Imperial College
London

Imperial College London
Department of Mathematics

State Space Modelling for Statistical Arbitrage

Author:
Philippe Remy

Supervised by:
Dr. Nikolas Kantas
Dr. Yanis Kiskiras

3rd September 2015

This report is submitted as part requirement for the MSc Degree in Statistics at Imperial College London. It is substantially the result of my own work except where explicitly indicated in the text. The report will be distributed to the internal and external examiners, but thereafter may not be copied or distributed except with permission from the author.

Abstract

Statistical Arbitrage is a computationally-intensive approach which involves the simultaneous buying and selling of securities according to statistical models. Statistical Arbitrage strategies are heavily based on the construction of stationary mean-reverting spreads and sophisticated models to identify opportunities. This thesis extends the classic cointegration-based pairs trading by considering two cases: triples and quadruples of securities. It is common, in pairs trading strategies to impose that the pairs belong to the same sector, for example in Chan (2009) and Dunis et al. (2010). Similar to Caldeira and Moura (2013) for pairs trading, we do not adopt this restriction for triple trading as the computational cost is still acceptable. It becomes much harder with quadruple trading with a dataset composed of the most liquid stocks traded on the US exchanges. Three strategies are discussed in this thesis: Bollinger bands with simple and complex volatility modelling and Z-score. In the complex modelling strategy, the volatility of the financial instruments is estimated with Stochastic Volatility models where the model parameters are estimated via *Particle Markov Chain Monte Carlo*. The profitability of the strategies is assessed on the US Equity markets for the period 1990-2014. Empirical analysis shows that the proposed strategy accounts for excess returns of xx% per year, Sharpe Ratio above 2 and low correlation with the market.

Acknowledgements

First and foremost I offer my deepest gratitude to my tutor, Dr. Nikolas Kantas, who has supported me throughout my thesis with his patience and knowledge whilst allowing me the room to work in my own way. Without his encouragements and efforts, this project would not have been completed. One simply could not wish for a better tutor.

Besides my tutor, I would like to thank Dr. Yanis Kiskiras for introducing me to the topic as well for the support on the way. His insightful comments have largely contributed to the success of this project.

I thank my fellow classmates of the *Further Topics in Statistics* group: Federico De Riu, Thomas Vogel, Johan Kestenare and Qing Liu for the stimulating discussions about *Particle Markov Chain Monte Carlo*. Also, the Department of Mathematics has provided me the support and equipment I have needed to produce and complete my thesis.

In my family, I thank my parents Catherine and Bertrand. Last but not least, I am grateful to the warm support Chieri Kii has sent me from Tokyo.

"We rise by lifting others."
–Robert Ingersoll

Table of Contents

List of Figures

List of Tables

Notation

The following notation is used throughout this thesis.

Notation	Definition
T	Sample size
1:T	State space
$\mathbf{X}_t \in \mathbb{R}^n$	A random time-indexed state vector with n components
$\mathbf{x}_t \in \mathbb{R}^n$	A realization of the random vector X_t, namely $\{x_1, ..., x_T\}$
$\mathbf{Y}_{1:T} \in \mathbb{R}^T \times \mathbb{R}^n$	A set of random vectors (observations), each with n components
$\mathbf{y}_{1:T} \in \mathbb{R}^T \times \mathbb{R}^n$	A set of observations, namely $\{y_1, ..., y_T\}$
$\sim$	Distributed as
$\propto$	Proportional to
i.i.d	Independent, identically distributed
L	Lag operator. Defined as $LX_t = X_{t-1}$
Δ	Difference operator. Defined as $\Delta X_t = (1 - L)X_t = X_t - X_{t-1}$
$E[\mathbf{X}_t \vert \mathcal{F}_t]$	Conditional expectation
$\text{Var}[\mathbf{X}_t \vert \mathcal{F}_t]$	Conditional variance
Cor	Correlation function
$\circ$	Composition function operator
$p(\cdot)$	General marginal probability density function
$p(\cdot \vert \cdot)$	Conditional probability density function
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$t(\nu)$	t-student distribution with ν degrees of freedom
erf	Error function defined as $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\ dt$
AR(p)	Auto Regressive process of lag $p \in \mathbb{N}^* \cup \infty$
MA(p)	Moving Average process of lag $p \in \mathbb{N}^* \cup \infty$

1 Introduction

For many years, the finance industry has used the concept of correlation in Statistical Arbitrage to detect opportunities. This widely use of short-term correlation on detrended non-stationary time series data turned out to be risky because a large amount of valuable information contained in the common trends of the prices was lost. Engle and Granger (1987) introduced a new concept, known as Cointegration. It has been widely used in the field of financial econometrics in the areas of multivariate time series analysis. This concept provides a way to identify the influence of common and stable long-term stochastic trends between assets. The variables are allowed to deviate from their inherent relationships in the short term but they are likely to revert to their long term equilibrium. Spot and Futures prices for a particular asset is an example of a bivariate cointegrated system.

Their pseudo difference is known as the spread and the analysis of its variance dynamics helps build trading signals. The GARCH model and the Stochastic Volatility model are competing - but non-nested models - to describe unobserved volatility in security returns. GARCH models the evolution of volatility deterministically. After the publications of Engle and Bollerslev (1986), these models have been generalized in numerous ways and applied to a vast amount of real-world problems. As an alternative, Taylor (1982) proposed in his seminal work to model the volatility probabilistically, i.e., through a state-space model where the logarithm of the squared volatilities - the latent states - follow an autoregressive process of order one. This specification became known as the Stochastic Volatility model. Kim et al. (1998) provided early evidence in favor of using this model. Kastner et al. (2014) analyzed foreign exchanges rates and showed that a standard Stochastic Volatility performs better than a vanilla GARCH(1,1) in terms of predictive accuracy. Chan and Grant (2015) compared a number of GARCH and Stochastic Volatility models on commodity markets and had the same conclusions. Even though several papers highlighted the fact that Stochastic Volatility generally compared favorably to their GARCH counterparts, these have found comparably little use in applied work. The main reason pointed out was the difficulty of estimating the parameters of such models.

Recently, new algorithms based on Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo have been designed. MCMC methods sample from a probability distribution by constructing a Markov chain with the same target distribution. Andrieu et al. (2010) introduced a new method which embed Sequential Monte Carlo filters within MCMC samplers for the joint estimation of static parameters and latent states in complex non-linear systems. These advanced particle methodologies belong to the class of

Feynman-Kac particle models and are called *Particle Markov Chain Monte Carlo*. Many aspects of their behavior in practical applications remain open research questions.

This thesis focuses on the development and the estimation of Stochastic Volatility models in order to output a better estimate of the volatility of cointegrated prices than traditional methods. This estimated volatility is later used as part of a trading strategy based on the Bollinger bands, a widely known technical trading indicator created in 1980. In a nutshell, it consists of a set of three curves drawn in relation to securities prices. The middle band represents the trend which is used for the upper and the lower bands. The interval between the upper and lower bands is determined by the recent volatility of the security prices. The purpose is to give systematic trading decisions by evaluating if the price is either high, low or in the range. This strategy is suitable for cointegration since it is based on the mean-reverting pattern of the security. Also, we investigate the risk and return of a portfolio consisting of various cointegrated tuples. For further discussions based on mean-reverting stationary spreads and illustrative numerical examples, the reader is referred to Vidyamurthy (2004). It is well known that those common strategies are popular among many hedge funds. However, there is not a significant amount of academic literature devoted to it, due to its proprietary nature. For a review of some of the existing academic models, see Gatev et al. (2006), Perlin (2009) and Broussard and Vaihekoski (2012).

The remainder of this thesis is organized as follows. In section 2, cointegration theory is presented in greater details. Section 3 introduces the state-space and the Stochastic Volatility models. Section 4 explains how to estimate the parameters using *Particle Markov Chain Monte Carlo*. In section 5, the best Stochastic Volatility model is selected based on real world data. The trading strategies are described in Section 6. Section 7 and 8 present the procedure and the results of the simulations. Finally, a conclusion based on the empirical results is presented, along with suggestions of future research.

1.1 Some Technical Definitions

Definition 1. *In the following, we will assume that a process $(X_t)_{t\in\mathbb{N}}$ is adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ which presents the accrual of information over time. We denote by $\mathcal{F}_t = \sigma\{X_s : s \leq t\}$ the σ-algebra generated by the history of X up to time t. The corresponding filtration is then called the natural filtration.*

Definition 2. *A n-unit-root tuple $\mathcal{T}$ is a finite ordered list of $I(1)$ processes $(\mathbf{X}_1, ..., \mathbf{X}_n)$ defined on $\mathbb{R}^{n\times t}$. Each component $\mathbf{X}_i$ can be modelled by an heteroskedastic random walk.*

Definition 3. *A spread $(S_t)_{t>0}$ is defined as a linear combination between the components $(\mathbf{X}_i)_{1<i\leq n}$ of an unit-root tuple $\mathcal{T}$. The combination is such that S_t is stationary (i.e $I(0)$). Let $\beta = (\beta_1, ...\beta_n)$ be the linear coefficients associated to the components of $\mathcal{T}$. In matrix notation, the spread can be written as $S_t = \beta'\mathbf{X}_t$. In practical applications, some constraints are added to this definition. It is also assumed that $\beta_1 = 1$*

and $\beta_2, ...\beta_n \leq 0$ (unless stated otherwise). The values of β are computed by considering the linear regression with first differenced variables: $\Delta\mathbf{X}_1 = f(\Delta\mathbf{X}_2, ...\Delta\mathbf{X}_n)$ where $\Delta\mathbf{X}_i \sim I(0)$ because $\mathbf{X}_i \sim I(1)$. The spread is defined as $\mathbf{S} = \mathbf{X}_1 + \sum_{i=2}^{n} \beta_i \mathbf{X}_i$. Note that with this construction, not all the spreads are stationary and additional tests must be performed.

Definition 4. *A rolling (or windowed) volatility process of lag p, $(r\sigma(t,p))_{t>0}$ is defined as a simple moving average process over the last p values of the volatility of S_t. Formally it can be written as*

$$r\sigma(t,p) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(S_{t-j} - \frac{1}{p}\sum_{u=1}^{p} S_{t-u}\right)^2}$$

When $p \to 1$, $r\sigma(t,p)$ converges to the instant volatility associated to S_t.

Definition 5. *Let X be a random variable taking values in $\mathbb{R}$. For $a \in \mathbb{R}$ and $b > 0$, if $Y = a + bX$ is equal in distribution to X, the random variables X and Y belong to the same location-scale family. Examples are the Normal, Cauchy, Uniform, Laplace, GEV and Student-t distributions.*

1.2 Presentation of the Data Sets

1.2.1 US Equities

The Equity dataset $\mathcal{EQTY}_{daily}$ consists of daily closing prices of the 1232 most liquid stocks traded on the US markets (NASDAQ, NYSE...). The sample period begins in January 1990 and ends in March 2014, summing up to 8844 observations. The liquidity feature is important for the strategies, since it greatly diminishes the slippage effect, reduces the transaction costs and permits unwinding any position without impacting the market too much. The data was adjusted for dividends and splits, avoiding false trading signals generated by these events, as pointed out by Broussard and Vaihekoski (2012).

1.2.2 Foreign Exchange Rates

The dataset $\mathcal{FX}_{daily}$ is composed of daily prices (FX Spot) for the pairs : AUD-USD, GBP-USD, CAD-USD, EUR-USD, NZD-USD, ZAR-USD, CHF-USD. The period spans the time from Jan, 01 1999 to Jan, 01 2013, summing up to 5116 observations.

2 Cointegration

Statistical arbitrage is based on the assumption that the patterns observed in the past are going to be repeated in the future. This is in opposition to the fundamental investment strategy that explores and tries to predict the behavior of economic forces that influence the share prices. Thereby, statistical arbitrage is a purely statistical approach designed to exploit market inefficiencies defined as the deviation from then long-term equilibrium across the stock prices in the past. Cointegration theory is the cornerstone of this approach.

2.1 Cointegration Theory

Cointegration is a statistical property possessed by some time series based on the concepts of stationary and the order of integration of the series. A series is considered stationary if its distribution is time invariant. In other words, the series will constantly return to its time invariant mean value as fluctuations occur. In contrast, a non-stationary series will exhibit a time varying mean. A series is said to be integrated of order d, denoted $I(d)$ if it must be differenced at least d times to produce a stationary series. Nelson and Plosser (1982) showed that most time series have stochastic trends and are $I(1)$.

The significance of cointegration analysis is its intuitive appeal for dealing with difficulties that arise when using non-stationary series, particularly those that are assumed to have a long-run equilibrium relationship. For instance, when non-stationary series are used in regression analysis, one as a dependent variable and the others as independent variables, statistical inference becomes problematic. Assume that y_t and x_t are two independent random walk, and let us consider the regression : $y_t = ax_t + b + \epsilon_t$, where ϵ_t is the residual at time t. It is obvious that the true value of a is 0, because $corr(x_t, y_t) = 0$. But in practical applications, the estimated value of a is often statistically different from 0. This is called a spurious regression, and was first noted by Monte Carlo studies by Granger and Newbold (1974). If x_t and y_t are both unit root processes, classical theory applies for the regression : $\Delta y_t = a\Delta x_t + b + \epsilon_t$ since both variables are stationary.

Cointegration is said to exist between two or more non-stationary time series if they possess the same order of integration and if a linear combination of these series is stationary. Let $\mathbf{X}_t = (x_{1t}, ..., x_{nt})^T$ be a n-$I(1)$ process. The vector $\mathbf{X}_t$ is said to be cointegrated if there exists at least one non trivial vector $\beta = (\beta_1, ..., \beta_n)^T$ such that $\epsilon_t = \beta^T \mathbf{X}_t$ is a stationary process $I(0)$. β is called a cointegrating vector and is defined up to a scaling

parameter k. Indeed, $k\beta^T \mathbf{X}_t \sim I(0)$ for any $k \neq 0$. There can be r different cointegrating vectors, where $0 \leq r < n$, i.e. r must be less than the number of variables n. In such a case, we can distinguish between a long-run relationship between the variables contained in $\mathbf{X}_t$, that is, the manner in which the variables drift upward together, and the short-run dynamics, that is the relationship between deviations of each variable from their corresponding long-run trend. The implication that non-stationary variables can lead to spurious regressions unless at least one cointegration vector is present means that some form of testing for cointegration is almost mandatory. In practical applications, the contribution of each security should be in the same proportions. If a coefficient of β is very large compared to the others, it means that the investor might be exposed to a high risk upon this asset if the vector came to lose its cointegrated property. Conversely, a coefficient close to zero requires almost no funds to invest in this asset.

2.2 Vector Auto Regressive Process (VAR)

The Vector Auto Regressive (VAR) process is a generalization of the univariate AR process to the multivariate case. It is defined as

$$\mathbf{X}_t = \nu + \sum_{j=1}^{k} \mathbf{A}_j \mathbf{X}_{t-j} + \epsilon_t,\ \epsilon_t \sim SWN(0, \Sigma) \tag{2.1}$$

where k is the lag, $\mathbf{X}_t = (x_{1t}, ..., x_{nt})^T$, each of the $\mathbf{A}_j$ is a $(n \times n)$ matrix of parameters, ν is a fixed vector of intercept terms. Finally ϵ_t is a n-dimensional strict white noise process of covariance matrix Σ. The process $\mathbf{X}_t$ is said to be stable if $\det|\mathbf{I}_n - \sum_{j=1}^{k} \mathbf{A}_j z^j| \neq 0$ for $|z| \leq 1$, i.e. there are no roots on the complex unit circle. If there are roots on the unit circle then some or all the variables in $\mathbf{X}_t$ are $I(1)$ and they may also be cointegrated. If $\mathbf{X}_t$ is cointegrated, then the VAR representation is not the most suitable representation because the cointegrating relations are not explicitly apparent. In this case, the VECM model is more adapted.

2.3 Vector Error Correction Model (VECM)

In an vector error correction model (VECM), the changes in a variable depend on the deviations from some equilibrium relation. Suppose the case $n = 2, \mathbf{X}_t = (x_t, y_t)^T$ where x_t represents the price of a Future contract on a commodity and y_t is the spot price of this same commodity traded on the same market. Assume furthermore that the equilibrium relation between them is given by $y_t = \beta x_t$ and that Δy_t depends on the deviation from this equilibrium at time $t-1$. A similar relation may also hold for x_t. The system is defined by

$$\Delta y_t = \alpha(y_{t-1} - \beta x_{t-1}) + \epsilon_{y_t} \tag{2.2}$$

$$\Delta x_t = \alpha(y_{t-1} - \beta x_{t-1}) + \epsilon_{x_t} \tag{2.3}$$

where α represents the speed of adjustments to disequilibrium and β is the long run coefficient of the equilibrium. In a more general error correction model, the Δy_t and Δx_t may in addition depend on previous changes in both variables as, for instance, in the following model with lag one

$$\Delta y_t = \alpha(y_{t-1} - \beta x_{t-1}) + \gamma_{11}\Delta y_{t-1} + \gamma_{12}\Delta x_{t-1} + \epsilon_{y_t} \tag{2.4}$$
$$\Delta x_t = \alpha(y_{t-1} - \beta x_{t-1}) + \gamma_{21}\Delta y_{t-1} + \gamma_{22}\Delta x_{t-1} + \epsilon_{x_t} \tag{2.5}$$

In matrix notation and in the general case, the VECM is written as

$$\Delta \mathbf{X}_t = \mathbf{\Phi D}_t + \mathbf{\Lambda X}_{t-1} + \sum_{j=1}^{k-1} \mathbf{\Gamma}_j \Delta \mathbf{X}_{t-j} + \epsilon_t \tag{2.6}$$

where $\mathbf{\Phi D}_t$ are the deterministic terms, $\mathbf{\Gamma}_j = -\sum_{i=j+1}^{k} \mathbf{A}_i$ and $\mathbf{\Lambda} = \left(\sum_{i=1}^{k} \mathbf{A}_i\right) - \mathbf{I}_n$. This way of specifying the system contains information on both the short-run and long run adjustments to changes in $\mathbf{X}_t$, via the estimates $\hat{\mathbf{\Gamma}}_j$ and $\hat{\mathbf{\Lambda}}$ respectively. In the VECM, $\Delta \mathbf{X}_t$ and its lags are $I(0)$. The term $\mathbf{\Lambda X}_{t-1}$ is the only one which includes potential $I(1)$ variables and for $\Delta \mathbf{X}_t$ to be $I(0)$, it must be the case that $\mathbf{\Lambda X}_{t-1}$ is also $I(0)$. Therefore, $\mathbf{\Lambda X}_{t-1}$ must contain the cointegrating relations provided that they exist. If the VAR(k) has unit roots on the complex unit circle, then

$$\det \left| \mathbf{I}_n - \sum_{j=1}^{k} \mathbf{A}_j z^j \right| = 0 \tag{2.7}$$
$$\det (\mathbf{\Lambda}) = 0 \tag{2.8}$$

which means that $\mathbf{\Lambda}$ is singular. A singular matrix has a reduced rank and $\text{rank}(\mathbf{\Lambda}) = r < n$. Two cases are to consider. If the rank is 0, it implies that $\mathbf{\Lambda} = 0$. In this case, $\mathbf{X}_t$ is not cointegrated and the VECM reduces to a VAR($k-1$) in first differences

$$\Delta \mathbf{X}_t = \mathbf{\Phi D}_t + \sum_{j=1}^{k-1} \mathbf{\Gamma}_j \Delta \mathbf{X}_{t-j} + \epsilon_t \tag{2.9}$$

If $0 < rank(\mathbf{\Lambda}) = r < n$, it implies that $\mathbf{X}_t$ is $I(1)$ with r linearly independent cointegrating vectors and $n - r$ common stochastic trends (unit roots). Since $\mathbf{\Lambda}$ has rank r, it can be written as the product $\mathbf{\Lambda} = \alpha\beta'$ where α and β are of dimension $(n \times r)$ and rank r. The rows of β' form a basis for the r cointegrating vectors and the elements of α distribute the impact of the cointegrating vectors to the evolution of $\Delta \mathbf{X}_t$. The VECM becomes

$$\Delta \mathbf{X}_t = \mathbf{\Phi D}_t + \alpha\beta' \mathbf{X}_{t-1} + \sum_{j=1}^{k-1} \mathbf{\Gamma}_j \Delta \mathbf{X}_{t-j} + \epsilon_t \tag{2.10}$$

where $\beta' \mathbf{X}_{t-1} \sim I(0)$ since β' is a matrix of cointegrating vectors. α corresponds to a matrix of error-correction speeds. It is also important to notice that the factorization of $\mathbf{\Lambda} = \alpha\beta'$ is not unique since for any $(r \times r)$ non singular matrix $\mathbf{H}$ we have

$$\alpha\beta' = \alpha\mathbf{H}\mathbf{H}^{-1}\beta' = (\mathbf{a}\mathbf{H})(\beta\mathbf{H}^{-1'})' = \mathbf{a}^*\beta^{*'}, \mathbf{a}^* = \mathbf{a}\mathbf{H}, \beta^* = \beta\mathbf{H}^{-1'} \tag{2.11}$$

Hence the factorization only identifies the space spanned by the cointegrating relations. To obtain unique values of α and β' requires further restrictions on the model.

2.4 Johansen Cointegration Test

The cointegration relations can be estimated with a Johansen test, as explained in Johansen (1988). The main advantage is the support of multi cointegrating relationships. It is generally more pertinent than the default Engle-Granger test which is based on the Dickey-Fuller test for unit roots in the residuals from a single cointegrating relation. The number of cointegrating vectors is determined through an iterative process of Likelihood Ratio Tests. Let the VECM with rank $(\mathbf{\Lambda}) < r$ be denoted $H(r)$. This creates a nested set of models $H(0) \in ... \in H(r) ... \in H(k)$. $H(0)$ means that there is no cointegrating relations. On the opposite, $H(k)$ means that we have a stationary VAR(k). This nested formulation is useful for developing an iterative procedure to test for r. The procedure begins by a test of $H_0(r_0 = 0)$ against $H_1(r_0 > 0)$. If this null is not rejected then it is concluded that there are no cointegrating vectors among the k variables in $\mathbf{X}_t$. If it is rejected, there is at least one cointegrating vector and we proceed to the test of $H_0(r_0 = 1)$ against $H_1(r_0 > 1)$. If the null is not rejected, then it is concluded that there is only one cointegrating vector. This iterative procedure is continued until the null is not rejected or that k is reached.

The LR tests are based on the estimated eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > ... > \hat{\lambda}_n$ of the matrix $\mathbf{\Lambda}$. Note that r is also equal to the number of non-zero eigenvalues of $\mathbf{\Lambda}$. If it is found that rank $(\mathbf{\Lambda}) = r, 0 < r < n$, then the cointegrated VECM becomes a reduced rank multivariate regression. Johansen (1988) derived this maximum likelihood estimation of the parameters under the reduced rank restriction. He showed that $\hat{\beta}_{mle} = (\hat{v_1}, ..., \hat{v_r})$ where $\hat{v_i}$ are the eigenvectors associated with the eigenvalues $\hat{\lambda}_i$. The columns of $\hat{\beta}'_{mle}$ are the estimators of the cointegrating vectors. The MLEs of the remaining parameters are obtained by least squares estimation of

$$\Delta\mathbf{X}_t = \mathbf{\Phi}\mathbf{D}_t + \alpha\hat{\beta}'_{mle}\mathbf{X}_{t-1} + \sum_{j=1}^{k-1}\mathbf{\Gamma}_j\Delta\mathbf{X}_{t-j} + \epsilon_t \tag{2.12}$$

The specification of the deterministic terms $\mathbf{\Phi}\mathbf{D}_t$ has to be taken into consideration. Following Johansen (1995), the deterministic terms are restricted to the form

$$\mathbf{\Phi}\mathbf{D}_t = \mu_t = \mu_0 + \mu_1 t \tag{2.13}$$

Restricted versions of the trend parameters μ_0 and μ_1 limit the trending nature of the series in $\mathbf{X}_t$. Johansen (1995) classified the trend behavior of $\mathbf{X}_t$ in five cases

- Model $H_2(r)$: $\mu_t = 0$ No constant.
  The series in $\mathbf{X}_t$ are $I(1)$ without drift and the cointegrating relation $\beta'\mathbf{X}_t$ have mean zero.

