Approximate Computation for Approximate Bayesian Models

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with Evgeny Levi (Toronto)

ICSDS, December 2022

MCMC at the crossroads

- Large data and/or intractable likelihoods have brought Bayesian computation at a crossroads.
- Consider observed data $\mathbf{y}_0 \in \mathcal{Y}$, likelihood function $L(\boldsymbol{\theta}|\mathbf{y}_0)$ (or sampling distribution $f(\mathbf{y}|\boldsymbol{\theta})$), prior $p(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \mathbf{R}^d$.
- Focus is on $\pi(\boldsymbol{\theta}|\mathbf{y}_0) \propto f(\mathbf{y}_0|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- The Metropolis-Hastings sampler is one of the most used algorithms in MCMC.
  - Given the current state of the chain θ, draw $\xi \sim q(\xi|\theta)$.
  - Accept ξ with probability $\min\left\{1, \frac{\pi(\xi|\mathbf{y}_0)q(\theta|\xi)}{\pi(\theta|\mathbf{y}_0)q(\xi|\theta)}\right\}$.
  - If ξ is accepted, the next state is ξ, otherwise it is (still) θ.
- Note that $\pi(\boldsymbol{\theta}|\mathbf{y}_0) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{y}_0)$ needs to be computed at each iteration. (hence $L(\theta|\mathbf{y}_0)$ must also be computable)

Challenge 1: Massive data set

- $L(\theta|\mathcal{D})$ is computable, but data is massive.
- Precomputing (Boland et al., EJS, 2018)
- Sequential processing (Bardenet et el. 2014; Korratikara et al. 2014)
- Divide and conquer (Neiswanger et al. 2013; Wang and Dunson 2013; Scott et al. 2016; Entezari et al. 2018; Nemeth and Sherlock 2018; Changye and Robert 2019)
- Subsampling (Quiroz et al. 2018; Campbell and Broderick 2019)

Challenge 2: Intractable likelihoods

- When the likelihood $L(\theta|\mathbf{y}_0)$ is not computable but one can sample from $f(\mathbf{y}|\theta)$ for all θ's....
- Approximate Bayesian Computation (ABC)
- Bayesian Synthetic Likelihood (BSL)

Double jeopardy: Large data and Intractable Likelihood

- The generation of pseudo-data can be expensive, e.g. climate change scenarios (Oyebamiji et al. 2015) or hurricane surges (Plumlee et al. 2021)
- Most of methods that address the challenge of large data cannot be used directly for intractable models.
- Today: discuss an approach that can be used with ABC and BSL.

ABC

- ABC:
  - Sample $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and $\mathbf{y} \sim f(\mathbf{y}|\boldsymbol{\theta})$;
  - Compute distance:

$$\delta(\mathbf{y}) := \|\mathbf{S}(\mathbf{y}), \mathbf{S}(\mathbf{y}_0)\| = \sqrt{[\mathbf{S}(\mathbf{y}) - \mathbf{S}(\mathbf{y}_0)]^T A \,[\mathbf{S}(\mathbf{y}) - \mathbf{S}(\mathbf{y}_0)]}$$

  - If $\delta(\mathbf{y}) < \epsilon$ retain $(\boldsymbol{\theta}, \mathbf{y})$ as a draw from

$$\pi_\epsilon(\boldsymbol{\theta}, \mathbf{y}|\mathbf{y}_0) \propto p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \mathbf{1}_{\{\delta(\mathbf{y}) < \epsilon\}}$$

- The marginal target (in $\boldsymbol{\theta}$) is

$$\begin{aligned}
\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}_0) &= \int_{\mathcal{Y}} \pi_\epsilon(\boldsymbol{\theta}, \mathbf{y}|\mathbf{y}_0) d\mathbf{y} \propto \\
&\propto p(\boldsymbol{\theta}) \underbrace{\int_{\mathcal{Y}} f(\mathbf{y}|\boldsymbol{\theta}) \mathbf{1}_{\{\delta(\mathbf{y}) \le \epsilon\}} d\mathbf{y}}_{\text{approximate likelihood}} = p(\boldsymbol{\theta}) \underbrace{\Pr(\delta(\mathbf{y}) \le \epsilon | \boldsymbol{\theta}, \mathbf{y}_0)}_{:= h(\boldsymbol{\theta})}
\end{aligned}$$

