed. Profits and losses from those transactions are calculated for the day's balance at time t. On the other hand, a cumulative balance shows the accrued profits and losses since the trading starts.

Figure 5.4 shows the estimated coefficients $|\hat{B}_t|$ by four different VFF algorithms when the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5) is applied to the spread time series of AEM-NEM. One of differences among the four plots in Figure 5.4 may be seen over some period between 16th Oct. 2012 (200^{th} trading day) and 13th Feb. 2013 (280^{th} trading day). The spread time series and the one-step ahead forecasts between 16th Oct. 2012 (200^{th} trading day) and 13th Feb. 2013 (280^{th} trading day) are shown in Figure 5.6 while the estimated coefficients $|\hat{B}_t|$ over the period are enlarged in Figure 5.5.

Table 5.4 shows the comparison of the forecasting over the period between 16th Oct. 2012 (200^{th} trading day) and 13th Feb. 2013 (280^{th} trading day) by the mean absolute deviation (MAD) and the mean squared error (MSE) where the MAD is defined by $\frac{\sum_{i=1}^{n}|e_i|}{n}$ and the MSE is by $\frac{\sum_{i=1}^{n}|e_i|^2}{n}$. In the table, $\hat{Y}_{t+1} = Y_t$ represents a naive forecast. The BBvFF(0.1,0.99) shows the lowest values for both of the MAD and the MSE, indicating the most accurate forecasting among the five. Even the BBvFF(0.1,0.5) is better than the GNvFF, the SDvFF, and the naive forecast.

5.4.1 Case A: Decision by $|\hat{B}_t|$

In Case A, we only consider the rule $|\hat{B}_t| < 1$ proposed by Triantafyllopoulos and Montana (2011) to detect mean-reversion.

Table 5.4: Comparisons of the forecasting over the period between 16th Oct. 2012 (200^{th} trading day) and 13th Feb. 2013 (280^{th} trading day) by the mean absolute deviation (MAD) and the mean squared error (MSE)

	$\hat{Y}_{t+1} = Y_t$	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
MAD	0.908	0.806	0.761	0.728	0.754
MSE	1.536	1.248	1.062	0.995	1.059

5.4.1.1 Case A with Trading Rule 1

Figure 5.7 and 5.8 shows the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms. According to the cumulative balances in the figures, all of them are in losses for most of time. Table 5.5 shows the daily earnings (D.E.) on average, the mean and the standard deviation (s.d.) of the cumulative balances over the period, and the final balance on the final trading day of 18th Oct. 2013 (452^{nd} trading day). Among them, the SDvFF seems to perform better than the other three, keeping the mean relatively at the highest level and the s.d. at the lowest level. Also, the SDvFF earns USD 1.691 daily on average while the GNvFF and the BBvFF(0.1,0.99) loses on average USD 1.098 and USD 0.058 per day respectively. Looking at the final balance, the SDvFF even ends up its cumulative balance at USD 764.19 while the BBvFF(0.1,0.5) at USD 103.29.

Table 5.5: Case A and Trading Rule 1: Comparisons of the daily earnings (D.E.) on average, the cumulative balances over the period between 3rd Jan. 2012 (1^{st} trading day) and 18th Oct. 2013 (452^{nd} trading day) by the mean, and the standard deviation (s.d.) and the final balance (F.B.) on 18th Oct. 2013 (452^{nd} trading day) in USD

	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	1.691	-1.098	-0.058	0.229
Mean	-358.122	-856.461	-746.100	-473.772
s.d.	519.001	649.662	560.472	541.552
F.B.	764.19	-496.20	-26.20	103.29

5.4.1.2 Case A with Trading Rule 2

Figure 5.9 and 5.10 shows the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms in Case A with Trading Rule 2 under $h = 0.01$. According to the cumulative balances in the figures, all of them struggle with some losses at the beginning of the period. During the rest of trading, they enjoy the cumulative balances in the black, still mostly for the SDvFF, ending up with the final cumulative balance at USD 922.29, USD 2,054.78, USD 2,208.65, and USD 1,710.18 respectively for the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5). The average cumulative balances are USD 648.662, USD 1,247.278, USD 1,818.474, and USD 1,685.890, and the daily earnings are USD 2.040, USD 4.546, USD 4.886, and USD 3.784 on average for each. According to the daily earnings and the cumulative balances both on average, the BBvFF(0.1,0.99) is the best performing in Case A with Trading Rule 2 under $h = 0.01$, followed by the GNvFF, the BBvFF(0.1,0.5), and the SDvFF.

Figure 5.11 and 5.12 shows the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms in Case A with Trading Rule 2 under $h = 0.03$. During most of the trading days, they enjoy the cumulative balances in the black, ending up with the final cumulative balance at USD 1,378.74, USD 1,325.09, USD 2,130.71, and USD 1,867.52 respectively for the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5) while the average cumulative balances are USD 1,604.252, USD 827.091, USD 1,933.781, and USD 1,871.015. The daily earnings are USD 3.050, USD 2.932, USD 4.714, and USD 4.132 on average for each. According to the daily earnings and the cumulative balances both on average, the BBvFF(0.1,0.99) produces the most profitable results in Case A with Trading Rule 2 under $h = 0.03$, followed by the BBvFF(0.1,0.5), the SDvFF, and the GNvFF.

Figure 5.13 and 5.14 shows the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms in Case A with Trading Rule 2 under $h = 0.05$. For a while since the

trading starts, all of them struggle with the cumulative balances in the red, but the recovery of the cumulative balances by both the BBvFF(0.1,0.99) and the BBvFF(0.1,0.5) is quicker to be back in the black than the SDvFF and the GNvFF. At the end, the final cumulative balance is at USD 1,207.48, USD 660.47, USD 1,918.13, and USD 1,147.95 respectively for the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5). The average cumulative balances are USD 838.991, USD 609.987, USD 1,497.844, and USD 1,224.440, and the daily earnings are USD 2.671, USD 1.461, USD 4.244, and USD 2.540 on average for each. According to the daily earnings and the cumulative balances both on average, the BBvFF(0.1,0.99) is the winner among the four, making the highest profits in Case A with Trading Rule 2 under $h = 0.05$, followed by the BBvFF(0.1,0.5), the SDvFF, and the GNvFF.

Table 5.6 summarises the daily earnings on average, the mean and the standard deviation of the cumulative balances, and the final balance on 18th Oct. 2013 (452^{nd} trading day) in USD for the DLM with each of four VFF algorithms.

Table 5.6: Case A and Trading Rule 2: Comparisons of the daily earnings (D.E.) on average, the cumulative balances by the mean and the standard deviation (s.d.) over the period between 3rd Jan. 2012 (1^{st} trading day) and 18th Oct. 2013 (452^{nd} trading day), and the final balance (F.B.) on 18th Oct. 2013 (452^{nd} trading day) in USD

$h = 0.01$	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	2.040	4.546	4.886	3.784
Mean	648.662	1,247.278	1,818.474	1,685.890
s.d.	537.441	772.829	964.357	951.765
F.B.	922.29	2,054.78	2,208.65	1,710.18
$h = 0.03$	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	3.050	2.932	4.714	4.132
Mean	1,064.252	827.091	1,933.781	1,871.015
s.d.	630.426	527.777	1,061.888	1,026.035
F.B.	1,378.74	1,325.09	2,130.71	1,867.52
$h = 0.05$	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	2.671	1.461	4.244	2.540
Mean	838.991	609.897	1,497.844	1,224.440
s.d.	748.599	583.855	980.317	864.153
F.B.	1,207.48	660.47	1,918.13	1,147.95

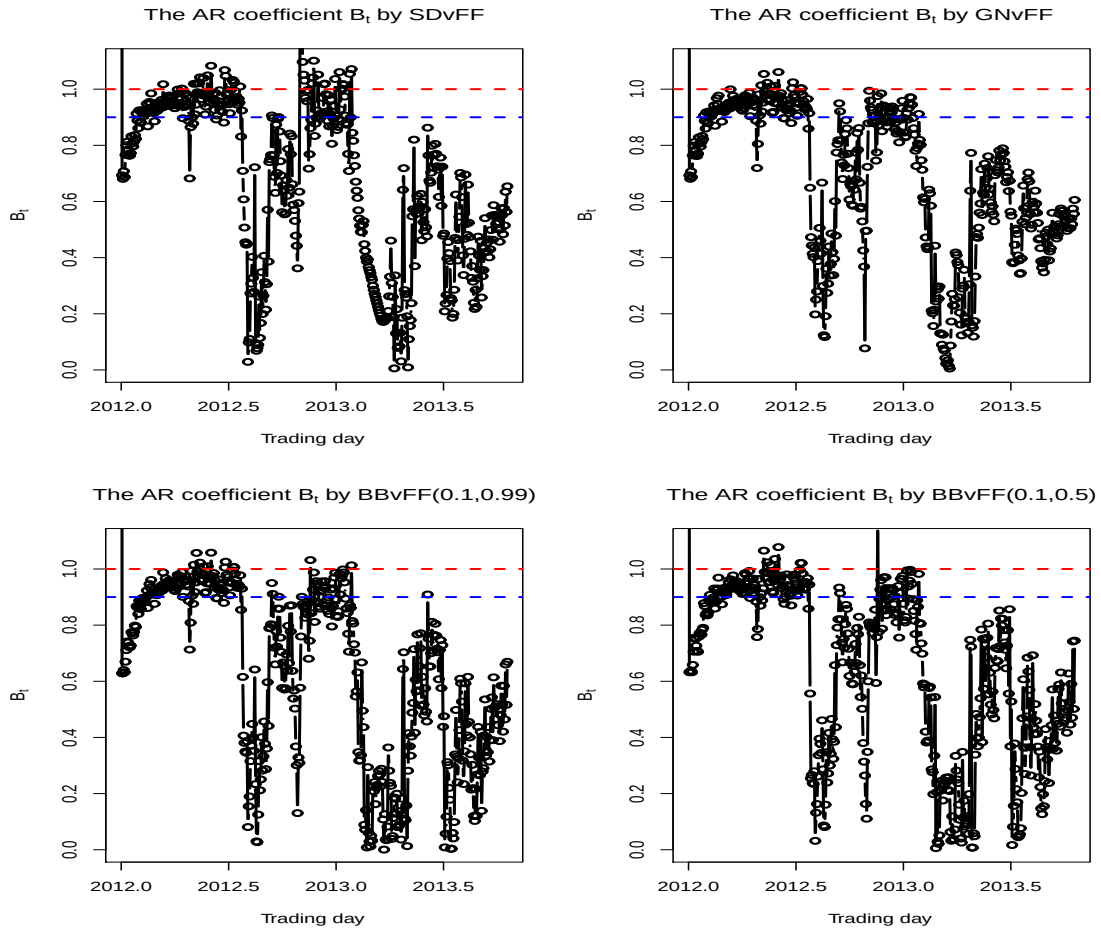

Figure 5.4: Comparison of the estimated coefficients $|\hat{B}_t|$ by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

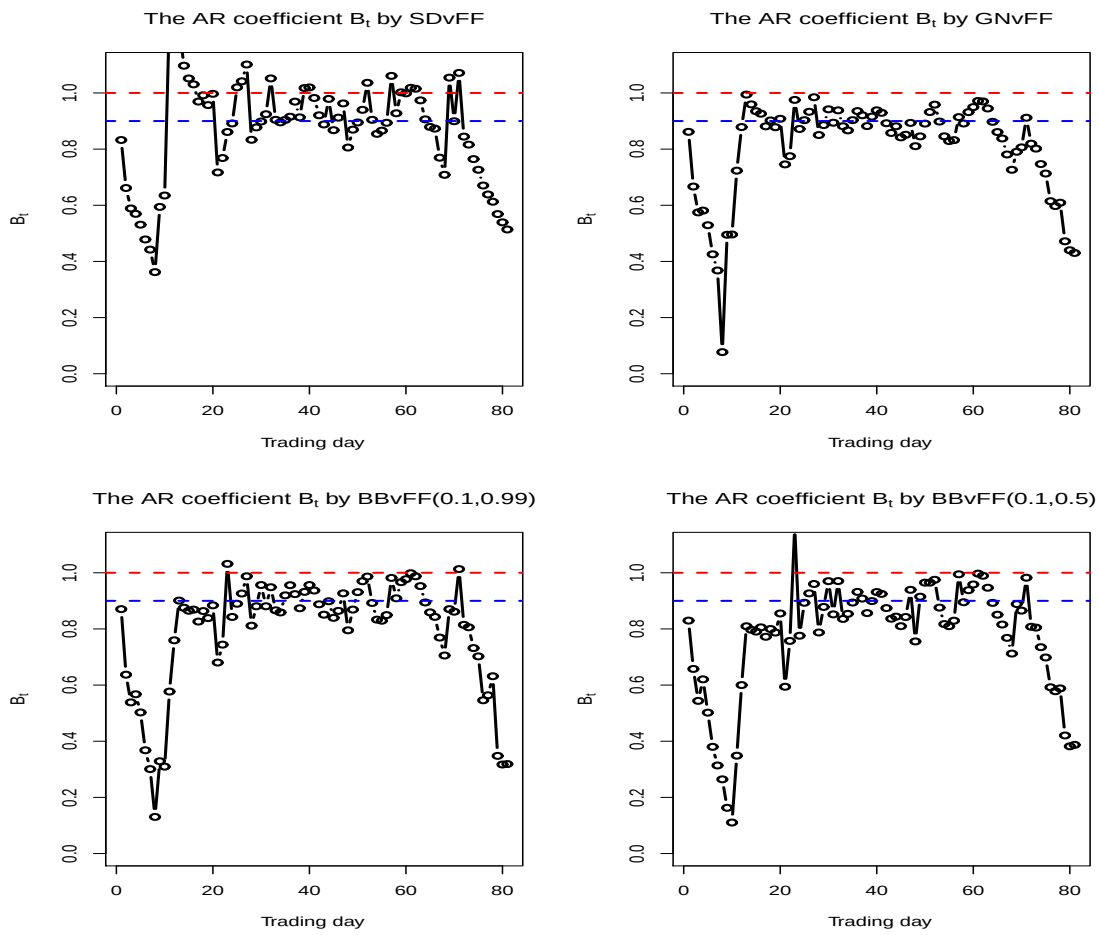

Figure 5.5: Comparison of the estimated coefficients $|\hat{B}_t|$ over the period from 16th Oct. 2012 to 13th Feb. 2013 by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

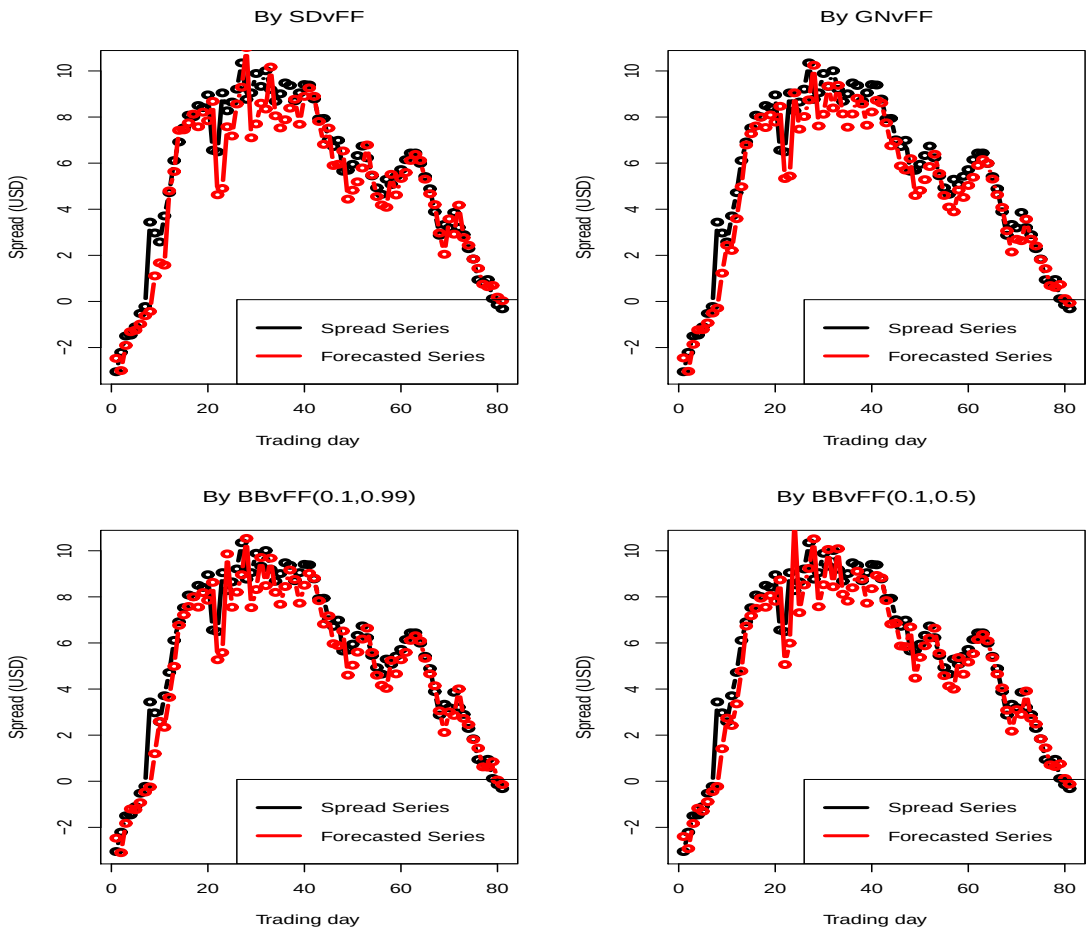

Figure 5.6: The spread time series of AEM-NEM with the one-step ahead forecast over the period from 16th Oct. 2012 to 13th Feb. 2013 by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

