Imperial College London

Imperial College London
Derpartment of Mathematics

State Space Modelling for Statistical Arbitrage

Philippe Remy

CID: 00993306

Supervised by Nikolas Kantas and Yanis Kiskiras

7th August 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.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Nothing added yet. Stay tuned.

Abstract

Statistical Arbitrage is a computationally-intensive approach which involves the simultaneous buying and selling of security portfolios according to predefined or adaptive statistical models. Statistical Arbitrage strategies are heavily based on the construction of stationary mean-reverting spreads and sophisticated models to identify opportunities. This thesis enriches the classic cointegration-based pairs trading by considering two cases: triples of assets, and quadruples where one is an index. 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 from the US exchanges. Two strategies are discussed in this thesis: Bollinger Bands and Z-score. The volatility of the traded assets is estimated using several stochastic volatility models where the parameters are estimated via Particle Markov Chain Monte Carlo. The profitability of the strategy is assessed with data from the S&P500 on 1232 stocks between 01-Jan-1990 and 19-Mar-2014. Empirical analysis shows that the proposed strategy accounts for excess returns of 17% per year, Sharpe Ratio above 2 and low correlation with the market.

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 realisation 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 $\Delta X_t = (1-L)X_t = X_t - X_{t-1}$
$E[\mathbf{X}_t\|\mathcal{F}_t]$	Conditional expectation
$\mathrm{Var}[\mathbf{X}_t\|\mathcal{F}_t]$	Conditional variance
Cor	Correlation function
$\circ$	Composition function operator
$p(\cdot)$	General marginal probability density function
$p(\cdot\|\cdot)$	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 $p \in \mathbb{N}^* \cup \infty$

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 state-space model $(\mathbf{X}_t, \mathbf{Y}_t)_{t\in\mathbb{N}}$ is adapted to a two steps filtration $(\mathcal{F}_t^2)_{t\in\mathbb{N}}$ if $\mathbf{X}_t$ and $\mathbf{Y}_t$ can be measured respectively at time t^- and t, where $t^- = t-\epsilon$. The concept of two steps is important here to understand that x_t and y_t are not measured at the same time but in a sequential way where the difference between the measurement times converges to 0. Therefore, the distribution $\mathcal{D}(\mathbf{Y}_t|\mathbf{X}_t = \mathbf{x}_t, \theta)$ becomes $\mathcal{D}(\mathbf{Y}_t|\mathcal{F}_{t^-})$ where $\mathcal{F}_{t^-} = \{\mathbf{x}_{1:t}, \theta\}$*

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 to address this problem. Cointegration is a concept that 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.

Markov Chain Monte Carlo (MCMC) methods are well known techniques for sampling from a probability distribution. It is based on constructing a Markov chain targeting this distribution. Andrieu et al. (2010) introduced a new method which embed SMC 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 behaviour in complex practical applications remain open research questions.

The GARCH model and the Stochastic Volatility model are competing but non-nested models to describe unobserved volatility in asset returns. The former 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) proposes 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 (SV) model. Even though several papers such as Kim et al. (1998) provide early evidence in favour of using SV, these have found comparably little use in applied work. The main discrepancy relied in the incapability of estimating the parameters of the SV models. It becomes now possible with techniques such as Particle Markov Chain Monte Carlo. Kastner et al. (2014) analysed exchanges rates from EUR to USD and showed that a standard SV performs better than a vanilla GARCH(1,1) in terms of predictive accuracy. Chan and Grant (2015) compare a number of GARCH and SV models on commodity markets. SV models generally compare favourably to their GARCH counterparts. The SV models have been retained as the default models for all the reasons specified above.

This thesis focuses on the development and the estimation of stochastic volatility models to output an accurate estimate of the volatility of the co-integrated prices. This 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 gauging 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 many tuple trades all selected based on criteria such as co-integration. 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 sample period starts in January 1990 and ends in March 2014, summing up to 8844 observations. The 1220 most liquid stocks of the US exchanges are used for the study. Daily equity closing prices are obtained from Bloomberg. The proposed statistical arbitrage strategy generated average excess returns of 17% per year in out-of-sample simulations, Sharpe Ratio of 2, low exposure to the equity markets and relatively low volatility.

The remainder of this thesis is organized as follows. In Section 2, co-integration is presented in greater details. Section 3 introduces the Particle Markov Chain Monte Carlo framework and the stochastic volatility models. In section 4, the trading strategies are described. In section 5, the data and the results obtained are discussed. In section 6, a conclusion based on the empirical results is presents, along with suggestions of future research.

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 behaviour of economic forces that influence the share prices. Thus statistical arbitrage is a purely statistical approach designed to exploit equity 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 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-tun 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 be two independent random walk for every t, and let's consider the regression : $y_t = ax_t + b + \epsilon_t$. It is obvious that the true value of a is 0 because $cor(x_t, y_t) = 0$. But the limiting distribution of $\hat{a}$ is such that $\hat{a}$ converges to a function of Brownian motions. 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 statistical applies for the regression : $\Delta y_t = b + a\Delta x_t + \epsilon_t$ since both are stationary variables. $\hat{a}$ is now a standard consistent estimator.

Cointegration is said to exist between two or more non-stationary time series if they possess the same order of integration and a linear combination of these series is stationary. Let $X_t = (x_{1t}, ..., x_{nt})_{t\geq 0}$ be n $I(1)$ processes. The vector $(X_t)_{t\geq 0}$ is said to be cointegrated if there exists at least one non trivial vector $\beta = (\beta_1, ..., \beta_n)$ such that $\epsilon_t = \beta^T X_t$ is a stationary process $I(0)$. β is called a cointegration vector, then so is $k\beta$ for any $k \neq 0$

since $k\beta^T X_t \sim I(0)$. There can be r different cointegrating vector, 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 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 cointegrating vector β must be well balanced. If a coefficient of β is very large compared to the others, it means that the investor is 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 Autoregressive (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 $\mathbf{X}_t = (x_{1t}, ..., x_{nt})_{t \geq 0}$, each of the A_j is a (nxn) 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 X_t is said to be stable if the roots of the determinant of the characteristic polynomial $|\mathbf{I}_n - \sum_{j=1}^{k} \mathbf{A}_j z^j| = 0$ lie outside 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 mat also be cointegrated. If $\mathbf{Y}_t$ 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 models (VECM)

In an error correction model (ECM), the changes in a variable depend on the deviations from some equilibrium relation. Suppose the case $n = 2, \mathbf{X}_t = (x_t, y_t)$ 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 further more that the equilibrium relation between them is given by $y_t = \beta x_t$ and the increments of y_t, Δy_t depend 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:

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

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

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

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

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

If $0 < rank(\prod) = r < n$. This implies that $\mathbf{Y}_t$ is $I(1)$ with r linearly independent cointegrating vectors and $n - r$ common stochastic trends (unit roots). Since $\prod$ has rank r, it can be written as the product $\prod = \alpha\beta'$ 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 lements of α distribute the impact of the cointegrating vectors to the evolution of $\Delta \mathbf{Y}_t$. The VECM becomes

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

where $\beta' \mathbf{Y}_{t-1} \sim I(0)$ 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 $\prod = \alpha\beta'$ is not unique since for any $r \times r$ nonsingular 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.

The cointegration relations can be estimated with a Johansen test, as explained in Johansen (1988). The main advantage is that it permits more than one cointegrating relationship and 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 $(\prod) < r$ 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 $\mathbf{Y}_t$. 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.

Since the rank of the long-run impact matrix $\prod$ gives the number of cointegrating relationships in $\mathbf{Y}_t$, Johansen (1988) formulates LR statistics to determine the rank of $\prod$. These LR tests are based on the estimated eigenvalues $\hat{\lambda_1} > \hat{\lambda_2} > ...\hat{\lambda_n}$ of the matrix $\prod$. Note that r is equal to the number of non-zero eigenvalues of $\prod$. If it is found that rank $(\prod) = r, 0 < r < n$, 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 MLEs of the remaining parameters are obtained by least squares estimation of

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

The columns of $\hat{\beta'_{mle}}$ are the estimators of the cointegrating vectors.

The specification of the deterministic terms 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 $\mathbf{Y}_t$. Johansen (1995) classified the trend behavior of $\mathbf{Y}_t$ in five cases:

- Model $H_2(r) : \mu_t = 0$. No constant. The series in $\mathbf{Y}_t$ are $I(1)$ without drift and the cointegrating relation $\beta'\mathbf{Y}_t$ have mean zero
- Model $H_1^*(r) : \mu_t = \mu_0 = \alpha\rho_0$. Restricted constant. The series in $\mathbf{Y}_t$ are $I(1)$ without drift and the cointegrating relation $\beta'\mathbf{Y}_t$ have non-zero mean ρ_0.
- Model $H_1(r) : \mu_t = \mu_0$. Unrestricted constant. The series in $\mathbf{Y}_t$ are $I(1)$ with drift vector μ_0 and the cointegrating relation $\beta'\mathbf{Y}_t$ may have a non-zero mean.
- Model $H^*(r) : \mu_t = \mu_0 + \alpha\rho_1 t$. Restricted trend. All the series in $\mathbf{Y}_t$ are $I(1)$ without drift and the cointegrating relation $\beta'\mathbf{Y}_t$ 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 $\mathbf{Y}_t$ are $I(1)$ with a linear trend and the cointegrating relation $\beta'\mathbf{Y}_t$ have a linear trend.

$H_1(r)$ seems to be definitely the most relevant model to use for spreads because there is drift in most of the assets composing $\mathbf{Y}_t$. This model eliminates both stochastic and deterministic trends in the cointegrating vectors.

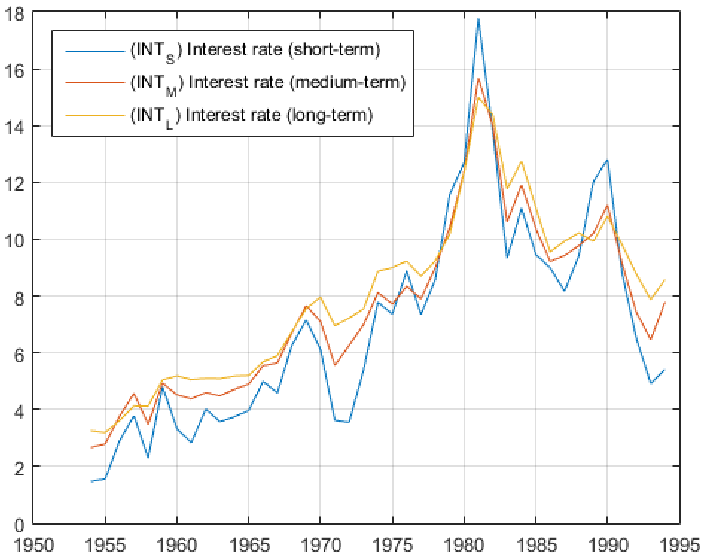


Figure 2.1: Cointegration of the interest rates (short, medium and long-term) in Canada from 1955 to 1995

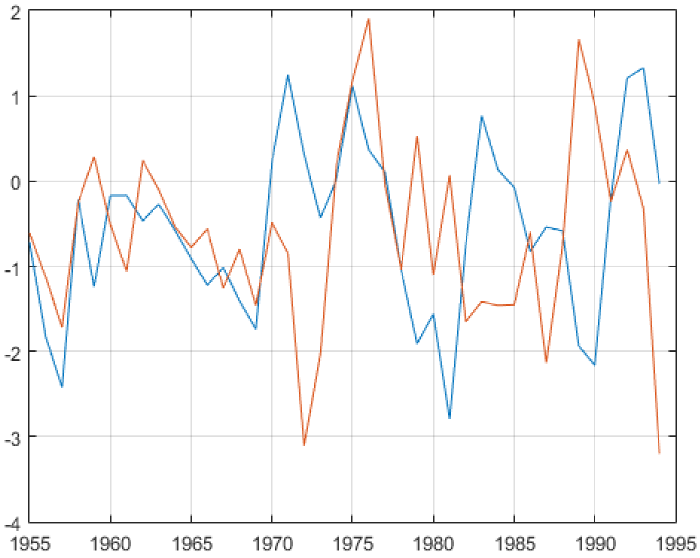


Figure 2.2: Estimated Cointegrating relations $\beta' y_{t-1} + c_0$

It seems that the existence of more than one cointegrating vectors (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.

2.4 Testing for Unit Roots in Stochastic Processes

Before testing for a unit root, i.e. the time series is $I(1)$, the time series must be transformed to its linear form. Usually, assets prices have an exponential growth and logarithm must be applied accordingly to satisfy this prerequisite. However, it is expected not to be the case for a spread instrument.

Once the data is transformed, we must choose the most pertinent model to use in the Augmented Dickey Fuller and Philipps-Perron tests. There are two basic models for economic data y_t with linear growth characteristics:

- Trend Stationary (TS) : $y_t = \gamma y_{t-1} + c + \delta t + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \epsilon_t$
- Auto Regressive with Drift (ARD) : same as TS with $\delta = 0$, and mean $\frac{c}{1-\gamma}$.

where the persistent parameter $\gamma = 1$ is tested against its alternative $\gamma < 1$. ϵ_t is an independent innovation process.

Which model to choose? The trend-stationary is characterized by its ability to revert quickly to its trend unlike the difference-stationary, which can drift away during a long time before reverting. In general, if a series is growing, the TS model provides a reasonable trend-stationary alternative to a unit-root process with drift. If a series is not growing, ARD model provide reasonable stationary alternatives to a unit-root process without drift. As we don't expect any drift component in the structure of a spread, the ARD model seems the most suitable model to use.