Zooming in on the target

- We consider building a chain with target $\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}_0) \propto p(\boldsymbol{\theta})h(\boldsymbol{\theta})$.
- Consider proposal $\xi_{t+1} \sim q(\xi|\boldsymbol{\theta}_t)$
- A Metropolis-Hastings sampler requires calculating

$$\frac{p(\xi_{t+1})h(\xi_{t+1})q(\boldsymbol{\theta}_t|\xi_{t+1})}{p(\boldsymbol{\theta}_t)h(\boldsymbol{\theta}_t)q(\xi_{t+1}|\boldsymbol{\theta}_t)}$$

A marginal yet important target

- Lee et al (2012) propose to use $\mathbf{w}_1, \ldots, \mathbf{w}_J \sim f(\mathbf{w}|\xi)$ to estimate

$$\widehat{h}(\xi) = J^{-1} \sum_{j=1}^{J} \mathbf{1}_{\{\delta(\mathbf{w}_j)<\epsilon\}}$$

- Wilkinson (2013) generalizes to smoothing kernels
- Bornn et al (2014) make the case of using $J = 1$.
- Idea in this talk: Recycle past proposals to estimate $h(\xi)$.

History repeating itself

- At time n the proposal is $(\zeta_{n+1}, \mathbf{w}_{n+1}) \sim q(\zeta|\theta_n) f(\mathbf{w}|\zeta)$
- At iteration n, all the proposals $\{\zeta_k\}_{k=1:n}$, either accepted or rejected, and distances $\delta_k = \delta(\mathbf{w}_k)$ are available.
- This is the history, denoted $\mathcal{Z}_n$, of the chain.

A selective memory helps

- Given a new proposal $\zeta_{n+1} \sim q(|\theta_n)$, we generate $\mathbf{w}_{n+1} \sim f(\cdot|\zeta_{n+1})$ and compute $\delta_{n+1} = \delta(\mathbf{w}_{n+1})$. Let $\mathcal{Z}_{n+1} = \mathcal{Z}_n \cup \{(\zeta_{n+1}, \delta_{n+1})\}$ and estimate $h(\zeta^*)$ using

$$\hat{h}(\zeta^*) = \frac{\sum_{k=1}^{n} W_k(\zeta_{n+1}) \mathbf{1}_{\delta_k < \epsilon}}{\sum_{k=1}^{n} W_k(\zeta_{n+1})}, \tag{1}$$

  where $W_k(\zeta_{n+1}) = W(\|\zeta_k - \zeta_{n+1}\|)$ are weights and $W : \mathbf{R} \to [0, \infty)$ is a decreasing function.

- Alternatively, use a subset of the K closest ζ_ks in $\mathcal{Z}_n$

Good news

- If $\delta_{n+1} > \epsilon \Rightarrow$ rejection for ABC-MCMC
- But if $\exists \zeta_k$ with a corresponding $\delta_k < \epsilon$ then $h(\zeta_{n+1}) \neq 0$
- Compare

$$\tilde{h}(\zeta^*) = \frac{1}{K} \sum_{j=1}^{K} \mathbf{1}_{\{\tilde{\delta}_j < \epsilon\}} \quad \Rightarrow \quad \text{unbiased}$$

$$\hat{h}(\zeta^*) = \frac{\sum_{n=1}^{N} W_{Nn}(\zeta^*) \mathbf{1}_{\{\tilde{\delta}_n < \epsilon\}}}{\sum_{n=1}^{N} W_{Nn}(\zeta^*)} \quad \Rightarrow \quad \text{consistent}$$

- When K is small - reduce variability.
- When K is large - reduce costs.

Complications

- If the past samples are used to modify the kernel $\Rightarrow$ Adaptive MCMC
- In order to avoid AMCMC conditions for validity, we separate the samples used as proposals from those used to estimate h
- At each time t:
  - We use the Independent Metropolis sampler, i.e. $q(\zeta|\theta^{(t)}) = q(\zeta)$
  - Generate two independent samples

$$\{(\zeta_{t+1}, \mathbf{w}_{t+1}), (\tilde{\zeta}_{t+1}, \tilde{\mathbf{w}}_{t+1})\} \overset{\text{iid}}{\sim} q(\zeta) f(\mathbf{w}|\zeta)$$

