Large Sample Asymptotics of the Pseudo-Marginal Method

Sebastian M Schmon, Arnaud Doucet, George Deligiannidis and Michael K Pitt

UNIVERSITY OF OXFORD DEPARTMENT OF STATISTICS

Introduction: Bayesian Inference and MCMC

- Observed data $y_{1:T} := (y_1, \dots, y_T)$, $T \geq 1$, likelihood function $p_\theta(y_{1:T})$ where $\theta \in \Theta \subseteq \mathbb{R}^d$ and prior density $p(\theta)$.
- Bayesian inference relies on the posterior

$$\pi_T(\theta) = p(\theta | y_{1:T}) = \frac{p_\theta(y_{1:T})\, p(\theta)}{\int_\Theta p_{\theta'}(y_{1:T})\, p(\theta')\, d\theta'}.$$

- For complex models inference relies usually on Markov chain Monte Carlo techniques, i.e. one simulates an ergodic Markov chain $\{\vartheta_i\}_{i\geq 1}$ with limiting distribution $\pi_T(\theta)$.
- **Problem**: Metropolis-Hastings (MH) cannot be implemented if $p_\vartheta(y_{1:T})$ cannot be evaluated.

Pseudo-Marginal Approach (I)

"Idea": Replace $p_\vartheta(y_{1:T})$ by an estimate $\widehat{p}_\vartheta(y_{1:T})$ in MH.
At iteration i

- Sample $\vartheta \sim q(\cdot | \vartheta_{i-1})$.
- Compute an estimate $\widehat{p}_\vartheta(y)$ of $p_\vartheta(y)$.
- With probability

$$\min\left\{1, \frac{\widehat{p}_\vartheta(y_{1:T})\, p(\vartheta)}{\widehat{p}_{\vartheta_{i-1}}(y)\, p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1}|\vartheta)}{q(\vartheta|\vartheta_{i-1})}\right\}$$

$$= \min\left\{1, \underbrace{\frac{p_\vartheta(y_{1:T})}{p_{\vartheta_{i-1}}(y_{1:T})} \frac{p(\vartheta)}{p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1}|\vartheta)}{q(\vartheta|\vartheta_{i-1})}}_{\text{exact MH ratio}} \times \underbrace{\frac{\widehat{p}_\vartheta(y_{1:T})/p_\vartheta(y_{1:T})}{\widehat{p}_{\vartheta_{i-1}}(y_{1:T})/p_{\vartheta_{i-1}}(y_{1:T})}}_{\text{noise}}\right\},$$

set $\vartheta_i = \vartheta$, $\widehat{p}_{\vartheta_i}(y_{1:T}) = \widehat{p}_\vartheta(y_{1:T})$ otherwise set $\vartheta_i = \vartheta_{i-1}$, $\widehat{p}_{\vartheta_i}(y_{1:T}) = \widehat{p}_{\vartheta_{i-1}}(y_{1:T})$.

Pseudo-Marginal Approach (II)

Proposition: If $\widehat{p}_\theta(y_{1:T})$ is a non-negative unbiased estimator of $p_\theta(y_{1:T})$ then the pseudo-marginal MH admits $\pi_T(\theta)$ as invariant density.

Let $Z_T(\theta) = \log\{\widehat{p}_\theta(y_{1:T})/p_\theta(y_{1:T})\}$ be the error in log-likelihood estimator and introduce an auxiliary target density on $\Theta \times \mathbb{R}$

$$\overline{\pi}_T(\theta, z) = \pi_T(\theta) \underbrace{\exp(z) g_T(z \mid \theta)}_{\text{unbiasedness} \Leftrightarrow \int(\cdot)dz=1}$$

where $Z_T(\theta) \sim g_T(\cdot \mid \theta)$; e.g. importance sampling or particle filter.
Pseudo-marginal MH is a standard MH of target $\overline{\pi}_T(\theta, z)$ and proposal $q(\vartheta|\theta)\, g_\vartheta(z)$ as

$$\frac{\overline{\pi}_T(\vartheta, w)}{\overline{\pi}_T(\theta, z)} \frac{q(\theta|\vartheta)\, g_\theta(z)}{q(\vartheta|\theta)\, g_\vartheta(w)} = \frac{\widehat{p}_\vartheta(y_{1:T})}{\widehat{p}_\theta(y_{1:T})} \frac{p(\vartheta)}{p(\theta)} \frac{q(\theta|\vartheta)}{q(\vartheta|\theta)}.$$

