More of the SAME?

Sequential and Pseudomarginal Monte Carlo for Point Estimation in Latent Variable Models

Adam M. Johansen

Collaborators: Manuel Davy, Arnaud Doucet and Axel Finke

a.m.johnsen@warwick.ac.uk

Imperial College — 6th February, 2014

Outline

- Background:
  - Marginal MLEs
  - SAME: An MCMC Scheme
- Sequential Monte Carlo
  - The SMC Method
  - A Population-Based SAME Method
  - Examples
- Pseudomarginal Methods
  - The Pseudomarginal Method
  - More of the SAME: multiple extensions of the space
  - Example
  - Even more of the SAME: complex extensions of the space
  - Examples

Background

- Marginal MLEs
- SAME: An MCMC Scheme

Maximum {Likelihood|*a Posteriori*} Estimation

- Consider a model with:
  - parameters, θ,
  - latent variables, x, and
  - observed data, y.
- Aim to maximise marginal likelihood

$$p(y|\theta) = \int p(x, y|\theta)dx$$

or posterior

$$p(\theta|y) \propto \int p(x, y|\theta)p(\theta)dx.$$

- Traditional approach is Expectation-Maximisation (EM)
  - Requires objective function in closed form.
  - Susceptible to trapping in local optima.

A Probabilistic Approach

- Optimization and probability $\rightsquigarrow$ simulated annealing.
- A distribution of the form

$$\pi(\theta|y) \propto p(\theta)p(y|\theta)^{\gamma}$$

  will become concentrated, as $\gamma \to \infty$ on the maximisers of $p(y|\theta)$ under weak conditions.
- Why not target $\pi(\theta|y)$ using MCMC?

Adapted from (Hwang, 1980; Theorem 2.1).

Assume:

- $p(\theta)$ and $p(y|\theta)$ are α-Lipschitz continuous in θ
- $\log(p(\theta)) \in \mathcal{C}^3(\mathbb{R}^n)$ and $\log p(y|\theta) \in \mathcal{C}^3(\mathbb{R}^n)$.
- Θ_{ML} is a non-empty, countable set which is nowhere dense;
- $p(\theta) \leq M < \infty$; $p(\theta) > 0 \forall \theta \in \Theta_{ML}$
- $p(y|\theta) \leq M' < \infty$
- For some $k < \sup p(y|\theta)$, $\{\theta : p(y|\theta) \geq k\}$ is compact.

Then:

$$\lim_{\gamma \to \infty} \pi_\gamma(dt) \propto \sum_{\theta_{ml} \in \Theta_{ML}} \alpha(\theta_{ml}) \delta_{\theta_{ml}}(dt), \tag{1}$$

$$\alpha(\theta_{ml}) = \det \left[- \left. \frac{\partial^2 \log p(y|\theta)}{\partial \theta_m \partial \theta_n} \right|_{\theta = \theta_{ml}} \right]^{-1/2} \tag{2}$$

State Augmentation for Maximisation of Expectations

Data Augmentation: Synthetic distributions of the form:

$$\bar{\pi}_\gamma(\theta, x_{1:\gamma}|y) \propto p(\theta) \prod_{i=1}^{\gamma} p(x_i, y|\theta)$$

admit the marginals

$$\bar{\pi}_\gamma(\theta|y) \propto p(\theta)p(y|\theta)^\gamma.$$

SAME Algorithm (Doucet, Godsill and Robert, 2002):

- $t = 0$: Initialise $(\theta_0, X_{0,1})$ arbitrarily.
- For $t = 1, \ldots, T$:
  - If $\gamma(t) > \gamma(t-1)$: Set $(X_{t-1,\gamma(t-1)+1}, \ldots, X_{t-1,\gamma(t)})$ arbitrarily.
  - Sample $(\theta_t, X_{t,1}, \ldots, X_{t,\gamma(t)}) \sim K_{\gamma(t)}(\theta_{t-1,1}, X_{t-1,1} \ldots, X_{t-1,\gamma(t)}, \cdot)$.

  Where K_γ is $\bar{\pi}_\gamma$-invariant.

NB An inhomogeneous Markov chain.

Sequential Monte Carlo

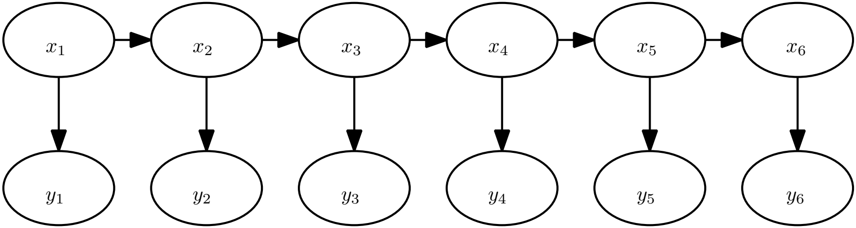
- The SMC Method
- A Population-Based SAME Method
- Examples

SMC: A Motivating Example — Filtering

- Let $X_1, \ldots$ denote the position of an object which follows Markovian dynamics.
- Let $Y_1, \ldots$ denote a collection of observations:

$$Y_i | \{X_i = x_i\} \sim g(\cdot | x_i).$$

- We wish to estimate, as observations arrive, $p(x_{1:t} | y_{1:t})$.
- A recursion obtained from Bayes rule exists but is intractable in most cases.



More Generally

- Really tracking a sequence of distributions, $p_t \ldots$
- on increasing state spaces.
- Other problems with the same structure exist.
- Any problem of sequentially approximating a sequence of such distributions, p_t, can be addressed in the same way.

Sequential Importance Resampling

At time t, $t \geq 2$. (Given $\{X_{1:t-1}^{(i)}\}_{i=1}^{N}$ approximating $p_{t-1}(x_{1:t-1})$).

Sampling Step

For $i = 1 : N$:

sample $X_t^{(i)} \sim q_t\left(\cdot \middle| X_{1:t-1}^{(i)}\right)$.