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. 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 suggested to use $L_{max} = 12\left(\frac{T}{100}\right)^{\frac{1}{4}}$ 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 averaged ADF t-statistics indicates the value of lag length that produces the most stationary residuals.

The general method to find the optimal lag for the Phillips-Perron test is to begin with few lags, and then evaluate the sensitivity of the results by adding more lags. Another rule of thumb is to look at sample autocorrelations of $\Delta y_t = y_t - y_{t-1}$. Slow rates of decay require more lags. It is less suitable for economic data because it is widely known that the returns Δy_t show no autocorrelation. Finally, when the optimal lags are known, multiple tests are run to avoid any possible inconsistency.

3 State-Space Models and Particle MCMC

4 Particle Markov Chain Monte Carlo

4.1 Introduction

Particle Markov Chain Monte Carlo (Particle MCMC) is a powerful technique for estimating parameters of a complex model where classical methods such as maximum likelihood estimation are limited. This is the case for State-Space models which incorporate latent variables. Particle MCMC embeds a particle filter within an MCMC scheme. The standard version uses a particle filter to propose new values for the stochastic process (basically $x_{0:T}$), and MCMC moves to propose new values for the parameters (usually named θ). One of the most challenging task in designing a PMCMC sampler is considering the trade-off between the Monte Carlo error of the particle filter and the mixing of the MCMC moves. Intuitively, when N, the number of particles grows to infinity, the variance of the error becomes very small and the mixing of the chain becomes very poor.

4.2 State-Space Models

The state-space models are parametrised by $\theta = (\theta_1, ..., \theta_n)$ and all components are considered to be independent one another. θ is associated a prior distribution $p(\theta) = \prod_i p(\theta_i)$ State-space models are usually defined in continuous time because physical laws are most often described in terms of differential equations. However, most of the time, a discrete-time representation exists. It is often written in the innovation form that describes noise. An example of such a process is describe in the Stochastic Volatility section. The model is composed of an unobserved process $X_{0:T}$ and $Y_{1:T}$, known as the observations. $X_{0:T}$ is assumed to be first order markovian, governed by a transition kernel $K(x_{t+1}|x_t)$. The probability density of a realization $x_{0:T}$ is written as

$$p(X_{0:T} = x_{0:T}|\theta) = p(x_1|\theta) \prod_{t=2}^{T} p(x_t|x_{t-1}, \theta) \tag{4.1}$$

The process X is not observed directly, but through $y_{1:T}$. The state-space model assumes that each y_t is dependent of x_t. As a consequence, the conditional likelihood of the observations, given the state process can be derived as

$$p(y_{1:T}|x_{1:T}, \theta) = \prod_{t=1}^{T} p(y_t|x_t, \theta) \tag{4.2}$$

The general idea is to find θ which maximize the marginal likelihood $p(y_{1:T}|\theta)$, x integrated out. It is interesting to begin by the approximation of $p(x_{1:T}, \theta|y_{1:T})$. By Bayes theorem

$$\begin{aligned} p(x_{1:T},\theta|y_{1:T}) &\propto p(\theta)p(x_{1:T}|\theta)p(y_{1:T}|x_{1:T},\theta) \\ &= p(\theta)p(x_1|\theta)\prod_{t=2}^{T} p(x_t|x_{t-1},\theta)\prod_{t=1}^{T} p(y_t|x_t,\theta) \end{aligned} \tag{4.3}$$

Usually, this probability density function is intractable since it becomes incredibly demanding in resources when T grows. That is where the particle filter comes in.

4.3 Particle Filter

The particle filter is an iterative Monte Carlo method for carrying out Bayesian inference on state-space models. The main idea is to assume that, at each time t, an approximation of $p(x_t|y_{1:t})$ can help generating approximate samples of $p(x_{t+1}|y_{1:t+1})$, using importance resampling.
More precisely, the procedure is initialised with a sample from $x_0^k \sim p(x_0),\ k = 1,\dots,M$ with uniform normalised weights $w'^k_0 = 1/M$. Then suppose that we have a weighted sample $\{x_t^k, w'^k_t | k = 1,\dots,M\}$ from $p(x_t|y_{1:t})$. First generate an equally weighted sample by resampling with replacement M times to obtain $\{\tilde{x}_t^k | k = 1,\dots,M\}$ (giving an approximate random sample from $p(x_t|y_{1:t})$). Note that each sample is independently drawn from $\sum_{i=1}^{M} w'^i_t \delta(x - x_t^i)$. Next propagate each particle forward according to the Markov process model by sampling $x_{t+1}^k \sim p(x_{t+1}|\tilde{x}_t^k),\ k = 1,\dots,M$ (giving an approximate random sample from $p(x_{t+1}|y_{1:t})$). Then for each of the new particles, compute a weight $w_{t+1}^k = p(y_{t+1}|x_{t+1}^k)$, and then a normalised weight $w'^k_{t+1} = w_{t+1}^k / \sum_i w_{t+1}^i$.

Sequential Importance Resampling (SIR) filters with transition prior probability distribution as importance function are commonly known as bootstrap filter. This choice is motivated by the facility of drawing particles and performing subsequent importance weight calculations. Here, $\pi(x_k|x_{0:k-1}, y_{0:k}) = p(x_k|x_{k-1})$ and the weights formula is now

$$w_k^{(i)} = w_{k-1}^{(i)} \frac{p(y_k|x_k^{(i)})p(x_k^{(i)}|x_{k-1}^{(i)})}{\pi(x_k^{(i)}|x_{0:k-1}^{(i)}, y_{0:k})} = w_{k-1}^{(i)} p(y_k|x_k^{(i)}) \tag{4.4}$$

It is clear from our understanding of importance resampling that these weights are appropriate for representing a sample from $p(x_{t+1}|y_{1:t+1})$, 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

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

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

$$p_\theta^N(y_{0:T}) = p(y_1)\prod_{t=2}^{T} p(y_t|y_{1:t-1}) \tag{4.6}$$

As T is usually large, it is preferred to work with the log likelihoods

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

$$\log \hat{p}_\theta^N(y_{1:t}) = \sum_{t=2}^{t} \log\left(\frac{1}{N}\sum_{k=1}^{N} w_t^{(t)}\right) \tag{4.8}$$

Algorithm 1 Bootstrap Particle Filtering Algorithm (SIR)

```
procedure INPUT($y_{1:T}$, $\theta$, N)
    for i from 1 to N do
        Sample $x_1^{(i)}$ independently from $p(x_1)$
        Calculate weights $w_1^{(i)} = p(y_1|x_1^{(i)})$
    end
    $x_1^* = \sum_{i=1}^N x_1^{(i)}.w_1^{(i)}$
    Set $\hat{p}(y_1) = \frac{1}{N}\sum_{i=1}^N w_1^{(i)}$
    for t from 1 to T do
        for i from 1 to N do
            Sample $j$ from 1:N with probabilities proportional to $\{w_{t-1}^{(1)}, ..., w_{t-1}^{(N)}\}$
            Sample $x_t^{(i)}$ from $p(x_t|x_{t-1})$
            Calculate weights $w_t^{(i)} = p(y_t|x_t^{(i)})$
        end
        $x_t^* = \sum_{i=1}^N x_t^{(i)}.w_t^{(i)}$
        Set $\hat{p}(y_{1:t}) = \hat{p}(y_{1:t-1})\left(\frac{1}{N}\sum_{i=1}^N w_t^{(i)}\right)$
    end
return $(x_{1:T}^*, \hat{p}(y_{1:T}))$
```

Again, from the importance resampling scheme, it should be reasonably clear that $p_\theta^N(y_{1:T})$ 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, especially for the particle marginal Metropolis-Hastings Algorithm.

4.4 Resampling phase

Sequential Monte Carlo (Particle filtering) can be decomposed in two main steps: sequential importance sampling (SIS) and resampling. The main drawback of SIS is that it becomes very unstable as T increases due to the discrepancy between the weights, a phenomenon known as weight degeneracy. To stabilize the algorithm and gain some accuracy, it is necessary to perform resampling sufficiently often. This step is also time-critical as it is on the critical path of the Component-Wise PMCMC algorithm. Benchmarks highlighted that it can represent up to half of the time spent in the filter with the Bootstrap scheme. Many different methods exist in the literature: multinomial, stratified, systematic and residuals resampling are such examples. In practical applications, they are generally found to provide comparable results. Despite the lack of complete theoretical analysis of its behaviour, multinomial resampling is probably the most used algorithm because almost all the softwares offer a default implementation of this method as default. We will focus on multinomial and stratified resampling in this section. The proposed 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 $\omega_i \leftarrow \left(\sum_{j=1}^n \omega_j\right)^{-1} \omega_i$. The resampling phase consists in selecting new particle positions and weights $(\xi_i^\sim, \omega_i^\sim)_{1\leq i\leq n}$ at time $t+1$ such that the discrepancy between the resampled weights $\omega_i^\sim$ 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 $\xi_i^\sim = \xi_{I^i}$ 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.

Theorem 3. *Stratified resampling has a lower variance, conditionally upon $\mathcal{F}_t$, than multinomial resampling.*

Proof. Douc and Cappé (2005)
For multinomial resampling, the selection indices $I^1, \cdots, I^n$ are conditionally iid given $\mathcal{F}_t$ and thus the conditional variance is given by:

$$\begin{aligned}\mathrm{Var}_M\left[\frac{1}{n}\sum_{i=1}^{n} f(\xi_i^{\sim})\middle|\mathcal{F}_t\right] &= \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} f(\xi_i^{\sim})\middle|\mathcal{F}_t\right]\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[f(\xi_i^{\sim})\middle|\mathcal{F}_t\right]\\ &= \frac{1}{n}\left\{\sum_{i=1}^{n}\omega_i f^2(\xi_i) - n\left(\sum_{i=1}^{n}\omega_i f(\xi_i)\right)^2\right\} \qquad (4.9)\end{aligned}$$

An important result for Stratified resampling is

$$\begin{aligned}E\left[\sum_{i=1}^{n} f(\xi_i^{\sim})\middle|\mathcal{F}_t\right] &= E\left[\sum_{i=1}^{n} f\circ\xi\circ D_{\omega}^{inv}(U^i)\middle|\mathcal{F}_t\right]\\ &= \sum_{i=1}^{n} E\left[f\circ\xi\circ D_{\omega}^{inv}(U^i)\middle|\mathcal{F}_t\right]\\ &= n\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_{\omega}^{inv}(u)\ du \qquad (4.10)\\ &= n\sum_{i=1}^{n}\omega_i f(\xi_i)\end{aligned}$$

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

$$\begin{aligned}\mathrm{Var}_S\left[\frac{1}{n}\sum_{i=1}^{n} f(\xi_i^{\sim})\middle|\mathcal{F}_t\right] &= \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} f(\xi_i^{\sim})\middle|\mathcal{F}_t\right]\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left\{E\left[f\circ\xi\circ D_{\omega}^{inv}(U^i)^2\middle|\mathcal{F}_t\right] - E\left[f\circ\xi\circ D_{\omega}^{inv}(U^i)\middle|\mathcal{F}_t\right]^2\right\}\\ &= \frac{1}{n^2}E\left[\sum_{i=1}^{n} f\circ\xi\circ D_{\omega}^{inv}(U^i)^2\middle|\mathcal{F}_t\right] - \frac{1}{n^2}E\left[f\circ\xi\circ D_{\omega}^{inv}(U^i)\middle|\mathcal{F}_t\right]^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\omega_i f^2(\xi_i) - \frac{1}{n^2}\sum_{i=1}^{n}\left[n\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_{\omega}^{inv}(u)\ du\right]^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\omega_i f^2(\xi_i) - \sum_{i=1}^{n}\left[\int_{(i-1)/n}^{i/n} f\circ\xi\circ D_{\omega}^{inv}(u)\ du\right]^2 \qquad (4.11)\end{aligned}$$

