On “Particle Filters” for Parameter Estimation

Adam M. Johansen

University of Warwick
a.m.johansen@warwick.ac.uk
www2.warwick.ac.uk/fac/sci/statistics/staff/academic/johansen/talks/

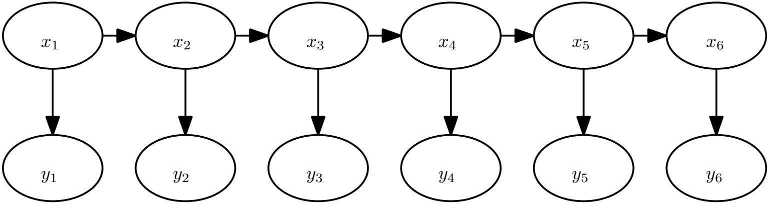
Christmas Workshop on SMC
Imperial College, December 22nd, 2015

Outline

- Background: SMC and PMCMC
- Iterative Lookahead Methods
- Tempering Blocks in SMC
- Hierarchical Particle Filters
- Conclusions

Discrete Time Filtering

Online inference for Hidden Markov Models:



- Given *transition* $f_\theta(x_{n-1}, x_n)$,
- and *likelihood* $g_\theta(x_n, y_n)$,
- use $p_\theta(x_n|y_{1:n})$ to characterize latent state, but,

$$p_\theta(x_n|y_{1:n}) = \frac{\int p_\theta(x_{n-1}|y_{1:n-1}) f_\theta(x_{n-1}, x_n) dx_{n-1} g_\theta(x_n, y_n)}{\int \int p_\theta(x_{n-1}|y_{1:n-1}) f_\theta(x_{n-1}, x'_n) dx_{n-1} g_\theta(x'_n, y_n) dx'_n}$$

isn't often tractable.

Particle Filtering

A (sequential) Monte Carlo (SMC) scheme to approximate the filtering distributions.

A Simple Particle Filter [6]

At $n = 1$:

- Sample $X_1^1, \ldots, X_1^N \sim \mu_\theta$.

For $n > 1$:

- Sample

$$X_n^1, \ldots, X_n^N \sim \frac{\sum_{j=1}^N g_\theta(X_{n-1}^j, y_{n-1}) f_\theta(X_{n-1}^j, \cdot)}{\sum_{k=1}^n g_\theta(X_{n-1}^k, y_{n-1})}$$

- Approximate $p_\theta(dx_n | y_{1:n}), p_\theta(y_{1:n})$ with

$$\widehat{p_\theta}(\cdot | y_{1:n}) = \frac{\sum_{j=1}^N g_\theta(X_n^j, y_n) \delta_{X_n^j}}{\sum_{k=1}^n g_\theta(X_n^k, y_n)}, \frac{\widehat{p_\theta}(y_{1:n})}{\widehat{p_\theta}(y_{1:n-1})} = \frac{1}{n} \sum_{j=1}^N g_\theta(X_n^k, y_n)$$

Online Particle Filters for Offline Systems Identification

Particle Markov chain Monte Carlo (PMCMC) [2]

- Embed SMC within MCMC,
- justified via explicit auxiliary variable construction,
- or in some cases by a pseudomarginal [1] argument.
- Very widely applicable,
- but prone to poor mixing when SMC performs poorly for some θ [9, Section 4.2.1].
- Is valid for *very* general SMC algorithms.

Twisting the HMM (a complement to [10])

Given (μ, f, g) and $y_{1:T}$, introducing
$\boldsymbol{\psi} := (\psi_1, \psi_2, \ldots, \psi_T), \psi_t \in \mathcal{C}_b(\mathsf{X}, (0, \infty))$ and

$$\tilde{\psi}_0 := \int_{\mathsf{X}} \mu(x_1)\, \psi_1(x_1)\, dx_1 \quad \tilde{\psi}_t(x_t) := \int_{\mathsf{X}} f(x_t, x_{t+1})\, \psi_{t+1}(x_{t+1})\, dx_{t+1}$$

we obtain $(\mu_1^{\boldsymbol{\psi}}, \{f_t^{\boldsymbol{\psi}}\}, \{g_t^{\boldsymbol{\psi}}\})$, with