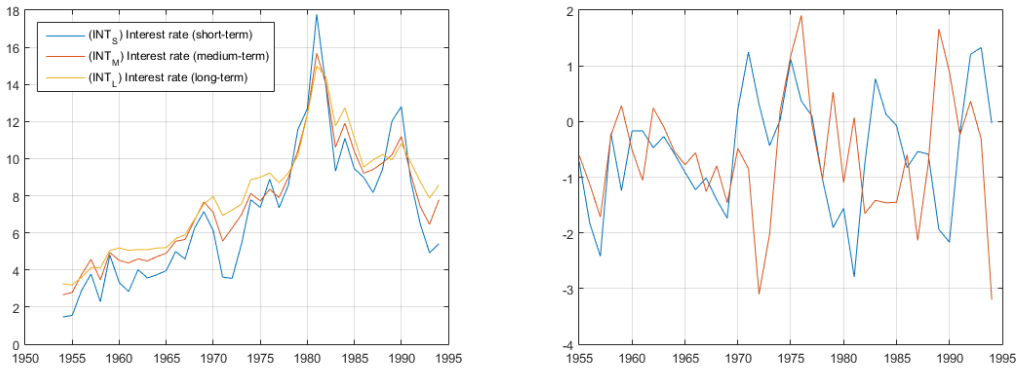
- Model $H_1^*(r)$: $\mu_t = \mu_0 = \alpha\rho_0$. Restricted constant.
  The series in $\mathbf{X}_t$ are $I(1)$ without drift and the cointegrating relation $\beta'\mathbf{X}_t$ have non-zero mean ρ_0.

- Model $H_1(r)$: $\mu_t = \mu_0$. Unrestricted constant.
  The series in $\mathbf{X}_t$ are $I(1)$ with drift vector μ_0 and the cointegrating relation $\beta'\mathbf{X}_t$ may have a non-zero mean.

- Model $H^*(r)$: $\mu_t = \mu_0 + \alpha\rho_1 t$. Restricted trend.
  All the series in $\mathbf{X}_t$ are $I(1)$ without drift and the cointegrating relation $\beta'\mathbf{X}_t$ have a linear trend term $\rho_1 t$.

- Model $H(r)$: $\mu_t = \mu_0 + \mu_1 t$. Unrestricted constant and trend.
  All the series in $\mathbf{X}_t$ are $I(1)$ with a linear trend and the cointegrating relation $\beta'\mathbf{X}_t$ have a linear trend.

$H_1(r)$ seems to be definitely the most relevant model to use for spread instruments because a drift is usually present in the components of $\mathbf{X}_t$. Moreover, this model eliminates both stochastic and deterministic trends in the cointegrating vectors.



(a) Cointegration of the interest rates (short, medium and long-term) in Canada from 1955 to 1995

(b) Estimated Cointegrating relations $\beta' y_{t-1} + c_0$

Figure 2.1: Cointegration property of the Canadian interest rates

It seems that the existence of more than one cointegrating vector (i.e. the long-run relationship) is not necessarily a good sign, since there is uncertainty as to which relationship the variables will obey in the long and short run. The dynamics may be unstable. The assumption will be discussed later.

2.5 Testing for Unit Roots in Stochastic Processes

Before testing for a unit root, (i.e. if the series is $I(1)$), the time series must be transformed to its linear form. Usually, assets prices have an exponential growth and logarithm should be applied accordingly to satisfy this prerequisite. Once the data is transformed, we have to choose the most pertinent model to use in the Augmented Dickey Fuller test. This test is an enhanced version of the Dickey-Fuller test, which assesses the presence of a unit root in an autoregressive model. The augmented version has a larger and more complicated set of time series models. There are three basic models for economic data $\mathbf{Y}_t$

- Trend Stationary model variant (TS)
  H0: $y_t = c + y_{t-1} + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \epsilon_t$
  H1: $y_t = c\delta t + \gamma y_{t-1} + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \epsilon_t$
  with drift coefficient c, deterministic trend coefficient δ and $AR(1)$ coefficient $\gamma < 1$.

- Auto Regressive with Drift variant (ARD)
  H0: $y_t = y_{t-1} + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \epsilon_t$
  H1: $y_t = c + \gamma y_{t-1} + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \epsilon_t$
  with drift coefficient c, and $AR(1)$ coefficient $\gamma < 1$.

- Auto Regressive variant (AR)
  H0: $y_t = y_{t-1} + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \epsilon_t$
  H1: $y_t = \gamma y_{t-1} + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \epsilon_t$
  with $AR(1)$ coefficient $\gamma < 1$.

ϵ_t is a mean zero innovation process. In general, if the series is growing, the TS model provides a reasonable trend-stationary alternative to a unit-root process with drift. If the series shows no trend but has a non zero mean, the ARD model provides reasonable stationary alternatives to a unit-root process without drift. Finally, if the series has no trend and a zero mean, the AR model is the most suitable. As the spread is a non zero mean without any drift, the ARD model is the best alternative model for testing.

The next step is to determine the number of lags to include in the model. Different criteria used for lag length often lead to different decisions regarding the optimal lag order that should be used in the model. DAO et al. (2014) suggested a general procedure for the ADF test

- Determine the optimal max lag value denoted L_{max}. It is clear that $L_{min} = 0$ is the minimum value of lag length that could be used. Schwert (2002) suggested to use $L_{max} = 12\,(T/100)^{1/4}$ where T is the length of the time series. It guarantees that L_{max} grows with T.

- When L_{min} and L_{max} are established, ADF t-statistics are calculated for all lag length values between the range (L_{min}, L_{max}). The most negative value from all ADF t-statistics indicates the value of lag length that produces the most stationary residuals.

2.6 Creation and Validation of the Spread Instruments

A rigorous testing for cointegration is performed to select the tuples for trading. It involves Johansen, Augmented Dickey Fuller and Variance Ratio tests. The Variance Ratio Test (VRT) is based on the idea that a stationary series does not have its variance increasing with time.

Algorithm 1 explains the procedure. The result of a test is denoted $h = p_{value} < 0.05$. Values of h equal to 1 indicate rejection of the null hypothesis in favor of the alternative model. Values of h equal to 0 indicate a failure to reject the null.

Algorithm 1 Formation of the Spread

procedure CREATESPREAD(M tuples of size N : $\{(\mathbf{X}_i)_{1<i\leq N}\}_m$)
  for m from 1 to M **do**
    Select the tuple $(\mathbf{X}_i)_{1<i\leq N}$ indexed by m
    for i from 1 to N **do**
      h = Test $\mathbf{X}_i \sim I(1)$ with an Augmented Dickey-Fuller test
      if h = 1 ($\mathbf{X}_i \not\sim I(1)$) **then**
        Break
    end
    $[h_1, ..., h_N]$ = Perform Johansen Cointegration Test on $(\mathbf{X}_i)_{1<i\leq N}$
    r = Determine Cointegration Rank of $[h_1, ..., h_N]$
    if $r \neq 1$ **then** //One cointegrating relation is enough
      Break
    for j from 1 to N **do** //Order is important in a tuple
      Regress $\Delta\mathbf{X}_j = f((\Delta\mathbf{X}_i)_{i\neq j})$
      Form the spread $\mathbf{S} = \beta'\mathbf{X} = \mathbf{X}_j - \sum_{i\neq j} \beta_i \mathbf{X}_i$
      Determine the optimal lag p for the ADF test (Section 2.5)
      h_1 = Test $\mathbf{S} \sim I(1)$ with an Augmented Dickey-Fuller test with lags $[0, p]$
      h_2 = Test $\mathbf{S} \sim RW(\cdot)$ with variance ratio test for random walk
      if $h_1 = 1$ **and** $h_2 = 1$ **then**
        S is a spread candidate for trading. Add to list $\mathcal{L}$
        Break //Success. Go to next tuple.
    end
  end
return $\mathcal{L}$: list of cointegrated spreads

3 State-Space Models and Stochastic Volatility

3.1 State-Space Models

The term *state-space* originated in 1960s in the area of control engineering (Kalman (1960)). State-space models refers to a class of probabilistic graphical model that describes the probabilistic dependence between the latent state variables $\mathbf{x}_{1:T}$ and the observed measurements $\mathbf{y}_{1:T}$. The system evolves according to

$$\mathbf{X}_t = f_\theta(\mathbf{X}_{t-1}, \mathbf{W}_{t-1}) \tag{3.1}$$

$$\mathbf{Y}_t = h_\theta(\mathbf{X}_t, \mathbf{V}_t) \tag{3.2}$$

where, $\mathbf{X}_t \in \mathbb{R}^n$ is the state vector and $\mathbf{Y}_t \in \mathbb{R}^n$ is the measurement vector. f_θ is a Markov process of order one and h_θ is any non-linear function. Both f_θ and h_θ are assumed to be time invariant and deterministic. $\mathbf{W}_t$ and $\mathbf{V}_t$ are respectively the i.i.d. system and i.i.d. measurement noise sequences. The Directed Acyclic Graph (DAG) for this state-space is given in Figure 3.1.

Equation (3.1) is known as the system equation and Equation (3.2) is known as the measurement equation. We assume that the process generating the system states $\mathbf{X}_t$ and thus the observed states $\mathbf{Y}_t$ starts from an initial value $\mathbf{x}_1$. The joint process $(\mathbf{W}_t, \mathbf{V}_t)$ is a zero mean, serially uncorrelated noise process with possibly time varying covariance matrices

$$\begin{pmatrix} \Sigma_{\mathbf{W}_t} & \Sigma_{\mathbf{W}_t,\mathbf{V}_t} \\ \Sigma_{\mathbf{V}_t,\mathbf{W}_t} & \Sigma_{\mathbf{V}_t} \end{pmatrix} \tag{3.3}$$

It is worth mentioning that some models make additional assumptions about the noise processes.

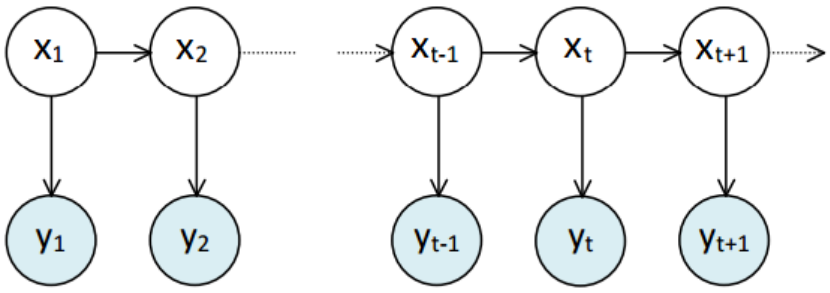


Figure 3.1: DAG for the state-space model with first order Markov latent dynamics

The functions f_θ and h_θ are parametrized by $\theta = (\theta_1, ..., \theta_n)^T$ and each θ_i is assumed to be independent from $(\theta_j)_{j \neq i}$. A joint prior distribution $p(\theta) = \prod_i p(\theta_i)$ is specified. With the assumptions mentioned above, the probability density of $\mathbf{x}_{1:T}$ is written as

$$p(\mathbf{x}_{1:T}|\theta) = p(\mathbf{x}_1|\theta) \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta) \tag{3.4}$$

The realization $\mathbf{x}_{1:T}$ is not observed directly, but through $\mathbf{y}_{1:T}$. The state-space model assumes that each observation $\mathbf{y}_t$ is statistically independent of every other quantity except $\mathbf{x}_t$ and θ, through Equation (3.2). As a consequence, the conditional likelihood of the observations, given the state process can be derived as

$$p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, \theta) = \prod_{t=2}^{T} p(\mathbf{y}_t|\mathbf{x}_t, \theta) = \int p(\mathbf{y}_T|\mathbf{x}_T, \theta) d\mathbf{y}_T \tag{3.5}$$

where $d\mathbf{y}_T$ is the Lebesgue measure. Here, θ is treated as unknown and the general idea is to estimate it using Maximum Likelihood Estimation (MLE) on the marginal likelihood $p(\mathbf{y}_{1:T}|\theta)$, with the latent variables $\mathbf{x}_{1:T}$ integrated out

$$p(\mathbf{y}_{1:T}|\theta) = p(\mathbf{y}_1|\theta) \prod_{t=2}^{T} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \theta) = \int p(\mathbf{y}_T|\mathbf{x}_T, \theta) p(\mathbf{x}_T|\mathbf{y}_{1:T-1}, \theta) d\mathbf{x}_T \tag{3.6}$$

It is also worth considering the approximation of the latent variables $p(\mathbf{x}_{1:T}, \theta|\mathbf{y}_{1:T})$. By Bayes theorem,

$$p(\mathbf{x}_{1:T}, \theta|\mathbf{y}_{1:T}) = \frac{p(\theta) p(\mathbf{x}_{1:T}|\theta) p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, \theta)}{p(\mathbf{y}_{1:T})} \tag{3.7}$$

where

$$p(\mathbf{y}_{1:T}) = \int p(\mathbf{y}_{1:T}|\theta') p(\theta') d\theta' = \iint p(\mathbf{y}_T|\mathbf{x}_T, \theta') p(\mathbf{x}_T|\mathbf{y}_{1:T-1}, \theta') p(\theta') d\mathbf{x}_T d\theta' \tag{3.8}$$

In most cases, $p(\mathbf{x}_{1:T}, \theta|\mathbf{y}_{1:T})$ is hard to compute because $p(\mathbf{y}_{1:T}|\theta)$ is analytically intractable. When θ is known, the problem of inference in the path space is effectively addressed using Sequential Monte Carlo methods. However, despite the success of standard SMC methods, the general case of the joint inference on θ and on $\mathbf{x}_{1:T}$ for a generic non-linear non-Gaussian state-space model is a very challenging problem, which, although extremely important for a wide variety of applications, is still somewhat unresolved. To attempt to overcome these difficulties, Andrieu et al. (2010) developed *Particle Markov Chain Monte Carlo* algorithms. These are MCMC algorithms which use a particle filter to estimate the intractable true value of (3.6). This framework is presented in more depth in Section 4.

3.2 Stochastic Volatility Models

The most important feature of the conditional return distribution $Y_t|\mathcal{F}_{t-1}$ is its variance dynamics. The first research on modelling this volatility was Engle (1982) with the famous heteroskedastic ARCH model. The main objective was to fit the fat tails of the return distributions and deal with volatility clustering. This section presents more recent models known as Stochastic Volatility models. The intrinsic feature of the SV model is that each observation y_t is assumed to have its own variance. In order to have a realistic model, this variance is not allowed to vary unrestrictedly with time. Rather, its logarithm is assumed to follow an autoregressive process of order one. It is worth noting that this feature is fundamentally different to GARCH models where the time-varying volatility is assumed to follow a deterministic instead of a stochastic evolution.

3.2.1 Model $\mathcal{M}_1$ - Standard Stochastic Volatility Model (SV)

The standard discrete-time SV model for the returns Y_t is defined as

$$X_t = \phi X_{t-1} + \sigma \epsilon_{X,t-1} \tag{3.9}$$

$$Y_t = \beta \exp\left(\frac{X_t}{2}\right) \epsilon_{Y,t} \tag{3.10}$$

where $\epsilon_{X,t}, \epsilon_{Y,t}$ are two independent and standard normally distributed processes. Let $\theta = (\rho, \sigma, \beta)$ be the parameters vector. This model is non-linear because of the nature of h_θ. X_t governs the volatility process of the observed returns Y_t, σ is the volatility of the volatility, ϕ the persistence parameter of X_t, and β is a scaling parameter. The condition $|\phi| < 1$ is imposed to have a stationary process, with initial condition $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right)$, where $\frac{\sigma^2}{1-\phi^2}$ is the unconditional variance of X_t. The next part explains the link between the stochastic volatility model and the Geometric Brownian Motion (GBM).

Definition 6. *A stochastic process S_t is said to follow a Geometric Brownian Motion if it satisfies the following stochastic differential equation $dS_t = \mu S_t dt + \sigma S_t dW_t$ where W_t is a Wiener process, μ the drift and σ the volatility. Both μ and σ are assumed to be constant.*

The process can be discretized by

$$\begin{aligned} S_{t+1} - S_t &= \mu S_t + \sigma S_t \epsilon_{t+1},\ \epsilon_t \sim \mathcal{N}(0,1) \\ S_{t+1} &= S_t + \mu S_t + \sigma S_t \epsilon_{t+1} \\ S_t &= S_{t-1} + \mu S_{t-1} + \sigma S_{t-1} \epsilon_t \end{aligned} \tag{3.11}$$

In the Stochastic Volatility model, Y_t represents the returns of the modelled asset. A general definition for computing the returns is $y_t = S_t/S_{t-1} - 1$, where S_t is the observed prices of the asset. When x_t is measured at time $t-1$ with regard to the filtration $\mathcal{F}_{t-1}$, $Y_t|x_t, \theta$ is normally distributed as

$$Y_t|x_t, \theta \sim \mathcal{N}(0, \beta^2 \exp(x_t))$$

$$S_t|x_t, \theta \sim \mathcal{N}(S_{t-1}, \underbrace{S_{t-1}^2 \beta^2 \exp(x_t)}_{\sigma^2(t)}) \tag{3.12}$$

The variance $\sigma^2(t)$ always exists as a product of square and exponential terms. Finally, Equation 3.12 can be rewritten as $S_t = S_{t-1} + \sigma(t) S_{t-1} \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$ which corresponds to the discretized Geometric Brownian Motion equation with $\mu = 0$ if and only if $\sigma(t) = \sigma$, $\forall t > 0$. The interest of using a Stochastic Volatility model essentially relies on the ability of modelling this volatility.

3.2.2 Model $\mathcal{M}_2$ - Stochastic Volatility Student-t (SVt)

The first extension is a stochastic volatility model with heavier tails where $\epsilon_{Y,t} \sim t(\nu)$. θ is enriched with the new parameter ν, supposed to be unknown.

Lemma 7. *Assume that X is a random variable of probability density function $f_X(x)$. The probability density function $f_Y(y)$ of $Y = g(X)$ where g is monotonic, is given by*

$$f_Y(y) = \left| \frac{d}{dy}(g^{-1}(y)) \right| \cdot f_X(g^{-1}(y)) \tag{3.13}$$

Applying this lemma on $y_t = \sigma(t)\epsilon_{Y,t}$ where $\sigma(t) = \beta \exp\left(\frac{x_t}{2}\right)$ and $g_t^{-1}(x) = \frac{x}{\sigma(t)}$ gives,

$$p(y_t|x_t, \theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \frac{1}{\sigma_t} \left(1 + \frac{y_t^2}{\sigma_t^2 \nu}\right)^{-\left(\frac{v+1}{2}\right)} \tag{3.14}$$

where $\Gamma(\cdot)$ is the gamma function. This result can also be retrieved by considering the t location-scale distribution with parameters $(\mu = 0, \sigma, \nu)$, whose probability density function is given by

$$p(x, \sigma, \nu, \mu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left[\frac{\nu + \left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right]^{-\left(\frac{\nu+1}{2}\right)} \tag{3.15}$$

The reasoning to find a closed form of $(S_t|x_t, \theta)$ is similar to the standard stochastic volatility model. Still under the assumption that $\epsilon_{Y,t} \sim t(\nu)$, if X has a t location-scale distribution, with parameters μ, σ, ν, then $\frac{X-\mu}{\sigma}$ has a Student's t distribution with ν degrees of freedom. Reverting the equation yields $X = \mu + \sigma\epsilon_{Y,t}$. Consequently,

$$S_t = \underbrace{S_{t-1}}_{\mu(t)} + \underbrace{\beta|S_{t-1}| \exp\left(\frac{x_t}{2}\right)}_{\sigma(t)} \epsilon_{Y,t} \tag{3.16}$$

Because the spread S_t can be negative, the absolute value $|.|$ is considered to ensure that $\sigma(t)$ is always positive definite. As a conclusion, $S_t|x_t, \theta$ follows a t location-scale distribution of parameters $(\mu(t), \sigma(t), \nu)$.

3.2.3 Model $\mathcal{M}_3$ - Stochastic Volatility Leverage (SVL)

In the second extension, a leverage effect is added. Black (1976) discovered that most measures of volatility of an asset are negatively correlated with the returns of that asset. It is considered nowadays as a stylized fact in econometrics series. Let ρ denote the correlation between the innovation processes $\epsilon_{X,t}$ and $\epsilon_{Y,t}$. θ is enriched with the new parameter ρ.

Lemma 8. *[Cholesky Decomposition] Let X, Y be two standard normally distributed random variables. The correlation between X and Y is ρ if and only if $Y = \rho X + \sqrt{1-p^2}Z$ where $Z \sim \mathcal{N}(0,1)$ and is independent of both X and Y.*

Applying the Cholesky decomposition on the innovations gives $\epsilon_{X,t} = \rho\epsilon_{Y,t} + \sqrt{1-\rho^2}Z$. This identity is helpful when it comes to generate artificial datasets from this model. Recall that $\epsilon_{Y,t}$ is first measured ($\mathcal{F}_{t-1}$-measurable) and $\epsilon_{X,t}$ is updated accordingly ($\mathcal{F}_t$-measurable). It means that $\epsilon_{Y,t}$ is independent from all the past values $(\epsilon_{X,s})_{s<t}$. Consequently, the conditional distributions of Y_t and S_t remain unchanged from Equation (3.12).

3.2.4 Model $\mathcal{M}_4$ - Stochastic Volatility Moving Average (SVMA)

The standard stochastic volatility model assumes that the errors in the measurement equation are serially independent. This is often an appropriate assumption for modelling financial data. To test this assumption, the plain model can be extended by allowing the errors in the measurement equation to follow a moving average (MA) process of order m. Here, we choose a simple specification and set $m = 1$. Hence, our model becomes

$$Y_t = \beta \exp\left(\frac{X_t}{2}\right)\epsilon_{Y,t} + \psi\beta \exp\left(\frac{X_{t-1}}{2}\right)\epsilon_{Y,t-1} \tag{3.17}$$

$$\begin{aligned} Y_t|\mathcal{F}_{t-1} &\sim \mathcal{N}\left(0, \beta^2\exp(x_t) + \psi^2\beta^2\exp(x_{t-1})\right) \\ S_t|\mathcal{F}_{t-1} &\sim \mathcal{N}\left(S_{t-1}, S_{t-1}^2\beta^2\exp(x_t) + S_{t-1}^2\psi^2\beta^2\exp(x_{t-1})\right) \end{aligned} \tag{3.18}$$

where the process X is defined in Equation (3.9). As before, we ensure that the root of the characteristic polynomial associated with the MA coefficient ψ, is outside the unit circle: $|\psi| < 1$. When $\psi = 0$, the SV-MA(1) model is reduced to the standard stochastic volatility model. The conditional variance of Y_t is given by $\mathrm{Var}(Y_t|\mathcal{F}_{t-1}) = \beta^2 e^{x_t} + \beta^2\psi^2 e^{x_{t-1}}$. The conditional variance is time-varying through two channels: a moving average composed of the two most recent variances $\beta^2 e^{x_t}$ and $\beta^2 e^{x_{t-1}}$ and secondly, according to the stationary $AR(1)$ process X_t.

3.2.5 Model $\mathcal{M}_5$ - Stochastic Mean (SVM)

Koopman and Hol Uspensky (2002) suggested an extension where the stochastic volatility also enters into the conditional mean equation. This model is known as the Stochastic Volatility in Mean (SVM). It is defined as

$$Y_t = \beta \exp\left(\frac{X_t}{2}\right) + \exp\left(\frac{X_t}{2}\right)\epsilon_{Y,t} \tag{3.19}$$

$$S_t|\mathcal{F}_{t-1} \sim \mathcal{N}\left(S_{t-1} + S_{t-1}\beta \exp\left(\frac{x_t}{2}\right), S_{t-1}^2 \exp\left(x_t\right)\right) \tag{3.20}$$

where X_t corresponds to the process of a standard stochastic volatility model defined in Equation (3.9). This model is pertinent if we believe that the conditional mean is somehow proportional to the conditional volatility.

