

Iterated filtering

Historical background and future developments

Erik Lindström

Mathematical Statistics, Lund University



Partially Observed Markov Processes

Consider models of the type

- ▶ Observation equation: $y_n = h_n(x_n, \eta_n, \theta)$
- ▶ Latent Markov state equation: $x_n = f_n(x_{n-1}, e_n, \theta)$

We are interested in estimating the parameters $\theta \in \Theta$.

Partially Observed Markov Processes

Consider models of the type

- ▶ Observation equation: $y_n = h_n(x_n, \eta_n, \theta)$
- ▶ Latent Markov state equation: $x_n = f_n(x_{n-1}, e_n, \theta)$

We are interested in estimating the parameters $\theta \in \Theta$.

Assuming that the functions $h_n()$ and $f_n()$ are sufficiently regular to ensure that the MLE is strongly consistent and efficient.

Partially Observed Markov Processes

Historical overview

- ▶ Linear models
 - ▶ Kalman filters
 - ▶ HMM

Partially Observed Markov Processes

Historical overview

- ▶ Linear models
 - ▶ Kalman filters
 - ▶ HMM
- ▶ Non-linear models
 - ▶ Extended Kalman filters, Iterated Extended Kalman filters

Partially Observed Markov Processes

Historical overview

- ▶ Linear models
 - ▶ Kalman filters
 - ▶ HMM
- ▶ Non-linear models
 - ▶ Extended Kalman filters, Iterated Extended Kalman filters
 - ▶ Unscented Kalman filters, SMC filters (and EnKF)

MLE

The Maximum Likelihood Estimator (MLE) is given by

$$\hat{\theta}^{MLE} = \arg \max_{\theta \in \Theta \subseteq \mathbb{R}^p} \log p_{Y_{1:n}}(y_{1:n}|\theta) = \arg \max_{\theta \in \Theta \subseteq \mathbb{R}^p} \ell(\theta). \quad (1)$$

MLE

The Maximum Likelihood Estimator (MLE) is given by

$$\hat{\theta}^{MLE} = \arg \max_{\theta \in \Theta \subseteq \mathbb{R}^p} \log p_{Y_{1:n}}(y_{1:n}|\theta) = \arg \max_{\theta \in \Theta \subseteq \mathbb{R}^p} \ell(\theta). \quad (1)$$

This is hard to compute as the log-likelihood is given by

$$\begin{aligned} \ell(\theta) &= \log \left(\int p_{X_0}(x_0; \theta) \prod_{k=1}^n p_{X_k, Y_k | X_{k-1}}(x_k, y_k | x_{k-1}; \theta) dx_{0:n} \right) \\ &= \log \left(\int p_{X_0}(x_0; \theta) \prod_{k=1}^n p_{Y_k | X_k}(y_k | x_k; \theta) p_{X_k | X_{k-1}}(x_k | x_{k-1}; \theta) dx_{0:n} \right) \end{aligned}$$

This is a $n \times d$ dimensional integral if $X_n \in \mathbb{R}^d$.

Different techniques

No exact method but plenty of approximations!

- ▶ Bayesian versus frequentist methods
- ▶ Likelihood or quasi Likelihood
- ▶ Moment matching (EF/GMM etc.)
- ▶ Likelihood free methods
 - ▶ Approximate Bayesian Computation
 - ▶ Artificial dynamics (including *Iterated Filtering*)

Likelihood methods

Sequential Monte Carlo methods can in principle be used to compute the likelihood function.

- ▶ Direct maximization (but the surface is often non-smooth, cf. Hürzeler and Künsch (2001))
- ▶ EM (requires smoothing), see Cappé et al. (2005)
- ▶ Bayesian methods, see Andrieu et al. (2010)
- ▶ Plug and Play (likelihood free) methods.

Artificial Dynamics, augmenting the state space,

A simple (adaptive?) estimator can be obtained by augmenting the state equation with the parameters (Åström & Eykhoff, 1971) [12], going from

- Observation equation:

$$y_n = h_n(x_n, \theta_n) + \eta_n,$$

- Latent Markov state equation:

$$x_n = f_n(x_{n-1}, \theta) + e_n^x$$

Artificial Dynamics, augmenting the state space,

A simple (adaptive?) estimator can be obtained by augmenting the state equation with the parameters (Åström & Eykhoff, 1971) [12], going from

- Observation equation:

$$y_n = h_n(x_n, \theta_n) + \eta_n,$$

- Latent Markov state equation:

$$x_n = f_n(x_{n-1}, \theta) + e_n^x$$

to

- Observation equation:

$$y_n = h_n(x_n, \theta_n) + \eta_n,$$

- Latent Markov state equation:

$$\begin{pmatrix} x_n \\ \theta_n \end{pmatrix} = \begin{pmatrix} f_n(x_{n-1}, \theta_{n-1}) \\ \theta_{n-1} \end{pmatrix} + \begin{pmatrix} e_n^x \\ e_n^\theta \end{pmatrix}$$

These estimators are well known in engineering, cf. Ljung (1979) who writes

These estimators are well known in engineering, cf. Ljung (1979) who writes

The recursive parameter estimation problem for linear systems is inherently a nonlinear filtering problem, and as pointed out in [12] there are in principle no differences between parameter estimation and state estimation

These estimators are well known in engineering, cf. Ljung (1979) who writes

The recursive parameter estimation problem for linear systems is inherently a nonlinear filtering problem, and as pointed out in [12] there are in principle no differences between parameter estimation and state estimation

Note: The resulting estimator $p(\theta_n|y_{1:n})$ is *in general not the MLE!*

These estimators are well known in engineering, cf. Ljung (1979) who writes

The recursive parameter estimation problem for linear systems is inherently a nonlinear filtering problem, and as pointed out in [12] there are in principle no differences between parameter estimation and state estimation

Note: The resulting estimator $p(\theta_n|y_{1:n})$ is *in general not the MLE!*

This type of estimators, when $\mathbf{Cov}(e^\theta) = \text{const}$, are known as linear forgetting type estimators.

These have been used in finance recently, cf. Lindström et al. (2008); Nystrup et al. (2015).