$$\mu_1^{\boldsymbol{\psi}}(x_1) := \frac{\mu(x_1)\psi_1(x_1)}{\tilde{\psi}_0}, \qquad f_t^{\boldsymbol{\psi}}(x_{t-1}, x_t) := \frac{f(x_{t-1}, x_t)\, \psi_t(x_t)}{\tilde{\psi}_{t-1}(x_{t-1})}$$

and the sequence of non-negative functions

$$g_1^{\boldsymbol{\psi}}(x_1) := g(x_1, y_1)\, \frac{\tilde{\psi}_1(x_1)}{\psi_1(x_1)} \tilde{\psi}_0, \qquad g_t^{\boldsymbol{\psi}}(x_t) := g(x_t, y_t)\, \frac{\tilde{\psi}_t(x_t)}{\psi_t(x_t)}.$$

Proposition

For any sequence of bounded, continuous and positive functions $\boldsymbol{\psi}$, let

$$Z_{\boldsymbol{\psi}} := \int_{\mathsf{X}^T} \mu_1^{\boldsymbol{\psi}}(x_1)\, g_1^{\boldsymbol{\psi}}(x_1) \prod_{t=2}^{T} f_t^{\boldsymbol{\psi}}(x_{t-1}, x_t)\, g_t^{\boldsymbol{\psi}}(x_t)\, dx_{1:T}.$$

Then, $Z_{\boldsymbol{\psi}} = p_\theta(y_{1:T})$ for any such $\boldsymbol{\psi}$.

The optimal choice is:

$$\psi_t^*(x_t) := g(x_t, y_t)\, \mathbb{E}\left[\prod_{p=t+1}^{T} g(X_p, y_p) \middle| \{X_t = x_t\} \right], \quad x_t \in \mathsf{X},$$

for $t \in \{1, \ldots, T-1\}$. Then, $Z_{\boldsymbol{\psi}^*}^N = p(y_{1:T})$ with probability 1.

Towards Iterative Auxiliary Particle Filters [7]

$\boldsymbol{\psi}$-Auxiliary Particle Filter

1. Sample $\xi_1^i \sim \mu^{\boldsymbol{\psi}}$ independently for $i \in \{1, \ldots, N\}$.
2. For $t = 2, \ldots, T$, sample independently

$$\xi_t^i \sim \frac{\sum_{j=1}^N g_{t-1}^{\boldsymbol{\psi}}(\xi_{t-1}^j) f_t^{\boldsymbol{\psi}}(\xi_{t-1}^j, \cdot)}{\sum_{j=1}^N g_{t-1}^{\boldsymbol{\psi}}(\xi_{t-1}^j)}, \qquad i \in \{1, \ldots, N\}.$$

Necessary features of $\boldsymbol{\psi}$

1. It is possible to sample from $f_t^{\boldsymbol{\psi}}$.
2. It is possible to evaluate $g_t^{\boldsymbol{\psi}}$.
3. To be useful: $\mathbb{V}(\widehat{Z}_{\boldsymbol{\psi}}^N)$ must be small.

A Recursive Approximtion

Proposition

The sequence $\boldsymbol{\psi}^*$ *satisfies* $\psi_T^*(x_T) = g(x_T, y_T)$, $x_T \in \mathsf{X}$ *and*

$$\psi_t^*(x_t) = g(x_t, y_t)\, f\left(x_t, \psi_{t+1}^*\right), \quad x_t \in \mathsf{X}, \quad t \in \{1, \ldots, T-1\}.$$

Algorithm 1 Recursive function approximations

For $t = T, \ldots, 1$:

1. Set $\psi_t^i \leftarrow g\left(\xi_t^i, y_t\right) f\left(\xi_t^i, \psi_{t+1}\right)$ for $i \in \{1, \ldots, N\}$.
2. Choose ψ_t as a member of Ψ on the basis of $\xi_t^{1:N}$ and $\psi_t^{1:N}$.

