Approximate MCMC for Approximate Bayesian Methods

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with Evgeny Levi (Toronto)

Ca'Foscari University
February, 2020

Outline

Introduction & Motivation
- Introduction
- Motivation

Approximate Bayesian Computation (ABC)
- Vanilla version
- ABC-MCMC
- Recycler ABC-MCMC
- Bayesian Synthetic Likelihood (BSL)

Theory
- Vanishing TV distance
- Vanishing MSE

Numerical Experiments
- Ricker's Model
- Stochastic Volatility

A Bayesian's Best Friend: MCMC

- Due to MCMC developments, for 30+ years Bayesian statisticians were *computationally liberated* when thinking about a statistical model.
- Large data and/or intractable likelihoods have brought Bayesian computation at a crossroads.

A Bayesian's Best Friend: MCMC

- Consider observed data $\mathbf{y}_0 \in \mathcal{Y}$ and likelihood function $L(\boldsymbol{\theta}|\mathbf{y})$ (or sampling distribution $f(\mathbf{y}|\boldsymbol{\theta})$) where $\boldsymbol{\theta} \in \mathbf{R}^d$.
- For a prior $p(\boldsymbol{\theta})$ the posterior is $\pi(\boldsymbol{\theta}|\mathbf{y}_0) \propto f(\mathbf{y}_0|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- The Metropolis-Hastings sampler is one of the most used algorithms in MCMC. It operates as follows:
  - 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)

Massive data set

- $L(\theta|\mathcal{D})$ is computable, but data is massive.
- Possible remedies: divide and conquer, sequential processing, pseudomarginal, precomputing, etc
- D &C: Divide data into batches, $\mathbf{y}^{(1)} \cup \ldots \mathbf{y}^{(K)}$, distribute the sampling from the K sub-posteriors

$$\pi_j(\theta) \propto [L_k(\theta|\mathbf{y}^{(j)})]^a [p_j(\theta)]^b$$

  among K processing units
- Depending on a, b values, design recombination strategies for the π_j-samples to recover the characteristics of the full posterior distribution.
- Entezari et al (2018) applied D&C to BART.

Motivation for ABC

- When the likelihood $L(\theta|\mathbf{y})$ is not computable but one can sample from $p(\mathbf{y}|\theta)$ for all θ's....

Motivation for ABC

- When the likelihood $L(\theta|\mathbf{y})$ is not computable but one can sample from $p(\mathbf{y}|\theta)$ for all θ's....
- Approximate Bayesian Computation (ABC - Marin et al., Comp & Stat. 2012) or Bayesian Synthetic Likelihood (BSL - Price et al, JCGS 2018) methods can be used.

A remarkable algorithm- ABC

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

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

Vanilla ABC

- Sampling candidate θ's from the prior is inefficient, especially if the prior is in conflict with the data (Evans and Moshonov, 2006).
- Marjoram et al (2003) propose an ABC-MCMC in which candidate moves are generated using a proposal $q(\theta|\theta_t)$ and they are accepted or rejected based on a MH-type rule.

Marjoram's ABC-MCMC

- Consider the joint target distribution in $(\boldsymbol{\theta}, \mathbf{y})$: $\pi_\epsilon(\boldsymbol{\theta}, \mathbf{y}|\mathbf{y}_0)$.
- A proposal $(\boldsymbol{\theta}, \mathbf{y}) \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) \times f(\mathbf{y}|\boldsymbol{\theta})$ is accepted using the Metropolis-Hastings acceptance ratio

$$
\begin{aligned}
\alpha &= \min\left\{1, \frac{\pi_\epsilon(\boldsymbol{\theta}, \mathbf{y}|\mathbf{y}_0) \times q(\boldsymbol{\theta}_t|\boldsymbol{\theta}) f(\mathbf{y}_t|\boldsymbol{\theta}_t)}{\pi_\epsilon(\boldsymbol{\theta}_t, \mathbf{y}_t|\mathbf{y}_0) \times q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) f(\mathbf{y}|\boldsymbol{\theta})}\right\} \mathbf{1}_{\{\delta(\mathbf{y}) \leq \epsilon\}} \\
&= \min\left\{1, \frac{p(\boldsymbol{\theta}) q(\boldsymbol{\theta}_t|\boldsymbol{\theta})}{p(\boldsymbol{\theta}_t) q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)}\right\} \mathbf{1}_{\{\delta(\mathbf{y}) \leq \epsilon\}}
\end{aligned}
$$