Simple case I

- ▶ Estimating a constant, $y_n = m + e_n$, $e_n \sim N(0, \Gamma)$

Simple case I

- ▶ Estimating a constant, $y_n = m + e_n$, $e_n \sim N(0, \Gamma)$
- ▶ Solving this with a Kalman filter (assuming $P_0 = \infty$).

Simple case I

- ▶ Estimating a constant, $y_n = m + e_n$, $e_n \sim N(0, \Gamma)$
- ▶ Solving this with a Kalman filter (assuming $P_0 = \infty$). Model:

$$y_n = x_n + \eta_n \quad (2)$$

$$x_n = x_{n-1} \quad (3)$$

Simple case I

- ▶ Estimating a constant, $y_n = m + e_n$, $e_n \sim N(0, \Gamma)$
- ▶ Solving this with a Kalman filter (assuming $P_0 = \infty$). Model:

$$y_n = x_n + \eta_n \quad (2)$$

$$x_n = x_{n-1} \quad (3)$$

- ▶ It can be shown that $\mathbf{E}[x_n|y_{1:n}]$ is the MLE!

Simple case I

- ▶ Estimating a constant, $y_n = m + e_n$, $e_n \sim N(0, \Gamma)$
- ▶ Solving this with a Kalman filter (assuming $P_0 = \infty$). Model:

$$y_n = x_n + \eta_n \quad (2)$$

$$x_n = x_{n-1} \quad (3)$$

- ▶ It can be shown that $\mathbf{E}[x_n | y_{1:n}]$ is the MLE!
- ▶ What about the adaptive recursive version?

$$y_n = x_n + \eta_n \quad (4)$$

$$x_n = x_{n-1} + q_n e_n \quad (5)$$

Simple case II

(Ljung, 1979) analyzed the model

$$y_n = x_n + \eta_n$$

$$x_n = ax_{n-1} + e_n^x$$

transformed into

$$y_n = x_n + \eta_n$$

$$\begin{pmatrix} x_n \\ a_n \end{pmatrix} = \begin{pmatrix} a_{n-1}x_{n-1} \\ a_{n-1} \end{pmatrix} + \begin{pmatrix} e_n^x \\ e_n^a \end{pmatrix}$$

The model is non-linear, so an EKF was used for filtering.

Simple case II

(Ljung, 1979) analyzed the model

$$y_n = x_n + \eta_n$$

$$x_n = ax_{n-1} + e_n^x$$

transformed into

$$y_n = x_n + \eta_n$$

$$\begin{pmatrix} x_n \\ a_n \end{pmatrix} = \begin{pmatrix} a_{n-1}x_{n-1} \\ a_{n-1} \end{pmatrix} + \begin{pmatrix} e_n^x \\ e_n^a \end{pmatrix}$$

The model is non-linear, so an EKF was used for filtering.

Result: The estimate will diverge with positive probability!

Iterated Filtering

It was shown in Ionides et al. (2006) that the score function can be approximated by moments. Introduce the time varying parameter process as

$$\check{\Theta}_0 = \theta + \tau \zeta_0, \quad (6)$$

$$\check{\Theta}_n = \check{\Theta}_{n-1} + \sigma \zeta_n. \quad (7)$$

where the random variables $\{\zeta_k\}_{k=1}^n$ are having compact support, $\mathbf{E}[\zeta] = 0$ and $\text{Var}(\zeta) = \Sigma$.

The joint distribution of the augmented model is given by

$$\begin{aligned} & g_{X_{0:N}, \Theta_{0:N}, Y_{1:N}}(x_{0:n}, \theta_{0:N}, y_{1:N} | \theta, \sigma, \tau, \Sigma) \\ &= f_{X_{0:N}, Y_{1:N}}(x_{0:n}, y_{1:N} | \theta_{0:N}) g_{\Theta_{0:N}}(\theta_{0:N} | \theta, \sigma, \tau, \Sigma) \end{aligned} \quad (8)$$

The joint distribution of the augmented model is given by

$$\begin{aligned} & g_{X_{0:N}, \Theta_{0:N}, Y_{1:N}}(x_{0:n}, \theta_{0:N}, y_{1:N} | \theta, \sigma, \tau, \Sigma) \\ &= f_{X_{0:N}, Y_{1:N}}(x_{0:n}, y_{1:N} | \theta_{0:N}) g_{\Theta_{0:N}}(\theta_{0:N} | \theta, \sigma, \tau, \Sigma) \end{aligned} \quad (8)$$

Define the moments

$$\check{\theta}_n^F = \mathbf{E}_{\theta, \sigma, \tau, \Sigma}[\check{\Theta}_n | Y_{1:n} = y_{1:n}], \quad (9)$$

$$\check{V}_n^P = \mathbf{Var}_{\theta, \sigma, \tau, \Sigma}[\check{\Theta}_n | Y_{1:n-1} = y_{1:n-1}]. \quad (10)$$

These moments depends implicitly on σ, τ, Σ .

Theorem 3 in Ionides et al. (2011)

- It can be shown that

$$\check{\theta}_n^F - \check{\theta}_{n-1}^F \approx \mathbf{Var}(\theta_n - \check{\theta}_{n-1}^F) \nabla \log p(y_n | \check{\theta}_{n-1}^F, \mathcal{F}_{n-1})$$

Theorem 3 in Ionides et al. (2011)

- It can be shown that

$$\check{\theta}_n^F - \check{\theta}_{n-1}^F \approx \mathbf{Var}(\theta_n - \check{\theta}_{n-1}^F) \nabla \log p(y_n | \check{\theta}_{n-1}^F, \mathcal{F}_{n-1})$$

- (Ionides et al., 2011, Theorem 3) Let K_1 be a compact subset of $\mathbb{R}^p$, C_1 is a constant, τ is small enough and $\lim_{\tau \rightarrow 0} \sigma(\tau)/\tau = 0$. It then holds that

$$\sup_{\theta \in K_1} \left| \sum_{n=1}^N \left(\check{V}_n^P \right)^{-1} \left(\check{\theta}_n^F - \check{\theta}_{n-1}^F \right) - \nabla \ell(\theta) \right| \leq C_1 \left(\tau + \frac{\sigma^2}{\tau^2} \right).$$

- This holds for *exact* filters.

