LUND UNIVERSITY

# Multilevel Monte Carlo methods for inference in multivariate diffusions

## Bayes @ Lund

Erik Lindström (joint work with my PhD student, Carl Åkerlindh)

Centre for Mathematical Sciences


# Outline - Title in reverse order

Introduction - Diffusion processes

Multilevel Monte Carlo

Analysis

Simulation study

References

# Bayesian angle

- Model selection is (increasingly) difficult!
- Introduce priors on models, not only parameters!

# Bayesian angle

- Model selection is (increasingly) difficult!
- Introduce priors on models, not only parameters!

**My view is that we should focus on physically plausible models, rather than considering all possible black box models!**

Prior information is often codes through differential relations classically described through ODEs and/or PDEs. Examples where stochastic generalizations have been successful includes

- Ebola outbreak - generalized SIR model, King et al. (2015)
- Energy efficiency in houses - thermodynamics, Bacher et al. (2013)
- Spread of Malaria, Bhadra et al. (2011)
- Wind speed forecasting - physics, Iversen et al. (2016)
- Glucose-insulin-glucagon pharmacodynamics modelling, Wendt et al. (2017)

# Stochastic differential equations

The formally correct generalization of

$$\textit{ODE} + \textit{White noise} \tag{1}$$

# Stochastic differential equations

The formally correct generalization of

$$\textit{ODE} + \textit{White noise} \tag{1}$$

is stochastic differential equations (SDEs)

$$\mathrm{d}X(t) = \mu_\theta(t, X(t))\mathrm{d}t + \sigma_\theta(t, X(t))\mathrm{d}W(t). \tag{2}$$

# Stochastic differential equations

The formally correct generalization of

$$\textit{ODE} + \textit{White noise} \tag{1}$$

is stochastic differential equations (SDEs)

$$\mathrm{d}X(t) = \mu_\theta(t, X(t))\mathrm{d}t + \sigma_\theta(t, X(t))\mathrm{d}W(t). \tag{2}$$

SDEs are Markov processes.

# Stochastic differential equations

The formally correct generalization of

$$\textit{ODE} + \textit{White noise} \tag{1}$$

is stochastic differential equations (SDEs)

$$\mathrm{d}X(t) = \mu_\theta(t, X(t))\mathrm{d}t + \sigma_\theta(t, X(t))\mathrm{d}W(t). \tag{2}$$

SDEs are Markov processes.

Interpretation of $\mu$ and $\sigma$:

$$\mu_\theta(t, X_t) = \lim_{h \to 0} \frac{1}{h} \mathbf{E}\left[X_{t+h} - X_t | \mathcal{F}_t\right] \tag{3}$$

$$\sigma_\theta(t, X_t)\sigma_\theta(t, X_t)^T = \lim_{h \to 0} \frac{1}{h} \mathbf{Var}\left[X_{t+h} - X_t | \mathcal{F}_t\right] \tag{4}$$

# Bayesian inference

Posterior distribution is given by

$$p(\theta|\vec{X}) \propto p(\vec{X}|\theta)p(\theta) \tag{5}$$

Problem is that $p(\vec{X}|\theta)$ is rarely given in closed form.

# Bayesian inference

Posterior distribution is given by

$$p(\theta|\vec{X}) \propto p(\vec{X}|\theta)p(\theta) \tag{5}$$

Problem is that $p(\vec{X}|\theta)$ is rarely given in closed form.

The transition kernel $p(\vec{X}|\theta) = \prod_{n=1}^{N} p_\theta(x_{t_n}|x_{t_{n-1}})$ can be approximated using

- PDE methods (Fokker-Planck)
- Monte Carlo
  - Pedersen sampler (1995)
  - Bridge sampler (2002)
  - Lindström sampler (2012)
  - Residual sampler (2016)

# Computing the transition density

Combine the *law of total probability* with Bayes formula and the fact that diffusions are *Markov processes*

$$p_\theta(x_T|x_0) = \int p_\theta(x_T, x_s|x_0)\mathrm{d}x_s \quad = \int p_\theta(x_T|x_s)p_\theta(x_s|x_0)\mathrm{d}x_s$$
$$= \mathbf{E}\left[p_\theta(x_T|x_s)|x_0\right].$$

Generate empirical version of $p_\theta(x_s|x_0)$ using Monte Carlo

$$p_\theta^K(x_s|x_0) = \frac{1}{K}\sum_{k=1}^{K}\delta(x_s - x_s^k) \tag{6}$$

resulting in, Pedersen (1995),

$$\hat{p}_\theta(x_T|x_0) = \frac{1}{K}\sum_{k=1}^{K}p_\theta(x_T|x_s^k) \tag{7}$$