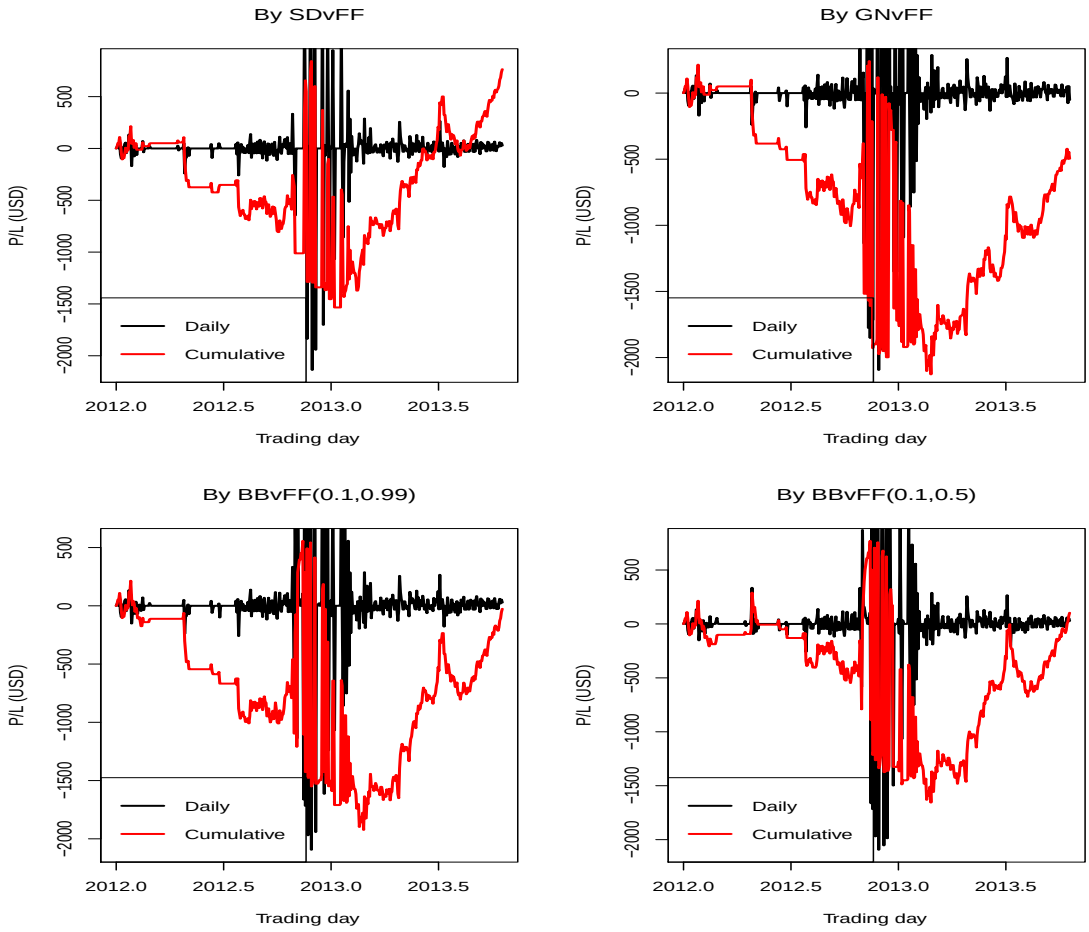

Figure 5.7: Case A with trading rule 1: Trading results by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

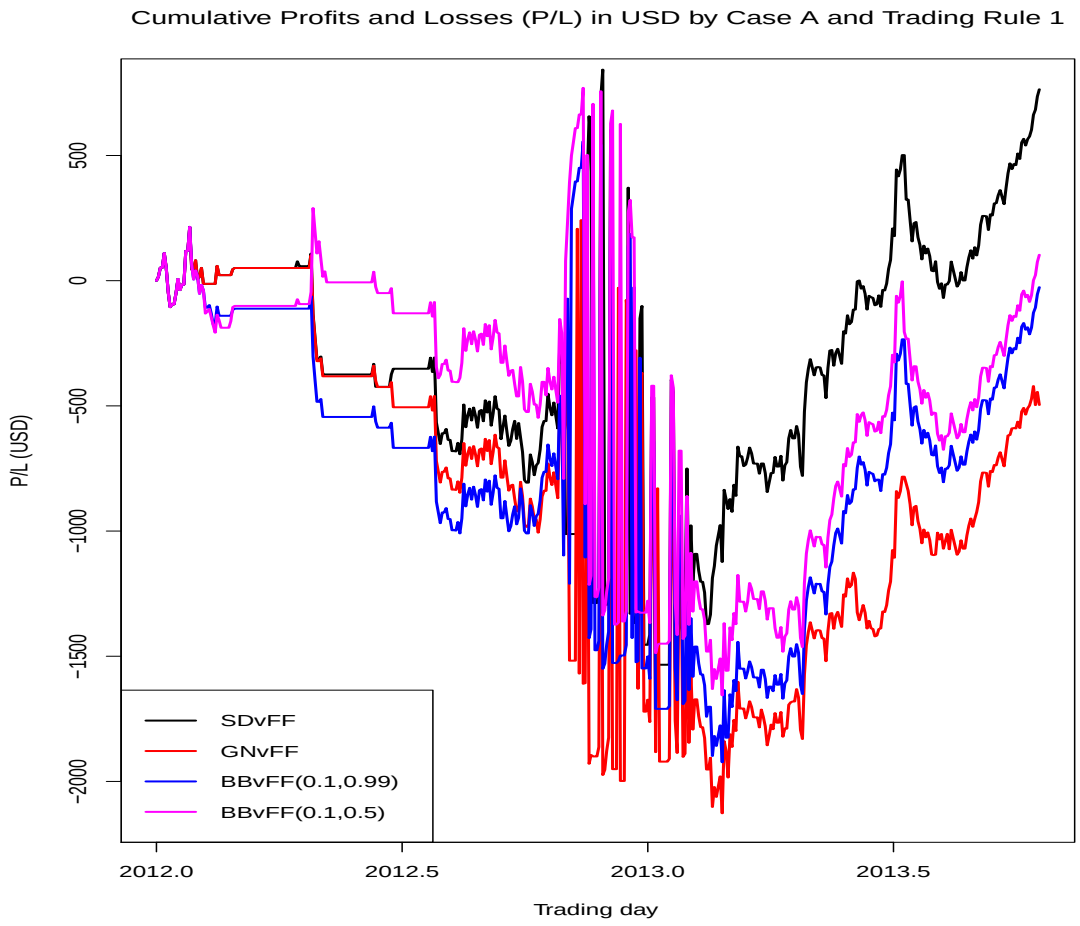

Figure 5.8: Case A with trading rule 1: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

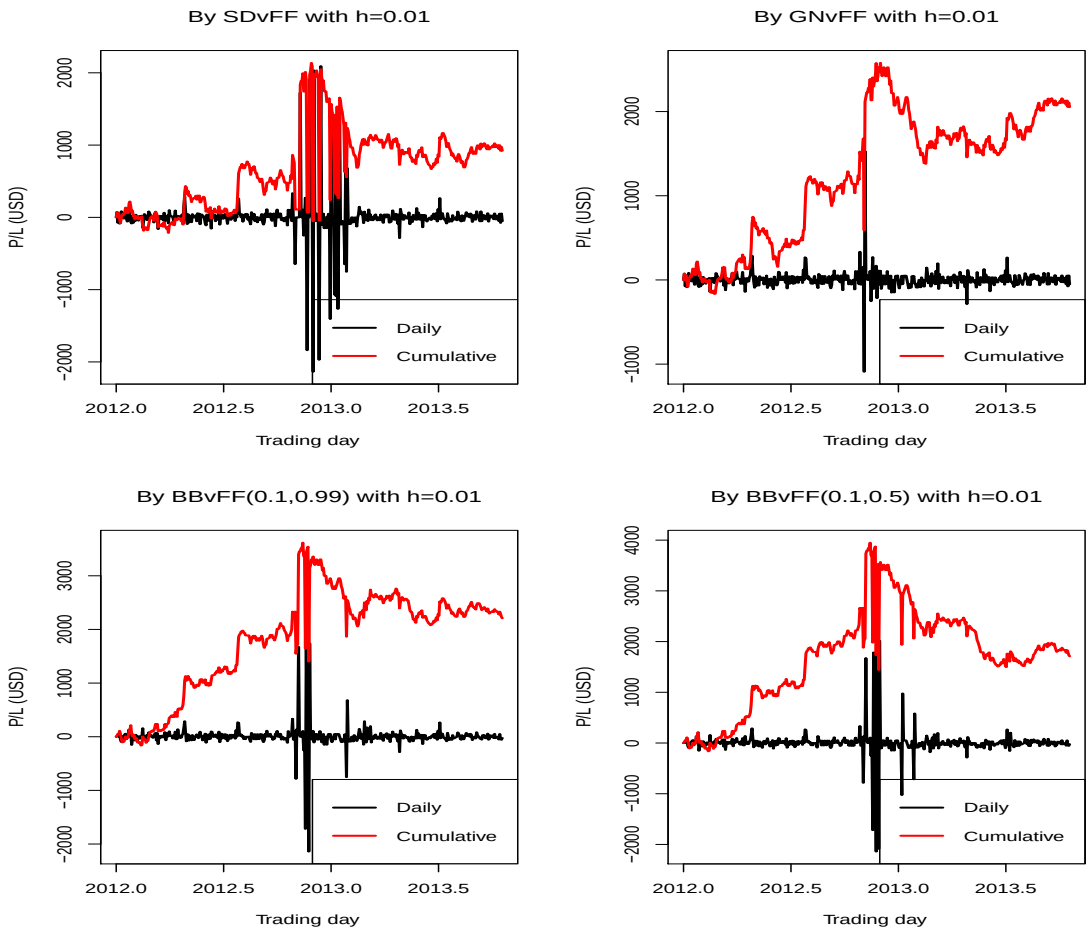

Figure 5.9: Case A with trading rule 2 with margin of 0.01: Trading results by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

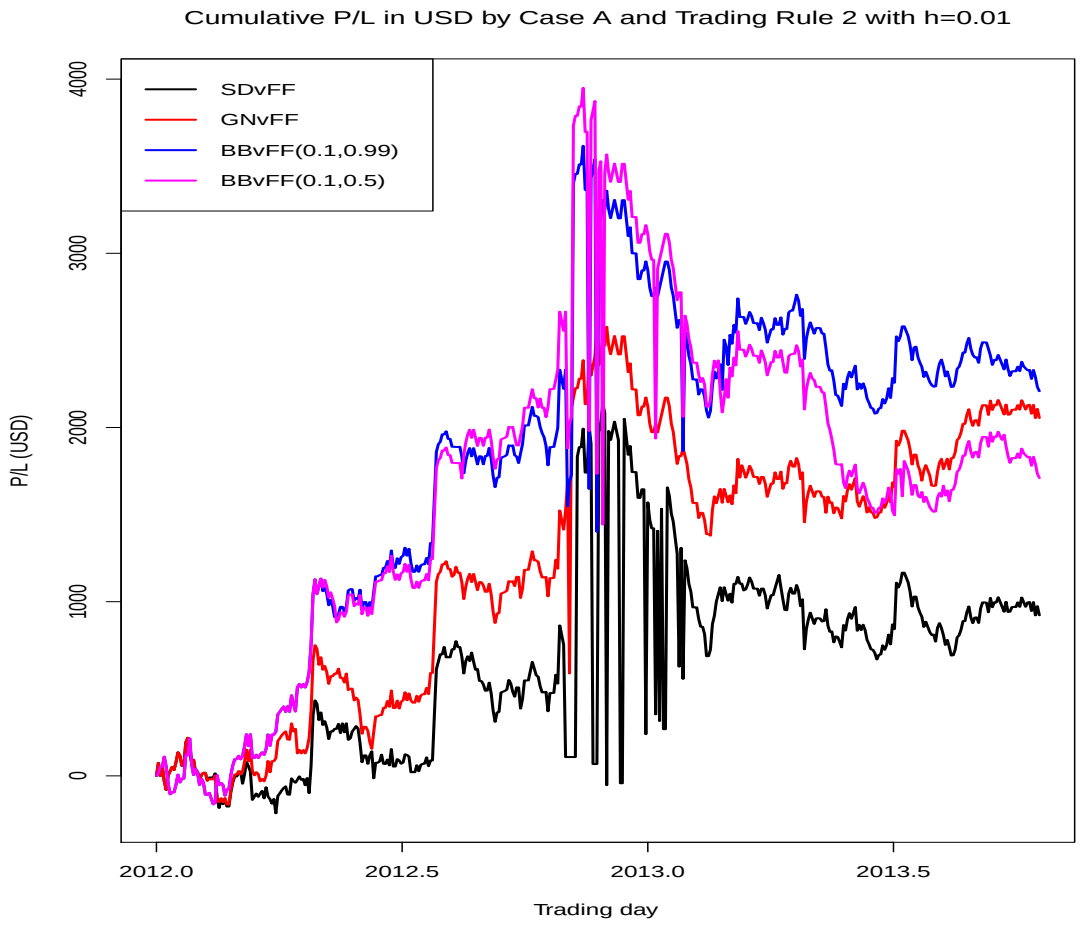

Figure 5.10: Case A with trading rule 2 with margin of 0.01: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

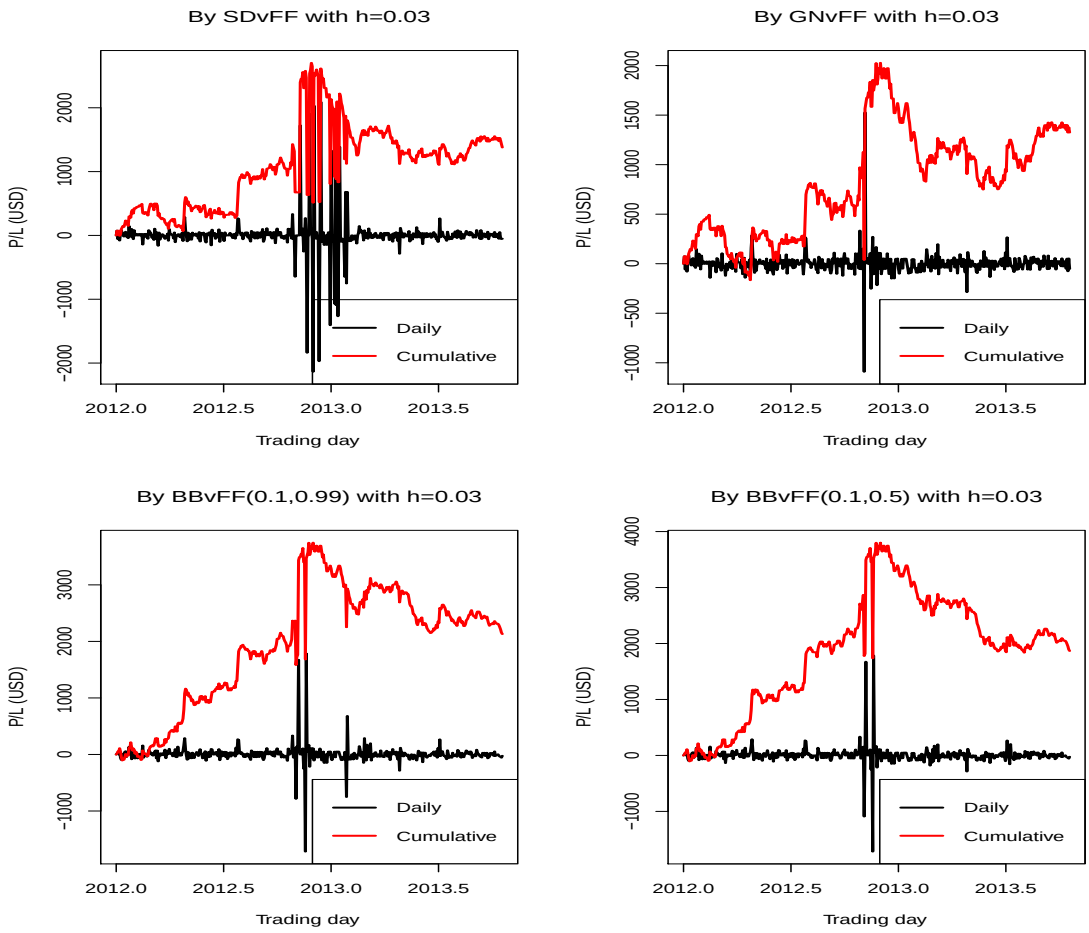

Figure 5.11: Case A with trading rule 2 with margin of 0.03: Trading results by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

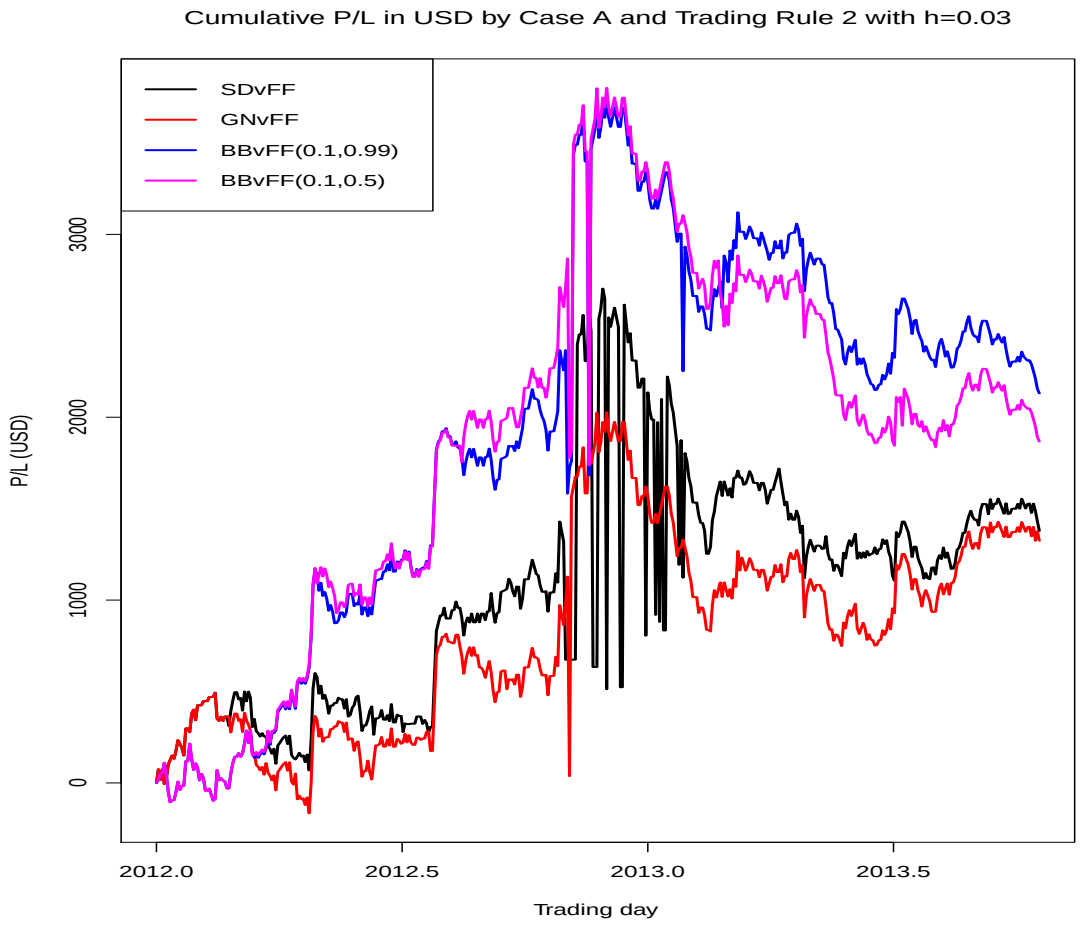

Figure 5.12: Case A with trading rule 2 with margin of 0.03: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