Zooming in on the target

- We consider building a chain with target $\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}_0)$.
- Set $h(\boldsymbol{\theta}) = \Pr(\delta(\mathbf{y}) < \epsilon|\boldsymbol{\theta}, \mathbf{y}_0)$ and proposal $\tilde{\boldsymbol{\theta}} \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$
- A Metropolis-Hastings sampler requires

$$\frac{p(\tilde{\boldsymbol{\theta}})h(\tilde{\boldsymbol{\theta}})q(\boldsymbol{\theta}_t|\tilde{\boldsymbol{\theta}})}{p(\boldsymbol{\theta}_t)h(\boldsymbol{\theta}_t)q(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}_t)}$$

A marginal yet important target

- Lee et al (2012) propose to use $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_J \sim f(\mathbf{y}|\tilde{\boldsymbol{\theta}})$ to estimate

$$\widehat{h}(\tilde{\boldsymbol{\theta}}) = J^{-1} \sum_{j=1}^{J} \mathbf{1}_{\{\delta(\tilde{\mathbf{y}}_j) < \epsilon\}}$$

- 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 $h(\tilde{\boldsymbol{\theta}})$.

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 ζ_n, the accepted and rejected ones, along with corresponding distances $\delta_n = \delta(\mathbf{w}_n)$ are available for $0 \leq n \leq N-1$.
- This is the history, denoted $\mathcal{Z}_{N-1}$, of the chain.

A selective memory helps

- Given a new proposal $\zeta^* \sim q(|\theta^{(t)})$, we generate $\mathbf{w}^* \sim f(\cdot|\zeta^*)$ and compute $\delta^* = \delta(S(\mathbf{w}^*))$. Set $\zeta_N = \zeta^*$, $\mathbf{w}_N = \mathbf{w}^*$, $\mathcal{Z}_N = \mathcal{Z}_{N-1} \cup \{(\zeta_N, \delta_N)\}$ and estimate $h(\zeta^*)$ using

$$\hat{h}(\zeta^*) = \frac{\sum_{n=1}^{N} W_{Nn}(\zeta^*)\mathbf{1}_{\delta_n < \epsilon}}{\sum_{n=1}^{N} W_{Nn}(\zeta^*)}, \tag{1}$$

  where $W_{Nn}(\zeta^*) = W(\|\zeta_n - \zeta^*\|)$ are weights and $W : \mathbf{R} \to [0, \infty)$ is a decreasing function.

- An alternative to (1) is to use a subset of size K of $\mathcal{Z}_N$

Good news

- If $\delta^* > \epsilon \Rightarrow$ rejection for ABC-MCMC
- But if $\exists \zeta^*$ with a corresponding $\delta < \epsilon$ then $h(\zeta^*) \neq 0$
- 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:
  - Generate two independent samples
$$\{(\zeta_{t+1}, \mathbf{w}_{t+1}), (\tilde{\zeta}_{t+1}, \tilde{\mathbf{w}}_{t+1})\} \overset{\text{iid}}{\sim} q(\zeta|\theta^{(t)})f(\mathbf{w}|\zeta)$$
  - Set $\mathcal{Z}_{N+1} = \mathcal{Z}_N \cup \{(\tilde{\zeta}_{N+1}, \tilde{\delta}_{N+1})\}$
- We use the Independent Metropolis sampler, i.e. $q(\zeta|\theta^{(t)}) = q(\zeta)$ so that the chain's trajectory is independent of $\mathcal{Z}$