3.2.6 Model $\mathcal{M}_6$ - Two Factors Stochastic Volatility (TFSV)

Chernov and Ghysels (2000) found that SV models with one volatility factor are not able to characterize all moments of asset return distributions. In particular, the fat tails of the return distribution are captured rather poorly. With a principal component analysis, Harvey et al. (1994) showed that a short-run and a long-run factors might be enough to explain the volatility of the returns. The study was performed on daily observations on several exchange rates. This model is known as the two factor stochastic volatility and relies on two different latent processes X and Z. It is defined as

$$X_t = \phi_X X_{t-1} + \sigma_X \epsilon_{X,t-1} \qquad |\phi_X| < 1, \epsilon_{X,t-1} \sim \mathcal{N}(0,1), X_1 \sim \mathcal{N}\left(0, \frac{\sigma_X^2}{1-\phi_X^2}\right) \tag{3.21}$$

$$Z_t = \phi_Z Z_{t-1} + \sigma_Z \epsilon_{Z,t-1} \qquad |\phi_Z| < 1, \epsilon_{Z,t-1} \sim \mathcal{N}(0,1), Z_1 \sim \mathcal{N}\left(0, \frac{\sigma_Z^2}{1-\phi_Z^2}\right) \tag{3.22}$$

$$Y_t = \beta \exp\left(\frac{X_t + Z_t}{2}\right) \epsilon_{Y,t} \qquad \epsilon_{Y,t} \sim \mathcal{N}(0,1) \tag{3.23}$$

Under these assumptions, the conditional distribution of the spread is $S_t|x_t, z_t, \theta \sim \mathcal{N}\left(S_{t-1}, S_{t-1}^2\beta^2 \exp\left(x_t + z_t\right)\right)$. The parameters vector θ is now $(\beta, \phi_X, \phi_Z, \sigma_X, \sigma_Z)$. It is of common knowledge that the returns are leptokurtic, i.e. with a positive kurtosis. Veiga (2006) showed that the second term introduced in the model helps generate extra kurtosis and accounts for short-run dynamics.

3.2.7 Model $\mathcal{M}_7$ - Two Factors Stochastic Volatility with Leverage (TFSVL)

In the final extension, we consider the two factors stochastic volatility with a correlation $\rho = corr(\epsilon_{X,t}, \epsilon_{Y,t})$. The idea is the same as the one developed for the model $\mathcal{M}_3$. We assume a non-zero correlation between the innovations of the returns and the long-run factor X from Equation (3.21). From the models presented before, this model is by far the most complex because 6 parameters are to be estimated: $\beta, \rho, \phi_X, \phi_Z, \sigma_X$ and σ_Z. Ruiz and Veiga (2008) studied a slightly different version with X_t defined as a fractional integrated Gaussian noise process (ARFIMA). They proved that the first order autocorrelation $corr(|y_t|, |y_{t+1}|)$ is smaller than the second order autocorrelation $corr(y_t^2, y_{t+1}^2)$ when $\rho < 0$. As explained by Cont (2005), it is usually the case in practical applications. From the same considerations as in model $\mathcal{M}_3$, the conditional distributions of Y_t and S_t remains unchanged compared to model $\mathcal{M}_6$.

4 Sequential Monte Carlo and Particle Markov Chain Monte Carlo

4.1 From SMC to PMCMC in Nonlinear Filtering

Many problems involve making inference on unknown parameters of complex models which have a sequential, if not explicitly temporal, basis. In the first part, Sequential Monte Carlo (SMC) methods are introduced followed by a presentation of the Particle Marginal Metropolis Hastings (PMMH), an implementation of the Particle MCMC scheme. This chapter keeps the same notations as in Chapter 3.1 where the state-space models are defined.

Sequential Monte Carlo (SMC) are a collection of simulation-based techniques to address nonlinear filtering problems arising in signal processing and Bayesian statistical inference. The filtering problem consists in estimating the latent states $\mathbf{x}_{1:T}$ when only the observations $\mathbf{y}_{1:T}$ are known. The objective of SMC is to compute a recursive series of posterior distributions over such complex models. SMC methods are very flexible, relatively easy to implement and parallelizable. Since computational resources have become so readily available, and due to certain recent advanced in applied statistics, these methods have recently become a mainstay of advanced research methods in this field.

Particle Markov Chain Monte Carlo (Particle MCMC) are powerful techniques for estimating parameters of a complex model where classical methods are limited. Examples of complex models include state-space models with latent variables. The main feature behind Particle MCMC is the use of a particle filter of size N inside a MCMC scheme. This filter provides estimates of the intractable marginal likelihood $p(\mathbf{y}_{1:T}|\theta)$.

4.2 Monte Carlo Methods

Sequential Monte Carlo (also known as Particle Filtering) paradigm is based on Monte Carlo methods such as rejection sampling and importance sampling. This section introduces both of them briefly.

4.2.1 Rejection Sampling

Rejection sampling is a technique which samples from a target distribution $p(\mathbf{X})$ (known up to a proportional constant) by sampling from another easy to sample proposal distribution $\pi(\mathbf{X})$. This assumes that there exists a known finite constant C such that

$p(\mathbf{X}) \leq C\pi(\mathbf{X})$ for every $\mathbf{x}$. The idea is to draw $\mathbf{X} \sim \pi$ and accept it as a sample of p with probability $p(\mathbf{X})/(C\pi(\mathbf{X}))$.

4.2.2 Importance Sampling

The idea of importance sampling is to choose a proposal distribution $\pi(\mathbf{X})$ in place of the true probability distribution $p(\mathbf{X})$, which is difficult to sample. The support of $\pi(\mathbf{X})$ is assumed to cover that of $p(\mathbf{X})$. Under these assumptions, the classic Monte Carlo integration problem is

$$p(f) = E[f(\mathbf{X})] = \int f(\mathbf{X})p(d\mathbf{X}) \tag{4.1}$$

for any suitable function f. It can be rewritten as

$$\int f(\mathbf{X})p(\mathbf{X})d\mathbf{X} = \int f(\mathbf{X})\frac{p(\mathbf{X})}{\pi(\mathbf{X})}\pi(\mathbf{X})d\mathbf{X} \tag{4.2}$$

Importance sampling is used to draw a number of independent samples from $\pi(\mathbf{X})$ to obtain an estimate of Equation (4.2). Each sample, $f(\mathbf{X}_i)$ is assigned an importance weight, $W(\mathbf{X}_i) \propto p(\mathbf{X}_i)/\pi(\mathbf{X}_i)$. In practice, the variance of the importance weights must be finite and the proposal distribution $\pi(\mathbf{X})$ should be as close to possible to $p(\mathbf{X})$ such that the variance of the weights is minimized.

4.3 Sequential Monte Carlo Methods

4.3.1 Sequential Importance Sampling Resampling

The Sequential Importance Sampling Resampling (SIS-R) algorithm is a sequential version of importance sampling with resampling which approximates the filtering probability density $p(\mathbf{x}_t|y_{1:t})$ of the latent variables by a weighted set of N samples $(w_t^{(i)}, \mathbf{x}_t^{(i)})_{1\leq i\leq N}$, where N is the number of particles. The importance weights $w_t^{(i)}$ are the approximations to the relative posterior distributions. This algorithm has been introduced in 1987 to tackle the degeneracy problem, unresolved by its previous version, known as Sequential Importance Sampling (SIS). The degeneracy happens when all but one of the importance weights are close to zero. The correction is done by adding a resampling step to eliminate samples with trivial importance weights and propagate particles with larger weights. As in importance sampling, the expectation of a function f can be approximated as a weighted average

$$\int f(\mathbf{x}_t)p(\mathbf{x}_t|y_{1:t})d\mathbf{x}_t \simeq \sum_{i=1}^{N} w_t^{(i)} f(\mathbf{x}_t^{(i)}) \tag{4.3}$$

This step is presented in Algorithm 2. The input is the set of particles at time t and the output is this same set resampled with the weights being re-initialized. N is the number of particles in the particle filter.

Algorithm 2 Multinomial Resampling

procedure MULTINOMIALRESAMPLING($(w_t^{(i)}, \mathbf{x}_t^{(i)})_{1 \leq i \leq N}$, N)
  for i from 1 to N **do**
    Sample an integer $j \in [1, N]$ with probabilities proportional to $\{w_t^{(1)}, ..., w_t^{(N)}\}$
    Replace the current particle i with this new one. $\mathbf{x}_t^{(i)} \leftarrow \mathbf{x}_t^{(j)}$
    Re initialize the weight $w_t^{(i)} = 1/N$
  end

The general algorithm SIR is presented in Algorithm 3. The input consists in the observations $y_{1:T}$, the parameters vector θ and the number of particles N. The output is the estimated latent states $\mathbf{x}_{1:T}^*$ and the estimator $\hat{p}_\theta^N(y_{1:T})$. Generally, the resampling step is not performed at each time but at some specific steps to improve the performance. It is controlled by the Effective Sample Size (ESS) parameter which gives an estimate of the effective number of particles. When it is too low, it means that the algorithm is performing poorly and a resampling is necessary. A threshold N_{thr} is arbitrary selected to trigger this resampling. Usually, it is equal to $N/3$ but can vary depending on the applications.

Algorithm 3 Sequential Importance Sampling Resampling (SISR)

procedure SISR($y_{1:T}$, θ, N, N_{thr})
  for i from 1 to N **do**
    Sample $\mathbf{x}_1^{(i)} \sim \pi(\mathbf{x}_1)$
    Calculate weight $w_1^{(i)} = \dfrac{p(y_1|\mathbf{x}_1)p(\mathbf{x}_1)}{\pi(\mathbf{x}_1|y_{1:t})}$
  end
  $\mathbf{x}_1^* = \sum_{i=1}^N \mathbf{x}_1^{(i)}.w_1^{(i)}$
  Set $\hat{p}_\theta^N(y_1) = \frac{1}{N}\sum_{i=1}^N w_1^{(i)}$
  for t from 2 to T **do**
    for i from 1 to N **do**
      Draw sample from the proposal distribution $\mathbf{x}_t^{(i)} \sim \pi\left(\mathbf{x}_t|\mathbf{x}_{1:t-1}^{(i)}, y_{1:t}\right)$
      Calculate weight $\hat{w}_t^{(i)} = w_{t-1}^{(i)} \dfrac{p(y_t|\mathbf{x}_t^{(i)})p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{\pi\left(\mathbf{x}_t^{(i)}|\mathbf{x}_{1:t-1}^{(i)}, y_{1:t}\right)}$
    end
    Compute the normalized importance weight for each i, $w_t^{(i)} = \dfrac{\hat{w}_t^{(i)}}{\sum_{j=1}^N \hat{w}_t^{(j)}}$
    Compute the ESS as $N_{eff} = 1/\left(\sum_{i=1}^N \left(w_t^{(i)}\right)^2\right)$
    if $N_{eff} < N_{thr}$ **then**
      //perform resampling step
    end
    $\mathbf{x}_t^* = \sum_{i=1}^N \mathbf{x}_t^{(i)}.w_t^{(i)}$
    Set $\hat{p}_\theta^N(y_{1:t}) = \hat{p}_\theta^N(y_{1:t-1})\left(\frac{1}{N}\sum_{i=1}^N w_t^{(i)}\right)$
  end
return ($\mathbf{x}_{1:T}^*$, $\hat{p}_\theta^N(y_{1:T})$)

It is worth noting that $\mathbf{x}$ can be multivariate. Models exist where the observations are generated from two or more latent processes (models $\mathcal{M}_6$ and $\mathcal{M}_7$ are such examples). In the case where $\mathbf{x}$ is bivariate, the particle filter is updated to draw two sets of particles instead of one (one for $\mathbf{x}_1$ and one for $\mathbf{x}_2$).

Particle filters with transition prior probability distribution f_θ as importance function $\pi(\mathbf{X})$ are commonly known as Bootstrap filter. This choice is motivated by the facility of drawing particles and performing subsequent importance weight calculations. Here, $\pi(\mathbf{x}_t|\mathbf{x}_{1:t-1}, y_{1:t}) = p(\mathbf{x}_t|\mathbf{x}_{1:t-1}, y_{1:t})$. Coupled with a resampling done at each step, set by $N_{thr} = \infty$, the weights formula is simplified to

$$w_t^{(i)} = \frac{p(y_t|\mathbf{x}_t^{(i)})p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{\pi(\mathbf{x}_t^{(i)}|\mathbf{x}_{1:t-1}^{(i)}, y_{1:t})} = p(y_t|\mathbf{x}_t^{(i)}) \tag{4.4}$$

It is clear from the understanding of importance resampling that these weights are appropriate for representing a sample from $p(\mathbf{x}_t|y_{1:t})$, and so the particles and weights can be propagated forward to the next time point. It is also clear that the average weight at each time gives an estimate of the marginal likelihood of the current data point given the data so far. So we define the conditional marginal of y_t by

$$\hat{p}_\theta^N(y_t|y_{1:t-1}) = \frac{1}{N}\sum_{k=1}^{N} w_t^k \tag{4.5}$$

and the conditional marginal estimator of $y_{1:T}$ over all the state-space is

$$\hat{p}_\theta^N(y_{1:T}) = \hat{p}_\theta^N(y_1)\prod_{t=2}^{T}\hat{p}_\theta^N(y_t|y_{1:t-1}) = \prod_{t=1}^{T}\left(\frac{1}{N}\sum_{k=1}^{N} w_t^k\right) \tag{4.6}$$

Again, from the importance resampling scheme, it should be reasonably clear that $\hat{p}_\theta^N(y_{1:T})$ is a consistent estimator of $p_\theta(y_{1:T})$. It is much less obvious, but nevertheless true that this estimator is also unbiased, according to Del Moral (2004). This result is the cornerstone of Particle MCMC models. As T is usually large, it is usually preferred to work with log likelihoods

$$\log p_\theta(y_{1:T}) = \log p_\theta(y_1) + \sum_{t=2}^{T}\log p_\theta(y_t|y_{1:t-1}) \tag{4.7}$$

$$\log \hat{p}_\theta^N(y_{1:T}) = \sum_{t=1}^{T}\log\left(\frac{1}{N}\sum_{k=1}^{N} w_t^k\right) \tag{4.8}$$

4.4 Resampling Methods

SISR can be decomposed in two main steps: sequential importance sampling (SIS) and resampling. To balance performance and accuracy, it is necessary to perform resampling

sufficiently often. This step is also time-critical as highlighted by the performance benchmarks we ran: resampling represents on average half of the time spent in the filter. Many resampling methods exist in the literature: multinomial, stratified, systematic, residuals... Despite the lack of complete theoretical analysis of its behavior, multinomial resampling is probably the most used algorithm because almost all software products offer a default implementation of this method. Recently, Douc and Cappé (2005) found an interesting result:

Theorem 9. *Stratified resampling has a lower conditional variance compared to multinomial resampling.*

Proof. See Appendix 10.2. □

From a mathematical point of view, the stratified resampling is compelling. A benchmark consisting in resampling 1000 weights numerous times was run and the results are gathered in Table 4.1. The stratified resampling seems to offer the best balance in terms of speed and variance. For this reason, we choose it as default resampling method inside our particle filters.

Resampling method (no parallel)	Min	Max	Avg	Std	95% Conf. Int.
Residual	16.37	25.33	18.59	1.68	[16.58,22.83]
Stratified	0.58	1.11	0.63	0.08	[0.58,0.94]
Systematic	0.57	1.17	0.64	0.10	[0.58,1.07]
Multinomial	1.65	3.23	1.84	0.26	[1.66,2.59]

Table 4.1: Time spent to resample 10^5 times 1000 weights (in seconds)

4.5 Particle Markov Chain Monte Carlo

4.5.1 Particle Marginal Metropolis-Hastings Algorithm

In the classic MCMC scheme, the Metropolis Hastings (MH) algorithm is used to target $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$ with the ratio

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)} \times \frac{p(\mathbf{y}|\theta^\star)}{p(\mathbf{y}|\theta)}\right) \tag{4.9}$$

where $q(\theta^\star|\theta)$ is the proposal density. As discussed before, the marginal likelihood $p(\mathbf{y}|\theta) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}|\theta)d\mathbf{x}$ is often intractable and the ratio becomes impossible to compute. The simple likelihood-free scheme targets the full joint posterior $p(\theta, \mathbf{x}|\mathbf{y})$. Usually the knowledge of the kernel f_θ makes $p(\mathbf{x_{1:T}}|\theta)$ tractable. For instance, a path $\mathbf{x}_{1:T}$ governed by a linear Gaussian process $\mathbf{X}_t = \rho\mathbf{X}_{t-1} + \tau\epsilon_{t-1}$, $\epsilon_t \sim \mathcal{N}(0,1)$ can be easily simulated as long as ρ, τ and $\mathbf{x}_1$ are known quantities.

The MH is built in two stages. First, a new candidate θ^* is proposed from $q(\theta^\star|\theta)$. Then, $\mathbf{x}^*$ is sampled from $p(\mathbf{x}^\star|\theta^\star)$. The generated pair $(\theta^\star, \mathbf{x}^\star)$ is accepted with the ratio

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)} \times \frac{p(\mathbf{y}|\mathbf{x}^\star, \theta^\star)}{p(\mathbf{y}|\mathbf{x}, \theta)}\right) \tag{4.10}$$

At each step, $\mathbf{x}^*$ is consistent with θ^* because it has been generated from $p(\mathbf{x}^\star|\theta^\star)$. The problem of this approach is that the sample $\mathbf{x}^*$ may not be consistent with $\mathbf{y}$. As T grows, it becomes nearly impossible to iterate over all possible values of $\mathbf{x}^\star$ to track $p(\mathbf{y}|\mathbf{x}^\star, \theta)$. This is why $\mathbf{x}^*$ should be sampled from $p(\mathbf{x}^\star|\theta^\star, \mathbf{y})$. Under this assumption, the ratio now becomes

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \frac{p(\mathbf{x}^\star|\theta^\star)}{p(\mathbf{x}|\theta)} \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)} \frac{p(\mathbf{y}|\mathbf{x}^\star, \theta^\star)}{p(\mathbf{y}|\mathbf{x}, \theta)} \frac{p(\mathbf{x}|\mathbf{y}, \theta)}{p(\mathbf{x}^\star|\mathbf{y}, \theta^\star)}\right) \tag{4.11}$$

Using the basic marginal likelihood identity described in Chib (1995), the ratio is simplified to

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \times \frac{p(\mathbf{y}|\theta^\star)}{p(\mathbf{y}|\theta)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)}\right) \tag{4.12}$$

It is now clear that a pseudo-marginal MCMC scheme for state-space models can be derived by substituting $\hat{p}_\theta^N(y_{1:T})$, computed from a particle filter, in place of $p_\theta(y_{1:T})$. This turns out to be a simple special case of the particle marginal Metropolis-Hastings (PMMH) algorithm described in Andrieu et al. (2010) and in Algorithm 4. Remarkably $\mathbf{x}$ is no more present and the ratio is exactly the same as the classical marginal scheme shown before in Equation (4.9). Indeed, the ideal marginal scheme corresponds to PMMH when $N \to \infty$. The likelihood-free scheme is obtained with just one particle in the filter. When N is intermediate, the PMMH algorithm is a trade-off between the ideal and the likelihood-free schemes, but is always likelihood-free when one Bootstrap particle filter is used. The PMMH algorithm proposed by Andrieu et al. (2010) is an MCMC algorithm for state-space models jointly updating θ and $\mathbf{x}_{1:T}$. A new $\theta^\star$ is proposed from $q(\theta^\star|\theta)$ and the corresponding $\mathbf{x}_{1:T}^\star$ is generated by running a Bootstrap particle filter with $\theta^\star$ The pair $(\theta^\star, x_{1:T}^\star)$ is accepted using the Metropolis-Hastings ratio

$$\min\left(1, \frac{\hat{p}_{\theta^\star}(y_{1:T})}{\hat{p}_\theta(y_{1:T})} \times \frac{p(\theta^\star)}{p(\theta)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)}\right) \tag{4.13}$$

where $\hat{p}_{\theta^\star}^N(y_{1:T})$ is the particle filter's unbiased estimate of marginal likelihood. Note that the terms $p(\cdot)$ and $q(\cdot|\cdot)$ cancel out when the proposal densities correspond to the respective prior distributions.

Due to the unbiasedness property of $\hat{p}_{\theta^\star}^N(y_{1:T})$, the PMMH algorithm works for any positive N. In practical applications, a critical issue resides in how to choose the number of particles N. A large N gives a more accurate estimate of the log likelihood at a greater computational cost, while a small N would lead to a larger variance.

Algorithm 4 Particle pseudo marginal Metropolis-Hastings Algorithm

procedure PMMH($y_{1:T}$, a proposal distribution $q(\cdot|\cdot)$, the number of particles N, the number of MCMC steps M)
  Set static parameter vector $\theta^{(1)}$ arbitrarily
  $\hat{p}^N_{\theta^{(1)}}(y_{1:T}), \mathbf{x}^{*(1)}_{1:T} \leftarrow$ Call Bootstrap Particle Filter with $(y_{1:T}, \theta^{(1)}, N)$
  for i from 2 to M **do**
    Sample θ' from $q(\theta'|\theta^{(i-1)})$
    $\hat{p}^N_{\theta'}(y_{1:T}), \mathbf{x}^{*'}_{1:T} \leftarrow$ Call Bootstrap Particle Filter with $(y_{1:T}, \theta', N)$
    With probability,

$$\min\left\{1, \frac{q(\theta^{(i-1)}|\theta')\hat{p}_N(y_{1:T}|\theta')p(\theta')}{q(\theta'|\theta^{(i-1)})\hat{p}_N(y_{1:T}|\theta^{(i-1)})p(\theta^{(i-1)})}\right\}$$

    Set $\mathbf{x}^{*(i)}_{1:T} \leftarrow \mathbf{x}^{'*}_{1:T}, \theta^{(i)} \leftarrow \theta', \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta'}(y_{1:T})$
    Otherwise $\mathbf{x}^{*(i)}_{1:T} \leftarrow \mathbf{x}^{*(i-1)}_{1:T}, \theta^{(i)} \leftarrow \theta^{(i-1)}, \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta^{(i-1)}}(y_{1:T})$
  end
return $(\mathbf{x}^{*(i)}_{1:T}, \theta^{(i)})_{i=1}^M$

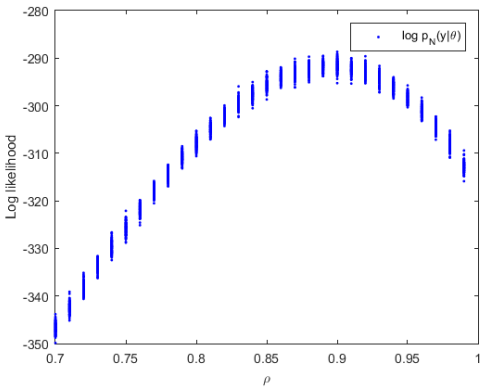
4.6 Tuning the Number of Particles N

Tran et al. (2014) showed that the efficiency of estimating an intractable likelihood using Bayesian inference and importance sampling is weakly sensitive to N around its optimal value. Furthermore, the loss of efficiency decreases at worse linearly when we choose N higher than the optimal value, whereas the efficiency can deteriorate exponentially when N is below the optimal. Pitt et al. (2012) showed that we should choose N so that the variance of the resulting log-likelihood is around 0.85. Of course, in practice this variance will not be constant, as it is a function of the parameters as well as a decreasing function of N. Pitt et al. (2012) suggests that a reasonable strategy is to estimate the posterior mean $\bar{\theta} = E[\theta|y_{1:T}]$ from an initial short run with N set to a large value. The value of N could then be adjusted such that the variance of the log-likelihood $\text{Var}(\log \hat{p}_N(y|\bar{\theta}))$ evaluated at $\bar{\theta}$ is around 0.85. The penalty for getting the variance wrong is not too severe within a certain range. Still from Pitt et al. (2012), their results indicated that although a value of $0.92^2 = 0.8464$ is optimal, the penalty is small provided the value is between 0.25 and 2.25. This allows for quite a large margin of error in choosing N and also suggests that the simple schemes advocated should work well.

