Correlated Particle Metropolis

Michael Pitt
King’s College London
Joint work George Deligiannidis & Arnaud Doucet (Oxford)

IMS, Singapore 2018
Organisation of the Talk

- Latent variable models
- The pseudo-marginal method
- The correlated pseudo-marginal method
- Illustrations
Assume \( \{X_t\}_{t \geq 1} \) is a latent Markov process, i.e. \( X_1 \sim \mu_\theta(\cdot) \) and
\[
X_{t+1} \mid (X_t = x) \sim f_\theta(\cdot \mid x), \quad Y_t \mid (X_t = x) \sim g_\theta(\cdot \mid x).
\]

Observations \( \{Y_t\}_{t \geq 1} \) are conditionally independent given \( \{X_t\}_{t \geq 0} \).

Likelihood of \( y_{1:T} = (y_1, ..., y_T) \) is
\[
p(y_{1:T} ; \theta) = \int_{X_{0:T+1}} p(x_{0:T}, y_{1:T} ; \theta) \, dx_{0:T}.
\]
Sequential Monte Carlo Estimator

- SMC provides an unbiased estimator of relative variance $O(T/N)$ where $N$ is the number of particles.
- Whatever being $N \geq 1$, the pseudo-marginal MH admits $\pi(\theta)$ as invariant distribution.
A directed acyclic graph (DAG) of the problem:

\[ y_t \sim g(y_t | x_t) \]

\[ x_t \sim f(x_t | x_{t-1}) \]
SMC: Likelihood Estimation

- Both $g(y_t | x_t)$, $f(x_t | x_{t-1})$ may be indexed by fixed parameters $\theta$.
- Filtering density
  \[ p(x_t | y_{1:t}; \theta). \]
- As an important byproduct we also obtain the one step predictive density
  \[ p(y_t | y_{1:t-1}; \theta). \]
- Yields the likelihood (KF)
  \[ p(y_{1:T} | \theta) = p(y_1 | \theta) \prod_{t=1}^{T-1} p(y_{t+1} | y_{1:t}; \theta). \]
Simulation based methods to perform filtering in nonlinear/non-Gaussian state space models.


We aim to have ‘particles’, $x_t^{1}, \ldots, x_t^{N}$ with associated discrete probability masses $\pi_t^{1}, \ldots, \pi_t^{N}$, drawn from the density $f(x_t|y_{1:t})$. 
We start at \( t = 0 \) with samples from \( x_0^k \sim p(x_0) \). For \( t=1,\ldots,T \):

We have samples \( x_t^k \sim p(x_t|y_1:t) \) for \( k = 1, \ldots, N \).

1. For \( k = 1 : N \), sample \( \tilde{x}_{t+1}^k \sim f(x_{t+1}|x_t^k) \).

2. For \( k = 1 : N \),

\[
\pi_t^k = \frac{g(y_{t+1}|\tilde{x}_{t+1}^k)}{\sum_{i=1}^N g(y_{t+1}|\tilde{x}_{t+1}^i)}.
\]

3. For \( j = 1 : N \), sample \( x_{t+1}^j \sim \sum_{k=1}^N \pi_t^k \delta(x_{t+1}^j - \tilde{x}_{t+1}^k) \).

Step 3 multinomial (or stratified) sampling (from the mixture). This will yield an approximate sample the desired posterior density, \( f(x_t|y_1:t) \) as \( t \) varies.
Parameter estimation using likelihood function, via prediction decomposition given by:

$$\log L(\theta) = \log p(y_1,\ldots,y_T|\theta) = \sum_{t=1}^{T} \log p(y_{t+1}|\theta; y_{1:t}).$$

We need to estimate the function:

$$\hat{p}(y_{1:T}|\theta) = \hat{p}(y_1|\theta) \prod_{t=1}^{T-1} \hat{p}(y_{t+1}|y_{1:t}; \theta),$$

$$\hat{p}(y_{t+1}|\theta; y_{1:t}) = \frac{1}{N} \sum_{i=1}^{N} p(y_{t+1}|\tilde{x}_{t+1}^i).$$

where $\tilde{x}_{t+1}^i \sim f(x_{t+1}|y_{1:t}; \theta)$, from step (2).

Remarkably (just like in IS) the lik estimator $\hat{p}(y_{1:T}|\theta)$ is unbiased for $p(y_{1:T}|\theta)$ regardless of $N$, Del Moral (04).
Bayesian Inference

- Likelihood function $p(y; \theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.
- Prior distribution of density $p(\theta)$.
- Bayesian inference relies on the posterior

$$\pi(\theta) = p(\theta | y) = \frac{p(y; \theta) p(\theta)}{\int_{\Theta} p(y; \theta') p(\theta') d\theta'}.$$ 

- For non-trivial models, inference relies typically on MCMC.
In numerous scenarios, \( p (y; \theta) \) cannot be evaluated pointwise; e.g.

\[
p (y; \theta) = \int p (x, y; \theta) \, dx
\]

where the integral cannot be evaluated.

A standard “solution” consists of using MCMC to sample from

\[
p (\theta, x | y) = \frac{p (x, y; \theta) p (\theta)}{p (y)}
\]
Standard MCMC Approaches

- Standard MCMC schemes target
  \[ p ( \theta, x_{1:T} | y_{1:T} ) \propto p ( \theta ) p_\theta ( x_{1:T}, y_{1:T} ) \]
  using Gibbs type strategy; i.e. sample alternately \( X_{1:T} \sim p_\theta ( \cdot | y_{1:T} ) \)
  and \( \theta \sim p ( \cdot | y_{1:T}, X_{1:T} ) \).
Standard MCMC Approaches

- Standard MCMC schemes target

\[ p(\theta, x_{1:T}|y_{1:T}) \propto p(\theta) p_{\theta}(x_{1:T}, y_{1:T}) \]

using Gibbs type strategy; i.e. sample alternately \( X_{1:T} \sim p_{\theta}(\cdot|y_{1:T}) \) and \( \theta \sim p(\cdot|y_{1:T}, X_{1:T}) \).