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 $\{\|\tilde{\zeta}_j - \zeta^*\| : \ 1 \le j \le N\}$
- Weights are defined as:
  - $W_n = 0$ for $n > K$

(U) The *uniform* kNN with $W_{Nn}(\zeta^*) = 1$ for all $n \le K$;

(L) The *linear* kNN with $W_{Nn}(\zeta^*) = W(\|\tilde{\zeta}_n - \zeta^*\|) = 1 - \|\tilde{\zeta}_n - \zeta^*\| / \|\tilde{\zeta}_K - \zeta^*\|$ for $n \le K$ so that the weight decreases from 1 to 0 as n increases from 1 to K.

Indirect inference - A David and Goliath story

- Indirect inference (Gallant and McCulloch, 2009)
- Complex model: $f(\mathbf{y}|\boldsymbol{\theta})$ with intractable f
- Simpler model $g(\mathbf{y}|\phi(\boldsymbol{\theta}))$ approximates well $f(\mathbf{y}|\boldsymbol{\theta})$, with $\dim(\phi) > \dim(\boldsymbol{\theta})$, g is tractable and $\phi : \Theta \to \Phi$ is unknown
- We can estimate $\hat{\phi}(\boldsymbol{\theta})$ by sampling $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, $\mathbf{y}_j \sim f(\mathbf{y}|\boldsymbol{\theta})$, $1 \leq j \leq K$ and estimate ϕ from $\mathbf{y}_1, \dots, \mathbf{y}_K$ using g - repeat
- Posterior $\pi_f(\boldsymbol{\theta}|\mathbf{y}_0) \propto p(\boldsymbol{\theta}) f(\mathbf{y}_0|\boldsymbol{\theta})$ is then approximated by

$$\pi_g(\boldsymbol{\theta}|\mathbf{y}_0) \propto p(\boldsymbol{\theta}) g(\mathbf{y}_0|\hat{\phi}(\boldsymbol{\theta}))$$

Bayesian Synthetic Likelihood (BSL)

- Alternative approach to bypass the intractability of the sampling distribution proposed by Wood (*Nature*, 2010).
- The simpler model (g): the conditional distribution for a user-defined statistic $S(\mathbf{y})$ given θ is Gaussian with parameters $\phi(\boldsymbol{\theta}) = (\mu_\theta, \Sigma_\theta)$
- The Synthetic Likelihood (SL) procedure assigns to each θ the likelihood $SL(\theta) = \mathcal{N}(s_0; \mu_\theta, \Sigma_\theta)$.
- The BSL posterior is $\pi(\theta|s_0) \propto p(\theta)\mathcal{N}(s_0; \mu_\theta, \Sigma_\theta)$.
- Acceptance ratios for a MH sampler are estimated from m statistics $(s_1, \cdots, s_m)$ sampled from their conditional distribution given θ.

Bayesian Synthetic Likelihood (BSL)

- Generate $\mathbf{y}_i \sim f(\mathbf{y}|\theta)$ and set $s_i = S(\mathbf{y}_i)$, $i = 1, \cdots, m$
- Estimate

$$\hat{\mu}_\theta = \frac{\sum_{i=1}^m s_i}{m},$$
$$\hat{\Sigma}_\theta = \frac{\sum_{i=1}^m (s_i - \hat{\mu}_\theta)(s_i - \hat{\mu}_\theta)^T}{m-1},$$

- The synthetic likelihood is

$$SL(\theta|\mathbf{y}_0) = \mathcal{N}(S(\mathbf{y}_0); \hat{\mu}_\theta, \hat{\Sigma}_\theta). \qquad (2)$$

- Acceptance probability requires repeated estimation of (2)

$$\min\left\{1, \frac{p(\theta)S(\theta|\mathbf{y}_0)q(\theta_t)}{p(\theta_t)S(\theta_t|\mathbf{y}_0)q(\theta)}\right\}$$