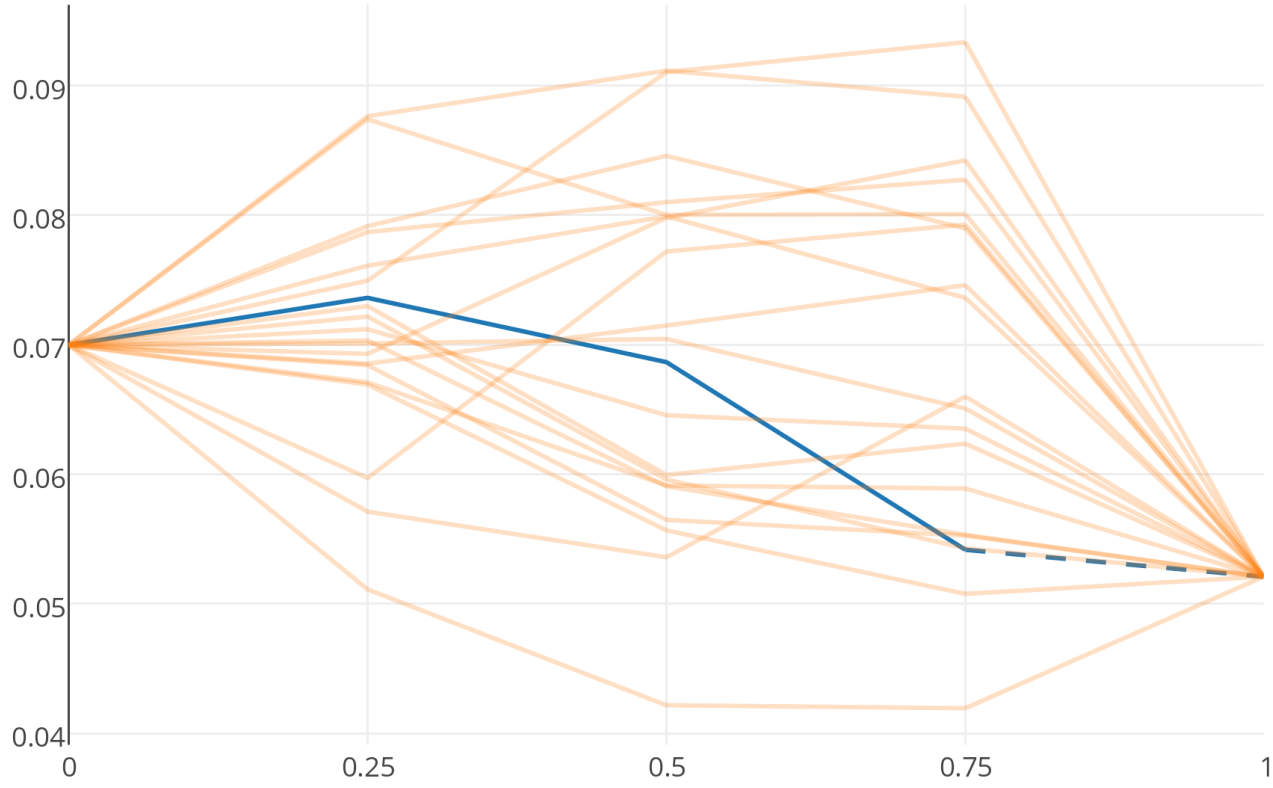



Figure: Example of the simulated maximum likelihood algorithm introduced by Pedersen (1995) using $K = 20$ trajectories on a grid with $R = 4$ equal intervals

# Complexity

In practice, there are two types of errors, discretization and variance

$$\begin{aligned}\epsilon &= \hat{p}_\theta(x_T^\delta|x_0) - p_\theta(x_T|x_0) \\ &= \underbrace{\hat{p}_\theta(x_T^\delta|x_0) - \hat{p}_\theta(x_T|x_0)}_{\textit{Bias}} + \underbrace{\hat{p}_\theta(x_T|x_0) - p_\theta(x_T|x_0)}_{\textit{Random error}}\end{aligned}$$

Simulating $K$ Monte Carlo trajectories, time partitioned in $R$ equal parts (s.t. $\delta = T/R$) gives