**Problem 1:** It can be difficult to sample \( p_{\theta}(x_{1:T}|y_{1:T}) \); e.g. non-Gaussian state-space models.
Standard MCMC Approaches

- Standard MCMC schemes target

\[ p(\theta, x_{1:T} \mid y_{1:T}) \propto p(\theta) p_{\theta}(x_{1:T}, y_{1:T}) \]

using Gibbs type strategy; i.e. sample alternately \( X_{1:T} \sim p_{\theta}(\cdot \mid y_{1:T}) \) and \( \theta \sim p(\cdot \mid y_{1:T}, X_{1:T}) \).

- **Problem 1**: It can be difficult to sample \( p_{\theta}(x_{1:T} \mid y_{1:T}) \); e.g. non-Gaussian state-space models.

- **Problem 2**: Even when it is implementable, Gibbs can converge very slowly e.g. diffusions.
Standard MCMC Approaches

- Standard MCMC schemes target
  \[ p(\theta, x_{1:T} | y_{1:T}) \propto p(\theta) p_\theta(x_{1:T}, y_{1:T}) \]
  using Gibbs type strategy; i.e. sample alternately \( X_{1:T} \sim p_\theta(\cdot | y_{1:T}) \)
  and \( \theta \sim p(\cdot | y_{1:T}, X_{1:T}) \).

- **Problem 1**: It can be difficult to sample \( p_\theta(x_{1:T} | y_{1:T}) \); e.g. non-Gaussian state-space models.

- **Problem 2**: Even when it is implementable, Gibbs can converge very slowly e.g. diffusions

- **Problem 3**: It may only be possible to generate from \( f(x_t | x_{t-1}) \) not to evaluate it, e.g. DSGE models.
Standard MCMC Approaches

- Standard MCMC schemes target

\[ p(\theta, x_{1:T} | y_{1:T}) \propto p(\theta) p_{\theta}(x_{1:T}, y_{1:T}) \]

using Gibbs type strategy; i.e. sample alternately \( X_{1:T} \sim p_{\theta}(\cdot | y_{1:T}) \) and \( \theta \sim p(\cdot | y_{1:T}, X_{1:T}) \).

**Problem 1:** It can be difficult to sample \( p_{\theta}(x_{1:T} | y_{1:T}) \); e.g. non-Gaussian state-space models.

**Problem 2:** Even when it is implementable, Gibbs can converge very slowly e.g. diffusions

**Problem 3:** It may only be possible to generate from \( f(x_t | x_{t-1}) \) not to evaluate it, e.g. DSGE models.

Pseudo-marginal methods mimic an algorithm targeting directly \( \pi(\theta) = p(\theta | y_{1:T}) \) instead of \( p(\theta, x_{1:T} | y_{1:T}) \).
Let \( \hat{p}(y; \theta, U) \) be an unbiased non-negative estimator of the likelihood where \( U \sim m_{\theta}(\cdot) \); i.e.

\[
p(y; \theta) = \int_U \hat{p}(y; \theta, u) m_{\theta}(u) \, du.
\]

Introduce a target distribution on \( \Theta \times U \) of density

\[
\pi(\theta, u) = \pi(\theta) \frac{\hat{p}(y; \theta, u)}{p(y; \theta)} m_{\theta}(u) = \frac{p(\theta) \hat{p}(y; \theta, u) m_{\theta}(u)}{p(y)}
\]

Then unbiasedness yields

\[
\int_U \pi(\theta, u) \, du = \pi(\theta)
\]

Any MCMC algorithm sampling from \( \pi(\theta, u) \) yields samples from \( \pi(\theta) \).
Pseudo-Marginal Metropolis-Hastings algorithm

- Can form an unbiased estimator, based on $N$ particles $\hat{p}(y; \theta, U)$.
- Set $\left( \vartheta^{(0)}, U^{(0)} \right)$ and iterate for $j = 1, 2, ...$

Sample $\vartheta \sim q \left( \cdot | \vartheta^{(j-1)} \right)$, $U \sim m_\vartheta (\cdot)$ to obtain $\hat{p}(y; \vartheta, U)$.

1. Compute

$$\alpha = 1 \wedge \frac{\hat{p}(y; \vartheta, U)}{\hat{p}(y; \vartheta^{(j-1)}, U^{(j-1)})} \frac{p(\vartheta)}{p(\vartheta^{(j-1)})} \frac{q(\vartheta^{(j-1)} | \vartheta)}{q(\vartheta | \vartheta^{(j-1)})}$$

2. With proba $\alpha$, set $\left( \vartheta^{(j)}, U^{(j)} \right) := (\vartheta, U)$ and stay where you are otherwise.
A Nonlinear State-Space Model

- Standard non-linear model

\[ X_t = \frac{1}{2} X_{t-1} + 25 \frac{X_{t-1}}{1+X_{t-1}^2} + 8 \cos(1.2t) + V_t, \quad V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_V^2), \]

\[ Y_t = \frac{1}{20} X_t^2 + W_t, \quad W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_W^2). \]

- \( T = 200 \) data points with \( \theta = (\sigma_V^2, \sigma_W^2) = (10, 10) \).

- Difficult to perform standard MCMC as \( p(\theta \mid y_{1:T}) \) is highly multimodal.

- We sample from \( p(\theta \mid y_{1:T}) \) using a random walk pseudo-marginal MH where \( p(y_{1:T} \mid \theta) \) is estimated using SMC with \( N \) particles.
Figure: Autocorrelation of $\{\sigma_V^{(i)}\}$ and $\{\sigma_W^{(i)}\}$ of the MH sampler for various $N$. 
How to Select the Number of Samples

- If $N$ is too small, then the algorithm mixes poorly and will require many MCMC iterations.
- If $N$ is too large, then each iteration is expensive due to estimating the likelihood.
- Equivalently we will examine $\sigma^2$, the variance of the estimator of the log-likelihood to trade off these two concerns.
- **Simple Guideline:** We find the optimal value for $\sigma$ around 1. 
Consider the error in the log-likelihood estimator

\[ Z = \log \hat{p}(y; \theta, U) - \log p(y; \theta) \sim g_\theta(\cdot) \]

In the \((\theta, Z)\) parameterization, the target is

\[ \pi(\theta, u) = \pi(\theta) \frac{\hat{p}(y; \theta, u)}{p(y; \theta)} m_\theta(u) \Rightarrow \pi(\theta, z) = \pi(\theta) \exp(z) g_\theta(z). \]
How precise should the log-likelihood estimator be?