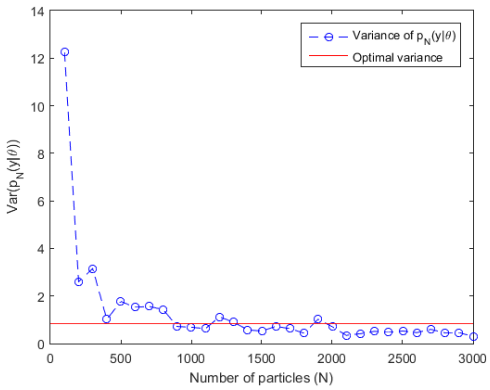
An analysis was carried out to measure the variations of the variance across the parameter space for different values of N. We use the state-space model $\mathcal{M}_2$ to generate $T = 1000$ synthetic daily returns with $\theta_{tr} = (\rho, \sigma, \nu) = (0.91, 1, 3)$. The Bootstrap filter is called repeatedly to estimate its intrinsic variance. Figure (4.1a) shows the behavior its variance when ρ varies over its domain of definition. It gives us a hint that the vari-

ance is not likely to oscillate in big proportions when θ change.

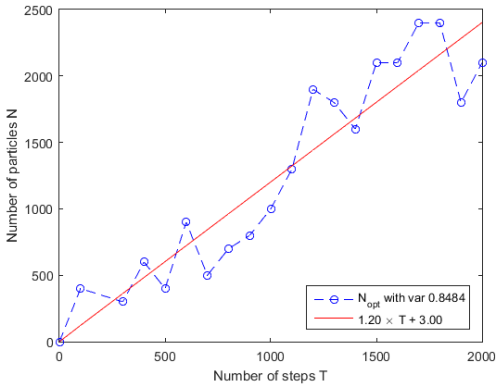
The reasonable strategy of Pitt et al. (2012) is not viable in practice as it requires to have a good estimate of $\bar{\theta}$ which is often difficult to achieve with a short run of PMCMC due to the burn-in phase. To some extent, it is more appropriate to derive a general rule on how to choose N optimal, provided that such a rule exists. We conduct a test on this same synthetic data set. For a given value of N, the Bootstrap filter of $\mathcal{M}_2$ is called several times and the variance of the log likelihood $\mathrm{Var}(\log \hat{p}_N(y|\bar{\theta}))$ is estimated. The process is repeated for different values of N. From Figure (4.1b), the optimal of N seems to be around 1000. The process is repeated for several values of T to detect a general rule. Figure 4.1c shows the results for $T \in [0, 2000]$ and $N \in [0, 2500]$. A linear trend can easily be identified. To reinforce this belief, a linear regression $N = aT + b$ is fitted. Both the values $b \simeq 0(3)$ and $a \simeq 1(1.2)$ suggest that the rule $T = N$ seems to hold, at least for $T < 2000$.



(a) $\mathrm{Var}(\log \hat{p}_N(y|\theta))$ when θ varies through ρ. Dataset generated from $\mathcal{M}_2$ with $(\rho, \sigma, \nu) = (0.91, 1, 3)$.



(b) $\mathrm{Var}(\log \hat{p}_N(y|\bar{\theta}))$ for different values of N. Dataset generated from $\mathcal{M}_2$ with $T = 1000$ and $(\rho, \sigma, \nu) = (0.91, 1, 3)$



(c) Behavior of N_{opt} when T varies

Figure 4.1: Finding the optimal number of particles N

5 Model Selection and Estimation

In practical applications, the true value of θ_{tr} is usually unknown and it makes the validation becoming harder. The validation is an important pre-task because it tests the implementation, the choice of the priors and the proposal distributions, and measures the dispersion of the estimator $\hat{\theta}$ to the true value θ_{tr}. The first step involves the sampling generation of both the process and the observations (X_t, Y_t) from a model $\mathcal{M}_x$. We choose an arbitrary realistic value for θ_{tr}. At this point, $x^{tr}_{1:T}$ and $y^{tr}_{1:T}$ are sampled. Each model takes $y^{tr}_{1:T}$ as argument and outputs an estimator $(\hat{x}_{1:T}, \hat{\theta})$. The estimated values are then compared to the true values using some dispersion measures such as the MSE defined by $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta_{tr})^2]$. It is also interesting to cross validate the models. The marginal likelihood of the data $p(y^*_{1:T})$ should be maximal for $\mathcal{M}_x$. If the parameters are estimated by another model $\mathcal{M}_y$ say, we should have $p(y^*_{1:T}|\mathcal{M}_x) > p(y^*_{1:T}|\mathcal{M}_y)$ according to the likelihood principles. Each of the 7 models presented in Section 3.2 has been successfully validated. The source code of the validation is available in the folder `models` of the repository (Appendix 10.1.1).

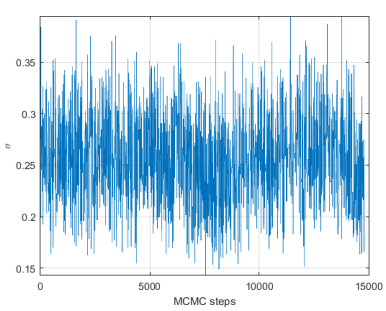
5.1 Parameter Estimation on Real Data

Once the models have been validated, they can be fitted to real world data. The dataset used here is $\mathcal{EQTY}_{daily}$, presented in 1.2.1. The number of steps of Particle MCMC is taken large enough to ensure that there is a sufficient number of samples available to form the Bayesian posterior distributions $\mathcal{D}(\theta|y_{1:T})$. Unless stated otherwise, the PMCMC scheme algorithm will loop 10000 times before stopping. The first 1000 samples are discarded for each parameter because the chains require several steps to reach their equilibrium distribution (burn-in). A component-wise scheme is used to update the parameters, i.e. one by one sequentially. Note that it is possible to parallel this scheme by introducing a bias. However, a more efficient way is to parallel the filter, still with a bias. Both algorithms have been implemented and are available in the folder `pmcmc` (Appendix 10.1.1). Because the bias has not been rigorously evaluated, a non parallel version was used for the computations throughout the scope of the thesis. Once the burn-in phase is performed, the mean value $\bar{\theta}$ is selected from the distribution $\mathcal{D}(\theta|y_{1:T})$ as the best estimation for θ_{tr}. Some statistics, moments and confidence intervals can be obtained from $\mathcal{D}(\theta|y_{1:T})$. It is also important to choose correctly the prior distributions $p(\theta)$ and the proposal densities $q(\theta|\theta')$ to maintain a good acceptance rate. Roberts et al. (1997) showed that the optimal acceptance rate is 0.234 under quite general conditions. Table 5.1 summarizes such an analysis for model $\mathcal{M}_5$ on the stock APPL for the period 2003-Sep-09 - 2006-Jun-04. Figure 5.1 exposes some checks on the posterior distribution

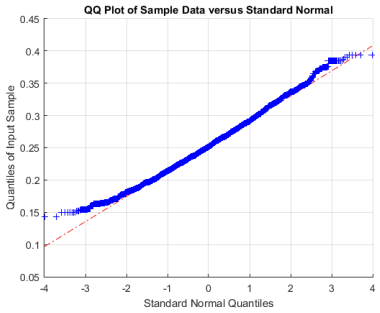
$p(\sigma|y_{1:T}, \mathcal{M}_5)$. The chain mixes well with an acceptance rate of 0.180, close to the 0.234 optimal value of Roberts et al. (1997). According to Figures 5.1b and 5.1c, the posterior distribution seems to be normally distributed, with a skewness of 0.224 and a kurtosis of 2.945. Finally, the autocorrelation function of the chain is fast decaying, which assesses the good performance of the sampling.

Parameter	ρ	σ	β
Mean	0.9981	0.2533	0.1475
Median	0.9982	0.2514	0.1448
Max	0.9991	0.3941	0.2189
Min	0.9865	0.1434	0.1100
Conf Int (95%)	[0.9904, 0.9989]	[0.1822, 0.3345]	[0.1242, 0.1839]
Acceptance Rate	0.11	0.18	0.15
MCMC Steps	10000	10000	10000
Burn-in	1000	1000	1000
$p(\theta)$	$\mathcal{U}[-1, 1]$	$\mathcal{IG}(1, 1)$	$\mathcal{IG}(1, 1)$
$q(\theta\|\theta')$	$\mathcal{N}(\theta', 0.1^2)$	$\mathcal{N}(\theta', 0.1^2)$	$\mathcal{N}(\theta', 0.1^2)$

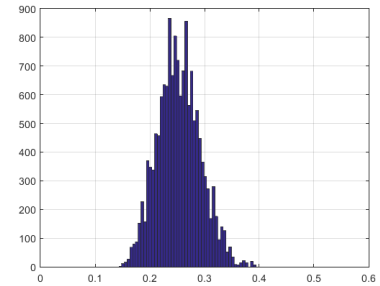
Table 5.1: Parameters estimation with model $\mathcal{M}_5$. Ticker is APPL. Sep, 09 2003 - Jun, 04 2006.



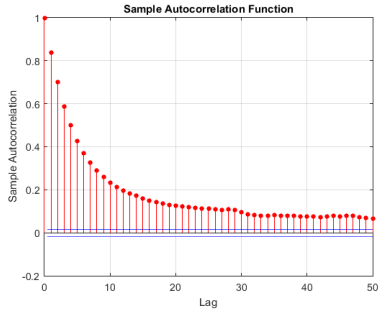
(a) MCMC Chain for σ



(b) QQ Plot - σ



(c) Posterior distribution $p(\sigma|y_{1:T})$



(d) Autocorrelation function of σ

Figure 5.1: MCMC Checks for $p(\sigma|y_{1:T}, \mathcal{M}_5)$. APPL - Sep, 09 2003 - Jun, 04 2006.

5.2 Model Selection

5.2.1 Methodology

The output of the particle filter is an unbiased estimate of $p(y_{1:T}|\theta)$, with the unobserved states integrated out. Although it is very tempting to use it as a measure to compare models, it is always preferred to use the true marginal likelihood $p(y_{1:T})$. According to Bayesian theory, the marginal likelihood for a model $\mathcal{M}$ is defined as

$$p(y_{1:T}|\mathcal{M}) = \int p(y_{1:T}|\theta, \mathcal{M})p(\theta|\mathcal{M})d\theta \tag{5.1}$$

Gelfand and Dey (1994) proposed a very general estimate for this marginal likelihood

$$\left(\frac{1}{M}\sum_{i=1}^{M}\frac{g(\theta_i)}{p(y_{1:T}|\theta_i)p(\theta_i)}\right)^{-1} \rightarrow p(y_{1:T}) \text{ as } M \rightarrow \infty \tag{5.2}$$

For this estimator to be consistent, $g(\theta_i)$ must be thin-tailed relative to the denominator. Gelfand and Dey (1994) argued that for most cases, a multivariate normal distribution $\mathcal{N}(\theta^*, \Sigma^*)$ can be used, where θ^* and Σ^* are equal to the empirical mean and sample unbiased variance, $\theta^* = \frac{1}{M}\sum_{i=1}^{M}\theta^i$ and $\Sigma^* = \frac{1}{M-1}\sum_{i=1}^{M}\left(\theta^i - \theta^*\right)\left(\theta^i - \theta^*\right)^T$.

The difficulty of this approach resides in its implementation. By its definition, $p(y_{1:T}|\theta)$ is usually either very close to 0 or very big as the size of the state-space T grows. The trick here is to consider the sum of the exponential of the logarithms and factorize by the maximum logarithm to avoid rounding errors. For example, let $M = 3$ and assume that the log-terms on the LHS are equal to -120, -121 and -122. Thereby, we have

$$\begin{aligned} p(Y_T)^{-1} &= e^{-120} + e^{-121} + e^{-122} \\ -\log p(Y_T) &= \log(e^{-120}(1 + e^{-1} + e^{-2})) \\ \log p(Y_T) &= 120 - \log(1 + e^{-1} + e^{-2})) \simeq 119.6 \end{aligned}$$

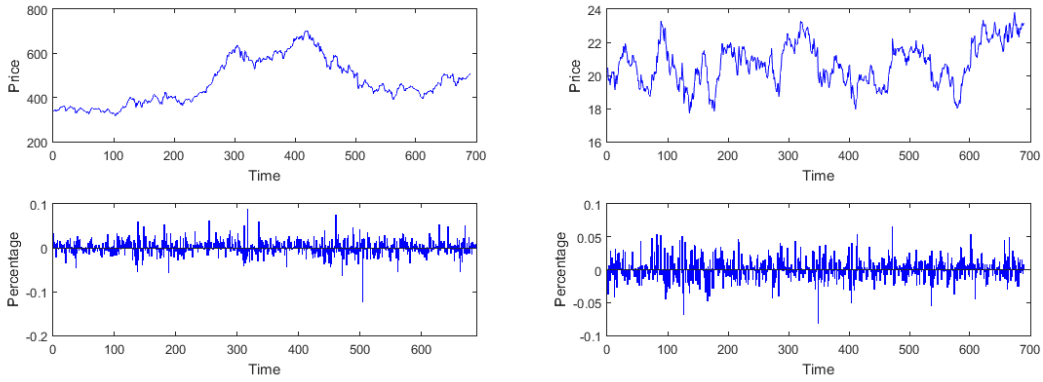
When $p(Y_T|\mathcal{M}_\mathcal{A})$ and $p(Y_T|\mathcal{M}_\mathcal{B})$ are estimated, Kass and Raftery (1995) suggests to use twice the logarithm of the Bayes factor for model comparison $2\log BF_{\mathcal{M}_{\mathcal{AB}}}$, where $\mathcal{M}_{\mathcal{AB}}$ is the Bayes Factor of $\mathcal{M}_\mathcal{A}$ to $\mathcal{M}_\mathcal{B}$. **The evidence of $\mathcal{M}_\mathcal{A}$ over $\mathcal{M}_\mathcal{B}$ is based on a rule-of-thumb: 0 to 2 not worth more than a bare mention, 2 to 6 positive, 6 to 10 strong, and greater than 10 as very strong.**

5.2.2 Results

It is interesting to see how models perform in practical applications. When it seems pretty obvious that using a leverage can be pertinent according to stylized facts, it seems less evident that the mean of the returns exhibits a stochastic mean proportional to its volatility. The best model is selected on a group composed of several cointegrated spreads on different periods. The computationally intensive property makes it

difficult to test every model for every spread. A sample of 10 spreads is considered randomly beforehand across different sectors such as Energy, Information Technology and Financial. The conclusion is fairly clear on the random group. It turns out that $\mathcal{M}_7$ outperforms all the other models in every situation, in terms of marginal likelihood and AIC. On average, the Kass Factor of $\mathcal{M}_7$ over $\mathcal{M}_6$ is between 2 and 10, showing a positive evidence. The $\mathcal{M}_2$ and $\mathcal{M}_3$ model are respectively ranked third and fourth. Therefore, model $\mathcal{M}_7$ is selected as the most compelling model for the rest of the analysis.

This section also introduces an example with detailed explanations about the procedure we used, for a particular spread and a particular stock. The results are confronted to detect if the inferences are in accordance.



(a) APPL stock

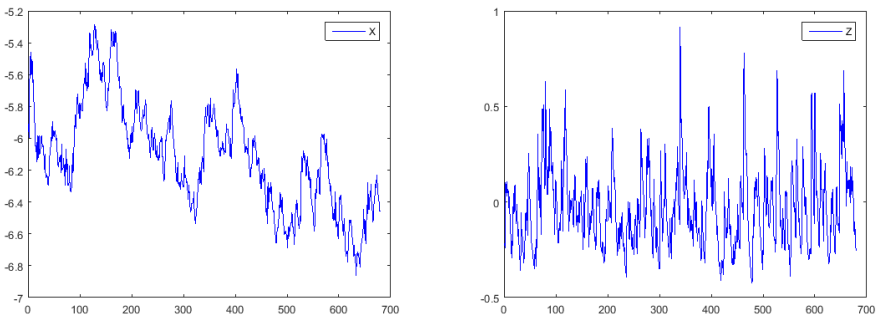
(b) Spread AMR CORP - CRANE CO - DOVER CORP. $\beta = (1, -0.0865, -0.3796)$

Figure 5.2: Stock APPL and Spread AMR-CRANE-DOVER. Period is from 09-Sep-2003 to 04-Jun-2006.

The stock at hand is Apple (APPL) on the period Sep, 9 2003 - Jun, 4 2006 (Figure 5.2a). The daily returns are computed according to the formula $Y_t = S_t/S_{t-1} - 1$ and are given as input to the stochastic volatility models. We set N, the number of particles to 1000 and run the different samplers for $M = 10000$ Metropolis Hastings iterations. After discarding the first 1000 iterations, we collect the final sample and compute the posterior mean $\bar{\theta}$, the posterior median, 95% credibility intervals, the log likelihoods that results from the particle filter, the logarithm of the marginal likelihood, the AIC criterion and the M-H acceptance ratio. The model with the highest marginal likelihood is taken as reference and the Bayes factors are computed relatively to this model. Table 5.2 and 5.2b report estimation of θ for the stochastic volatility models $(\mathcal{M}_1, ..., \mathcal{M}_7)$. $\log(L)$ is the log marginal likelihood $\hat{p}_N(y|\mathcal{M})$. We find that the Gaussian TFSVL model $(\mathcal{M}_7)$ performs best in terms of marginal likelihood and AIC criteria. The Kass factor $2 \log BF$ of SVTFL $\mathcal{M}_7$ versus SVTF $\mathcal{M}_6$ is 2.8 which indicates a positive evidence in favor of the SVTF model and its leverage ρ. Compared to the SV with leverage $\mathcal{M}_3$ with one factor, the Kass factor in favor of SVL is 10.0 which is a strong evidence. The distribution of

the parameters are also fairly concentrated around their means. Overall, the values of ϕ are very close to one and confirm strong daily volatility persistence, in accordance to the volatility clustering fact seen in econometrics. The values of (ϕ_X, σ_X) and (ϕ_Z, σ_Z) are very interesting. ϕ_X is very close to 1 and σ_X is small whereas ϕ_Z is close to 0 and σ_Z is high. It seems clear now that the volatility of the returns can be decomposed into two distinct processes: a long-run stochastic trend $\mathbf{X}$ and a process $\mathbf{Z}$ accounting for short-run dynamics.

The same procedure was conducted on a spread, composed of three stocks: AMR CORP, CRANE CO and DOVER CORP with associated cointegrating vector $\beta = (1, -0.0865, -0.3796)$. The considered period is the same as the one used for the stock. Table 5.3 reports estimation of θ for the stochastic volatility models $(\mathcal{M}_1, ..., \mathcal{M}_7)$. We find again that the Gaussian TFSVL model $(\mathcal{M}_7)$ performs best in terms of the marginal likelihood and AIC criteria. This time, the Kass factor of SVTFL $\mathcal{M}_7$ versus SVTF $\mathcal{M}_6$ is 10.8 which indicates a very strong positive evidence in favor of the SVTF model and its leverage ρ. Figure 5.3 shows the estimation of the latent processes $\mathbf{X}$ and $\mathbf{Z}$ of model $\mathcal{M}_7$ for the spread.



(a) X_t (b) Z_t

Figure 5.3: Estimation of the latent processes X_t and Z_t of $\mathcal{M}_7$. Data is Spr AMR CORP - CRANE CO - DOVER CORP.

Parameter	$\bar{\theta}_{\mathcal{M}1}$	$\bar{\theta}_{\mathcal{M}2}$	$\bar{\theta}_{\mathcal{M}3}$	$\bar{\theta}_{\mathcal{M}4}$	$\bar{\theta}_{\mathcal{M}5}$	$\bar{\theta}_{\mathcal{M}6}$	$\bar{\theta}_{\mathcal{M}7}$
ϕ	0.9991	0.9989	0.9960	0.9981	0.9986		
σ	0.2395	0.1983	0.2728	0.1694	0.2533		
β	0.8783	0.3705	0.1	0.2359	0.1625	0.7427	0.6992
ν		7.6850					
ρ			-0.4397				-0.4178
ψ				0.0060			
ϕ_X						0.9995	0.9989
ϕ_Z						0.3181	0.2477
σ_X						0.1222	0.1906
σ_Z						0.6657	0.5219
$\log(L)$	2646.3	2659.7	2660.9	2649.2	2649.3	2664.5	2665.9
AIC	-5286.6	-5311.4	-5313.8	-5290.4	-5292.6	-5319.0	-5319.8
$2\log\mathcal{BF}(\cdot,\mathcal{M}7)$	33.9	12.4	10.0	33.4	33.2	2.8	0

Table 5.2: Estimation of the SV parameters. Data is APPL.

Parameter	$\bar{\theta}_{\mathcal{M}1}$	$\bar{\theta}_{\mathcal{M}2}$	$\bar{\theta}_{\mathcal{M}3}$	$\bar{\theta}_{\mathcal{M}4}$	$\bar{\theta}_{\mathcal{M}5}$	$\bar{\theta}_{\mathcal{M}6}$	$\bar{\theta}_{\mathcal{M}7}$
ϕ	0.9981	0.9993	0.9986	0.9981	0.9986		
σ	0.2238	0.1752	0.2188	0.1694	0.2533		
β	0.4419	0.5722	0.4559	0.2359	0.1625	0.3478	0.3690
ν		7.6850					
ρ			-0.3017				-0.8532
ψ				0.0852			
ϕ_X						0.9995	0.9996
ϕ_Z						0.1926	0.7554
σ_X						0.1268	0.0725
σ_Z						0.4913	0.3443
$\log(L)$	1792.3	1797.8	1795.1	1793.5	1788.5	1801.3	1806.7
AIC	-3578.6	-3587.6	-3582.2	-3579.0	-3571.0	-3592.6	-3601.4
$2\log\mathcal{BF}(\cdot,\mathcal{M}7)$	28.8	17.8	23.2	26.4	36.4	10.8	0

Table 5.3: Estimation of the SV parameters. Data is Spr AMR CORP - CRANE CO - DOVER CORP.

5.3 Estimation of Rolling Volatility of Spread Instruments

Once the best Stochastic Volatility model has been selected ($\mathcal{M}_7$ according to Section 5.2.2) and its parameters being estimated, the volatility of the spread can be approximated. The main idea behind using these stochastic volatility models is to catch the dynamics of the spread through a better estimation of its hidden volatility. According to Definition 3, a spread is a particular linear combination of assets where each asset price is one observation of a more general process, over a time interval. For a given first

order Markov n-process $\mathbf{X}_t$, the returns $y_{1:T}$ modelled by a SV model, are usually of the form $y_t|\mathbf{x}_t, \theta \sim \mathcal{D}\left(\mu_\theta(t), \sigma_\theta^2(t)\right)$, where $\mathcal{D}$ can represent any suitable distribution in a location-scale family (Definition 5). By definition, $Y_t = S_t/S_{t-1} - 1$. We then have

$$S_t|S_{t-1}, \mathbf{x}_t, \theta \sim \mathcal{D}\left(S_{t-1}\mu_\theta(t) + S_{t-1}, |S_{t-1}|^2\sigma_\theta^2(t)\right) \tag{5.3}$$

where the volatility $\sigma_\theta^2(t)$ and the mean $\mu_\theta(t)$ are known quantities because they depend on $\mathcal{F}_{t-1}$-measurable quantities.

In order to estimate the rolling volatility of the spread S_t, we generate many Monte Carlo paths according to Equation (5.3). Algorithm 5 explains the procedure when model $\mathcal{M}_7$ (TFSVL) is considered[1]. In the most general case, M paths $\{S_{t,n}\}_{0<n\leq M, t\in\mathbb{N}^*}$ are generated from Equation (5.3). Let $f_a : \mathbb{R}^{+M} \to \mathbb{R}^+$ be a positive-definite aggregating function. The aggregated rolling volatility of lag p for all the M paths is defined as $r\sigma_p(t) = f_a(r\sigma_p(t)_1, ..., r\sigma_p(t)_M)$ for $t > 0$. If f_a is simply the sample mean estimator, the equation is simplified to $r\sigma_p(t) = \frac{1}{M}\sum_{i=1}^{M} r\sigma_p(t)_i$. Depending on the context, f_a can be any measurable function satisfying the conditions above. Throughout the scope of this thesis, f_a will be equal to the sample mean. Without any more information, the sample mean is a pertinent measure to summarize a set of points in one single point.