Theorem 3 in Ionides et al. (2011)

- It can be shown that

$$\check{\theta}_n^F - \check{\theta}_{n-1}^F \approx \mathbf{Var}(\theta_n - \check{\theta}_{n-1}^F) \nabla \log p(y_n | \check{\theta}_{n-1}^F, \mathcal{F}_{n-1})$$

- (Ionides et al., 2011, Theorem 3) Let K_1 be a compact subset of $\mathbb{R}^p$, C_1 is a constant, τ is small enough and $\lim_{\tau \rightarrow 0} \sigma(\tau)/\tau = 0$. It then holds that

$$\sup_{\theta \in K_1} \left| \sum_{n=1}^N \left(\check{V}_n^P \right)^{-1} \left(\check{\theta}_n^F - \check{\theta}_{n-1}^F \right) - \nabla \ell(\theta) \right| \leq C_1 \left(\tau + \frac{\sigma^2}{\tau^2} \right).$$

- This holds for *exact* filters.
- The means and covariance can be computed using Seq. Monte Carlo methods. We denote the empirical versions of the moments by $\check{\theta}_n^F$ and $\check{V}_n^P$.

Theorem 3 in Ionides et al. (2011)

- It can be shown that

$$\check{\theta}_n^F - \check{\theta}_{n-1}^F \approx \mathbf{Var}(\theta_n - \check{\theta}_{n-1}^F) \nabla \log p(y_n | \check{\theta}_{n-1}^F, \mathcal{F}_{n-1})$$

- (Ionides et al., 2011, Theorem 3) Let K_1 be a compact subset of $\mathbb{R}^p$, C_1 is a constant, τ is small enough and $\lim_{\tau \rightarrow 0} \sigma(\tau)/\tau = 0$. It then holds that

$$\sup_{\theta \in K_1} \left| \sum_{n=1}^N \left(\check{V}_n^P \right)^{-1} \left(\check{\theta}_n^F - \check{\theta}_{n-1}^F \right) - \nabla \ell(\theta) \right| \leq C_1 \left(\tau + \frac{\sigma^2}{\tau^2} \right).$$

- This holds for *exact* filters.
- The means and covariance can be computed using Seq. Monte Carlo methods. We denote the empirical versions of the moments by $\check{\theta}_n^F$ and $\check{V}_n^P$.
- The Monte Carlo error can be controlled by the number of Monte Carlo samples or particles J .

Theorem 5 in (Ionides et al., 2011)

- Let $\{a_m\}$, $\{\tau_m\}$, $\{\sigma_m\}$ and $\{J_m\}$ be positive sequences such that $\tau_m \rightarrow 0$, $\sigma_m \tau_m^{-1} \rightarrow 0$, $\tau_m J_m \rightarrow \infty$, $\sum_m a_m = \infty$ and $\sum_m a_m^2 J_m \tau_m^{-2} < \infty$ and define $\hat{\theta}_m$ according to:

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} \left(\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F \right) \quad (11)$$

The estimate will then converge $\lim_{m \rightarrow \infty} \hat{\theta}_m = \hat{\theta}^{MLE}$ with probability one.

Theorem 5 in (Ionides et al., 2011)

- ▶ Let $\{a_m\}$, $\{\tau_m\}$, $\{\sigma_m\}$ and $\{J_m\}$ be positive sequences such that $\tau_m \rightarrow 0$, $\sigma_m \tau_m^{-1} \rightarrow 0$, $\tau_m J_m \rightarrow \infty$, $\sum_m a_m = \infty$ and $\sum_m a_m^2 J_m \tau_m^{-2} < \infty$ and define $\hat{\theta}_m$ according to:

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} \left(\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F \right) \quad (11)$$

The estimate will then converge $\lim_{m \rightarrow \infty} \hat{\theta}_m = \hat{\theta}^{MLE}$ with probability one.

- ▶ (Ionides et al., 2011) noted that $a_m = m^{-1}$, $\tau_m^2 = m^{-1}$, $\sigma_m^2 = m^{-(1+\delta)}$ and $J_m = m^{(1/2+\delta)}$ where $\delta > 0$ satisfies the conditions for convergence (a.s.).

Some papers using the estimator

- ▶ King, A. A., Ionides, E. L., Pascual, M. and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature* 454 877-880.
- ▶ 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 Levy noise. *Journal of the American Statistical Association* 106 440-451.

References: King et al. (2008); Bhadra et al. (2011)

Discussion

Pros:

- ▶ It is a plug and play method: Only requirement (if a Bootstrap filter is used) is the ability to
 - ▶ Simulate the system
 - ▶ Relate those simulations to data
- ▶ Performance often competitive compared to PMCMC.
- ▶ Implemented in the `pomp` R package, see King et al. (2016).

Discussion

Pros:

- ▶ It is a plug and play method: Only requirement (if a Bootstrap filter is used) is the ability to
 - ▶ Simulate the system
 - ▶ Relate those simulations to data
- ▶ Performance often competitive compared to PMCMC.
- ▶ Implemented in the `pomp` R package, see King et al. (2016).

Cons:

- ▶ The algorithm is a stochastic gradient method...
- ▶ It is plug and play
- ▶ The curse of dimensionality

Example: Rate of convergence

- ▶ A statistical model for the constant is given by

$$Y_n = \mu + \eta_n, \quad \eta \sim N(0, 1) \quad (12)$$

The MLE (μ_{MLE}) is given by the sample mean.

Example: Rate of convergence

- ▶ A statistical model for the constant is given by

$$Y_n = \mu + \eta_n, \quad \eta \sim N(0, 1) \quad (12)$$

The MLE (μ_{MLE}) is given by the sample mean.

- ▶ The augmented model is given by

$$Y_n = \mu_{m,n} + \eta_n$$

$$\mu_{m,n} = \mu_{m,n-1} + \sigma_m e_n$$

with $\mu_{m,0} \sim N(\hat{\mu}_m, \tau_m^2)$.