Algorithm 2 An iterated auxiliary particle filter with parameters (N_0, k, τ)

1. Initialize: set ${\psi_t}^0 = \mathbf{1}$. $l \leftarrow 0$.
2. Repeat:
   - 2.1 Run a $\boldsymbol{\psi}^l$-APF with N_l particles; set $\hat{Z}_l \leftarrow Z_{\boldsymbol{\psi}^l}^{N_l}$.
   - 2.2 If $l > k$ and $\mathrm{sd}(\hat{Z}_{l-k:l})/\mathrm{mean}(\hat{Z}_{l-k:l}) < \tau$, go to 3.
   - 2.3 Compute $\boldsymbol{\psi}^{l+1}$ using Algorithm 1.
   - 2.4 If $N_{l-k} = N_l$ and the sequence $\hat{Z}_{l-k:l}$ is not monotonically increasing, set $N_{l+1} \leftarrow 2N_l$.
     Otherwise, set $N_{l+1} \leftarrow N_l$.
   - 2.5 Set $l \leftarrow l + 1$. Go to 2a.
3. Run a $\boldsymbol{\psi}^l$-APF. Return $\hat{Z} := Z_{\boldsymbol{\psi}}^{N_l}$.

An Elementary Implementation

Function Approximation .

- Numerically obtain:

$$(m_t^*, \Sigma_t^*, \lambda_t^*) = \arg\min_{(m,\Sigma,\lambda)} \sum_{i=1}^{N} (\mathcal{N}(x_t^i, m, \Sigma) - \lambda \psi_t^i)^2$$

- Set:

$$\psi_t(x_t) := \mathcal{N}(x_t; m_t^*, \Sigma_t^*) + c(N, m_t^*, \Sigma_t^*).$$

Stopping Rule .

- $k = 3$ or $k = 5$ in the following examples
- $\tau = 0.5$

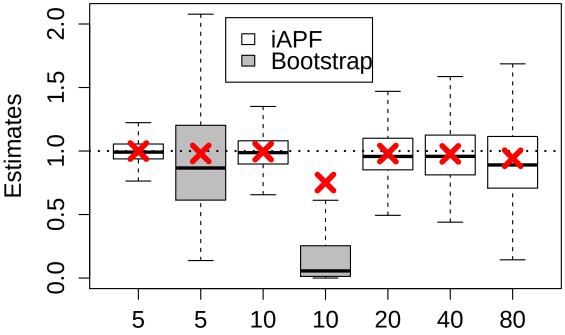
Resampling .

- Multinomial when ESS $< N/2$.

A Linear Gaussian Model: Behaviour with Dimension

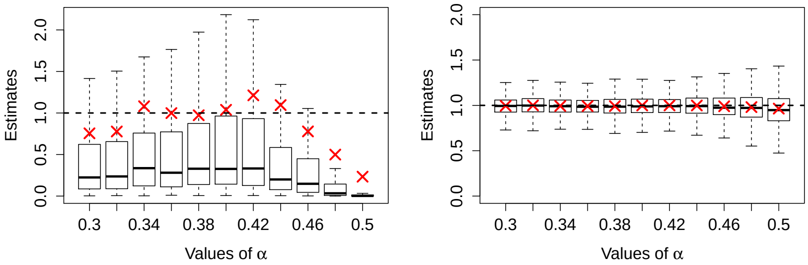
$$\mu = \mathcal{N}\left(\cdot; \mathbf{0}, I_d\right) \qquad f(x,\cdot) = \mathcal{N}\left(\cdot; Ax, I_d\right)$$
$$\text{and } g(x,\cdot) = \mathcal{N}\left(\cdot; x, I_d\right) \qquad \text{where } A_{ij} = 0.42^{|i-j|+1},$$

Box plots of $\hat{Z}/Z$ for different $|\mathsf{X}|$ (1000 replicates; $T = 100$).



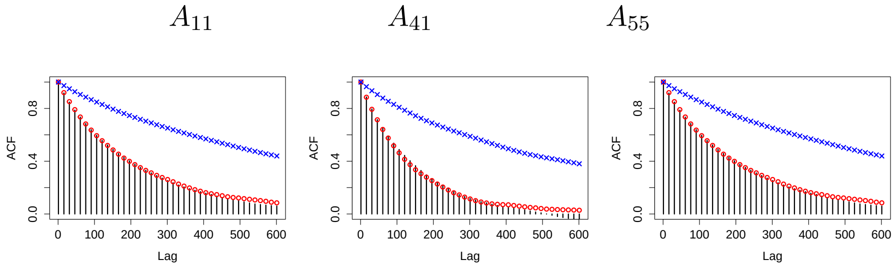
Linear Gaussian Model: Sensitivity to Parameters

Fixing $d = 10$: Bootstrap ($N = 50,000$) / iAPF ($N_0 = 1,000$)



Box plots of $\frac{\hat{Z}}{Z}$ for different values of the parameter α using 1000 replicates.