- **Aim**: Minimize the “computational time”

\[
CT(Q, h) = \frac{\text{IACT}(Q, h)}{\sigma^2}
\]

as \( \sigma^2 \propto 1/N \) and computational efforts proportional to \( N \)
How precise should the log-likelihood estimator be?

- **Aim**: Minimize the “computational time”

\[
CT(Q, h) = \frac{IACT(Q, h)}{\sigma^2}
\]

as \(\sigma^2 \propto 1/N\) and computational efforts proportional to \(N\)

- where

\[
IACT(Q, h) = \text{Integrated Autocorrelation Time of } \{h(\theta_i)\}_{i \geq 1}
\]
How precise should the log-likelihood estimator be?

- The IACT is

\[
\text{IACT} \left( Q, h \right) = 1 + 2 \sum_{\tau=1}^{\infty} \text{corr}_{\pi,Q} \left\{ h \left( \theta_0 \right), h \left( \theta_\tau \right) \right\}
\]

- where \( Q \) is the pseudo-marginal kernel with acceptance criterion

\[
\min \left\{ 1, r \left( \theta, \vartheta \right) \exp \left( w - z \right) \right\}.
\]

- where \( r \left( \theta, \vartheta \right) \) is exact Metropolis ratio,

- and

\[
z = \log \left\{ \hat{p}_\theta(y_{1:T}) / p_\theta(y_{1:T}) \right\}, \quad w = \log \left\{ \hat{p}_\theta(y_{1:T}) / p_\theta(y_{1:T}) \right\}.
\]
Simplifying Assumption: The noise $Z$ is independent of $\theta$ and Gaussian; i.e. $g(z|\sigma) = \mathcal{N}(z; -\sigma^2/2; \sigma^2)$:

$$\bar{\pi}(\theta, z) = \pi(\theta) \exp(z) g(z) = \pi(\theta) \mathcal{N}(z; \sigma^2/2; \sigma^2).$$

Justified empirically and theoretically (a CLT and concentration).
Pseudo-proof:

IID data, fixed $\theta$ (IS)

$$Z = \log \hat{p}(y|\theta) - \log p(y|\theta) = \sum_{t=1}^{T} \log \left\{ 1 + \frac{\gamma_t}{\sqrt{N}} \varepsilon_t \right\},$$

where $\varepsilon_t$ white noise, variance 1 and $N = \beta T$,

$$\sim \sum_{t=1}^{T} \frac{\gamma_t}{\sqrt{\beta T}} \varepsilon_t - \frac{1}{2} \frac{\gamma^2_t}{\beta T} \varepsilon^2_t \rightarrow \mathcal{N} \left( -\frac{1}{2} \frac{\bar{\gamma}^2}{\beta}, \frac{\bar{\gamma}^2}{\beta} \right),$$

where $\bar{\gamma}^2 = T^{-1} \sum_{t=1}^{T} \gamma^2_t$.

- More detail (and more generality, PFs) in Bérard, Del Moral, Doucet (EJP, 2014).
- Need concentration (in $\theta$ as $T \rightarrow \infty$), unproven.
Simulation:

SSF (from JoE(2012, Kohn, Giordani, Silva))

Figure: AR(1) plus noise model with fixed parameters. replications is 10,000. SIR likelihood estimator (divided by the true likelihood) on LEFT and for the error in the log of the SIR likelihood estimator on RIGHT. Both $N$ and $T$ vary
Special cases

- When \( q(\theta|\theta) = \pi(\theta) \), \( \sigma_{\text{opt}} = 0.92 \) (Pitt et al., JoE 2012). The acf simply reduces to:

\[
\phi_n(\theta, Q) = \int \Pr(R | z; \sigma)^n \pi_Z(z|\sigma) \, dz
\]

\[
\text{IACT}(\sigma) = \int \frac{1 + \Pr(R | z; \sigma)}{1 - \Pr(R | z; \sigma)} \pi_Z(z|\sigma) \, dz.
\]

\( \Pr(R | z; \sigma) \) analytically available.

- Sherlock, Thiery, Roberts and Rosenthal, (Annals 2015) consider the (joint) optimisation for a limiting target in the important case of RWM. Again analytically available and exact. Similar results.
Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^Q (\sigma) = IF_h^Q (\sigma) / \sigma^2$ impossible so minimize an upper bound over it.
Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT^Q_h(\sigma) = IF^Q_h(\sigma) / \sigma^2$ impossible so minimize an upper bound over it.
- We introduce an auxiliary $\pi(\theta, z)$—reversible kernel

$$Q^* \{ (\theta, z), (d\theta, dw) \} = q(\theta|\theta)g(w)\alpha_{EX}(\theta, \theta)\alpha_Z(z, w)\, d\theta \, dw$$

$$+ \{ 1 - q_{EX}(\theta)q_Z(z) \} \delta(\theta, z) (d\theta, dw),$$

where we have a product acceptance criterion:

$$\alpha_{EX}(\theta, \theta) = \min \{ 1, r(\theta, \theta) \}, \quad \alpha_Z(z, w) = \min \{ 1, \exp(w - z) \}. $$
Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^Q(\sigma) = IF_h^Q(\sigma)/\sigma^2$ is impossible so minimize an upper bound over it.