Example: Rate of convergence

- ▶ A statistical model for the constant is given by

$$Y_n = \mu + \eta_n, \quad \eta \sim N(0, 1) \quad (12)$$

The MLE (μ_{MLE}) is given by the sample mean.

- ▶ The augmented model is given by

$$\begin{aligned} Y_n &= \mu_{m,n} + \eta_n \\ \mu_{m,n} &= \mu_{m,n-1} + \sigma_m e_n \end{aligned}$$

with $\mu_{m,0} \sim N(\hat{\mu}_m, \tau_m^2)$. We have generated $N_{sim} = 100$ independent realizations, each consisting of $N = 50$ observations of $N(\mu, 1)$ random variables with $\mu = 1$.

- ▶ These are parametrized according to

$$\begin{aligned} a_m &= m^{-1} & \tau_m^2 &= m^{-1} \\ \sigma_m^2 &= m^{-(1+0.1)} & J_m &= \text{round} \left(J_e \exp(-m) + J_0 + J_1 m^{(0.5+0.1)} \right) \end{aligned}$$

with $J_e = 500$, $J_0 = 100$ and $J_1 = 1$.

Result - Very slow convergence!

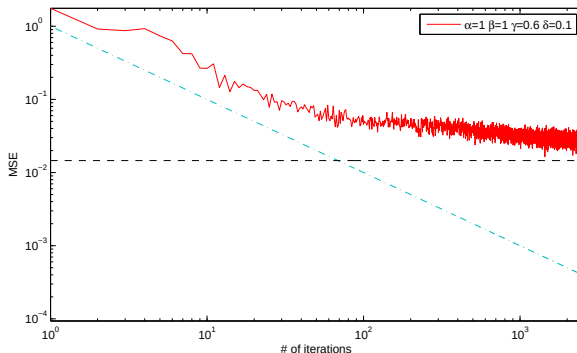


Figure: Convergence (computed as MSE) between the Iterated Filtering estimate and the MLE (solid line). The dashed black line is the MSE between the MLE and the true parameter while the dashed-dotted line is $1/m$.

Need for improvements!

- ▶ Changing the algorithm, cf Lindström et al. (2012); Ionides et al. (2015)
- ▶ Better exploration of the parameter space Lindström (2013)
- ▶ Optimizing the algorithm

A second order algorithm

Heuristic arguments (Quantity Calculus) suggests that the algorithm is suboptimal.

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} \left(\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F \right) \quad (13)$$

A second order algorithm

Heuristic arguments (Quantity Calculus) suggests that the algorithm is suboptimal.

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} \left(\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F \right) \quad (13)$$

The a_m (matrix) has to reset the quantity distortion caused by V^P .

A second order algorithm

Heuristic arguments (Quantity Calculus) suggests that the algorithm is suboptimal.

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} \left(\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F \right) \quad (13)$$

The a_m (matrix) has to reset the quantity distortion caused by V^P .

Second order Stochastic Approximation indicates that a_m should be related to the Fisher information.

Second order Iterated Filtering

- ▶ From Lindström et al. (2012)
 - ▶ Approximate second order information include in the SA algorithm, taking $a_m = V^P$.
 - ▶ "Recursive formulation" (avoiding the batch asymptotics, better for "big data")
- ▶ From Ionides et al. (2015)
 - ▶ Similar, but batch mode.
 - ▶ Derivation more general.

IF2 algorithm, see Ionides et al. (2015)

Comparisons with the IF1 algorithm

- ▶ IF1 is an approximate Fisher score algorithm
- ▶ IF2 combines Bayes maps with data cloning, cf. Lele et al. (2007) which is closely related to the SAME algorithm, see Doucet et al. (2002).
- ▶ Another approach, that relies on smoothing rather than filtering, is presented in Doucet et al. (2013) (still not published).

IF2 algorithm

Pseudo code

- ▶ Define M , J and σ_m
- ▶ For $m = 1 : M$
 - ▶ Initialize $\theta_{0,j}^{F,m} \sim h_0(\theta|\theta_j^{m-1})$ and $X_{0,j}^{F,m} \sim f(x_0)$
 - ▶ For $n = 1 : N$
 - ▶ Predict to obtain $\theta_{n,j}^{P,m} \sim h_n(\theta|\theta_{n-1,j}^{F,m})$ and $X_{n,j}^{P,m} \sim f(x_n|X_{n-1,j}^{F,m})$
 - ▶ Compute weights $w_{n,j}^m = p(y_n|\theta_{n,j}^{P,m}, X_{n,j}^{P,m})$
 - ▶ Resample $P(l_k = l) = \frac{w_{n,l}^m}{\sum_r w_{n,r}^m}$
 - ▶ Update $\theta_{n,j}^{F,m} = \theta_{n,l_j}^{P,m}$ and $X_{n,j}^{F,m} = X_{n,l_j}^{P,m}$
 - ▶ end
 - ▶ Define $\theta_j^m = \theta_{N,j}^{F,m}$
- ▶ end

Note: It is not obvious how h_0 relates to h_n , cf. σ_m vs τ_m in IF1 (King et al. (2016) sets $\tau_m^2 = 5\sigma_m^2$).