Linear Gaussian Model: PMMH Empirical Autocorrelations



In this case:

$$
\begin{aligned}
d &= 5 \\
\mu &= \mathcal{N}\left(\cdot; \mathbf{0}, I_d\right) \\
f(x, \cdot) &= \mathcal{N}\left(\cdot; Ax, I_d\right) \\
\text{and } g(x, \cdot) &= \mathcal{N}\left(\cdot; x, 0.25 I_d\right)
\end{aligned}
$$

$$
A = \begin{pmatrix} 0.9 & 0 & 0 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 & 0 \\ 0.1 & 0.2 & 0.6 & 0 & 0 \\ 0.4 & 0.1 & 0.1 & 0.3 & 0 \\ 0.1 & 0.2 & 0.5 & 0.2 & 0 \end{pmatrix},
$$

treated as unknown lower triangular matrix.

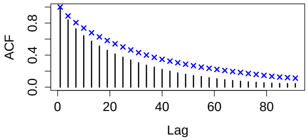
Stochastic Volatility

- A simple stochastic volatility model is defined by:

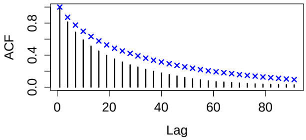
$$\begin{aligned} \mu(\cdot) =& \mathcal{N}(\cdot; 0, \sigma^2/(1-\alpha)^2) \\ f(x,\cdot) =& \mathcal{N}(\cdot; \alpha x, \sigma^2) \\ \text{and } g(x,\cdot) =& \mathcal{N}(\cdot; 0, \beta^2 \exp(x)), \end{aligned}$$

  where $\alpha \in (0,1)$, $\beta > 0$ and $\sigma^2 > 0$ are unknown.
- Considered $T = 945$ observations $y_{1:T}$ corresponding to the mean-corrected daily returns for the GBP/USD exchange rate from 1/10/81 to 28/6/85.