- We introduce an auxiliary $\pi(\theta, z)$—reversible kernel

\[Q^* \{(\theta, z), (d\theta, dw)\} = q(\vartheta|\theta)g(w)\alpha_{EX}(\theta, \vartheta)\alpha_Z(z, w) d\theta dw + \{1 - q_{EX}(\theta)q_Z(z)\} \delta(\theta, z) (d\theta, dw),\]

where we have a product acceptance criterion:

\[\alpha_{EX}(\theta, \vartheta) = \min \{1, r(\theta, \vartheta)\}, \quad \alpha_Z(z, w) = \min \{1, \exp(w - z)\}.\]

- Peskun’s theorem (1973) guarantees that $IF_h^Q(\sigma) \leq IF_h^{Q^*}(\sigma)$ so that $CT_h^Q(\sigma) \leq CT_h^{Q^*}(\sigma)$. 

(IMS, Singapore 2018)
**pseudo-proof: Jump chain**

**IACT:** $E_{\pi}[x] = 0$, on J.C.s Douc and Robert (Annals, 2011)

$$\tilde{\pi}(x) = \frac{\pi(x)p(x)}{P_A}$$

$$(IACT + 1)E_{\pi}[x^2_o] = 2\sum_{t=0}^{\infty} E_{\pi}[x_0 x_t]$$

$$= 2\sum_{\tau=0}^{\infty} E_{\pi,\tilde{\pi}}[x_0 \tau \tilde{x}_\tau]$$

$$= 2\sum_{\tau=0}^{\infty} E_{\pi,\tilde{\pi}} \left[ x_0 \frac{\tilde{x}_\tau}{p(\tilde{x}_\tau)} \right] \text{ because sojourns Geometric}$$

$$= 2P_A \sum_{\tau=0}^{\infty} E_{\tilde{\pi}} \left[ \frac{\tilde{x}_o}{p(\tilde{x}_0)} \frac{\tilde{x}_\tau}{p(\tilde{x}_\tau)} \right] \text{ change of measure for}$$

$$= 2P_A (IACT + 1) E_{\tilde{\pi}} \left[ \frac{\tilde{x}_o^2}{p(\tilde{x}_0)^2} \right].$$
Simpler Bounds on the Relative Inefficiency

- We obtain an explicit expression for $IF_h^Q (\sigma)$.
- If $IF_{h/\rho_{EX}}^Q \geq 1$, e.g. $\tilde{Q}^{EX}$ is a positive kernel, then

$$\frac{IF_h^Q (\sigma)}{IF_{h}^{EX}} \leq \frac{IF_{h}^{Q^*} (\sigma)}{IF_{h}^{EX}} \leq \frac{1}{2} \left(1 + \frac{1}{IF_{h}^{EX}}\right) \pi_{Z}^{\sigma} \left(1/\rho_{Z}^{\sigma}\right) - \frac{1}{IF_{h}^{EX}}$$

and the bound is tight as $IF_{h}^{EX} \to 1$ or $\sigma \to 0$.
- As $IF_{J,h/\rho_{EX}}^{EX} \to \infty$,

$$\frac{IF_{h}^{Q^*} (\sigma)}{IF_{h}^{EX}} \to \frac{1}{\pi_{Z}^{\sigma} \left(\rho_{Z}^{\sigma}\right)} = \frac{1}{P_A(\sigma)}.$$

- Results used to minimize w.r.t $\sigma$ upper bounds on $CT_h^Q (\sigma) = IF_h^Q (\sigma) / \sigma^2$. 

(IMS, Singapore 2018)
Left: upper bound on $CT_h^{Q^*}(\sigma)$ as a function of $\sigma$ for $IF_h^{EX} = 1$ (square), 4 (crosses), 20 (circles), 80 (triangles). Right: upper bounds on $CT_h^{Q^*}(\sigma)$ as a function of $\sigma$ for $IF_{J,h//\varrho_{EX}}^{EX} = 1$ for $IF_{J,h//\varrho_{EX}}^{EX} = 1, 4, 20, 80$ and lower bound (solid line).

\[
\begin{align*}
\text{dv}_1 (t) &= -k_1 \{ \nu_1 (t) - \mu_1 \} \, dt + \sigma_1 dW_1 (t), \\
\text{dv}_2 (t) &= -k_2 \nu_2 (t) + \{ 1 + \beta_{12} \nu_2 (t) \} \, dW_2 (t), \\
\text{d} \log P (t) &= \mu_y \, dt + s \cdot \exp \left[ \{ \nu_1 (t) + \beta_2 \nu_2 (t) \} / 2 \right] \, dB (t),
\end{align*}
\]

with \( \phi_1 = \text{corr}\{ B (t), W_1 (t) \} \) and \( \phi_2 = \text{corr}\{ B (t), W_2 (t) \} \).

- Euler discretization of the volatilities \( \nu_1 (t) \) and \( \nu_2 (t) \) provides closed form expression for \( Y (s) = \log P (\tau_{s+1}) - \log P (\tau_s) \).
- Straightforward to estimate the likelihood and simulate forward.
Daily returns \( y = (y_1, \ldots, y_T) \) of the S&P 500 index.

Bayesian Inference on \( \theta = (k_1, \mu_1, \sigma_1, k_2, \beta_{12}, \beta_2, \mu_y, \phi_1, \phi_2) \).

Performance of the pseudo-marginal for RW proposal w.r.t \( \sigma \), standard deviation of \( \log \hat{p}_\theta(y) \) at posterior mean \( \bar{\theta} \).
Empirical vs Assumed Distributions of $Z$ for SV model:

Figure: Empirical distributions (dashed) vs assumed Gaussians (solid) of $Z$. $T = 300$ and $T = 2700$ at $\bar{\theta}$ (left) and marginalized over $\pi(\theta)$. 
Figure: Average over the 9 parameter components of the log-integrated autocorrelation time of pseudo-marginal chain as a function of $\sigma$ for $T = 300$. 
Computational time for the SV model

Figure: Computational time as a function of $\sigma$
**Guideline:** Optimal $\sigma$ depends on efficiency of the ideal MH algorithm but $\sigma \approx 1.2 - 1.3$ is a sweet spot.
Guideline and Discussion

- **Guideline**: Optimal $\sigma$ depends on efficiency of the ideal MH algorithm but $\sigma \approx 1.2 – 1.3$ is a sweet spot.
- Pseudo-marginal scales in $O(T^2)$ at each iteration as we require $N \propto T$. 
Guideline and Discussion