By Jensen's inequality,

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

$$= \left[\sum_{i=1}^{n} w_i f(\xi_i) \right]^2 \tag{4.12}$$

Finally,

$$\mathrm{Var}_M \left[\frac{1}{n} \sum_{i=1}^{n} f(\xi_i^{\sim}) \middle| \mathcal{F}_t \right] \geq \mathrm{Var}_S \left[\frac{1}{n} \sum_{i=1}^{n} f(\xi_i^{\sim}) \middle| \mathcal{F}_t \right] \tag{4.13}$$

which closes the proof. □

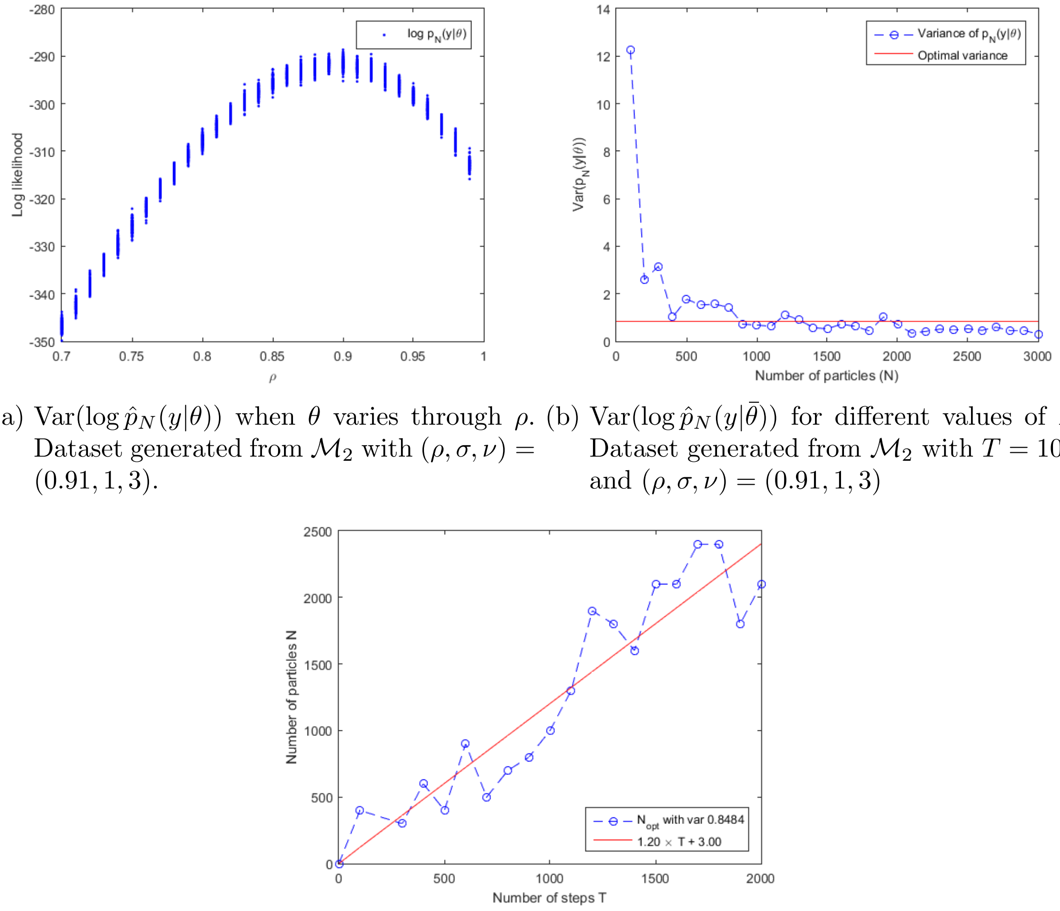
From a mathematical point of view, the stratified resampling is pertinent. A benchmark study consisting in resampling 1000 weights a large number of times was performed and the results are promising according to Table 4.1. The stratified is among the algorithms which performs the best in terms of speed. Coupled with a lower variance when resampling, it is the most compelling method to use inside the particle filters.

Resampling method	Elapsed Time (average)
Residual	18.90 s
Stratified	0.62 s
Systematic	0.63 s
Multinomial	1.87 s

Table 4.1: Time spent to resample 10^5 times 1000 weights

4.5 Tuning the number of particles

A critical issue in practice is the choice of the number of particles N. A large N gives a more accurate estimate of the log likelihood at greater computational cost, while a small N would lead to a very large estimator variance. 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. Figure 4.1a gives a hint that the variance is not likely to oscillate in big proportions when the model parameters θ changes. Pitt et al. (2012) suggests that a reasonable strategy is to estimate the posterior mean $\bar{\theta} = E[\theta | y_{0:T}]$ from an initial short run of the PMCMC scheme with N set to a large value. The value of N could then be adjusted such that the variance of the log-likelihood $\mathrm{Var}(\log p_N(y|\bar{\theta}))$ 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.



(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) Behaviour of N_{opt} when T varies

Figure 4.1

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. It is much more relevant to derive a general rule on how to choose N optimal, provided that such a rule exists. A test is conducted on an artificial dataset where the true value $\theta_{tr} = \bar{\theta}$ is known. It is composed of $T = 1000$ daily returns, generated from model $\mathcal{M}_2$. For a given value of N, the bootstrap filter of $\mathcal{M}_2$ is called several times and the variance of the log likelihoods $\mathrm{Var}(\log p_N(y|\bar{\theta}))$ 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 performed.

Both the values $b \simeq 0$ and $a \simeq 1$ suggest that the rule $T = N$ seems to hold.

4.6 Particle marginal Metropolis-Hastings Algorithm

Before explaining in details how the Particle marginal Metropolis-Hastings Algorithm (PMMH) works, a more general context is presented. The Metropolis Hastings MCMC scheme is used to target $p(\theta|y) \propto p(y|\theta)p(\theta)$ with the ratio:

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)} \times \frac{p(y|\theta^\star)}{p(y|\theta)}\right) \tag{4.14}$$

where $q(\theta^\star|\theta)$ is the proposal density. As discussed before, in hidden Markov models, the marginal likelihood $p(y|\theta) = \int_{\mathbb{R}^T} p(y|x)p(x|\theta)dx$ is often intractable and the ratio becomes impossible to compute. The simple likelihood-free scheme targets the full joint posterior $p(\theta, x|y)$. Usually the knowledge of the kernel $K(x_t|x_{t-1})$ makes $p(x|\theta)$ tractable. For instance, for a path $x_{0:T}$ governed by a linear Gaussian process $x_t = \rho x_{t-1} + \tau\epsilon_t$, $\epsilon_t \sim \mathcal{N}(0,1)$ can be simulated as long as ρ, τ and x_0 are known quantities. The MH is built in two stages. First, a new θ^* is proposed from $q(\theta^\star|\theta)$. Then, x^* is sampled from $p(x^\star|\theta^\star)$. The generated pair $(\theta^\star, x^\star)$ is accepted with the ratio

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \times \frac{q(\theta|\theta^\star)}{q(\theta^\star|\theta)} \times \frac{p(y|x^\star, \theta^\star)}{p(y|x, \theta)}\right) \tag{4.15}$$

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

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \frac{p(x^\star|\theta^\star)}{p(x|\theta)} \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)} \frac{p(y|x^\star, \theta^\star)}{p(y|x, \theta)} \frac{p(x|y, \theta)}{p(x^\star|y, \theta^\star)}\right) \tag{4.16}$$

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

$$\min\left(1, \frac{p(\theta^\star)}{p(\theta)} \frac{p(y|\theta^\star)}{p(y|\theta)} \frac{f(\theta|\theta^\star)}{f(\theta^\star|\theta)}\right) \tag{4.17}$$

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). Remarkably x is no more present and the ratio is exactly the same as the marginal scheme shown before. 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 (Algorithm 2) is an MCMC algorithm for state space models jointly updating θ and $x_{0:T}$. First, a proposed new $\theta^\star$ is generated from a proposal $f(\theta^\star|\theta)$, and then a corresponding $x^\star_{0:T}$ is generated by running a bootstrap particle filter using the proposed new model parameters, $\theta^\star$, and selecting a single trajectory by sampling once from the final set of particles using the final set of weights. This proposed pair $(\theta^\star, x^\star_{0:T})$ is accepted using the Metropolis-Hastings ratio

$$\frac{\hat{p}_{\theta^\star}(y_{1:T})p(\theta^\star)q(\theta|\theta^\star)}{\hat{p}_{\theta}(y_{1:T})p(\theta)q(\theta^\star|\theta)} \tag{4.18}$$

where $\hat{p}^N_{\theta^\star}(y_{1:T})$ is the particle filter's unbiased estimate of marginal likelihood.

Algorithm 2 Particle pseudo marginal Metropolis-Hastings Algorithm

```
procedure INPUT($y_{1:T}$, a proposal distribution $q(\cdot|\cdot)$, the number of particles $N$, the number of MCMC steps $M$)
    $\hat{p}^N_{\theta^{(0)}}(y_{1:T}), x^{*(0)}_{1:T}$ ← Call Bootstrap Particle Filter with $(y_{1:T}, \theta^{(0)}, N)$
    for i from 1 to M do
        Sample $\theta'$ from $q(\theta|\theta^{(i-1)})$
        $\hat{p}^N_{\theta'}(y_{1:T}), x^{*'}_{1:T}$ ← 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 $x^{(i)*}_{1:T} \leftarrow x'^{*}_{1:T}, \theta^{(i-1)} \leftarrow \theta', \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta'}(y_{1:T})$
        Otherwise $x^{(i)*}_{1:T} \leftarrow x^{(i-1)*}_{1:T}, \theta^{(i-1)} \leftarrow \theta^{(i-1)}, \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta^{(i-1)}}(y_{1:T})$
    end

return $(x^{(i)*}_{1:T}, \theta^{(i)})^M_{i=1}$
```

5 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 ARCH model. The main objective was to fit volatility clustering and the fat tails of the return distributions. In this section, we introduce the standard stochastic volatilities in a first time and its different extensions. The first extension consists in replacing the gaussian errors with Student-t errors. In the second extension, we incorporate a leverage effect by modelling a correlation parameter between measurement and state errors. In the third extension, we implement a model to check that the measurement errors are serially independent. Finally, the last extensions incorporate two factors with and without leverage to model the volatility of the returns.

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

The standard discrete-time stochastic volatility model for the asset prices returns $(Y_t)_{t>0}$ is defined as:

$$X_t = \phi X_{t-1} + \sigma \epsilon_{X,t} \tag{5.1}$$

$$Y_t = \beta \exp\left(\frac{X_t}{2}\right) \epsilon_{Y,t} \tag{5.2}$$

where $(\epsilon_{X,t})_{t>0}, (\epsilon_{Y,t})_{t>0}$ are two independent and standard normally distributed processes. Let $\theta = (\rho, \sigma^2, \beta)$ be the parameters vector. This model is non-linear because of the non-additive noise of the transition kernel. $(X_t)_{t>0}$ governs the volatility process of the observed returns $(Y_t)_{t>0}$, σ is the volatility of the volatility, and ϕ the persistence 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 $(X_t)_{t>0}$. The next part explains the link between the stochastic volatility model and the Geometric Brownian Motion (GBM).

Definition 4. *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, and μ the drift and σ the volatility. Both are constants.*

The process can be discretized by

$$S_{t+1} - S_t = \mu S_t + \sigma S_t \epsilon_{t+1},\ \epsilon_t \sim \mathcal{N}(0,1)$$
$$S_{t+1} = S_t + \mu S_t + \sigma S_t \epsilon_{t+1}$$

$$S_t = S_{t-1} + \mu S_{t-1} + \sigma S_{t-1}\epsilon_t \tag{5.3}$$

In the Stochastic Volatility model (SV), $(Y_t)_{t>0}$ represents the returns of the modelled asset. A general definition for computing the returns is $y_t = \frac{S_t}{S_{t-1}} - 1$, where $(S_t)_{t>0}$ is the asset observed prices. When x_t is measured at time t^- with regard to the filtration $\mathcal{F}_t$, $(Y_t|X_t = x_t)_{t>0}$ is normally distributed as

$$\begin{aligned} Y_t|X_t = x_t, \theta &\sim \mathcal{N}(0, \beta^2 \exp(x_t)) \\ S_t|X_t = x_t, \theta &\sim \mathcal{N}(S_{t-1}, \underbrace{S_{t-1}^2 \beta^2 \exp(x_t)}_{\sigma^2(t)}) \end{aligned} \tag{5.4}$$

The variance $\sigma^2(t)$ always exists as a product of square and exponential terms. Finally, $S_t = S_{t-1} + \sigma(t) S_{t-1}\epsilon_R$, $\epsilon_R \sim \mathcal{N}(0,1)$ 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 capability of modelling this volatility.

5.2 Model $\mathcal{M}_2$ - Stochastic Volatility Student-t (SVT)

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

Theorem 5. *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*

$$f_Y(y) = \left| \frac{d}{dy}(g^{-1}(y)) \right| \cdot f_X(g^{-1}(y)) \tag{5.5}$$

Applying this theorem 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,

$$p(y_t|X_t = x_t, \theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \frac{1}{\sigma_t} \left(1 + \frac{y_t^2}{\sigma_t^2 \nu}\right)^{-\left(\frac{v+1}{2}\right)} \tag{5.6}$$

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

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left[\frac{\nu + \left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right]^{-\left(\frac{\nu+1}{2}\right)} \tag{5.7}$$

The reasoning to find a closed form of $(S_t|x_t)$ 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,

$$S_t = \underbrace{S_{t-1}}_{\mu(t)} + \underbrace{\beta S_{t-1} \exp\left(\frac{x_t}{2}\right)}_{\sigma(t)} \epsilon_{Y,t} \tag{5.8}$$

As a conclusion, $(S_t|x_t, \theta)_{t>0}$ follows a t location-scale distribution of parameters $(\mu(t), \sigma(t), \nu)$.

5.3 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 $(\epsilon_{X,t})_{t>0}$ and $(\epsilon_{Y,t})_{t>0}$. θ is enriched with the new parameter ρ.