Theoretical results

Introduce the likelihood function

$$L(\theta_{0:N}) = \int p_{X_0}(x_0; \theta) \prod_{k=1}^n p_{Y_k|X_k}(y_k|x_k; \theta) p_{X_k|X_{k-1}}(x_k|x_{k-1}; \theta) dx_{0:n} \quad (14)$$

Then define the Bayes map

$$T_\sigma f(\theta) = \frac{\int L(\theta_{0:N}) h(\theta_{0:N}) p(\psi) d\psi d\theta_{0:N-1}}{\int L(\theta_{0:N}) h(\theta_{0:N}) p(\psi) d\psi d\theta_{0:N}} \quad (15)$$

This targets $p(\theta_N|\mathcal{F}_N)$

Theoretical results

Introduce the likelihood function

$$L(\theta_{0:N}) = \int p_{X_0}(x_0; \theta) \prod_{k=1}^n p_{Y_k|X_k}(y_k|x_k; \theta) p_{X_k|X_{k-1}}(x_k|x_{k-1}; \theta) dx_{0:n} \quad (14)$$

Then define the Bayes map

$$T_\sigma f(\theta) = \frac{\int L(\theta_{0:N}) h(\theta_{0:N}) p(\psi) d\psi d\theta_{0:N-1}}{\int L(\theta_{0:N}) h(\theta_{0:N}) p(\psi) d\psi d\theta_{0:N}} \quad (15)$$

This targets $p(\theta_N|\mathcal{F}_N)$

Ionides et al. (2015) shows

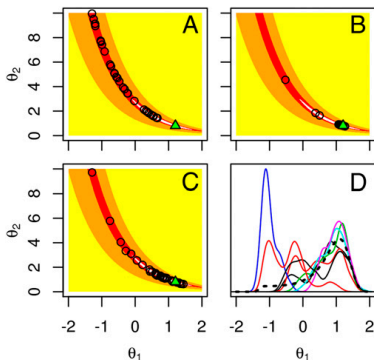
- ▶ For every fixed $\sigma > 0$, $\lim_{m \rightarrow \infty} T_\sigma f(\theta) = f_\sigma(\theta)$ exist.
- ▶ The IF2 output approximates $f_\sigma(\theta)$ as $M, J \rightarrow \infty$
- ▶ $\lim_{\sigma \rightarrow 0} f_\sigma$ approaches a point mass at the MLE if it exists.

This makes it very plausible (but it is not a complete formal proof) that the estimate will have nice properties.

Discussion

Ionides et al. (2015) compares the IF1 (A), IF2 (B) and PMCMC (C) for a simple example. Model $X_n = [\exp(\theta_1), \theta_2 \exp(\theta_1)]$ with $\theta = (1, 1)$. The measurements are drawn from

$$p(Y_n|X_n) \in N\left(X_n, \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}\right)$$



Similarly, Cholera model

Model from King et al. (2008).

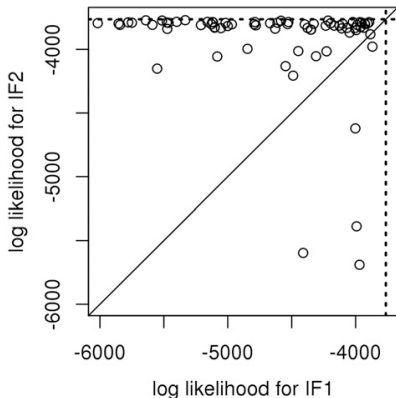


Figure borrowed from Ionides et al. (2015).

The Efficient Iterated Filtering algorithm

From Lindström et al. (2012). **Let** $\hat{\theta}_n = \tilde{\theta}_n^F$.

The Efficient Iterated Filtering algorithm

From Lindström et al. (2012). **Let** $\hat{\theta}_n = \tilde{\theta}_n^F$.

It then follows that $\hat{\theta}_n \rightarrow \theta^*$ as $n \rightarrow \infty$.

The Efficient Iterated Filtering algorithm

From Lindström et al. (2012). **Let** $\hat{\theta}_n = \tilde{\theta}_n^F$.

It then follows that $\hat{\theta}_n \rightarrow \theta^*$ as $n \rightarrow \infty$.

$$\begin{aligned}\hat{\theta}_n &= \tilde{\theta}_n^F \pm \tilde{\theta}_{n-1}^F \\ &= \hat{\theta}_{n-1} + \left(\tilde{\theta}_n^F - \tilde{\theta}_{n-1}^F \right) \\ &= \hat{\theta}_{n-1} + (\tilde{V}_n^P)(\tilde{V}_n^P)^{-1} \left(\tilde{\theta}_n^F - \tilde{\theta}_{n-1}^F \right) \\ &= \theta_{n-1} + \frac{1}{n} (n \tilde{V}_n^P) [(\tilde{V}_n^P)^{-1} \left(\tilde{\theta}_n^F - \tilde{\theta}_{n-1}^F \right) \\ &\quad \pm \nabla \log p(y_n | \tilde{\theta}_{n-1}^F, y_{1:n-1})]\end{aligned}$$

The Efficient Iterated Filtering algorithm

From Lindström et al. (2012). **Let** $\hat{\theta}_n = \tilde{\theta}_n^F$.

It then follows that $\hat{\theta}_n \rightarrow \theta^*$ as $n \rightarrow \infty$.

$$\begin{aligned}\hat{\theta}_n &= \tilde{\theta}_n^F \pm \tilde{\theta}_{n-1}^F \\ &= \hat{\theta}_{n-1} + \left(\tilde{\theta}_n^F - \tilde{\theta}_{n-1}^F \right) \\ &= \hat{\theta}_{n-1} + (\tilde{V}_n^P)(\tilde{V}_n^P)^{-1} \left(\tilde{\theta}_n^F - \tilde{\theta}_{n-1}^F \right) \\ &= \theta_{n-1} + \frac{1}{n} (n \tilde{V}_n^P) [(\tilde{V}_n^P)^{-1} \left(\tilde{\theta}_n^F - \tilde{\theta}_{n-1}^F \right) \\ &\quad \pm \nabla \log p(y_n | \tilde{\theta}_{n-1}^F, y_{1:n-1})]\end{aligned}$$

Define $a_n = 1/n$, $\Phi_n = n \tilde{V}_n^P$ and

$$R_n = (\tilde{V}_n^P)^{-1} \left(\tilde{\theta}_n^F - \tilde{\theta}_{n-1}^F \right) - \nabla \log p(y_n | \theta_{n-1}, y_{1:n-1}).$$

The new algorithm 2

The algorithm is then given by

$$\hat{\theta}_n = \hat{\theta}_{n-1} + a_n \Phi_n(\nabla \log p(y_n | \hat{\theta}_{n-1}, y_{1:n-1}) + R_n) \quad (16)$$

It is possible to show that (these are upper bounds, not sharp bounds!)

$$E[|R_n|] \leq \frac{C_2}{\tau_n J_n} + C_1 \left(\tau_n + \frac{\sigma_n^2}{\tau_n^2} \right) \quad (17)$$

$$\text{Var}_{MC}(R_n) \leq \frac{C_3}{\tau_n^2 J_n} \quad (18)$$

where C_2 and C_3 are constants.

