A Monte Carlo EM algorithm for discretely observed Diffusions, Jump-diffusions and Lévy-driven Stochastic Differential Equations

Erik Lindström

Abstract— Stochastic differential equations driven by standard Brownian motion(s) or Lévy processes are by far the most popular models in mathematical finance, but are also frequently used in engineering and science. A key feature of the class of models is that the parameters are easy to interpret for anyone working with ordinary differential equations, making connections between statistics and other scientific fields far smoother.

We present an algorithm for computing the (historical probability measure) maximum likelihood estimate for parameters in diffusions, jump-diffusions and Lévy processes. This is done by introducing a simple, yet computationally efficient, Monte Carlo Expectation Maximization algorithm. The smoothing distribution is computed using resampling, making the framework very general.

The algorithm is evaluated on diffusions (CIR, Heston), jump-diffusion (Bates) and Lévy processes (NIG, NIG-CIR) on simulated data and market data from S & P 500 and VIX, all with satisfactory results.

Keywords— Bates model, Heston model, Jump-Diffusion, Lévy process, parameter estimation, Monte Carlo Expectation Maximization, NIG, Stochastic differential equation.

I. Introduction

Stochastic differential equations (SDEs) driven by standard Brownian motion(s) is an increasingly popular class of models in diverse scientific fields such as mathematics, [27], biology [24], hydrology, engineering, economics, [35], [41] and finance, [21], [11], [25]. The textbook [34] contains additional examples and references for this class. A strong argument for using SDEs is that the parameters are easy to interpret even for non-experts, contrary to most other non-linear time series models or neural nets.

Parameter estimation for diffusion processes is an established science, where many different strategies are known. The Euler-Maruyama scheme can be used to compute the maximum likelihood estimate if the data is sampled frequently, see [39]. More advanced methods for approximating the transition kernel are reviewed in [22]. These include series expansions, numerical solution of the Fokker-Planck equation and Monte Carlo methods.

- Series expansions of the transition kernel was developed in [1], [2] for diffusion processes. The series expansion is correcting for deviations from a Gaussian distribution. Processes must be transformed such that the diffusion term is independent of the state, which is feasible for all univariate models but not for all multivariate models. [3] extends the framework to non-Gaussian expansions.
- Numerical methods for the Fokker-Planck equation was introduced in [33] and explored further by [28], [38]. This approach is more general than the series expansion, and is computationally efficient for low-dimensional systems, cf. [28]. However, generalizing the approach to jump-diffusions is rather difficult as the corresponding Fokker-Planck equation is a Partial Integro-Differential Equation (PIDE) which is considerably more complicated to solve numerically.
- Monte Carlo methods are very general. Early methods, such as the Pedersen method, [37] are computationally inefficient but later versions, [13], [18], [29] are several magnitudes faster. MCMC methods have been proposed by [15], [14], [19], [42]. Another approach is exact simulation, see [8]. That method use rejection sampling to simulate trajectories with no systematic bias, but the computational complexity can be costly for sparsely sampled data and their algorithm is also less general than the Pedersen method.

However, the recent financial crisis have clearly shown that diffusion processes may be insufficient, as the continuous path of the Brownian motion is unable to generate rapid changes. Recent financial models are routinely including jumps in addition to the diffusion to account for this, see [11].

The statistical analysis of these models is less well understood as it is difficult to find a closed form expression for the transition kernel for many relevant models. Models for which we can approximate the transition kernel have therefore been popular, but this reminds us of the story of the drunk looking for his lost keys under the lamp post. How do we know that we are not missing something important when restricting ourselves due to primarily computational considerations?

Research on jump-diffusion often assume that the dynamics can be approximated by the Euler-Maruyama scheme, see e.g. [16], [23], although some work on asymptotically unbiased maximum likelihood estimation have been done, [20], [7], [30]. These methods are quite restrictive in terms of applicability.

The purpose of this paper is to derive a simple, general and computational efficient algorithm for computing the maximum likelihood estimate for parameters in diffusion, jump-diffusions and Lévy driven stochastic differential equations in a unified framework, thereby extending the work by [30]. This extension introduces additional issues as Lévy processes are far more general than the standard Brownian motion that drives

Manuscript received May 4, 2012.

Erik Lindström is with the Centre for Mathematical Sciences, Lund University, Sölvegatan 18, 221 00 Lund, Sweden, (email: erikl@maths.lth.se)

diffusions (the Brownian motion is a special case of a Lévy process, the Poisson process is another).

This is achieved by deriving an EM-algorithm where the smoothing distribution is computed using resampling of paths generated by sampling from the naive dynamics. The framework is very general and can be applied to most semimartingales. Another approach would be to use importance sampling to derive the smoothing distribution, but we have refrained from doing so as it is practically difficult to ensure that the theoretical support (e.g. the tails of the sampling distribution must be approximately as heavy as the target) for importance sampling is valid for Lévy-driven SDEs.

The remainder of the paper is organized as follows: Section II introduces the EM-algorithm and explains how resampling can be used to compute the smoothing distribution. Section III evaluates the derived estimator on simulated and real data (S&P 500 and VIX) and compares the estimates to the results in [4]. Finally, Section IV concludes.

II. Parameter estimation

We denote the observations sampled at discrete time points $t_1, \ldots, t_N$ by $\boldsymbol{y}_{t_1}, \ldots, \boldsymbol{y}_{t_N}$. The maximum likelihood estimates are then defined as

$$\hat{\theta}_{MLE} = \arg\max \log p_\theta(\boldsymbol{y}_{t_1}, \ldots, \boldsymbol{y}_{t_N}). \tag{1}$$

The MLE is often the optimal parameter estimator in terms of variance, cf. the Cramer-Rao inequality. The optimality is shared by approximate maximum likelihood estimators, cf. [36] if the approximation fulfils some conditions.

A particularly nice feature of the Pedersen (Monte Carlo) method, see [37], is that it is very flexible and does not (contrary to many extensions) impose any implicit assumption on the dynamics of the model. The only property used is that the model is a Markov process, which is why we base our work on the ideas behind that method.