Resampling Step

For $i = 1 : N$:

$$\text{compute } w_t\left(X_{1:t}^{(i)}\right) = \frac{p_t\left(X_{1:t}^{(i)}\right)}{p_{t-1}\left(X_{1:t-1}^{(i)}\right) q_t\left(X_t^{(i)} \middle| X_{1:t-1}^{(i)}\right)}$$

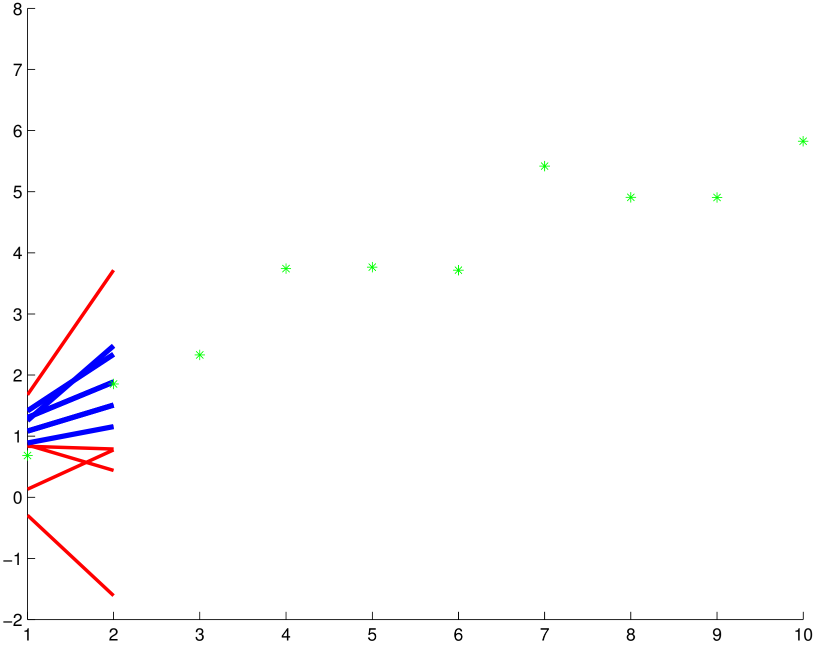
and $W_t^{(i)} = \frac{w_t\left(X_{1:t}^{(i)}\right)}{\sum_{j=1}^{N} w_t\left(X_{1:t}^{(j)}\right)}$

For $i = 1 : N$:

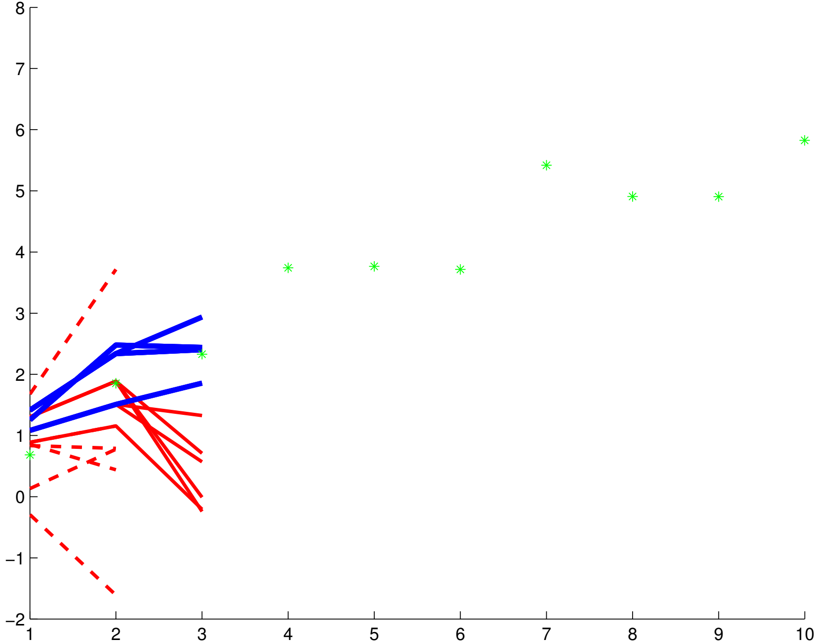
sample $A_t^{(i)} \sim \sum_{j=1}^{N} W_t^{(j)} \delta_j$