It then follows (...) that $\theta_n \xrightarrow{a.s.} \theta$.

Consider a linear Gaussian model

$$Y_n = X_n + \eta_n, \quad \eta \sim N(0, 1), \quad (19)$$

$$X_n = \phi X_{n-1} + e_n^X, \quad e^X \sim N(0, 1). \quad (20)$$

Augmenting the state vector by introducing ϕ as a state variable leads to the statistical model

$$Y_n = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} X_n \\ \phi_n \end{bmatrix} + \eta_n, \quad (21)$$

$$X_n = X_{n-1} \phi_{n-1} + e_n^X, \quad (22)$$

$$\phi_n = \phi_{n-1} + e_n^\phi, \quad e^\phi \sim N(0, \tau_m^2), \quad (23)$$

where $\tau_m^2 = 10^{-2}/m^2$.

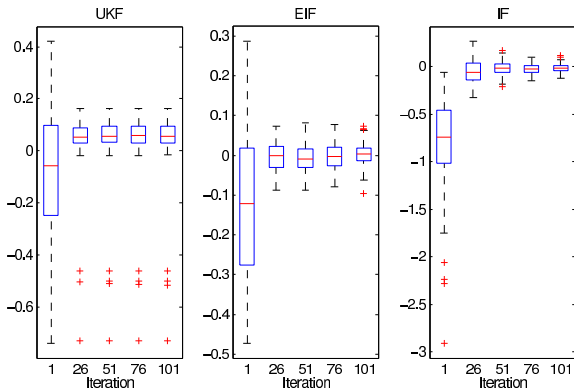


Figure: Convergence to the MLE $\hat{\theta}_m - \hat{\theta}_{MLE}$ of the UKF (left), EIF (middle) and IF(right) for 100 independent realizations from the linear model. Note the different scales.

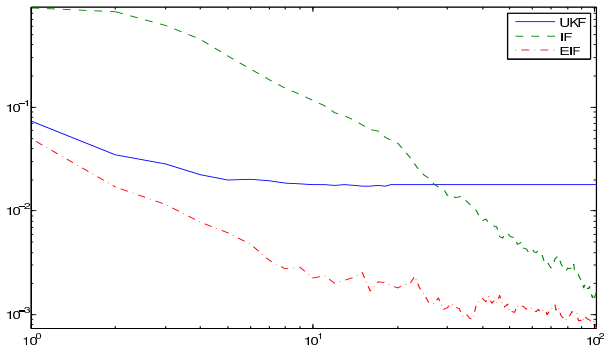


Figure: RMSE between the MLE and the UKF, the IF and EIF for 100 independent realizations from the linear model.

Tuning

- ▶ Metropolis Hastings algorithms are very powerful
- ▶ *and* often easy to implement with a random walk proposal.

However, the resulting Markov chain may not mix very well.

Tuning

- ▶ Metropolis Hastings algorithms are very powerful
- ▶ *and* often easy to implement with a random walk proposal.

However, the resulting Markov chain may not mix very well.

Solution: Adapt the random walk to better explore the parameter space.

Tuning the Iterated Filter algorithm

This is based on Lindström (2013)

- ▶ Knowing the joint density of the augment model, see equation (8), suggests that Maximum Likelihood methods should be used.

Tuning the Iterated Filter algorithm

This is based on Lindström (2013)

- ▶ Knowing the joint density of the augment model, see equation (8), suggests that Maximum Likelihood methods should be used.
- ▶ How? SMC filters can estimate the joint posterior density of the latent parameter process. (EM)

Tuning the Iterated Filter algorithm

This is based on Lindström (2013)

- ▶ Knowing the joint density of the augment model, see equation (8), suggests that Maximum Likelihood methods should be used.
- ▶ How? SMC filters can estimate the joint posterior density of the latent parameter process. (EM)
- ▶ We use a fix-lag SMC-EM algorithm, see Olsson et al. (2008)

Tuning the Iterated Filter algorithm

This is based on Lindström (2013)

- ▶ Knowing the joint density of the augment model, see equation (8), suggests that Maximum Likelihood methods should be used.
- ▶ How? SMC filters can estimate the joint posterior density of the latent parameter process. (EM)
- ▶ We use a fix-lag SMC-EM algorithm, see Olsson et al. (2008)
- ▶ We assume (erroneously) that the latent parameter evolves according to a Gaussian random walk,

$$g_{\Theta_n|\Theta_{n-1}}(\theta_n|\theta_{n-1}; \Sigma) = N(\theta_n; \theta_{n-1}, \sigma_m^2 \Sigma).$$

The Σ matrix is then be estimated with the EM algorithm.

E-step

$$Q(\Sigma, \hat{\Sigma}_{m-1}) = \mathbf{E}[\log f_{X_{0:n}, Y_{1:N}}(X_{0:N}, Y_{1:N}; \theta_{0:N}) \quad (24) \\ + \sum_{n=1}^N \log g_{\Theta_n | \Theta_{n-1}}(\Theta_n | \Theta_{n-1}; \Sigma) | \hat{\Sigma}_{m-1}, Y_{1:N}].$$

M-step

$$\hat{\Sigma}'_m = \frac{1}{N} \mathbf{E} \left[\frac{(\Theta_1 - \Theta_0)(\Theta_1 - \Theta_0)^T}{\tau_m^2 + \sigma_m^2} \quad (25) \right. \\ \left. + \sum_{n=2}^N \frac{(\Theta_n - \Theta_{n-1})(\Theta_n - \Theta_{n-1})^T}{\sigma_m^2} \right] | \hat{\Sigma}_{m-1}, Y_{1:N}]$$

where the expectation is computed using the Monte Carlo smoothing distribution.

There is a non-zero probability that the covariance matrix does not have full rank (cf. Persistent excitation).