It follows from the law of total probability and the Markov property that the transition kernel, $p_\theta(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{y}_{t_n})$, for any Markov process can be computed from

$$p_\theta(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{y}_{t_n}) = \int p_\theta(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{y}_\tau) p_\theta(\boldsymbol{y}_\tau|\boldsymbol{y}_{t_n}) \mathrm{d}\boldsymbol{y}_\tau \tag{2}$$

where τ is some time point satisfying $t_n < \tau < t_{n+1}$. A (point wise) Monte Carlo approximation of the transition kernel can therefore be obtained by sampling $\boldsymbol{u}_\tau^{(k)},\ k = 1, \ldots, K$ from $p(\boldsymbol{y}_\tau|\boldsymbol{y}_{t_n})$ and computing

$$\hat{p}_\theta(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{y}_{t_n}) \approx \frac{1}{K}\sum_{k=1}^{K} p_\theta(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{u}_\tau^{(k)}). \tag{3}$$

We need to discretize the model to perform the computations in practice. This is done by introducing a partitioning of the time scale between two observations into J smaller time intervals $t_n = \tau_{(n-1)J} < \tau_{(n-1)J+1} < \ldots < \tau_{nJ} = t_{n+1}$, such that the discretization error over a short time interval $|\tau_{j+1} - \tau_j|$ is small (it can be controlled by making the partitioning finer as the bias typically is a function of the time interval). The properties of this approximation are well known, cf. [36], [13], [43]. The bias decreases with finer partitioning but the variance increases at the same time.

The transition density can also be computed using importance sampling. Generating samples $\tilde{u}_\tau^k$ from $q(\boldsymbol{y}_\tau)$ leads to

$$\hat{p}_\theta(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{y}_{t_n}) \approx \frac{1}{K}\sum_{k=1}^{K} p_\theta(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{u}_\tau^{(k)}) \frac{p_\theta(\boldsymbol{u}_\tau^{(k)}|\boldsymbol{y}_{t_n})}{q(\boldsymbol{u}_\tau^{(k)})}. \tag{4}$$

A practical problem (apart from the computational issues) when applying the Pedersen sampler to jump-diffusions is that the sampler only generates point wise estimates of the transition kernel. This is not a problem for pure diffusions as common random numbers can be used, but the number of jumps in an interval cannot be simulated using common random numbers. It would then be necessary to combine the Pedersen method with stochastic approximation to get reliable parameter estimates, making the method even more computationally intensive.

A. EM-algorithm

The EM-algorithm is commonly used in statistics when dealing with latent or missing variables. We use $\boldsymbol{x}$ to denote any latent variable. The purpose of the EM-algorithm is to find the maximum likelihood estimate by iteratively applying the E-step and M-step

○ **E-step:** Compute the expectation

$$Q(\theta, \theta_m) = \boldsymbol{E}[\log p_\theta(\boldsymbol{y}, \boldsymbol{x})|\boldsymbol{y},\ \theta_m]. \tag{5}$$

This is called the *intermediate quantity* by [9].

○ **M-step**: Maximize

$$\theta_{m+1} = \arg\max Q(\theta, \theta_m). \tag{6}$$

A feature of the EM-algorithm that is important for our purpose is that the smoothing distribution is fixed when computing $Q(\theta, \theta_m)$. This makes the $Q(\theta, \theta_m)$ function continuous in the θ-parameter, simplifying the optimization problem.

It has been shown under weak conditions that the EM-algorithm has a monotonic behaviour and that it converges to a local maximum of the log-likelihood function, see [9]. This does not hold when the expectation is approximated with a Monte Carlo technique. The Monte Carlo EM (MCEM) converge if the size of the random sample K_m is increasing sufficiently fast as $m \to \infty$, i.e replacing the E-step with

$$Q_m(\theta, \theta_m) = \frac{1}{K_m}\sum_{k=1}^{K_m} \log p_\theta(\boldsymbol{y}, \boldsymbol{x}^{(k)}). \tag{7}$$

see [9] for further details.

It is natural to identify the observations $\boldsymbol{y}_1,\ \ldots, \boldsymbol{y}_N$ as $\boldsymbol{y}$, and the process at all intermediate time points as $\boldsymbol{x} = \boldsymbol{y}_{\tau_{(n-1)J+1}:\tau_{nJ-1}}, n = 1, \ldots, N$, cf. [19]. The intermediate quantity $Q_m(\theta, \theta_m)$ can then be written as

$$Q_m(\theta, \theta_m) = \frac{1}{K_m}\sum_{k=1}^{K_m}\sum_{r=1}^{(N-1)J} \log p_\theta(\boldsymbol{y}_{\tau_{r+1}}^{(k)}|\boldsymbol{y}_{\tau_r}^{(k)}). \tag{8}$$

This is a (non-linear) sum of quadratic terms as $p_\theta(\boldsymbol{y}_{\tau_{r+1}}^{(k)}|\boldsymbol{y}_{\tau_r}^{(k)})$ is approximately Gaussian for diffusions. The expression is unfortunately more complex when considering Lévy-driven SDEs.

1) Computation of the smoothing distribution: We are interested in the smoothing distribution

$$p_{\theta_m}(\boldsymbol{y}_{\tau_{(n-1)J+1}:\tau_{nJ-1}}|\boldsymbol{y}_{t_n},\boldsymbol{y}_{t_{n+1}}) \tag{9}$$

in order to compute expectations. Let $\phi(\cdot)$ be a test function. The expectation of the test function with respect to the smoothing distribution is given by

$$\boldsymbol{E}_{\theta_m}[\phi(\boldsymbol{y}_{\tau_{(n-1)J+1}:\tau_{nJ-1}})|\boldsymbol{y}_{t_n},\boldsymbol{y}_{t_{n+1}}] \tag{10}$$

$$= \int \phi(\boldsymbol{y}_{\tau_{(n-1)J+1}:\tau_{nJ-1}})\frac{p_{\theta_m}(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{y}_{\tau_{nJ-1}})}{p_{\theta_m}(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{y}_{t_n})} \tag{11}$$

$$\cdot\ p_{\theta_m}(\boldsymbol{y}_{\tau_{(n-1)J+1}:\tau_{nJ-1}}|\boldsymbol{y}_{t_n})\mathrm{d}\boldsymbol{y}_{\tau_{(\mathrm{n}-1)\mathrm{J}+1}:\tau_{\mathrm{nJ}-1}}. \tag{12}$$

Sampling $\boldsymbol{u}^{(k)} \equiv \boldsymbol{u}_{\tau_{(n-1)J+1}:\tau_{nJ-1}},\ k = 1,\dots,K_m$ from $p_{\theta_m}(\boldsymbol{y}_{\tau_{(n-1)J+1}:\tau_{nJ-1}}|\boldsymbol{y}_{t_n})$ and approximating the expectation leads to

$$\approx \frac{1}{K_m}\sum_{k=1}^{K_m}\phi(\boldsymbol{u}^{(k)})\frac{p_{\theta_m}(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{u}^{(k)})}{\sum_i p_{\theta_m}(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{u}^{(i)})}. \tag{13}$$