  - Set $\mathcal{Z}_{N+1} = \mathcal{Z}_N \cup \{(\tilde{\zeta}_{N+1}, \tilde{\delta}_{N+1})\}$

Friendly neighbors

- The k-Nearest-Neighbor (kNN) regression approach has a property of uniform consistency
- Set $K = \sqrt{n}$ and relabel history so that $(\tilde{\zeta}_1, \tilde{\delta}_1)$ and $(\tilde{\zeta}_n, \tilde{\delta}_n)$ corresponds to the smallest and largest among all distances $\{\|\tilde{\zeta}_j - \zeta_{n+1}\| : \ 1 \leq j \leq n\}$
- Weights are defined as:
  - $W_k = 0$ for $k > K$

  (U) The *uniform* kNN with $W_k(\zeta_{n+1}) = 1$ for all $k \leq K$;

  (L) The *linear* kNN with $W_k(\zeta_k) = W(\|\tilde{\zeta}_k - \zeta_{n+1}\|) = 1 - \|\tilde{\zeta}_k - \zeta_{n+1}\|/\|\tilde{\zeta}_K - \zeta_{n+1}\|$ for $k \leq K$ so that the weight decreases from 1 to 0 as k increases from 1 to K.

A bit of theory

(B1) Θ is a compact set.

(B2) $q(\theta) > 0$ is a continuous density (proposal).

(B3) $p(\theta) > 0$ is a continuous density (prior).

(B4) $h(\theta)$ continuous function of θ.

(B5) In kNN estimation assume that $K(n) = \sqrt{n}$ with uniform or linear weights.

Some comfort

- Let $P(\cdot, \cdot)$ denote the transition kernel of our AABC sampler, when $h(\boldsymbol{\theta})$ is computed exactly.
- μ denotes stationary distribution for $P(\cdot, \cdot)$
- The approximate kernel at time t is denoted $\hat{P}_t$
- The distribution of $\boldsymbol{\theta}_t$ is denoted $\mu_t := \nu \hat{P}_1 \ldots ... \hat{P}_t$

Some comfort

Vanishing TV Theorem

Suppose that (**A1**)- (**A3**) are satisfied . Let π denote the invariant measure of P and ν be any probability measure on $(\Theta, \mathcal{F}_0)$, then

$$\left\| \mu - \frac{\sum_{t=0}^{M-1} \nu \hat{P}_1 \cdots \hat{P}_t}{M} \right\|_{TV} \leq O(M^{-1}) + O(M^{-1}\epsilon) + O(\epsilon),$$

More Comfort

Vanishing MSE Theorem

Let π denote the invariant measure of P, $f(\theta)$ be a bounded function and $\theta^{(0)} \sim \nu$. Then

$$E\left[\left(\mu f - \frac{1}{M}\sum_{t=0}^{M-1} f(\theta^{(t)})\right)^2\right] \leq |f|^2[O(M^{-1}) + O(\epsilon^2) + O(M^{-1}\epsilon)]$$

where $\mu f = E_\mu f$.

Numerical Experiments: Ricker's Model

- A particular instance of hidden Markov model:

$$x_{-49} = 1; \quad z_i \overset{iid}{\sim} \mathcal{N}(0, \exp(\theta_2)^2); \quad i = \{-48, \cdots, n\},$$
$$x_i = \exp(\exp(\theta_1))x_{i-1} \exp(-x_{i-1} + z_i); \quad i = \{-48, \cdots, n\},$$
$$y_i = Pois(\exp(\theta_3)x_i); \quad i = \{-48, \cdots, n\},$$

  where $Pois(\lambda)$ is Poisson distribution
- Only $\mathbf{y} = (y_1, \cdots, y_n)$ sequence is observed, because the first 50 values are ignored.