Assumptions

Assumption 1: The posterior $\{\pi_T(d\theta);\, T \geq 1\}$ concentrates in a "Bernstein-von Mises" sense:

$$\int \left|\pi_T(\theta) - \varphi(\theta; \widehat{\theta}_T, \Sigma/T)\right| d\theta \xrightarrow{\mathbb{P}^Y} 0, \qquad \widehat{\theta}_T \xrightarrow{\mathbb{P}^Y} \bar{\theta},$$

with covariance matrix Σ and estimators $\widehat{\theta}_T$.
Assumption 2: The proposal distributions $\{q_T(\theta, d\theta');\, T \geq 1\}$ are scaled with the data

$$\theta' = \theta + \frac{\xi}{\sqrt{T}}, \quad \xi \sim \nu(\cdot)$$

where ν is a continuous probability density on $\mathbb{R}^d$ with $\mathbb{E}_\nu[\|\xi\|] < \infty$.
Assumption 3: There exists an ε-ball $B(\bar{\theta})$ around $\bar{\theta}$ such that the distributions of the error in the log-likelihood estimator $\{g_T(dz \mid \theta);\, T \geq 1\}$ satisfy

$$\sup_{\theta \in B(\bar{\theta})} d_{\mathrm{BL}}\left(g_T(\cdot \mid \theta), \varphi(\cdot\,; -\sigma^2(\theta)/2, \sigma^2(\theta))\right) \xrightarrow{\mathbb{P}^Y} 0,$$

where d_{BL} denotes the bounded Lipschitz metric, $\sigma\colon \Theta \to [0, \infty)$ is continuous at $\bar{\theta}$ with $0 < \sigma(\bar{\theta}) < \infty$. An analogous result holds for $\bar{g}_T(z \mid \theta) = \exp(z) g_T(z \mid \theta)$, the distribution of this error at equilibrium, that is

$$\sup_{\theta \in B(\bar{\theta})} d_{\mathrm{BL}}\left(\bar{g}_T(\cdot \mid \theta), \varphi(\cdot\,; \sigma^2(\theta)/2, \sigma^2(\theta))\right) \xrightarrow{\mathbb{P}^Y} 0.$$

References and related work

Deligiannidis, G.; Doucet, A. & Pitt, M. K. *The Correlated Pseudo-Marginal Method*, Journal of the Royal Statistical Society Series B, to appear

Doucet, A.; Pitt, M. K.; Deligiannidis, G. & Kohn, R. *Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator* Biometrika, 2015, 102 (2), 295-313

Pitt, M. K., dos Santos Silva, R., Giordani, P., & Kohn, R. *On some properties of Markov chain Monte Carlo simulation methods based on the particle filter* Journal of Econometrics, 2012, 171(2), 134-151.

Sherlock, C.; Thiery, A. H.; Roberts, G. O.; Rosenthal, J. S. & others *On the efficiency of pseudo-marginal random walk Metropolis algorithms* The Annals of Statistics, 2015, 43, 238-275

Parameter Rescaling

Consider the pseudo-marginal chain $\{(\vartheta_{T,k}, Z_{T,k});\, k \geq 0\}$ started at $(\vartheta_{T,0}, Z_{T,0}) \sim \overline{\pi}_T$. Update $(\vartheta_{T,k}, Z_{T,k}) \sim P_T(\vartheta_{T,k-1}, Z_{T,k-1}; \cdot)$ for $k \geq 1$. Let $\chi_T = \{(\widetilde{\vartheta}_{T,k}, Z_{T,k});\, k \geq 0\}$ where $\widetilde{\vartheta}_{T,k} := \sqrt{T}(\vartheta_{T,k} - \widehat{\theta}_T)$ is the Markov chain arising from rescaling the parameter component of the pseudo-marginal chain. The transition kernel is

$$\widetilde{P}_T(\widetilde{\theta}, z; d\widetilde{\theta}', dz') = \widetilde{q}_T(\widetilde{\theta}, d\widetilde{\theta}')\widetilde{g}_T(dz'|\widetilde{\theta}')\widetilde{\alpha}_T(\widetilde{\theta}, z; \widetilde{\theta}', z') + \widetilde{\rho}_T(\widetilde{\theta}, z)\delta_{(\widetilde{\theta}, z)}(d\widetilde{\theta}', dz'),$$

where

$$\widetilde{\alpha}_T(\widetilde{\theta}, z; \widetilde{\theta}', z') = \min\left\{1, \frac{\widetilde{\pi}_T(d\widetilde{\theta}')}{\widetilde{\pi}_T(d\widetilde{\theta})} \frac{\widetilde{q}_T(\widetilde{\theta}', d\widetilde{\theta})}{\widetilde{q}_T(\widetilde{\theta}, d\widetilde{\theta}')} \exp(z' - z)\right\},$$

$\widetilde{\rho}_T(\theta, z)$ is the corresponding rejection probability, $\widetilde{\pi}_T(\widetilde{\theta}) := \pi_T(\widehat{\theta}_T + \widetilde{\theta}/\sqrt{T})/\sqrt{T}$, $\widetilde{q}_T(\widetilde{\theta}, \widetilde{\theta}') := q_T(\widehat{\theta}_T + \widetilde{\theta}/\sqrt{T}, \widehat{\theta}_T + \widetilde{\theta}'/\sqrt{T})/\sqrt{T}$ and $\widetilde{g}_T(z|\widetilde{\theta}) := g_T(z|\widehat{\theta}_T + \widetilde{\theta}/\sqrt{T})$.