- Bias $\mathcal{O}(1/R)$
- Variance $\mathcal{O}(1/K)$
- Mean square error $\mathbf{E}[\epsilon^2]$ is then $\mathcal{O}(1/R^2 + 1/K)$
- Computational complexity for RMSE error of $\mathcal{O}(\epsilon)$ in $\mathcal{O}(\epsilon^{-3})$

# Complexity

In practice, there are two types of errors, discretization and variance

$$\begin{aligned}\epsilon &= \hat{p}_\theta(x_T^\delta|x_0) - p_\theta(x_T|x_0) \\ &= \underbrace{\hat{p}_\theta(x_T^\delta|x_0) - \hat{p}_\theta(x_T|x_0)}_{\textit{Bias}} + \underbrace{\hat{p}_\theta(x_T|x_0) - p_\theta(x_T|x_0)}_{\textit{Random error}}\end{aligned}$$

Simulating $K$ Monte Carlo trajectories, time partitioned in $R$ equal parts (s.t. $\delta = T/R$) gives

- Bias $\mathcal{O}(1/R)$
- Variance $\mathcal{O}(1/K)$
- Mean square error $\mathbf{E}[\epsilon^2]$ is then $\mathcal{O}(1/R^2 + 1/K)$
- Computational complexity for RMSE error of $\mathcal{O}(\epsilon)$ in $\mathcal{O}(\epsilon^{-3})$

Exact simulation would give a computational complexity $\mathcal{O}(\epsilon^{-2})$

# Multilevel Monte Carlo

Giles (2008) showed organizing computations in a clever way reduces complexity. The MLMC method reduces the cost for obtaining an RMSE of $\mathcal{O}(\epsilon)$ to $\mathcal{O}(\epsilon^{-2}(\log \epsilon)^2)$.

# Multilevel Monte Carlo

Giles (2008) showed organizing computations in a clever way reduces complexity. The MLMC method reduces the cost for obtaining an RMSE of $\mathcal{O}(\epsilon)$ to $\mathcal{O}(\epsilon^{-2}(\log \epsilon)^2)$.

The idea is basically *control variates*, but refined.

# Control variates

- Suppose we want to estimate $\mathbf{E}[X]$, and that we also have $Y$ as well as $\mathbf{E}[Y]$.
- Define

$$\xi = X + \gamma(Y - \mathbf{E}[Y]). \tag{8}$$

  Then $\mathbf{E}[\xi] = \mathbf{E}[X]$.
- $\gamma$ can be chosen such that the variance is minimized

$$\mathbf{Var}[\xi] = \mathbf{Var}[X](1 - \rho^2) \tag{9}$$

# Example - fig tree

My sister can grow figs in her garden - Can I grow figs in my garden?

# Example - fig tree

My sister can grow figs in her garden - Can I grow figs in my garden?
Estimate the average temperature in my garden, $\mathbf{E}[X]$, given the measurements last year $X$ and the temperatures $Y$ and average temperature where she lives $\mathbf{E}[Y]$.

# Multi Level Monte Carlo

- Compute the expectation as a sequence of correction terms at levels $l = 0, \ldots, L$
- Denote the approximation at level $l$ by $P_l$.
- Then $\xi$ is defined as

$$\xi = P_0 + \sum_{l=1}^{L} (P_l - P_{l-1}) \tag{10}$$

where $\mathbf{E}[\xi] = \mathbf{E}[P_L]$.