- **Guideline:** Optimal $\sigma$ depends on efficiency of the ideal MH algorithm but $\sigma \approx 1.2 - 1.3$ is a sweet spot.
- Pseudo-marginal scales in $\mathcal{O}(T^2)$ at each iteration as we require $N \propto T$.
- For i.i.d. data, simulated ML estimator is efficient as long as $N$ increases at a rate faster than $\sqrt{T}$; e.g. Lee, *Econometric Theory*, 1999.
Guideline and Discussion

- **Guideline:** Optimal $\sigma$ depends on efficiency of the ideal MH algorithm but $\sigma \approx 1.2 - 1.3$ is a sweet spot.
- Pseudo-marginal scales in $O(T^2)$ at each iteration as we require $N \propto T$.
- For i.i.d. data, simulated ML estimator is efficient as long as $N$ increases at a rate faster than $\sqrt{T}$; e.g. Lee, *Econometric Theory*, 1999.
- **Problem:** the ratio $p_\theta(y_{1:T}) / p_\theta(y_{1:T})$ is estimated by estimating independently the numerator and denominator in pseudo-marginal.
Correlated Pseudo-Marginal Algorithm

- Previously, we consider the likelihood estimator \( \hat{p}_\theta (y_{1:T}; U) \) where \( U \sim m_\theta (\cdot) \).
- Reparameterize the likelihood estimator so that \( U \sim \mathcal{N} (0, I) \).
- Correlate estimators of \( p_\theta (y_{1:T}) \) and \( p_\vartheta (y_{1:T}) \) by setting
  \[
  \hat{p}_\theta (y_{1:T}) = \hat{p}_\vartheta (y_{1:T}; V)
  \]
  where
  \[
  V = \rho U + \sqrt{1 - \rho^2} \varepsilon, \varepsilon \sim \mathcal{N} (0, I).
  \]
- In practice, \( \rho \) will be selected close to 1.
- The invariance of the PMH unaffected.
- Can be seen as a solution of a standard O-U over \([0, \delta]\), \( \rho = \exp(-\delta) \).
Correlated Pseudo-Marginal Algorithm

- Correlated pseudo-marginal simulates a Markov chain \( \{\theta_i, U_i\}_{i \geq 1} \) of limiting distribution \( \pi(\theta, u) \).
Correlated pseudo-marginal simulates a Markov chain \( \{\theta_i, U_i\}_{i \geq 1} \) of limiting distribution \( \overline{\pi}(\theta, u) \).
Correlated pseudo-marginal simulates a Markov chain \( \{\vartheta_i, U_i\}_{i \geq 1} \) of limiting distribution \( \bar{\pi}(\theta, u) \).

At iteration \( i \)

- Sample \( \vartheta \sim q(\cdot | \vartheta_{i-1}) \) and \( U = \rho U_{i-1} + \sqrt{1-\rho^2} \varepsilon, \varepsilon \sim \mathcal{N}(0, I) \).
Correlated Pseudo-Marginal Algorithm

- Correlated pseudo-marginal simulates a Markov chain \( \{\vartheta_i, U_i\}_{i \geq 1} \) of limiting distribution \( \pi(\theta, u) \).

**At iteration \( i \)**

- Sample \( \vartheta \sim q(\cdot | \vartheta_{i-1}) \) and \( U = \rho U_{i-1} + \sqrt{1 - \rho^2} \varepsilon, \varepsilon \sim \mathcal{N}(0, I) \).
- Compute the estimate \( \hat{p}_\theta(y_{1:T}; U) \) of \( p_\theta(y_{1:T}) \).
Correlated Pseudo-Marginal Algorithm

- Correlated pseudo-marginal simulates a Markov chain \( \{\vartheta_i, U_i\}_{i \geq 1} \) of limiting distribution \( \overline{\pi}(\theta, u) \).

At iteration \( i \)

- Sample \( \vartheta \sim q(\cdot | \vartheta_{i-1}) \) and \( U = \rho U_{i-1} + \sqrt{1 - \rho^2} \varepsilon, \varepsilon \sim \mathcal{N}(0, I) \).
- Compute the estimate \( \hat{p}_\theta(y_{1:T}; U) \) of \( p_\theta(y_{1:T}) \).
- With probability

\[
\min\{1, \frac{\hat{p}_\theta(y_{1:T}; U)}{\hat{p}_{\vartheta_{i-1}}(y_{1:T}; U_{i-1})} \frac{p(\vartheta)}{p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1}| \vartheta)}{q(\vartheta | \vartheta_{i-1})} \}
\]

set \( \vartheta_i = \vartheta, U_i = U \), otherwise set \( \vartheta_i = \vartheta_{i-1}, U_i = U_{i-1} \).
If the likelihood is computed using standard SMC, \((\theta, u) \mapsto \hat{p}_\theta(y; u)\) is not continuous and displays large variations for moderate \(N\).
If the likelihood is computed using standard SMC, \((\theta, u) \mapsto \hat{p}_\theta(y; u)\) is not continuous and displays large variations for moderate \(N\).

Discontinuities arise from the resampling step: you can end up picking resampling particles very far from each other even with small variations in \((\theta, u)\).
If the likelihood is computed using standard SMC, \((\theta, u) \mapsto \hat{p}_\theta(y; u)\) is not continuous and displays large variations for moderate \(N\).

Discontinuities arise from the resampling step: you can end up picking resampling particles very far from each other even with small variations in \((\theta, u)\).