Theorem (Weak Convergence):

Under Assumptions 1, 2 and 3, the sequence of stationary Markov chains $(\chi_T; T \geq 1)$ converges weakly in $\mathbb{P}^Y$-probability as $T \to \infty$ to the law of a stationary Markov chain of initial distribution

$$\widetilde{\pi}(d\widetilde{\theta}, dz) := \varphi(d\widetilde{\theta}; 0, \Sigma)\varphi\left(dz; \sigma^2/2, \sigma^2\right)$$

and transition kernel

$$\widetilde{P}(\widetilde{\theta}, z; d\widetilde{\theta}', dz') = \widetilde{q}(\widetilde{\theta}, d\widetilde{\theta}')\varphi\left(dz'; -\sigma^2/2, \sigma^2\right)\widetilde{\alpha}(\widetilde{\theta}, z; \widetilde{\theta}', z') + \widetilde{\rho}(\widetilde{\theta}, z)\delta_{(\widetilde{\theta}, z)}(d\widetilde{\theta}', dz')$$

where $\sigma := \sigma(\bar{\theta})$,

$$\widetilde{\alpha}(\widetilde{\theta}, z; \widetilde{\theta}', z') = \min\left\{1, \frac{\varphi(\widetilde{\theta}'; 0, \Sigma)}{\varphi(\widetilde{\theta}; 0, \Sigma)} \frac{\widetilde{q}(\widetilde{\theta}', \widetilde{\theta})}{\widetilde{q}(\widetilde{\theta}, \widetilde{\theta}')} \exp(z' - z)\right\},$$

and $\widetilde{\rho}(\theta, z)$ is the corresponding rejection probability.

Simulation Study: Tuning the Pseudo-Marginal algorithm

We optimize the performance of the limiting pseudo-marginal chain identified above as a proxy for the optimization of the original pseudo-marginal chain. We consider a Gaussian random walk proposal parameterized by ℓ

$$q(\theta, \theta') = \varphi\left(\theta'; \theta, \frac{\ell^2}{d} I_d\right).$$

Denoting τ the integrated autocorrelation time, we minimize the computing time

$$\mathrm{ct}(h, \widetilde{P}_{\ell,\sigma}) = \frac{\tau(h, \widetilde{P}_{\ell,\sigma})}{\sigma^2}$$

over a grid $(\ell, \sigma) \in \{1.8, 1.9, \dots, 2.7\} \times \{1, 1.1, \dots, 2\}$. We restrict attention here to the case where $h(\theta, z) = \theta_1$, the first component of θ.

Dimension d	$\hat{\ell}_{\text{opt}}$	$\hat{\sigma}_{\text{opt}}$	$\mathrm{ct}(\hat{\sigma}_{\text{opt}}, \hat{\ell}_{\text{opt}})$	$p_{\text{jump}}(\hat{\sigma}_{\text{opt}}, \hat{\ell}_{\text{opt}})$
$d = 1$	2.05 (0.25)	1.16 (0.07)	8.47	25.73%
$d = 2$	1.97 (0.14)	1.21 (0.06)	12.71	22.92%
$d = 3$	2.11 (0.07)	1.24 (0.05)	16.79	19.97%
$d = 5$	2.17 (0.12)	1.30 (0.05)	23.18	17.35%
$d = 10$	2.20 (0.08)	1.44 (0.05)	37.93	14.27%
$d = 15$	2.33 (0.08)	1.50 (0.00)	53.43	12.07%
$d = 20$	2.34 (0.10)	1.54 (0.05)	65.62	11.44%
$d = 30$	2.36 (0.11)	1.61 (0.03)	90.46	10.41%
$d = 50$	2.41 (0.10)	1.74 (0.05)	136.38	8.66%

Table 1: Optimal values for scaling ℓ and noise σ and associated value of computing time and average acceptance probability.

Real Data: Indonesian Preschool Children

We now consider a Bayesian logistic mixed effects model applied to a real data set with linear predictor

$$\eta_{t,j} = x_{t,j}^T \beta + U_t, \quad U_t \sim \mathcal{N}(0, \tau),$$

where U_t denotes the random intercept for children $t = 1, \dots, T$ and β the regression parameters. The observations are assumed conditionally independent given the random effects and the likelihood of the population parameters is

$$p(y_{1:T} \mid \beta, \tau) = \prod_{t=1}^{T} \int \prod_{j=1}^{J} \frac{\exp(y_{t,j}\eta_{t,j})}{1 + \exp(\eta_{t,j})} \varphi(du_t, 0, \tau).$$