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

15th July 2015

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

Abstract

This project is aimed to investigate the practical benefit of using more complex modelling than what is currently standard practice in applications related to statistical arbitrage. The underlying assets will be modelled using appropriate mean-reverting time series or state space models. In order to fit these models to real data the project will involve using advanced particle methods such as Particle Markov Chain Monte Carlo. The primary aim of the project is to assess whether using more advanced modelling and model calibration will result to better performance than simple models used often in practise. This will be illustrated in numerical examples, where the computed portfolio is used for a realistic scenario obtained by popular trading platforms. Simulations will be mainly run in Matlab, but embedding C/C++ routines may be required to speed up computations. The project is a challenging computational Statistics application to finance and is this suitable for a student with an interest in finance, very good aptitute to computing and understanding of the material in the course related to Monte Carlo methods and Time Series.

Throughout, $p(\cdot)$ and $p(\cdot|\cdot)$ are used to denote general marginal and conditional probability density functions, with the arguments making it clear which distributions these relate to.
A realization of a stochastic process X on a finite discrete space $\{0, ..., T\}$ is denoted $x_{0:T} = (x_1, ..., x_T)$.

1 Cointegration

1.1 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. Charles Nelson and Charles 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 asumed to have a long-tun equilibrum 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. Recall Δ is the first difference operator defined by : $\Delta x_t = x_t - x_{t-1}$.

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-rtun 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 correspondinng 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.

1.2 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:

$$X_t = \nu + \sum_{j=1}^{k} A_j X_{t-j} + \epsilon_t,\ \epsilon_t \sim SWN(0, \Sigma)$$

where $X_t = (x_{1t}, ..., x_{nt})_{t \geq 0}$, 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 $|I_n - \sum_{j=1}^{k} A_j z^j| = 0$ lie outside the complex unit circle.

1.3 Vector Error Correction models (VECM)

In an error correction model, the changes in a variable depend on the deviations from some equilibrium relation. Suppose the case $n = 2, X_t = (x_t, y_t)$ 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:

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

where α represents the speed of adjustements 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:

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

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

$$\Delta y_t = \prod y_{t-1} + \sum_{j=1}^{k-1} \Gamma_j \Delta_{t-j} + \epsilon_t$$

where $\Gamma_j = -\sum_{i=j+1}^{k} A_i$ and $\prod = -\sum_{i=1}^{k} A_i$. This way of specifying the system contains information on both the short-run and long run adjustements to changes in y_t, via the estimates of $\hat{\Gamma}_j$ and $\hat{\prod}$ respectively. To be continued...

2 Particle MCMC

2.1 Introduction

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.

2.2 State-Space Models

The state-space models are parameterised 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:

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

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:

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

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:

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

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.

2.3 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 $x_0^k \sim p(x_0),\ k = 1, \ldots, M$ with uniform normalised weights $w'^k_0 = 1/M$. Then suppose that we have a weighted sample $\{x_t^k, w'^k_t | k = 1, \ldots, M\}$ from $p(x_t|y_{1:t})$. First generate an equally weighted sample by resampling with replacement M times to obtain $\{\tilde{x}_t^k | k = 1, \ldots, M\}$ (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 $x_{t+1}^k \sim p(x_{t+1}|\tilde{x}_t^k),\ k = 1, \ldots, M$ (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:

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

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:

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

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

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

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

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

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

Algorithm 1 Bootstrap Particle Filtering Algorithm (SIR)

procedure INPUT($y_{1:T}$, θ, 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. This result is the cornerstone of Particle MCMC models, especially for the particle marginal Metropolis-Hastings Algorithm explained in the next section.

2.4 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:

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

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 is hard 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 linear Gaussian process $x_{t+1} = \rho x_t + \tau\epsilon_{t+1}$, a path $x_{0:T}$ 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:

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

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 the reason why x^* should be sampled from $p(x^\star|\theta^\star, y)$. With the remark, the ratio now becomes:

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

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

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

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 Andreiu 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 is depicted below.

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)
  $\hat{p}^N_{\theta^{(0)}}(y_{1:T}), {x^*_{1:T}}^{(0)} \leftarrow$ Call Bootstrap Particle Filter with ($y_{1:T}$, $\theta^{(0)}$, N)
  for i from 1 to M **do**
    Sample θ' from q($\theta|\theta^{(i-1)}$)
    $\hat{p}^N_{\theta'}(y_{1:T}), {x^*_{1:T}}' \leftarrow$ Call Bootstrap Particle Filter with ($y_{1:T}$, θ', 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 ${x^*_{1:T}}^{(i)} \leftarrow {x^*_{1:T}}', \theta^{(i-1)} \leftarrow \theta', \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta'}(y_{1:T})$
    Otherwise ${x^*_{1:T}}^{(i)} \leftarrow {x^*_{1:T}}^{(i-1)}, \theta^{(i-1)} \leftarrow \theta^{(i-1)}, \hat{p}^N_{\theta^{(i)}}(y_{1:T}) \leftarrow \hat{p}^N_{\theta^{(i-1)}}(y_{1:T})$
  end
  return $({x^*_{1:T}}^{(i)}, \theta^{(i)})_{i=1}^M$