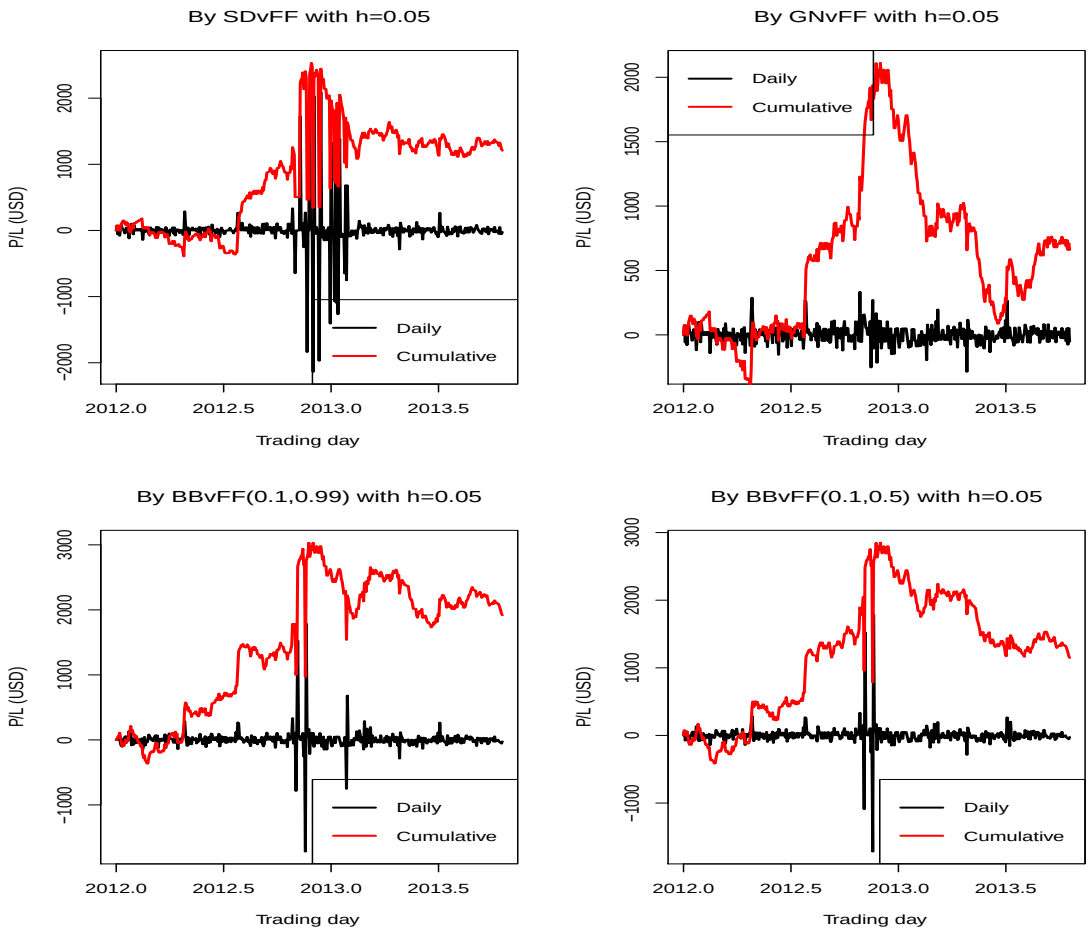

Figure 5.13: Case A with trading rule 2 with margin of 0.05: Trading results by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

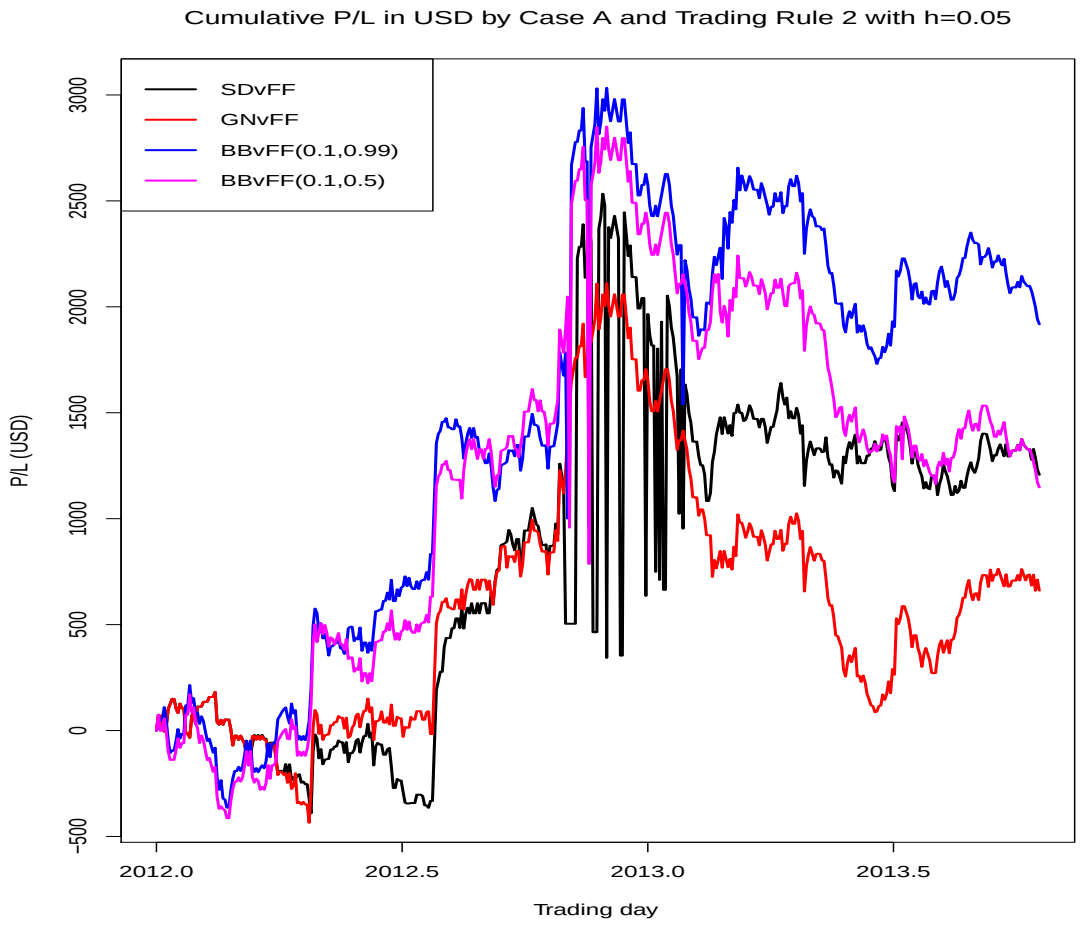

Figure 5.14: Case A with trading rule 2 with margin of 0.05: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

5.4.2 Case B: Decision by Monitoring Results

In Case B, we consider online monitoring results by the dynamic generalised linear model with the particle filter developed for multi-categorical time series in Chapter 4 of this thesis. An online monitoring process is achieved by sequentially applying the dynamic generalised linear model (DGLM) to the categorical time series generated at each time t. The particle filter is adopted for inference for multi-categorical time series. For illustrative purposes, three categories are assumed in this section according to the value of $|\hat{B}_t|$. With two thresholds of 0.9 and 1.0, there are three categories: Category 1 for $0 \leq |\hat{B}_t| < 0.9$, Category 2 for $0.9 \leq |\hat{B}_t| < 1$, and Category 3 for $1.0 \leq |\hat{B}_t|$. For example, when $|\hat{B}_t| = 0.7$ is obtained from the DLM with a VFF algorithm at t, it is counted as 1 for Category 2, and 0s for the other categories. As more observations are made for a category, the posterior probability of that category would increase. Sequential application of the DGLM with the particle filter to the multi-categorical time series aims to monitor the behaviour of $|\hat{B}_t|$ in real time. Now that the online monitoring process applies, a position is opened at t only when the posterior probability for Category 1 is greater than 0.5. Figure 5.15, 5.16, 5.17, and 5.18 show the resulting posterior probabilities of three categories, obtained from the application of the online monitoring process. Each of them takes the values of $|\hat{B}_t|$ by the DLM with each of VFF algorithms for the observations of multi-categorical time series.

5.4.2.1 Case B with Trading Rule 1

Figure 5.19 and 5.20 show the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms. Table 5.4.2.1 shows the daily earnings (D.E.) on average, the mean and the standard deviation (s.d.) of the cumulative balances over the period, and the final balance on the final trading day of 18th Oct. 2013 (452^{nd} trading day). Among the four, the BBvFF(0.1,0.99) outperforms the other three, making more daily earnings on average at USD 4.135. The BBvFF(0.1,0.99) keeps the cumulative balances at USD 292.654 on average, ending up with its cumulative balance at USD 1,869.12 on the final trading day. For the daily earnings, the BBvFF(0.1,0.99) is followed by the BBvFF(0.1,0.5) at USD 2.627, the GNvFF at USD 2.074, and the

SDvFF at USD 1.954.

Table 5.7: Case B and Trading Rule 1: Comparisons of the daily earnings (D.E.) on average, the cumulative balances over the period between 3rd Jan. 2012 (1^{st} trading day) and 18th Oct. 2013 (452^{nd} trading day) by the mean, and the standard deviation (s.d.) and the final balance (F.B.) on 18th Oct. 2013 (452^{nd} trading day) in USD

	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	1.954	2.074	4.135	2.627
Mean	-243.451	-78.915	292.654	23.755
s.d.	482.010	454.475	681.782	484.235
F.B.	883.01	937.39	1,869.12	1,187.37

5.4.2.2 Case B with Trading Rule 2

Figure 5.21 and 5.22 shows the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms in Case B with Trading Rule 2 under $h = 0.01$. The BBvFF(0.1,0.99) ends up with the biggest cumulative balance on the final trading day at USD 1,343.37, followed by the BBvFF(0.1,0.5) at USD 1,270.24, the GNvFF at USD 547.53, and the SDvFF at USD -156.02. Daily earning on average, shown in Table 5.4.2.2, is the highest with the BBvFF(0.1,0.99) at USD 2.972, followed by the BBvFF(0.1,0.5) at USD 2.810, the GNvFF at USD 1.211, and the SDvvFF, where the SDvvFF loses USD 0.345.

Figure 5.23 and 5.24 shows the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms in Case B with Trading Rule 2 under $h = 0.03$. The final cumulative balance at USD -16.13, USD 779.20, USD 1,024.12, and USD 954.10 respectively for the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5) while the average cumulative balances are USD 101.098, USD 432.415, USD 658.227, and USD 847.556. The daily earnings are USD 1.724, USD 2.266, and USD 2.111 on average for the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5) while the SDvFF loses USD 0.036. According to the daily earnings on average, the BBvFF(0.1,0.99) produces the most profitable results in Case B with Trading Rule 2 under $h = 0.03$,

followed by the BBvFF(0.1,0.5), the GNvFF, and the SDvFF.

Figure 5.25 and 5.26 shows the daily and the cumulative balances from the trading at each time t and only the cumulative balances by the DLM with each of four VFF algorithms in Case B with Trading Rule 2 under $h = 0.05$. The BBvFF(0.1,0.99) ends up with the biggest cumulative balance on the final trading day at USD 1,477.58, followed by the GNvFF at USD 882.38, the BBvFF(0.1,0.5) at USD 548.34, and the SDvFF at USD -211.24. Daily earning on average, shown in Table 5.4.2.2, is the highest with the BBvFF(0.1,0.99) at USD 3.269, followed by the GNvFF at USD 1.952, the BBvFF(0.1,0.5) at USD 1.213, and the SDvvFF, where the SDvvFF loses USD 0.467.

Table 5.4.2.2 summarises the daily earnings on average, the mean and the standard deviation of the cumulative balances, and the final balance on 18th Oct. 2013 (452^{nd} trading day) in USD for the DLM with each of four VFF algorithms.

5.5 Conclusion of Chapter 5

For illustrative purposes, two trading rules of Trading Rule 1 and Trading Rule 2 for algorithmic pairs trading are suggested in this chapter. As for decision rules whether to open a position at t, two different criteria are also suggested as Case A and Case B. A key difference between Case A and B is the condition to open a position, and Trading Rules propose a rule on which asset to buy and short-sell between the two. For both cases A and B, the DLM with VFF algorithms are applied to the spread time series of AEM-NEM over the period of 452 trading days from 3rd Jan. 2012 (1^{st} trading day) to 18th Oct. 2013 (452^{nd} trading day). In Case A, the decision is made by the rule $|\hat{B}_t| < 1$ proposed by Triantafyllopoulos and Montana (2011) to detect mean-reversion. In Case B, the behaviour of $|\hat{B}_t|$ is monitored online and the position is opened at t when the posterior probability for Category 1 ($0 \leq |\hat{B}_t| < 0.9$) is greater than 0.5. In Case B, number of categories and the threshold, specified as 0.5 in this chapter, to decide when to exercise the trading can be differently set by the modeller. In Trading Rule 1, Y_t is compared with the level or the mean response μ_t at t while f_{t+1} is chosen for comparison with Y_t in Trading Rule 2. In Case A,

Table 5.8: Case B (threshold=0.5 for category 1) and Trading Rule 2: Comparisons of the daily earnings (D.E.) on average, the cumulative balances by the mean and the standard deviation (s.d.) over the period between 3rd Jan. 2012 (1^{st} trading day) and 18th Oct. 2013 (452^{nd} trading day), and the final balance (F.B.) on 18th Oct. 2013 (452^{nd} trading day) in USD

$h = 0.01$	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	-0.345	1.211	2.972	2.810
Mean	-145.788	226.472	633.767	900.780
s.d.	279.660	364.439	856.141	831.830
F.B.	-156.02	547.53	1,343.37	1,270.24
$h = 0.03$	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	-0.036	1.724	2.266	2.111
Mean	101.098	432.415	658.227	847.556
s.d.	292.086	343.961	818.804	740.414
F.B.	-16.13	779.20	1,024.12	954.10
$h = 0.05$	SDvFF	GNvFF	BBvFF(0.1,0.99)	BBvFF(0.1,0.5)
D.E.	-0.467	1.952	3.269	1.213
Mean	-92.484	374.920	739.997	460.875
s.d.	284.180	380.639	922.310	688.686
F.B.	-211.24	882.38	1,477.58	548.34

the BBvFF(0.1,0.99) produces the highest daily earnings at USD 4.886 on average by Trading Rule 2 with $h = 0.01$ while the BBvFF(0.1,0.99) earns USD 4.135 by Trading Rule 1 in Case B. In both cases, the BBvFF(0.1,0.99) is found to be the most profitable algorithm.