# Multiple levels

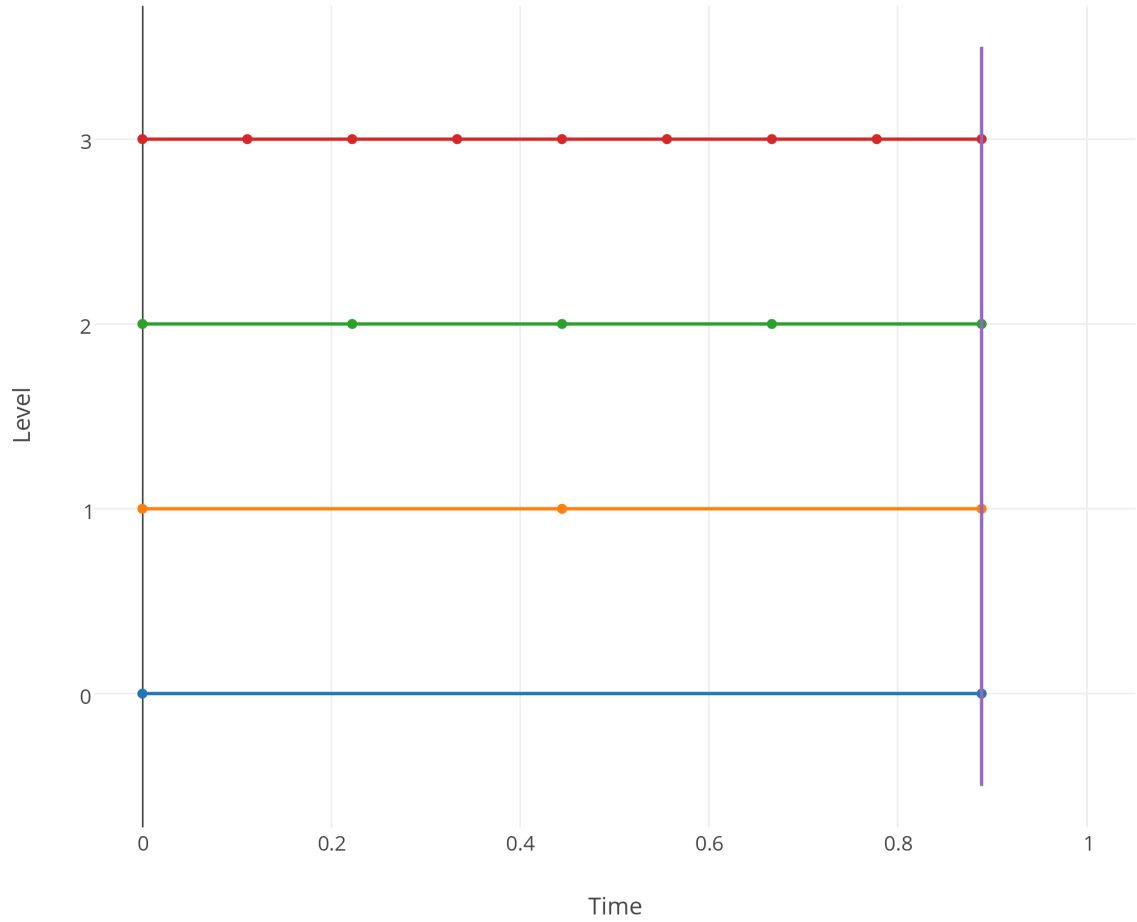



Figure: Multilevel grid example for the Pedersen algorithm. Here we use $L = 3$ levels with $M = 2$. $s$ is indicated by the vertical line.

# MLMC 2

- A naive implementation *increases* the variance
- However, the variance of each $P_l - P_{l-1}$ is very small, and decays with the step size if the use the *same* random elements.

# MLMC 2

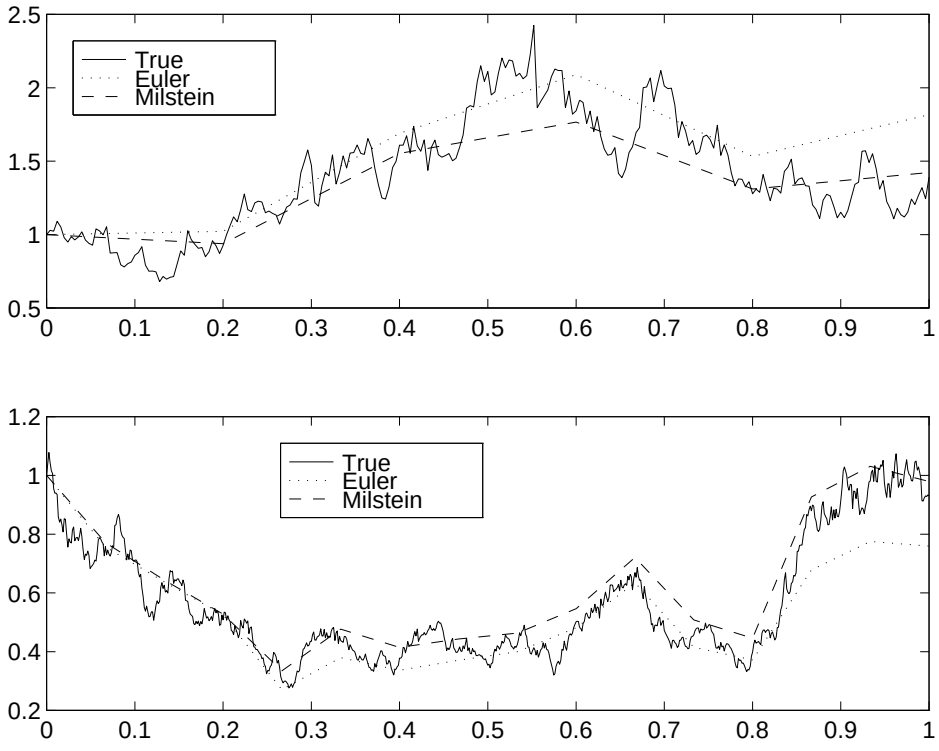
- A naive implementation *increases* the variance
- However, the variance of each $P_l - P_{l-1}$ is very small, and decays with the step size if the use the *same* random elements. Draw a figure to see this!