The empirical smoothing distribution is therefore a weighted version of empirical predictive distribution.

$$p_{\theta_m,K_m}(\boldsymbol{u}) = \sum_{k=1}^{K_m}\frac{p_{\theta_m}(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{u}^{(k)})}{\sum_i p_{\theta_m}(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{u}^{(i)})}\delta(\boldsymbol{u}-\boldsymbol{u}^{(k)}) \tag{14}$$

$$= \sum_{k=1}^{K_m} w_k\delta(\boldsymbol{u}-\boldsymbol{u}^{(k)}) \tag{15}$$

where $\delta(\boldsymbol{u}-\boldsymbol{u}^{(k)})$ is a delta-Dirac measure centered at $\boldsymbol{u}^{(k)}$ and $\sum w_l = 1$. It is straightforward to introduce an importance sampler. We refer to [13], [29] for more information on these samplers. However, it should be pointed out that importance samplers can increase the variance, rather than decrease it, cf. [26] and that it may be difficult to ensure that the conditions needed for variance reduction hold in practice.

It is well known from [13] that most of the mass in the distribution is concentrated to a small number of samples - this is the reason why the ordinary Pedersen sampler is so inefficient. We avoid this by resampling from the distribution, generating a new sample with equal weights with $K'_m \ll K_m$ elements.

$$\tilde{p}_{\theta_m,K'_m}(\boldsymbol{u}) = \frac{1}{K'_m}\sum_{k=1}^{K'_m}\delta(\boldsymbol{u}-\tilde{\boldsymbol{u}}^{(k)}). \tag{16}$$

It is common within the particle filter community to compute the effective sample size (ESS) as a measure of how many equally weighted elements the unequally weighted sample corresponds to. The ESS is defined as

$$ESS = \frac{(\sum_k w_k)^2}{\sum_k w_k^2}. \tag{17}$$

We have used the ESS as a guideline for selecting K'_m, as there is little to gain from resampling additional elements.

The proposed MCEM-algorithm is presented in Algorithm 1.

Algorithm 1 Pseudo code for the MCEM-algorithm.

```
Initiate $\theta_1$
for $m = 1 : (M-1)$ do
  % E-step
  for $n = 1 : (N-1)$ do
    for $k = 1 : K_m$ do
      $\boldsymbol{u}^{(k)}_{\tau_{(n-1)J}} = \boldsymbol{y}_{t_n}$
      for $j = 1 : (J-1)$ do
        $\boldsymbol{x}^{(k)}_{\tau_{(n-1)J+j}} \sim p_{\theta_m}(\cdot|\boldsymbol{u}^{(k)}_{\tau_{(n-1)J+j-1}})$
      end for
      Compute $\tilde{w}_k = p_{\theta_m}(\boldsymbol{y}_{t_{n+1}}|\boldsymbol{u}^{(k)}_{\tau_{nJ-1}})$
    end for
    Compute $w_k = \frac{\tilde{w}_k}{\sum_l \tilde{w}_l}$
    % Compute the smoothing distribution
    for $k = 1 : K'_m$ do
      Sample $I_k \sim P(I_k = j) = w_j$
      Set $\tilde{\boldsymbol{x}}^{(k)}_{\tau_{(n-1)J+1}:\tau_{nJ-1}} = \boldsymbol{u}^{(I_k)}_{\tau_{((n-1)J+1)}:\tau_{nJ-1}}$
    end for
  end for
  % M-step
  Compute $\theta_{m+1} = \arg\max Q_m(\theta,\theta_m;)$ using Equation (8) and $\tilde{x}$.
end for
```

2) Complexity of the algorithm: Let us compare the computational complexity of the algorithm with the computational complexity of the Pedersen method as they are similar (no variance reduction etc.). Both algorithms are being applied to a data set with N observations, with J substeps taken to reduce the bias of the discretization. The cost for computing a single estimate of the likelihood function using the Pedersen method and for computing the smoothing distribution is the same ,see Table I.

	Pedersen	Proposed MCEM
Comp. smooth	-	KNJ
Comp. loss fcn	KNJ	$K'NJ$
Optim. cost	$R_{Ped}KNJ$	$M(KNJ+$ $R_{EM}K'NJ)$

TABLE I

COMPLEXITY OF THE PEDERSEN METHOD WHEN COMMON RANDOM NUMBERS CAN BE USED, AND THE PROPOSED MCEM ALGORITHM.

The difference is that the optimization in the MCEM uses far less samples ($K' \ll K$) and few iterations ($M \ll R_{Ped}$). Additionally, the number of iterations needed during the M-step, R_{EM}, is often quite small ($R_{EM} \ll R_{Ped}$), and it is possible to compute the estimate in closed form for a large class of processes when the distributions belong to the exponential family, cf. [9]. That would lead to $R_{EM} = 1$.

III. Simulations

We evaluate the proposed estimator on simulated data, which eliminates the problem of model mis-specification.

We also compare the estimator to the series expansion estimators in [4] when applicable. That paper use daily returns from the S & P 500 from January 2nd, 1990, to September 30th, 2003, and also VIX data from the Chicago Board of Options Exchange (CBOE) which is a measure of the volatility computed from option prices for the corresponding time period. The S & P 500 and VIX data from 1990 to 2011 are presented in Figure 1.

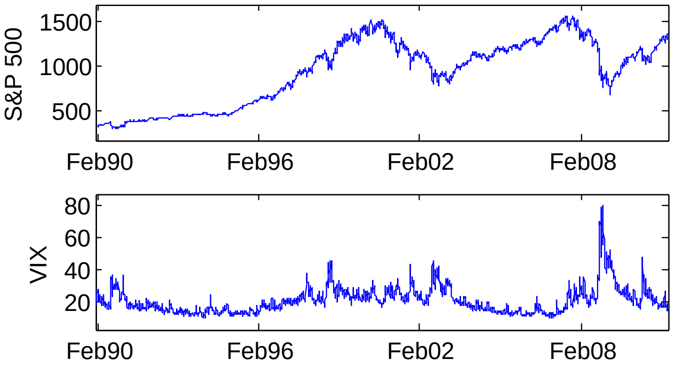

Fig. 1. S & P 500 (top) and VIX (bottom) from January 1990 to April 2011.