Introducing a regularized version

$$\hat{\Sigma}_m = (1 - \varepsilon)\hat{\Sigma}'_m + \varepsilon\Sigma. \quad (26)$$

There is a non-zero probability that the covariance matrix does not have full rank (cf. Persistent excitation).

Introducing a regularized version

$$\hat{\Sigma}_m = (1 - \varepsilon)\hat{\Sigma}'_m + \varepsilon\Sigma. \quad (26)$$

Similar ideas are used for adaptive Metropolis Hastings.

There is a non-zero probability that the covariance matrix does not have full rank (cf. Persistent excitation).

Introducing a regularized version

$$\hat{\Sigma}_m = (1 - \varepsilon)\hat{\Sigma}'_m + \varepsilon\Sigma. \quad (26)$$

Similar ideas are used for adaptive Metropolis Hastings.

Two closely related tuning strategies:

- Replace Σ by $\hat{\Sigma}_m$.

There is a non-zero probability that the covariance matrix does not have full rank (cf. Persistent excitation).

Introducing a regularized version

$$\hat{\Sigma}_m = (1 - \varepsilon)\hat{\Sigma}'_m + \varepsilon\Sigma. \quad (26)$$

Similar ideas are used for adaptive Metropolis Hastings.

Two closely related tuning strategies:

- ▶ Replace Σ by $\hat{\Sigma}_m$.
- ▶ The structure of $\hat{\Sigma}_m$ can be preserved by normalizing the matrix. Define $\hat{\hat{\Sigma}}_m = \hat{\Sigma}_m \cdot \|\Sigma\|_2 / \|\hat{\Sigma}_m\|_2$.

- We have studied a partially observed linear, Gaussian model, given by

$$Y_k = X_k + \eta_k, \quad (27)$$

$$X_k = aX_{k-1} + \sigma e_k, \quad (28)$$

where $\{\eta_k\}$ and $\{e_k\}$ are sequences of pairwise independent standard Gaussian random variables.

- We have simulated $N_{sim} = 100$ independent realizations, each consisting of $N_{obs} = 200$ observations. The true parameters used were $a = 0.7$ and $\sigma = 2$. The filter was iterated $M = 200$ times, using the filter parameters used were $a_m = m^{-1}$, $\tau_m^2 = m^{-1}$, $\sigma_m^2 = m^{-1.3}$ and $\Sigma = I_2$.

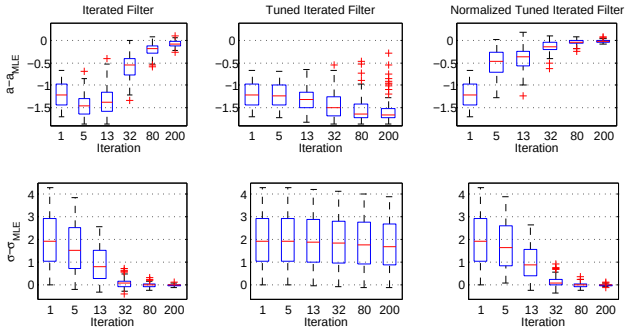


Figure: Difference between the estimate and MLE for the a and σ parameter as a function of the number of iterations.

Scaling is needed to avoid inferring with the stochastic approximation sequence!

Preliminary work on weak convergence

- ▶ (Ionides et. al, 2011) only showed strong consistency (LLN)
- ▶ This does not tell us anything about rates of convergence!!

Preliminary work on weak convergence

- ▶ (Ionides et. al, 2011) only showed strong consistency (LLN)
- ▶ This does not tell us anything about rates of convergence!!
- ▶ Example: α – *stable* random variables with $\alpha \in (1, 2)$.

Preliminary work on weak convergence

- ▶ (Ionides et. al, 2011) only showed strong consistency (LLN)
- ▶ This does not tell us anything about rates of convergence!!
- ▶ Example: α – *stable* random variables with $\alpha \in (1, 2)$.
- ▶ **We think we have shown convergence for the IF1 with CLT properties.** Yet unpublished!

$$\sqrt{N}(\theta_N - \theta) \xrightarrow{d} N(0, \Xi) \quad (29)$$

- ▶ These are somewhat more restrictive than the LLN conditions.

The Gompertz model is given by

$$Y_n = X_n \exp(\eta_n), \quad \eta \sim N(0, \Gamma), \quad (30)$$

$$X_n = K^{1-a} X_{n-1}^a \exp(\epsilon_n), \quad \eta \sim N(0, Q). \quad (31)$$

This non-linear model can be rewritten as a linear model, making likelihood inference based on the Kalman filter feasible.

$$\log Y_n = \log X_n + \eta_n, \quad (32)$$

$$\log X_n = (1 - a) \log K + a \log X_{n-1} + \epsilon. \quad (33)$$

We used the non-linear version of the model with parameters ($K = 2$, $a = 0.8$, $Q = 0.5^2$, $\Gamma = 0.25^2$) to test the algorithm, comparing the estimates from the modified Iterated Filtering algorithms to the Maximum Likelihood estimate. 50 independent realizations were generated, each consisting of 250 observations.

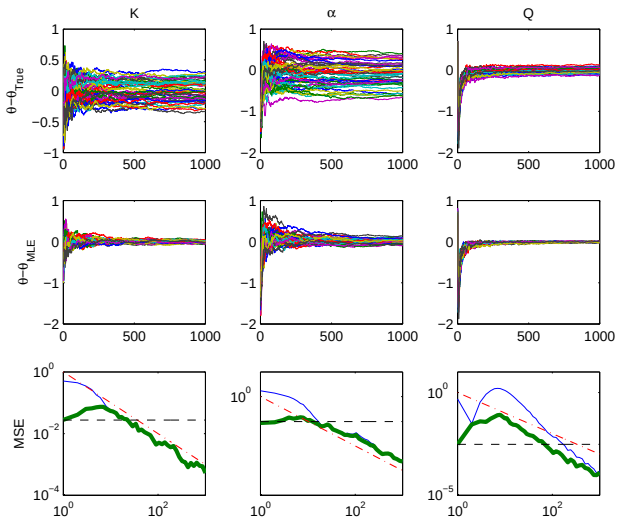


Figure: Convergence for the fast algorithm applied to the Gompertz model. Difference between the estimate and the true parameter (top row), the estimate and the MLE (middle row)

Bibliography I

- Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.
- 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.
- Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Verlag.
- Doucet, A., Godsill, S. J., and Robert, C. P. (2002). Marginal maximum a posteriori estimation using markov chain monte carlo. *Statistics and Computing*, 12(1):77–84.