# MLMC 2

- A naive implementation *increases* the variance
- However, the variance of each $P_l - P_{l-1}$ is very small, and decays with the step size if the use the *same* random elements. Draw a figure to see this!
- This follows from the strong convergence of the discretization scheme.
- Both weak and strong rates of converges are important!

Use Pederson idea in MLMC framework, see (Lindström & Åkerlindh, 2016)**?**.

# Why does it work?




Note the difference between different Brownian motions vs. different schemes/time steps.

LUND UNIVERSITY

# Comparison

Keep the bias fixed, while comparing variances

## Standard Pedersen

- Cost: $K_p M^L$
- Variance: $\sigma^2 / K_p$

## Multilevel Pedersen

- Cost: $\sum_{l=0}^{L} K_l M^l = K(L+1)$
- Variance: $(L+1)\sigma^2 / K$

Equal cost gives $K = K_p M^L / (L+1)$

# Comparison

Keep the bias fixed, while comparing variances

## Standard Pedersen

- Cost: $K_p M^L$
- Variance: $\sigma^2/K_p$

## Multilevel Pedersen

- Cost: $\sum_{l=0}^{L} K_l M^l = K(L+1)$
- Variance: $(L+1)\sigma^2/K$

Equal cost gives $K = K_p M^L/(L+1)$. That leads to

$$\frac{\mathbf{Var}[MLMC]}{\mathbf{Var}[MC]} = \frac{(L+1)^2}{M^L} \tag{11}$$

We are currently working on an improved version that will make the MLMC estimator even more efficient.