A. *Diffusions*

We have studied the Cox-Ingersoll-Ross (CIR) model (for which the transition kernel is known, see [12]) and the Heston model, see [21], which is probably the most popular continuous time stochastic volatility model.

1) Cox-Ingersoll-Ross: The CIR model, see [12], is defined by the SDE

$$\mathrm{dY_t} = \kappa(\theta - \mathrm{Y_t})\mathrm{dt} + \sigma\sqrt{\mathrm{Y_t}}\mathrm{dW_t}. \tag{18}$$

The model is commonly used to model positive, mean reverting quantities such as interest rates, see [12], energy price, see [31], [40] and unobserved volatility, see [21]. The need for time varying volatility is obvious when studying Figure 1.

It is known that the transition kernel for the CIR model is a non-central χ^2 distribution, see [28]. The MCEM algorithm was compared to the MLE for a large data set, corresponding to 20 years of daily data. The results are presented in Table II.

Iter. no	κ	θ	σ
Start	4.5	.04	.5
1	6.1037	0.0562	0.4537
2	5.5002	0.0561	0.4292
3	5.1590	0.0561	0.4165
4	5.0126	0.0561	0.4100
5	4.9292	0.0561	0.4066
10	4.8431	0.0561	0.4032
20	4.8057	0.0561	0.4019
30	4.8065	0.0561	0.4022
MLE	4.7969	0.0561	0.4047

TABLE II

CONVERGENCE OF THE MCEM ALGORITHM FOR THE CIR PROCESS. THE FIRST 15 ITERATIONS USED $K = 200,\ K' = 30$ WHILE THE LAST 15 USED $K = 500,\ K' = 100$.

Convergence to the MLE is rapid, as the difference is negligible after a few iterations.

2) Heston: We proceeded by applying the estimator to the Heston model, [21]. The model is given by

$$\mathrm{dX_t} = \left(\mu - \frac{V_t}{2}\right)\mathrm{dt} + \sqrt{\mathrm{V_t}}\mathrm{dW_t^{(S)}}, \tag{19}$$

$$\mathrm{dV_t} = \kappa(\theta - V_t)\mathrm{dt} + \sigma_\mathrm{V}\sqrt{\mathrm{V_t}}\mathrm{dW_t^{(V)}}. \tag{20}$$

The μ parameter is notoriously difficult to estimate, which is why it has been set to $\mu = 0.05$ throughout the paper. We examined the model by simulating 3000 daily observations (again corresponding to 20 years of daily data), and fitting the model using the proposed algorithm. The results are presented in Table III, again showing satisfactory results.

Iter. no	κ	θ	σ	ρ
Start	3.5	0.05	.5	-.5
1	11.1297	0.0560	0.3914	-0.5353
2	7.9429	0.0557	0.3261	-0.5789
3	6.3109	0.0555	0.2877	-0.6229
4	5.3994	0.0553	0.2658	-0.6567
5	4.9740	0.0552	0.2543	-0.6778
10	4.5519	0.0550	0.2421	-0.7024
20	4.5211	0.0550	0.2413	-0.7024
25	4.5260	0.0550	0.2413	-0.7026
30	4.5203	0.0550	0.2413	-0.7026
True	4	0.05	.25	-.7

TABLE III

CONVERGENCE OF THE MCEM ALGORITHM FOR THE SIMULATED HESTON PROCESS. THE FIRST 15 ITERATIONS USED $K = 200,\ K' = 80$ WHILE THE LAST 15 USED $K = 1000,\ K' = 200$.

The parameters were also estimated using data from S & P 500 and CBOE (VIX), cf. [4]. It is possible to find approximate parameters by fitting a CIR model to VIX data. These estimates, along with the estimates for the full model, are presented in Table IV. These can be compared to the (AS-K) estimates in [4] that uses the same data but another approximation to compute the MLE.

Iter. no	κ	θ	σ	ρ
CIR	5.69	0.0455	0.407	-
Start	3.5	0.05	.5	-.5
1	3.0017	0.0484	0.3285	-0.5196
2	3.3950	0.0483	0.3499	-0.5365
3	3.6776	0.0482	0.3662	-0.5523
4	3.9303	0.0482	0.3851	-0.5683
5	4.2364	0.0481	0.3997	-0.5826
10	4.8232	0.0483	0.4457	-0.6422
20	5.0307	0.0487	0.4809	-0.7090
30	4.9674	0.0489	0.4921	-0.7337
AS-K	5.07	0.0457	0.48	-0.767

TABLE IV

CONVERGENCE OF THE MCEM ALGORITHM FOR THE S&P 500 AND VIX DATA. THE FIRST 15 ITERATIONS USED $K = 500,\ K' = 50$ WHILE THE LAST 15 USED $K = 1000,\ K' = 200$. AS-K ARE THE ESTIMATES FROM [4].

The estimates from the [4] paper and MCEM are similar, and both are different from estimating the (κ, θ, σ) parameters using only the V process, indicating that the CIR estimates should primarily be considered as starting values for estimation of the full model.

B. Jump-Diffusion

A popular extension of the Heston model is the Bates model, see [6]. The model is defined as

$$\mathrm{dX_t} = \left(\mu - \frac{V_t}{2}\right)\mathrm{dt} + \sqrt{\mathrm{V_t}}\mathrm{dW}_\mathrm{t}^{(\mathrm{S})} + \mathrm{dL_t}, \tag{21}$$

$$\mathrm{dV_t} = \kappa(\theta - V_t)\mathrm{dt} + \sigma_\mathrm{V}\sqrt{\mathrm{V_t}}\mathrm{dW}_\mathrm{t}^{(\mathrm{V})} \tag{22}$$

where L_t is a compound Poisson process with jumps arriving with intensity λ and the jumps are Gaussian with mean μ_{Jump} and variance σ^2_{Jump}. We expect the estimates to be similar to the Heston estimates, but with a smaller latent volatility as the jump component will capture some of the dynamics.

The parameter estimates, when using the same data as for the Heston model and computing the median over all iterations (the estimates are rather noisy), cf. [9], are presented in Figure 2.