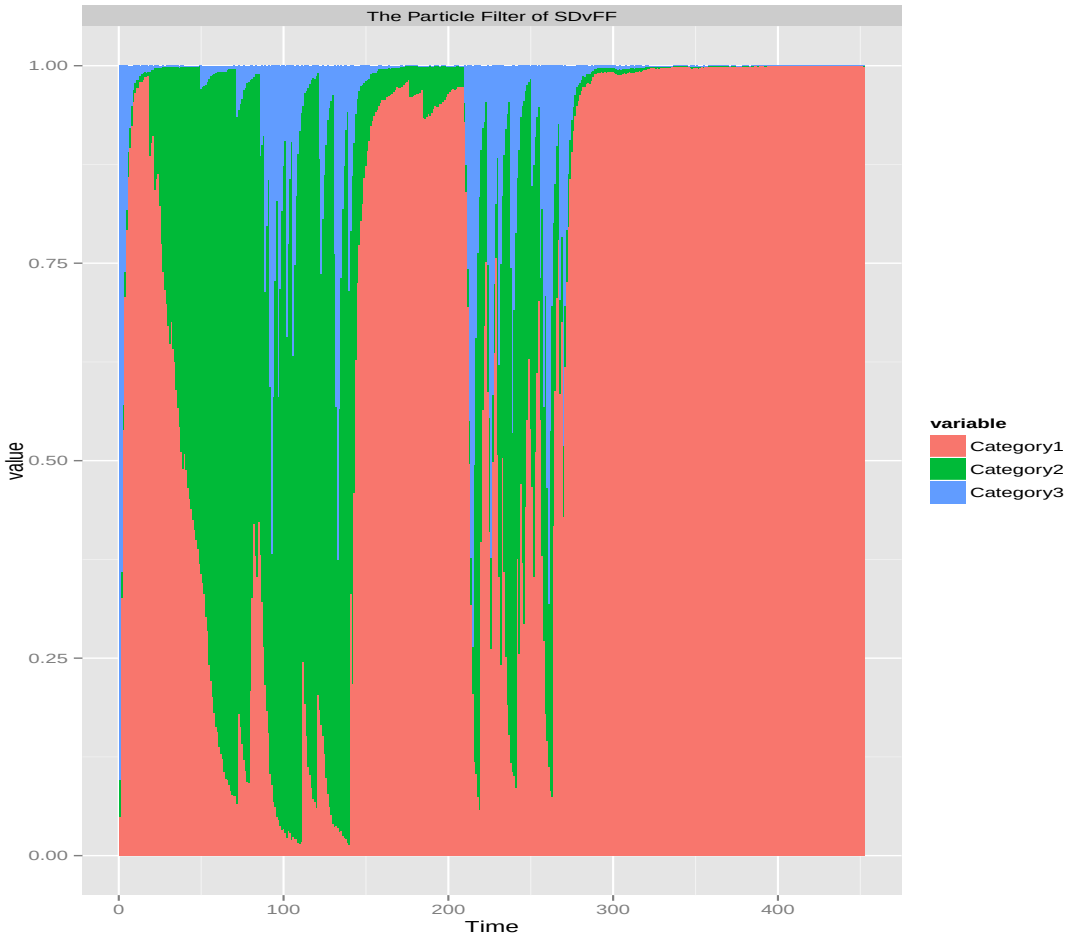

Figure 5.15: The posterior probabilities by applying the particle filter to multi-categorical time series of the counts, based on the values of $|\hat{B}_t|$ obtained from the DLM with the SDvFF

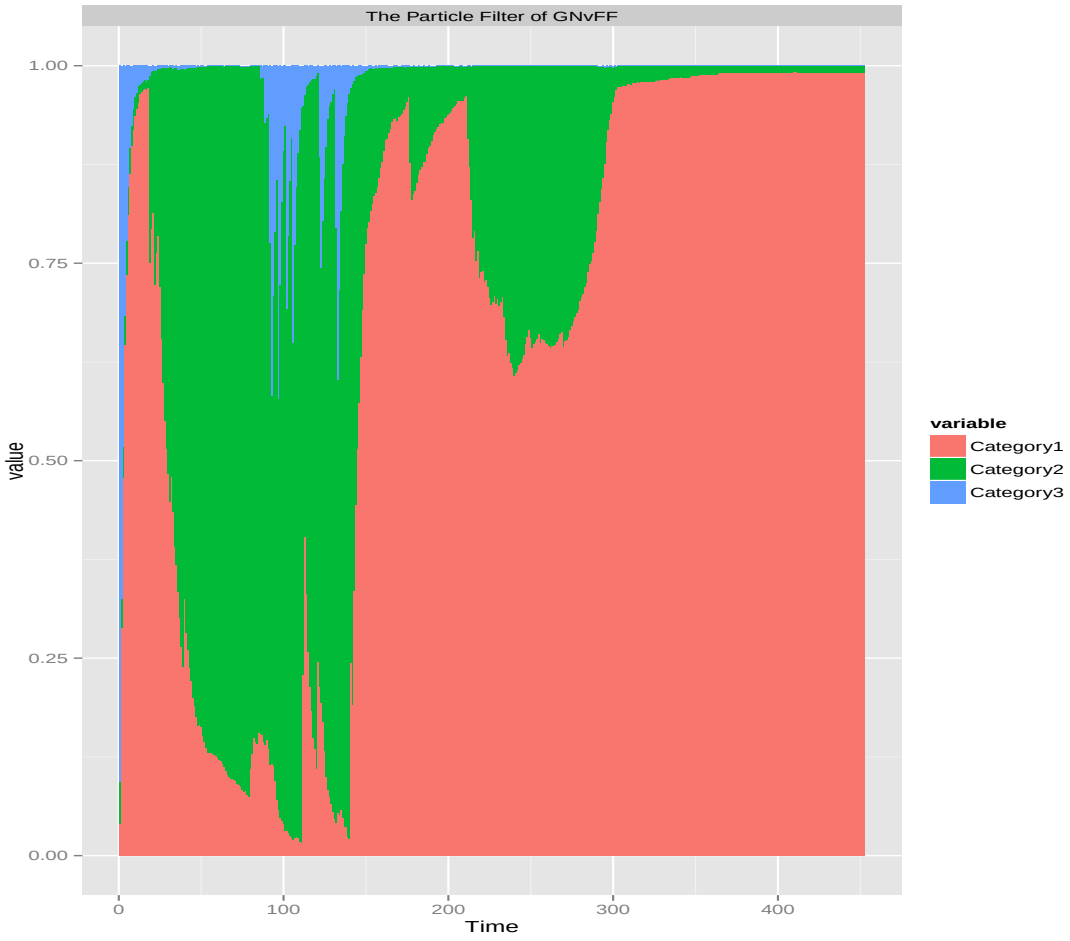

Figure 5.16: The posterior probabilities by applying the particle filter to multi-categorical time series of the counts, based on the values of $|\hat{B}_t|$ obtained from the DLM with the GNvFF

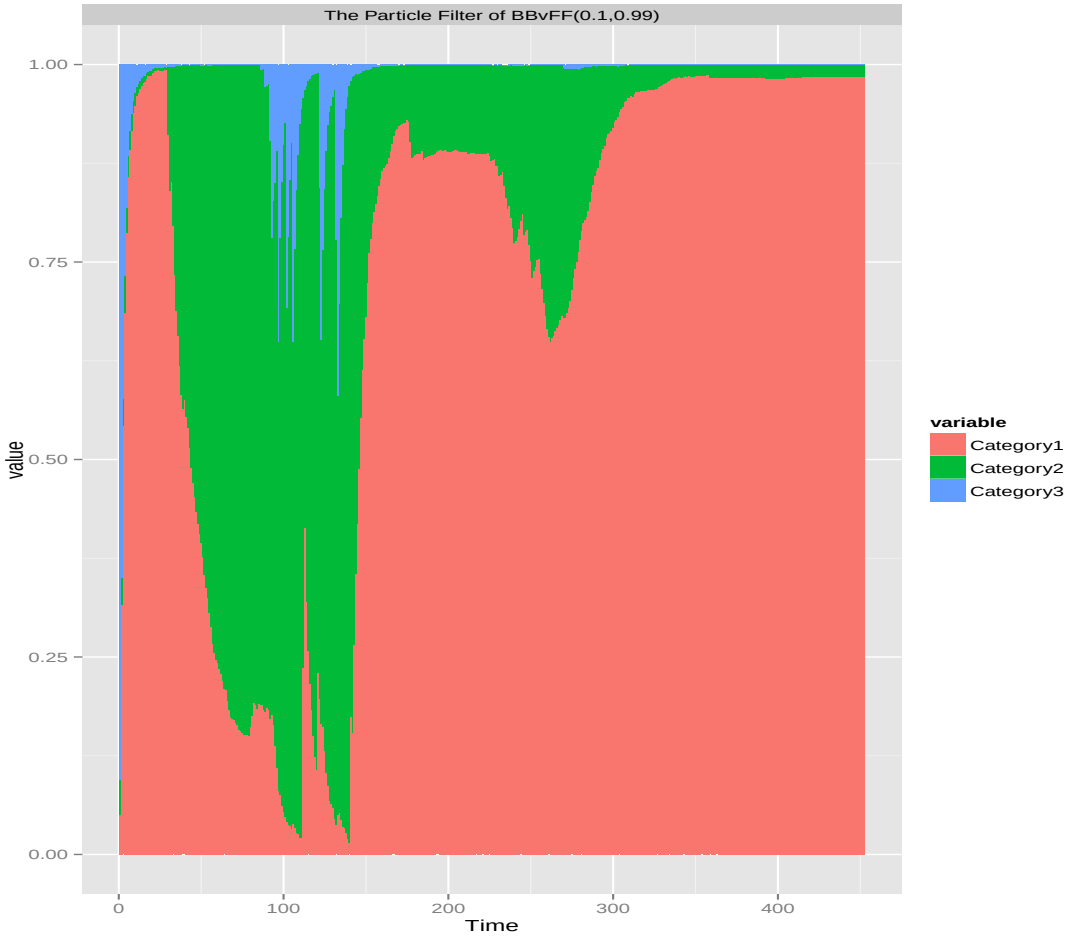

Figure 5.17: The posterior probabilities by applying the particle filter to multi-categorical time series of the counts, based on the values of $|\hat{B}_t|$ obtained from the DLM with the BBvFF(0.1,0.99)

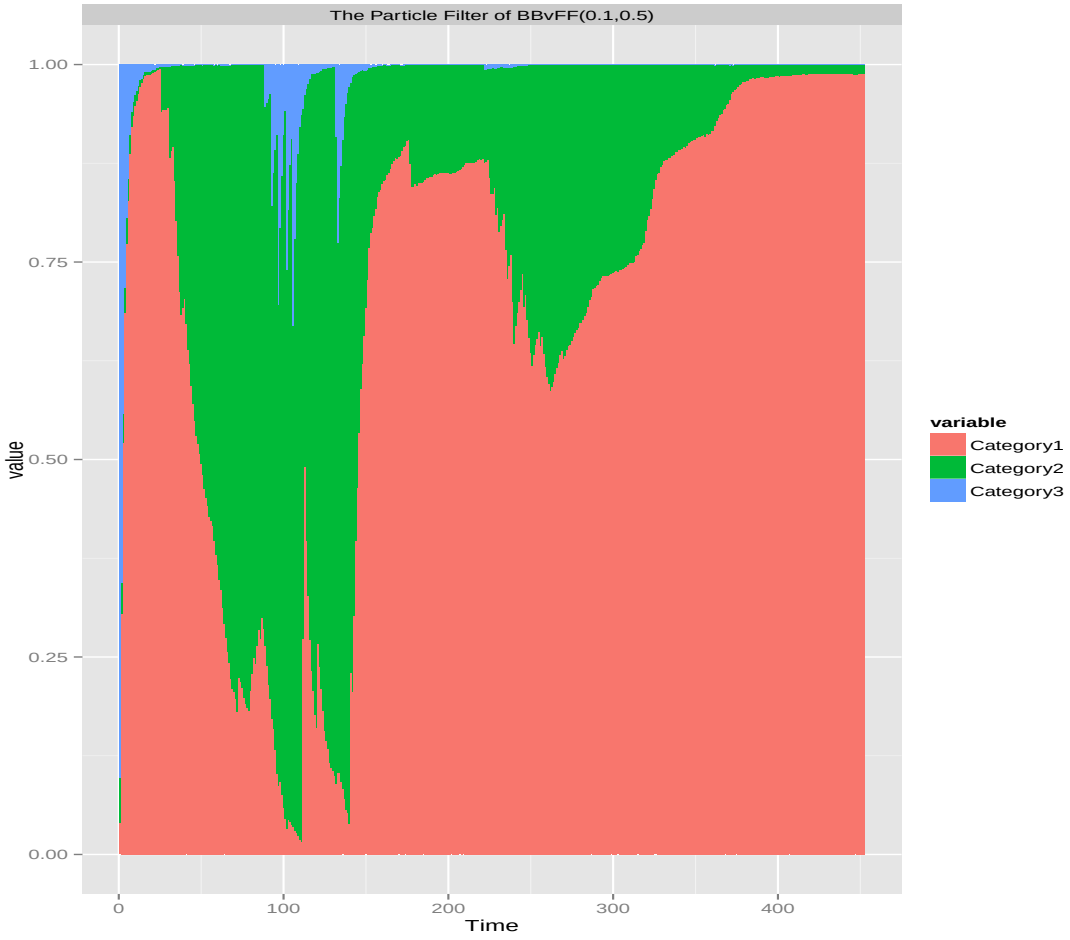

Figure 5.18: The posterior probabilities by applying the particle filter to multi-categorical time series of the counts, based on the values of $|\hat{B}_t|$ obtained from the DLM with the BBvFF(0.1,0.5)

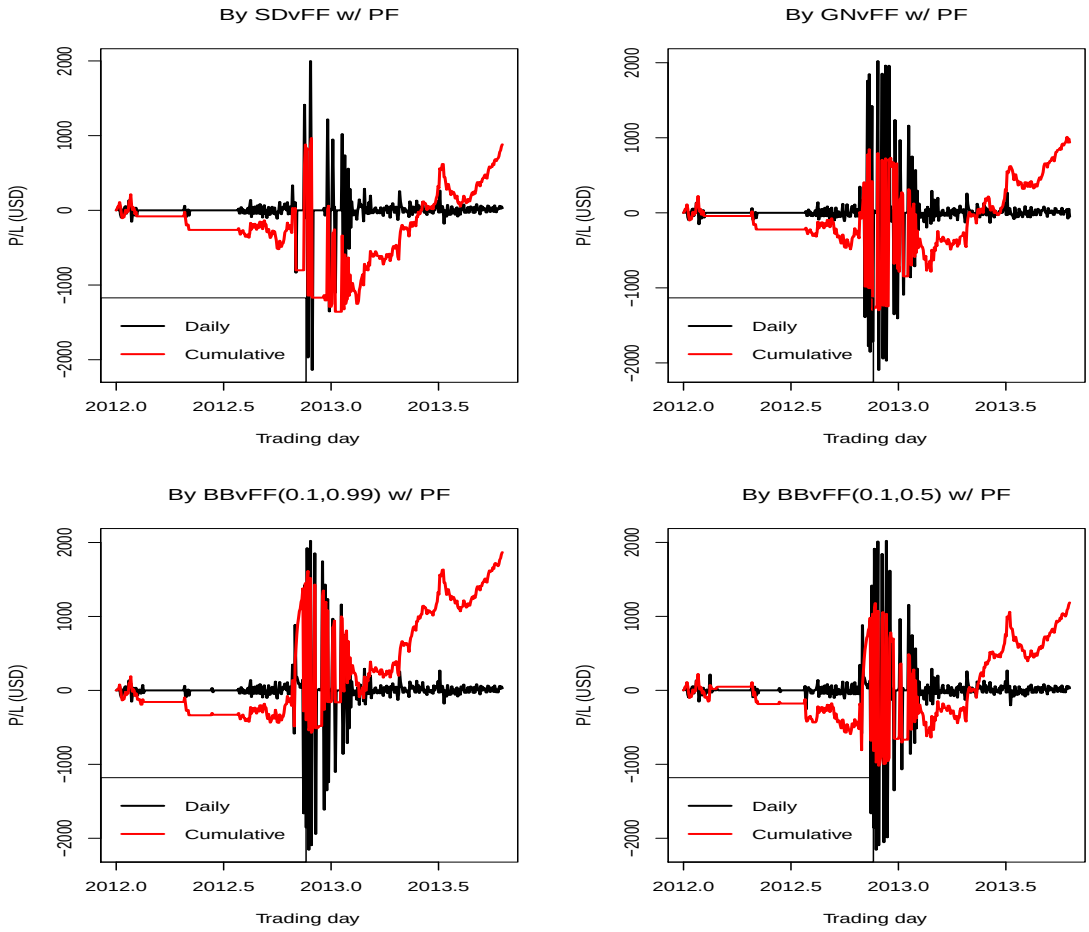

Figure 5.19: Case B with trading rule 1: Trading results by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

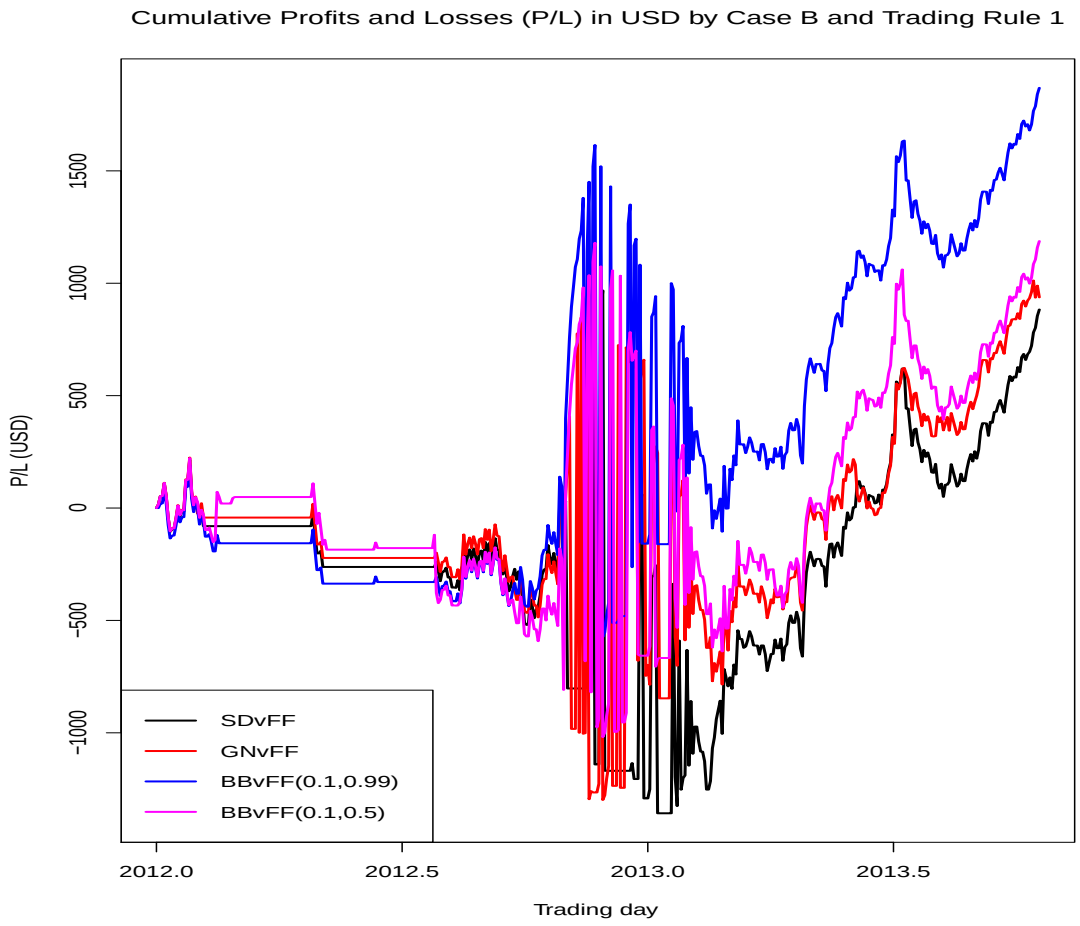

Figure 5.20: Case B with trading rule 1: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

