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

20th June 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 Particle MCMC

1.1 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.

1.2 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.

1.3 Particle Filter

The particle filter is an iterative MC method for carrying out bayesian inference on state-space models. The method is a very simple application of the importance resampling technique. At each time, t, we assume that we have a (approximate) sample from $p(x_t|y_{1:t})$ and use importance resampling to generate an approximate sample from $p(x_{t+1}|y_{1:t+1})$.

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 $w'^k_{t+1} = w_{t+1}^k / \sum_i w_{t+1}^i$.

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

$$\hat{p}(y_t|y_{1:t-1}) = \frac{1}{M} \sum_{k=1}^{M} w_t^k$$

and

$$\hat{p}(y_{1:T}) = \hat{p}(y_1) \prod_{t=2}^{T} \hat{p}(y_t|y_{1:t-1}).$$

Again, from our understanding of importance resampling, it should be reasonably clear that $\hat{p}(y_{1:T})$ is a consistent estimator of $p(y_{1:T})$. It is much less clear, but nevertheless true that this estimator is also unbiased. The standard reference for this fact is Del Moral (2004), but this is a rather technical monograph. A much more accessible proof (for a very general particle filter) is given in Pitt et al (2011).

It should therefore be clear that if one is interested in developing MCMC algorithms for state space models, one can use a pseudo-marginal MCMC scheme, substituting in $_\theta(y_{1:T}) from a bootstrap particle filter in place of \mathrm{p}(\mathrm{y}_{1:T}|\theta)$. This turns out to be a simple special case of the particle marginal Metropolis-Hastings (PMMH) algorithm described in Andreiu et al (2010).

In bootstrap particle filter, $\pi(x_k^{(i)}|x_{0:k-1}^{(i)}, y_{0:k}) = p(x_k^{(i)}|x_{k-1}^{(i)})$. When the transition prior probability is used as the importance function, the weights update formula is simplified:

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

In the bootstrap particle filter, it is 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:

$$p_\theta^N(y_t|y_{0:t-1}) = \frac{1}{N}\sum_{k=1}^{N} w_t^k$$

The marginal likelihood at time T is:

$$p_\theta(y_{0:T}) = p(y_0)\prod_{t=1}^{T} p(y_t|y_{1:t-1})$$

Again, from our understanding of importance resampling, it should be reasonably clear that $\hat{p}_\theta^N(y_{0:T})$ is a consistent estimator of $p_\theta(y_{0:T})$. It is much less clear, but nevertheless true that this estimator is also unbiased according to Del Moral (2004).

The marginal log likelihood and its estimator are:

$$\log(p_\theta(y_{0:t})) = \log(p(y_0)) + \sum_{t=1}^{t} \log\left(p(y_t|y_{1:t-1})\right)$$

$$\log(p_\theta^N(y_{0:t})) = -\log(N) + \sum_{t=1}^{t} \log\left(\frac{1}{N}\sum_{k=1}^{N} w_t^k\right)$$

Algorithm 1 Bootstrap Particle Filtering Algorithm (SIR)

procedure

  for i from 1 to N **do**
    $x_k^{(i)} \sim \pi(x_k|x_{0:k-1}^{(i)}, y_{0:k})$
    $w_k^{(i)} = \hat{w}_k^{(i-1)} p(y_k|x_k^{(i)})$
  end

  for i from 1 to N **do**
    $w_k^{(i)} = \hat{w}_k^{(i)} / \sum_{j=1}^{N} \hat{w}_k^{(j)}$
  end

  x_k = resampling(x_k, w_k)
  for i from 1 to N **do**
    $w_k^{(i)} = 1/N$
  end

return (x_k, w_k)

1.4 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.

1.4.1 Simple SV Model

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

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

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 $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\rho^2}\right)$.

1.4.2 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.

1.4.3 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.

1.4.4 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.

1.4.5 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 $p(y_t|x_t, Y_{t-1}, \theta) \sim N(\beta \exp(x_t/2), exp(x_t))$.

1.4.6 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 drawa two sets of particles (one for X_t and one for Z_t) instead of one.

1.5 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:

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

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 $\Sigma^* = \frac{1}{N-1}\sum_{i=1}^{N} (\theta^i - \theta^*)(\theta^i - \theta^*)^T$. 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.

1.6 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 1.1: 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.

1.7 PMMH

In a more general context, a Metropolis Hastings scheme can be used to target $p(\theta|y)$ 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. 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, in the linear Gaussian case, a path $x_{0:T}$ can be simulated when ρ and τ are known. The MH is built in two stages. First, a new θ^* is proposed from $q(\theta^\star|\theta)$ and then, x^* is sampled from $p(x^\star|\theta^\star)$. The 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, the x^* is consistent of θ^* thanks to the tractability of $p(x^\star|\theta^\star)$. 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 $p(x^\star|\theta^\star, y)$. The ratio 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:

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

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.