Theorem 6. *[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 $\epsilon_{Y,t} = \rho\epsilon_{X,t} + \sqrt{1-\rho^2}Z$.

$$\begin{aligned}
Y_t &= \beta \exp\left(\frac{x_t}{2}\right) \cdot \left(\rho\epsilon_{X,t} + \sqrt{1-\rho^2}Z\right) \\
Y_t &= \beta \exp\left(\frac{x_t}{2}\right) \rho\epsilon_{X,t} + \beta \exp\left(\frac{x_t}{2}\right)\sqrt{1-\rho^2}Z \\
\frac{S_t}{S_{t-1}} - 1|X_t = x_t &\sim \mathcal{N}\left(\beta\rho \exp\left(\frac{x_t}{2}\right)\epsilon_{X,t}, \beta^2 \exp(x_t) \cdot (1-\rho^2)\right) \\
\frac{S_t}{S_{t-1}}|X_t = x_t &\sim \mathcal{N}\left(1 + \rho\beta \exp\left(\frac{x_t}{2}\right)\epsilon_{X,t}, \beta^2 \exp(x_t) \cdot (1-\rho^2)\right) \\
S_t|X_t = x_t &\sim \mathcal{N}\left(S_{t-1} + S_{t-1}\rho\beta \exp\left(\frac{x_t}{2}\right)\epsilon_{X,t}, S_{t-1}^2\beta^2 \exp(x_t) \cdot (1-\rho^2)\right) \qquad (5.9)
\end{aligned}$$

The differences between the normal leverage model and the standard model where $\rho = 0$, are the correcting drift term $S_{t-1}\rho\beta \exp\left(\frac{x_t}{2}\right)\epsilon_{X,t}$ and the factor $(1-\rho^2) \leq 1$ reducing the volatility.

5.4 Model $\mathcal{M}_4$ SV-MA(1) - Moving Average

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 more simple specification and set $m = 1$. Hence, our model becomes

$$\begin{aligned}
X_t &= \phi X_{t-1} + \sigma\epsilon_{X,t} && (5.10) \\
Y_t &= \beta \exp\left(\frac{X_t}{2}\right)\epsilon_{Y,t} + \psi\beta \exp\left(\frac{X_{t-1}}{2}\right)\epsilon_{Y,t-1} && (5.11) \\
Y_t|\mathcal{F}_{t^-} &\sim \mathcal{N}\left(0, \beta^2 \exp(x_t) + \psi^2\beta^2 \exp(x_{t-1})\right) \\
S_t|\mathcal{F}_{t^-} &\sim \mathcal{N}\left(S_{t-1}, S_{t-1}^2\beta^2 \exp(x_t) + S_{t-1}^2\psi^2\beta^2 \exp(x_{t-1})\right) && (5.12)
\end{aligned}$$

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 $\text{Var}\left(Y_t|\mathcal{F}_{t^-},\theta\right) = \beta^2 e^{x_t} + \beta^2\psi^2 e^{x_{t-1}}$. 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 $(X_t)_{t>0}$.

5.5 Model $\mathcal{M}_5$ Stochastic Mean

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

$$Y_t = \beta \exp\left(\frac{X_t}{2}\right) + \exp\left(\frac{X_t}{2}\right)\epsilon_{Y,t} \tag{5.13}$$

$$S_t|\mathcal{F}_{t^-} \sim \mathcal{N}\left(S_{t-1} + S_{t-1}\beta\exp\left(\frac{x_t}{2}\right), S_{t-1}^2\exp\left(x_t\right)\right) \tag{5.14}$$

where $(X_t)_{t>0}$ corresponds to the process of a standard stochastic volatility model defined in Equation (5.1). This model is pertinent if we believe that the conditional mean is somehow proportional to the conditional volatility. This can be the case in financial data, where high volatility appears in clusters where the absolute conditional mean is high.

5.6 Model $\mathcal{M}_6$ Two Factors Stochastic Volatility

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 returns volatility. 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. It is defined as

$$X_t = \phi_X X_{t-1} + \sigma_X\epsilon_{X,t} \qquad |\phi_X| < 1, \epsilon_{X,t} \sim \mathcal{N}(0,1), X_0 \sim \mathcal{N}\left(0, \frac{\sigma_X^2}{1-\phi_X^2}\right) \tag{5.15}$$

$$Z_t = \phi_Z Z_{t-1} + \sigma_Z\epsilon_{Z,t} \qquad |\phi_Z| < 1, \epsilon_{Z,t} \sim \mathcal{N}(0,1), Z_0 \sim \mathcal{N}\left(0, \frac{\sigma_Z^2}{1-\phi_Z^2}\right) \tag{5.16}$$

$$Y_t = \beta\exp\left(\frac{X_t + Z_t}{2}\right)\epsilon_{Y,t} \qquad \epsilon_{Y,t} \sim \mathcal{N}(0,1) \tag{5.17}$$

$S_t|\theta, X_t = x_t, Z_t = z_t \sim \mathcal{N}\left(S_{t-1}, S_{t-1}^2\beta^2\exp\left(x_t + z_t\right)\right)$. The parameters vector θ is now defined as $\theta = (\beta, \phi_X, \phi_Z, \sigma_X, \sigma_Z)$ where β is a scaling term. It is of common knownledge 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. Also, 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.

Estimating these parameters using Particle Markov Chain Monte Carlo is fairly straightforward. The particle filter must be updated such that two sets of particles (one for X_t and one for Z_t) must be drawn instead of one. Because of the symmetry between X_t and Z_t in Y_t, some conditions on the parameters have to be set to ensure the convergence, such that $\phi_X > \phi_Z$.

5.7 Model $\mathcal{M}_7$ Two Factors Stochastic Volatility with Leverage

One final extension considers the two factors stochastic volatility and assume that the correlation $\rho = cor(\epsilon_{X,t}, \epsilon_{Y,t})$ is statistically different from zero. The idea is the same as the one developped for the model $\mathcal{M}_3$. Recall that $(X_t)_{t>0}$ is the long-run factor from Equation (5.15) which corresponds to the stochastic trend of the returns volatility. From the models presented before, this model is by far the most complex because 6 parameters are to be estimated: $\theta = (\beta, \rho, \phi_X, \phi_Z, \sigma_X, \sigma_Z)$. Ruiz and Veiga (2008) studied a slightly different version where the $(X_t)_{t>0}$ is a fractional integrated Gaussian noise process but the overall behaviour remains the same. They proved that the first order autocorrelation $cor(|y_t|, |y_{t+1}|)$ is smaller than the second order autocorrelation $cor(y_t^2, y_{t+1}^2)$ when $\rho < 0$. As explained by Cont (2005), it is usually the case in practical applications. If $\rho = 0$, there is no more asymmetry in the model. Still with the Cholesky decomposition and with the same methodology presented for model $\mathcal{M}_3$, Y_t can be expressed as

$$Y_t|x_t, z_t, \theta \sim \mathcal{N}\left(\rho\beta \exp\left(\frac{x_t + z_t}{2}\right)\epsilon_{X,t}, \beta^2 \exp(x_t + z_t)(1 - \rho^2)\right) \tag{5.18}$$

$$S_t|x_t, z_t, S_{t-1}, \theta \sim \mathcal{N}\left(S_{t-1} + S_{t-1}\rho\beta \exp\left(\frac{x_t + z_t}{2}\right)\epsilon_{X,t}, S_{t-1}^2\beta^2 \exp(x_t + z_t)(1 - \rho^2)\right) \tag{5.19}$$

6 Validation, Estimation and Selection of the best SV model

6.1 Validation of the models

In practical applications, the true value of θ_{tr} is usually unknown and it makes the validation harder. The validation is an important pre-task because it tests the implementation, the choice of the priors and the proposal distributions, measures the dispersion of the estimator $\hat{\theta}$ to the true value θ_{tr}. The first step involves the sample generation of both the process and the observations $(X_t, Y_t)_{t>0}$. We choose an arbitrary value for θ_{tr}. From this point, $x^*_{1:T}$ and $y^*_{1:T}$ are sampled. Each model takes $(y^*_t)_{1:T}$ as argument and outputs an estimator $(x_{\hat{1:T}}, \hat{\theta})$. The estimated values are then compared to the true values using some 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. In this example, $x^*_{1:T}$ and $y^*_{1:T}$ have been sampled from the model $\mathcal{M}_x$. 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.

6.2 Parametrisation and estimation

Once the model has been validated, it can be fitted to sample data. The number of steps required in Particle MCMC is taken large enough to ensure that enough samples are available for analysis. Unless stated otherwise, the PMCMC scheme algorithm will loop 10000 times before stopping. The first 1000 samples are discarded for each parameter. This is because the chains require several steps to reach their equilibrium distribution. A component-wise scheme is used to update the parameters, one by one sequentially. Note that it is possible to parallel this scheme it 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 appendix. Because the bias has not been rigourously, a simple version with no parallel was used for the computations. Once the burn-in phase was performed, the mean value $\bar{\theta}$ is selected from the distribution $\mathcal{D}(\theta)$ as the best estimation for θ_{tr}. Some statistics, moments and confidence intervals can be obtained from $\mathcal{D}(\theta)$.

6.3 Model Comparison

The output of the particle filter is an unbiased estimate of $p(y|\theta)$, with the unobserved states integrated out. However this estimation is not sufficient to compare two 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

$$p(Y_{1:T}|\mathcal{M}) = \int_{\theta} p(Y_{1:T}|\theta, \mathcal{M})p(\theta|\mathcal{M})d\theta \tag{6.1}$$

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

$$\left(\frac{1}{N}\sum_{i=1}^{N}\frac{g(\theta_i)}{p(Y_{1:T}|\theta_i)p(\theta_i)}\right)^{-1} \to p(Y_{1:T}) \text{ as } N \to \infty \tag{6.2}$$

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 $N(\theta^*, \Sigma^*)$ can be used, where θ^* and Σ^* are equal to the empirical mean and sample unbiased variance, $\theta^* = \frac{1}{N}\sum_{i=1}^{N}\theta^i$ and $\Sigma^* = \frac{1}{N-1}\sum_{i=1}^{N}\left(\theta^i - \theta^*\right)\left(\theta^i - \theta^*\right)^T$.

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 $N = 3$ and assume that the log-terms on the LHS are equal to -120, -121 and -122 :

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

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.

6.3.1 Case study with a stock

A practical case is considered both on a stock and a spread to see if the results are in accordance. The stock at hand is Apple (APPL) and the period is Sep, 09 2003 - Jun, 04 2006 (Figure 6.1). The daily returns are computed according to $Y_t = S_t/S_{t-1} - 1$ and are given as input to the stochastic volatility models. Table **??** reports estimation of θ for the stochastic volatility models $(\mathcal{M}_1, ..., \mathcal{M}_6)$. $\log(L)$ is the log marginal likelihood $p_N(y|\bar{\theta}, \mathcal{M})$. We find that the Gaussian two factor SV model performs best in terms of the marginal likelihood and AIC criteria. The Kass factor $2\log BF$ of SVTFL $\mathcal{M}_7$

versus SVTF $\mathcal{M}_6$ is 10.8 which indicates very strong evidence in favour of the SVTF model and its leverage ρ. Compared to the SV with leverage $\mathcal{M}_3$ with one factor, the Kass factor in favour of SVL is 23.3 which is very 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 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 much smaller whereas ϕ_Z is almost 0 and σ_Z is higher. It seems clear now that the volatility of the returns can be decomposed into two distinct processes: a long-run stochastic trend $(X_t)_{t>0}$ and a process $(Z_t)_{t>0}$ accounting for short-run dynamics.

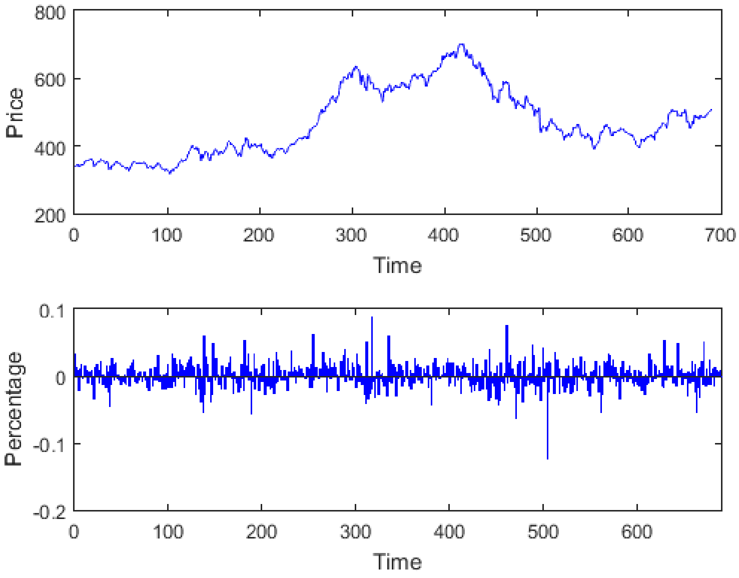


Figure 6.1: APPL stock. Period is from 09-Sep-2003 to 04-Jun-2006

Parameter	$\bar{\theta}_{\mathcal{M}1}$	$\bar{\theta}_{\mathcal{M}2}$	$\bar{\theta}_{\mathcal{M}3}$	$\bar{\theta}_{\mathcal{M}4}$	$\bar{\theta}_{\mathcal{M}5}$	$\bar{\theta}_{\mathcal{M}6}$
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	0.1
ν		7.6850				
ρ			-0.4397			
ψ				0.0060		
ϕ_X						0.9978
ϕ_Z						0.1443
σ_X						0.1162
σ_Z						0.6398
μ						0
$\log(L)$	2646.3	2659.7	2660.9	2649.2	2649.3	2663.6
AIC	-5286.6	-5311.4	-5313.8	-5290.4	-5292.6	-5319.2
$2\log\mathcal{BF}(\cdot,\mathcal{M}6)$	33.6	6.9	4.5	31.5	26.1	0
N	1000	1000	1000	1000		1000
T	1000	1000	1000	1000		1000
Steps	10000	10000	10000	10000		10000
Burn-in	1000	1000	1000	1000		1000

Table 6.1: Estimation of the parameters for the SVM model. Data is APPL.