# Comparison

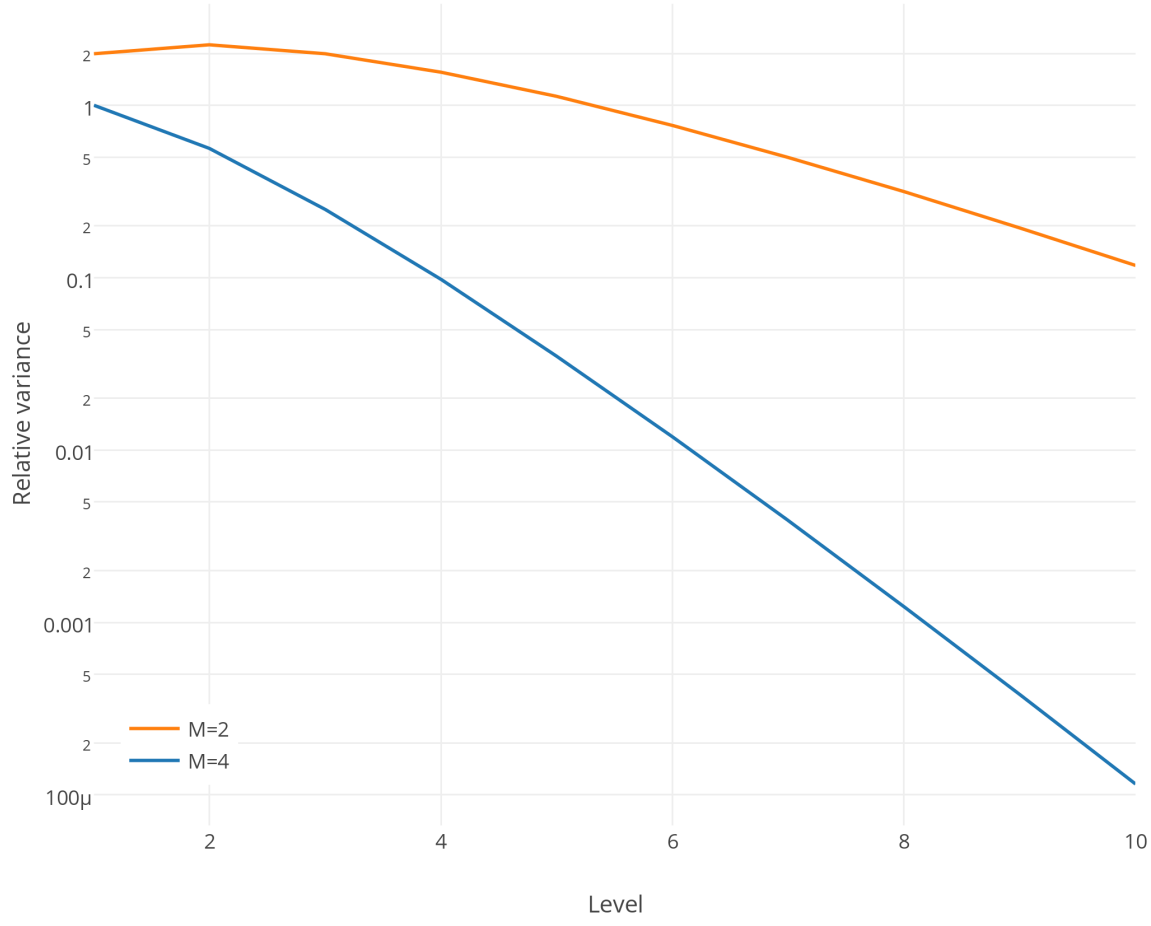



Figure: Theoretical relative variance of standard and multilevel Pedersen with identical compuational cost and bias.

# Simulation study

We consider complete (bivariate) observations from the Heston (1993) model,

$$\mathrm{d}S_t = 0.05 S_t \mathrm{d}t + \sqrt{V_t} S_t \mathrm{d}W_t^{(S)} \tag{12a}$$

$$\mathrm{d}V_t = 2\left(0.04 - V_t\right)\mathrm{d}t + 0.25\sqrt{V_t}\mathrm{d}W_t^{(V)} \tag{12b}$$

with $\mathrm{d}W_t^{(S)}\mathrm{d}W_t^{(V)} = -0.5\mathrm{d}t$

# Simulation study

- 1000 observations
- Standard Pedersen using $K_p = 2^8 = 256$
- Multilevel Pedersen using $M = 4$ and $L = 3$
- Ground truth, standard Pedersen using $K_{\text{true}} = 10^6$
- Compute

$$\text{VR} = \frac{\sum_{n=1}^{N} \left(\hat{p}_{\text{ML}}(x_n|x_{n-1}) - \hat{p}_{\text{True}}(x_n|x_{n-1})\right)^2}{\sum_{n=1}^{N} \left(\hat{p}_{\text{Pedersen}}(x_n|x_{n-1}) - \hat{p}_{\text{True}}(x_n|x_{n-1})\right)^2} \tag{13}$$