2.5 Heston

2.5.1 Simulation of Probability Densities

By Ito calculus, and more precisely the Euler-Maruyama method, the Heston stochastic process can be discretized and results in:

$$\begin{aligned} S_t &= S_{t-1} + rS_{t-1}dt + \sqrt{V_{t-1}}S_{t-1}\sqrt{dt}Z^S_t \\ V_t &= V_{t-1} + \kappa(\theta - V_{t-1})dt + \sigma\sqrt{V_{t-1}}\sqrt{dt}Z^V_t \end{aligned}$$

where the innovations $\{Z^S_t\}_{t\geq 0}$ and $\{Z^V_t\}_{t\geq 0}$ are standard normal random variables with correlation ρ. The generation is made simple by considering the Cholesky decomposition,

$$\begin{aligned} Z^S_t &= \phi^S_t \\ Z^V_t &= \rho\phi^V_t + \sqrt{1-\rho^2}\phi^V_t \end{aligned}$$

where $\{\phi^S_t\}_{t\geq 0}$ and $\{\phi^S_t\}_{t\geq 0}$ are independent standard normal random variables.

2.6 Stochastic Volatility

In this section, we introduce the standard stochastic volatility with Gaussian errors. Next, we consider different well-known extensions of the SV model. The first extension is a SV model with Student-t errors. In the second extension, we incorporate a leverage effect by modeling a correlation parameter between measurement and state errors. In the

third extension, we implement a model that has both stochastic volatility and moving average errors.

2.6.1 Simple SV Model

The standard discrete-time stochastic volatility model for the returns Y_n is defined as:

$$\begin{aligned} X_{n+1} &= \rho X_n + \sigma V_n \\ Y_n &= \beta \exp\left(\frac{X_n}{2}\right) W_n \end{aligned}$$

where $\{V_n\}$,$\{W_n\}$ are two independent sequences of independent standard normal random variables. Let $\theta = (\rho, \sigma^2, \beta)$. Notice that the non-linearity of the models relies in the non-additive noise of the transition Kernel. X_n is the unobserved log-volatility associated to the observed log-returns Y_n, σ is the volatility of the log-volatility and ρ is the persistence parameter. The condition $|\rho| < 1$ is imposed to have a stationary process with the initial condition $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\rho^2}\right)$.

2.6.2 SVt - Student-t innovations

The first extension is a stochastic volatility model with $W_n \sim St(\nu)$ where St stands for the Student-t distribution with $\nu > 2$. The conditional density becomes (pdf variable substitution):

$$p(y_n|x_n, Y_{n-1}, \theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\nu-2)\pi}} \frac{1}{\sigma_n} \left(1 + \frac{y_n^2}{\sigma_n^2 \nu}\right)^{-\frac{v+1}{2}}$$

where $\sigma_n = \beta \exp\left(\frac{x_n}{2}\right)$. We then follow the sampling steps as before.

2.6.3 SVL - Stochastic Volatility Leverage

In the second extension, we incorporate a leverage effect by letting c denote the correlation between V_n and W_n. Here, we use the fact that $V_n = cW_n + \sqrt{1-c^2}\Psi_n$ where $\Psi_n \sim N(0,1)$:

$$\begin{aligned} X_{n+1} &= \rho X_n + \sigma \left(cW_n + \sqrt{1-c^2} B_n\right) \\ X_{n+1} &= \rho X_n + \sigma \left(cY_n \exp\left(-\frac{X_n}{2} - \log(\beta)\right) + \sqrt{1-c^2} B_n\right) \end{aligned}$$

Notice that we need to sample the additional parameter c.

2.6.4 SV-MA(1) - Moving Average

We can also expand the plain stochastic volatility model by allowing the errors in the measurement equation to follow a moving average (MA) process of order m. This means that the errors in the measurement equation are no longer serially independent as for the plain SV model. Here, we choose a more simple specification and set $m = 1$. Hence, our model becomes:

$$\begin{aligned} Y_n &= \beta \exp\left(\frac{X_n}{2}\right) W_n + \Psi\beta \exp\left(\frac{X_{n-1}}{2}\right) W_{n-1} \\ X_{n+1} &= \rho X_n + \sigma V_n \end{aligned}$$

We ensure that the root of the characteristic polynomial assocaited with the MA coefficient Ψ is outside the unit circle, $|\Psi|$

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.

$$Var(y_t|\mathcal{F}_{t-1}) = \exp(X_t) + \Psi^2 \exp(X_{t-1})$$

because X_t is measurable with regard to $\mathcal{F}_{t-1}$. It turns out that the conditional variance of y_t is varying through two channels. Estimating this model is straightfoward as again we only need to make small adjustments in the codes.