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, if $h(\boldsymbol{\theta})$ were computable exactly.
- The invariant distribution of $P(\cdot, \cdot)$ is denoted μ
- 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 $\theta^{(0)} \sim \nu$, where ν is a probability distribution. 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: General Setup

- Efficiency measures

$$\text{Diff in mean (DIM)} = Mean_{r,s}(|Mean_t(\theta_{rs}^{(t)}) - Mean_t(\tilde{\theta}_{rs}^{(t)})|),$$

$$\text{Diff in covariance (DIC)} = Mean_{r,s}(|Cov_t(\theta_{rs}^{(t)}) - Cov_t(\tilde{\theta}_{rs}^{(t)})|),$$

$$\text{Total Variation (TV)} = Mean_{r,s}\left(0.5\int |D_{rs}(x) - \tilde{D}_{rs}(x)|dx\right),$$

$$\text{Bias}^2 = Mean_s\left(\left(Mean_{tr}(\theta_{rs}^{(t)}) - \theta_s^{true}\right)^2\right),$$

$$\text{VAR} = Mean_s(Var_r(Mean_t(\theta_{rs}^{(t)}))),$$

$$\text{MSE} = \text{Bias}^2 + \text{VAR},$$

where r is the replicate and s is the parameter component.

- We account for CPU time using

$$\begin{aligned} \text{ESS} &= Mean_{rs}((M-B)/\text{ACT}_{rs}), \\ \text{ESS/CPU} &= Mean_{rs}((M-B)/\text{ACT}_{rs}/CPU_r), \end{aligned} \tag{3}$$

where $M - B$ is the number of chain iterations.

Numerical Experiments: Ricker's Model

- A particular instance of hidden Markov model:

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

  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 $\mathbf{y}$, $\bar{y}$,

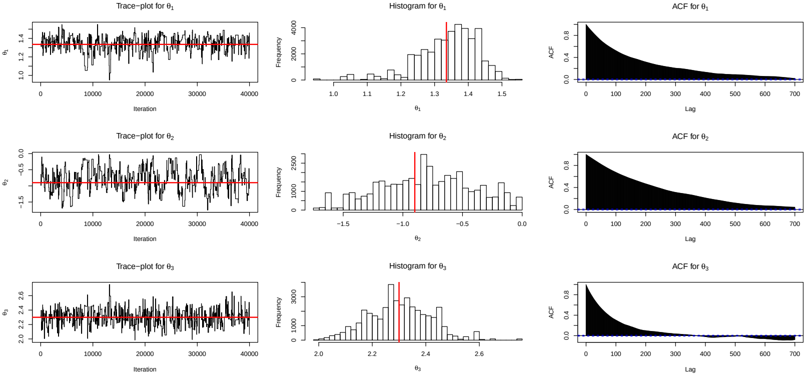
(C3:C7) Sample auto-correlations at lags 1 through 5,

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

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

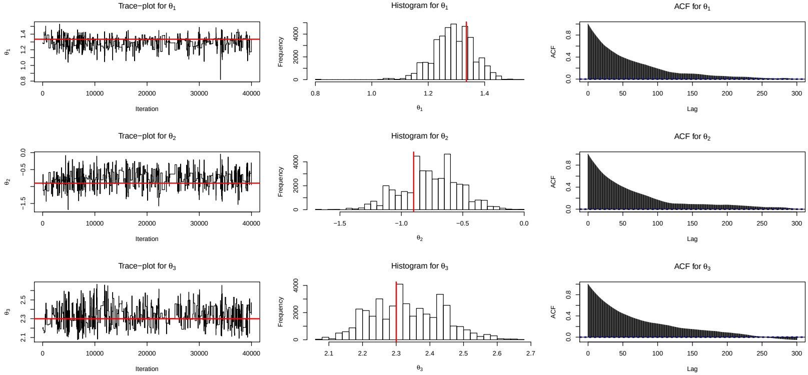
Numerical Experiments: Ricker's Model - ABC/RWM

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.