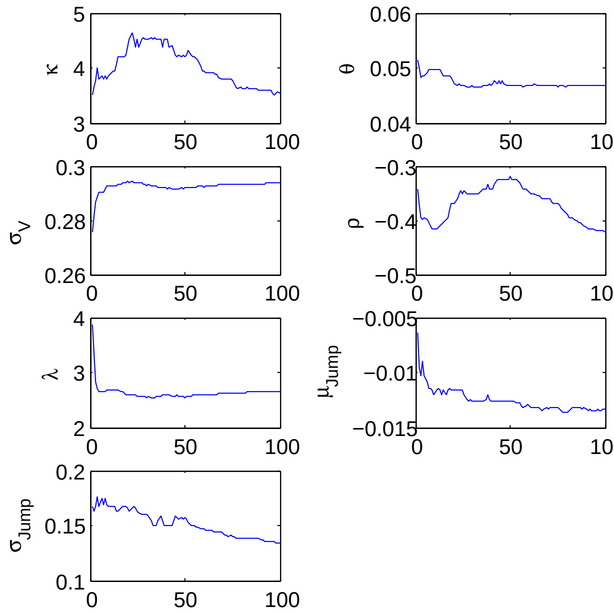

Fig. 2. Median (computed over iterations) of the estimates for the Bates model when applied to S&P 500 and VIX.

The parameters are similar to the Heston estimates, but adjusted for the contribution from the jump component.

C. Levy-driven SDEs

An increasingly popular generalization of diffusion processes and jump-diffusions are Levy-driven SDEs, i.e. stochastic differential equations driven by a general white noise process. These are particularly popular in finance, see [11], [32]. We start with an exponential Lévy process.

1) NIG: One of the most popular Lévy processes is the Normal Inverse Gaussian (NIG) process, see [5]. The process is a subordinated Brownian motion. Formally, let u be a random variable that follows an inversion Gaussian probability law

$$u \sim IG(\delta, \sqrt{\alpha^2 + \beta^2}). \tag{23}$$

Furthermore, assume that w conditional on u is normally distributed with mean $\mu + \beta u$ and variance u

$$v|u \sim N(\mu + \beta u, u). \tag{24}$$

It can then be shown that the unconditional density for w is given by

$$v \sim NIG(\alpha, \beta, \mu, \delta). \tag{25}$$

The α parameter controls the tails, the β parameter the asymmetry, the μ parameter is a location parameter and δ parameter is a scale parameter. A very convenient property of the NIG distribution is that it is closed under convolution. Let $v_1 \sim NIG(\alpha, \beta, \mu_1, \delta_1)$ and $v_2 \sim NIG(\alpha, \beta, \mu_2, \delta_2)$. It can then be shown the NIG distribution is closed under convolution, i.e. that $v_1 + v_2 \sim NIG(\alpha, \beta, \mu_1 + \mu_2, \delta_1 + \delta_2)$.

The NIG process is a generalization of the NIG-distribution, cf. [5]. The process is used to model stocks as an exponential NIG process

$$S(t) = \exp(L(t)) \tag{26}$$

where $L(t)$ is a NIG process.

The parameters governing the NIG process can be estimated using moment matching, see [17] or Maximum likelihood estimation. We have compared these methods to Algorithm 1 in a simulation study. The parameters that we used in the simulation were estimated from the S & P 500 data that were used in Section III-A and III-B.

We have simulated 10 independent realizations, each consisting of 20 years of weekly data using the MLE from the S & P 500 data. The results are presented in Figure 3 (the estimates), Figure 4 (estimates minus the true parameters) and Figure 5 (estimates minus the MLE). The simulations were started from the Moment matching estimates and used $K_m = 100(1 + \sqrt{m}))$ and $K'_m = 20(1 + \sqrt{m}))$ samples when approximating the expectation in intermediate quantity.

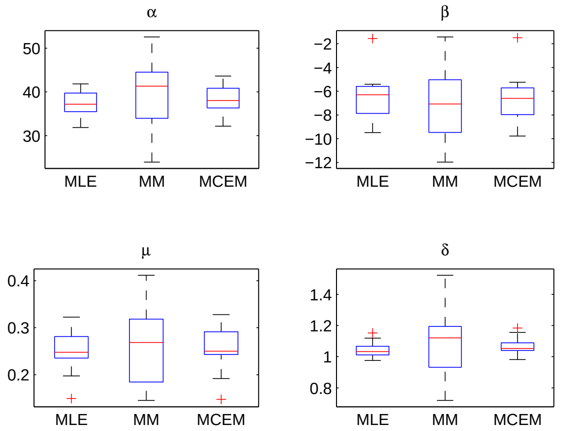

Fig. 3. Parameter estimates for the NIG model when using MLE, Moment matching (MM) and Monte Carlo EM (MCEM) in this paper.

There is little doubt that MLE or the proposed Monte Carlo EM algorithm is generating better estimates that the moment matching estimator. This result is consistent with the

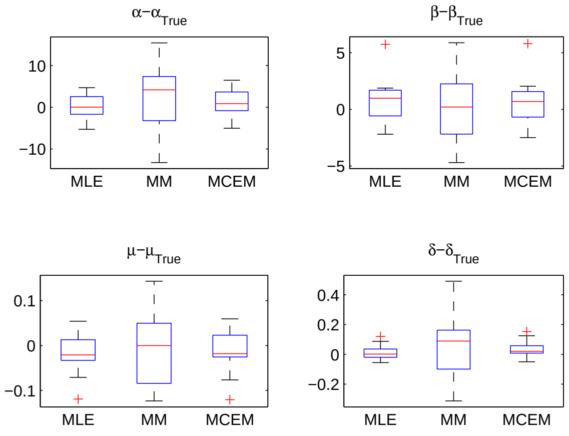

Fig. 4. Parameter estimates for the NIG model minus the true parameters when using MLE, Moment matching (MM) and Monte Carlo EM (MCEM).

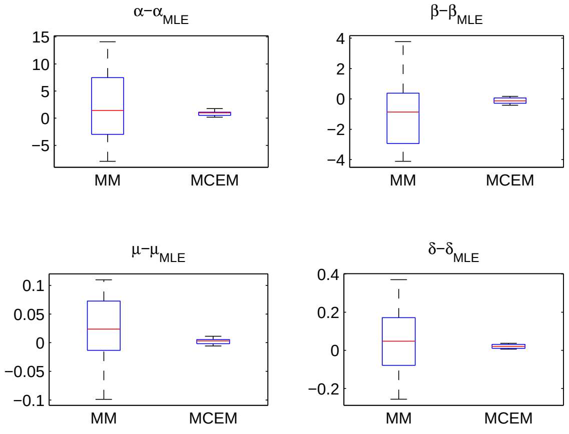

Fig. 5. Parameter estimates for the NIG model minus the MLE when using Moment matching (MM) and Monte Carlo EM (MCEM).