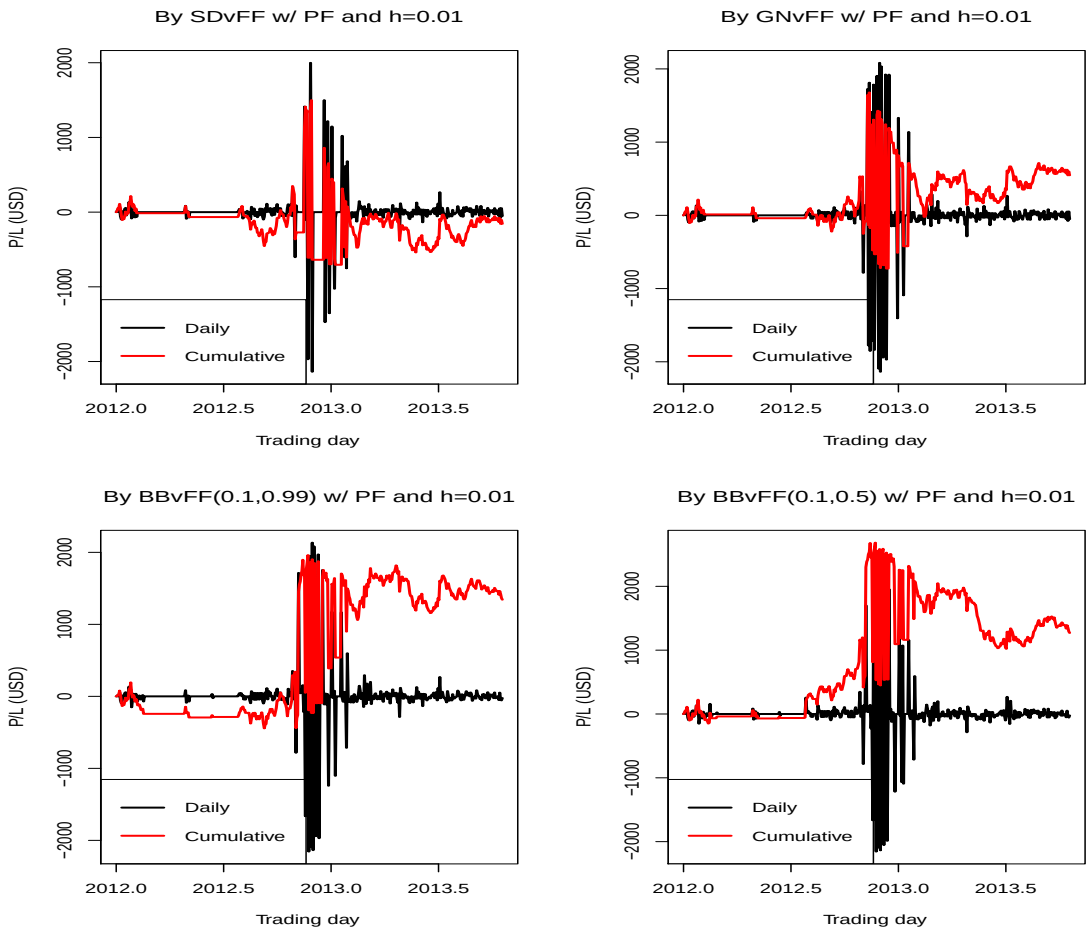

Figure 5.21: Case A with trading rule 2 with margin of 0.01: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

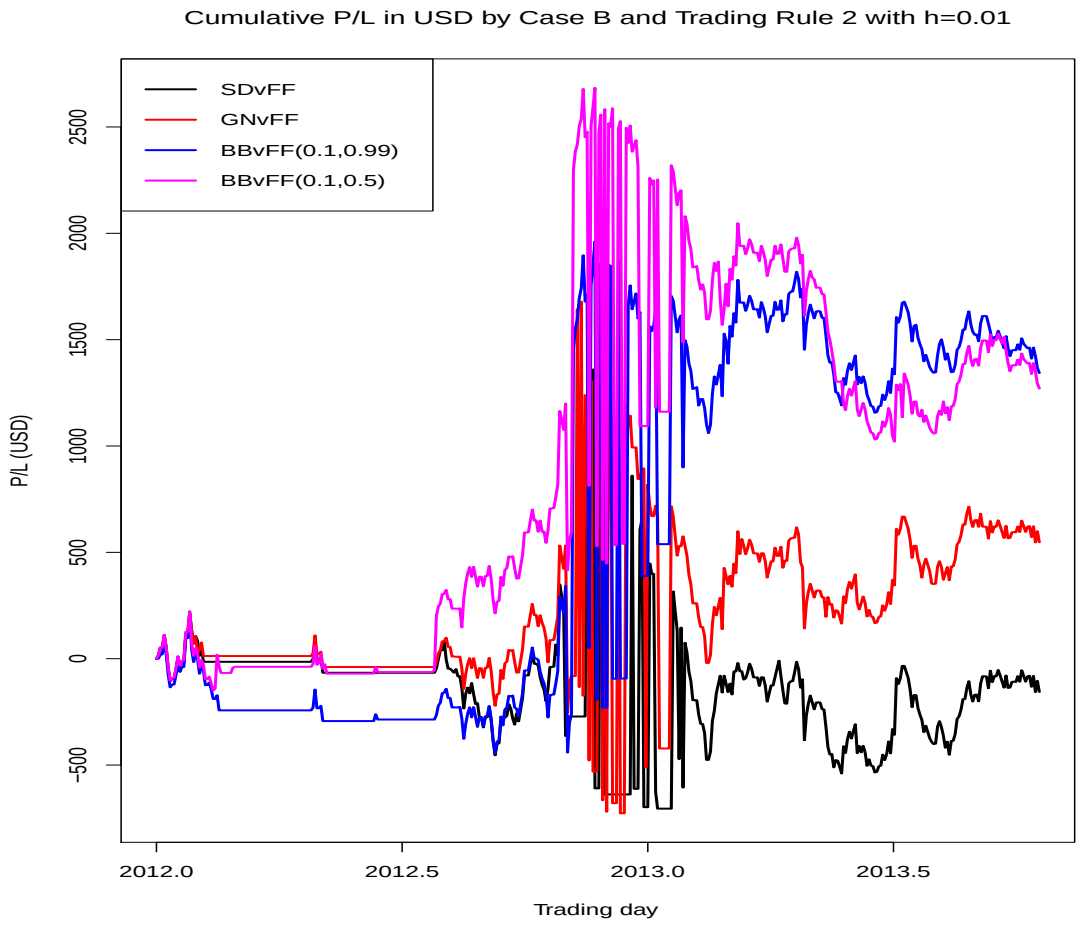

Figure 5.22: Case A with trading rule 2 with margin of 0.01: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

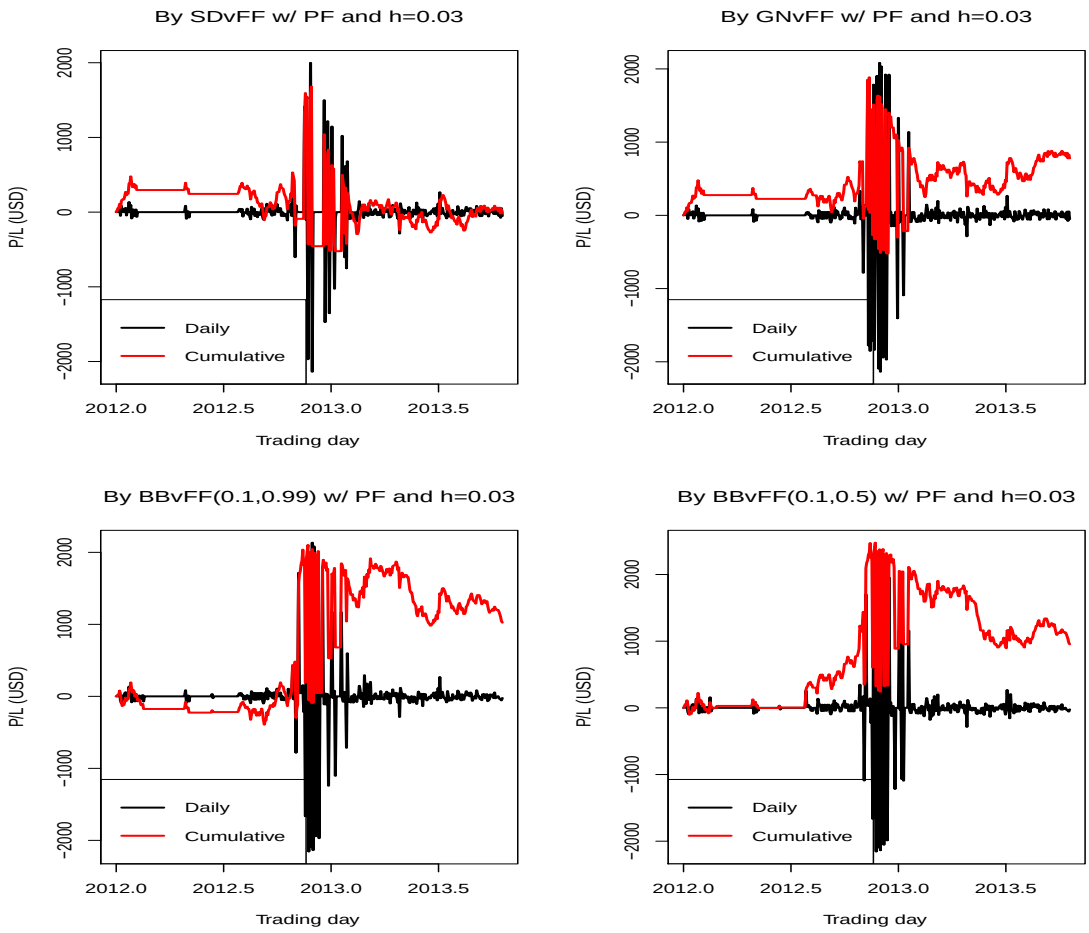

Figure 5.23: Case A with trading rule 2 with margin of 0.03: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

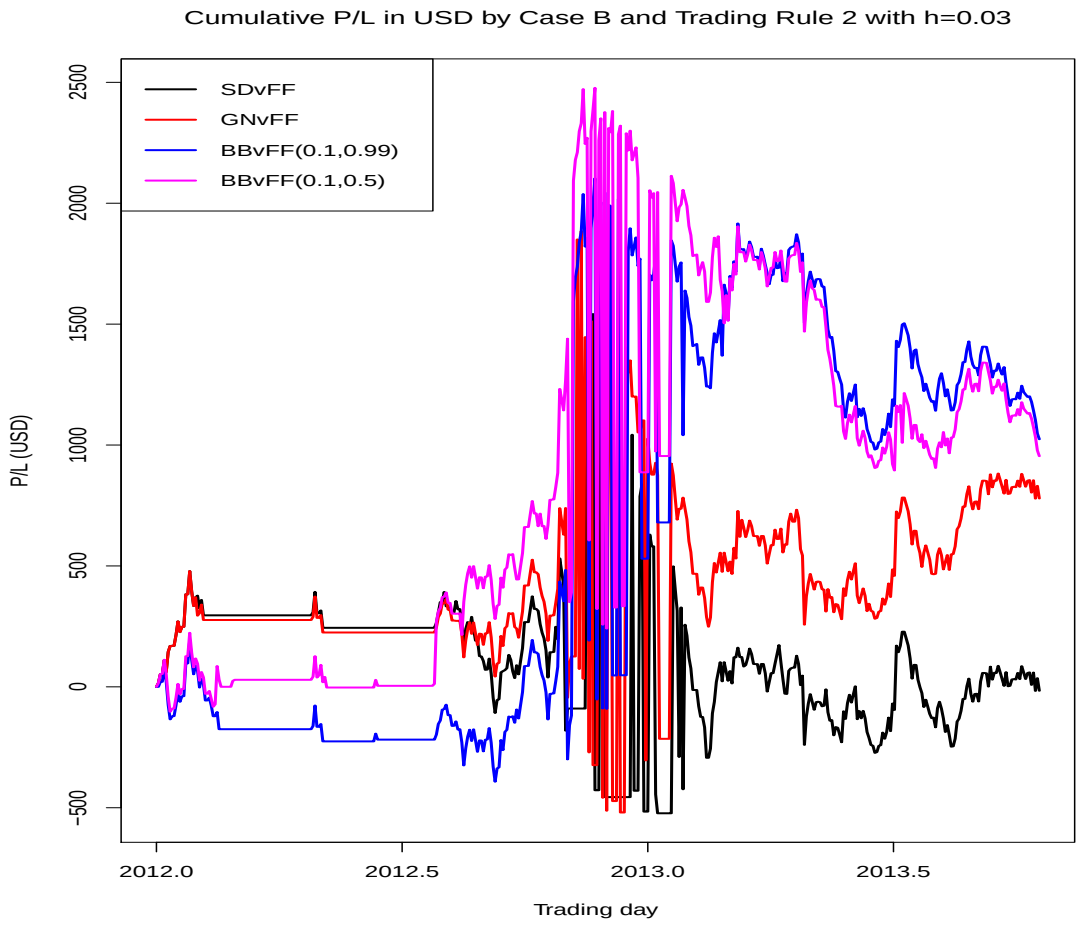

Figure 5.24: Case A with trading rule 2 with margin of 0.03: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

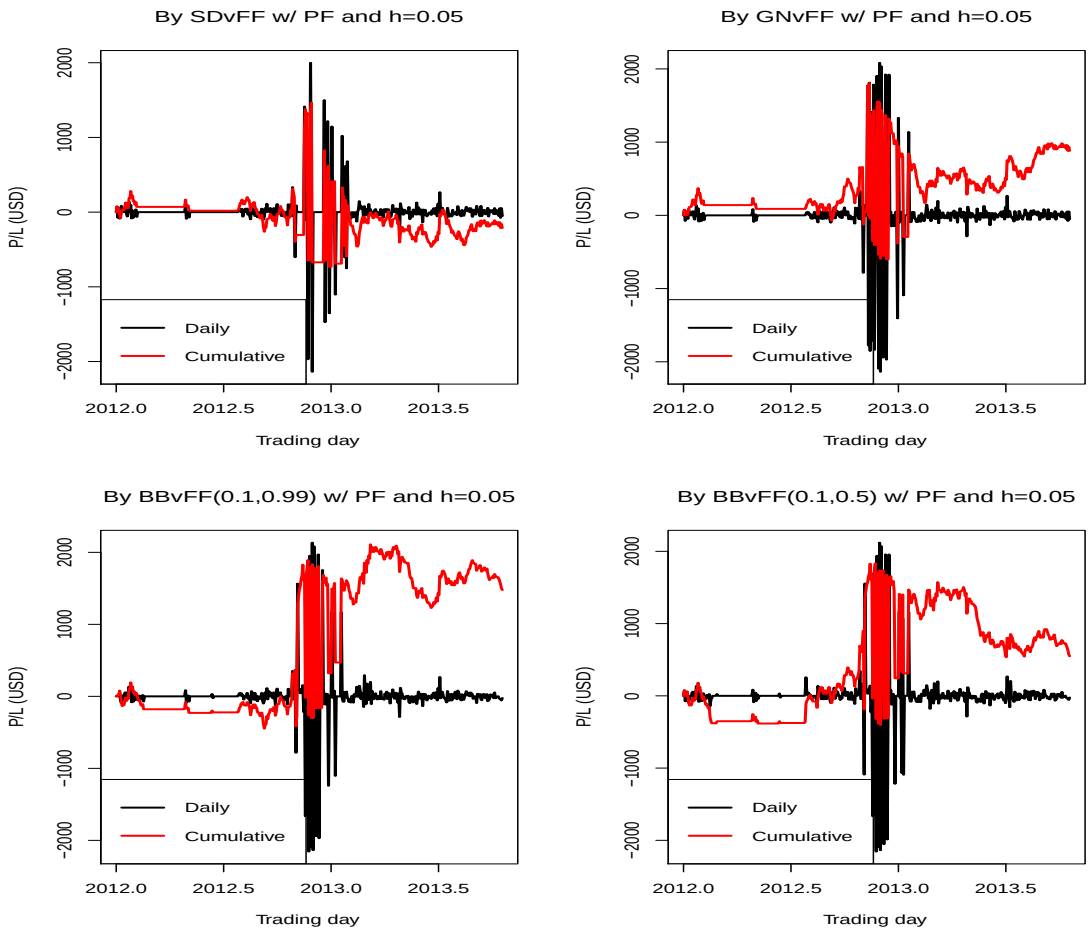

Figure 5.25: Case A with trading rule 2 with margin of 0.05: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

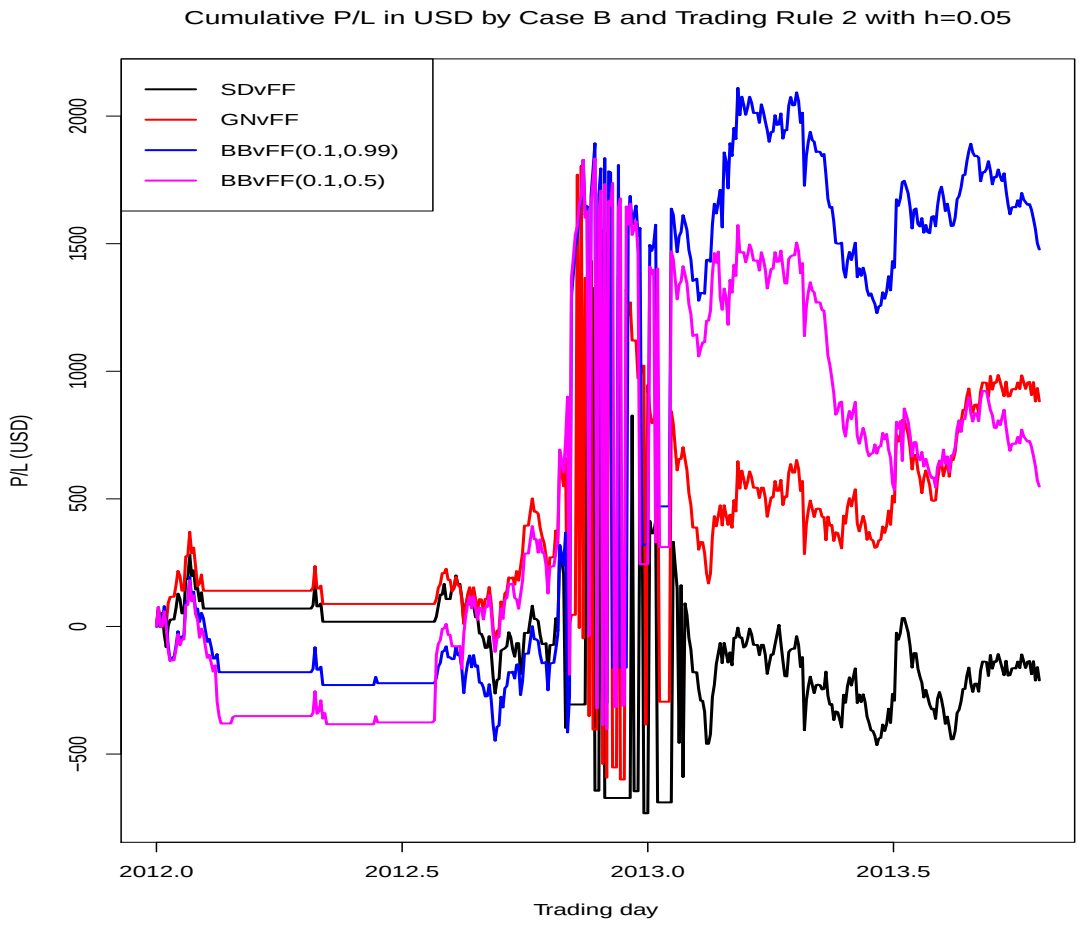

Figure 5.26: Case A with trading rule 2 with margin of 0.05: Cumulative profits and losses (P/L) comparison by the DLM with each of the SDvFF, the GNvFF, the BBvFF(0.1,0.99), and the BBvFF(0.1,0.5)

Chapter 6

Conclusions

This thesis is concerned with online detection of mean-reversion in algorithmic pairs trading. In pairs trading, a pair of assets is chosen when their prices are expected to show similar movements. Examples can be a pair of stocks from the same industry, an index and an exchange-traded-fund (ETF) tracking the index, an ETF and the asset held by the ETF, and so on. Although pairs trading is heavily dependent on the assumption of mean-reversion of the spread, the mean-reversion of any pair does not hold for good. Thus, the mean-reversion is detected locally rather than globally.