Algorithm 5 Rolling volatility computation for model $\mathcal{M}_7$ (TFSVL)

procedure RollingVolatility($x_{1:T}, z_{1:T}, S_{1:T}$, $\theta = \beta$, M, $f_a = M^{-1}\sum_{i=1}^{M}\cdot$, p)
  for t from 1 to T **do**
    for i from 1 to M **do**
      Sample the t^{th} value of the i^{th} path, $S_{ti} \sim \mathcal{M}(S_{t-1}, S_{t-1}^2\beta^2\exp(x_t + z_t))$
    end
  end
  for i from 1 to M **do**
    Compute the default rolling volatility $(r\sigma_p(t)_i)_{t>0}$ for the i^{th} path, $(S_{ti})_{t>0}$
  end
  for t from 1 to T **do**
    $r\sigma_p(t) = M^{-1}\sum_{i=1}^{M} r\sigma_p(t)_i$
  end
return $r\sigma_p(t)$

[1]The volatility computed in this approach is of the same shape as the one computed in the default Bollinger bands (Equation 6.3).

6 Statistical Arbitrage Strategies

Statistical arbitrage conjectures statistical mis-pricings or price relationships that are true in expectation, in the long run when repeating a trading strategy. It describes a variety of automated trading systems which commonly make use of data mining, statistical methods and artificial intelligence techniques. A popular strategy is pairs trade, in which stocks are put into pairs by fundamental or market-based similarities. When one stock in a pair outperforms the other, the poorer performing stock is bought (long) with the expectation that it will climb towards its outperforming partner, and the other is sold (short). This approach has the advantage of eliminating the market exposure. The idea can be easily generalized to n stocks or assets where an asset can be a sector index. The investment strategy we aim at implementing is market neutral, thus we will hold a long and a short position both having approximately the same value in local currency. It is important to understand that the quantity to interest is the pseudo difference between the assets, better known as the spread. The approach abstracts the trading on the underlying assets and primarily focuses on spread trading. The common strategy is to evaluate if the spread instrument is either underpriced or overpriced. A typical way is to open a position once the spread deviates far from its long-run equilibrium, and unwind it when it reverts. Dealing with spreads instead of non-stationary assets is beneficial because stationary series are on average more reverting due to their nature.

In this section, two strategies are presented: Bollinger bands and Z-score. The first one models a time-varying mean for the spread and time-varying volatility bands (with simple or complex modelling) to gauge the spread deviation, whereas the second one assumes a fixed non-zero mean and fixed volatility bands.

6.1 Bollinger Bands

Bollinger bands is a widely used technical volatility indicator invented by John Bollinger in the 1970s which consists of using a moving average $m_\theta(t)$ of lag p with two volatility bands $B_\theta^+(t), B_\theta^-(t)$, above and below it. The computation of the volatility bands involves a windowed standard deviation of lag p (Definition 4). The shift between the bands and the stochastic mean is proportional to a parameter called α. The bands will expand and contract as the prices become volatile or bound into a tight trading pattern. When prices continually touch the upper bound $B_\theta^+(t)$, the spread is considered to be overbought. Conversely, when they continually touch the lower band, it is oversold. The indicator is

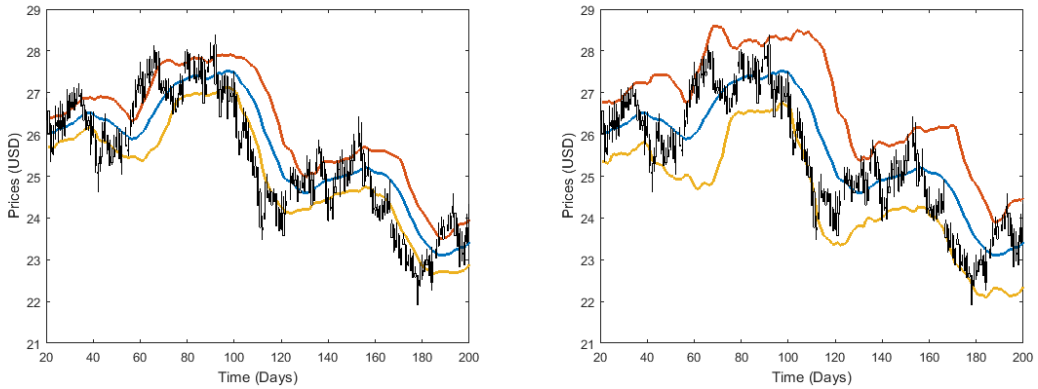
calculated by

$$m_{SMA}(t,p) = \frac{1}{p}\sum_{j=1}^{p} S_{t-j} \tag{6.1}$$

$$m_{EMA}(t,p) = k \times S_t + (1-k) \times m_{EMA}(t-1,p),\ k = 2/(p+1) \tag{6.2}$$

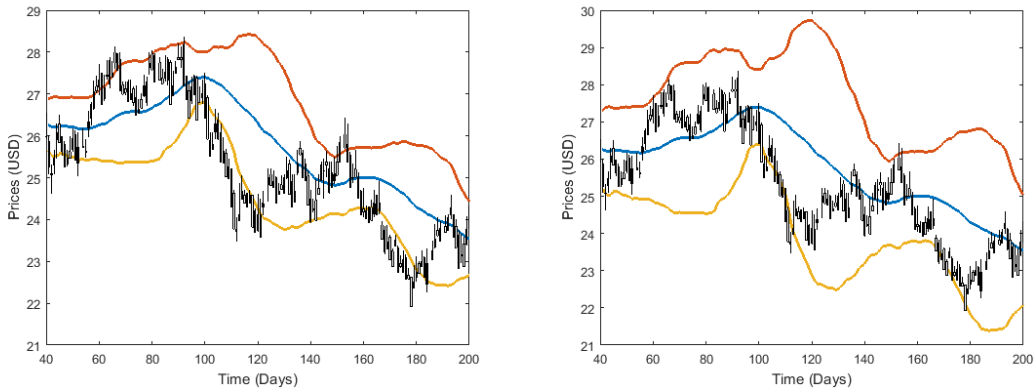
$$B^{\pm}(t,p,\alpha) = m(t,p) \pm \alpha \underbrace{\sqrt{\frac{1}{p}\sum_{j=1}^{p}(S_{t-j} - m(t,p))^2}}_{r\sigma_B(t,p)} \tag{6.3}$$

where S_t is the price of the spread, p is the lag and α is the number of standard deviations to shift the Bollinger bands. According to John Bollinger, the default values are $p = 20$ and $\alpha = 2$. $m_\theta(t)$ is the mid band used as a relative mean value. The exponential moving average (EMA) gives more weights to new values and may be faster to detect opportunities. $B_\theta^+(t)$ and $B_\theta^-(t)$ are respectively the upper and lower bands. Their intrinsic purpose is to measure how far the price deviates from its mean. Under the mild assumption that the returns are normally distributed and independent, approximately 95% of the prices should appear within the bands when $\alpha = 2$. Figures 6.1 and 6.2 show different configurations of the Bollinger bands applied to Walt Disney Co NYSE for the year 2002, all computed with Equations (6.1), (6.2) and (6.3).



(a) $(p, \alpha, t) = (20, 1.0, SMA)$ (b) $(p, \alpha, t) = (20, 2.0, SMA)$

Figure 6.1: Simple Bollinger bands strategy applied to Walt Disney Co NYSE for the year 2002. Lag is 20 days. B_θ^+ is red, B_θ^- yellow and m_θ Navy blue.



(a) $(p, \alpha, t) = (40, 1.5, SMA)$ (b) $(p, \alpha, t) = (40, 2.5, SMA)$

Figure 6.2: Simple Bollinger bands strategy applied to Walt Disney Co NYSE for the year 2002. Lag is 40 days. B_θ^+ is red, B_θ^- yellow and m_θ navy blue.

A trading rule determines the correct timing when to open and close a position. With the Bollinger bands strategy, the rule is

- LONG - Open a long position when there is an upward crossing between the spread and the lower band. Unwind this position when there is an upward crossing between the spread and the upper band;
- SHORT - Open a short position when there is a downward crossing between the spread and the upper band; Unwind this position when there is a downward crossing between the spread and the lower band.

6.2 Z-score

Z-score is a strategy based on mean-reverting patterns but unlike the Bollinger bands, Z-score assumes a non-zero constant mean and constant volatility bands. For this reason, Z-score is only suitable for stationary processes such as spread instruments. Z-score is a dimensionless indicator defined as $z_t = (S_t - \mu_S)/\sigma_S$, where μ_S and σ_S are respectively the unconditional mean and variance of the spread. z_t measures the distance to the long-term mean in units of long-term standard deviation. The basic rule is to open/close a position when the Z-score hits a predefined n-quantile of the standard normal distribution $\Phi^{-1}(q_n)$. If the Z-score hits a low threshold, it means that the spread is underpriced and a long position should be opened. When the spread reverts to its mean, the position is unwound. A same reasoning is done for short positions. Caldeira and Moura (2013) suggested the basic trading strategy signals: Open long position if $z_t \leq \Phi^{-1}(q_{OL}) = -2.00$, open short position if $z_t \geq \Phi^{-1}(q_{OS}) = 2.00$, close short position if $z_t \leq \Phi^{-1}(q_{CS}) = 0.75$ and close long position if $z_t \geq \Phi^{-1}(q_{CL}) = -0.50$. Figure 6.3 shows the spread S_t, the Z-score z_t and the thresholds $\Phi^{-1}(q)$ of the strategy.

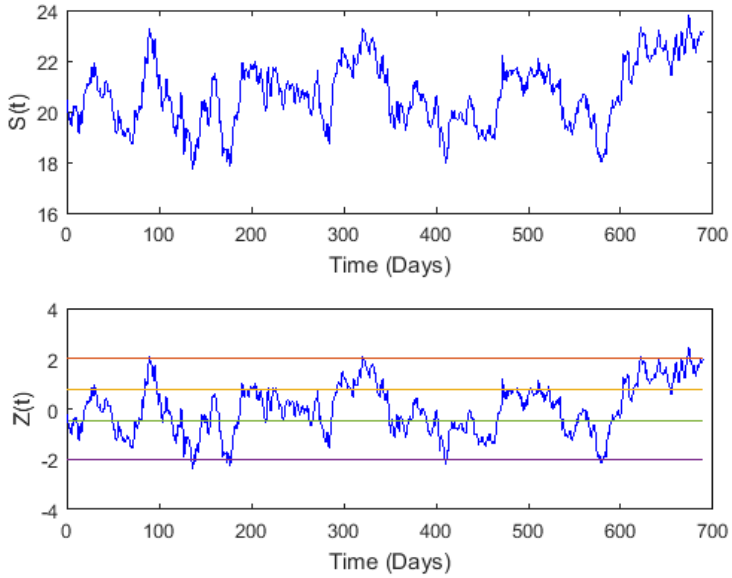


Figure 6.3: Spread S_t (defined in Section 5.2.2) and its Z-score z_t. From top to bottom: $\Phi^{-1}(q_{OS}), \Phi^{-1}(q_{CS}), \Phi^{-1}(q_{CL}), \Phi^{-1}(q_{OL})$.

Unlike the Bollinger bands, the Z-score is highly sensitive to stochastic trends because the mean is assumed to be strictly constant. In other words, it could be dangerous if the spread lost its cointegrated property and became divergent. In practical applications and according to the risk policy, a stop loss threshold is usually set to avoid any huge losses.

6.3 Selection of the Cointegrated Tuples

The assumption that the pairs should belong to the same sector is very common in the literature. Some examples include Chan (2009) and Dunis et al. (2010). Other such as Caldeira and Moura (2013), did not adopt this restriction but bound their study to pairs trading applied to the Brazilian markets.

6.3.1 Complexity Reduction with Correlation

In the general case and according to Vidyamurthy (2004), cointegration implies correlation but the reverse is not true. Spurious regression is a very good example where the reverse is not true. The idea is to filter the uncorrelated tuples to limit the number of candidates for cointegration testing. This assertion holds because a cointegration test is more time-consuming than a correlation test (Table 6.1).

Test	Min	Max	Avg	Std	95% Conf. Int.
Correlation *corr*	0.26	1.00	0.33	0.10	[0.26 , 0.70]
Correlation R^2 (fast)	0.47	1.45	0.58	0.18	[0.48 , 1.19]
Johansen	16.77	37.96	19.23	3.74	[16.93 , 34.04]
Aug. Dickey Fuller	1.68	3.42	1.96	0.33	[1.68 , 3.31]
Phillips-Perron	2.62	5.93	2.99	0.67	[2.63 , 5.66]

Table 6.1: Average time spent to test a bivariate time series $\mathbf{X}_t$ (in milliseconds)

When it comes to pairs trading, a simple correlation test is enough. When $n \geq 3$, it is preferred to use the multiple correlation coefficient, better known as R^2. It can be computed using the vector $c = (r_{\mathbf{x}_1\mathbf{y}}, r_{\mathbf{x}_2\mathbf{y}}, ..., r_{\mathbf{x}_N\mathbf{y}})^T$ of correlation $r_{\mathbf{x}_n\mathbf{y}}$ between the predictor variables $(\mathbf{x}_n)_{1\leq n\leq N}$ and the target variable $\mathbf{y}$. The correlation matrix $R_{\mathbf{xx}}$ of inter-correlations (Equation 6.4) between predictor variables also comes into play. It is given by $R^2 = c^T R_{\mathbf{xx}}^{-1} c$.

$$R_{\mathbf{xx}} = \begin{pmatrix} r_{\mathbf{x}_1\mathbf{x}_1} & r_{\mathbf{x}_1\mathbf{x}_2} & \cdots & r_{\mathbf{x}_1\mathbf{x}n} \\ r_{\mathbf{x}_2\mathbf{x}_1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ r_{\mathbf{x}_n\mathbf{x}_1} & \cdots & & r_{\mathbf{x}_n\mathbf{x}_n} \end{pmatrix} \tag{6.4}$$

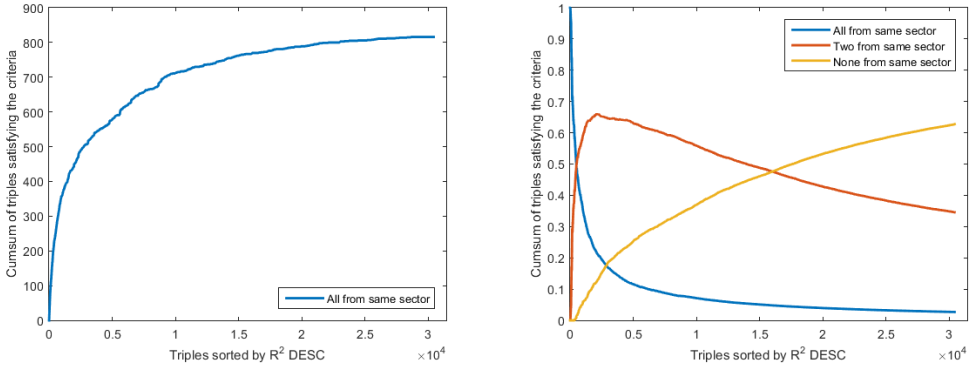
It is worth noting that R^2 is order-dependent. To provide convincing evidence of this fact, let us consider a simple example. A regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}$ will in general have a different R^2 that will a regression of $\mathbf{z}$ on $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{z}$ be uncorrelated with both $\mathbf{x}$ and $\mathbf{y}$ while $\mathbf{x}$ and $\mathbf{y}$ are linearly related to each other. A regression of $\mathbf{z}$ on $\mathbf{y}$ and $\mathbf{x}$ will yield a R^2 of zero, while a regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}$ will yield a strictly positive R^2. It means that the ordering inside a tuple has its importance at least from a statistical point of view, as highlighted in Definition 2. This assertion is also true for most cointegrations tests. This notion of ordering is however less obvious from a pure financial point of view.

An example is presented with $n = 4$ in Appendix 10.3. In this case, the assumption of the same sector is almost inevitable. It becomes interesting to question it for triples trading, which is the purpose of the next section.

6.3.2 Sector Analysis for Triple Trading

Chan (2009) and Dunis et al. (2010) argued that the pairs should belong to the same sector, otherwise the cointegration and the correlation would be purely fortuitous. To check the veracity of this assumption for triple trading, all the possible cointegrated triples ($n = 3$) are formed on the whole period of the dataset - from January, 01 1990 to March, 14 2014 - and the R^2 is computed using the methodology exposed in Section 6.3.1. The cointegrated triples are then sorted according to their R^2 from the highest to the lowest value. Each triple is characterized by the sector criteria: *All*, *Partial* or *None*. *All* means the three assets composing the triple belong to the same sector, *Partial* that

exactly two belong to the same sector, *None* that no one belongs to the same sector. As a result, 30418 cointegrated triples were formed. 816 belonged to *All*, 10517 to *Partial* and the remaining 19085 to *None*. Figure 6.4b shows that for very high R^2 on daily returns, almost all the cointegrated triples belong to the same sector. Then for high R^2, the proportion of partial triples becomes higher than two other groups until the half of the set. The conclusion is that when the number of selected cointegrating triples is not very large (less than 500 or 1.5% here for the whole period), it might be reasonable to consider the assumption of the same sector for increased execution speed. However, the same sector approach will only be retained for quadruples trading and not for triple trading because the computational cost is still affordable.



(a) Cumulative sum of the cointegrated triples from the same sector sorted by R^2 from highest to lowest.

(b) Distribution of the cointegrated triples sorted by R^2 from highest to lowest and regarding their belonging to sectors.

Figure 6.4: Distributions of the cointegrated triples. Period is Jan, 01 1990 - Mar, 14 2014.

7 Procedure

The workflow involves the preparation of the datasets and the selection of the cointegrated tuples. From this point, the three strategies are applied and backtested on the data. Finally, the cumulative returns and risk measures are computed to assess the performance of the strategies.

7.1 General Framework

7.1.1 Training and Simulation Sets

We use the dataset $\mathcal{EQTY}_{daily}$ defined in Section 1.2.1. The whole sample period is divided into sets of length **two** years. Each set is split into two sets: the in-sample set denoted $\mathcal{I}$ and the out-of-sample $\mathcal{O}$ with a 1:2 ratio: $(\mathcal{I}_i, \mathcal{O}_i)_{1 \leq i \leq 12}$. The choice of the ratio is motivated by the fact that the Bollinger bands has a stochastic mean and are therefore less sensitive to a break of the cointegration property, contrary to the Z-score. Detection of cointegrated tuples, selection of the best tuples and tuning of the Bollinger bands parameters (p, t, α) is done on $\mathcal{I}$. The purpose of $\mathcal{O}$ is to assess the performance of the strategy on unseen data with the parameters computed in the training period. This technique is known as cross validation and is used to avoid overfitting during the calibration.

7.1.2 Framework Hypotheses

Only two types of transactions are considered: move into a new position and unwind a previously opened position. The operations on the spread are assumed to be atomic, i.e. they have a succeed-or-fail property. No stop loss protections are considered. At the end of each trading period, all the outstanding positions are closed. We consider 5 point basis of transaction costs. This choice was made for pairs trading in Dunis et al. (2010), Dunis and Ho (2005) and Alexander and Dimitriu (2002). For simplicity, no rental costs are considered for short positions but the capital invested in short selling cannot exceed 50% of the total capital, either invested or in cash. The strategies are self-financing, i.e. profits are reinvested and no deposits or withdrawals are permitted. When a long position is initiated, the first asset is bought with quantity 1 and the remaining assets of the tuple are sold with the respective quantities indicated by the cointegrating vector β. This same position is closed by selling one unit of the first asset and buying the remaining assets, still in the same proportions. It is assumed that the trader can buy a portion of an asset.

7.1.3 Portfolio Approach

The first approach consists in ranking the cointegrated tuples based on the best in-sample Sharpe Ratios $\mathcal{SR}$ for every period training-simulation. For each period, the 20 highest-ranked tuples are used to compose the portfolio. The first motivation of considering a portfolio is to lower the volatility associated to each tuple trading by smoothing the net value over time. The asset allocation in the portfolio follows an investing weighting scheme with no dynamic rebalancing. At each time, the portfolio valuation is a linear combination of each valuation, where the weights are constant and equally distributed. For any spread, the number of opened positions is limited to two (one long and one short at the same time). An example is provided with a portfolio composed of two assets: X_1 and X_2. Denote their respective cumulative returns by $CR_{X_1}(t)$ and $CR_{X_2}(t)$. At each time, the valuation of the portfolio is simply given by $p(t) = 0.5 \times CR_{X_1}(t) + 0.5 \times CR_{X_2}(t)$.

7.2 Optimization of the Strategies

The optimization of the Bollinger bands strategy requires to tune three parameters: the number of periods p to compute the bands, the type t of moving average (EMA or SMA) used in the mid band and α which controls the interval between the volatility bands. John Bollinger suggests $p = 20, \alpha = 2$ and exponential moving average as default values. To get the best out of the strategy, a cross validation is performed on the in-sample set $\mathcal{I}$. The criterion of optimization is the in-sample Sharpe Ratio $\mathcal{SR}$. The cross validation parameter space is denoted $\Omega_{CV} = \mathcal{P}_n \times \mathcal{P}_t \times \mathcal{P}_\alpha$. An exhaustive search was made possible because of the efforts invested into the parallelization of the task. This search provides good visual results and is therefore our recommended choice to reveal the topology of $f(\mathcal{P}_n \times \mathcal{P}_t \times \mathcal{P}_\alpha) \to \mathbb{R}$. The dataset is split into two sets $\mathcal{I}$ and $\mathcal{O}$ with a ratio 2:1. The parameters of the Bollinger bands (p, α, t) are tuned exhaustively, as explained in Section 7.2. The parameter space is defined as $\Omega_{CV} = [5, 6, ..., 60] \times \{\text{SMA}, \text{EMA}\} \times [1, 1.1, ..., 2.9, 3.0]$ and the topology function $f(\Omega_{CV}) \to \mathbb{R}$ is the Sharpe Ratio (Definition 7.3). The idea is to find regions, not peaks where f has a stable global maximum. As f is multivariate, the idea is to detect a region where $\nabla F(\rho_{max}, \sigma_{max}, t_{max}) = 0$. The gradient ∇F is defined by

$$\nabla F = \frac{\partial F}{\partial \rho}\vec{i} + \frac{\partial F}{\partial \alpha}\vec{j} + \frac{\partial F}{\partial t}\vec{k} \tag{7.1}$$

on a 3D space $(\vec{i}, \vec{j}, \vec{k})$. In a metric space, $M = (\Omega, d)$, a Borel σ-algebra $\mathcal{V} = \mathcal{B}(\Omega)$ is a region of a point p if there exists an open ball with center p and radius $r > 0$, such that $B_r(p) = \{\omega \in \Omega | d(\omega, p) < r\}$ is contained in $\mathcal{V}$. To find a suitable global maximum, an exhaustive search is performed for every $\omega \in \Omega$. From this point, the Sharpe ratios are sorted in the descending order $s_1 > s_2 > ... > s_{card(\Omega)}$. The highest Sharpe ratio s_1 is considered and its region evaluated with a suitable r. The gradient is then evaluated and checked if it is close to 0. If this is the case, $(p_1, \alpha_1, t_1, s_1)$ is selected. If not, s_1

is dropped and s_2 is now considered. The procedure goes on until a Sharpe ratio is selected.

7.3 Performance Assessment

Once the strategy was optimized on $\mathcal{I}$, it can be assessed on the out-sample test $\mathcal{O}$. The performance of the portfolios are examined in terms of cumulative return (CR), Sharpe Ratio (SR) and Maximum Drawdown (MDD). Let y_t denote the simple daily return on a particular asset at time t.

