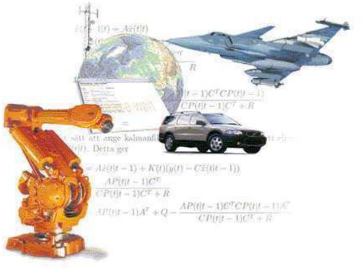
Second-order particle MCMC for Bayesian parameter inference

IFAC World Congress 2014, Cape Town, South Africa, August 28, 2014.



Johan Dahlin
johan.dahlin@liu.se

Division of Automatic Control,
Linköping University,
Sweden.

This is collaborative work with

Dr. Fredrik Lindsten (University of Cambridge, United Kingdom)
Prof. Thomas B. Schön (Uppsala University, Sweden)

Summary

Aim

Efficient Bayesian parameter inference in nonlinear SSMs.

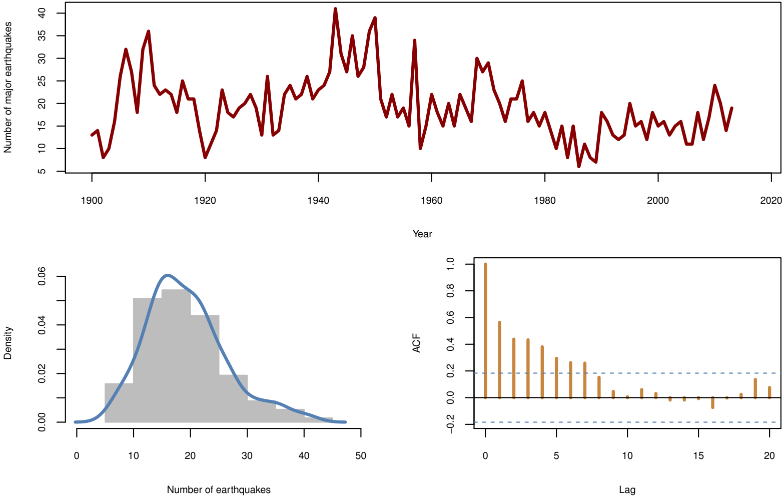
Methods

Markov chain Monte Carlo.
Sequential Monte Carlo methods.

Contributions

Efficient estimation of first/second order information.
Inclusion of first/second order in the proposal.
PMH1 and PMH2.

Example: Earthquakes between 1900 and 2013

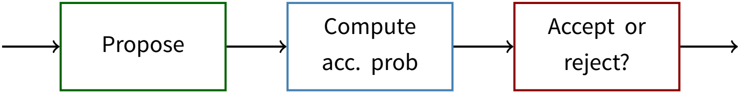


Example: A simple model of annual earthquake counts

$$x_{t+1}|x_t \sim \mathcal{N}\Big(x_{t+1}; \phi x_t, \boldsymbol{\sigma}^2\Big),$$
$$y_t|x_t \sim \mathcal{P}\Big(y_t; \boldsymbol{\beta} \exp(x_t)\Big).$$

Task: Estimate $\pi(\theta) = p(\theta|y_{1:T}) \propto p_\theta(y_{1:T})p(\theta)$ with $\theta = \{\phi, \boldsymbol{\sigma}\}$.

Metropolis-Hastings algorithm

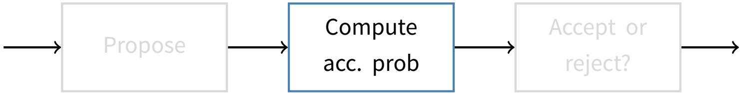


- Propose: $\theta' \sim q(\theta' | \theta_{k-1})$.
- Compute acceptance probability:

$$\alpha(\theta', \theta_{k-1}) = \min\left\{1, \frac{\pi(\theta')}{\pi(\theta_{k-1})} \frac{q(\theta_{k-1} | \theta')}{q(\theta' | \theta_{k-1})}\right\}$$

- Accept or reject? $\theta' \to \theta_k$ w.p. $\alpha(\theta', \theta_{k-1})$.

Particle Metropolis-Hastings algorithm (cont.)