For \(n = 1\), (Pitt et al, 2012) propose sorting particles \(X_{t}^{\sigma_t(1)} \leq \ldots \leq X_{t}^{\sigma_t(N)}\).
Likelihood estimation for state-space models

- If the likelihood is computed using standard SMC, \((\theta, u) \mapsto \hat{p}_\theta(y; u)\) is not continuous and displays large variations for moderate \(N\).
- Discontinuities arise from the resampling step: you can end up picking resampling particles very far from each other even with small variations in \((\theta, u)\).
- For \(n = 1\), (Pitt et al, 2012) propose sorting particles \(X_{t\sigma_t(1)} \leq \ldots \leq X_{t\sigma_t(N)}\).
- For \(n \geq 2\), (Gerber & Chopin, 2015) use Hilbert curve; e.g. map particles to \([0, 1]^n\) using \(\psi : \mathbb{R}^n \rightarrow [0, 1]^n\) (e.g. logistic) and \(h : [0, 1]^n \rightarrow [0, 1]\) (pseudo-inverse Hilbert curve) and sort projected particles on \([0, 1]\).
If the likelihood is computed using standard SMC, $(\theta, u) \mapsto \hat{p}_\theta(y; u)$ is not continuous and displays large variations for moderate $N$.

Discontinuities arise from the resampling step: you can end up picking resampling particles very far from each other even with small variations in $(\theta, u)$.

For $n = 1$, (Pitt et al, 2012) propose sorting particles $X^{\sigma_t(1)}_t \leq \ldots \leq X^{\sigma_t(N)}_t$.

For $n \geq 2$, (Gerber & Chopin, 2015) use Hilbert curve; e.g. map particles to $[0, 1]^n$ using $\psi : \mathbb{R}^n \rightarrow [0, 1]^n$ (e.g. logistic) and $h : [0, 1]^n \rightarrow [0, 1]$ (pseudo-inverse Hilbert curve) and sort projected particles on $[0, 1]$.

Alternative coupling ideas have been used to mitigate these fluctuations (Jacob et al., arXiv 2016).
Assumption 1 - Asymptotic Normality: \( \exists \hat{\theta}^T \xrightarrow{P} \bar{\theta} \) and \( \Sigma \) a p.d. matrix s.t.

\[
\int \left| p(\theta \mid Y_{1:T}) - \phi(\theta; \hat{\theta}^T, \Sigma / T) \right| \, d\theta \xrightarrow{P} 0.
\]
Assumption 1 - Asymptotic Normality: \( \exists \hat{\theta}^T \xrightarrow{P} \bar{\theta} \) and \( \Sigma \) a p.d. matrix s.t.

\[
\int \left| p(\theta|Y_{1:T}) - \phi(\theta;\hat{\theta}^T,\Sigma/T) \right| d\theta \xrightarrow{P} 0.
\]

Assumption 2 - Proposal: \( \vartheta = \theta + \varepsilon/\sqrt{T} \) where \( \varepsilon \sim \nu(\cdot) \) with \( \nu(\varepsilon) = \nu(-\varepsilon) \).
Large sample analysis of the correlated PM - i.i.d. case

**Proposition.** Let $N \to \infty$ as $T \to \infty$ with $N = o(T)$. When $U \sim \pi_T(\cdot|\theta)$ and $U' = \rho U + \sqrt{1 - \rho^2} \epsilon$ with $\rho = \exp\left(-\psi \frac{N}{T}\right)$ then as $T \to \infty$

\[
\log \left\{ \frac{\hat{p}_{\theta+\xi} / \sqrt{T} (Y_1:T; U')}{\hat{p}_{\theta} (Y_1:T; U)} \right\} \left\{ \frac{p_{\theta+\xi} / \sqrt{T} (Y_1:T)}{p_{\theta} (Y_1:T)} \right\} \bigg| \mathcal{Y}^T, \mathcal{U}^T \Rightarrow \mathcal{N}\left(-\frac{\kappa^2(\theta)}{2}, \kappa^2(\theta)\right)
\]

where

$$
\kappa^2(\theta) = 2\psi \mathbb{E}\left[\|\partial_u \omega (U, Y; \theta)\|^2\right]
$$

- This CLT is conditional on the observation sequence and the current auxiliary variables.
Large sample analysis of the correlated PM - i.i.d. case

**Proposition.** Let $N \to \infty$ as $T \to \infty$ with $N = o(T)$. When $U \sim \pi_T(\cdot | \theta)$ and $U' = \rho U + \sqrt{1 - \rho^2} \epsilon$ with $\rho = \exp\left(-\frac{N}{T}\right)$ then as $T \to \infty$

$$
\log \left\{ \frac{\hat{p}_{\theta + \xi / \sqrt{T}} (Y_{1:T}; U')} {p_{\theta} (Y_{1:T}; U)} \right\} \left/ \frac{p_{\theta + \xi / \sqrt{T}} (Y_{1:T})} {p_{\theta} (Y_{1:T})} \right\} \Rightarrow \mathcal{N} \left( - \frac{\kappa^2(\theta)}{2}, \kappa^2(\theta) \right)
$$

where

$$
\kappa^2(\theta) = 2\psi \mathbb{E} \left[ \| \partial_u \omega (U, Y; \theta) \|^2 \right]
$$

- This CLT is conditional on the observation sequence and the current auxiliary variables.
- Asymptotically the distribution of the log-ratio decouples from the current location of the Markov chain.
Large sample analysis of the correlated PM - i.i.d. case

**Proposition.** Let $N \to \infty$ as $T \to \infty$ with $N = o(T)$. When $U \sim \pi_T(\cdot | \theta)$ and $U' = \rho U + \sqrt{1 - \rho^2} \epsilon$ with $\rho = \exp(-\psi \frac{N}{T})$ then as $T \to \infty$

$$
\log \left\{ \frac{\hat{p}_{\theta+\xi} / \sqrt{T} (Y_{1:T}; U')} {\hat{p}_{\theta} (Y_{1:T}; U)} \frac{p_{\theta+\xi} / \sqrt{T} (Y_{1:T})} {p_{\theta} (Y_{1:T})} \right\} \mathcal{Y}^T, \mathcal{U}^T \Rightarrow \mathcal{N} \left( - \frac{\kappa^2(\theta)}{2}, \kappa^2(\theta) \right)
$$

where

$$
\kappa^2(\theta) = 2\psi \mathbb{E} \left[ \| \partial_u \omega(U, Y; \theta) \|^2 \right]
$$

- This CLT is conditional on the observation sequence and the current auxiliary variables.
- Asymptotically the distribution of the log-ratio decouples from the current location of the Markov chain.
- The asymptotic variance is $O(1)$ even for $N \sim \log(T)$. 
Let $\Theta_T := \{\theta^T_i\}_{i \geq 0}$ be the stationary non-Markovian sequence of the correlated PM targetting $p(\theta | Y_{1:T})$. 
Let $\Theta_T := \{\theta_i^T\}_{i \geq 0}$ be the stationary \textit{non-Markovian} sequence of the correlated PM targetting $p(\theta|Y_{1:T})$.