Chapter 2 reviews literature on time series and Bayesian forecasting, pairs trading, the spread model in dynamic linear model, variable forgetting factor, and dynamic generalised linear model (DGLM). Assuming mean-reversion of the spread, the time-varying and non-stationary dynamics of the spread can be implemented in a time-varying autoregressive model of order 1 (TV-AR(1)), represented in state space model. As the spread model, a TV-AR(1) with constant forgetting factor is proposed in dynamic linear model by Triantafyllopoulos and Montana (2011). According to the conditions for mean-reversion or a state of the model by Triantafyllopoulos and Montana (2011), the detection of mean-reversion in dynamic linear model depends on the value of autoregressive coefficient B_t.

Chapter 3 introduces the variability of forgetting factor and two algorithms for variable forgetting factor from the field of signal processing and control engineering using two widely applied methods of the steepest descent method and the Gauss-Newton

method. Also, the beta-Bernoulli variable forgetting factor algorithm (BBvFF), named after the conjugacy of beta and Bernoulli distributions, is developed and proposed. For Monte Carlo simulation, several sets of time series are generated and iterated for 1,000 times where the assessment is based on the mean square error. The results show that the BBvFF$(0.1, k)$ outperforms the other variable forgetting factor algorithms.

In algorithmic pairs trading, in addition to the detection of mean-reversion and the forecast of the spread, the trading rules also need to be implemented and executed online. As discussed earlier in Chapter 2 and 3, the detection of mean-reversion of the spread at time t solely relies on the value of $|\hat{B}_t|$ at that time. Whatever happens before and after the time t, the algorithm detects mean-reversion as long as $|\hat{B}_t| < 1$. A trader or an investor may think that this is too dangerous to take the risks of algorithmic pairs trading. In particular, when the spread shows volatile movements, algorithmic pairs trading may end up with huge loss, and there is no way to avoid the extreme. Thus, we find the need to monitor the behaviour of $|\hat{B}_t|$, slicing the range which $|\hat{B}_t|$ can be located into categories. For this, in Chapter 4, DGLM for multi-categorical time series is developed and the states are approximated by the moments.

In Chapter 4, DGLM for multi-categorical time series is developed, and recursions are based on the approximated moments using the linear Bayes estimates and sequential Monte Carlo methods. West et al. (1985) show the approximation for DGLM of a univariate time series, using Bayes linear methods. In Chapter 4, the approximation for DGLM using the linear Bayesian methods is extended to a multivariate case. Assuming the multivariate normal distribution for the states in the evolution model, the particle filters are applied to multi-categorical time series for approximation of the posterior distribution of the states by the moments. The particle filter using the optimal importance kernel is shown to outperform the bootstrap filter using the prior distribution as the importance density. It proves that the importance density plays a key role in importance sampling, and the optimal importance kernel is a better choice.

In Chapter 5, an opportunity by algorithmic pairs trading is proposed, applying

the methodologies developed in Chapter 3 and 4. An illustration aims to show how the algorithms from the previous chapters can be applied for the algorithmic pairs trading. It is shown that algorithmic pairs trading can be successful even with simple trading rules. For an illustration of algorithmic pairs trading, a pair of stocks, Agnico-Eagle Mines Limited and Newmont Mining Corporation, listed on New York Stock Exchange (NYSE), are presented. Two different decision rules are considered of Case A and Case B. Case A considers only the condition for the mean-reversion by Triantafyllopoulos and Montana (2011) while Case B takes into account of the online monitoring process. Trading Rule 1 compares Y_t and μ_t to see which of a pair to buy and short-sell while Trading Rule 2 uses Y_t and f_{t+1} with a prediction margin h. In Case A, the BBvFF(0.1,0.99) produces the highest daily earnings at USD 4.886 on average by Trading Rule 2 with $h = 0.01$ while the BBvFF(0.1,0.99) earns USD 3.484 by Trading Rule 1 in Case B. In both cases, the BBvFF(0.1,0.99) is found to be the most profitable algorithm.

In DGLM for multi-categorical time series developed in Chapter 4, the evolution of the states is assumed to be a random walk where G_t is set up to be an identity matrix I. However, in some other applications, the evolution model of the states can imply the components of the states such as trend, seasonality, and irregular variation. For further research, these components of G_t can be treated as parameters to estimate. With regard to algorithmic pairs trading, how long the detected mean-reversion lasts can be a topic for research. Others may be how to select a pair or a portfolio of the pairs, and more sophisticated trading strategy to guarantee the earnings.

Appendix A

A.1 Generating Functions

By definition, the generating function is such that $f(x) = \sum_{i=0}^{\infty} a_i x^i$ for the sequence of $a_0, a_1, \cdots$. The generating function $f(x)$ transforms problems about the sequences into problems of the function. For example, the sequence of $a_0, a_1, \cdots$ are generated by the generating function $f(x)$ and the function $f(x)$ has a_n as the coefficients of x^n in its power series representation. For example, suppose that $f(x) = (1-x-x^2)^{-1}$. This function $f(x)$ can be expanded as a power series so that $f(x) = 1+x+2x^2+3x^3+5x^4+8x^5+13x^6+\cdots$ where the sequence of $a_0, a_1, a_2, \cdots$ becomes $a_0 = 1$, $a_1 = 1$, $a_2 = 2$ and so on.

If a univariate and discrete random variable of interest is X, the moment generating function is defined by $M_X(t) = \mathrm{E}(e^{tX})$ and the cumulant generating function by $K_X(t) = \log\{M_X(t)\}$. Both of them can be represented by the power series expansion.

For the moment generating function,

$$\begin{aligned} M_X(t) &= \mathrm{E}(e^{tX}) \\ &= \sum_{n=0}^{\infty} \mathrm{E}(X^n)\frac{t^n}{n!} \quad \text{by power series expansion} \end{aligned}$$

By differentiation, the moments about the origin can be found as

$$M_X^{(r)}(0) = \mathrm{E}(X^r) = \left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0}$$

For the cumulant generating function,

$$\begin{aligned} K_X(t) &= \log\{M_X(t)\} \\ &= \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} \quad \text{by power series expansion} \end{aligned}$$

where κ_n, the coefficients of $\frac{t^n}{n!}$ is called as the cumulant of X.

The moments about the origin can be found by differentiation as

$$K_X^{(r)}(0) = \kappa_r = \left.\frac{d^r K_X(t)}{dt^r}\right|_{t=0}$$

where κ_r is called as the cumulant, having the first and the second moment of a distribution for a univariate random variable X as $\kappa_1 = \left.\frac{dK_X(t)}{dt}\right|_{t=0} = \mathrm{E}(X)$ and $\kappa_2 = \left.\frac{d^2K_X(t)}{dt^2}\right|_{t=0} = \mathrm{Var}(X)$.

Comparing the coefficients of $\left(\frac{t^n}{n!}\right)$ in both the moment generating function and the cumulant generating function shows the basic relationship between the two as

$$\begin{aligned} \kappa_1 &= M_X^{(1)}(t) = \mathrm{E}(X) \\ \kappa_2 &= M_X^{(2)}(t) - \kappa_1^2 = \mathrm{Var}(X) \end{aligned}$$

Thus, we aim to find κ_1 and κ_2.

A.2 Digamma and Trigamma Functions

By definition, $\Gamma(z+1) = z!$ and $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}dt$ in Euler's integral.

$\Gamma(z)$ can be approximated by Stirling's formula as

$$\Gamma(z) \approx \mathrm{e}^{-z}z^{z-1/2}(2\pi)^{1/2}\left(1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{2488320z^4}+\cdots\right)$$

as $z \to \infty$ in $|arc(z)| < \pi$.

In Abramowitz and Stegun (1965), $\Psi(z)$ and $\Psi^{(1)}(z)$ is referred to as a digamma function and a trigamma function each, and defined as $\Psi(z) = \frac{d}{dz}\{\log\Gamma(z)\}$ and $\Psi^{(1)}(z) = \frac{d}{dz}\left\{\Psi(z) = \frac{d^2}{dz^2}\log\Gamma(z)\right\}$ respectively.

From an approximation by Stirling's formula of $\Gamma(z)$, $\log\Gamma(z)$ can be written by

$$\begin{aligned}\log\Gamma(z) &\approx -z+(z-1/2)\log(z)+\log(2\pi)^{1/2} \\ &\quad +\log\left(1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{2488320z^4}+\cdots\right)\end{aligned}$$

Thus, a digamma and a trigamma function are obtained as

$$\begin{aligned}\Psi(z) &= \frac{d}{dz}\{\log\Gamma(z)\} \\ &\approx \log(z)-\frac{1}{2z}+\cdots \qquad (1) \\ \Psi^{(1)}(z) &= \frac{d}{dz}\Psi(z) = \frac{d^2}{dz^2}\{\log\Gamma(z)\} \\ &\approx \frac{1}{z}+\frac{1}{2z^2}+\cdots \qquad (2)\end{aligned}$$

As an approximation of a digamma and a trigamma function, $\Psi(z) \approx \log(z)$ and $\Psi^{(1)}(z) \approx \frac{1}{z}$ are adopted in this thesis.

A.3 Distributions

A.3.1 Continuous Distribution

A.3.1.1 Beta Distribution

The beta distribution is defined on the interval $[0, 1]$ and commonly used as the conjugate prior distribution for the binomial probability in Bayesian inference. The probability density function of the beta distribution is characterised by two positive shape parameters, usually denoted by α and β. When a random variable X is distributed by the beta distribution with parameters of α and β, it is denoted by $X \sim \text{Beta}(\alpha, \beta)$ or $p(X) = \text{Beta}(X \mid \alpha, \beta)$. If $\alpha = \beta = 1$, the distribution becomes the standard uniform distribution. In Bayesian statistics, the beta distribution with $\alpha = \beta = 0.5$ is sometimes used as a noninformative prior density. The density function is

$$p(X) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} X^{\alpha-1}(1 - X)^{\beta-1}$$

where $\alpha > 0$, $\beta > 0$, and $X \in [0, 1]$. The mean, the variance, and the mode are $\text{E}(X) = \frac{\alpha}{\alpha+\beta}$, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, and $\text{mode}(X) = \frac{\alpha-1}{\alpha+\beta-2}$ respectively.

A.3.1.2 Dirichlet Distribution

The Dirichlet distribution is a multivariate generalisation of the beta distribution, named after Johann Peter Gustav Lejeune Dirichlet, a German mathematician. It is commonly used as the the conjugate prior distribution for the parameters of the multinomial distribution in Bayesian inference. The probability density function of the Dirichlet distribution is

$$p(\mathbf{X}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} X_1^{\alpha_1-1} \cdots X_k^{\alpha_k-1}$$

where $\mathbf{X} = (X_1, \ldots, X_k)$, $x_1, \ldots, X_k \geq 0$ with $\sum_{i=1}^{k} X_i = 1$, and $\alpha_i > 0$. It is denoted by $\mathbf{X} \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ or $p(\mathbf{X}) = \text{Dirichlet}(\mathbf{X} \mid \alpha_1, \ldots, \alpha_k)$. The

mean, the variance, the covariance, and the mode are respectively

$$\begin{aligned}
\mathrm{E}(X_i) &= \frac{\alpha_i}{\alpha_0} \\
\mathrm{Var}(X_i) &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \\
\mathrm{Cov}(X_i, X_j) &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \\
\mathrm{mode}(X_i) &= \frac{\alpha_i - 1}{\alpha_0 - k}
\end{aligned}$$

where $\alpha_0 \equiv \sum_{i=1}^{k} \alpha_i$.

A.3.2 Discrete Distribution

A.3.2.1 Binomial Distribution

The binomial distribution represents the number of 'success' in n independent Bernoulli trials where π is a 'success' probability in each trial. Each trial belongs to one of two categories where a category is regarded commonly as a 'success' with a probability of π_1 and the other as a 'failure' with $\pi_2 = 1 - \pi_1$, having $\sum_{i=1}^{2} \pi_i = 1$. The binomial distribution can be represented as $X \sim \mathrm{Bin}(n; \pi_1, 1 - \pi_1)$ or $p(X) = \mathrm{Bin}(X \mid n; \pi_1)$ where n is the number of Bernoulli trials and a positive integer.

The probability mass function of the binomial distribution is

$$\begin{aligned}
p(X \mid \pi_1) &= \left(\begin{array}{c} n \\ X \end{array} \right) \pi_1^X \cdot (1 - \pi_1)^{n-X} \\
&= \frac{n!}{X! \cdot (n - X)!} \pi_1^X \cdot (1 - \pi_1)^{n-X}
\end{aligned}$$

where X is the number of 'success' in n independent Bernoulli trials, $X \in \{0, 1, \ldots, n\}$, and π_1 is a 'success' probability in each trial with $1 - \pi_1$ as a 'failure' probability, satisfying $0 \leq \pi_1 < 1$. The mean and the variance is $\mathrm{E}(X) = n \cdot \pi_1$ and $\mathrm{Var}(X) = n \cdot \pi_1 \cdot (1 - \pi_1)$ respectively.

A.3.2.2 Multinomial Distribution

The multinomial distribution is a multivariate generalisation of the binomial distribution. Assuming that there are n independent trials and k categories, each trial belongs to one of k categories where each category has a 'success' probability of $\mathbf{\Pi} = (\pi_1, \ldots, \pi_k)$ and $\sum_{i=1}^{k} \pi_i = 1$. Suppose that $\mathbf{X} = (X_1, \ldots, X_k)$ and X_i represents number of events for a category i in n trials with a probability π_i where $X_i \in \{0, 1, \ldots, n\}$ and $\sum_{i=1}^{k} X_i = n$. The multinomial distribution can be represented as $\mathbf{X} \sim \text{Multin}(n; \pi_1, \ldots, \pi_k)$ or $p(\mathbf{X}) = \text{Multin}(\mathbf{X} \mid n; \pi_1, \ldots, \pi_k)$ where n is a positive integer.

The probability mass function of the multinomial distribution is

$$\begin{aligned} p(\mathbf{X} \mid \mathbf{\Pi}) &= \begin{pmatrix} & n & \\ X_1 & \cdots & X_k \end{pmatrix} \pi_1^{X_1} \cdots \pi_k^{X_k} \\ &= \frac{n!}{X_1! \cdots X_k!} \pi_1^{X_1} \cdots \pi_k^{X_k} \end{aligned}$$

where $\mathbf{X} = (X_1, \ldots, X_k)$, $X_i \in \{0, 1, \ldots, n\}$, $\sum_{i=1}^{k} X_i = n$ and $\mathbf{\Pi} = (\pi_1, \ldots, \pi_k)$ with $\sum_{i=1}^{k} \pi_i = 1$ satisfying $0 \leq \pi_i < 1$. The mean and the variance is $\text{E}(X_i) = n \cdot \pi_i$ and $\text{Var}(X_i) = n \cdot \pi_i \cdot (1 - \pi_i)$ respectively.

The moment generating function is

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= M_{\mathbf{X}}(t_1, \ldots, t_k) = \text{E}\left\{ \exp\left(\sum_{i=1}^{k} t_i \cdot X_i \right) \right\} \\ &= \left\{ \sum_{i=1}^{k} \pi_i \cdot e^{t_i} \right\}^n \end{aligned}$$

The cumulant generating function is

$$\begin{aligned} K_{\mathbf{X}}(\mathbf{t}) &= K_{\mathbf{X}}(t_1, \ldots, t_k) = \log\left\{M_{\mathbf{X}}(t_1, \ldots, t_k)\right\} \\ &= n \cdot \log\left\{\sum_{i=1}^{k} \pi_i \cdot e^{t_i}\right\} \end{aligned}$$

Appendix B

B.1 The Derivations of The Updating Equations in Dynamic Linear Model

This appendix is for Section 2.5 of Chapter 2, Section 3.2 of Chapter 3, and Section 4.2 of Chapter 4. However, the derivations are based on the specification for a univariate time series in Section 2.5 of Chapter 2, assuming the Gaussianity.