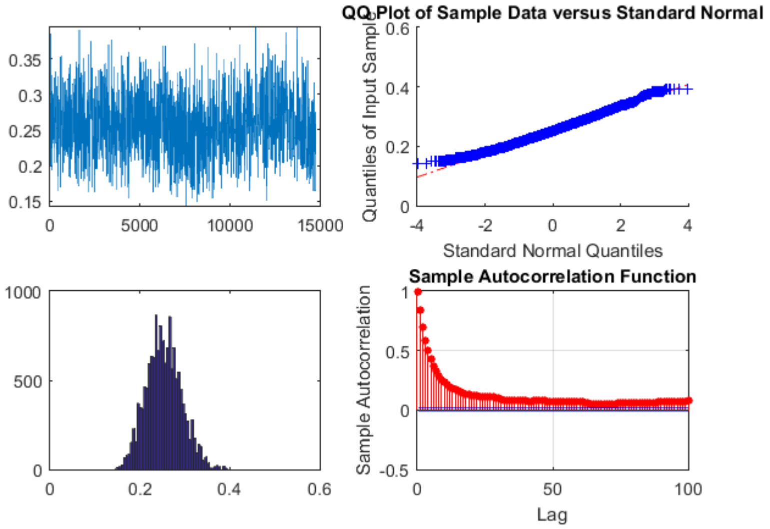


Figure 6.2: Estimation of σ with the SVM model $\mathcal{M}_5$

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	0.08	0.14	0.11

Table 6.2: Estimation of the parameters for the SV model $\bar{\theta}_{\mathcal{M}5}$. Data is APPL.

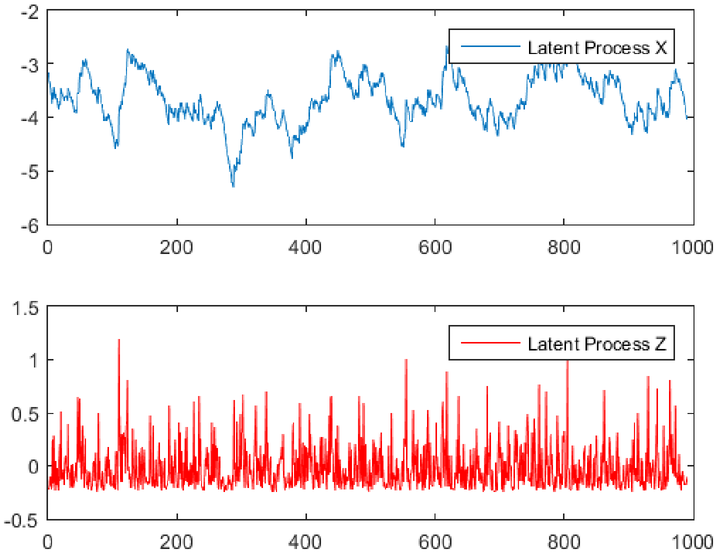


Figure 6.3: Estimation of the latent processes X and Z in the Two Factors SV model

6.3.2 Case study with a spread

The same procedure was conducted for a Spread, composed of three stocks: AMR CORP, CRANE CO and DOVER CORP with associated cointegrating vector $\beta = (1, -0.0865, -0.3796)$. Period is from 09-Sep-2003 to 04-Jun-2006. Table **??** reports estimation of θ for the stochastic volatility models $(\mathcal{M}_1, ..., \mathcal{M}_7)$. We find that the Gaussian two factor SVL model performs best in terms of the marginal likelihood and AIC criteria. The Kass factor $2\log BF$ of SVTFL $\mathcal{M}_7$ versus SVTF $\mathcal{M}_6$ is 10.8 which indicates very strong evidence in favour of the SVTF model and its leverage ρ. Compared to the SV with leverage $\mathcal{M}_3$ with one factor, the Kass factor in favour of SVL is 23.3 which is very 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 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 much smaller whereas ϕ_Z is almost 0 and σ_Z is higher. It seems clear now that the volatility of the returns can be decomposed into two distinct processes: a long-run stochastic trend $(X_t)_{t>0}$ and a process $(Z_t)_{t>0}$ accounting for short-run dynamics.

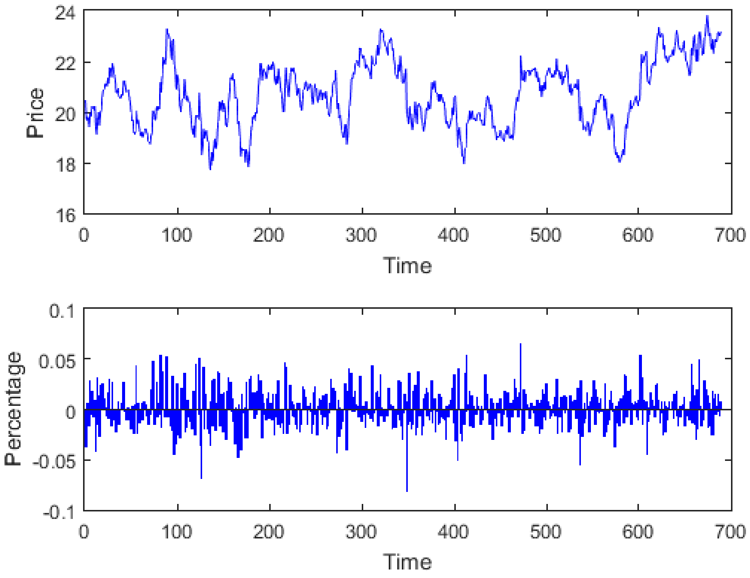


Figure 6.4: Spread AMR CORP - CRANE CO - DOVER CORP. $\beta = (1, -0.0865, -0.3796)$. Period is from 09-Sep-2003 to 04-Jun-2006

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				0.0060			
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
N	1000	1000	1000	1000	1000	1000	1000
T	689	689	689	689	689	689	689
Steps	10000	10000	10000	10000	10000	10000	10000
Burn-in	1000	1000	1000	1000	1000	1000	1000

Table 6.3: Estimation of the parameters for the SVM model. Data is Spr AMR CORP - CRANE CO - DOVER CORP.

6.4 Volatility Modelling

Once the best model has been selected, validated and its parameters estimated, the volatility of the asset can be approximated. For a given process $(\mathbf{X}_t)_{t>0} \in \mathbb{R}^N$ on a state-space, the returns $(Y_t)_{t>0}$ modelled by a SV model, are usually of the form $y_t|\mathbf{x}_t, \theta \sim \mathcal{D}(\mu(t), \sigma^2(t))$. By definition, $Y_t = S_t/S_{t-1} - 1$. We then have $S_t|S_{t-1}, \mathbf{x}_t, \theta \sim \mathcal{D}(S_{t-1}\mu(t) + S_{t-1}, S_{t-1}^2\sigma^2(t))$.

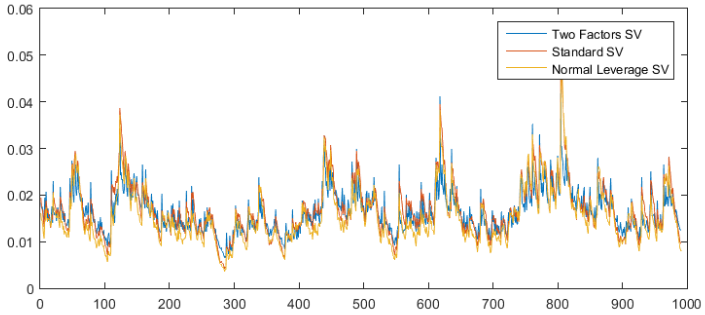


Figure 6.5: Volatility associated to the returns y_t for model $\mathcal{M}_1$, $\mathcal{M}_3$ and $\mathcal{M}_6$

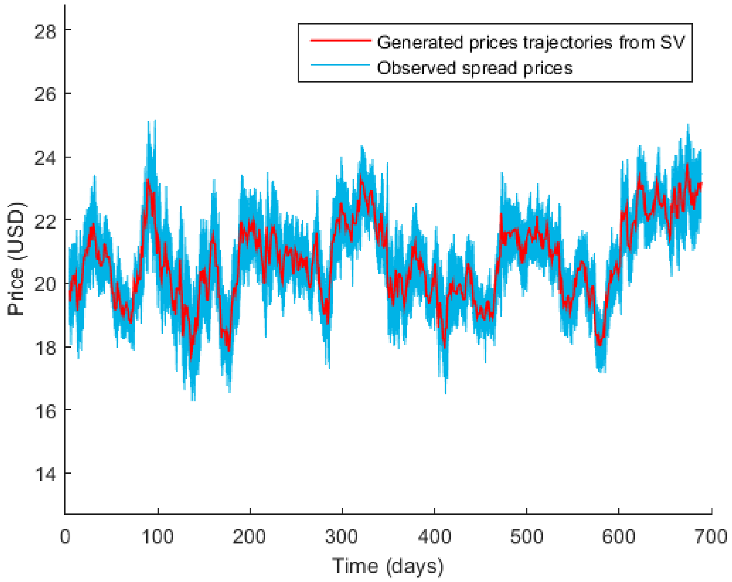


Figure 6.6: Generation of $M = 1000$ MC prices trajectories with model $\mathcal{M}_7$

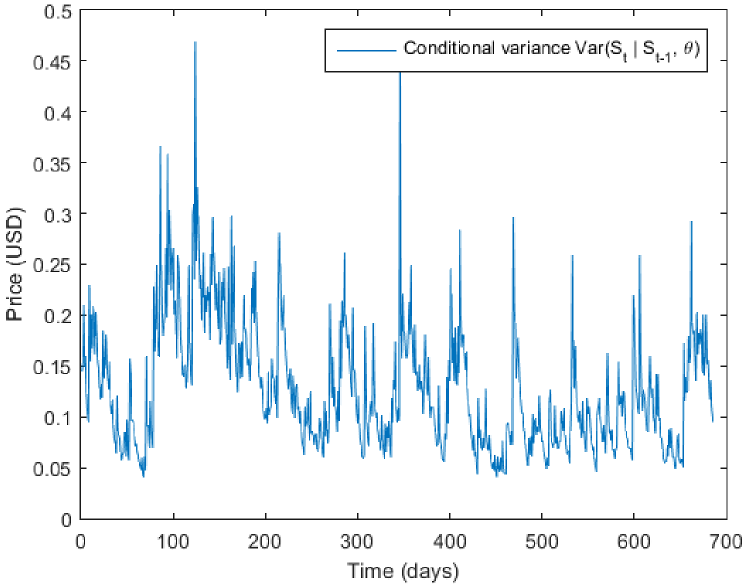


Figure 6.7: Conditional variance of $S_t|S_{t-1}, x_t, \theta$ with model $\mathcal{M}_7$

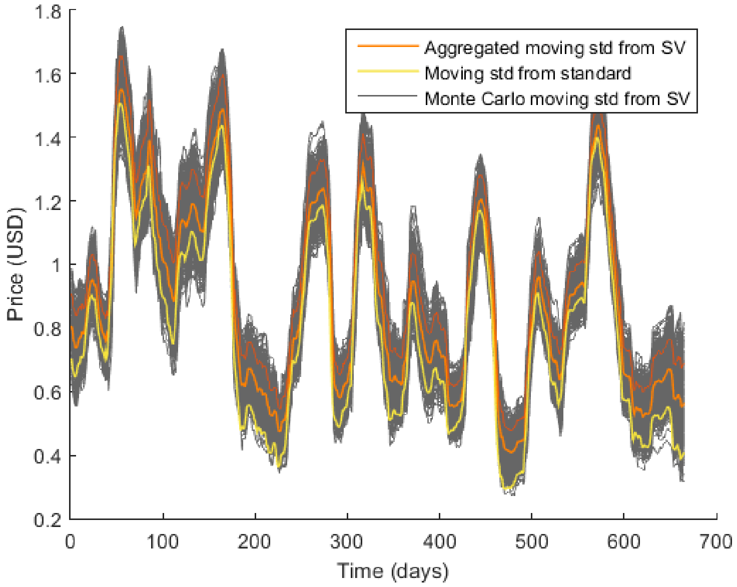


Figure 6.8: Rolling volatility processes for the generated prices trajectories and the spread prices