# Simulation study

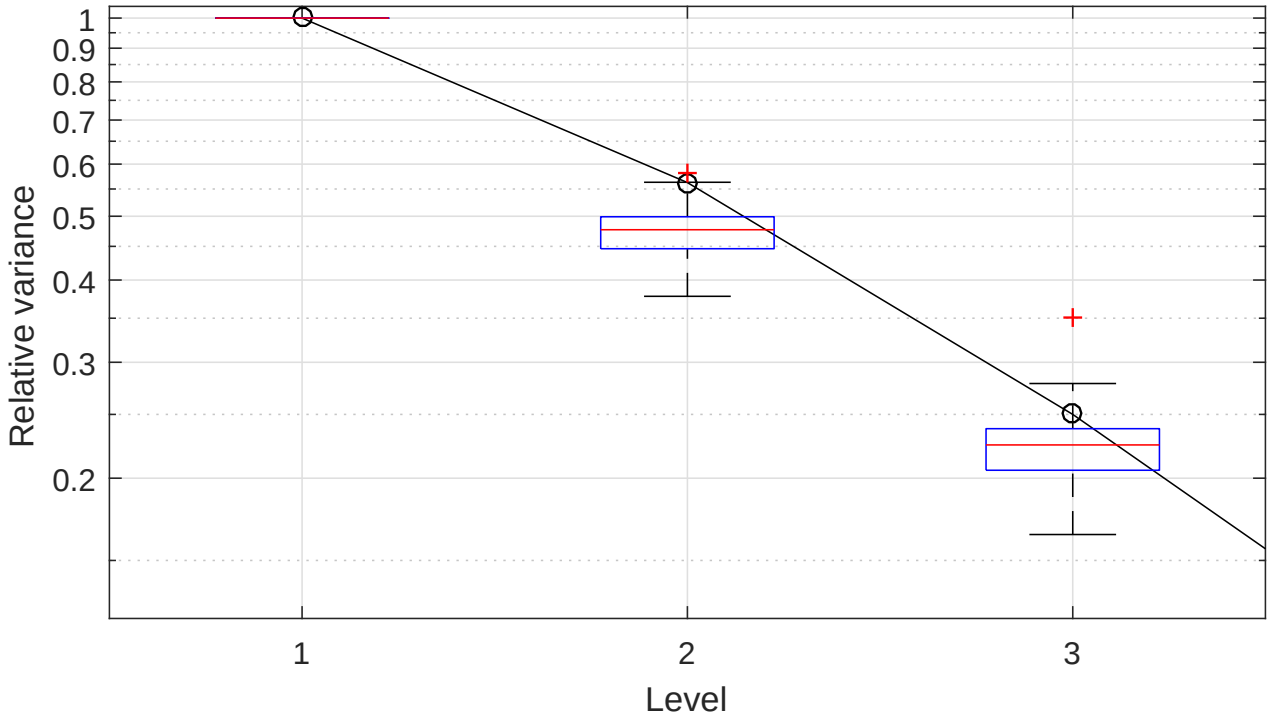



Figure: Relative variance between multilevel and standard Pedersen. Bootstrapped estimates (boxplot) compared to theoretical value (solid line).

# Parameter estimation

Can be done, but

- The computations provide a point wise estimate

# Parameter estimation

Can be done, but

- The computations provide a point wise estimate
- Can use importance sampling

Simulations indicate that it works fairly well in practice.

# Parameter estimation

Can be done, but

- The computations provide a point wise estimate
- Can use importance sampling

Simulations indicate that it works fairly well in practice.

- We could also use the Multi-Level Monte Carlo estimates within an adaptive PMMH algorithm

# Cox-Ingersoll-Ross model

Standard interest rate model until recent years (neg rates)

$$\mathrm{d}r(t) = \kappa\left(\xi - r(t)\right)\mathrm{d}t + \sigma\sqrt{r(t)}\mathrm{d}W(t). \qquad (14)$$

Compare PMMH results

- Exact likelihood (best mixing)
- MC - Pedersen
- MC - Bridge sampler (Durham-Gallant, 2002)
- MLMC - Pedersen

Adaptive MCMC, 2000 iterations as burnin, 2000 additional iterations.

# Cox-Ingersoll-Ross model

Standard interest rate model until recent years (neg rates)

$$\mathrm{d}r(t) = \kappa\left(\xi - r(t)\right)\mathrm{d}t + \sigma\sqrt{r(t)}\mathrm{d}W(t). \qquad (14)$$

Compare PMMH results

- Exact likelihood (best mixing)
- MC - Pedersen
- MC - Bridge sampler (Durham-Gallant, 2002)
- MLMC - Pedersen

Adaptive MCMC, 2000 iterations as burnin, 2000 additional iterations.The simulation used $M = 4$, $L = 2$ and $K_p = 20$.

# Cox-Ingersoll-Ross model

Standard interest rate model until recent years (neg rates)