- Propose: $\theta' \sim q(\theta' | \theta_{k-1}, u')$ and $u' \sim \mathrm{PF}(\theta')$.
- Compute $\widehat{p}_{\theta'}(y_{1:T} | u')$ and the acceptance probability:

$$\alpha(\theta', \theta_{k-1}) = 1 \wedge \frac{p(\theta')}{p(\theta_{k-1})} \frac{\widehat{p}_{\theta'}(y_{1:T} | u')}{\widehat{p}_{\theta_{k-1}}(y_{1:T} | u_{k-1})} \frac{q(\theta_{k-1} | \theta', u')}{q(\theta' | \theta_{k-1}, u_{k-1})}.$$

- Accept or reject? $\theta' \to \theta_k$ and $u' \to u_k$ w.p. $\alpha(\theta', \theta_{k-1})$.

Particle filtering

Given the particle system (the random variables u)

$$u = \left\{ x_{1:T}^{(i)}, w_{1:T}^{(i)} \right\}_{i=1}^{N},$$

we can obtain (consistent) estimates of:

- the likelihood $p_\theta(y_{1:T})$.
- the first and second order information of $\pi(\theta)$.

Zeroth order proposal (PMH0)

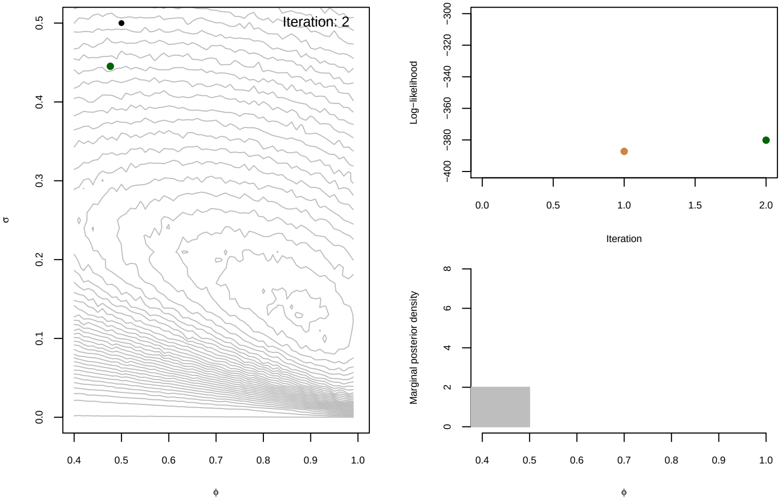
Gaussian random walk

$$\theta'' = \theta' + \epsilon z, \quad z \sim \mathcal{N}(z; 0, 1).$$

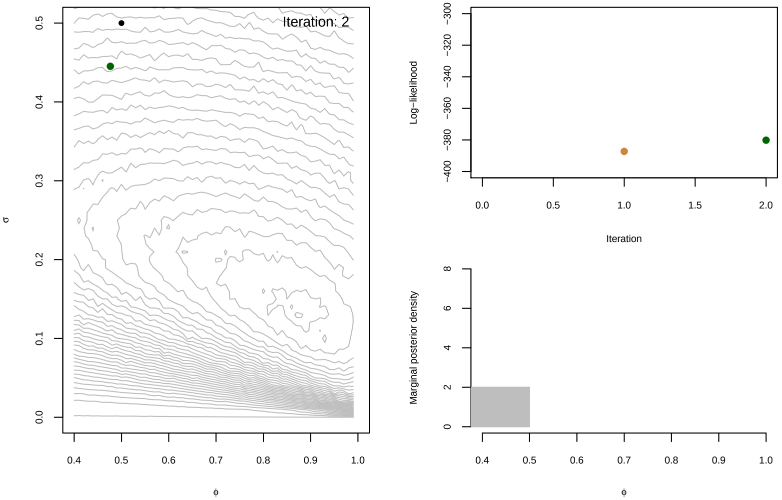
gives the zeroth order (marginal) proposal

$$q(\theta''|\theta', u') = \mathcal{N}\left(\theta''; \theta', \epsilon^2 I_d\right).$$

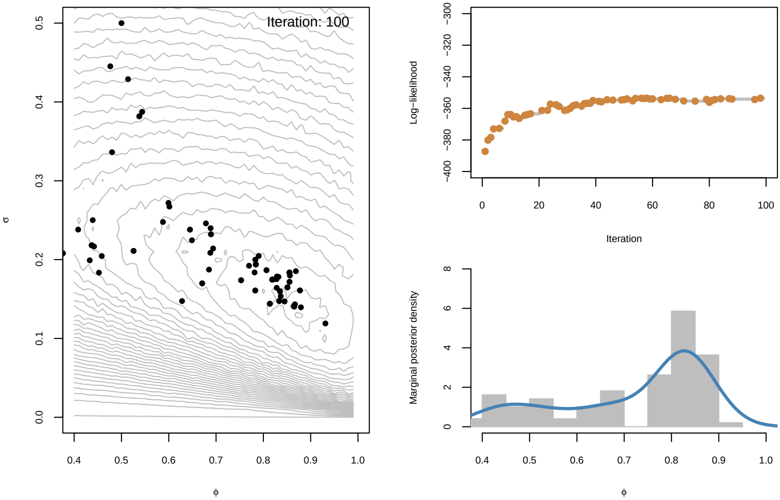
Example: Parameter inference in earthquake model