B.1.1 When $\tau = \frac{1}{V}$ is known as a constant V

Supposing that **(a1)** is true, **(a2)** holds from (2.6) as shown by

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\theta}_t \mid V, D_{t-1}) &= \mathrm{E}(G\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \mid V, D_{t-1}) \;=\; G\mathrm{E}(\boldsymbol{\theta}_{t-1} \mid V, D_{t-1}) \\
&= G\mathbf{m}_{t-1} \;=\; \mathbf{a}_t \\
\mathrm{Var}(\boldsymbol{\theta}_t \mid V, D_{t-1}) &= \mathrm{Var}(G\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \mid V, D_{t-1}) \\
&= G\mathrm{Var}(\boldsymbol{\theta}_{t-1} \mid V, D_{t-1})G' + \mathrm{Var}(\boldsymbol{\omega}_t \mid V, D_{t-1}) \\
&= GVC^*_{t-1}G' + VW^*_t \;=\; V(GC^*_{t-1}G' + W^*_t) \\
&= VR^*_t
\end{aligned}$$

Therefore, we can say that $(\boldsymbol{\theta}_t \mid V, D_{t-1}) \sim N_2(\mathbf{a}_t, VR^*_t)$ as in **(a2)**.

Similarly, **(a3)** holds from (2.5) and (2.6) as shown by

$$\begin{aligned}
\mathrm{E}(Y_t \mid V, D_{t-1}) &= \mathrm{E}(\mathbf{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t \mid V, D_{t-1}) \;=\; \mathrm{E}(\mathbf{F}_t'G\boldsymbol{\theta}_{t-1} + \mathbf{F}_t'\boldsymbol{\omega}_t + \boldsymbol{\nu}_t \mid V, D_{t-1}) \\
&= \mathbf{F}_t'G\mathrm{E}(\boldsymbol{\theta}_{t-1} \mid V, D_{t-1}) + \mathbf{F}_t'\mathrm{E}(\boldsymbol{\omega}_t \mid V, D_{t-1}) + \mathrm{E}(\boldsymbol{\nu}_t \mid V, D_{t-1}) \\
&= \mathbf{F}_t'G\mathbf{m}_{t-1} \;=\; \mathbf{f}_t
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(Y_t \mid V, D_{t-1}) &= \mathrm{Var}(\mathbf{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t \mid V, D_{t-1}) \;=\; \mathrm{Var}(\mathbf{F}_t'G\boldsymbol{\theta}_{t-1} + \mathbf{F}_t'\boldsymbol{\omega}_t + \boldsymbol{\nu}_t \mid V, D_{t-1}) \\
&= \mathbf{F}_t'G\mathrm{Var}(\boldsymbol{\theta}_{t-1} \mid V, D_{t-1})G'\mathbf{F}_t + \mathbf{F}_t'\mathrm{Var}(\boldsymbol{\omega}_t \mid V, D_{t-1})\mathbf{F}_t \\
&\quad + \mathrm{Var}(\boldsymbol{\nu}_t \mid V, D_{t-1}) \\
&= \mathbf{F}_t'GVC_{t-1}^*G'\mathbf{F}_t + \mathbf{F}_t'VW_t^*\mathbf{F}_t + V \\
&= V(\mathbf{F}_t'GC_{t-1}^*G'\mathbf{F}_t + \mathbf{F}_t'W_t^*\mathbf{F}_t + 1) \\
&= V\{\mathbf{F}_t'(GC_{t-1}^*G' + W_t^*)\mathbf{F}_t + 1\} \\
&= V(\mathbf{F}_t'R_t^*\mathbf{F}_t + 1) \\
&= VQ_t^*
\end{aligned}$$

Now it is seen that $(Y_t \mid V, D_{t-1}) \sim N_2(\mathbf{f}_t, VQ_t^*)$ as in **(a3)**.

For **(a4)**, it can be proved either using the normal distribution theory or using Bayes' theorem. Firstly, using the normal distribution theory, the joint distribution of $(Y_t, \boldsymbol{\theta}_t \mid D_{t-1})$ at time t is calculated and then the conditional distribution of

$(\boldsymbol{\theta}_t \mid Y_t, D_{t-1})$ is derived from the joint distribtuion of $(Y_t, \boldsymbol{\theta}_t \mid D_{t-1})$.

$$\begin{aligned}
\mathrm{Cov}(Y_t, \boldsymbol{\theta}_t \mid D_{t-1}) &= \mathrm{Cov}(\mathbf{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \boldsymbol{\theta}_t \mid D_{t-1}) \\
&= \mathbf{F}_t'\mathrm{Cov}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t \mid D_{t-1}) + \mathrm{Cov}(\boldsymbol{\nu}_t, \boldsymbol{\theta}_t \mid D_{t-1}) \\
&= \mathbf{F}_t'\mathrm{Var}(\boldsymbol{\theta}_t \mid D_{t-1}) \;=\; \mathbf{F}_t'(VR_t^*) \\
&= V\mathbf{F}_t'R_t^* \\
\mathrm{Cov}(\boldsymbol{\theta}_t, Y_t \mid D_{t-1}) &= \mathrm{Cov}(\boldsymbol{\theta}_t, \mathbf{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t \mid D_{t-1}) \\
&= \mathrm{Cov}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t \mid D_{t-1})\mathbf{F}_t + \mathrm{Cov}(\boldsymbol{\theta}_t, \boldsymbol{\nu}_t \mid D_{t-1}) \\
&= \mathrm{Var}(\boldsymbol{\theta}_t \mid D_{t-1})\mathbf{F}_t + 0 \\
&= VR_t^*\mathbf{F}_t
\end{aligned}$$

With the results above, the joint distribution of $(Y_t, \boldsymbol{\theta}_t \mid D_{t-1})$ at time t is

$$\begin{pmatrix} Y_t \\ \boldsymbol{\theta}_t \end{pmatrix} \;\middle|\; D_{t-1} \sim N_2\left[\begin{pmatrix} \mathbf{f}_t \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} VQ_t^* & V\mathbf{F}_t'R_t^* \\ VR_t^*\mathbf{F}_t & VR_t^* \end{pmatrix}\right] \tag{3}$$

From (3), the conditional distribution of $(\boldsymbol{\theta}_t \mid Y_t, D_{t-1})$, or $(\boldsymbol{\theta}_t \mid D_t)$, can be found as

$$(\boldsymbol{\theta}_t \mid D_t) \sim N_2(\mathbf{m}_t, VC_t^*)$$

$$\begin{aligned}
\text{where} \quad \mathbf{m}_t &= \mathbf{a}_t + VR_t^*\mathbf{F}_t(VQ_t^*)^{-1}(Y_t - \mathbf{f}_t) \\
&= \mathbf{a}_t + K_t e_t \\
&\quad \text{and} \\
VC_t^* &= VR_t^* - VR_t^*\mathbf{F}_t(VQ_t^*)^{-1}V\mathbf{F}_t'R_t^* \\
&= VR_t^* - VR_t^*\mathbf{F}_t(VQ_t^*)^{-1}V\mathbf{F}_t'R_t^* \\
&= VR_t^* - VK_tQ_t^*K_t' \quad \text{by multiplying } \frac{VQ_t^*}{VQ_t^*} \text{ to the latter term}
\end{aligned}$$

K_t can be referred to as the regression matrix of $\boldsymbol{\theta}\mathbf{a}_t$ on Y_t and also as the Kalman gain while $C_t = R_t - K_tQ_tK_t'$ is a Riccati equation.

Secondly, using Bayes' theorem, **(a4)** also can be proved to hold as follows. The likelihood function of the observed series $\{Y_t\}$ is $(Y_t \mid \boldsymbol{\theta}_t, D_{t-1}) \sim N(\mathbf{F}_t'\boldsymbol{\theta}_t, V)$ from the observation equation (2.5) and from **(a2)** the prior distribution of $\boldsymbol{\theta}_t$ is $(\boldsymbol{\theta}_t \mid V, D_{t-1}) \sim N_2(\mathbf{a}_t, VR_t^*)$, which can be written up to proportionality as

$$\begin{aligned} p(Y_t \mid \boldsymbol{\theta}_t, D_{t-1}) &\propto \exp\left\{-\frac{(Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t)'(Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t)}{2V}\right\} \\ &\text{and} \\ p(\boldsymbol{\theta}_t \mid V, D_{t-1}) &\propto \exp\left\{-\frac{(\boldsymbol{\theta}_t - \mathbf{a}_t)'R_t^{*-1}(\boldsymbol{\theta}_t - \mathbf{a}_t)}{2V}\right\} \end{aligned}$$

Therefore, by Bayes' theorem,

$$\begin{aligned} p(\boldsymbol{\theta}_t \mid Y_t, D_{t-1}) &= \frac{p(\boldsymbol{\theta}_t \mid D_{t-1})p(Y_t \mid \boldsymbol{\theta}_t, D_{t-1})}{p(Y_t \mid D_{t-1})} \\ &\propto p(\boldsymbol{\theta}_t \mid D_{t-1})p(Y_t \mid \boldsymbol{\theta}_t, D_{t-1}) \quad \text{as a function of } \boldsymbol{\theta}_t \\ &\propto \exp\left\{-\frac{(Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t)'(Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t)}{2V} - \frac{(\boldsymbol{\theta}_t - \mathbf{a}_t)'R_t^{*-1}(\boldsymbol{\theta}_t - \mathbf{a}_t)}{2V}\right\} \end{aligned}$$

Taking the logarithm and multiplying by -2 for both,

$$\begin{aligned} -2\ ln(p(\boldsymbol{\theta}_t \mid Y_t, D_{t-1}) &\propto \frac{(Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t)'(Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t)}{V} + \frac{(\boldsymbol{\theta}_t - \mathbf{a}_t)'R_t^{*-1}(\boldsymbol{\theta}_t - \mathbf{a}_t)}{V} \\ &\propto (Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t)'(Y_t - \mathbf{F}_t'\boldsymbol{\theta}_t) + (\boldsymbol{\theta}_t - \mathbf{a}_t)'R_t^{*-1}(\boldsymbol{\theta}_t - \mathbf{a}_t) \\ &= \boldsymbol{\theta}_t'(R_t^{*-1} + \mathbf{F}_t\mathbf{F}_t')\boldsymbol{\theta}_t - 2\boldsymbol{\theta}_t'(R_t^{*-1}\mathbf{a}_t + \mathbf{F}_tY_t) + L_1 \qquad (4) \end{aligned}$$

By the matrix inversion lemma,

$$\begin{aligned} (R_t^{*-1} + \mathbf{F}_t\mathbf{F}_t')(C_t^*) &= I \\ \text{or} \quad R_t^{*-1} + \mathbf{F}_t\mathbf{F}_t' &= C_t^{*-1} \qquad (5) \end{aligned}$$

$$
\begin{aligned}
C_t^{*-1}\mathbf{m}_t &= (R_t^{*-1} + \mathbf{F}_t\mathbf{F}'_t)(\mathbf{a}_t + K_t e_t) \\
&= R_t^{*-1}\mathbf{a}_t + R_t^{*-1}K_t e_t + \mathbf{F}_t\mathbf{F}'_t\mathbf{a}_t + \mathbf{F}_t\mathbf{F}'_t K_t e_t \\
&= R_t^{*-1}\mathbf{a}_t + R_t^{*-1}\frac{R_t\mathbf{F}_t}{Q_t^*}e_t + \mathbf{F}_t\mathbf{F}'_t\mathbf{a}_t + \mathbf{F}_t\mathbf{F}'_t\frac{R_t^*\mathbf{F}_t}{Q_t^*}e_t \\
&= R_t^{*-1}\mathbf{a}_t + \frac{\mathbf{F}_t}{Q_t^*}e_t + \mathbf{F}_t\mathbf{F}'_t\frac{R_t^*\mathbf{F}_t}{Q_t^*}e_t + \mathbf{F}_t\mathbf{F}'_t\mathbf{a}_t \\
&= R_t^{*-1}\mathbf{a}_t + \mathbf{F}_t\left(\frac{e_t}{Q_t^*} + \mathbf{F}'_t\frac{R_t^*\mathbf{F}_t}{Q_t^*}e_t + \mathbf{F}'_t\mathbf{a}_t\right) \\
&= R_t^{*-1}\mathbf{a}_t + \mathbf{F}_t\left\{\frac{e_t}{Q_t^*}(1 + \mathbf{F}'_t R_t^*\mathbf{F}_t) + \mathbf{F}'_t\mathbf{a}_t\right\} \\
&= R_t^{*-1}\mathbf{a}_t + \mathbf{F}_t\left(e_t + \mathbf{F}'_t\mathbf{a}_t\right) \\
&= R_t^{*-1}\mathbf{a}_t + \mathbf{F}_t Y_t \qquad (6)
\end{aligned}
$$

By taking (5) and (6) into (4), the density function of $-2\ ln(p(\boldsymbol{\theta}_t \mid Y_t, D_{t-1}))$ reduces to

$$
\begin{aligned}
&= \boldsymbol{\theta}'_t C_t^{*-1}\boldsymbol{\theta}_t - 2\boldsymbol{\theta}'_t C_t^{*-1}\mathbf{m}_t + L_1 \\
&= (\boldsymbol{\theta}_t - \mathbf{m}_t)' C_t^{*-1}(\boldsymbol{\theta}_t - \mathbf{m}_t) + L_2 \quad \text{as a function of } \theta_t
\end{aligned}
$$

where L_1 and L_2 are constants with regard to a function of $\boldsymbol{\theta}_t$.

Therefore, $(\boldsymbol{\theta}_t \mid D_t) \sim N(\mathbf{m}_t, VC_t^*)$ since $p(\boldsymbol{\theta}_t \mid D_t) \propto \exp\left\{-\frac{(\boldsymbol{\theta}_t - \mathbf{m}_t)' C_t^{*-1}(\boldsymbol{\theta}_t - \mathbf{m}_t)}{2V}\right\}$.

B.1.2 $\tau = \frac{1}{V}$ is unknown, but assumed as a constant V

The updating equations are derived by marginalisation of the distributions when $\tau = \frac{1}{V}$ is known as V with respect to the appropriate prior/posterior gamma distribution for τ. The only difference is that the variance-covariance matrices are in the student T distributions including the relevant estimates of V.

B.1.3 Derivation of The Updating Equations for τ

By applying Bayes' theorem, the posterior for τ is

$$p(\tau \mid D_t) \propto p(\tau \mid D_{t-1}) p(Y_t \mid \tau, D_{t-1})$$

From **(c1)** in Section 2.5.3, the prior distribution at time t of τ is given by

$$p(\tau \mid D_{t-1}) = \frac{\left(\frac{d_{t-1}}{2}\right)^{\frac{n_{t-1}}{2}}}{\Gamma(\frac{n_{t-1}}{2})} \tau^{\frac{n_{t-1}}{2}-1} e^{-\tau \frac{d_{t-1}}{2}}$$

where n_t, d_t, and $\tau > 0$.

On the other hand, from **(a3)**,

$$p(Y_t \mid \tau, D_{t-1}) = \left(\frac{\tau}{2\pi Q_t^*}\right)^{1/2} e^{\left(-\frac{\tau (Y_t - \mathbf{f}_t)^2}{2Q_t^*}\right)}$$

Therefore, the posterior distribution at time t of τ reduces to

$$\begin{aligned} p(\tau \mid D_t) &\propto \tau^{\frac{(n_{t-1}+1)}{2}-1} e^{-\frac{\tau}{2}\left(d_{t-1} + \frac{e_t^2}{Q_t^*}\right)} \\ &\propto \tau^{\frac{n_t}{2}-1} e^{-d_t \frac{\tau}{2}} \end{aligned} \tag{7}$$

by taking $n_t = n_{t-1} + 1$ and $d_t = d_{t-1} + e_t^2 / Q_t^*$.

The posterior distribution of τ with parameters $\frac{n_t}{2}$ and $\frac{d_t}{2}$ is denoted by $(\tau \mid D_t) \sim Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)$, and the density function is

$$p(\tau \mid D_t) = \frac{\left(\frac{d_t}{2}\right)^{\frac{n_t}{2}}}{\Gamma(\frac{n_t}{2})} \tau^{\frac{n_t}{2}-1} e^{-\tau \frac{d_t}{2}}$$

where n_t, d_t, and $\tau > 0$. Since n_t indicates the degrees of freedom, each observation increases the degrees of freedom n_t by 1. Thus, $n_t = n_{t-1} + 1$ while d_t is determined by $d_{t-1} + \frac{e_t^2}{Q_t^*}$.

Appendix C

C.1 Computation of a Jacobian in Dynamic Generalised Linear Model

This appendix is about the computation of a Jacobian in Section 4.4.3.1. The result of $|J| = \frac{\prod_{i=1}^{n} e^{\eta_i}}{(1+\sum_{i=1}^{n} e^{\eta_i})^{n+1}}$ in (4.6) is obtained from the followings.

From the definition, the Jacobian $|J|$ is obtained by

$$|J| = \begin{vmatrix} \frac{\partial \pi_1}{\partial \eta_1} & \frac{\partial \pi_1}{\partial \eta_2} & \cdots & \frac{\partial \pi_1}{\partial \eta_n} \\ \frac{\partial \pi_2}{\partial \eta_1} & \frac{\partial \pi_2}{\partial \eta_2} & \cdots & \frac{\partial \pi_2}{\partial \eta_n} \\ \vdots & \cdots & \ddots & \vdots \\ \frac{\partial \pi_n}{\partial \eta_1} & \frac{\partial \pi_n}{\partial \eta_2} & \cdots & \frac{\partial \pi_n}{\partial \eta_n} \end{vmatrix} = \begin{vmatrix} \frac{e^{\eta_1}\mathrm{A} - e^{\eta_1}e^{\eta_1}}{\mathrm{A}^2} & \frac{-e^{\eta_1}e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_1}e^{\eta_n}}{\mathrm{A}^2} \\ \frac{-e^{\eta_2}e^{\eta_1}}{\mathrm{A}^2} & \frac{e^{\eta_2}\mathrm{A} - e^{\eta_2}e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_2}e^{\eta_n}}{\mathrm{A}^2} \\ \vdots & \cdots & \ddots & \vdots \\ \frac{-e^{\eta_n}e^{\eta_1}}{\mathrm{A}^2} & \frac{-e^{\eta_n}e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_n}\mathrm{A} - e^{\eta_n}e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix}$$

where $\mathrm{A} = 1 + \sum_{i=1}^{n} e^{\eta_i}$.