retain $\left\{X_{1:t}^{(A_t^i)}\right\}_{i=1}^{N}$

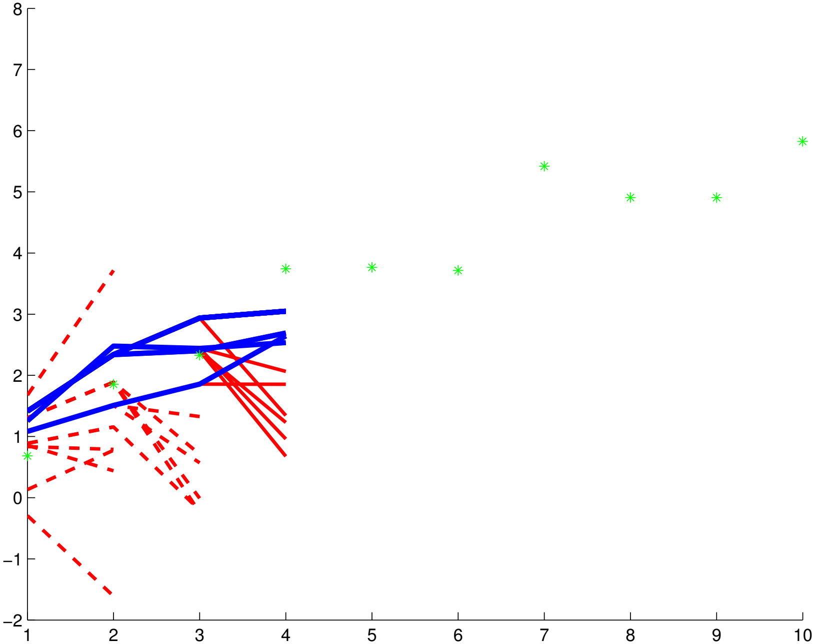
Iteration 2



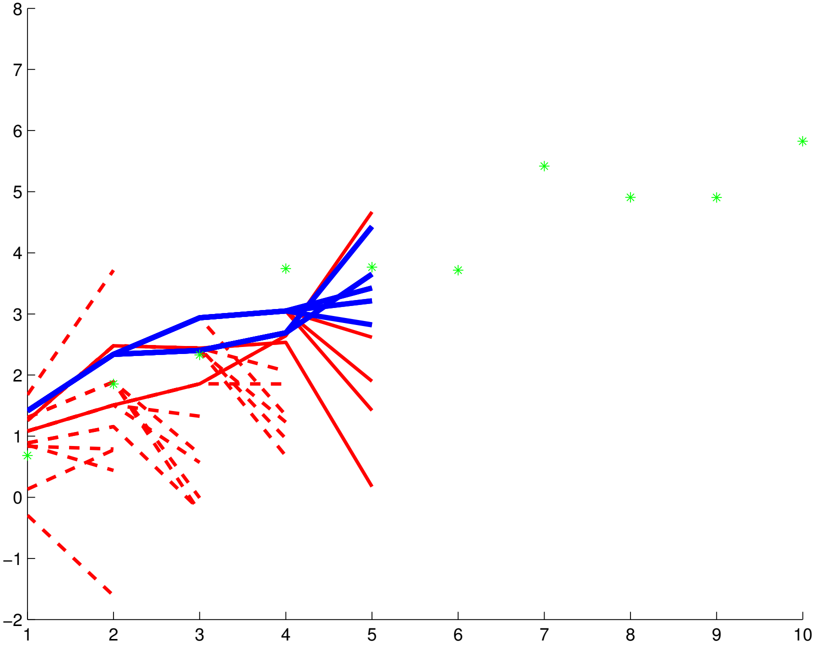
Iteration 3



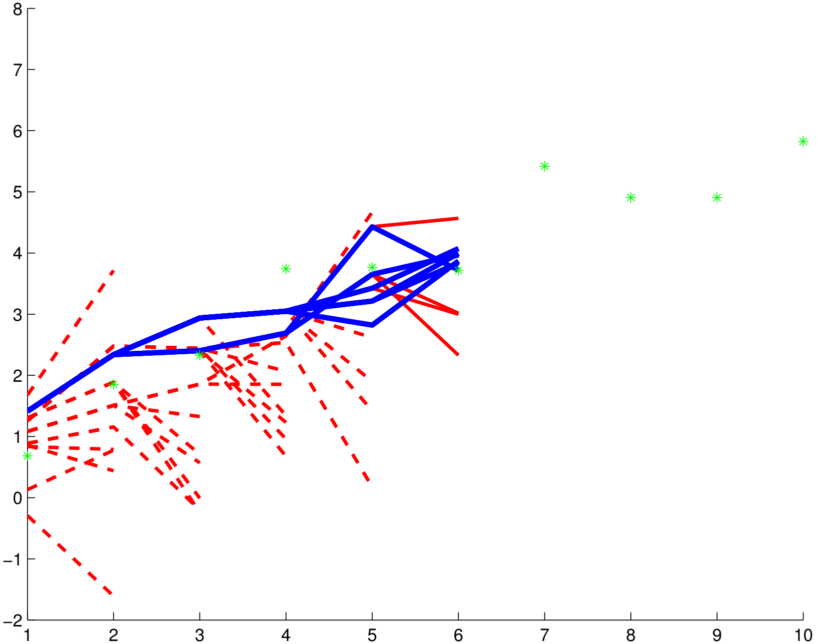
Iteration 4



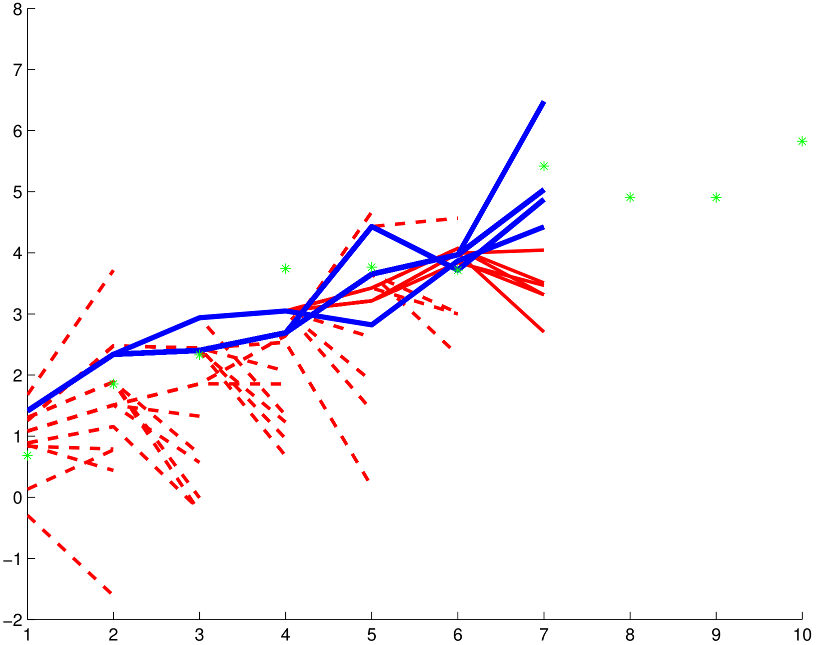
Iteration 5



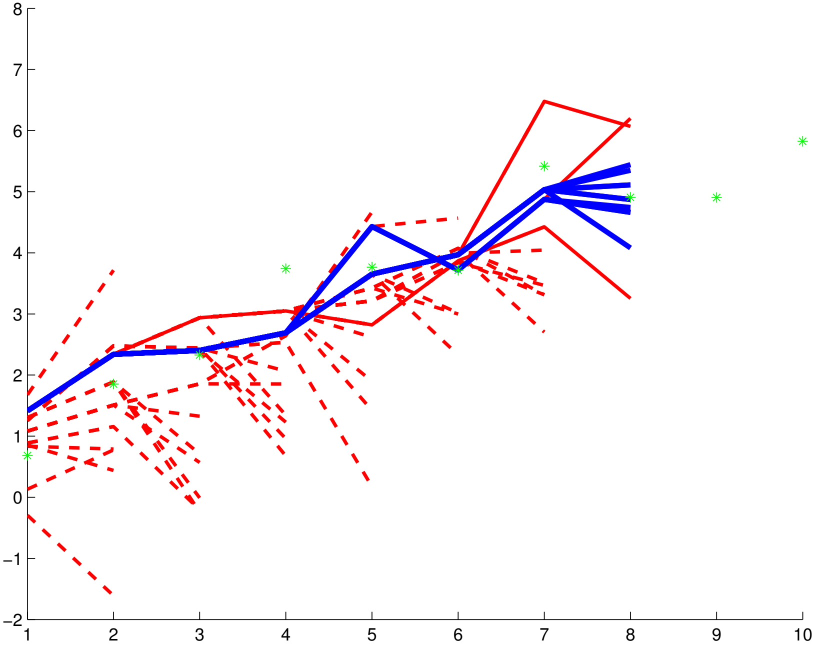
Iteration 6



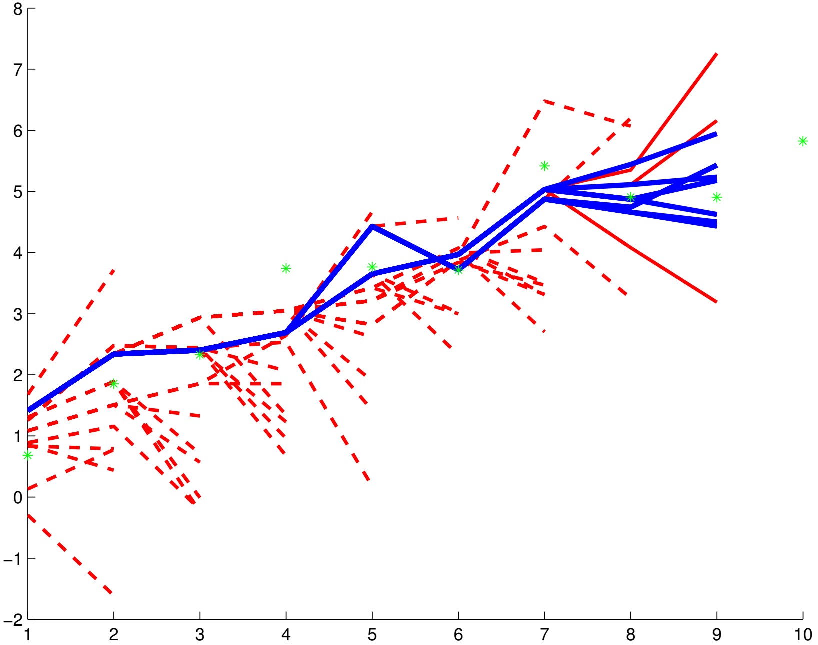
Iteration 7



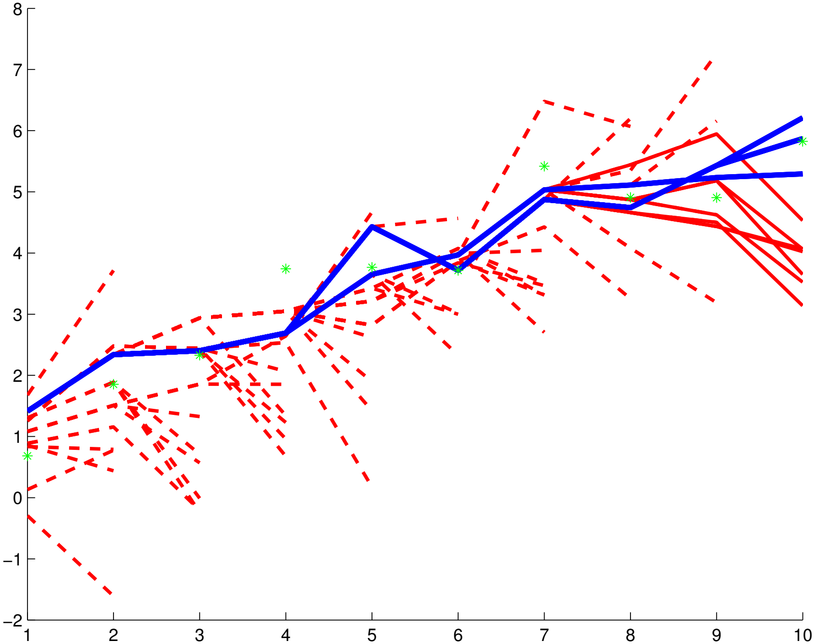
Iteration 8



Iteration 9



Iteration 10



SMC Samplers (Del Moral et al., 2006)

Can be used to sample from *any* sequence of distributions:

- Given a sequence of *target* distributions, η_n, on $E_n \ldots$,
- construct a synthetic sequence $\widetilde{\eta}_n$ on spaces $\bigotimes_{p=1}^{n} E_p$
- by introducing Markov kernels, L_p from E_{p+1} to E_p:

$$\widetilde{\eta}_n(x_{1:n}) = \eta_n(x_n) \prod_{p=1}^{n-1} L_p\left(x_{p+1}, x_p\right),$$

- These distributions
  - have the target distributions as final time marginals,
  - have the correct structure to employ SMC techniques.

SMC Outline

- Given a sample $\{X_{1:n-1}^{(i)}\}_{i=1}^{N}$ targeting $\widetilde{\eta}_{n-1}$,
- sample $X_n^{(i)} \sim K_n(X_{n-1}^{(i)}, \cdot)$,
- calculate

$$W_n(X_{1:n}^{(i)}) = \frac{\eta_n(X_n^{(i)}) L_{n-1}(X_n^{(i)}, X_{n-1}^{(i)})}{\eta_{n-1}(X_{n-1}^{(i)}) K_n(X_{n-1}^{(i)}, X_n^{(i)})}.$$

- Resample, yielding: $\{X_{1:n}^{(i)}\}_{i=1}^{N}$ targeting $\widetilde{\eta}_n$.