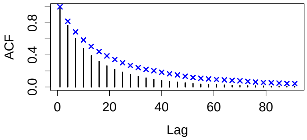
Estimated PMCMC Autocorrelation



a

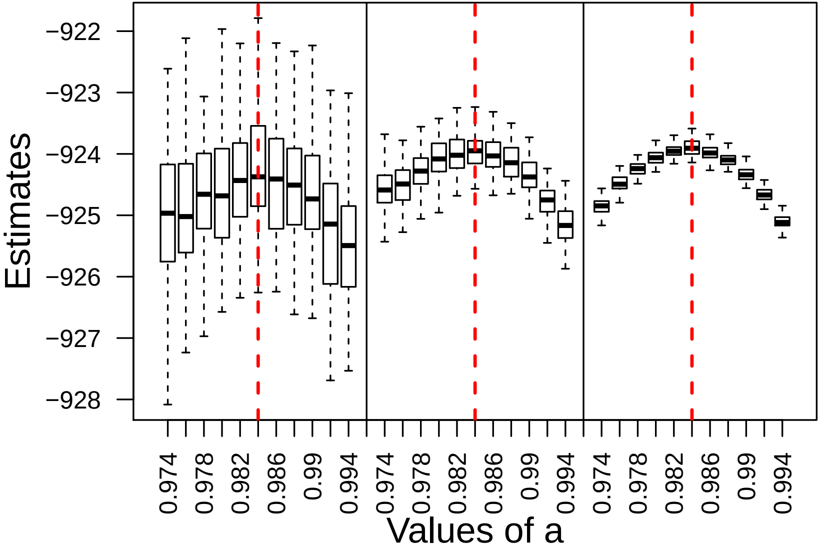


σ

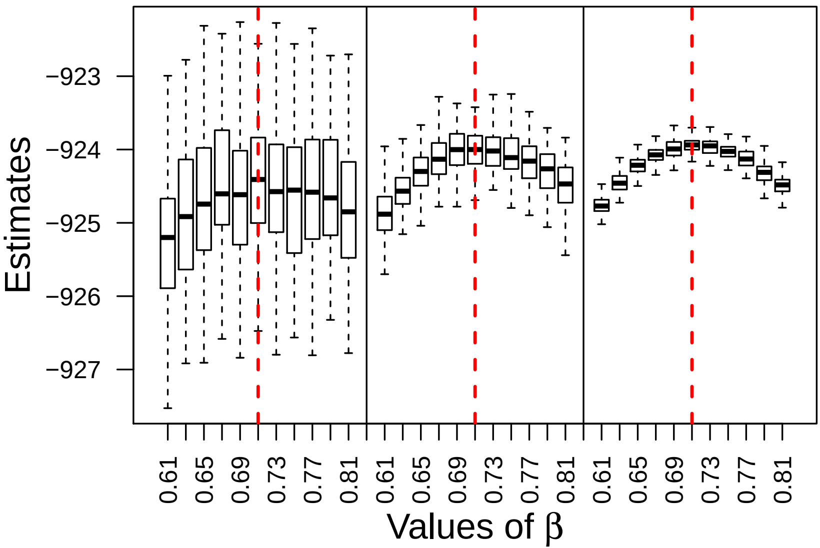


β

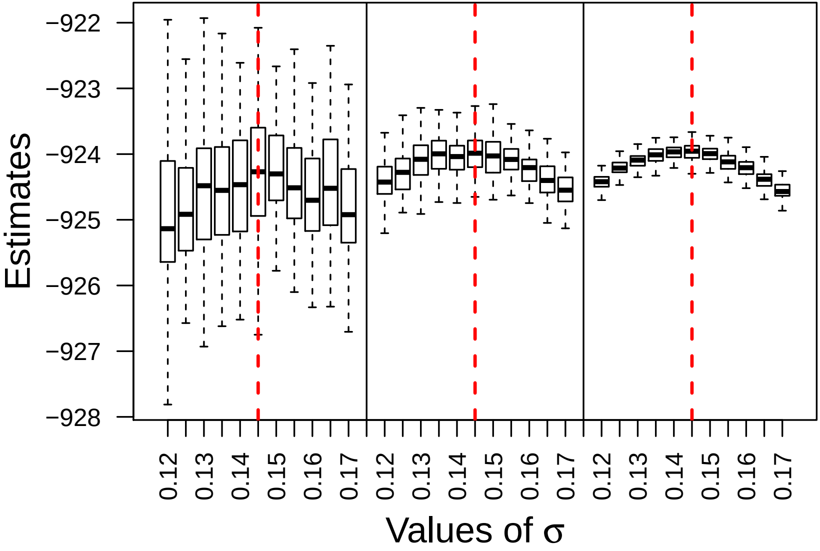
Boostrap $N = 1,000$.
iAPF $N_0 = 100$.
Comparable cost.
150,000 PMCMC iterations.



$Bootstrap : N = 1,000 \ / \ N = 10,000 \ / \ iAPF, N_0 = 100$



$Bootstrap : N = 1,000 \; / \; N = 10,000 \; / \; iAPF, N_0 = 100$



$Bootstrap : N = 1,000 \ / \ N = 10,000 \ / \ iAPF, N_0 = 100$

Block-Sampling and Tempering in Particle Filters

Tempered Transitions [5]

- Introduce each likelihood term gradually, targetting:

$$\pi_{n,m}(x_{1:n}) \propto p(x_{1:n}|y_{1:n-1})p(y_n|x_n)^{\beta_m}$$

  between $p(x_{1:n-1}|y_{1:n-1})$ and $p(x_{1:n}|y_{1:n})$.
- Can improve performance — but to a limited extent.

Block Sampling [4]

- Essentially uses $y_{n:n+L}$ in proposing x_n.
- Can *dramatically* improve performance,
- but requires good analytic approximation of $p(x_{n:n+L}|x_{n-1}, y_{n:n+L})$.