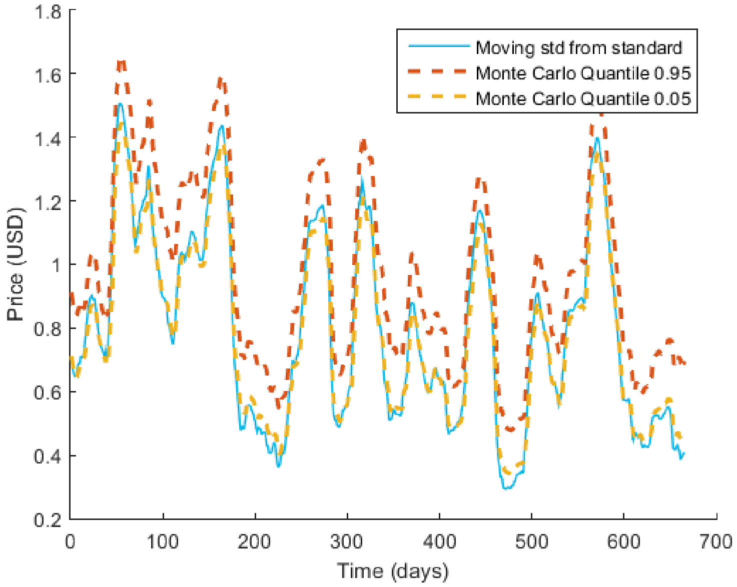


Figure 6.9: 90 % Confidence intervals for the generated prices trajectories

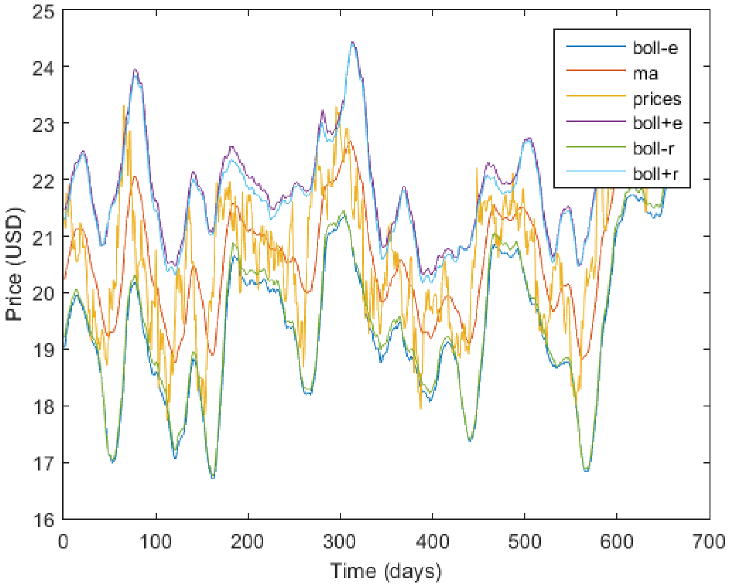


Figure 6.10: Bollinger Bands

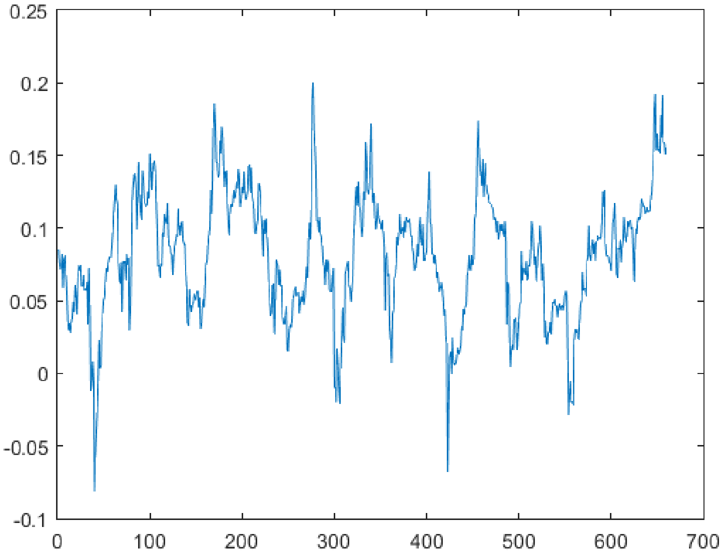


Figure 6.11: Normalized difference $d(t) = (r\sigma(t)_E - r\sigma(t)_C)/S_t$between the evaluated $r\sigma(t)_E$ and common $r\sigma(t)_C$ windowed volatilities used in the Bollinger Bands

7 Statistical Arbitrage

7.1 Introduction

Statistical arbitrage conjectures statistical mis-pricings or price relationships that are true in expectation, in the long run when repeating a trading strategy. Statistical arbitrage is a heavily quantitative and computational approach to equity trading. 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, the other is sold short. This hedges risk from whole-market movements. This 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 the same value in local currency. The difference between this long and short position is known as the spread. Once the spread deviates far from its long-run equilibrium, a position is opened and is unwind when the spread reverts. Dealing with spreads instead of non-stationarity stocks is beneficial because stationary series are on average much more reverting. This approach has the advantage of eliminating the market exposure (memo corr cumsum with SP500 should be around 0).

7.2 Presentation of the dataset

The sample period used starts in January 1990 and ends in March 2014 summing up to 8844 observations. Daily equity closing prices obtained from Bloomberg. The analysis covers all stocks in the SP500 index from the American stock markets. The proposed statistical arbitrage generated average excess returns of 12% per year in out-of-samples simulations, Sharpe ratio of 1.70, 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.

7.3 Composition of the portfolio

The first motivation of considering a portfolio approach is to lower the volatility associated to each tuple trading by smoothing the net value over time. The approach consists in selection the tuples for trading based on the best in-sample Sharpe Ratios. We form the portfolio of 20 best trading pairs that present the greatest SR in the in-sample simulations and use them to compose a pairs trading portfolio to be employed out-of-sample. Once a trade is initiated, the portfolio is not rebalanced. Only two types of transactions are considered: move into a new position, or the total unwind of a previously opened position. Any opened position is closed at the end of the study.

7.4 Strategies

7.4.1 Bollinger Bands

Bollinger Bands is a widely used technical volatility indicator which consists in placing volatility bands $\{Boll_t^+, Boll_t^-\}$ above and below the moving average prices $\{m_t\}$. Volatility is based on the standard deviation, which changes as volatility increases and decreases. The bands automatically widen when volatility increases and narrow when volatility decreases. They are calculated by

$$m(t) = \frac{1}{n}\sum_{j=1}^{n} S_j \text{ (SMA)}$$

$$Boll^{\pm}(t) = m(t) \pm \alpha\sqrt{\frac{1}{n}\sum_{j=1}^{n}(S_j - m(t))}$$

where $(S_t)_{t\geq 0}$ is the price of the asset, n is the number of time periods in the moving average and α is the number of standard deviations to shift the Bollinger bands. The default values are $n = 20$ and $\alpha = 2$. $m(t)$ is called the mid band and is used as a relative mean value. $Boll^+(t)$ and $Boll^-(t)$ are respectively the upper and lower bands. Their purpose is to measure how far the price deviates from its mean. Under the assumption that the returns are normally distributed, 95% of the prices should appear within the bands when $\alpha = 2$. The simple moving averages used in the computation of the bands can be replaced by exponential moving averages which gives more weights to new values and may increase the accuracy.

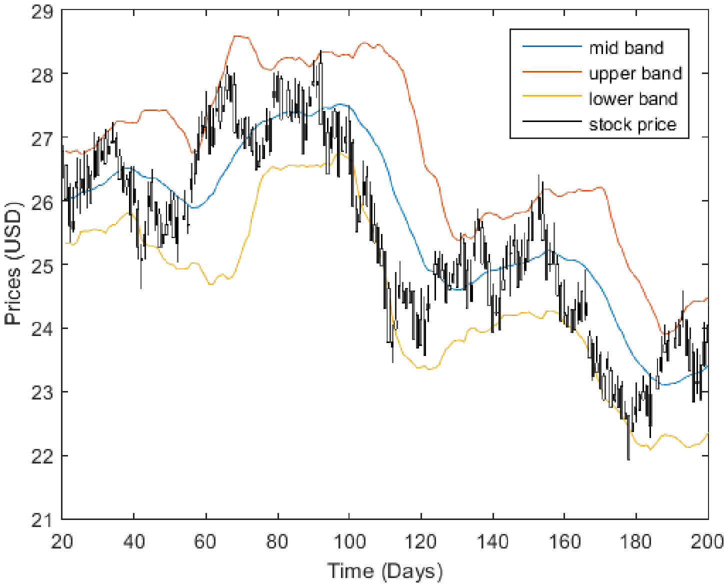


Figure 7.1: Example of Bollinger bands strategy applied to Walt Disney Co NYSE (2002). The default values are $n = 20$ and $\alpha = 2$.

The SV model is firstly calibrated on the returns y_t of the spread. It is computed as $y_t = \frac{S_t}{S_{t-1}} - 1$ where S_t is the spread process. x_t represents the volatility associated to y_t. The log-returns are not used here for a specific reason discussed later. Let's consider the general case where S_t is of the form

$$S_t|\mathcal{F}_{t^-} \sim \mathcal{D}\left(\mu(t), \sigma^2(t)\right)$$

where $\mathcal{F}_{t^-}$ contains all the information up to time t^-, i.e. the volatility $\sigma(t)$ and the mean $\mu(t)$ are known because the process x_t and S_{t-1} were measured. $\mathcal{D}$ can represent any suitable distribution such as the normal distribution or the t-student distribution. 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. A spread is a linear combination of assets where each asset price is one observation of a more general process, over a time interval. From this idea, two main approaches are discussed.

The first approach consists in working directly with the returns. This contrasts with the default Bollinger bands strategy where the price series is used to compute the volatility. The idea is to detect large movements in returns and bet for a reversion to the long-range mean. This strategy employs a large quantity of trades because it is highly sensitive to price movements. A simple moving average of lag p on the volatility $\sigma(t)$ is considered to detect such movements:

$$SMA_\sigma(t,p) = \frac{1}{p}\sum_{i=0}^{p-1}\sigma(t-i) \tag{7.1}$$

The mid band is generated from a moving average over the returns. The upper and lower bands are generated by adding and subtracting this rolling volatility SMA_σ to the mid band.

The second method is based on drawing a large of number of sample paths from the model to estimate the volatility of the spread. The volatility computed in this approach is of the same shape as the one computed in the default Bollinger bands strategy. Once every volatility was computed, each of them must be aggregated to form the final rolling volatility $(r\sigma(t)_{t>0}$ from the different generated paths. This quantity is the one used in the computation of the upper and lower bands. The way the rolling volatility is computed and its associated lag are omitted here, for a better clarity. Algorithm 3 summarizes the procedure for the standard stochastic volatility whose equation is given below as a recall

$$S_t|X_t = x_t, S_{t-1}, \theta \sim \mathcal{N}(\mu(t) = S_{t-1}, \sigma^2(t) = S_{t-1}^2\beta^2 \exp(x_t)) \tag{7.2}$$

Algorithm 3 Rolling volatility computed with the standard stochastic volatility model

```
procedure INPUT($(x_t)_{t>0}$, $(S_t)_{t>0}$, $\theta = \beta$, $N$, $f_a = n^{-1}\sum_{i=1}^{N} \cdot$)
    for t from 1 to T do
        for i from 1 to N do
            Sample the $t^{th}$ value of the $i^{th}$ path, $S_{ti} \sim \mathcal{N}(S_{t-1}, S_{t-1}^2\beta^2 \exp(x_t))$
        end
    end
    for i from 1 to N do
        Compute the default rolling volatility $(r\sigma_i(t))_{t>0}$ for the $i^{th}$ path, $(S_{ti})_{t>0}$
    end
    for t from 1 to T do
        $r\sigma(t) = n^{-1}\sum_{i=1}^{N} r\sigma_i(t)$
    end
return $(r\sigma(t))_{t>0}$
```

In the most general case, N paths $\{S_{t,n}\}_{0<n\leq N, t\in\mathbb{N}}$ are generated from an equation involving $S_t|\mathcal{F}_{t^-}$. Let $f_a : \mathbb{R}^{+N} \to \mathbb{R}^+$ be a positive-definite aggregating function. The aggregated rolling volatility of lag p for all the N paths is defined as $r\sigma(t,p) = f_a(r\sigma(t,p)_1, ..., r\sigma(t,p)_N)$. If f_a is simply the sample mean estimator, the equation is simplified to $\sigma(t,p) = \frac{1}{N}\sum_{i=1}^{N} SMA_\sigma(t,p)_i$. It is a well known fact that this estimator is also unbiased and consistent. Depending on the context and on the cross validation phase, f_a can be any measurable function satisfying the conditions above, such as the median or the quantile function. Note that a rolling volatility is similar to a moving average on the volatility process. When $p \to 1$, $r\sigma(t,p)$ converges to $\sigma(t)$, the instant volatility associated to $S_t|\mathcal{F}_{t^-}$.

7.5 Z-score

Once the spread $(\epsilon_t)_{t\geq 0}$ is formed, Caldeira and Moura (2013) suggests to compute the dimensionless z-score. Defined as $z_t = \frac{\epsilon_t - \mu_\epsilon}{\sigma_\epsilon}$, it measures the distance to the long-term mean in units of long-term standard deviation. The basic rule is to open a position when the z-score hits the n-quantile of the standard normal distribution $\Phi^{-1}(q_n)$. According to the 68-95-99.7 rule, having a two standard deviation thresholds from above and below seems relevant. If the z-score hits the 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 has to be unwind. The same reasoning for the high threshold holds for short positions. Caldeira and Moura (2013) suggested the basic trading strategy signals

$$\begin{aligned}
\text{Open long position if } &\leq \Phi^{-1}(q_{OL}) = -2.00\\
\text{Open short position if } &\geq \Phi^{-1}(q_{OS}) = 2.00\\
\text{Close short position if } &\leq \Phi^{-1}(q_{CS}) = 0.75\\
\text{Close long position if } &\geq \Phi^{-1}(q_{CL}) = -0.50
\end{aligned}$$

Note that unlike the Bollinger Bands, the Z-score is highly sensitive to stochastic trends because the mean is supposed to be constant. For the strategy to work, the spread should not be divergent. This shows how important it is to choose the right alternative hypothesis of ARD instead of TS (Trend Stationary) in the unit root tests. In practical applications and according to the risk policy of the firm, a stop loss threshold is set to avoid any huge losses.

7.6 Dataset

The data used in this study consists of daily closing prices of the 1232 stocks from the US markets. All the stocks used are listed in stock exchanges such as Dow Jones or NASDAQ, which means they are among the most liquid stocks traded within the US markets. This characteristic is important for the strategies, since it greatly diminishes the slippage effect, reduces the transaction costs and permits to unwind any position without impacting the market too much. The data were obtained from Bloomberg, taken from the period of January 1990 to March 2014. The data are adjusted for dividends and splits, avoiding false trading signals generated by these events, as pointed out by Broussard and Vaihekoski (2012).