The following steps show the detailed computation for the Jacobian $|J|$.

Step 1
(Step 1-1) Add 1^{st} row to 2^{nd} row
(Step 1-2) Add new 2^{nd} row from (Step 1-1) to 3^{rd} row
(Step 1-3) Add new 3^{rd} row from (Step 1-2) to 4^{th} row
.....
(Step 1-(n-1)) Add new $(n-1)^{th}$ row from (Step 1-(n-2)) to n^{th} row.

These operations reveals that

$$\begin{vmatrix} \frac{e^{\eta_1}\mathrm{A}-e^{\eta_1}e^{\eta_1}}{\mathrm{A}^2} & \frac{-e^{\eta_2}e^{\eta_1}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_n}e^{\eta_1}}{\mathrm{A}^2} \\ \frac{e^{\eta_1}\mathrm{A}-e^{\eta_1}e^{\eta_1}-e^{\eta_1}e^{\eta_2}}{\mathrm{A}^2} & \frac{e^{\eta_2}\mathrm{A}-e^{\eta_2}e^{\eta_1}-e^{\eta_2}e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_n}e^{\eta_1}-e^{\eta_n}e^{\eta_2}}{\mathrm{A}^2} \\ \frac{e^{\eta_1}\mathrm{A}-e^{\eta_1}e^{\eta_1}-e^{\eta_1}e^{\eta_2}-e^{\eta_1}e^{\eta_3}}{\mathrm{A}^2} & \frac{e^{\eta_2}\mathrm{A}-e^{\eta_2}e^{\eta_1}-e^{\eta_2}e^{\eta_2}-e^{\eta_2}e^{\eta_3}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_n}e^{\eta_1}-e^{\eta_n}e^{\eta_2}-e^{\eta_n}e^{\eta_3}}{\mathrm{A}^2} \\ \vdots & \cdots & \ddots & \vdots \\ \frac{e^{\eta_1}\mathrm{A}-e^{\eta_1}e^{\eta_1}-\cdots-e^{\eta_1}e^{\eta_n}}{\mathrm{A}^2} & \frac{e^{\eta_2}\mathrm{A}-e^{\eta_2}e^{\eta_1}-\cdots-e^{\eta_2}e^{\eta_n}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_n}\mathrm{A}-e^{\eta_n}e^{\eta_1}-\cdots-e^{\eta_n}e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix}$$

Since $\mathrm{A} = 1 + \sum_{i=1}^{n} e^{\eta_i}$, the entries of the n^{th} row reduce to $\{a_{n,i}\} = \frac{e^{\eta_i}}{\mathrm{A}^2}$ for $i = 1, \ldots, n$. The entries of the $(n-1)^{th}$ row also become $\{a_{n-1,i}\} = e^{\eta_i}(1 + e^{\eta_n})$ for $i = 1, \ldots, n-1$ and $\{a_{n-1,n}\} = -e^{\eta_n}(e^{\eta_1} + \cdots + e^{\eta_{n-1}})$. Accordingly, the matrix gets simpler as follows.

$$\begin{vmatrix} \frac{e^{\eta_1}\mathrm{A}-e^{\eta_1}e^{\eta_1}}{\mathrm{A}^2} & \frac{-e^{\eta_2}e^{\eta_1}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_n}e^{\eta_1}}{\mathrm{A}^2} \\ \frac{e^{\eta_1}\mathrm{A}-e^{\eta_1}e^{\eta_1}-e^{\eta_1}e^{\eta_2}}{\mathrm{A}^2} & \frac{e^{\eta_2}\mathrm{A}-e^{\eta_2}e^{\eta_1}-e^{\eta_2}e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_n}e^{\eta_1}-e^{\eta_n}e^{\eta_2}}{\mathrm{A}^2} \\ \frac{e^{\eta_1}\mathrm{A}-e^{\eta_1}e^{\eta_1}-e^{\eta_1}e^{\eta_2}-e^{\eta_1}e^{\eta_3}}{\mathrm{A}^2} & \frac{e^{\eta_2}\mathrm{A}-e^{\eta_2}e^{\eta_1}-e^{\eta_2}e^{\eta_2}-e^{\eta_2}e^{\eta_3}}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_n}e^{\eta_1}-e^{\eta_n}e^{\eta_2}-e^{\eta_n}e^{\eta_3}}{\mathrm{A}^2} \\ \vdots & \cdots & \ddots & \vdots \\ \frac{e^{\eta_1}(1+e^{\eta_n})}{\mathrm{A}^2} & \frac{e^{\eta_2}(1+e^{\eta_n})}{\mathrm{A}^2} & \cdots & \frac{-e^{\eta_n}e^{\eta_1}-e^{\eta_n}e^{\eta_2}-\cdots-e^{\eta_n}e^{\eta_{n-1}}}{\mathrm{A}^2} \\ \frac{e^{\eta_1}}{\mathrm{A}^2} & \frac{e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix}$$

Step 2

(Step 2-1) Add (n^{th} row) $\times(e^{\eta_1} + \cdots + e^{\eta_{n-1}})$ to $(n-1)^{th}$ row

(Step 2-2) Add (n^{th} row) $\times(e^{\eta_1} + \cdots + e^{\eta_{n-2}})$ to $(n-2)^{th}$ row

(Step 2-3) Add (n^{th} row) $\times(e^{\eta_1} + \cdots + e^{\eta_{n-3}})$ to $(n-3)^{th}$ row

.....

(Step 2-(n-1)) Add (n^{th} row) $\times(e^{\eta_1})$ to 1^{st} row

The result of this operation is obtained by

$$\begin{vmatrix} \frac{e^{\eta_1}}{\mathrm{A}} & 0 & 0 & \cdots & 0 \\ \frac{e^{\eta_1}}{\mathrm{A}} & \frac{e^{\eta_2}}{\mathrm{A}} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ \frac{e^{\eta_1}}{\mathrm{A}} & \frac{e^{\eta_2}}{\mathrm{A}} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}} & 0 \\ \frac{e^{\eta_1}}{\mathrm{A}^2} & \frac{e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}^2} & \frac{e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix}$$

Step 3 Determinants computation

$$\begin{aligned} |J| &= \begin{vmatrix} \frac{e^{\eta_1}}{\mathrm{A}} & 0 & 0 & \cdots & 0 \\ \frac{e^{\eta_1}}{\mathrm{A}} & \frac{e^{\eta_2}}{\mathrm{A}} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ \frac{e^{\eta_1}}{\mathrm{A}} & \frac{e^{\eta_2}}{\mathrm{A}} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}} & 0 \\ \frac{e^{\eta_1}}{\mathrm{A}^2} & \frac{e^{\eta_2}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}^2} & \frac{e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix} = \frac{e^{\eta_1}}{A} \cdot \begin{vmatrix} \frac{e^{\eta_2}}{\mathrm{A}} & 0 & 0 & \cdots & 0 \\ \frac{e^{\eta_2}}{\mathrm{A}} & \frac{e^{\eta_3}}{\mathrm{A}} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ \frac{e^{\eta_2}}{\mathrm{A}} & \frac{e^{\eta_3}}{\mathrm{A}} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}} & 0 \\ \frac{e^{\eta_2}}{\mathrm{A}^2} & \frac{e^{\eta_3}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}^2} & \frac{e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix} \\ &= \frac{e^{\eta_1}}{A} \cdot \frac{e^{\eta_2}}{A} \cdot \begin{vmatrix} \frac{e^{\eta_3}}{\mathrm{A}} & 0 & 0 & \cdots & 0 \\ \frac{e^{\eta_3}}{\mathrm{A}} & \frac{e^{\eta_4}}{\mathrm{A}} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ \frac{e^{\eta_3}}{\mathrm{A}} & \frac{e^{\eta_4}}{\mathrm{A}} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}} & 0 \\ \frac{e^{\eta_3}}{\mathrm{A}^2} & \frac{e^{\eta_4}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}^2} & \frac{e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix} \\ &= \frac{e^{\eta_1}}{A} \cdot \frac{e^{\eta_2}}{A} \cdot \frac{e^{\eta_3}}{A} \cdot \begin{vmatrix} \frac{e^{\eta_4}}{\mathrm{A}} & 0 & 0 & \cdots & 0 \\ \frac{e^{\eta_4}}{\mathrm{A}} & \frac{e^{\eta_5}}{\mathrm{A}} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ \frac{e^{\eta_4}}{\mathrm{A}} & \frac{e^{\eta_5}}{\mathrm{A}} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}} & 0 \\ \frac{e^{\eta_4}}{\mathrm{A}^2} & \frac{e^{\eta_5}}{\mathrm{A}^2} & \cdots & \frac{e^{\eta_{n-1}}}{\mathrm{A}^2} & \frac{e^{\eta_n}}{\mathrm{A}^2} \end{vmatrix} \\ &= \cdots = \frac{e^{\eta_1}}{A} \cdot \frac{e^{\eta_2}}{A} \cdots \frac{e^{\eta_{n-1}}}{A} \cdot \frac{e^{\eta_n}}{A^2} = \frac{e^{\eta_1} \cdots e^{\eta_n}}{A^{n+1}} \\ &= \frac{\prod_{i=1}^{n} e^{\eta_i}}{(1 + \sum_{i=1}^{n} e^{\eta_i})^{n+1}} \end{aligned} \tag{8}$$

Thus, assuming that $(\mathbf{\Pi}_t \mid D_{t-1}) \sim Dirichlet(r_{1,t}, \ldots, r_{n+1,t})$, the density function

of $(\boldsymbol{\eta}_t \mid D_{t-1})$ is obtained by

$$p(\boldsymbol{\eta}_t \mid D_{t-1}) = \frac{\Gamma(\sum_{i=1}^{n+1} r_{i,t})}{\prod_{i=1}^{n+1} \Gamma(r_{i,t})} \frac{e^{\sum_{i=1}^{n} r_{i,t}\eta_{i,t}}}{(1+\sum_{i=1}^{n} e^{\eta_{i,t}})^{\sum_{i=1}^{n+1} r_{i,t}}}$$

as shown in (4.5).

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. New York: Dover Publications.

Box, G. and Kramer, T. (1992). Statistical process monitoring and feedback adjustment - a discussion. *Technometrics*, 34(3):251–267.

Box, G. E. P. and Jenkins, G. M. (1962). Some Statistical Aspects of Adaptive Optimization and Control. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):297–343.

Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. New York; London: McGraw-Hill.

Brown, R. G. (1963). *Smoothing, Forecasting, and Prediction of Discrete Time Series*. Englewood Cliffs, N.J.: Prentice-Hall.

Cappé, O., Godsill, S. J., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924.

Cargnoni, C., Müller, P., and West, M. (1997). Bayesian Forecasting of Multinomial Time Series through Conditionally Gaussian Dynamic Models. *Journal of the American Statistical Association*, 92(438):640–647.

Chan, E. P. (2008). *Quantitative Trading*. Hoboken, N.J.: Wiley; Chichester: John Wiley distributor.

Chan, E. P. (2013). *Algorithmic Trading*. Hoboken, N.J.: Wiley.

Chun, B., Kim, B., and Lee, Y.-H. (1998). Generalization of exponentially weighted rls algorithm based on a state-space model. In *Circuits and Systems, 1998. ISCAS '98. Proceedings of the 1998 IEEE International Symposium on*, volume 5, pages 198–201.

DeBondt, W. F. M. and Thaler, R. H. (1985). Does the stock market overreact? *Journal of Finance*, 40(3):793–805.

Douc, R., Cappé, O., and Moulines, E. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium*, pages 64–69.

Doucet, A., de Freitas, N., and Gordon, N. (2010a). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Doucet, A., Godsill, S., and Andrieu, C. (2010b). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12:656–704.

Drenick, R. F. and Shahbender, R. A. (1957). Adaptive servomechanisms. *American Institute of Electrical Engineers, Part II: Applications and Industry, Transactions of the*, 76(5):286–292.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Ehrman, D. S. (2005). *The Handbook of Pairs Trading*. Hoboken, N.J.: Wiley.

Elliott, R. J., Van der Hoek, J., and Malcolm, W. P. (2005). Pairs Trading. *Quantitative Finance*, 5(3):271–276.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 55(2):251–276.

Engle, R. F. and Yoo, B. S. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics*, 35(1):143–159.

Fahrmeir, L. (1992). Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models. *Journal of the American Statistical Association*, 87(418):501–509.

Fahrmeir, L. and Kaufmann, H. (1991). On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. *Metrika*, 38(1):37–60.

Fahrmeir, L. and Tutz, G. (2010). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York; London: Springer-Verlag.

Fama, E. F. (1965). Random walks in stock market prices. *Financial Analysts Journal*, 21:55–60.

Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

Fama, E. F. and French, K. (1988). Permanent and temporal components of stock prices. *Journal of Political Economics*, 96(2):246–273.

Fortescue, T., Kershenbaum, L., and Ydstie, B. (1981). Implementation of self-tuning regulators with variable forgetting factors. *Automatica*, 17(6):831–835.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika*, 85(1):215–227.

Gatev, E., Goetzmann, W. N., and Rouwenhorst, K. G. (2006). Pairs Trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827.

Goldstein, M. (1976). Bayesian analysis of regression problems. *Biometrika*, 63(1):51–58.

Goldstein, M. and Wooff, D. (2007). *Bayes Linear Statistics*. Chichester: John Wiley.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113.

Grunwald, G. K., Raftery, A. E., and Guttorp, P. (1993). Time series of continuous proportions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):103–116.

Harrison, P. J. and Stevens, C. F. (1971). A Bayesian approach to Short-Term Forecasting. *Operational Research Quarterly*, 22(4):341–362.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian Forecasting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):205–247.

Hartigan, J. A. (1969). Linear Bayesian Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(3):446–454.

Haykin, S. (1996). *Adaptive filter theory*. Upper Saddle River, N.J.: Prentice Hall.

Haykin, S. (2001). *Adaptive filter theory*. Upper Saddle River, N.J.: Prentice Hall.

Hol, J., Schon, T., and Gustafsson, F. (2006). On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop, 2006 IEEE*, pages 79–82.

Holt, C. C. (1957). Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. Technical report, Carnegie Institute of Technology, Pittsburgh, Pennsylvania.

Joe Qin, S. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, 17(8-9):480–502.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of basic Engineering*, 83(1):95–108.

Kendall, M. G. and Hill, A. B. (1953). The Analysis of Economic Time-Series–Part I: Prices. *Journal of the Royal Statistical Society. Series A (General)*, 116(1):11–34.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag.

Ljung, L. and Soderstrom, T. (1983). *Theory and Practice of Recursive Identification*. Cambridge, Mass.; London: MIT Press.

Lowenstein, R. (2002). *When Genius Failed: The Rise and Fall of Long Term Capical Management*. London: Fourth Estate.

Malik, M. B. (2006). State-space recursive least-squares with adaptive memory. *Signal processing*, 86(7):1365–1374.

Malkiel, B. G. (2004). *A Random Walk Down Wall Street*. London: W. W. Norton & Co.

Montana, G. and Parrella, F. (2008). Learning to Trade with Incremental Support Vector Regression Experts. In *Lecture Notes in Artificial Intelligence*, volume 5271, pages 591–598. Springer.

Montana, G., Triantafyllopoulos, K., and Tsagaris, K. (2009). Flexible least squares for temporal data mining and statistical arbitrage. *Expert Systems with Applications*, 36(2):2819–2830.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. Dordrecht; London: Springer.

Pole, A. (2007). *Statistical Arbitrage*. Hoboken, N.J.: Wiley; Chichester: John Wiley distributor.

Poterba, J. M. and Summers, L. (1988). Mean reversion in stock returns: Evidence and Implications. *Journal of Financial Mathematics*, 22(1):27–59.

Prado, R. and West, M. (2010). *Time Series Modeling, Computation, and Inference*. Boca Raton, Fla.: Chapman & Hall/CRC.

Åström, K. J., Borisson, U., Ljung, L., and Wittenmark, B. (1977). Theory and applications of self-tuning regulators. *Automatica*, 13(5):457–476.

Åström, K. J. and Wittenmark, B. (1973). On self tuning regulators. *Automatica*, 9(2):185–199.

Smith, J. Q. (1979). A Generalization of the Bayesian Steady Forecasting Model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(3):375–387.

Song, S., Lim, J.-S., Baek, S., and Sung, K.-M. (2000). Gauss Newton variable forgetting factor recursive least squares for time varying parameter tracking. *Electronics Letters*, 36(11):988–990.

Triantafyllopoulos, K. (2009). Inference of Dynamic Generalized Linear Models: On-Line Computation and Appraisal. *International Statistical Review*, 77(3):430–450.

Triantafyllopoulos, K. and Han, S. (2013). Detecting Mean-Reverted Patterns in Algorithmic Pairs Trading. In *Mathematical Methodologies in Pattern Recognition and Machine Learning*, volume 30 of *Springer Proceedings in Mathematics and Statistics*, pages 127–147. Springer.

Triantafyllopoulos, K. and Montana, G. (2011). Dynamic modelling of mean-reverting spreads for statistical arbitrage. *Computational Management Science*, 8(1-2):23–49.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of Brownian motion. *Physical Review*, 36(5):823–841.

Vidyamurthy, G. (2004). *Pairs Trading*. Hoboken, N.J.; Great Britain: John Wiley.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83.

Whistler, M. (2004). *Trading Pairs*. Hoboken, N.J.: Wiley; Chichester: John Wiley.

Zhang, H. and Zhang, Q. (2008). Trading a mean reverting asset: Buy low and sell high. *Automatica*, 44(6):1511–1518.