- Hints that we'd like to use

$$L_{n-1}(x_n, x_{n-1}) = \frac{\eta_{n-1}(x_{n-1}) K_n(x_{n-1}, x_n)}{\int \eta_{n-1}(x'_{n-1}) K_n(x'_{n-1}, x_n) dx_{n-1}}.$$

Recall: Maximum {Likelihood|*a Posteriori*} Estimation

- A model with:
  - parameters, θ,
  - latent variables, x, and
  - observed data, y.
- Aim to maximise Marginal likelihood

$$p(y|\theta) = \int p(x, y|\theta) dx$$

or posterior

$$p(\theta|y) \propto \int p(x, y|\theta) p(\theta) dx$$

- Using

$$\bar{\pi}_\gamma(\theta, x_{1:\gamma}|y) \propto p(\theta) \prod_{i=1}^{\gamma} p(x_i, y|\theta)$$

Maximum Likelihood via SMC

- Use a sequence of distributions $\eta_n = \bar{\pi}_{\gamma_n}$ for some $\{\gamma_n\}$.
- The MCMC approach (Doucet et al., 2002).
  - Requires slow "annealing".
  - Separation between distributions is large.
  - Mixes poorly as γ increases.
- Using SMC has some substantial advantages:
  - Introducing bridging distributions, for $\gamma = \lfloor \gamma \rfloor + \langle \gamma \rangle$, of:

$$\bar{\pi}_\gamma(\theta, x_{1:\lfloor \gamma \rfloor + 1} | y) \propto p(\theta) \boldsymbol{p(x_{\lfloor \gamma \rfloor + 1}, y | \theta)^{\langle \gamma \rangle}} \prod_{i=1}^{\lfloor \gamma \rfloor} p(x_i, y | \theta)$$

    is straightforward.
  - Population of samples improves robustness.
  - It is less dependent upon mixing of K_γ.