7.7 Procedure

A typical trading strategy is made of three parts: selection of the suitable tuples satisfying some criteria like cointegration, create trading signals based on define predefined investment decision rules and finally assess the performance of the strategy.

7.7.1 Tuples selection

It is common in pair trading and more generally in tuple trading to require that the tuples belong to the same sector, for example in Chan (2009) and Dunis et al. (2010). Other did not adopt this restriction, for example Caldeira and Moura (2013). It is harder but nevertheless possible to bypass this restriction at a greater computational cost when the number of assets grows. Several tricks are performed to diminish this combinatorial explosion; one is based on correlation. In the general case, cointegration usually implies correlation but correlation doesn't always imply cointegration. 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. This assertion holds because a correlation test can be performed much faster than a cointegration test (Table 7.1).

Test	Elapsed Time (average)
Correlation *corr*	0.33 ms
Correlation R^2 (fast)	0.57 ms
Johansen	19.08 ms
Aug. Dickey Fuller	2.33 ms
Phillips-Perron	3.04 ms

Table 7.1: Average time spent to test a bivariate time series $X_t = (x_{t1}, x_{t2})$

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 $c = (r_{x1y}, r_{x2y}, ..., r_{xNy})^T$ of correlation r_{xny} between the predictor variables x_n and the target variable y, and the correlation matrix R_{xx} of inter-correlations between predictor variables. It is given by $R^2 = c^T R_{xx}^{-1} c$ where R_{xx}^{-1} is the inverse of the matrix

$$R_{xx} = \begin{pmatrix} r_{x1x1} & r_{x1x2} & \cdots & r_{x1xn} \\ r_{x2x1} & \ddots & & \vdots \\ \vdots & & \ddots & \\ r_{xnx1} & \cdots & & r_{xnxn} \end{pmatrix} \tag{7.3}$$

One problem arises: the value of the coefficient depends on the ordering of the tuple. To provide convincing evidence of this fact, let's consider a simple example. A regression of y on x and z will in general have a different R that will a regression of z on x and y. Let z be uncorrelated with both x and y while x and y are linearly related to each other. A regression of z on y and x will yield a R of zero, while a regression of y on x and z will yield a strictly positive R. It means that the ordering inside a tuple has its importance, at least from a statistical point of view. This assertion is also true for all cointegrations tests, except for the Johansen test where the ordering does not matter. This notion of ordering is much less obvious from a pure financial point of view.

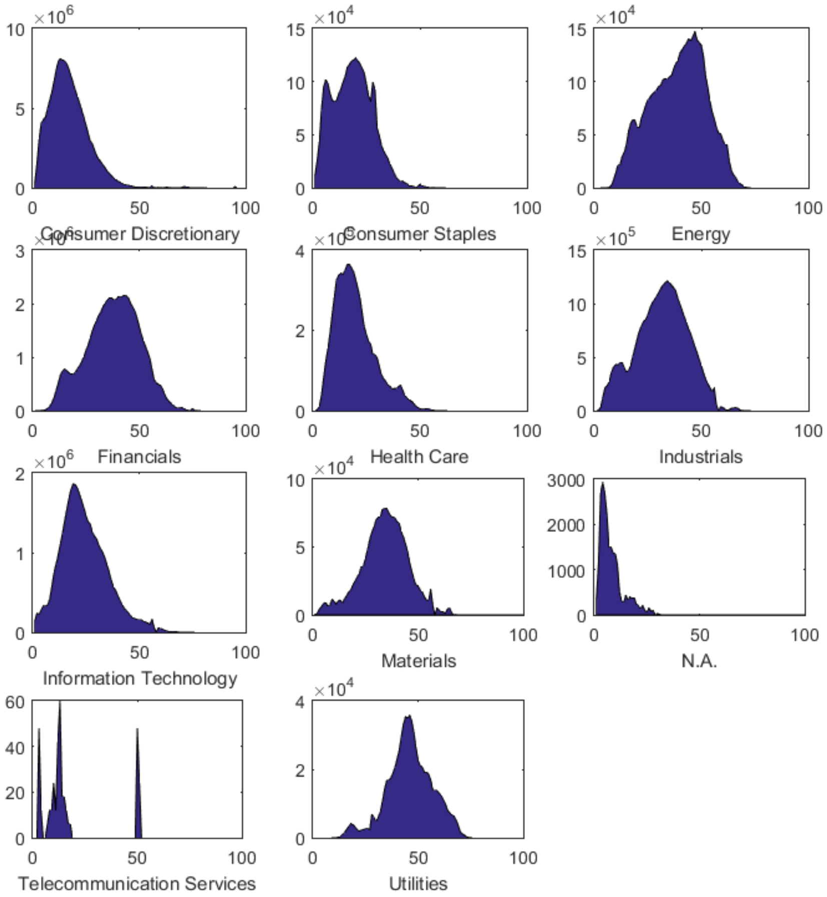


Figure 7.2: Distribution of $100 \times R^2$ for the quadruples (not all are cointegrated). Period is from Jan 01, 2012 to May 27, 2013

A rigorous testing of cointegration is performed to select the candidates. First, all the time series are tested for a unit root with an Augmented Dickey Fuller test. For a fixed n, all possible tuples are formed and R^2 is evaluated for each of them. A threshold R_{th} is arbitrary picked up (default is 0.80) and the tuples whose $R^2 > R_{th}$ are selected. For every selected tuple, form the spread $S_t = \beta_0 P_{t0} - \sum_{i=1}^{n-1} \beta_i P_{ti}$ and apply a triple Johansen, Dickey Fuller and Phillips-Perron test to check for cointegration. If the tuple is cointegrated i.e. all the three tests have positive outcome, form the spread and mark it as cointegrated.

7.7.2 Assumption of the same sector

Chan (2009) and Dunis et al. (2010) argued that the tuples should belong to the same sector, otherwise the cointegration and the correlation would be purely fortuitous. To check the veracity of this assumption, all the possible cointegrated triples are formed on the whole period of the dataset (from 01-Jan-1990 to 14-Mar-2014) and the R^2 is computed using the methodology exposed before. 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 all belong to different sectors. As a result, 30418 cointegrated triples were formed. 816 belonged to All, 10517 to Partial and the remaining 19085 to None. Figure 7.3 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 or more generally tuples is not very large (less than 500 or 1.5% here for the whole period), it is reasonable to consider the assumption of the same sector.

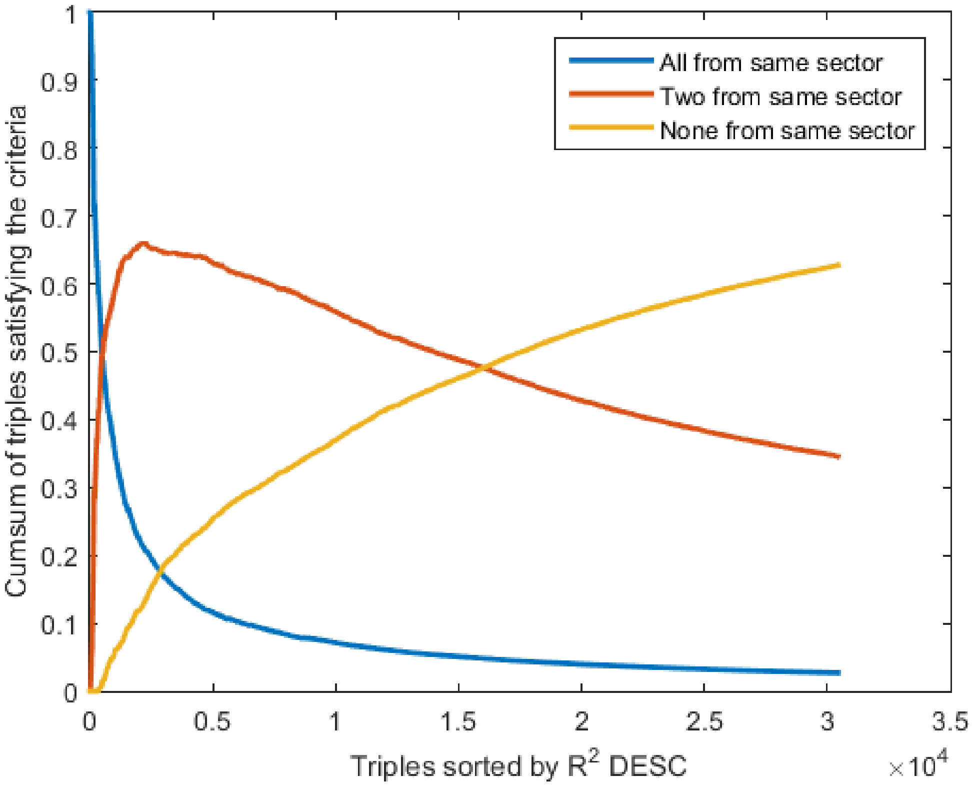


Figure 7.3: Repartition of the cointegrated triples sorted by R^2 from highest to lowest and regarding their belonging to sectors. Period is from 01-Jan-1990 to 14-Mar-2014

7.7.3 Creation of the Trading Signals

For cointegrated marked spreads, the second part of the algorithm creates trading signals based on predefined investment decision rules. A trading rule determines when to open and close a position. With Bollinger bands strategy, the basic rule is

- Open a long position when there is an upward crossing between the spread and the lower band;
- Unwind (close) this position when there is an upward crossing between the spread and the upper band;
- 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.

Empirical studies showed that this strategy is one of the most profitable with the use of Bollinger Bands. 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 cointegrated vector β. It is assumed that the trader can buy a portion of an asset. This same position is closed by selling one unit of the first asset and buying the remaining assets, still in the same proportions.

7.7.4 General parameters of the strategy

Throughout the strategies, we consider 0.5% of the total nominal as transaction costs for tuple trading. This choice was discussed 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, invested or in cash. The asset allocation in the portfolio follows a invested weighting scheme with no dynamic rebalancing. Each tuple is given the same weight throughout the study. If there are no open positions, the money is not invested and remain as cash in the portfolio. For a particular tuple, the number of open positions is limited to only one on the spread. The strategy is self-financing, i.e. profits are reinvested and no deposits or withdrawals are permitted.

7.8 Cross Validation

The whole sample period is divided into sets of length a year. Each set is split into training (in-sample) and testing (out-sample) periods with a ratio of 2:1 (8 and 4 months). The training period is used to select the tradable tuples and to tune the parameters of the strategy. The testing period follows and its purpose is to assess the strategy by running the experiments with the parameters computed in the first period. Cointegration tests are performed for all possible combinations. Of the 1220 stocks, it is possible to form 1.8 billion spreads per period. The resulting cointegrated tuples are then ranked based on the in-sample Sharpe Ratio, similarly to Gatev et al. (2006). After selecting 20 pairs with the highest SR, four months of trading are carried out on the out-sample period. At the end of each trading period, all open positions are closed. The procedure continues in a rolling window fashion until the end of the sample.

7.8.1 Optimization of the strategy

Bollinger bands strategy requires to estimate three parameters: the number of periods p to compute the bands, the type of moving average used in the mid band and α which controls the interval between the volatility bands. John Bollinger suggests $p = 20, \alpha = 2$ and simple moving average by default. To optimize those parameters, a cross-validation scheme is performed. The criterion of optimization is the in-sample Sharpe Ratio.

7.8.2 Performance Assessment

The performance of the portfolios are examined in terms of cumulative return, variance of returns (σ^2), Sharpe Ratio (SR) and Maximum Drawdown (MDD). The maximum drawdown (MDD) is defined as the maximum percentage drop incurred from a peak to a bottom up to time T. Drawdowns help determine an investment's financial risk.

$$MDD(T) = \max_{\tau\in(0,T)} [\max_{\tau\in(0,T)} X(t) - X(\tau)] \tag{7.4}$$

The Sharpe Ratio (RP) based on daily returns is defined as

$$SR = \sqrt{252}.\frac{\bar{R_t}}{\sqrt{T^{-1}\sum_{t=1}^{T}(R_t - \bar{R_t})^2}}, \text{ where } \bar{R_T} = T^{-1}\sum_{t=1}^{T} R_t \tag{7.5}$$

One of the techniques to assess the performance of a strategy is to compare it to the very simple Buy and Hold strategy where the holder buys various assets at time 0 and keep them until time T. Gatev et al. (2006) also considered a bootstrap approach to generate random trading signals to assess the performance of a strategy over pure randomness. This approach is not discussed here since such a strategy has a negative expectation because of the trading costs and assuming the fact that you cannot beat the market with a random approach in the long run. So better not trade at all in this case.

7.9 Estimation and out-of-sample results

The sample is split into several training (in-sample) and testing sets (out-of-sample). Cross validation is performed on the training sets to tune the parameters of the strategies. The performance is evaluated on the testing set. We suggest a period of one year for testing and four months for testing.