Cramer-Rao inequality. We have also studied the convergence of the Monte Carlo EM algorithm. The results are presented in Figure 6.

The estimates from the MCEM algorithm is converging nicely for all realizations.

2) NIG-CIR: It is by now well known that stochastic volatility is needed when modelling financial assets, cf. [21], [11]. A simple extension of the NIG model is the NIG-CIR model, see [10], that introduces a stochastic time shift to the model. Let V be a CIR-process independent of the NIG-process. The model is then defined as

$$Z(t) = NIG(I(t)) \tag{27}$$

where $I(t) = \int_0^t V_s \mathrm{d}s$. That means that the distribution of Z, conditional on the I process, is a NIG-distributed process. We will exploit this fact in our algorithm. Furthermore, it follows

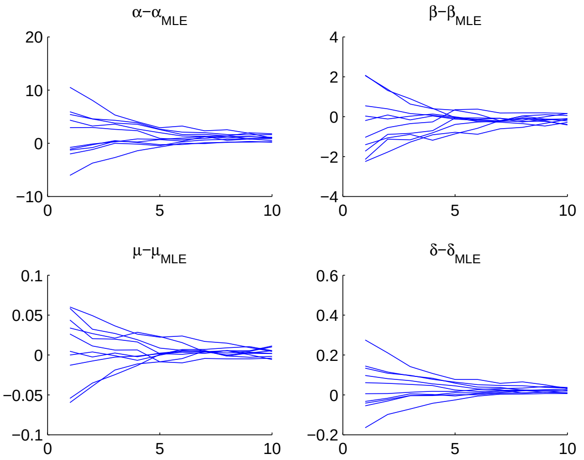

Fig. 6. Convergence of the MCEM parameter estimates minus the MLE for the NIG model.

that

$$\begin{aligned} &Z(t_{n+1}) - Z(t_n) = NIG(I(t_{n+1})) - NIG(I(t_n)) \qquad (28)\\ &\overset{d}{=} NIG(I(t_{n+1}) - I(t_n)) = NIG\left(\int_{t_n}^{t_{n+1}} V_s \mathrm{d}s\right) \end{aligned}$$

from the fact that the NIG process is a process having independent increments. The V process can easily be recovered from option prices, cf. [32] while the relationship with the VIX is more complicated.

The straightforward implementation of the Pedersen sampler for this model is given by

$$\begin{aligned} &p\left(Z_{t_{n+1}}, V_{t_{n+1}} | Z_{t_n}, V_{t_n}\right) = &(29)\\ &= \int p\left(Z_{t_{n+1}}, V_{t_{n+1}}, V_\tau | Z_{t_n}, V_{t_n}\right) \mathrm{d}V_\tau &(30)\\ &= \int p\left(Z_{t_{n+1}} | V_{t_{n+1}}, V_\tau, V_{t_n}, Z_{t_n}\right) \cdot &(31)\\ &p\left(V_{t_{n+1}} | V_\tau\right) p\left(V_\tau | V_{t_n}\right) \mathrm{d}V_\tau &\\ &\approx \int p\left(Z_{t_{n+1}} | I(t_{n+1}) - I(t_n), Z_{t_n}\right) \cdot &(32)\\ &p\left(V_{t_{n+1}} | V_\tau\right) p\left(V_\tau | V_{t_n}\right) \mathrm{d}V_\tau & \end{aligned}$$

while a simple extension would be to implement a Durham-Gallant type importance sampler for the hidden V process (the process is a diffusion, indicating that the regularity conditions are likely to hold).

The smoothing distribution can be derived by computing the expectation of a test function $f(V_\tau)$ with respect to the smoothing distribution.

$$
\begin{aligned}
&E[f(V_\tau)|V_{t_n}, V_{t_{n+1}}, X_{t_n}, X_{t_{n+1}}] = && (33)\\
&= \int f(V_\tau) p(V_\tau | V_{t_n} V_{t_{n+1}}, X_{t_n}, X_{t_{n+1}}) \mathrm{d}V_\tau && (34)\\
&= \int f(V_\tau) \frac{p(X_{t_{n+1}}, V_{t_{n+1}}, V_\tau | V_{t_n}, X_{t_n})}{p(X_{t_{n+1}}, V_{t_{n+1}} | V_{t_n}, X_{t_n})} \mathrm{d}V_\tau && (35)\\
&= \int f(V_\tau) \frac{p(X_{t_{n+1}} | V_{t_{n+1}}, V_\tau, V_{t_n}, X_{t_n})}{p(X_{t_{n+1}}, V_{t_{n+1}} | V_{t_n}, X_{t_n})} \cdot && (36)\\
&\cdot \frac{p(V_{t_{n+1}} | V_\tau) p(V_\tau | V_{t_n})}{q(V_\tau)} q(V_\tau) \mathrm{d}V_\tau
\end{aligned}
$$

where we introduced an importance sampler for the V process. The V process is a pure diffusion process with known initial and final values, implying that the Durham-Gallant sampler is a good choice of sampler. Evaluating the integral using samples $V_\tau^{(k)}$ from the importance sampler $q(V_\tau)$ results in

$$
E[f(V_\tau)|V_{t_n}, V_{t_{n+1}}, X_{t_n}, X_{t_{n+1}}] \approx \sum_{k=1}^{K} f(V_\tau^{(k)}) w_k \qquad (37)
$$

This holds for any function $f(V)$, including $f(V) = 1$. This is used to derive the self-normalized importance sampling weights as

$$
\begin{aligned}
\tilde{w}_k &= \frac{p(X_{t_{n+1}} | \Delta I^{(k)}, X_{t_n}) p(V_{t_{n+1}} | V_\tau^{(k)}) p(V_\tau^{(k)} | V_{t_n})}{K p(X_{t_{n+1}}, V_{t_{n+1}} | V_{t_n}, X_{t_n}) q(V_\tau^{(k)})}, && (38)\\
w_k &= \frac{\tilde{w}_k}{\sum_l \tilde{w}_l}. && (39)
\end{aligned}
$$

where $\Delta I^{(k)}$ is an approximation of $\int_{t_n}^{t_{n+1}} V_s \mathrm{d}s$ using $V_{t_{n+1}}, V_\tau^{(k)}$ and V_{t_n}. The corresponding intermediate quantity is given by

$$
\begin{aligned}
Q_m(\theta, \theta_m) = \frac{1}{K_m} \Bigg(&\sum_{k=1}^{K_m} \sum_{r=1}^{(N-1)J} \log p_\theta (V_{\tau_{r+1}}^{(k)} | V_{\tau_r}^{(k)}) && (40)\\
&+ \sum_{n=1}^{N-1} \log p_\theta (X_{t_{n+1}} | X_{t_n}, I(t_{n+1}) - I(t_n)) \Bigg).
\end{aligned}
$$

The connection between the VIX and the time shift process is rather complex (it depends on both the historical *and* the risk neutral measure), it is possible to estimate the hidden time shift process directly using the sequential calibration method in [32] but that approach requires observations from several options at each time point in addition to the returns. We have therefore used simulated data to evaluate the MCEM estimator for NIG-CIR model.