Numerical Experiments: Ricker's Model

Define summary statistics $S(\mathbf{y})$ as the 14-dimensional vector whose components are:

(C1) $\#\{i : y_i = 0\}$,

(C2) Average of $\mathbf{y}$, $\bar{y}$,

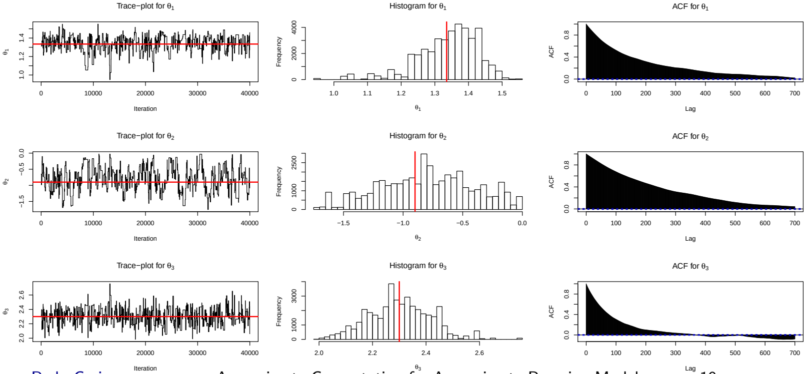
(C3:C7) Sample auto-correlations at lags 1 through 5,

(C8:C11) Coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ of cubic regression $(y_i - y_{i-1}) = \beta_0 + \beta_1 y_i + \beta_2 y_i^2 + \beta_3 y_i^3 + \epsilon_i$, $i = 2, \dots, n$,

(C12-C14) Coefficients $\beta_0, \beta_1, \beta_2$ of quadratic regression $y_i^{0.3} = \beta_0 + \beta_1 y_{i-1}^{0.3} + \beta_2 y_{i-1}^{0.6} + \epsilon_i$, $i = 2, \dots, n$.

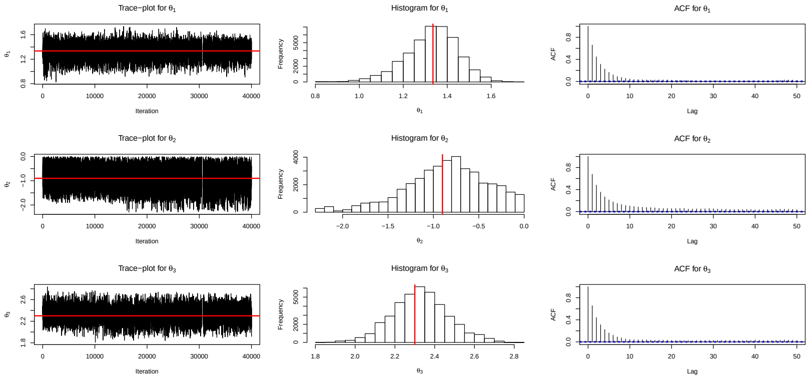
Numerical Experiments: Ricker's Model - ABC/RWM

Figure: Ricker's model: ABC-RW Sampler. Each row corresponds to parameters θ_1 (top row), θ_2 (middle row) and θ_3 (bottom row) and shows in order from left to right: Trace-plot, Histogram and Auto-correlation function. Red lines represent true parameter values.



Numerical Experiments: Ricker's Model - ABC

Figure: Ricker's model: AABC-U Sampler.



Numerical Experiments: Ricker's Model - ABC

	Diff with exact			Diff with true parameter			Efficiency	
Sampler	DIM	DIC	TV	$\sqrt{\text{Bias}^2}$	$\sqrt{\text{VAR}}$	$\sqrt{\text{MSE}}$	ESS	ESS/CPU
ABC-RW	0.135	0.0201	0.389	0.059	0.180	0.189	87	0.199
AABC-U	0.147	0.0279	0.402	0.076	0.190	0.204	3563	4.390
AABC-L	0.141	0.0258	0.392	0.070	0.189	0.201	4206	5.193
BSL-RW	0.129	0.0080	0.382	0.038	0.206	0.209	131	0.030
ABSL-U	0.103	0.0054	0.377	0.023	0.170	0.171	284	0.180
ABSL-L	0.106	0.0051	0.382	0.012	0.173	0.173	207	0.135

Table: Summaries based on 40K samples

Concluding remarks

- Precomputation! Useful also for Bayesian synthetic likelihood methods.
- We obtain good results even if $q(\xi|\theta) = \mathcal{N}(\theta, \Sigma)$ but more theory needed.
- The computational burden can prohibit the full reach of approximate methods so more solutions are needed.
- Computation $\stackrel{\heartsuit}{\rightarrow}$ Statistics.
- Is it time for more Statistics $\stackrel{\heartsuit}{\rightarrow}$ Computation?

All papers available at:
`http://www.utstat.toronto.edu/craiu/`