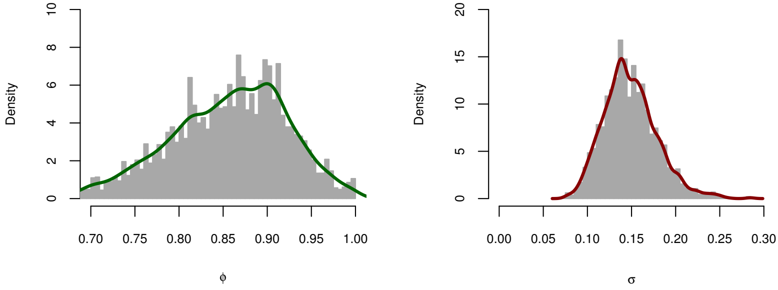
Example: Parameter inference in earthquake model



Example: Parameter inference in earthquake model

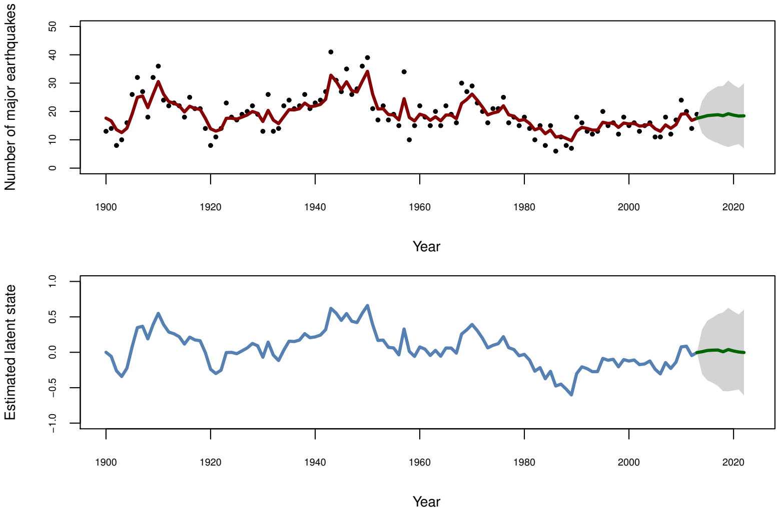


Example: Parameter inference in earthquake model



	ϕ	σ
Posterior mean	0.86	0.15
Posterior median	0.86	0.15
Posterior mode	0.90	0.14

Example: State inference in the earthquake model



Fixed-lag particle smoothing: overview

The first and second order information can be estimated using

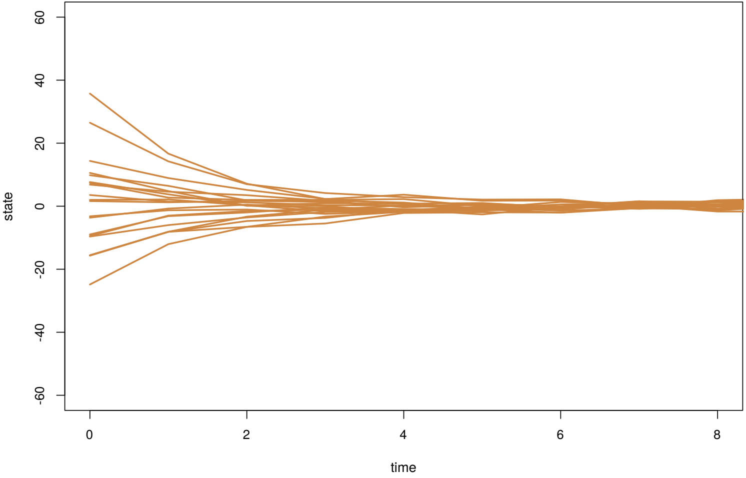
$$u = \left\{ x_{1:T}^{(i)}, w_{1:T}^{(i)} \right\}_{i=1}^{N},$$

and the fixed-lag particle smoother approximation,

$$\widehat{p}_\theta(\mathrm{d}x_{t:t-1} | y_{1:T}) \approx \widehat{p}_\theta(\mathrm{d}x_{t:t-1} | y_{1:\kappa_t}), \quad \kappa_t = \min\{T, t + \Delta\},$$

with no additional computational complexity.