The parameters in our simulation study was chosen as a combination of the parameters estimated in Section III-C.1 and [32]. The parameters used are $\kappa = 10$, $\theta = 0.1$, $\sigma = 1$, $\alpha = 30$, $\beta = -7$, $\mu = 1$ and $\delta = 4$. The data corresponds to 10 years of monthly data. The MCEM algorithm was run using $J = 10$ subintervals and approximating the expectation with $K_m = 100(1 + \sqrt{m})$ and $K_m' = 20(1 + \sqrt{m})$ samples. The MCEM algorithm used a Durham-Gallant sampler for the V process in order to reduce the variance of the approximation of the intermediate quantity.

We have compared the MCEM estimates to other estimators. It is possible to compute the MLE for the κ, θ and σ parameters as the likelihood function is known in closed form. We refer to this estimator as the exact estimator, in-spite knowing that it is not based on the full data set and therefore not maximum likelihood estimator for the full model. The other comparison we use is the Discrete Maximum Likelihood (DML) estimator, cf [22], approximating the transition probabilities with a single Euler step. The parameter estimates for all methods are presented in Figure 7.

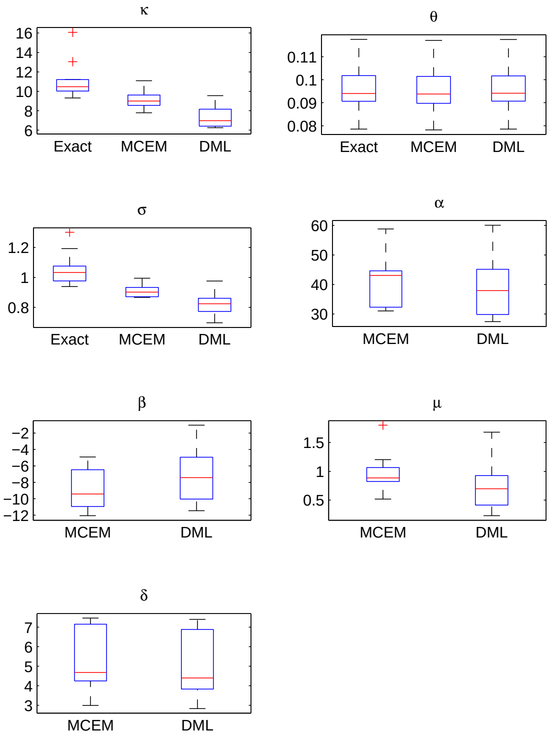

Fig. 7. Parameter estimates generated by the MLE (only CIR parameters), MCEM and DML methods.

Subtracting the true parameters from the estimates gives an indication of the quality (unbiased etc) of the estimates. This is presented in Figure 8.

We can see that some of the DML estimates are biased (e.g. κ and σ), while the Exact and MCEM are unbiased or nearly unbiased (we expect to bias in the MCEM to be much smaller than the DML estimates considering that many intermediate time steps ($J = 10$) were used). Another difference between the MCEM and DML estimator is that the variance of the MCEM estimates is lower than the variance of the DML estimates.

IV. Conclusion

We have introduced a Monte Carlo EM-algorithm for maximum likelihood estimation for diffusions, jump-diffusions and Lévy driven stochastic differential equations. The algorithm

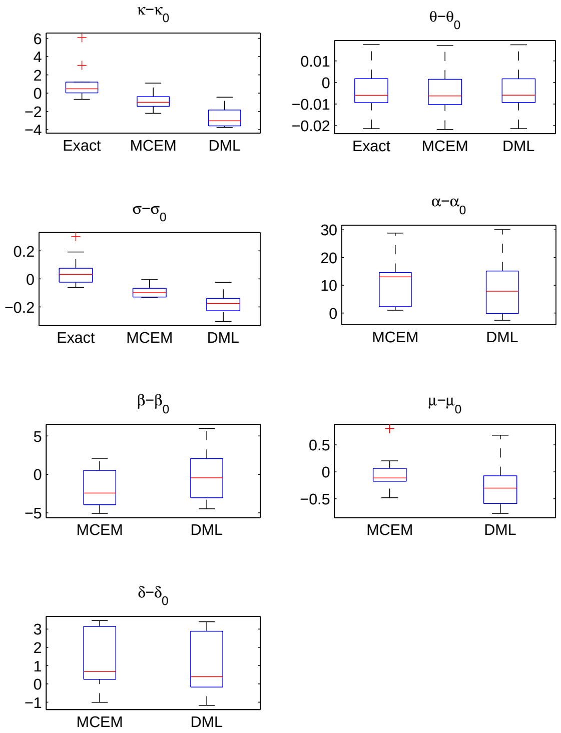

Fig. 8. Difference between the parameter estimates generated by the MLE(CIR), MCEM and DML methods and the true parameters used when generating the data.

is simple, only marginally for complex than the Pedersen method, [37], general (applicable for scalar and multivariate discretely observed models) and computationally efficient compared to standard methods, cf. Section II-A.2.

The algorithm can be used to estimate parameters in complex models, beyond what the financial industry is currently using, and it will be useful when estimating parameters governing the historical probability needed for financial risk management.

Acknowledgement

We are grateful to Philipp Nowak, Magnus Wiktorsson and the participants of the ACMOS' 12 conference for valuable comments on the algorithm.

References

[1] Y. Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1):223–262, 2002.

[2] Y. Ait-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906–937, 2008.

[3] Y. Aït-Sahalia and Yu Jialin. Saddlepoint approximations for continuous-time markov processes. *Journal of Econometrics*, 134(2):507–551, 2006.