7.10 Model selection

For each training period, the spread is computed and all the stochastic volatility models presented are applied to its returns. The model with the highest marginal likelihood is taken as reference and the Bayes factors are computed relatively to this model. AIC is also used to reinforce our decisions. 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. We find that the Gaussian Stochastic Volatility Leverage model performs best in terms of the marginal likelihood criterion. An advantage of using Bayes Factors and AIC is that they automatically penalize highly parametrized models that do not deliver improved content. Figure 1 displays the SPX500 index for the period 1/1/2010-1/1/2014 followed by the returns and the filtered estimates of $(X_t)_{t>0}$. We find that the Gaussian SVL model performs best in terms of the marginal likelihood and AIC criteria. The Kass factor $2 \log BF$ of SVL versus SV is 8 and this indicates strong evidence in favour of the SVL model. Compared to SVt, the Kass factor in favour of SVL is 13 which is also very strong evidence. The distribution of the parameters are also fairly concentrated around their means. The values of ϕ very close to one confirm strong daily volatility persistence, in accordance with stylized facts in econometrics known as volatility clustering.

8 Results

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 in vestment 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 deleverage 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.

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	14.88%	
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.

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. 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.

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.

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.

9 Appendix

9.1 SVL

Again by the Cholesky decomposition, y_t can be written as:

$$y_t|x_t = \rho\beta \exp(x_t/2)\epsilon_{X,t} + \beta \exp(x_t/2)\sqrt{1-\rho^2}Z \tag{9.1}$$

The only random quantity here is $Z \sim \mathcal{N}(0,1)$. Both factors on the right hand side are measurable at time t^-. Therefore, $y_t|x_t$ is normally distributed:

$$y_t|x_t \sim \mathcal{N}\left(\mathcal{A} = \rho\beta \exp(x_t/2)\epsilon_{X,t}, \mathcal{B} = \beta^2 \exp(x_t)(1-\rho^2)\right) \tag{9.2}$$

Using the fact that any AR(1) admits an infinite MA representation,

$$\begin{aligned} x_t &= \phi x_{t-1} + \sigma\epsilon_{X,t} \\ &= \phi(\phi x_{t-2} + \sigma\epsilon_{X,t-1}) + \sigma\epsilon_{X,t} \\ &= \sigma \sum_{j=0}^{\infty} \phi^j \epsilon_{X,t-j} \end{aligned} \tag{9.3}$$

and using this new representation into $\mathcal{A}$ gives:

$$\begin{aligned} \mathcal{A} &= \rho\beta \exp(x_t/2)\epsilon_{X,t} \\ &= \rho\beta \exp\left(\frac{\sigma}{2}\sum_{j=1}^{\infty}\phi^j\epsilon_{X,t-j}\right)\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t} \\ &= \rho\beta \exp\left(\frac{\phi}{2}x_{t-1}\right)\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t} \end{aligned} \tag{9.4}$$

At time $t-1$, only $\mathcal{C} = \exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t}$ is random. Because $\epsilon_{X,t}$ is independent from x_{t-1},

$$\begin{aligned} E[\mathcal{A}] &= \rho\beta E\left[\exp\left(\frac{\sigma}{2}\sum_{j=1}^{\infty}\phi^j\epsilon_{X,t-j}\right)\right] E\left[\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t}\right] \\ &= \rho\beta E\left[\prod_{j=1}^{\infty}\exp\left(\frac{\sigma}{2}\phi^j\epsilon_{X,t-j}\right)\right] E\left[\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t}\right] \\ &= \rho\beta \prod_{j=1}^{\infty} E\left[\exp\left(\frac{\sigma}{2}\phi^j\epsilon_{X,t-j}\right)\right] E\left[\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t}\right] \end{aligned} \tag{9.5}$$

$$
\begin{aligned}
E\left[\exp\left(\frac{\sigma}{2}\phi^j\epsilon_{X,t-j}\right)\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(\frac{-x^2}{2}+\frac{\sigma\phi^j}{2}x\right)dx \\
&= \left[\frac{1}{2}\exp\left(\frac{(\sigma\phi^j)^2}{8}\right)erf\left(\frac{2x-\sigma\phi^j}{2\sqrt{2}}\right)\right]_{-\infty}^{+\infty} \\
&= \frac{1}{2}\exp\left(\frac{(\sigma\phi^j)^2}{8}\right)(1-(-1)) \\
&= \exp\left(\frac{(\sigma\phi^j)^2}{8}\right) \qquad (9.6)
\end{aligned}
$$

$$
\begin{aligned}
E\left[\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(\frac{-x^2}{2}+\frac{\sigma}{2}x\right)dx \\
&= \frac{\sigma}{4}\exp\left(\frac{\sigma^2}{8}\right)\left[erf\left(\frac{2x-\sigma}{2\sqrt{2}}\right)-\frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}x(\sigma-x)\right)\right]_{-\infty}^{+\infty} \\
&= \frac{\sigma}{4}\exp\left(\frac{\sigma^2}{8}\right)(1-(-1)) \\
&= \frac{\sigma}{2}\exp\left(\frac{\sigma^2}{8}\right) \qquad (9.7)
\end{aligned}
$$

Because $\frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}x(\sigma-x)\right)\sim e^{-x^2}\to 0\ (x\to\infty)$. Therefore,

$$
\begin{aligned}
E[\mathcal{A}] &= \rho\beta\prod_{j=1}^{\infty}\exp\left(\frac{(\sigma\phi^j)^2}{8}\right)E\left[\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t}\right] \\
&= \rho\beta\frac{\sigma}{2}\exp\left(\frac{\sigma^2}{8}\right)\prod_{j=1}^{\infty}\exp\left(\frac{(\sigma\phi^j)^2}{8}\right) \\
&= \rho\beta\frac{\sigma}{2}\exp\left(\frac{\sigma^2}{8}\right)\exp\left(\frac{\sigma^2}{8}\sum_{j=1}^{\infty}\phi^{2j}\right) \\
&= \rho\beta\frac{\sigma}{2}\exp\left(\frac{\sigma^2}{8}\right)\exp\left(\frac{\sigma^2}{8}\left(\frac{1}{1-\phi^2}-1\right)\right) \\
&= \rho\beta\frac{\sigma}{2}\exp\left(\frac{\sigma^2}{8}\frac{1}{1-\phi^2}\right) \qquad (9.8)
\end{aligned}
$$

$$
\begin{aligned}
E[\mathcal{B}] &= \beta^2\exp(x_t)(1-\rho^2) \\
&= \beta^2(1-\rho^2)E\left[\exp\left(\sigma\sum_{j=0}^{\infty}\phi^j\epsilon_{X,t-j}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \beta^2(1-\rho^2)\prod_{j=0}^{\infty} E\left[\exp\left(\sigma\phi^j \epsilon_{X,t-j}\right)\right] \\
&= \beta^2(1-\rho^2)\prod_{j=0}^{\infty} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \exp\left(\sigma\phi^j x - \frac{x^2}{2}\right) dx \\
&= \beta^2(1-\rho^2)\prod_{j=0}^{\infty} \left[\frac{1}{2}\exp\left(\frac{(\sigma\phi^j)^2}{2}\right) erf\left(\frac{x-\sigma\phi^j}{\sqrt{2}}\right)\right]_{-\infty}^{+\infty} \\
&= \beta^2(1-\rho^2)\prod_{j=0}^{\infty} \exp\left(\frac{(\sigma\phi^j)^2}{2}\right) \\
&= \beta^2(1-\rho^2)\exp\left(\sum_{j=0}^{\infty}\frac{(\sigma\phi^j)^2}{2}\right) \\
&= \beta^2(1-\rho^2)\exp\left(\frac{\sigma^2}{2}\sum_{j=0}^{\infty}\phi^{2j}\right) \\
&= \beta^2(1-\rho^2)\exp\left(\frac{\sigma^2}{2}\frac{1}{1-\phi^2}\right) \qquad (9.9)
\end{aligned}
$$

9.2 TWSVL

The MA(∞) representation of X_t and Z_t are respectively

$$X_t = \sigma_X \sum_{j=0}^{\infty} \phi_X^j \epsilon_{X,t-j} \qquad (9.10)$$

$$Z_t = \sigma_Z \sum_{j=0}^{\infty} \phi_Z^j \epsilon_{X,t-j} \qquad (9.11)$$

We use the fact that X_t and Z_t are independent for each $t > 0$ i.e. $E[X_t Z_t] = E[X_t]E[Z_t]$. From the law of total expectation, $E(Y) = E_Y\left[E_{Y|X,Z}[Y|X,Z]\right]$. This assertion holds because $(X_t)_{t>0}$ and $(Z_t)_{t>0}$ are AR(1) processes and by their stationary property, $E[|X|], E[|Z|] < \infty$. Y is any random variable, not necessarily integrable but belonging to the same probability space.

$$
\begin{aligned}
E\left[Y_t\right] &= E\left[E[Y_t|x_t, z_t]\right] \\
&= \rho\beta E\left[\exp\left(\frac{\sigma}{2}\epsilon_{X,t}\right)\epsilon_{X,t}\right]\prod_{i=1}^{\infty} E\left[\exp\left(\frac{\sigma_X}{2}\phi_X^i\epsilon_{X,t-i}\right)\right]\prod_{j=0}^{\infty} E\left[\exp\left(\frac{\sigma_Z}{2}\phi_Z^j\epsilon_{X,t-j}\right)\right] \\
&= \frac{\sigma_X}{2}\exp\left(\frac{\sigma_X^2}{8}\right)\exp\left(\frac{\sigma_X^2}{8}\left(\frac{1}{1-\phi_X^2}-1\right)\right)\exp\left(\frac{\sigma_Z^2}{8}\left(\frac{1}{1-\phi_Z^2}\right)\right) \qquad (9.12)
\end{aligned}
$$

Similarly, the law of total variance is used to compute the unconditional variance of the stochastic process $(Y_t)_{t>0}$. It is assumed that $\mathrm{Var}(Y) < \infty$ which is the case in practice as the returns are finite almost surely. By definition,

$$\mathrm{Var}(Y) = \underbrace{E_{X,Z}(Var[Y|X,Z])}_{(1)} + \underbrace{\mathrm{Var}_{X,Z}(E[Y|X,Z])}_{(2)} \tag{9.13}$$

The term (1) is computed using the same logic as seen for model $\mathcal{M}_3$ with the independence of X_t and Z_t for every $t > 0$.

$$E_{X,Z}\left[\mathrm{Var}[Y_t|x_t,z_t]\right] = \beta^2(1-\rho^2)\exp\left(\frac{\sigma_X^2}{2}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{2}\frac{1}{1-\phi_Z^2}\right) \tag{9.14}$$

The term (2) is rewritten as

$$\begin{aligned}\mathrm{Var}_{X,Z}\left[E[Y_t|x_t,z_t]\right] &= \rho^2\beta^2\,\mathrm{Var}\left(\exp\left(\frac{x_t+z_t}{2}\right)\right)\\ &= \rho^2\beta^2\left(E\left[\exp\left(x_t+z_t\right)\right] - E\left[\left(\exp\left(\frac{x_t+z_t}{2}\right)\right)\right]^2\right)\end{aligned} \tag{9.15}$$

Recall from the calculus for model $\mathcal{M}_3$, $E\left[\exp\left(\frac{\sigma}{2}\phi^j\epsilon_{X,t-j}\right)\right] = \exp\left(\frac{(\sigma\phi^j)^2}{8}\right)$. By a trivial substitution, $\sigma' = 2\sigma$, $E\left[\exp\left(\sigma'\phi^j\epsilon_{X,t-j}\right)\right] = \exp\left(\frac{(\sigma'\phi^j)^2}{2}\right)$. Therefore,

$$\begin{aligned}(2) &= \rho^2\beta^2\left(E[\exp(x_t)]E[\exp(z_t)] - E\left[\left(\exp\left(\frac{x_t+z_t}{2}\right)\right)\right]^2\right)\\ &= \rho^2\beta^2\left(\prod_{j=0}^{\infty}\exp\left(\frac{(\sigma_X\phi_X^j)^2}{2}\right)\prod_{j=0}^{\infty}\exp\left(\frac{(\sigma_Z\phi_Z^j)^2}{8}\right) - E\left[\left(\exp\left(\frac{x_t+z_t}{2}\right)\right)\right]^2\right)\\ &= \rho^2\beta^2\left(\exp\left(\frac{\sigma_X^2}{2}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{2}\frac{1}{1-\phi_Z^2}\right) - E\left[\left(\exp\left(\frac{x_t+z_t}{2}\right)\right)\right]^2\right)\\ &= \rho^2\beta^2\left(\exp\left(\frac{\sigma_X^2}{2}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{2}\frac{1}{1-\phi_Z^2}\right) - E\left[\left(\exp\left(\frac{x_t}{2}\right)\right)\right]^2 E\left[\left(\exp\left(\frac{z_t}{2}\right)\right)\right]^2\right)\\ &= \rho^2\beta^2\left(\exp\left(\frac{\sigma_X^2}{2}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{2}\frac{1}{1-\phi_Z^2}\right) - \exp\left(\frac{\sigma_X^2}{4}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{4}\frac{1}{1-\phi_Z^2}\right)\right)\end{aligned} \tag{9.16}$$

Finally by adding (1) and (2),

$$\begin{aligned}Var(Y) = \rho^2\beta^2&\left(\exp\left(\frac{\sigma_X^2}{2}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{2}\frac{1}{1-\phi_Z^2}\right) - \exp\left(\frac{\sigma_X^2}{4}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{4}\frac{1}{1-\phi_Z^2}\right)\right)\\ &+ \beta^2(1-\rho^2)\exp\left(\frac{\sigma_X^2}{2}\frac{1}{1-\phi_X^2}\right)\exp\left(\frac{\sigma_Z^2}{2}\frac{1}{1-\phi_Z^2}\right)\end{aligned} \tag{9.17}$$