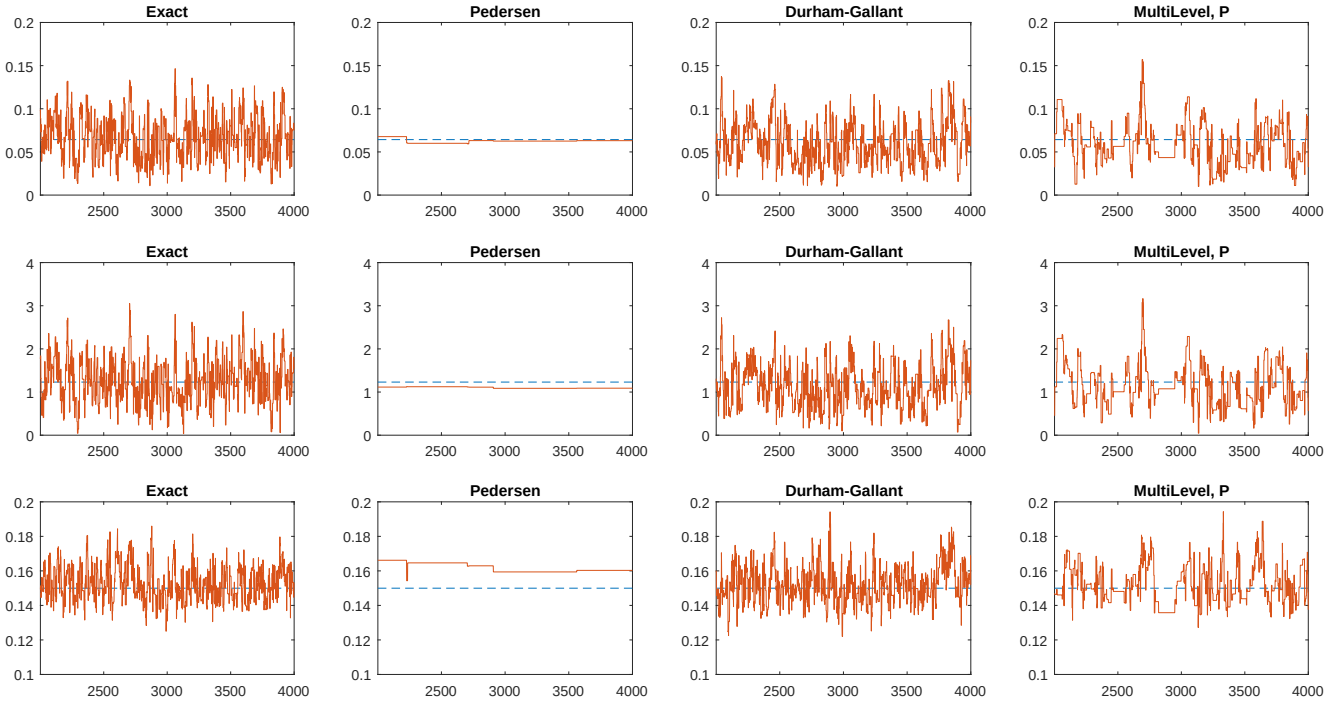
$$\mathrm{d}r(t) = \kappa\left(\xi - r(t)\right)\mathrm{d}t + \sigma\sqrt{r(t)}\mathrm{d}W(t). \qquad (14)$$

Compare PMMH results

- Exact likelihood (best mixing)
- MC - Pedersen
- MC - Bridge sampler (Durham-Gallant, 2002)
- MLMC - Pedersen

Adaptive MCMC, 2000 iterations as burnin, 2000 additional iterations.The simulation used $M = 4$, $L = 2$ and $K_p = 20$.
**Prior:** All parameters positive and Feller condition holds.

# Path plots for the parameters




It is known that noisy estimates of the log-likelihood function degrades the mixing of the simulated parameters.

LUND UNIVERSITY

# References I


Bacher, P., Madsen, H., Nielsen, H. A., and Perers, B. (2013). Short-term heat load forecasting for single family houses. *Energy and buildings*, 65:101–112.

Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., and Dhiman, R. C. (2011). Malaria in northwest india: Data analysis via partially observed stochastic differential equation models driven by lévy noise. *Journal of the American Statistical Association*, 106(494):440–451.

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

Giles, M. B. (2008). Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617.


# References II


Giles, M. B. (2015). Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328.

Giles, M. B., Szpruch, L., et al. (2014). Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. *The Annals of Applied Probability*, 24(4):1585–1620.

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

Iversen, E. B., Morales, J. M., Møller, J. K., and Madsen, H. (2016). Short-term probabilistic forecasting of wind speed using stochastic differential equations. *International Journal of Forecasting*, 32(3):981–990.


# References III


King, A. A., de Cellès, M. D., Magpantay, F. M., and Rohani, P. (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. In *Proc. R. Soc. B*, volume 282, page 20150347. The Royal Society.

Lindström, E. (2012). A regularized bridge sampler for sparsely sampled diffusions. *Statistics and Computing*, 22(2):615–623.

Lindström, E. and Åkerlindh, C. (2016). Multilevel monte carlo methods for simulated maximum likelihood inference in multivariate diffusions. In *9th World Congress of the Bachelier Finance Society*.

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian journal of statistics*, pages 55–71.


# References IV


Rhee, C.-H. and Glynn, P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043.

Stramer, O. and Yan, J. (2007). Asymptotics of an efficient Monte Carlo estimation for the transition density of diffusion processes. *Methodology and Computing in Applied Probability*, 9(4):483–496.

Wendt, S. L., Ranjan, A., Møller, J. K., Schmidt, S., Knudsen, C. B., Holst, J. J., Madsbad, S., Madsen, H., Nørgaard, K., and Jørgensen, J. B. (2017). Cross-validation of a glucose-insulin-glucagon pharmacodynamics model for simulation using data from patients with type 1 diabetes. *Journal of Diabetes Science and Technology*, page 1932296817693254.

Contact details


- Erik Lindström — erik.lindstrom@matstat.lu.se



LUND
UNIVERSITY