7.3.1 Maximum Drawdown

The maximum drawdown (MDD) is defined as the maximum percentage drop incurred from a peak to a bottom up to time T. It is the worst possible scenario up to time T. By construction, it is higher in absolute values than the maximum loss. Indeed, this is the most pessimistic scenario. Figure 7.1 provides a graphical interpretation of the Maximum Drawdown.

$$MDD(T) = \max_{s\in(0,T)} [\max_{s\in(0,T)} y_t - y_s] \tag{7.2}$$

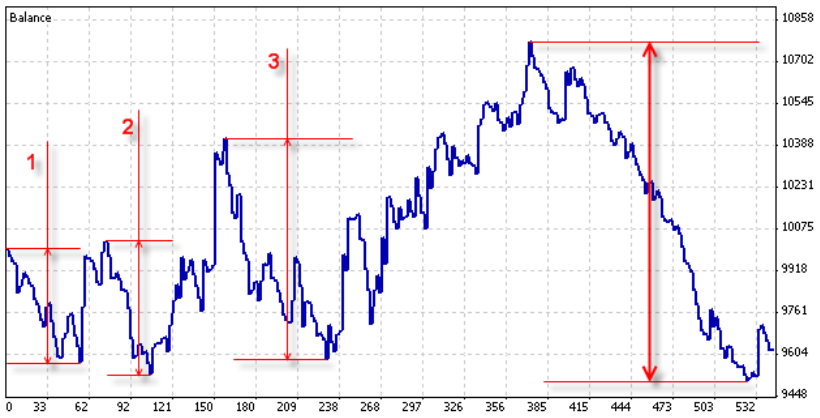


Figure 7.1: Evolution of the Maximum Drawdown for a synthetic portfolio

7.3.2 Sharpe Ratio

The Sharpe Ratio ($\mathcal{SR}$) based on daily returns is defined as

$$\mathcal{SR} = \frac{\hat{\mu}}{\hat{\sigma}} \text{ where, } \hat{\mu} = 252 \times \frac{1}{T}\sum_{t=1}^{T} y_t \text{ and } \hat{\sigma} = \sqrt{252} \times \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{\mu})^2 \tag{7.3}$$

A variant is to subtract the daily risk free return y_{rf}, given by the US Treasury bills, to the returns of the strategy. This version is expected to output a slightly lower value. $\hat{\mu}$

represents the mean asset return and $\hat{\sigma}$ the standard deviation of the returns. The idea of the Sharpe Ratio is to quantify how much additional return the investor is receiving for the additional volatility of holding the risky asset over a risk-free asset. **A general rule of thumb is: a ratio of 1 or better is considered good, 2 and better is very good, and 3 and better is considered excellent.**

7.3.3 Buy and Hold Strategy

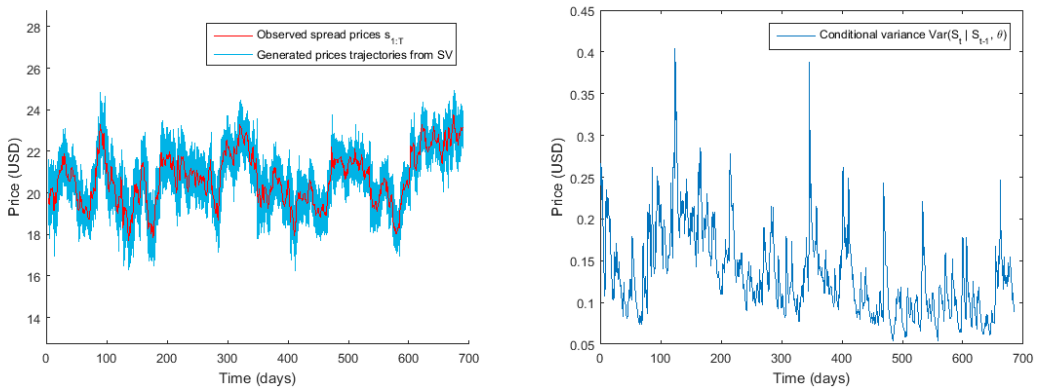
Finally, one technique to assess the performance of a strategy is to compare it to the standard Buy and Hold (B&H) strategy where the holder has a portfolio replicating the stock index from time 0 to T.

8 Empirical Analysis

This chapter is structured in three main parts. Firstly, a case study is examined to show some interesting facts about volatility modelling. Next, the cross validation of the Bollinger bands parameters is presented in further details. Finally, the Bollinger bands with and without the complex volatility estimation are compared. Also, the Z-score and the Buy and Hold strategies are considered as a benchmark.

8.1 Volatility Modelling of Spread Instruments with $\mathcal{M}_7$

Throughout this section, the same spread S_t as that of Section 5.2.2 is considered. This procedure is not restricted to this particular spread and is valid for any stationary spread and any stochastic volatility models. This section focused on the main points described in Section 5.3. The model $\mathcal{M}_7$ is used to estimate the conditional variance $S_t|S_{t-1}, \mathbf{x}_t, \theta$. From there and from Equation (5.3), $M = 1000$ Monte Carlo trajectories are generated. Figure 8.1 shows the generation of the trajectories according to the estimated conditional variance.



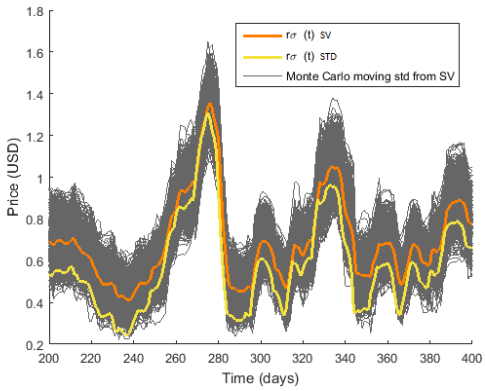
(a) Generation of $M = 1000$ MC prices trajectories with model $\mathcal{M}_7$

(b) Conditional variance of $S_t|S_{t-1}, \mathbf{x}_t, \theta$ with model $\mathcal{M}_7$

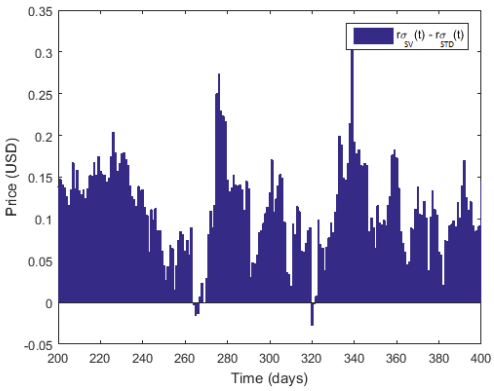
Figure 8.1: Generation of the trajectories of the spread process S_t

The next step is to compute the rolling volatility for every trajectory. The aggregated function f_a is the sample mean and $r\sigma_{SV}(t)$ is the resulting aggregation of the MC trajectories. The standard rolling volatility on the observed prices $s_{1:T}$ is denoted $r\sigma_{STD}(t)$. From this point, 90% and 95% confidence intervals are derived for every $t \in [1, T]$ using

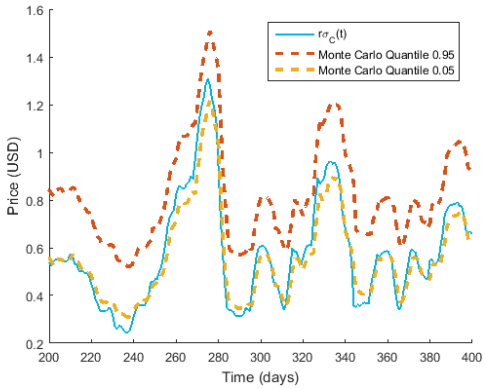
Monte Carlo. The results are presented in Figure 8.2[1]. It turns out that $r\sigma_{STD}(t)$ is clearly underestimated most of the time. Only 63.9% of $r\sigma_{STD}(t)$ is contained inside the 0.90 confidence intervals and 78.3% inside the 0.95 confidence intervals. Moreover, Figure 8.2b summarizes the difference $\delta(t) = r\sigma_{SV}(t) - r\sigma_{STD}(t)$. With $E[\delta] = 0.1035$ (USD), the standard volatility estimator is clearly biased.



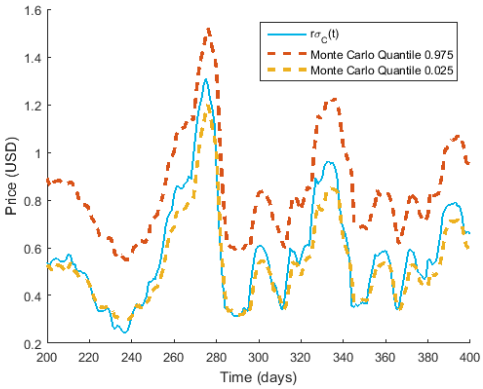
(a) Generated prices trajectories with rolling volatility processes



(b) $\delta(t) = r\sigma_{SV}(t) - r\sigma_{STD}(t)$



(c) $r\sigma_{STD}(t)$ and estimated 90% confidence intervals (MC)



(d) $r\sigma_{STD}(t)$ and estimated 95% confidence intervals (MC)

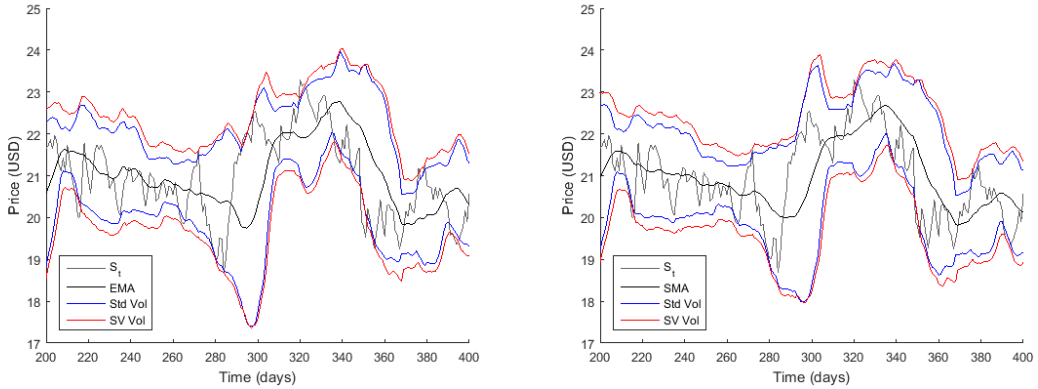
Figure 8.2: Confidence intervals of the rolling volatilities of S_t

When the rolling volatilities have been estimated, the Bollinger bands can finally be computed. Equation (6.3) and (8.1) are used to form the bands respectively for the standard and stochastic volatility estimators $r\sigma_{STD}(t)$ and $r\sigma_{SV}(t)$.

$$B^{\pm}(t, p, \alpha) = m(t, p) \pm \alpha \cdot r\sigma_{SV}(t, p) \tag{8.1}$$

[1]It is worth noting that the time axis has been truncated to improve readability. The analysis and the conclusions are carried and made on the whole period $[1, T]$.

where the notations of Section 6.1 apply. Figure 8.3 shows the results for $p = 20$ and $\alpha = 2$. The mid band was computed using a simple moving average (SMA) and an exponential moving average (EMA). Not surprisingly, the bands computed with the standard methods are narrower than the ones estimated from stochastic volatility models. It becomes now interesting to see if this more complex modelization can help build more accurate trading signals.



(a) $(p, \alpha, t) = (20, 2.0, \text{EMA})$ (b) $(p, \alpha, t) = (20, 2.0, \text{SMA})$

Figure 8.3: Bollinger bands computed with $r\sigma_{STD}(t)$ (blue) and $r\sigma_{SV}(t)$ (red)

8.2 Stochastic Volatility Modelling for Triple Trading

This section presents the results of the comparisons between Bollinger bands strategies with simple and complex Stochastic Volatility modelling. The aim is to assess whether using more advanced modelling will result to better performance than simple models used often in practice.

8.2.1 Selection and Calibration Step

For each training set $(\mathcal{I}_i)_{1 \leq i \leq 12}$, the cointegrated triples are formed and the models calibrated. The same sector assumption is not considered but the sets of cointegrated triples are truncated to ensure a reasonable size. Table 8.1 summarizes the selected triples for each period and to which sector they belong. The Bollinger bands strategy with simple modelling does not require a calibration phase. The function `bollinger` from the MATLAB Financial toolbox is used to compute the three bands[2]. Complex modelling requires the estimation of each parameter of model $\mathcal{M}_7$ with *Particle Markov Chain Monte Carlo*. As the number of spreads is big (46484 according to Table 8.1), a slightly different version of Algorithm 4 is used (Appendix 10.6).

[2]For more information, the reader can consult the MATLAB documentation: `http://mathworks.com/help/finance/bollinger.html`

		Sector Belonging Criterion		
Period	Coint triples	All	Partial	None
Jan, 1 1990 - Jan, 1 1992	3503	622 (0.177)	1637 (0.467)	1244 (0.355)
Jan, 1 1992 - Jan, 1 1994	2147	323 (0.150)	1187 (0.552)	637 (0.297)
Jan, 1 1994 - Jan, 1 1996	3928	449 (0.114)	1748 (0.445)	1731 (0.440)
Jan, 1 1996 - Jan, 1 1998	4640	471 (0.101)	2207 (0.475)	1962 (0.422)
Jan, 1 1998 - Jan, 1 2000	3254	285 (0.083)	1574 (0.483)	1395 (0.429)
Jan, 1 2000 - Jan, 1 2002	2643	228 (0.085)	1355 (0.512)	1060 (0.401)
Jan, 1 2002 - Jan, 1 2004	5718	338 (0.059)	2340 (0.409)	3040 (0.531)
Jan, 1 2004 - Jan, 1 2006	3539	186 (0.052)	1508 (0.426)	1845 (0.521)
Jan, 1 2006 - Jan, 1 2008	3541	159 (0.044)	1350 (0.381)	2032 (0.573)
Jan, 1 2008 - Jan, 1 2010	5967	190 (0.031)	2205 (0.369)	3572 (0.599)
Jan, 1 2010 - Jan, 1 2012	3521	167 (0.047)	1147 (0.410)	1907 (0.541)
Jan, 1 2012 - Jan, 1 2014	4083	122 (0.029)	1592 (0.389)	2371 (0.580)
Total	46484	3540 (0.083)	19850 (0.427)	22796 (0.490)

Table 8.1: Cointegrated triples selected on the in-sample sets. Percentages are shown in brackets.

8.2.2 Results with Default Bollinger Bands Parameters

In this section, the tuning of the parameters is discarded and the default parameters $\theta(p = 20, \alpha = 2, t = \text{EMA})$, as recommended by John Bollinger are considered. The cross validation phase could add a bias to the results because we aim to compare the two Bollinger bands strategies for a fixed θ. The default parameters are widely used in practice among traders and it is therefore pertinent to base the comparison for those values.

Portfolio Approach

For each period - indexed by i - and for each strategy, the 20 best in-sample $\mathcal{SR}_{\mathcal{I}_i}$ triples are selected to compose the portfolio of the strategy. The assessment is done on the corresponding out-sample set $\mathcal{O}_i$. The procedure is explained for the period Jan, 01 2008 - Dec, 31 2009. Of the possible triples candidates, 5967 passed the cointegration tests described in Algorithm 1 (Table 8.1). Tables 8.2, 8.3 and 8.4 show the 20 tradable triples used in each strategy. $\beta = (\beta_1, \beta_2, \beta_3)$ is the cointegrating vector associated to the triple. $\mathcal{SR}_{\mathcal{O}}$ corresponds to the Sharpe ratio on the out-sample set. The cumulative returns (CR) with costs and the number of trades are recorded on the out-sample step. A closer look shows that all the triples exhibit a very high Sharpe ratio on the in-sample period, as expected, and a lower but still promising Sharpe ratio on the out-sample period. Table 8.5 summarizes some statistics for the three portfolios. The results are in favor of the Z-score and the Bollinger bands with complex modelling. It is worth noting that the Z-score strategy had two triples 7 and 19 which clearly underperformed the others. Indeed, both triples lost their cointegration property.

Stock 1	Stock 2	Stock 3	β_1	β_2	β_3	$\mathcal{SR}_{\mathcal{I}}$	$\mathcal{SR}_{\mathcal{O}}$	Net Return	Annualized Return	Max Drawdown	Number of Trades
SYY	PNC	PRU	1.00	-0.04	-0.13	3.90	1.07	27394.60	130.45 %	0.29	10 × 2
SEE	SHLD	UNM	1.00	-0.21	-0.28	3.75	1.80	16396.51	47.97 %	0.09	12 × 2
HIG	BCR	BRK/B	1.00	-0.65	-1.58	3.55	0.23	12665.36	19.99 %	0.62	8 × 2
ETFC	DDR	DE	1.00	-0.31	-0.47	3.51	0.87	13869.04	29.01 %	0.29	10 × 2
FCX	DIS	DO	1.00	-0.43	-0.84	3.48	0.46	19763.68	73.22 %	0.74	8 × 2
MA	LSI	OXY	1.00	-0.22	-0.31	3.47	0.88	24489.13	108.67 %	0.31	11 × 2
AMT	AMP	CCI	1.00	-0.16	-0.49	3.44	1.14	15091.08	38.18 %	0.14	11 × 2
TIE	0772031D	CVH	1.00	-0.47	-0.24	3.44	0.34	10992.36	7.44 %	0.19	10 × 2
SEE	SCI	SHLD	1.00	-0.34	-0.27	3.38	1.19	14839.25	36.29 %	0.15	10 × 2
LUK	MAC	MHK	1.00	-0.31	-0.43	3.37	0.20	11666.71	12.50 %	0.49	8 × 2
NSC	PG	R	1.00	-0.50	-0.37	3.35	1.12	32770.77	170.78 %	0.48	12 × 2
FII	FCX	MSI	1.00	-0.22	-0.34	3.35	1.21	20997.83	82.48 %	0.25	12 × 2
MKC	JNJ	KSU	1.00	-0.57	-0.06	3.28	0.00	9989.55	-0.10 %	0.22	10 × 2
SEE	SHLD	TSN	1.00	-0.28	-0.29	3.26	0.56	11882.95	14.12 %	0.25	8 × 2
MA	LSI	PXD	1.00	-0.23	-0.24	3.24	0.34	12827.76	21.20 %	0.35	10 × 2
JNS	JDSU	MDT	1.00	-0.51	-0.76	3.24	0.26	11255.74	9.42 %	0.33	13 × 2
SYY	PRU	PTC	1.00	-0.09	-0.25	3.21	0.83	22991.15	97.43 %	0.41	10 × 2
FOSL	CMS	COL	1.00	-0.27	-0.73	3.21	0.37	11676.18	12.57 %	0.24	7 × 2
MTW	MTG	SUNE	1.00	-0.08	-0.59	3.21	-0.12	9551.03	-4.48 %	0.27	10 × 2
HOG	HNZ	MDT	1.00	-0.97	-0.42	3.21	0.15	11108.11	8.31 %	0.42	9 × 2

Table 8.2: The 20 triples composing the portfolio of the simple Bollinger bands (2008-2009)

Stock 1	Stock 2	Stock 3	β_1	β_2	β_3	$\mathcal{SR}_{\mathcal{I}}$	$\mathcal{SR}_{\mathcal{O}}$	Net Return	Annualized Return	Max Drawdown	Number of Trades
SYY	PNC	PRU	1.00	-0.04	-0.13	4.08	0.58	18069.64	60.52 %	0.33	6 × 2
TIE	0772031D	CVH	1.00	-0.47	-0.24	4.05	0.70	12159.16	16.19 %	0.17	6 × 2
MAT	KLAC	KO	1.00	-0.30	-0.47	3.87	0.51	19505.18	71.28 %	0.41	6 × 2
MTW	MTG	SUNE	1.00	-0.08	-0.59	3.66	0.28	10968.99	7.26 %	0.19	4 × 2
ETFC	ETN	GCI	1.00	-0.95	-0.23	3.49	0.59	10912.62	6.84 %	0.10	2 × 2
MKC	COST	COV	1.00	-0.24	-0.20	3.47	0.06	10240.28	1.80 %	0.16	4 × 2
TWX	TIF	TNB	1.00	-0.23	-0.39	3.41	0.35	13045.36	22.84 %	0.38	5 × 2
WY	MET	MHFI	1.00	-0.19	-0.44	3.41	1.04	23042.24	97.81 %	0.30	8 × 2
CCK	CCI	GRA	1.00	-0.25	-0.18	3.38	0.60	12199.33	16.49 %	0.16	6 × 2
MAT	IP	IR	1.00	-0.19	-0.26	3.38	0.84	16451.04	48.38 %	0.33	3 × 2
BF/B	BJS	BNI	1.00	-0.18	-0.17	3.35	0.32	19287.20	69.65 %	0.53	9 × 2
SEE	SHLD	WYN	1.00	-0.20	-0.22	3.35	0.85	11306.68	9.80 %	0.08	4 × 2
INTC	CSX	CVG	1.00	-0.41	-0.19	3.21	0.15	12892.33	21.69 %	0.91	5 × 2
CCK	CHRS	CSCO	1.00	-0.03	-0.60	3.14	0.56	11523.00	11.42 %	0.14	5 × 2
ROP	MTW	MUR	1.00	-0.19	-0.34	3.04	1.46	46642.74	274.82 %	0.79	5 × 2
HPQ	HSP	PPG	1.00	-0.22	-0.48	3.04	0.97	16047.82	45.35 %	0.21	3 × 2
ETFC	DDR	DE	1.00	-0.31	-0.47	2.99	1.17	15966.98	44.75 %	0.29	6 × 2
MSFT	EL	EMN	1.00	-0.22	-0.36	2.98	0.63	14874.22	36.55 %	0.20	5 × 2
AA	0848680D	AKS	1.00	-0.09	-0.53	2.96	0.72	13474.59	26.05 %	0.23	6 × 2
THC	PVH	SIAL	1.00	-0.41	-0.42	2.95	0.93	12894.73	21.71 %	0.14	8 × 2

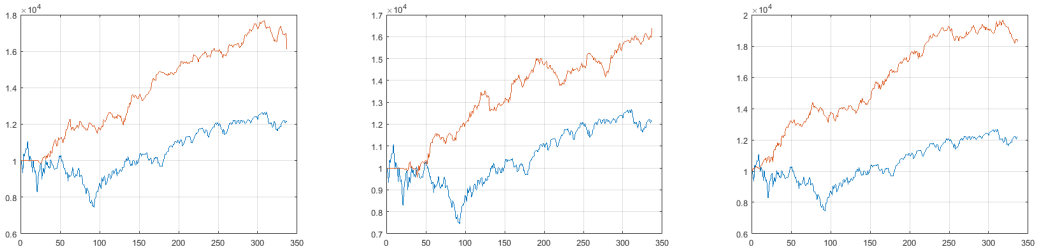
Table 8.3: The 20 triples composing the portfolio of the complex Bollinger bands (2008-2009)

Stock 1	Stock 2	Stock 3	β_1	β_2	β_3	$\mathcal{SR}_{\mathcal{I}}$	$\mathcal{SR}_{\mathcal{O}}$	Net Return	Annualized Return	Max Drawdown	Number of Trades
LNC	LLTC	MHK	1.00	-0.71	-0.79	4.53	1.73	47928.66	284.46%	0.27	8 × 2
PNR	HLS	HNZ	1.00	-0.28	-0.66	4.15	0.66	14050.95	30.38%	0.20	4 × 2
NTAP	DTV	NEU	1.00	-0.59	-0.20	3.99	1.49	16051.81	45.38%	0.14	6 × 2
SLB	CEG	CF	1.00	-0.24	-0.44	3.97	1.08	30520.06	153.90%	0.98	3 × 2
L	KSU	MDLZ	1.00	-0.44	-0.67	3.95	0.32	12355.60	17.66%	0.27	2 × 2
MHFI	MET	NEE	1.00	-0.25	-0.56	3.94	0.64	18726.76	65.45%	0.35	5 × 2
EMR	EIX	FCX	1.00	-0.52	-0.26	3.91	-0.30	2537.76	-74.62%	0.92	7 × 2
HSH	HRS	OKE	1.00	-0.16	-0.41	3.91	0.74	15225.56	39.19%	0.23	4 × 2
ETFC	AYE	BCR	1.00	-0.80	-0.51	3.90	1.11	18337.12	62.52%	0.22	4 × 2
L	KIM	R	1.00	-0.31	-0.24	3.88	0.09	11723.15	12.92%	0.77	2 × 2
WY	RTN	SLM	1.00	-0.60	-0.18	3.87	1.36	23280.39	99.60%	0.31	4 × 2
JNS	JDSU	MDT	1.00	-0.51	-0.76	3.85	1.62	16043.91	45.32%	0.16	6 × 2
FII	FISV	MSI	1.00	-0.67	-0.27	3.78	0.98	21794.70	88.46%	0.35	4 × 2
KO	KMX	VIAB	1.00	-0.05	-0.26	3.75	0.63	14000.93	30.00%	0.18	5 × 2
SUNE	AIG	AIZ	1.00	-0.06	-0.53	3.73	0.94	14008.95	30.06%	0.16	3 × 2
HPQ	HSP	ITW	1.00	-0.15	-0.60	3.71	0.05	11105.64	8.29%	0.57	4 × 2
NI	NU	OMC	1.00	-0.56	-0.34	3.66	1.14	14343.82	32.57%	0.16	5 × 2
ETFC	CCL	CEG	1.00	-0.90	-0.24	3.63	0.84	14664.55	34.98%	0.19	3 × 2
SLM	RX	SWY	1.00	-0.77	-0.69	3.63	-0.12	8622.02	-13.77%	0.87	7 × 2
FCX	DIS	DO	1.00	-0.43	-0.84	3.56	1.11	40656.10	229.92%	0.53	5 × 2

Table 8.4: The 20 triples composing the portfolio of the Z-score strategy (2008-2009)

Statistics	Bollinger Simple	Bollinger Complex	Z-score Strategy	SPX Buy&Hold
# of observations in the sample	505	505	505	505
# of observations in $\mathcal{I}$	169	169	169	169
# of days in the trading period ($\mathcal{O}$)	336	336	336	336
# of triples in the trading period	20	20	20	N.A.
# of cointegrated spreads to analyze	5967	5967	5976	N.A.
Average annualized return	45.83%	47.93%	62.05%	12.92%
Annualized Sharpe Ratio	2.176	2.227	2.341	0.412
Annualized Sharpe Ratio (w. risk free 2%)	2.110	2.293	2.394	0.362
Largest daily return	3.59%	4.49%	6.44%	10.78%
Lowest daily return	-5.04%	-4.45%	-3.47%	-8.92%
Cumulative profit on the trading period	61.11%	63.91%	82.74%	17.26%
Correlation with the market returns	-0.0246	0.0348	0.0804	1
Skewness of the returns	-0.0702	0.3212	0.8913	0.1331
Kurtosis of the returns	7.3164	5.5508	7.4581	6.4213
Maximum Drawdown	0.0888	0.0839	0.0894	0.3273

Table 8.5: Summary of triple trading for the period 2008-2009



(a) Bollinger Simple (b) Bollinger Complex (c) Z-score

Figure 8.4: Portfolio valuation of each strategy on the period 2008-2009

The valuations for the other periods are provided in Appendices 10.7 and 10.8.