**Proposition** (Deligiannidis et al., 2016): The sequences $\{\Theta_T\}_{T \geq 1}$ converge weakly as $T \to \infty$ to a stationary Markov chain of invariant density $\phi(\tilde{\theta}; 0, \Sigma)$ and kernel given for $\tilde{\theta} \neq \tilde{\vartheta}$ by

$$\tilde{Q}(\tilde{\theta}, d\tilde{\vartheta}') = v(\tilde{\vartheta}' - \tilde{\theta}) \mathbb{E}_{R \sim \mathcal{N}(-\kappa^2/2, \kappa^2)} \left[ \min \left\{ 1, \frac{\phi(\tilde{\vartheta}'; 0, \Sigma)}{\phi(\tilde{\theta}; 0, \Sigma)} \exp R \right\} \right] d\tilde{\vartheta}'$$

where $\kappa := \kappa(\tilde{\theta})$. 

It is tempting to use this result to provide guidelines on the optimization of CPM... but one has to be careful.
Let $\Theta_T := \{\theta_i^T\}_{i \geq 0}$ be the stationary non-Markovian sequence of the correlated PM targeting $p(\theta \mid Y_{1:T})$.

**Proposition** (Deligiannidis et al., 2016): The sequences $\{\Theta_T\}_{T \geq 1}$ converge weakly as $T \to \infty$ to a stationary Markov chain of invariant density $\phi(\tilde{\theta}; 0, \Sigma)$ and kernel given for $\tilde{\theta} \neq \tilde{\theta}'$ by

$$
\tilde{Q}(\tilde{\theta}, d\tilde{\theta}') = \nu(\tilde{\theta}' - \tilde{\theta}) \mathbb{E}_{R \sim \mathcal{N}(-\kappa^2/2, \kappa^2)} \left[ \min \left\{ 1, \frac{\phi(\tilde{\theta}'; 0, \Sigma)}{\phi(\tilde{\theta}; 0, \Sigma)} \exp R \right\} \right] d\tilde{\theta}'
$$

where $\kappa := \kappa(\tilde{\theta})$.

It is tempting to use this result to provide guidelines on the optimization of CPM... but one has to be careful.
For a stationary CPM chain \((\vartheta_i, U_i)\), decompose

\[
h(\vartheta_i) = \mathbb{E}[h(\vartheta) | U_i] + h(\vartheta_i) - \mathbb{E}[h(\vartheta) | U_i].
\]

\(U_i\) is proposed according to AR scheme with persistency

\[
\psi_N / T\:
\]

when \(N\) grows to slowly with \(T\), IACT of \(h(\vartheta_i)\) is dominated by IACT of \(\mathbb{E}[h(\vartheta) | U_i]\) where for scalar \(h(\vartheta) = \vartheta\) \(\mathbb{E}[\vartheta | U_i] = b_{\theta} T + \Sigma T_r \theta \log b_p Y_1: T_j b_{\theta} T + \mathbb{E}[h(\vartheta) | U_i]\).

Proposition. Let \(N \propto T^\alpha\) for \(0 < \alpha < 1\) then

\[
\text{IACT}(Q, \psi) \& T_1^2 \alpha.
\]

This result suggests we need at least

\[
p T / N = O(1).
\]
Decomposition in fast and slow components

- For a stationary CPM chain \((\vartheta_i, U_i)\), decompose

\[
h(\vartheta_i) = \mathbb{E}[h(\vartheta)|U_i] + h(\vartheta_i) - \mathbb{E}[h(\vartheta)|U_i].
\]

- \(U_i\) is proposed according to AR scheme with persistency \(\approx 1 - \psi N / T\) : when \(N\) grows to slowly with \(T\), IACT of \(h(\vartheta_i)\) is dominated by IACT of \(\mathbb{E}[h(\vartheta)|U_i]\) where for scalar \(h(\vartheta) = \vartheta\)

\[
\mathbb{E}[\vartheta|U_i] = \hat{\vartheta}_T + \frac{\sum}{T} \nabla_{\vartheta} \log \frac{\hat{p}(Y_{1:T} | \hat{\vartheta}_T, U_i)}{p(Y_{1:T} | \hat{\vartheta}_T)} + O_P(T^{-2}).
\]
Decomposition in fast and slow components

- For a stationary CPM chain \((\vartheta_i, U_i)\), decompose
  \[
  h(\vartheta_i) = \underbrace{\mathbb{E}[h(\vartheta)|U_i]}_{\text{slow}} + \underbrace{h(\vartheta_i) - \mathbb{E}[h(\vartheta)|U_i]}_{\text{fast}}.
  \]

- \(U_i\) is proposed according to AR scheme with persistency
  \(\approx 1 - \psi N/T\): when \(N\) grows to slowly with \(T\), IACT of \(h(\vartheta_i)\) is
  dominated by IACT of \(\mathbb{E}[h(\vartheta)|U_i]\) where for scalar \(h(\vartheta) = \vartheta\)
  \[
  \mathbb{E}[\vartheta|U_i] = \hat{\vartheta}_T + \frac{\sum}{T} \nabla_\vartheta \log \frac{\hat{p}\left(Y_{1:T} | \hat{\vartheta}_T, U_i\right)}{p\left(Y_{1:T} | \hat{\vartheta}_T\right)} + O_P(T^{-2}).
  \]

- Proposition. Let \(N \propto T^\alpha\) for \(0 < \alpha < 1\) then
  \(\text{IACT}(Q, \Psi) \gtrsim T^{1-2\alpha}\).
Decomposition in fast and slow components