[4] Y. Aıt-Sahalia and R. Kimmel. Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics*, 83:413–452, 2007.

[5] O.E. Barndorff-Nielsen. Processes of normal inverse gaussian type. *Finance and stochastics*, 2(1):41–68, 1997.

[6] D.S. Bates. Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *Review of financial studies*, 9(1):69–107, 1996.

[7] D.S. Bates. Maximum likelihood estimation of latent affine processes. *Review of Financial Studies*, 19(3):909–965, 2006.

[8] A. Beskos, O. Papaspiliopoulos, and G. Roberts. Monte carlo maximum likelihood estimation for discretely observed diffusion processes. *The Annals of Statistics*, 37(1):223–245, 2009.

[9] O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Verlag, 2005.

[10] P. Carr, H. Geman, D.B. Madan, and M. Yor. Stochastic volatility for Lévy processes. *Mathematical Finance*, 13(3):345–382, 2003.

[11] R. Cont and P. Tankov. *Financial modelling with jump processes*, volume 2. Chapman & Hall, 2004.

[12] J.C. Cox, J.E. Ingersoll Jr, and S.A. Ross. A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, pages 385–407, 1985.

[13] G.B. Durham and A.R. Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business and Economic Statistics*, 20(3):297–338, 2002.

[14] O. Elerian, S. Chib, and N. Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993, 2001.

[15] B. Eraker. MCMC analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19(2):177–191, 2001.

[16] B. Eraker. Do stock prices and volatility jump? reconciling evidence from spot and option prices. *The Journal of Finance*, 59(3):1367–1404, 2004.

[17] A. Eriksson, E. Ghysels, and F. Wang. The normal inverse gaussian distribution and the pricing of derivatives. *The Journal of Derivatives*, 16(3):23–37, 2009.

[18] P. Fearnhead. Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Statistics and Computing*, 18(2):151–171, 2008.

[19] A. Golightly and DJ Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693, 2008.

[20] O. Hellquist, E. Lindström, and J. Ströjby. Likelihood Inference in Jump Diffusion driven SDE's. In *Symposium i anvendt Statistik*, volume 32, pages 269–278, 2010. ISBN: 978-87-501-1832-9.

[21] S.L. Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of financial studies*, 6(2):327–343, 1993.

[22] AS Hurn, JI Jeisman, and KA Lindsay. Seeing the wood for the trees: a critical evaluation of methods to estimate the parameters of stochastic differential equations. *Journal of Financial Econometrics*, 5(3):390–455, 2007.

[23] M.S. Johannes, N.G. Polson, and J.R. Stroud. Optimal filtering of jump diffusions: Extracting latent states from asset prices. *Review of Financial Studies*, 22(7):2759–2799, 2009.

[24] A.A. KELLER. Population biology models with time-delay in a noisy environment. *WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE*, 8:113–134, 2011.

[25] K. Kiriakopoulos and G. Kaimakamis. Optimal interest rate derivatives portfolio with constrained greeks-a stochastic control approach. In *Proceedings of the 9th WSEAS international conference on Simulation, modelling and optimization*, pages 217–222. World Scientific and Engineering Academy and Society (WSEAS), 2009.

[26] S.J. Koopman, N. Shephard, and D. Creal. Testing the assumptions behind importance sampling. *Journal of Econometrics*, 149(1):2–11, 2009.

[27] K. Laas, R. Mankin, and E. Reiter. Influence of memory time on the resonant behavior of an oscillatory system described by a generalized Langevin equation. *INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES*, 5:280–289, 2011.

[28] E. Lindström. Estimating parameters in diffusion processes using an approximate maximum likelihood approach. *Annals of Operations Research*, 151(1):269–288, 2007.

[29] E. Lindström. A regularized bridge sampler for sparsely sampled diffusions. *Statistics and Computing*, pages 1–9, 2011.

[30] E. Lindström. Inference for Non-Linear Diffusions and Jump-Diffusions: A Monte Carlo EM approach. In *Recent Researches in Automatic Control and Electronics*, volume 14 of *International Conference on Automatic Control, Modelling and Simulation (ACMOS)*, pages 110–115. WSEAS Press, 2012.

[31] E. Lindström and F. Regland. Modelling extreme dependence between european electricity markets. *Energy Economics*, 2012. DOI:10.1016/j.eneco.2012.04.006.

[32] E. Lindström, J. Ströjby, M. Brodén, M. Wiktorsson, and J. Holst. Sequential calibration of options. *Computational Statistics & Data Analysis*, 52(6):2877–2891, 2008.

[33] A.W. Lo. Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data. *Econometric Theory*, 4(02):231–247, 1988.

[34] Holst J. Madsen, H. and E. Lindström. Modelling non-linear and non-stationary time series. *Lecture Notes, Technical University of Denmark, Dpt. of Informatics and Mathematical Modeling, Kgs. Lyngby, Denmark*, 2010.

[35] G. Mircea, M. Neamţu, and D. Opriş. Deterministic, uncertainty and stochastic models of kaldor-kalecki model of business cycles. *WSEAS Transactions on Mathematics*, 9(8):638–647, 2010.

[36] A.R. Pedersen. Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, pages 257–279, 1995.

[37] A.R. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, pages 55–71, 1995.

[38] M.W. Pedersen, U.H. Thygesen, and H. Madsen. Nonlinear tracking in a diffusion process with a bayesian filter and the finite element method. *Computational Statistics & Data Analysis*, 55(1):280–290, 2011.

[39] D. Prathumwan, Y. Lenbury, P. Satiracoo, and C. Rattanakul. Euler-maruyama approximation and maximum likelihood estimatorfor a stochastic differential equation model of the signal transduction process. *INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES*, 6:323–331, 2012.

[40] F. Regland and E. Lindström. Independent spike models: Estimation and validation. *Czech Journal of Economics and Finance*, 62(2), 2012.

[41] N. Sîrghi, M. Neamu, and D. Opri. Effects of changes in some parameters on the deterministic and stochastic dynamic economic model with wealth and human capital accumulation. *INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES*, 6:63–71, 2012.

[42] O. Stramer and M. Bognar. Bayesian inference for irreducible diffusion processes using the pseudo-marginal approach. *Bayesian Analysis*, 6:231–258, 2011.

[43] O. Stramer and J. Yan. On simulated likelihood of discretely observed diffusion processes and comparison to closed-form approximation. *Journal of Computational and Graphical Statistics*, 16(3):672–691, 2007.