Block-Tempering [8]

- We could combine blocking and tempering strategies.
- Run a simple SMC sampler [3] targetting:

$$\pi^{\theta}_{t,r}(x_{1:t\wedge T}) = \mu_\theta(x_1)^{\beta^1_{(t,r)}} g_\theta(y_1|x_1)^{\gamma^1_{(t,r)}} \cdot \prod_{s=2}^{T\wedge t} f_\theta(x_s|x_{s-1})^{\beta^s_{(t,r)}} g_\theta(y_s|x_s)^{\gamma^s_{(t,r)}}, \qquad (1)$$

  where $\{\beta^s_{(t,r)}\}$ and $\{\gamma^s_{(t,r)}\}$ are $[0,1]$-valued
- for $s \in [\![1,T]\!]$, $r \in [\![1,R]\!]$ and $t \in [\![1,T']\!]$, with $T' = T + L$
- for some $R, L \in \mathbb{N}$.
- Can be *validly* embedded within PMCMC:
  - Terminal likelihood estimate is unbiased.
  - Explicit auxiliary variable construction is possible.

Two Simple Block-Tempering Strategies

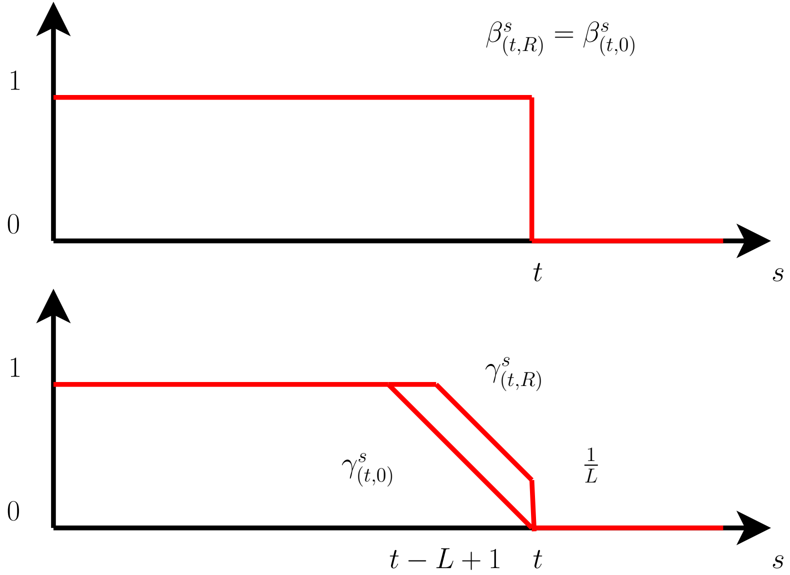
Tempering both likelihood and transition probabilities

$$\beta^s_{(t,r)} = \gamma^s_{(t,r)} = \left(1 \wedge \frac{R(t-s)+r}{RL}\right) \vee 0 \tag{2}$$

Tempering only the observation density

$$\beta^s_{(t,r)} = \mathbb{I}\{s \leq t\} \qquad \gamma^s_{(t,r)} = \left(1 \wedge \frac{R(t-s)+r}{RL}\right) \vee 0, \tag{3}$$

Tempering Only the Observation Density

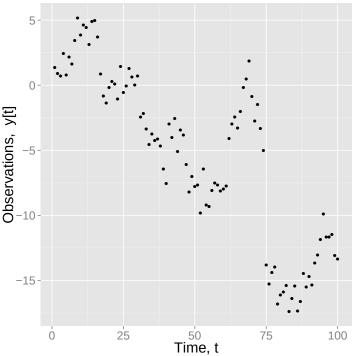


Illustrative Example

- Univariate Linear Gaussian SSM:

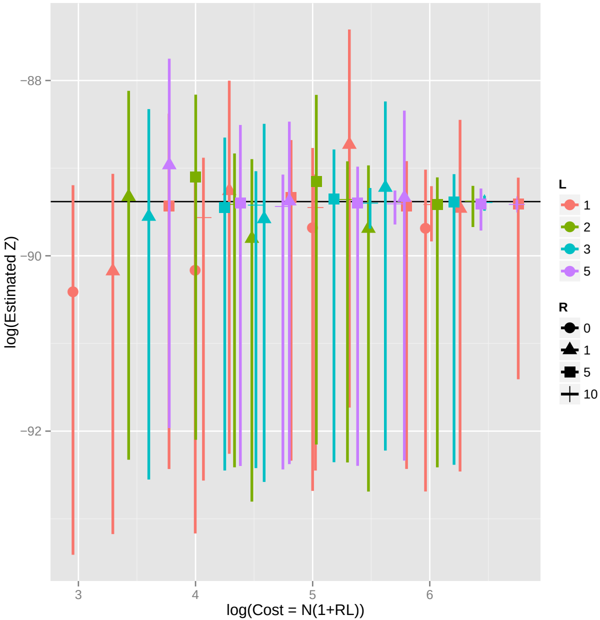
$$\begin{aligned} \text{transition} \quad f(x'|x) =& \mathcal{N}(x'; x, 1) \\ \text{likelihood} \quad g(y|x) =& \mathcal{N}(y; x, 1) \end{aligned}$$