Bibliography II

- Doucet, A., Jacob, P. E., and Rubenthaler, S. (2013).
Derivative-free estimation of the score vector and observed
information matrix with application to state-space models.
arXiv preprint arXiv:1304.5768.
- Hürzeler, M. and Künsch, H. R. (2001). Approximating and
maximising the likelihood for a general state-space model. In
Sequential Monte Carlo Methods in Practice. Springer Verlag.
- Ionides, E., Bretó, C., and King, A. (2006). Inference for
nonlinear dynamical systems. *Proceedings of the National
Academy of Sciences*, 103(49):18438–18443.
- Ionides, E. L., Bhadra, A., Atchade, Y. F., and King, A. A.
(2011). Iterated filtering. *Annals of Statistics*, 39:1776–1802.
- Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King,
A. A. (2015). Inference for dynamic and latent variable
models via iterated, perturbed bayes maps. *Proceedings of the
National Academy of Sciences*, 112(3):719–724.

Bibliography III

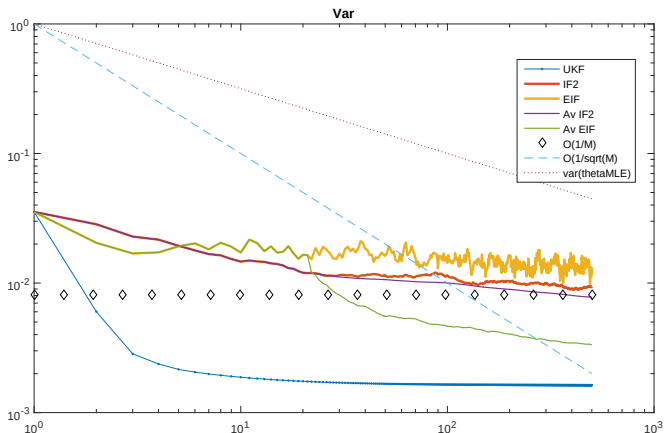
- King, A., Ionides, E., Pascual, M., and Bouma, M. (2008). Inapparent infections and cholera dynamics. *Nature*, 454(7206):877–880.
- King, A. A., Nguyen, D., and Ionides, E. L. (2016). Statistical inference for partially observed markov processes via the r package pomp. *Journal of Statistical Software*.
- Lele, S. R., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. *Ecology letters*, 10(7):551–563.
- Lindström, E. (2013). Tuned iterated filtering. *Statistics & Probability Letters*, 83(9):2077–2080.
- Lindström, E., Ionides, E., Frydendall, J., and Madsen, H. (2012). Efficient Iterated Filtering. In *Proceedings IFAC System Identification*, volume 16, pages 1785–1790.

Bibliography IV

- Lindström, E., Ströjby, J., Brodén, M., Wiktorsson, M., and Holst, J. (2008). Sequential calibration of options. *Computational Statistics & Data Analysis*, 52(6):2877–2891.
- Ljung, L. (1979). Asymptotic behavior of the extended kalman filter as a parameter estimator for linear systems. *Automatic Control, IEEE Transactions on*, 24(1):36–50.
- Nystrup, P., Hansen, B. W., Madsen, H., and Lindström, E. (2015). Regime-based versus static asset allocation: Letting the data speak. *The Journal of Portfolio Management*, 42(1):103–109.
- Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential monte carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179.

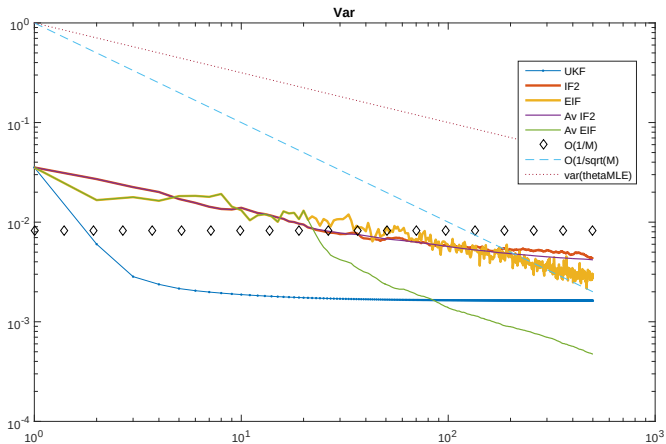
IF2 and EIF, I

Latent AR(1), estimate a_1 parameter. 100 observations, 200 simulations, 500 iterations



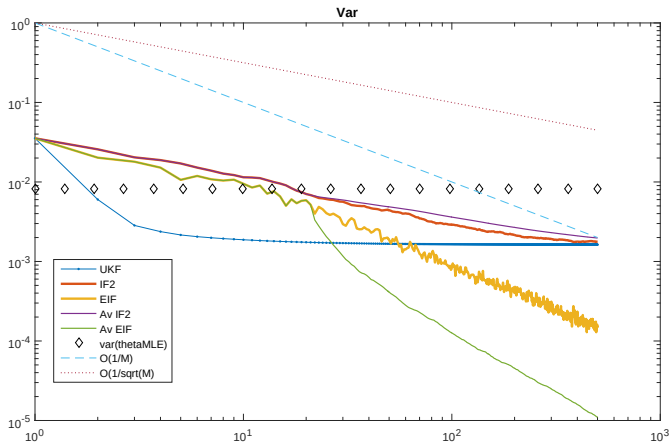
IF2 and EIF, II

Latent AR(1), estimate a_1 parameter. 100 observations, 200 simulations, 500 iterations



IF2 and EIF, III

Latent AR(1), estimate a_1 parameter. 100 observations, 200 simulations, 500 iterations





LUND
UNIVERSITY