Sharpe Ratio Distributions

Building a portfolio is a good approach to highlight the positive impact of using more complex models. Nevertheless, this cannot be considered as a full proof because the inference is only done on a small set of selected triples and not on the entire population. To reinforce our thoughts, we consider the study without any portfolios on the whole population of the cointegrated triples listed in Table 8.1. The model calibration is performed for each spread with the Cauchy PMCMC algorithm (Algorithm 6). Table 8.6 presents the results of the calibration step on the in-sample sets. Some distributions are presented in Appendix 10.5. The facts $\rho < 0$, $\phi_X \to 1$, $\phi_Z < \phi_X$, $\sigma_Z > \sigma_X$ are in accordance with the stylized facts and the specifications of the $\mathcal{M}_7$ (TFSVL). Not surprisingly, the two crisis periods 2000-2002 and 2008-2010 are the ones which have the highest values of β and σ_X, σ_Z. Recall that β models the amplitude of the absolute daily returns and σ_X, σ_Z represent respectively the long-run and short-run volatilities. It is of common knowledge that periods of crisis come along with high absolute returns and volatility.

Period	$\mu(\phi_X)$	$\mu(\sigma_X)$	$\mu(\phi_Z)$	$\mu(\sigma_Z)$	$\mu(\beta)$	$\mu(\rho)$
Jan, 1 1990 - Jan, 1 1992	0.9853	0.3590	0.4002	0.4987	0.4761	-0.5256
Jan, 1 1992 - Jan, 1 1994	0.9860	0.3468	0.4109	0.4737	0.4552	-0.5419
Jan, 1 1994 - Jan, 1 1996	0.9834	0.3932	0.4102	0.5243	0.4762	-0.5160
Jan, 1 1996 - Jan, 1 1998	0.9808	0.4560	0.4116	0.6149	0.5241	-0.5673
Jan, 1 1998 - Jan, 1 2000	0.9802	0.4780	0.4041	0.4041	0.5426	-0.5729
Jan, 1 2000 - Jan, 1 2002	0.9780	0.5204	0.3970	0.6375	0.5913	-0.5654
Jan, 1 2002 - Jan, 1 2004	0.9842	0.4048	0.3822	0.5600	0.5079	-0.5237
Jan, 1 2004 - Jan, 1 2006	0.9857	0.3759	0.3635	0.6377	0.5102	-0.4270
Jan, 1 2006 - Jan, 1 2008	0.9847	0.4211	0.3620	0.6905	0.5551	-0.4334
Jan, 1 2008 - Jan, 1 2010	0.9813	0.4795	0.3809	0.6424	0.5607	-0.5078
Jan, 1 2010 - Jan, 1 2012	0.9829	0.4144	0.3649	0.6558	0.5460	-0.5052
Jan, 1 2012 - Jan, 1 2014	0.9865	0.3685	0.3366	0.6508	0.4909	-0.4515

Table 8.6: Means of the parameters of $\mathcal{M}_7$ per period

Period	Mean	Median	Min	Max
Jan, 1 1990 - Jan, 1 1992	0.4442 (0.4699)	0.3855 (0.3841)	-2.1859 (-2.6415)	4.2892 (4.4382)
Jan, 1 1992 - Jan, 1 1994	0.5627 (0.5844)	0.5195 (0.5353)	-2.9373 (-3.0505)	3.5681 (3.9564)
Jan, 1 1994 - Jan, 1 1996	0.6751 (0.6219)	0.6075 (0.5751)	-2.3025 (-2.4409)	4.7169 (4.7520)
Jan, 1 1996 - Jan, 1 1998	0.4010 (0.3596)	0.3726 (0.3705)	-3.0290 (-3.1060)	4.1024 (4.0582)
Jan, 1 1998 - Jan, 1 2000	0.2073 (0.1601)	0.1991 (0.0959)	-2.5701 (-2.5804)	4.0478 (5.0003)
Jan, 1 2000 - Jan, 1 2002	0.4488 (0.3212)	0.4118 (0.2567)	-2.2615 (-2.0821)	3.7445 (3.9255)
Jan, 1 2002 - Jan, 1 2004	0.5194 (0.4646)	0.5367 (0.4736)	-3.1040 (-3.5439)	4.2277 (4.1472)
Jan, 1 2004 - Jan, 1 2006	0.5582 (0.5185)	0.5827 (0.5214)	-2.4348 (-2.7721)	4.1926 (4.1905)
Jan, 1 2006 - Jan, 1 2008	0.4027 (0.4089)	0.3539 (0.3878)	-2.5793 (-2.5666)	4.5596 (3.8987)
Jan, 1 2008 - Jan, 1 2010	0.4226 (0.3697)	0.3845 (0.3473)	-2.7525 (-2.7811)	4.0152 (3.8975)
Jan, 1 2010 - Jan, 1 2012	0.6071 (0.5296)	0.5850 (0.5020)	-2.3119 (-2.2534)	4.3517 (4.0816)
Jan, 1 2012 - Jan, 1 2014	0.5137 (0.5021)	0.4877 (0.4958)	-2.8241 (-2.8924)	4.8451 (4.8451)
Total Average	0.4802 (0.4425)	0.4521 (0.4121)	-2.6077 (-2.7259)	4.2217 (4.2659)

Table 8.7: Statistics about $\mathcal{SR}_O$-distributions for simple and complex volatility modelling. Plain is complex model. Brackets is simple model.

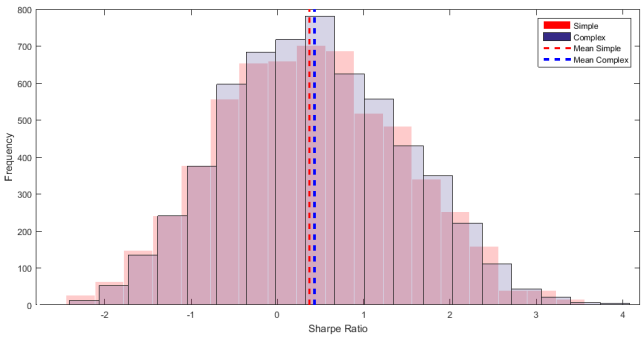
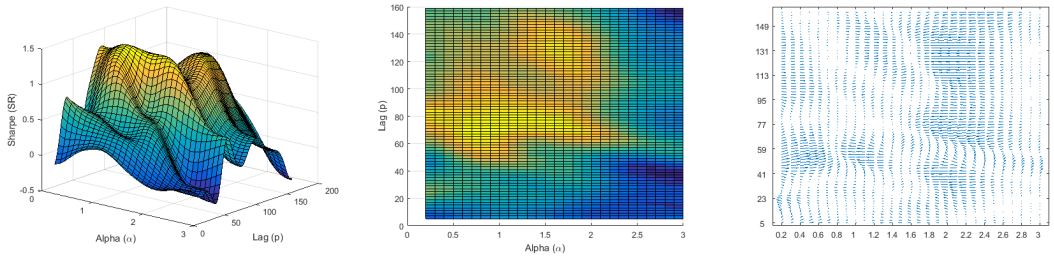


Figure 8.5: Histograms of the Sharpe Ratios on the out-sample set (2008-2009)

The Sharpe Ratios $\mathcal{SR}_{\mathcal{O}}$ statistics are presented in Table 8.7. The values in plain are for complex modelling and the ones in brackets for simple modelling. **It turns out that using complex modelling results in an average increase of 8.52% in the Sharpe-Ratio.** A t-test[3] is conducted to assert that the pairwise difference between the means of simple and complex modelling Sharpe Ratios has a mean equal to 0. With a t-score of 2.815 with 11 degrees of freedom and a pvalue of 0.0168, the Null hypothesis is rejected at the 5% significance level. Figure 10.4 shows the distributions of the Sharpe Ratios with simple and complex modelling on the period 2008-2009. We can see clearly that using complex modellings leads to an improvement in this measure.

8.3 Gradient and Optimization of the Bollinger bands

It becomes now interesting to see if the optimization of the default Bollinger bands parameters can lead to an increase of the profitability. To have an overview, 50 spreads are randomly selected and cross validated on each set $(\mathcal{I}_i, \mathcal{O}_i)_{1 \leq i \leq 12}$. Figure 8.6 introduces the topology of f after the validation on $\mathcal{O}$ and the gradient ∇f. The type of moving average is $t = \{EMA\}$ and is fixed. At first sight, the region where the global maximum seems to be around $(80, 1)$. In this region, the gradient seems to be fairly constant and close to 0. It gives us a hint that the bands are optimal for very large values of p and flexible for α. This result is in accordance with the good performance of the Z-score strategy that assumes a non-zero fixed mean. In fact, this case happens when $p \to \infty$. However, the whole population should be analyzed before making any hasty conclusions.



(a) Topology of f (3D) (b) Topology of f (2D) (c) Gradient ∇f

Figure 8.6: Detection of the stable global maximum of f with ∇f

[3]For more information, the reader can consult the MATLAB documentation on `http://mathworks.com/help/stats/ttest.html`

9 Conclusion and Future Work

The strategy is compared to the traditional buy and hold strategy where the investor buys a basket of stocks to reproduce the S&P500 index and holds it until the end of the period where the position is unwound. Table xx presents the results of both strategies. Figure xx compares the cumulative excess returns and volatility of the strategy with the ones of the SPX index. The portfolio composed of the tuples shows very little volatility compared to the Buy and Hold strategy of the S&P500 index. The second panel presents the implied volatility of the returns for both strategies computed with a standard stochastic volatility model. The strategy accounts for a low and stable volatility for the whole period. A very low correlation with the market returns attests the market neutral property of the strategy. Table 3 shows the performance year by year of the strategy and it is worth noticing that the excess returns is very high during the crisis where the volatility was very high. As highlighted by Khandani and Lo (2007) and Avellaneda and Lee (2010), the second semester of 2007 and first semester of 2008 were quite complicated for quantitative investment funds. Particularly for statistical arbitrage strategies that experienced significant losses during the period, with subsequent recovery in some cases. Many managers suffered losses and had to sell out their portfolios, not benefiting from the subsequent recovery. We obtain results which are consistent with Khandani and Lo (2007) and Avellaneda and Lee (2010) and validate their unwinding theory for the quant fund drawdown. Note that in Figure 3, the proposed pairs trading strategy presented significant losses in the first semester of 2008, starting its recovery in the second semester. Khandani and Lo (2007) and Avellaneda and Lee (2010) suggest that the events of 2007-2008 may be a consequence of a lack of liquidity, caused by funds that had to undo their positions. The proposed statistical arbitrage generated average excess returns of xx% per year in out-of-samples simulations, Sharpe ratio of xx, low exposure to the equity market and relatively low volatility and 5pt basis for transaction costs. Even in market crashes, it turns out that the strategy is still highly profitable, reinforcing the usefulness of co-integration in quantitative strategies.

STILL UNDER PROGRESS

Summary Statistics of the tuple Trading strategy	Strategy	SPX (Buy and Hold)
# of observations in the sample	8844	
# of observations in the training window	170	
# of days in the trading period	84	
# of trading periods	1	
# of pairs in each trading period	20	
# min of cointegrated pairs in a trading period	35000	
# max of cointegrated pairs in a trading period	35000	
Average annualized return	17.88%	
Annualized volatility	6.92%	
Annualized Sharpe Ratio	2.54	
Largest daily return	2.80%	
Lowest daily return	-1.94%	
Cumulative profit	844.48%	
Correlation with the market returns	0.061	
Skewness	1.09	
Kurtosis	19.89	
Maximum Drawdown	3.80%	

Bibliography

C. Alexander and A. Dimitriu. The cointegration alpha: Enhanced index tracking and long-short equity market neutral strategies. 2002.

C. Andrieu, A. Doucet, and R. Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3): 269–342, 2010.

M. Avellaneda and J.-H. Lee. Statistical arbitrage in the us equities market. *Quantitative Finance*, 10(7):761–782, 2010.

F. Black. Studies of stock price volatility changes. *Proceedings of the Meetings of the American Statistical Association*, 1976.

J. P. Broussard and M. Vaihekoski. Profitability of pairs trading strategy in an illiquid market with multiple share classes. *Journal of International Financial Markets, Institutions and Money*, 22(5):1188–1201, 2012.

J. Caldeira and G. V. Moura. Selection of a portfolio of pairs based on cointegration: A statistical arbitrage strategy. *Available at SSRN 2196391*, 2013.

E. Chan. *Quantitative trading: how to build your own algorithmic trading business*, volume 430. John Wiley & Sons, 2009.

J. C. Chan and A. L. Grant. Modeling energy price dynamics: Garch versus stochastic volatility. 2015.

M. Chernov and E. Ghysels. A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *Journal of financial economics*, 56(3):407–458, 2000.

S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

R. Cont. Long range dependence in financial markets. In *Fractals in Engineering*, pages 159–179. Springer, 2005.

P. B. DAO, W. J. STASZEWSKI, A. KLEPKA, and F. AYMERICH. Impact damage detection in composites using nonlinear vibro-acoustic wave modulations and cointegration analysis. 2014.

P. Del Moral. *Feynman-Kac Formulae Genealogical and Interacting Particle Systems with Applications*. Springer-Verlag, New York, USA, 2004.

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

C. L. Dunis and R. Ho. Cointegration portfolios of european equities for index tracking and market neutral strategies. *Journal of Asset Management*, 6(1):33–52, 2005.

C. L. Dunis, G. Giorgioni, J. Laws, and J. Rudy. Statistical arbitrage and high-frequency data with an application to eurostoxx 50 equities. *Liverpool Business School, Working paper*, 2010.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

R. F. Engle and T. Bollerslev. Modelling the persistence of conditional variances. *Econometric reviews*, 5(1):1–50, 1986.

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

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

A. E. Gelfand and D. K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514, 1994.

C. W. Granger and P. Newbold. Spurious regressions in econometrics. *Journal of econometrics*, 2(2):111–120, 1974.

A. Harvey, E. Ruiz, and N. Shephard. Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264, 1994.

S. Johansen. Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, 12(2):231–254, 1988.

S. Johansen. Likelihood-based inference in cointegrated vector autoregressive models. *OUP Catalogue*, 1995.

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

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

G. Kastner, S. Frühwirth-Schnatter, and H. F. Lopes. Analysis of exchange rates via multivariate bayesian factor stochastic volatility models. In *The Contribution of Young Researchers to Bayesian Statistics*, pages 181–185. Springer, 2014.

A. Khandani and A. Lo. What happened to the quants in august 2007. *Journal of investment management*, 5(4):29–78, 2007.

S. Kim, N. Shephard, and S. Chib. Stochastic volatility: likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65(3):361–393, 1998.

S. J. Koopman and E. Hol Uspensky. The stochastic volatility in mean model: empirical evidence from international stock markets. *Journal of applied Econometrics*, 17(6): 667–689, 2002.

C. R. Nelson and C. R. Plosser. Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of monetary economics*, 10(2):139–162, 1982.

M. S. Perlin. Evaluation of pairs-trading strategy at the brazilian financial market. *Journal of Derivatives & Hedge Funds*, 15(2):122–136, 2009.

M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn. On some properties of markov chain monte carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

G. O. Roberts, A. Gelman, W. R. Gilks, et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

E. Ruiz and H. Veiga. Modelling long-memory volatilities with leverage effect: A-lmsv versus fiegarch. *Computational Statistics & Data Analysis*, 52(6):2846–2862, 2008.

G. W. Schwert. Tests for unit roots: A monte carlo investigation. *Journal of Business & Economic Statistics*, 20(1):5–17, 2002.

S. J. Taylor. Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time series analysis: theory and practice*, 1: 203–226, 1982.

M.-N. Tran, M. Scharth, M. K. Pitt, and R. Kohn. Importance sampling squared for bayesian inference in latent variable models. *Available at SSRN 2386371*, 2014.

H. Veiga. A two factor long memory stochastic volatility model. 2006.

G. Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.

10 Appendices

10.1 Implementation

10.1.1 Source Code

All the source code (algorithms, scripts) has been written in Octave using MATLAB 2015a. Since it is open source, it is available to everyone although you have to follow the licenses as defined in the LICENSE file.

Statistics	
Repository URL	`https://github.com/philipperemy/Statistical-Arbitrage`
Number of commits	165
Number of files	195 (MATLAB extension: .m)
Codebase	8727 Lines
Author	Philippe Remy
First commit	May, 19 2015

Table 10.1: Statistics about the repository

10.1.2 Hierarchical Structure

- `coint/`
  Files related to cointegration tests and research on spreads (triples and quadruples).
- `data/`
  Contains the datasets.
- `filters/`
  Sequential Monte Carlo filters.
- `helpers/`
  Library of useful functions to manipulate data and perform common computations.
- `likelihoods/`
  Set of functions related to model comparisons.
- `models/`
  Stochastic Volatility model classes used for validation.
- `pmcmc/`
  Generic Particle Markov Chain Monte Carlo framework.

- `profiling/`
  Optimization Functions (number of particles, simulated annealing).
- `sandbox/`
  Experimental folder.
- `scripts/`
  Routine Scripts to run tests, validate models and interact with the git remote repository.
- `strategy/`
  Trading framework gathering strategies (Bollinger bands, Z-score).
- `test/`
  Test folder. Non regression and validation tests.

10.1.3 How to Get Started

The codebase has thousands of lines of code. Therefore, getting started is not easy. The Particle MCMC framework, implemented for this thesis, is a highly extensible, multi-threaded and customizable framework designed to estimate parameters in non linear state-space models. The source code is provided under the MIT license and is available on Github. Contributions are welcome.

To define a new PMCMC scheme, the user must inherit from the base abstract class and implement the basic functions. The user must define each of its MC chains as protected member variables, define its priors and proposals distributions and finally link a Particle Filter to the class. The convention used for Particle Filter classes is to return the marginal likelihood and the estimated hidden states.

10.2 Multinomial and Stratified Resampling

In this section, we focus on multinomial and stratified resampling. The mathematical framework is taken from Douc and Cappé (2005).

Denote by $(\xi_i, \omega_i)_{1 \leq i \leq n, t>0}$ the set of particle positions and associated weights at time t. The filtration $(\mathcal{F}_t)_{t>0}$ is used to model the information known of the particles and the weights up to time t. The weights are assumed to be normalized, i.e. $\forall t > 0, \sum_{i=1}^{n} \omega_i = 1$. Otherwise, consider $\omega_i \leftarrow \omega_i \,/\, \sum_{j=1}^{n} \omega_j$. The resampling step consists in selecting new particle positions and weights $\left(\widetilde{\xi}_i, \widetilde{\omega}_i\right)_{1 \leq i \leq n}$ at time $t+1$ such that the discrepancy between the resampled weights $\widetilde{\omega}_i$ is reduced. There are many possible ways to resample. Two methods are discussed in this section: multinomial and stratified resampling.

Multinomial resampling is at the core of the Bootstrap method that consists in drawing, conditionally upon $\mathcal{F}_t$, the new positions $(\xi_i)_{1\leq i\leq n}$ independently. In practice, this is achieved by repeated uses of the inversion method

- Draw n independent uniforms $(U^i)_{1\leq i\leq n}$ on the interval $(0,1]$.
- Set $I^i = D_\omega^{inv}(U^i)$ and $\widetilde{\xi}_i = \xi_{I^i}$ where D_ω^{inv} is the inverse of the cumulative distribution associated with the normalized weights $(\omega_i)_{1\leq i\leq n}$, that is $D_\omega^{inv}(u) = i$ for $u \in \left(\sum_{j=1}^{i-1}\omega_j, \sum_{j=1}^{i}\omega_j\right)$. For better clarity, the function $\xi(i) = \xi_i$ is written as $\xi \circ D_\omega^{inv}(U^i)$.

This form of resampling is known as multinomial since the duplication counts are by definition distributed according to the multinomial distribution.

Stratified resampling is based on concepts used in survey sampling and consists in pre-partitioning the $(0,1]$ interval into n disjoint sets, $(0,1] = (0,1/n]\cup\cdots\cup(1-1/n,1]$. The uniform random variables U^i are then drawn independently in each of these sub-intervals: $U^i \sim \mathcal{U}\left(\frac{i-1}{n}, \frac{i}{n}\right)$. Then, the inversion method is used as in multinomial resampling.