Fixed-lag particle smoothing: motivation



First order proposal (PMH1)

Noisy gradient-based ascent update

$$\theta'' = \theta' + \frac{\epsilon^2}{2}\mathcal{S}(\theta') + \epsilon z, \quad z \sim \mathcal{N}(z; 0, 1),$$

with the first order information

$$\mathcal{S}(\theta') = \nabla_\theta \log \pi(\theta)\big|_{\theta=\theta'},$$

gives the first order proposal

$$q(\theta''|\theta', u') = \mathcal{N}\left(\theta''; \theta' + \frac{\epsilon^2}{2}\widehat{\mathcal{S}}(\theta'|u'), \epsilon^2 I_d\right).$$

Second order proposal (PMH2)

Noisy Newton update

$$\theta'' = \theta' + \frac{\epsilon^2}{2}\big[\mathcal{J}(\theta')\big]^{-1}\mathcal{S}(\theta') + \epsilon\big[\mathcal{J}(\theta')\big]^{-1/2} z, \quad z \sim \mathcal{N}(z; 0, 1),$$

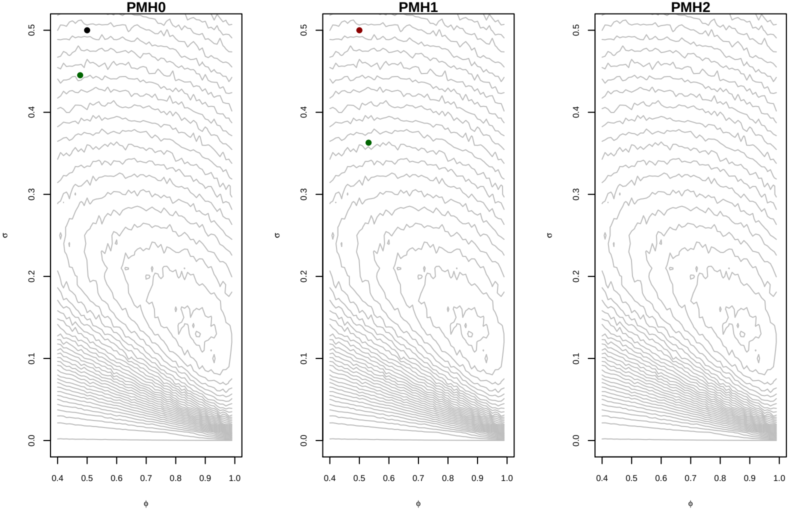
with the second order information

$$\mathcal{J}(\theta') = -\nabla^2_{\theta} \log \pi(\theta)\big|_{\theta=\theta'},$$

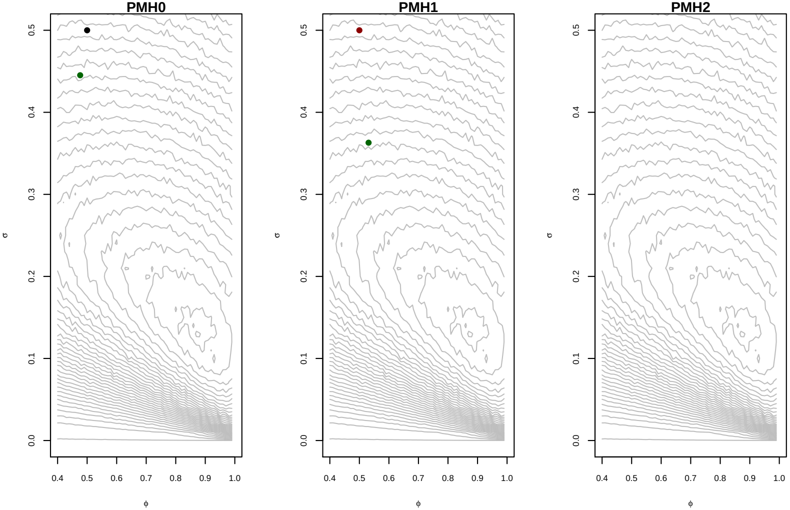
gives the second order proposal

$$q(\theta''|\theta', u') = \mathcal{N}\left(\theta''; \theta' + \frac{\epsilon^2}{2}\widehat{\mathcal{S}}(\theta'|u')\big[\widehat{\mathcal{J}}(\theta'|u')\big]^{-1}, \epsilon^2\big[\widehat{\mathcal{J}}(\theta'|u')\big]^{-1}\right).$$