- Artificial jump in observation seqeuence at time 75.
- Cartoon of *model misspecification*
- — a key difficulty with PMCMC.
- Temper only likelihood.
- Use single-site Metropolis-Within Gibbs (standard normal proposal) MCMC moves.

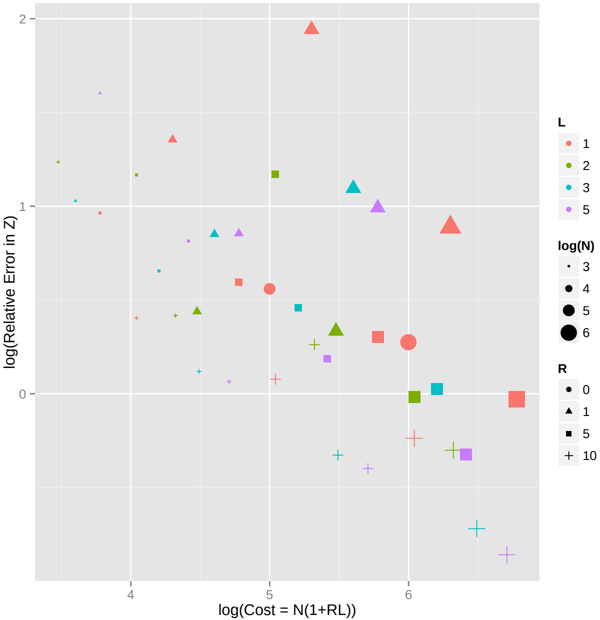


True Filtering and Smoothing Distributions

Estimating the Normalizing Constant, $\widehat{Z}$



Relative Error in $\widehat{Z}$ Against Computational Effort



Performance:

$$(R = 1, L = 1)$$
$$\prec (R > 1, L = 1)$$
$$\prec (R > 1, L > 1)$$

Conclusions

- To fully realise the potential of PMCMC we should exploit its flexibility.
- Even very simple variants on the standard particle filter can significantly improve performance.
- The iAPF can improve performance substantially in some settings.
- Extending the extent of its applicability is ongoing work.
- In principle any *function approximation* scheme can be employed: provided that f_t^{ψ} can be sampled from and g_t^{ψ} evaluated.
- Other [standard and less standard] ideas including blocking and tempering can also be readily employed.

References

[1] C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725, 2009.

[2] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo. *Journal of the Royal Statistical Society B*, 72(3):269–342, 2010.

[3] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B*, 63(3):411–436, 2006.

[4] A. Doucet, M. Briers, and S. Sénécal. Efficient block sampling strategies for sequential Monte Carlo methods. *Journal of Computational and Graphical Statistics*, 15(3):693–711, 2006.

[5] S. Godsill and T. Clapp. Improvement strategies for Monte Carlo particle filters. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, pages 139–158. Springer Verlag, New York, 2001.

[6] N. J. Gordon, S. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, April 1993.

[7] P. Guarniero, A. M. Johansen, and A. Lee. The iterated auxiliary particle filter. ArXiv mathematics e-print 1511.06286, ArXiv Mathematics e-prints, 2015.

[8] A. M. Johansen. On blocks, tempering and particle MCMC for systems identification. In *Proceedings of 17th IFAC Symposium on System Identification*, Beijing, China., 2015. IFAC.

[9] J. Owen, D. J. Wilkinson, and C. S. Gillespie. Likelihood free inference for markov processes: a comparison. *Statistical Applications in Genetics and Molecular Biology*, 2015. In press.

[10] N. Whiteley and A. Lee. Twisted particle filters. *Annals of Statistics*, 42(1):115–141, 2014.