Algorithms

- A generic SMC sampler can be written down directly...
- An easy case:
  - Sample from $p(x_t|y,\theta_{t-1})$ and $p(\theta_t|x_t,y)$.
  - Weight according to $p(y|\theta_{t-1})^{\gamma_t-\gamma_{t-1}}$.
- The general case:
  - Sample existing variables from a π_t-invariant kernel:

$$(\theta_t, X_{t,1:\gamma_{t-1}}) \sim K_t((\theta_{t-1}, X_{t-1}), \cdot).$$

  - Sample new variables from an arbitrary proposal:

$$X_{t,\lceil\gamma_{t-1}\rceil+1:\lceil\gamma_t\rceil} \sim q(\cdot|\theta_t).$$

  - Use combination of time-reversal and optimal auxiliary kernel.
  - Weight expression does not involve the marginal likelihood.

initialisation: $t = 1$**:**

sample $\left\{\left(\theta_1^{(i)}, X_1^{(i)}\right) \overset{\text{iid}}{\sim} \nu\right\}_{i=1}^{N}$

calculate $W_1^{(i)} \propto \frac{\pi_{\gamma_1}(\theta_1^{(i)}, X_1^{(i)})}{\nu(\theta_1^{(i)}, X_1^{(i)})}$ $\qquad \left(\sum_{i=1}^{N} W_1^{(i)} = 1\right)$

for $t = 2$ to T **do**

resample

sample
$$\left\{\begin{array}{l} \left(\theta_t^{(i)}, X_{t,1:\lceil\gamma_{t-1}\rceil}^{(i)}\right) \sim K_{t-1}\left(\theta_{t-1}^{(i)}, X_{t-1}^{(i)}; \cdot\right) \\ \left\{X_{t,j}^{(i)} \sim q(\cdot|\theta_t^{(i)})\right\}_{j=\lceil\gamma_{t-1}\rceil+1}^{\lfloor\gamma_t\rfloor} \quad \text{if } \lceil\gamma_{t-1}\rceil < \lfloor\gamma_t\rfloor \\ X_{t,\lceil\gamma_t\rceil}^{(i)} \sim q_{\langle\gamma_t\rangle}(\cdot|\theta_t^{(i)}) \text{ if } \lceil\gamma_{t-1}\rceil < \lceil\gamma_t\rceil \neq \gamma_t \end{array}\right\}_{i=1}^{N}$$

calculate

$$W_t^{(i)} \propto \frac{p(y, X_{t,\lceil\gamma_{t-1}\rceil}|\theta)^{1\wedge\gamma_t - \lfloor\gamma_{t-1}\rfloor}}{p(y, X_{t,\lceil\gamma_{t-1}\rceil}|\theta)^{\langle\gamma_t\rangle}} \prod_{j=\lceil\gamma_{t-1}\rceil+1}^{\lfloor\gamma_t\rfloor} \frac{p(y, X_{t,j}|\theta_t)}{q(X_{t,j}|\theta_t)} \left(\frac{p(y, X_{t,\lceil\gamma_t\rceil}|\theta_t)^{\langle\gamma_t\rangle}}{q_{\langle\gamma_t\rangle}(X_{t,\lceil\gamma_t\rceil}|\theta_t)}\right)^{I}$$

with $I = \mathbb{I}(\lceil\gamma_t\rceil > \lfloor\gamma_t\rfloor \geq \lceil\gamma_{t-1}\rceil)$.

end for

Toy Example (using known marginal likelihood)

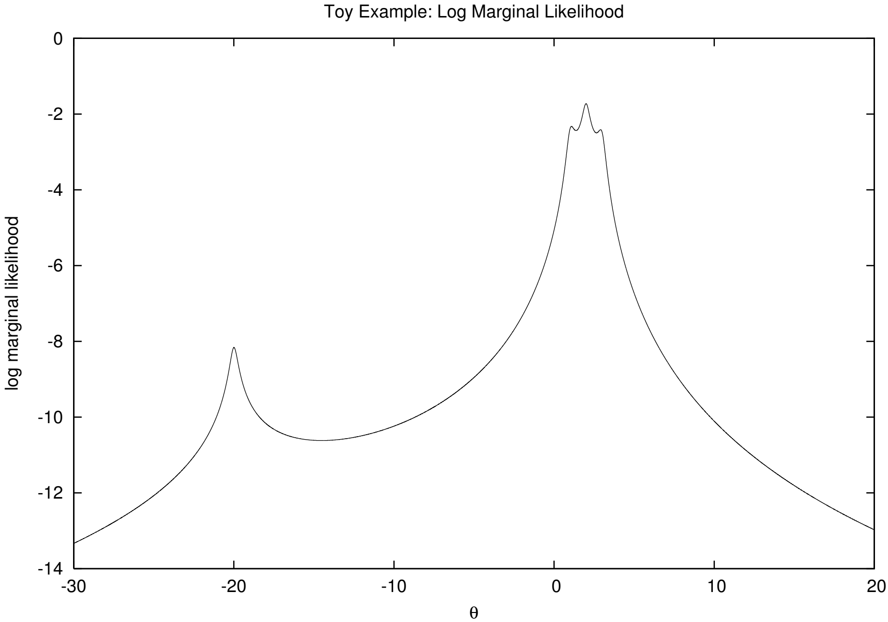
- Student t-distribution of unknown location parameter θ with $\nu = 0.05$.
- Four observations are available, $y = (-20, 1, 2, 3)$.
- Log likelihood is:

$$\log p(y|\theta) = -0.525 \sum_{i=1}^{4} \log\left(0.05 + (y_i - \theta)^2\right).$$