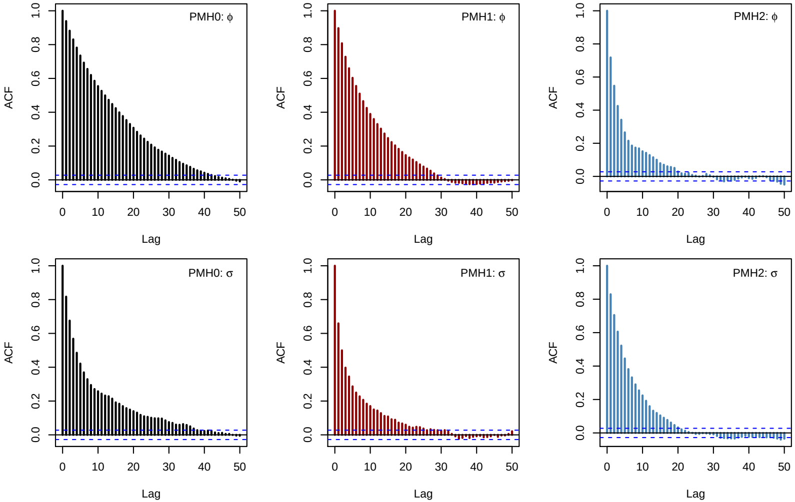
Example: Parameter inference in earthquake model



Example: Parameter inference in earthquake model



Integrated autocorrelation time



Integrated autocorrelation time (cont.)

IACT: the number of iterations between two uncorrelated samples.

	Acceptance rate	max IACT
PMH0	0.47	31.82
PMH1	0.38	21.38
PMH2	0.54	**14.15**

Conclusions

Results

Shorter burn-in phase.
Increased efficiency (about 2 times).
Simplified tuning.
Retained linear computational complexity in N.

Methods

Include u into the proposal.
Particle fixed-lag smoothing (almost for free).
Laplace approximation / Random walk on a Riemann manifold.

Future work

Non-reversible Markov chains.
Adaptive methods.
Approximate Bayesian computations.

Thank you for your attention!

Questions, comments and suggestions are most welcome.

Further developments

Extended version available at arXiv.

The paper and code to replicate the results within it are found at:

`http://work.johandahlin.com/.`

Particle Metropolis-Hastings algorithm

The *target distribution* is given by the parameter proposal

$$\pi(\theta) = \frac{p_\theta(y_{1:T})p(\theta)}{p(y_{1:T})}.$$

An *unbiased estimator of the likelihood* is given by

$$\mathbb{E}_m\left[\widehat{p}_\theta(y_{1:T}|u)\right] = \int \widehat{p}_\theta(y_{1:T}|u)m_\theta(u)\,\mathrm{d}u = p_\theta(y_{1:T}).$$

An *extended target* is given by

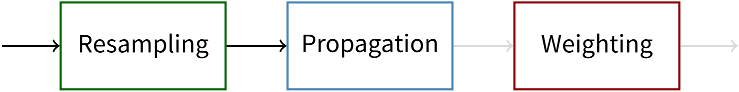
$$\pi(\theta, u) = \frac{\widehat{p}_\theta(y_{1:T}|u)m_\theta(u)p(\theta)}{p(y_{1:T})} = \frac{\widehat{p}_\theta(y_{1:T}|u)m_\theta(u)\pi(\theta)}{p_\theta(y_{1:T})}.$$

Particle Metropolis-Hastings algorithm (cont.)

$$\begin{aligned}
\int \pi(\theta, u)\,\mathrm{d}u &= \int \frac{\widehat{p}_\theta(y_{1:T}|u) m_\theta(u) \pi(\theta)}{p_\theta(y_{1:T})}\,\mathrm{d}u \\
&= \frac{\pi(\theta)}{p_\theta(y_{1:T})} \underbrace{\int \widehat{p}_\theta(y_{1:T}|u) m_\theta(u)\,\mathrm{d}u}_{=p_\theta(y_{1:T})}, \\
&= \pi(\theta).
\end{aligned}$$

That is, the marginal is the desired target distribution and the Markov chain is kept invariant.