Proof. For multinomial resampling, the selection indices $I^1, \cdots, I^n$ are conditionally i.i.d. given $\mathcal{F}_t$ and thus the conditional variance is given by

$$\begin{aligned}
\mathrm{Var}_M\left[\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] &= \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n}\omega_i f^2(\xi_i) - n\left(\sum_{i=1}^{n}\omega_i f(\xi_i)\right)^2\right\} \qquad (10.1)
\end{aligned}$$

An important result for Stratified resampling is

$$\begin{aligned}
E\left[\sum_{i=1}^{n} f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] &= E\left[\sum_{i=1}^{n} f\circ\xi\circ D_\omega^{inv}(U^i)\middle|\mathcal{F}_t\right] \\
&= \sum_{i=1}^{n} E\left[f\circ\xi\circ D_\omega^{inv}(U^i)\middle|\mathcal{F}_t\right] \\
&= n\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_\omega^{inv}(u)\,du \qquad (10.2) \\
&= n\sum_{i=1}^{n}\omega_i f(\xi_i)
\end{aligned}$$

$U^1, \cdots, U^n$ are still conditionally independent given $\mathcal{F}_t$ for the stratified resampling

$$\begin{aligned}
\mathrm{Var}_S\left[\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] &= \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left\{E\left[f\circ\xi\circ D_\omega^{inv}(U^i)^2\middle|\mathcal{F}_t\right] - E\left[f\circ\xi\circ D_\omega^{inv}(U^i)\middle|\mathcal{F}_t\right]^2\right\} \\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n} f\circ\xi\circ D_\omega^{inv}(U^i)^2\middle|\mathcal{F}_t\right] - \frac{1}{n^2}E\left[f\circ\xi\circ D_\omega^{inv}(U^i)\middle|\mathcal{F}_t\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\omega_i f^2(\xi_i) - \frac{1}{n^2}\sum_{i=1}^{n}\left[n\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_\omega^{inv}(u)\;du\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\omega_i f^2(\xi_i) - \sum_{i=1}^{n}\left[\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_\omega^{inv}(u)\;du\right]^2 \qquad (10.3)
\end{aligned}$$

By Jensen's inequality,

$$\sum_{i=1}^{n}\left[\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_\omega^{inv}(u)\;du\right]^2 \geq \left[\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_\omega^{inv}(u)\;du\right]^2 = \left[\sum_{i=1}^{n} w_i f(\xi_i)\right]^2 \qquad (10.4)$$

Finally,

$$\mathrm{Var}_M\left[\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] \geq \mathrm{Var}_S\left[\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{\xi}_i)\middle|\mathcal{F}_t\right] \qquad (10.5)$$

which closes the proof. □

10.3 Correlation Analysis of Quadruples

Figure 10.1 presents the distributions of R^2 for each stock sector for quadruples. The period spans the time from Jan 01, 2012 to May 27, 2013. Most distributions exhibit a bell shape with thin right tails and are therefore candidates for a filtering selection based on an arbitrary threshold R^2_{th}. It is worth noting that for $n = 4$, each sector has its own threshold R^2_{th}.

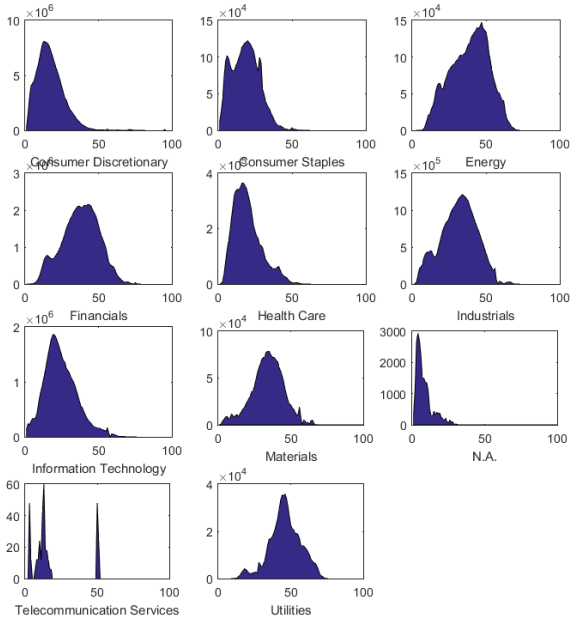


Figure 10.1: Densities of $100 \times R^2$ for the quadruples (not all are cointegrated). Period is from Jan 01, 2012 to May 27, 2013

Sector name	Count before filtering	Measure R^2_{thr}	Count after filtering
Consumer Discretionary	165986922	0.95	15936
Consumer Staples	3025246	0.54	1074
Energy	4651592	0.70	1536
Financials	69777874	0.77	2730
Health Care	7567468	0.56	2982
Industrials	36063822	0.70	1338
Information Technology	44043326	0.72	1080
Materials	1972014	0.66	2070
N.A	23232	0.21	1080
Telecommunication Services	360	0.00	360
Utilities	760164	0.72	1290

Table 10.2: Correlation filtering for the quadruples on Jan 01, 2012 - May 27, 2013

Table 10.2 shows the number of quadruples before and after the filtering. R^2_{thr} has been selected in such a way that roughly between 1000 and 10000 quadruples are selected for each sector for cointegration tests.

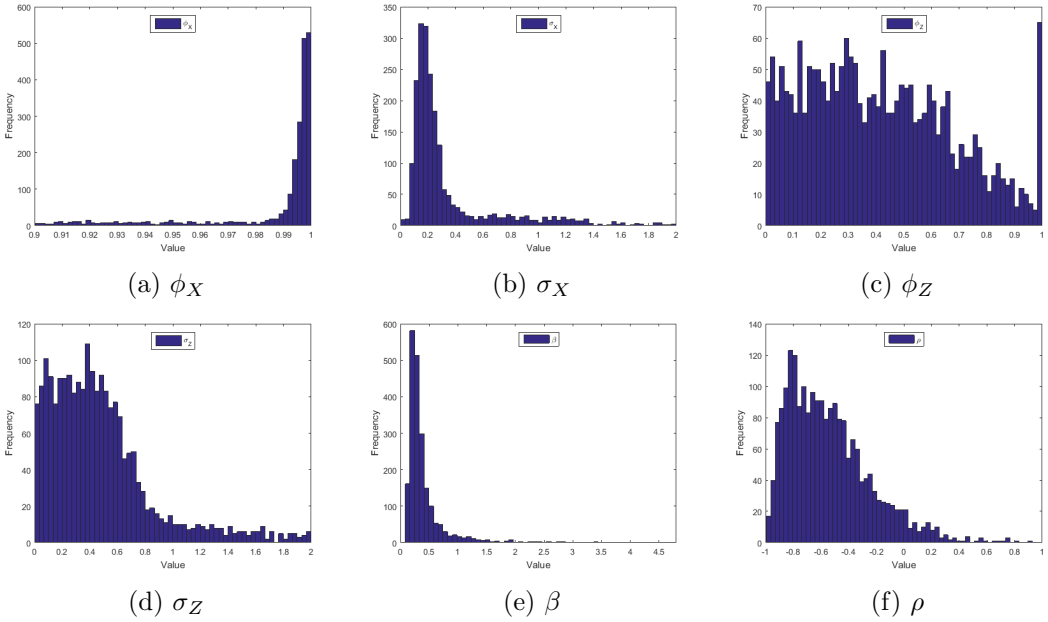
10.4 Cointegration on Foreign Exchange Rates (FX)

The dataset $\mathcal{FX}_{daily}$ presented in Section 1.2.1 is used in this section. Table 10.3 presents an overview of the cointegrated prices between the different currency pairs for triple trading.

Currency 1	Currency 2	Currency 3	β_1	β_2	β_3
01-Jan-1999	- 01-Jan-2001				
CADUSD	AUDUSD	EURUSD	1.000000	-0.169999	0.041106
CADUSD	AUDUSD	CHFUSD	1.000000	-0.168860	0.050292
CADUSD	EURUSD	NZDUSD	1.000000	-0.042766	-0.136377
CADUSD	EURUSD	CHFUSD	1.000000	-0.060119	0.070569
CADUSD	NZDUSD	CHFUSD	1.000000	-0.137284	0.055910
01-Jan-2001	- 01-Jan-2003				
EURUSD	AUDUSD	GBPUSD	1.000000	-0.243406	-0.858606
EURUSD	GBPUSD	CADUSD	1.000000	-0.942491	-0.167773
CHFUSD	GBPUSD	CADUSD	1.000000	-0.773401	-0.197333
EURUSD	GBPUSD	NZDUSD	1.000000	-0.843365	-0.251468
EURUSD	GBPUSD	CHFUSD	1.000000	-0.508161	-0.566663
CHFUSD	GBPUSD	NZDUSD	1.000000	-0.706983	-0.182011
CHFUSD	CADUSD	EURUSD	1.000000	-0.045434	-0.871733
01-Jan-2003	- 01-Jan-2005	(Nothing)			
01-Jan-2005	- 01-Jan-2007				
EURUSD	AUDUSD	GBPUSD	1.000000	-0.225093	-0.672751
EURUSD	GBPUSD	ZARUSD	1.000000	-0.738529	-0.102339
01-Jan-2007	- 01-Jan-2009	(Nothing)			
01-Jan-2009	- 01-Jan-2011	(Nothing)			
01-Jan-2011	- 01-Jan-2013				
GBPUSD	AUDUSD	CADUSD	1.000000	-0.308276	-0.168456
GBPUSD	CADUSD	NZDUSD	1.000000	-0.268217	-0.227796
GBPUSD	CADUSD	ZARUSD	1.000000	-0.276975	-0.172995
GBPUSD	CADUSD	CHFUSD	1.000000	-0.432923	-0.177363

Table 10.3: Cointegration on FX Rates between Jan, 1 1999 and Jan, 1 2013

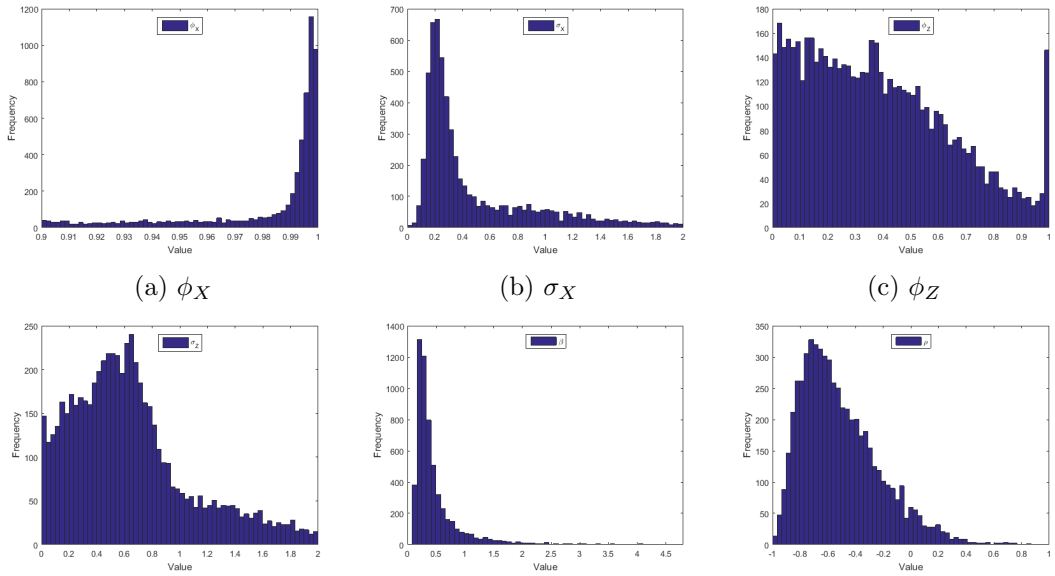
10.5 Distribution of model parameters



(a) ϕ_X (b) σ_X (c) ϕ_Z

(d) σ_Z (e) β (f) ρ

Figure 10.2: Distributions of the Stochastic Volatility parameters of $\mathcal{M}_7$ for 2147 spreads (Jan 1992 - Jan 1994).



(a) ϕ_X (b) σ_X (c) ϕ_Z

(d) σ_Z (e) β (f) ρ

Figure 10.3: Distributions of the Stochastic Volatility parameters of $\mathcal{M}_7$ for 5967 spreads (Jan 2008 - Jan 2010).

10.6 Cauchy PMMH

In order to balance the high computational cost of estimating thousands of parameters, we derived a slightly different version of the PMMH algorithm of Andrieu et al. (2010) that we named Cauchy PMMH. The purpose of Cauchy PMMH is not to sample from the posterior distributions but to estimate the parameters with MLE in the most efficient way. As the name would suggest, the algorithm is based on Cauchy sequences to assess the convergence of the conditional marginal likelihood estimator $\log \hat{p}_\theta^N(y_{1:T})$. A cauchy sequence is a sequence whose elements become arbitrarily close to each other as the sequence progresses. The termination criterion is satisfied when $\log \hat{p}_\theta^N(y_{1:T})$ has converged at precision ϵ over the last k values. Algorithm 6 describes it.

In practical applications, $N = T$, $\epsilon = 1$, $k = 50$ and q is the centered multivariate normal distribution. On average, the convergence is reached for i between 75 and 150, which is respectively 133x and 67x faster than the default implementation with $N = 10000$. However, this huge performance gain is balanced by the possible termination onto local maxima. Our benchmarks revealed that it is rarely the case. Initial values of θ were randomly drawn and the Cauchy PMMH always converged to the true values.

Algorithm 6 Particle pseudo marginal Metropolis-Hastings Algorithm (Cauchy)

procedure CauchyPMMH($y_{1:T}$, a proposal distribution $q(\cdot|\cdot)$, the number of particles N, precision ϵ, k)
  Set static parameter vector $\theta^{(1)}$ arbitrarily
  $\hat{p}^N_{\theta^{(1)}}(y_{1:T}), \mathbf{x}^{*(1)}_{1:T} \leftarrow$ Call Bootstrap Particle Filter with $(y_{1:T}, \theta^{(1)}, N)$
  $i \leftarrow 2$
  while last k values of $\log \hat{p}_N(y|\theta^{(i-1:k)})$ are not steady at precision ϵ **do**
    Sample θ' from $q(\theta'|\theta^{(i-1)})$
    $\hat{p}^N_{\theta'}(y_{1:T}), \mathbf{x}^{*'}_{1:T} \leftarrow$ Call Bootstrap Particle Filter with $(y_{1:T}, \theta', N)$
    If,

$$\frac{q(\theta^{(i-1)}|\theta')\hat{p}_N(y_{1:T}|\theta')p(\theta')}{q(\theta'|\theta^{(i-1)})\hat{p}_N(y_{1:T}|\theta^{(i-1)})p(\theta^{(i-1)})} \geq 1$$

    Set $\mathbf{x}^{*(i)}_{1:T} \leftarrow \mathbf{x}^{'*}_{1:T}, \theta^{(i)} \leftarrow \theta', \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta'}(y_{1:T})$
    Otherwise $\mathbf{x}^{*(i)}_{1:T} \leftarrow \mathbf{x}^{*(i-1)}_{1:T}, \theta^{(i)} \leftarrow \theta^{(i-1)}, \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta^{(i-1)}}(y_{1:T})$
    $i \leftarrow i + 1$
  end
return $\{\mathbf{x}^{*(i)}_{1:T}, \hat{\theta}_{mle} = \theta^{(i)}\}$

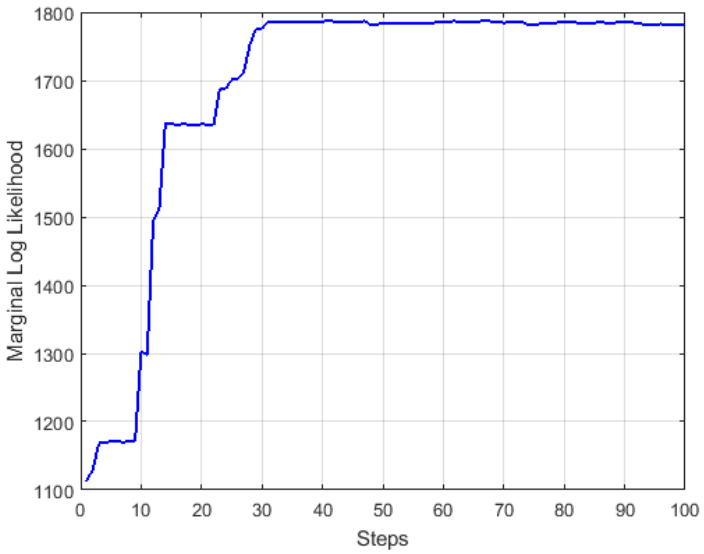
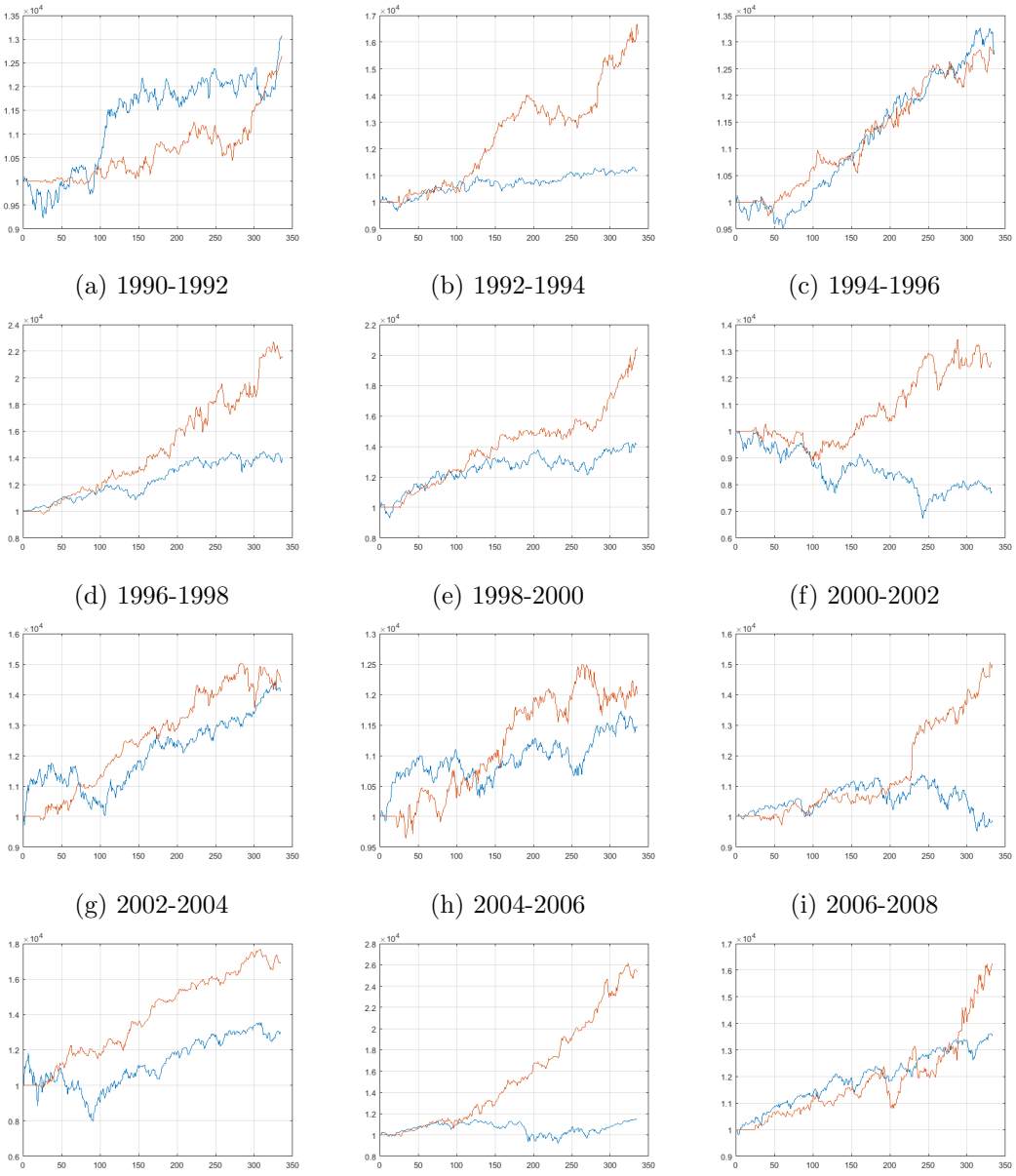


Figure 10.4: Convergence of $\log \hat{p}_N(y|\theta^{(i)})$ on synthetic data generated from $\mathcal{M}_2$. Convergence is at $i = 80$ for $\epsilon = 2, k = 50, N = 2T$.

10.7 Bollinger bands Strategy - Cumulative Returns



(a) 1990-1992 (b) 1992-1994 (c) 1994-1996

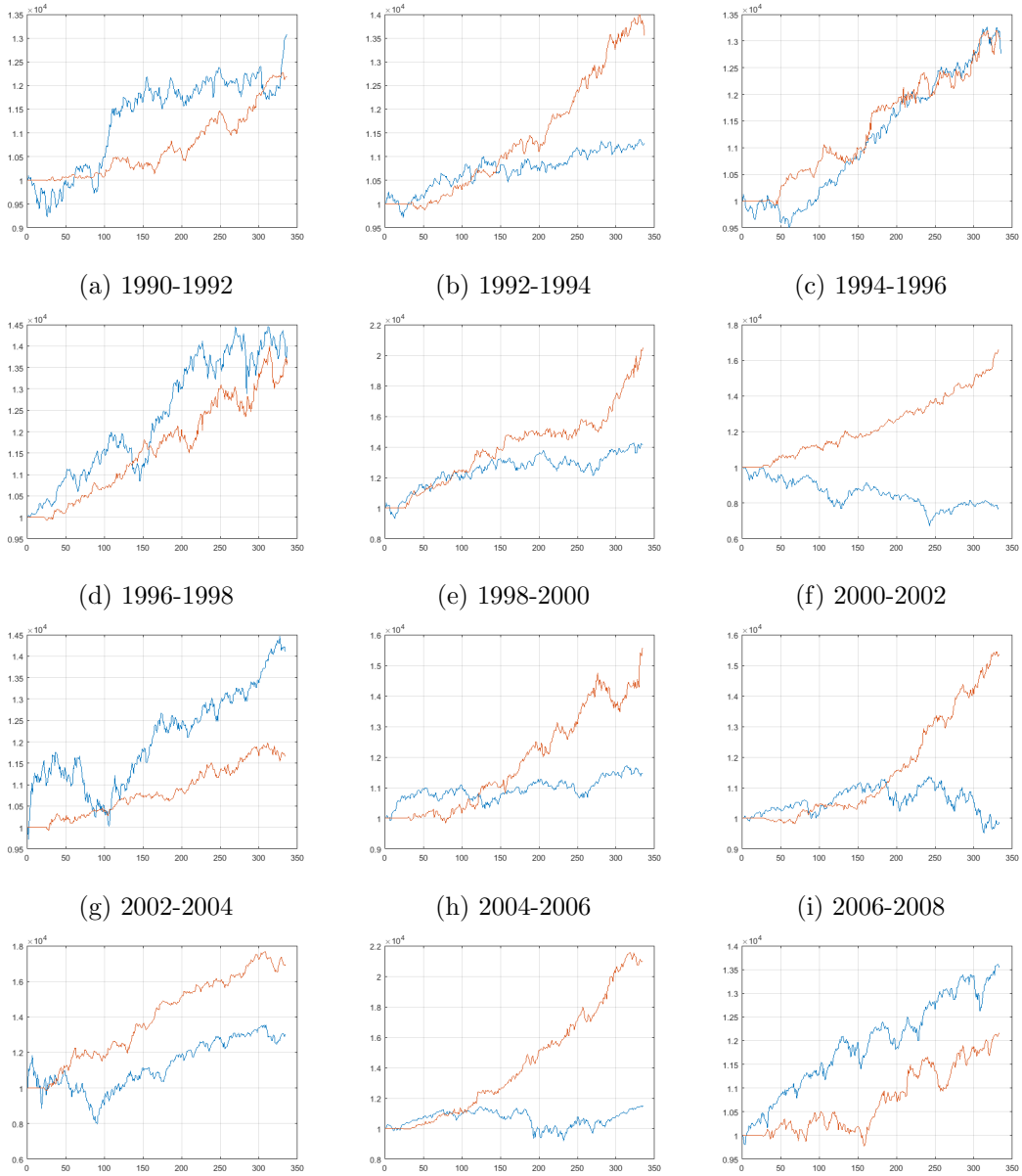
(d) 1996-1998 (e) 1998-2000 (f) 2000-2002

(g) 2002-2004 (h) 2004-2006 (i) 2006-2008

(j) 2008-2010 (k) 2010-2012 (l) 2012-2014

Figure 10.5: Cumulative returns for Bollinger bands strategy with simple modelling

10.8 Bollinger bands Strategy - Cumulative Returns (complex)



(a) 1990-1992 (b) 1992-1994 (c) 1994-1996

(d) 1996-1998 (e) 1998-2000 (f) 2000-2002

(g) 2002-2004 (h) 2004-2006 (i) 2006-2008

(j) 2008-2010 (k) 2010-2012 (l) 2012-2014

Figure 10.6: Cumulative returns for Bollinger bands strategy with simple modelling