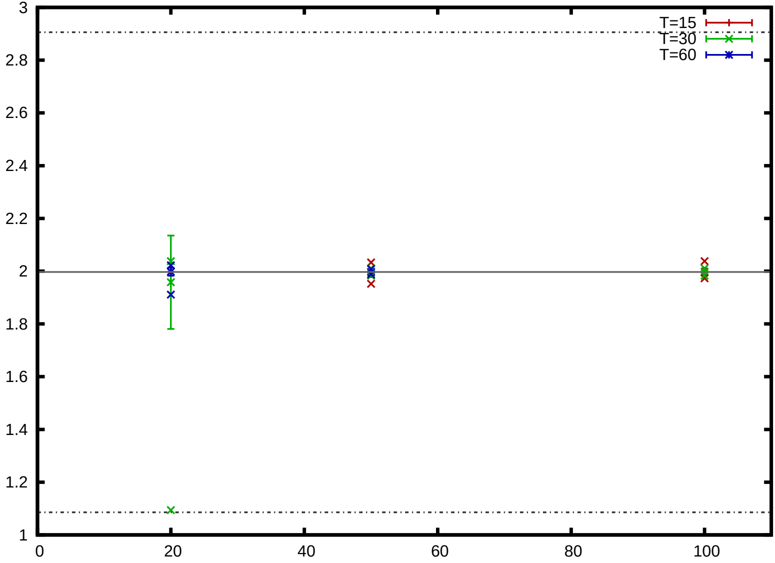
- Global maximum is at 1.997.
- Local maxima at $\{-19.993, 1.086, 2.906\}$.
- Complete log likelihood ($X_i \sim \mathcal{G}\text{a}$):

$$\log p(y, z|\theta) = -\sum_{i=1}^{4} \left[0.475 \log x_i + 0.025 x_i + 0.5 x_i (y_i - \theta)^2\right].$$

Toy Example: Marginal Likelihood



Toy Example: SMC Method using Gibbs Kernels



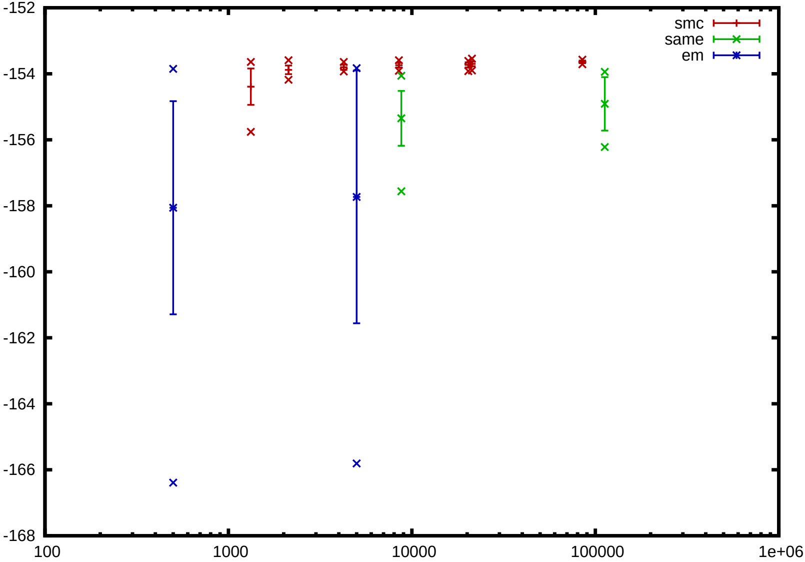
Example: Gaussian Mixture Model – MAP Estimation

- Likelihood $p(y|x, \omega, \mu, \sigma) = \mathcal{N}(y|\mu_x, \sigma_x^2)$.
- Marginal likelihood $p(y|\omega, \mu, \sigma) = \sum_{j=1}^{3} \omega_j \mathcal{N}(y|\mu_j, \sigma_j^2)$.
- Diffuse conjugate priors were employed:

$$\begin{aligned} \omega &\sim \mathcal{D}i\left(\delta\right) \\ \sigma_i^2 &\sim \mathcal{IG}\left(\frac{\lambda_i + 3}{2}, \frac{\beta_i}{2}\right) \\ \mu_i|\sigma_i^2 &\sim \mathcal{N}\left(\alpha_i, \sigma_i^2/\lambda_i\right), \end{aligned}$$

- **All full conditional distributions of interest are available.**
- **Marginal posterior can be calculated.**

3 Component GMM (Roeders Galaxy Data Set)



Pseudomarginal Monte Carlo

- The Pseudomarginal Method
- More of the SAME: multiple extensions of the space
- Example
- Even more of the SAME: complex extensions of the space
- Examples

The Pseudomarginal Method

- Marginal MH-Acceptance Probability:

$$1 \wedge \frac{\pi(\theta')Q(\theta',\theta)}{\pi(\theta)Q(\theta,\theta')}$$

- But $\pi(\theta)$ isn't tractable: how about using:

$$1 \wedge \frac{\widehat{\pi}(\theta')Q(\theta',\theta)}{\widehat{\pi}(\theta)Q(\theta,\theta')}$$

where

$$\widehat{\pi}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \frac{\pi(\theta, X_i)}{q(X_i)} \qquad X_i \overset{\text{iid}}{\sim} q$$

- Suggests two algorithms (Beaumont, 2003):
  - Monte Carlo within Metropolis
  - Grouped Independence Metropolis Hastings

Pseudomarginal Methods: GIMH is "Exact"

- Extended Target (Andrieu & Roberts, 2009):

$$\widetilde{\pi}(\theta, x_1, \ldots, x_m) = \sum_{j=1}^{m} \frac{1}{m} \pi(\theta, x_j) \prod_{k \neq j} q(x_k)$$

$$= \frac{1}{m} \sum_{j=1}^{m} \frac{\pi(\theta, x_j)}{q(x_j)} \cdot \prod_{k=1}^{m} q(x_k) = \hat{\pi}(\theta) \prod_{k=1}^{m} q(x_k)$$

- The acceptance probability becomes:

$$1 \wedge \frac{\widetilde{\pi}(\theta', x'_1, \ldots, x'_m) Q(\theta', \theta) \prod_{j=1}^{m} q(x_j)}{\widetilde{\pi}(\theta, x_1, \ldots, x_m) Q(\theta, \theta') \prod_{j=1}^{m} q(x'_j)} = 1 \wedge \frac{\hat{\pi}(\theta') Q(\theta', \theta)}{\hat{\pi}(\theta) Q(\theta, \theta')}$$

- NB MCWM is *not* exact… but perhaps we don't care.