- For a stationary CPM chain \((\vartheta_i, U_i)\), decompose

\[
h(\vartheta_i) = \underbrace{\mathbb{E}[h(\vartheta)|U_i]}_{\text{slow}} + \underbrace{h(\vartheta_i) - \mathbb{E}[h(\vartheta)|U_i]}_{\text{fast}}.\]

- \(U_i\) is proposed according to AR scheme with persistency

\[
\approx 1 - \psi N/T : \text{when } N \text{ grows to slowly with } T, \text{ IACT of } h(\vartheta_i) \text{ is dominated by IACT of } \mathbb{E}[h(\vartheta)|U_i] \text{ where for scalar } h(\vartheta) = \vartheta
\]

\[
\mathbb{E}[\vartheta|U_i] = \hat{\vartheta}_T + \frac{\sum}{T} \nabla_\vartheta \log \frac{\hat{p}(Y_{1:T} | \hat{\vartheta}_T, U_i)}{p(Y_{1:T} | \hat{\vartheta}_T)} + O_P(T^{-2}).
\]

- **Proposition.** Let \(N \propto T^\alpha\) for \(0 < \alpha < 1\) then

\[
\text{IACT}(Q, \Psi) \gtrsim T^{1-2\alpha}.
\]

- This result suggests we need at least \(\sqrt{T}/N = O(1)\).
Graphical Illustration

Modified score error

acf (score error)

θ

acf (θ)

θ - modified score error

acf (θ - modified score error)
Example: Gaussian Latent Variable Model

<table>
<thead>
<tr>
<th>MH ($T = 8192$)</th>
<th>IACT($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PM ($\rho = 0.0$)</th>
<th>IACT($\theta$)</th>
<th>RCT($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>RIACT($\theta$)</td>
<td>RCT($\theta$)</td>
</tr>
<tr>
<td>5000</td>
<td>2.2</td>
<td>11210</td>
</tr>
<tr>
<td>CPM ($\rho = 0.9963$)</td>
<td>RIACT($\theta$)</td>
<td>RCT($\theta$)</td>
</tr>
<tr>
<td>$N$</td>
<td>$\kappa$</td>
<td>RIACT($\theta$)</td>
</tr>
<tr>
<td>10</td>
<td>3.1</td>
<td>14.0</td>
</tr>
<tr>
<td>20</td>
<td>2.2</td>
<td>4.7</td>
</tr>
<tr>
<td>25</td>
<td>2.0</td>
<td>2.8</td>
</tr>
<tr>
<td>35</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>56</td>
<td>1.3</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Here $\text{RIACT} = \frac{\text{IACT}}{\text{IACT}_{\text{MH}}}$ and $\text{RCT} = N \times \text{RIACT}$. Improvement by 180 fold.
**Example: Noisy Autoregressive Model**

<table>
<thead>
<tr>
<th>MH ($T = 16,000$)</th>
<th>IACT($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PM ($\rho = 0.0$)</th>
<th>RIACT($\theta$)</th>
<th>RCT($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$ 2500</td>
<td>3.1</td>
<td>8427.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CPM ($\rho = 0.9965$)</th>
<th>RIACT($\theta$)</th>
<th>RCT($\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$\kappa$</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6.7</td>
<td>43.8</td>
</tr>
<tr>
<td>10</td>
<td>3.3</td>
<td>8.7</td>
</tr>
<tr>
<td>16</td>
<td>1.9</td>
<td>6.0</td>
</tr>
<tr>
<td>22</td>
<td>1.3</td>
<td>3.9</td>
</tr>
<tr>
<td>35</td>
<td>0.8</td>
<td>2.4</td>
</tr>
<tr>
<td>40</td>
<td>0.7</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Improvement by 100 fold.
Heston Stochastic Volatility Model

- Inference for a discretized one-dimensional SDE, 40,000 latent variables. 100-fold gain compared to PM.
Heston Stochastic Volatility Model

- Inference for a discretized one-dimensional SDE, 40,000 latent variables. 100-fold gain compared to PM.
- Real data: 4,000 returns from the S&P 500 index from 15/08/1990 to 03/07/2006.
Higher-dimensional SSM

- CLT appears to hold only for $N$ at least of order $T^{n/n+1}$ where $n$ is state-dimension.
Higher-dimensional SSM

- CLT appears to hold only for $N$ at least of order $T^{n/n+1}$ where $n$ is state-dimension.
- Long-range effect is consequently limited.
CLT appears to hold only for $N$ at least of order $T^{n/n+1}$ where $n$ is state-dimension.

Long-range effect is consequently limited.

Still significant gains over PM: over 50 fold for 2-d complex SV model, over 70-fold for 4-d model in (Jacob et al., 2016).
Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O(T^2)$.
Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O(T^2)$.

Correlated pseudo-marginal can achieve very substantial improvement.
Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O\left(T^2\right)$.

Correlated pseudo-marginal can achieve very substantial improvement.

In i.i.d. case, analysis shows $\sqrt{T}/N = O(1)$ is necessary and we conjecture it is sufficient leading to complexity $O\left(T^{3/2}\right)$ vs $O\left(T^2\right)$. 

Implementation for state-space models in state dimension $n > 1$ relies on non-standard particle scheme (e.g., Gerber & Chopin, 2015): our analysis does not capture these cases, experimental results suggest $O\left(T^{1+n/n+1}\right)$. 

(IMS, Singapore 2018)
• Large sample analysis of pseudo-marginal algorithm provides useful guidelines, overall complexity is $O(T^2)$.

• Correlated pseudo-marginal can achieve very substantial improvement.

• In i.i.d. case, analysis shows $\sqrt{T}/N = O(1)$ is necessary and we conjecture it is sufficient leading to complexity $O(T^{3/2})$ vs $O(T^2)$.

• Implementation for state-space models in state dimension $n > 1$ relies on non-standard particle scheme (e.g., Gerber & Chopin, 2015): our analysis does not capture these cases, experimental results suggest $O(T^{1+ \frac{n}{n+1}})$. 
Some References