2.6.5 SV-M

Let's consider the population stochastic volatility in mean (SVM) model where $\exp(X_t/2)$ appears in both the conditional mean and the conditional volatility. We follow the same notation as before and define the SVM model as:

$$y_t = \beta \exp\left(\frac{X_t}{2}\right) + \exp\left(\frac{X_t}{2}\right) + W_t, \ , W_t \sim N(0,1)$$

where X_t is ruled by the dynamics of a simple SV model. The conditional probability density of y_t is $p(y_t|x_t, Y_{t-1}, \theta) \sim N(\beta \exp(x_t/2), exp(x_t))$.

2.6.6 TFSV - Two Factors

Finally, we estimate a two factor SV model. It is defined as:

$$\begin{aligned} X_{n+1} &= \rho_1 X_n + \sigma_2 V_n, \ |\rho_1| < 1, V_n \sim N(0,1) \\ Z_{n+1} &= \rho_2 Z_n + \sigma_2 P_n, \ |\rho_2| < 1, P_n \sim N(0,1) \\ Y_n &= \exp\left(\frac{\mu}{2} + \frac{X_n + Z_n}{2}\right) W_n, \ |\rho_2| < 1, W_n \sim N(0,1) \end{aligned}$$

θ is enriched with the new parameters. Thus, we only need to modify the particle filter such that we draws two sets of particles (one for X_t and one for Z_t) instead of one.

2.6.7 Some calculus

It turns out that the model with the highest likelihood is the Normal Leverage Stochastic Volatility model. The model is equipped with a two steps filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. x_t and y_t are both 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. Again by the Cholesky decomposition, y_t can be written as:

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

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

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

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

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

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

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

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},

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

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

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

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,

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

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

2.6.8 Models and Bollinger Bands

The Stochastic Volatility models are fitted to the prices. y_t represents the daily returns of the prices. 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. Assume the standard stochastic volatility model,

$$
\begin{aligned}
y_t | x_t &= \mathcal{N}(0, \beta^2 \exp(x_t)) \\
\frac{S_t}{S_{t-1}} - 1 | x_t &= \mathcal{N}(0, \beta^2 \exp(x_t)) \\
\frac{S_t}{S_{t-1}} | x_t &= \mathcal{N}(1, \beta^2 \exp(x_t)) \\
S_t | x_t &= \mathcal{N}(S_{t-1}, \beta^2 \exp(x_t))
\end{aligned}
$$

This trick does not work for log-returns. The idea behind using these models is to catch the dynamics of the spread. A stock price is only one observation of a more general process over a time interval. Two general approaches can be derived. The first approach

consists in working with returns and not price curves like used usually with Bollinger bands. Denote $x_t = \sigma(t)$. The idea is to use a moving average on $\sigma(t)$ to compute the bands:

$$SMA_\sigma(n) = \frac{1}{n}\sum_{i=0}^{n-1}\sigma(t-i)$$

The mid band is generated from a moving average over the returns.

The second method is based on drawing many paths from the model to estimate the volatility of the spread. The volatility is associated to the prices and not to the returns. It corresponds to the usual volatility concept used in Bollinger Bands. An aggregation function must be selected to compute the volatility from the different generated paths.

2.7 Model Comparison

The output of the particle filter is an estimate of $p(y|\theta)$, with the unobserved states integrated out. The marginal likelihood for a model $\mathcal{M}$ is defined as:

$$p(Y_T|\mathcal{M}) = \int_\theta p(Y_T|\theta,\mathcal{M})p(\theta|\mathcal{M})d\theta$$

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

$$\frac{1}{N}\sum_{i=1}^{N}\frac{g(\theta_i)}{p(Y_T|\theta_i)p(\theta_i)} \rightarrow \frac{1}{p(Y_T)} \text{ as } \mathrm{It}_{mcmc} \rightarrow +\infty$$

For this estimator to be consistent, $g(\theta_i)$ must be thin-tailed relative to the denominator. For most cases, a multivariate Normal distribution $N(\theta^*, \Sigma^*)$ can be used, where $\theta^* = \frac{1}{N}\sum_{i=1}^{N}\theta^i$ 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 the implementation. As a matter of fact, $p(Y_T|\theta)$ is usually very small as 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, when N = 3 and let 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 $p(Y_T|\mathcal{M}_\mathcal{B})$ have been estimated, Kass and Raftery (1995) suggest to use twice the logarithm of the Bayes factor for model comparison, $2\log BF_{\mathcal{M}_{\mathcal{AB}}}$. 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.