(bootstrap) Particle filtering



- Resampling: $\mathbb{P}(a_t^{(i)} = j) = \widetilde{w}_{t-1}^{(j)}$ and set $\widetilde{x}_{t-1}^{(i)} = x_{t-1}^{a_t^{(i)}}$.
- Propagation: $x_t^{(i)} \sim R_\theta\left(x_t | \widetilde{x}_{t-1}^{(i)}\right) = f_\theta(x_t | \widetilde{x}_{t-1}^{(i)})$.
- Weighting: $w_t^{(i)} = W_\theta\left(x_t^{(i)}, \widetilde{x}_{t-1}^{(i)}\right) = g_\theta(y_t | x_t)$.

Likelihood estimation using the APF

The likelihood for an SSM can be decomposed by

$$\mathcal{L}(\theta) = p_\theta(y_{1:T}) = p_\theta(y_1) \prod_{t=2}^{T} p_\theta(y_t | y_{1:t-1}),$$

where the *one-step ahead predictor* can be computed by

$$\begin{aligned} p_\theta(y_t | y_{1:t-1}) &= \int f_\theta(x_t | x_{t-1}) g_\theta(y_t | x_t) p_\theta(x_{t-1} | y_{1:t-1}) \, \mathrm{d}x_t \\ &= \int W_\theta(x_t | x_{t-1}) R_\theta(x_t | x_{t-1}) p_\theta(x_{t-1} | y_{1:t-1}) \, \mathrm{d}x_t. \end{aligned}$$

$$p_\theta(y_t | y_{1:t-1}) \approx \frac{1}{N} \sum_{i=1}^{N} \int W_\theta(x_t | x_{t-1}) \delta_{x_t^{(i)}, \widetilde{x}_{t-1}^{(i)}} \, \mathrm{d}x_t = \frac{1}{N} \sum_{i=1}^{N} w_t^{(i)}.$$

Fixed-lag particle smoothing (cont.)

Assume that

$$p_\theta(x_t|y_{1:T}) \approx p_\theta(x_t|y_{1:\kappa_t}), \qquad \kappa_t = \min\{T, t+\Delta\},$$

for some $0 \leq \Delta \leq T$. It follows that

$$\widehat{p}_\theta(x_{t-1:t}|y_{1:T}) = \sum_{i=1}^{N} \widetilde{w}_{\kappa_t}^{(i)} \delta_{\widetilde{x}_{t-1:t,\kappa_t}^{(i)}}(\mathrm{d}x_{t-1:t})$$

which can be used to estimate the gradient and Hessian information about the log-target.

Score estimation using the FL smoother

The score can be estimated using *Fisher's identity* given by

$$\nabla_\theta \log p_\theta(y_{1:T})\big|_{\theta=\theta'} = \int \nabla_\theta \log p_\theta(x_{1:T}, y_{1:T}) p_{\theta'}(x_{1:T} | y_{1:T}) \mathrm{d}x_{1:T}$$
$$\approx \int \nabla_\theta \log p_\theta(x_{1:T}, y_{1:T}) \widehat{p}_{\theta'}(x_{1:T} | y_{1:T}) \mathrm{d}x_{1:T}$$

We also know that

$$\nabla_\theta \log p_\theta(x_{1:T}, y_{1:T}) = \sum_{t=1}^{T} \underbrace{\left[\nabla_\theta \log f_\theta(x_t | x_{t-1}) + \nabla_\theta \log g_\theta(y_t | x_t)\right]}_{\triangleq \eta(x_t, x_{t-1})},$$

which gives

$$\nabla_\theta \log p_\theta(y_{1:T})\big|_{\theta=\theta'} \approx \sum_{t=1}^{T} \sum_{i=1}^{N} \widetilde{w}_{\kappa_t}^{(i)} \eta(\widetilde{x}_{t-1,\kappa_t}^{(i)}, \widetilde{x}_{t,\kappa_t}^{(i)}).$$

Mixing

Let $\varphi(\theta)$ denote a *test function*, then

$$\sqrt{M}\left[\widehat{\varphi}_{\mathsf{MH}} - \mathbb{E}[\varphi(\theta)]\right] \xrightarrow{d} \mathcal{N}(0, \sigma^2_\varphi).$$

Here, σ^2_φ depends on the *integrated autocorrelation time* (IACT)

$$\mathsf{IACT}(\theta_{1:M}) = 1 + 2\sum_{k=1}^{\infty} \rho_k(\theta_{1:M}).$$