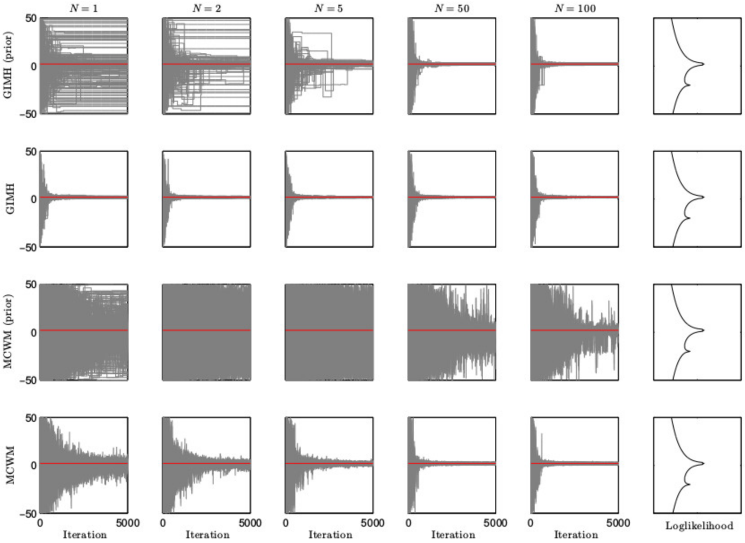
A Pseudomarginal SAME Algorithm

- We'd like to target $\pi_\gamma(\theta|y) \propto p(\theta)p(y|\theta)^\gamma$.
- Why not use the pseudomarginal approach, considering instead:

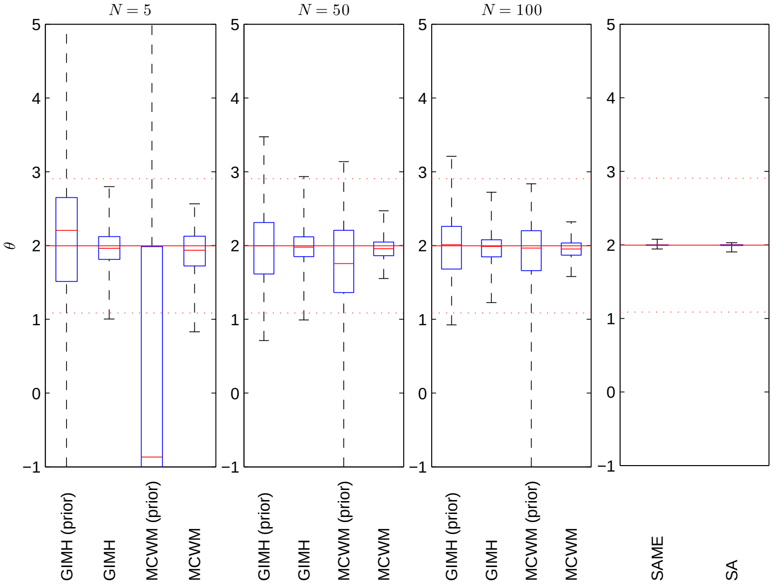
$$\widetilde{\pi}_\gamma(\pi, x^1_{1:m}, \ldots, x^\gamma_{1:m}) = p(\theta) \prod_{i=1}^{\gamma} \sum_{j=1}^{m} \frac{1}{m} \frac{p(x^i_j, y|\theta)}{q(x^i_j|\theta)} \prod_{k=1}^{m} q(x^i_k|\theta)$$

- Expect behaviour like simulated annealing for large m.

The Student t Model Revisited



Summary of 200 Runs



What about complicated latent variable structures?

- Actually, pseudomarginal algorithms are more flexible.
- We're especially interested in particle MCMC implementations (Andrieu et al., 2010):
  - Particle Marginal Metropolis-Hastings(PMMH)
  - MCWM variant of PMMH
  - Particle Gibbs (with ancestor sampling)
- State-space models are the real motivation for this methodology.
- Many other complex models could be addressed using this technique.

Linear Gaussian Hidden Markov Model

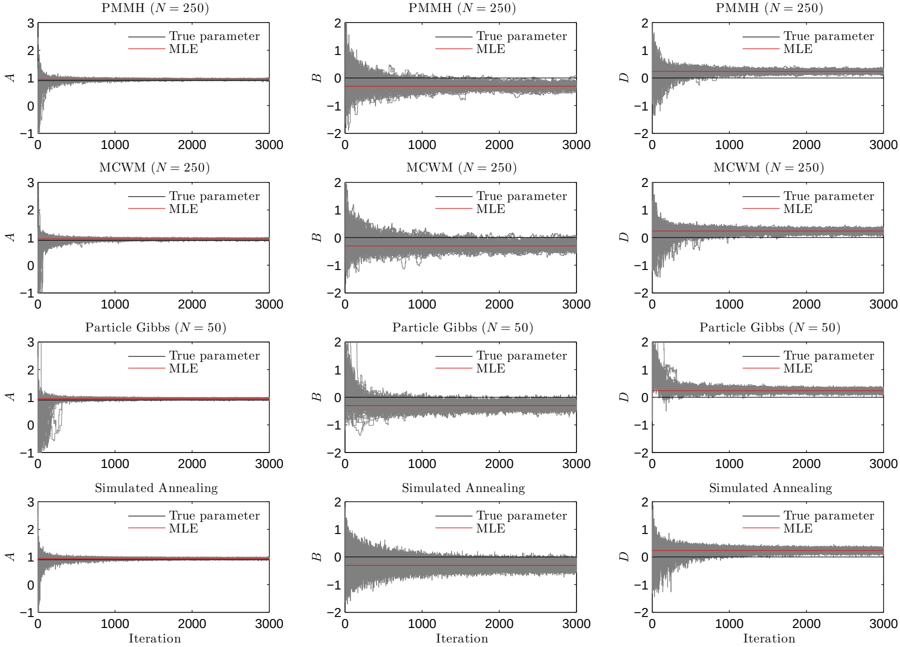
- Model:

$$\begin{aligned} X_t =& AX_{t-1} + BU_t \\ Y_t =& X_t + DV_t \end{aligned}$$

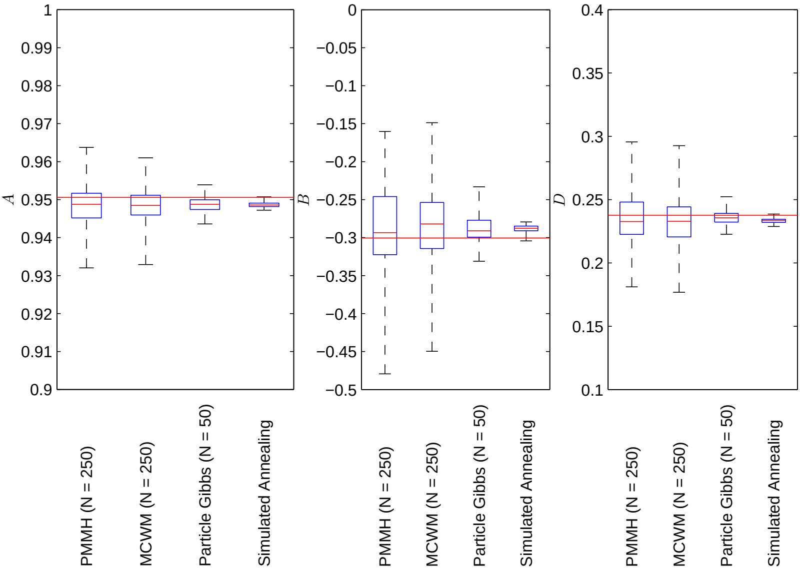
- Data 50 observations simulated using:

$$A = 0.9, B = 1, \text{ and } D = 1$$

- Algorithms
  - The PMMH/MCWM algorithms use $N = 250$ particles;
  - The PG algorithm (with ancestor sampling) uses $N = 50$ particles but attempts 100 static parameter updates per iteration.
  - Inverse temperature increases linearly from 0.1 to 10.
  - Final 1000 iterations $\gamma_t = 10$.
  - Compare with exact marginal simulated annealing algorithm.


PMMH (N = 250)
MCWM (N = 250)
Particle Gibbs (N = 50)
Simulated Annealing
True parameter
MLE
A
B
D
Iteration

Summary of 100 Runs



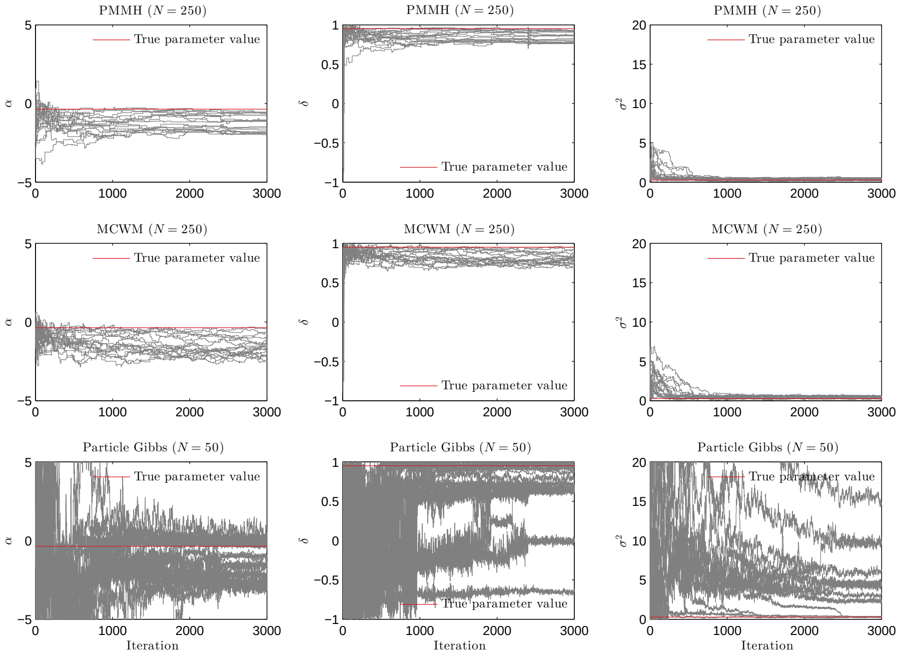
A Simple Stochastic Volatility Model

- Model:

$$X_i = \alpha + \delta X_{i-1} + \sigma_u u_i \qquad X_1 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$Y_i = \exp\left(\frac{X_i}{2}\right)\epsilon_i$$

  where u_i and ϵ_i are uncorrelated standard normal random variables, and $\theta = (\alpha, \delta, \sigma_u)$.
- 200 Observations; simulated with $\delta = 0.95$, $\alpha = -0.363$ and $\sigma = 0.26$.
- Diffuse instrumental prior distributions:
  - $\delta \sim U(-1, 1)$
  - $\alpha \sim \mathcal{N}(0, 1)$
  - $\sigma^{-2} \sim \mathcal{G}\text{a}(1, 0.1)$

  are quickly forgotten.
- Inverse temperature increases linearly from 0.1 to 10.
- Final 500 iterations $\gamma_t = 10$.
- A more complex multi-factor model is also under investigation.



PMMH ($N = 250$)
True parameter value
α
δ
σ^2
MCWM ($N = 250$)
Particle Gibbs ($N = 50$)
Iteration

In Conclusion

- Monte Carlo isn't just for calculating posterior expectations.
- SMC and Pseudomarginal methods are effective for ML and MAP estimation.
- Still work in progress…
- Scope for embedding Pseudomarginal target within SMC algorithm…
- and adaptation.

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] M. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

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

[5] A. Doucet, S. J. Godsill, and C. P. Robert. Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12:77–84, 2002.

[6] A. Finke. *On Extended State-Space Constructions for Monte Carlo Methods.* Ph.D. thesis, University of Warwick, 2015. In preparation.

[7] A. Finke and A. M. Johansen. More of the SAME? Pseudomarginal methods for point estimation in latent variable models. In preparation, 2015.

[8] C.-R. Hwang. Laplace's method revisited: Weak convergence of probability measures. *Annals of Probability*, 8(6):1177–1182, December 1980.

[9] A. M. Johansen, A. Doucet, and M. Davy. Maximum likelihood parmeter estimation for latent models using sequential Monte Carlo. In *Proceedings of ICASSP*, volume III, pages 640–643. IEEE, May 2006.

[10] A. M. Johansen, A. Doucet, and M. Davy. Particle methods for maximum likelihood parameter estimation in latent variable models. *Statistics and Computing*, 18(1):47–57, March 2008.