2.8 Resampling

Resampling is a key component of the Particle Filter. Different methods exist: stratified, systematic and residuals resampling. In practical applications, they are generally found to provide comparable results. Despite the lack of complete theoretical analysis of its behavior, systematic resampling is often preferred because it is the simplest method to implement. Randal Douc proved that residual and stratified resampling methods dominate the basic multinomial approach, in the sense of having lower conditional variance for all configurations of the weights.

Resampling method	Residual	Stratified	Systematic	Multinomial
Time (in seconds)	18.90	0.62	0.63	1.87

Table 2.1: Time spent to resample 100K times 1000 weights

The multinomial implementation is the MATLAB default version. According to Randal Douc and the performance results, the stratified resampling seems the most compelling method to use inside the particle filters. This part is critical because it can represent up to 50% of the total time spent in the filter.

3 Trading Strategy

3.1 Statistical Arbitrage

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 strategy is a mean-reverting strategy. Once the spread (define it) is far from its long-run equilibrium, enter position and unwind (timing is important)

3.2 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 $\{Ma_t\}$. 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:

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

$$Boll^{\pm}(t) = Ma(t) \pm d * \sqrt{\frac{1}{n}\sum_{j=1}^{n}(p_j - Ma(t))}$$

where n is the number of time periods in the moving average and d is the number of standard deviations to shift the bollinger bands. The default values are $n = 20$ and $d = 2$. The moving average can be replaced by an exponential moving average that gives more weights to new values and increase the accuracy eventually.

3.3 Z-score

Once the spread $(\epsilon_t)_{t\geq 0}$ is formed, Caldeira 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 reasonment for the high threshold holds for short positions. Caldeira Moura (2013) suggested the basic trading strategy signals:

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

3.4 Strategy

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. This approach has the advantage of eliminating the market exposure (memo corr cumsum with SP500 should be around 0). A typical trading strategy is made of three parts: selection of the suitable tuples satisfying some criterias like the cointegration, create trading signals based on define predefined investment decision rules and finally assess the performance of the strategy.

3.4.1 Tuples selection

Cointegrated tuples

It is common in pair trading and more generally in basket 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 Moura (2013). However it is possible for pair trading but becomes impossible when the number of assets grows. Having in mind that increasing the size n of the tuples leads to an combinatorial explosion, it is necessary to make some trade-offs. First, cointegration usually implies correlation but correlation doesn't always imply cointegration. Spurious regression is a very good example where the reverse is not true. However, a correlation test is usually much faster than for instance a Johansen cointegration test (cf. table 2).

A simple correlation test is used for pair trading. 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

Test	Correlation	Johansen	Aug. Dickey Fuller	Phillips-Perron
Time	0.33 ms	19.08 ms	2.33 ms	3.04 ms

Table 3.1: Average time spent to test a bivariate time series $X_t = (x_{t1}, x_{t2})$

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

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

One problem arises: the value of the coefficient depends on the order of the tuple. 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. To convince ourselves, 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 order inside a tuple has its importance, at least from a statistical point of view. It is much less obvious from a pure financial point of view.

First, an Augmented Dickey Fuller test is performed to check that all series are integrated of the same order, $I(1)$, i.e, they contain a unit root. For each tuple candidate, R^2 is evaluated. When the empirical distribution of all R^2 is known, a threshold R_{th} is derived and the tuples whose $R^2 > R_{th}$ are chosen. For every selected tuple, apply a triple Johansen, Dickey Fuller and Phillips-Perron test to check for cointegration. If the tuple is cointegrated, form the spread and mark it as tradable.

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. Recall that Sharpe ratio is calculated as the ratio of annualized return to annualized standard deviation. 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.

4 Performance Assessment

In order to reduce risk in the strategies, it is interesting to open many trades inside a portfolio, all with a very short holding time, hoping to diversify the risk of each trade. The performance of the portfolios are examined in terms of cumulative return, variance of returns, Sharpe Ratio and Maximum Drawdown (MDD). The maximum drawdown (MDD) is defined as the maximum peak to trough decline and represents the worst scenario up to time T. Drawdowns help determine an investment's financial risk.

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

The Sharpe Ratio (RP) is defined as:

$$SR = \sqrt{252}.\frac{T^{-1}\sum_{t=1}^{T} R_t}{\sqrt{T^{-1}\sum_{t=1}^{T}(R_t - T^{-1}\sum_{t=1}^{T} R_t)^2}}$$

One of the technique 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. Gatel 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.

4.1 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.

4.2 Presentation of the dataset - Big introduction

The dataset is composed ... 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 cointegration in quantitative strategies.