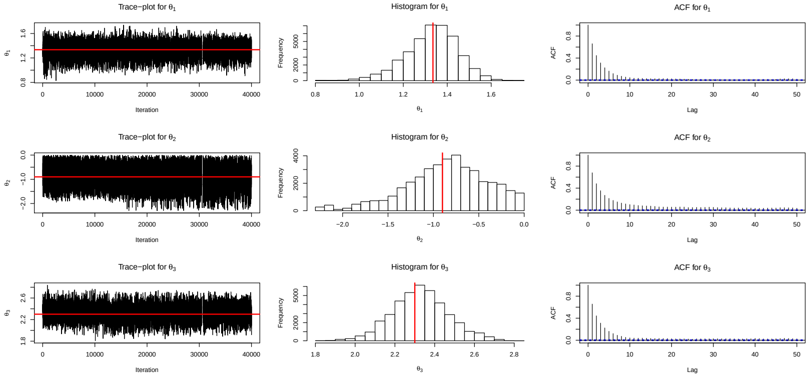
Numerical Experiments: Ricker's Model - BSL

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



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
SMC	0.152	0.0177	0.378	0.086	0.201	0.219	472	0.521
ABC-RW	0.135	0.0201	0.389	0.059	0.180	0.189	87	0.199
ABC-IS	0.139	0.0215	0.485	0.063	0.195	0.205	47	0.099
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
BSL-IS	0.122	0.0082	0.455	0.022	0.197	0.198	33	0.007
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

Example: Stochastic Volatility

Stochastic volatility model with stable errors:

$$x_1 \sim \mathcal{N}(0, 1/(1-\theta_1^2)); \quad v_i \overset{iid}{\sim} \mathcal{N}(0,1); \quad w_i \overset{iid}{\sim} Stab(\theta_4, -1); \quad i = \{1, \cdots$$
$$x_i = \theta_1 x_{i-1} + v_i; \quad i = \{2, \cdots, n\},$$
$$y_i = \sqrt{\exp(\theta_2 + \exp(\theta_3) x_i)} w_i; \quad i = \{1, \cdots, n\}.$$

Here $St(\alpha, \beta)$ is a stable distribution with parameters $\theta_4 \in [0, 2]$ and skew parameter.

Example: Stochastic Volatility

For summary statistics we use a 7-dimensional vector whose components are:

(C1) $\#\{i : y_i^2 > \text{quantile}(\mathbf{y}_0^2, 0.99)\}$,

(C2) Average of $\mathbf{y}^2$,

(C3) Standard deviation of $\mathbf{y}^2$,

(C4) Sum of the first 5 auto-correlations of $\mathbf{y}^2$,

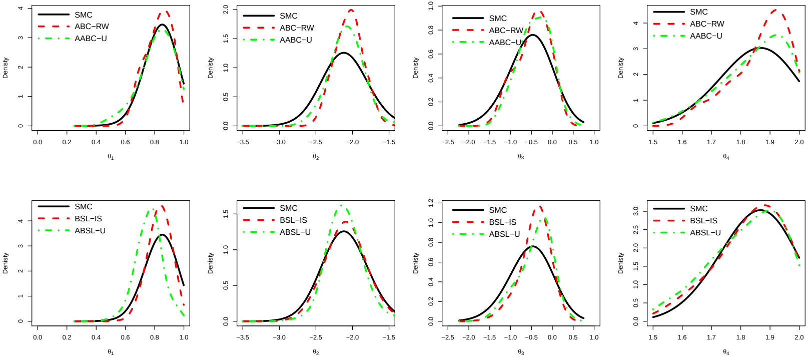
(C5) Sum of the first 5 auto-correlations of $\{\mathbf{1}_{\{y_i^2 < \text{quantile}(\mathbf{y}^2, 0.1)\}}\}_{i=1}^n$,

(C6) Sum of the first 5 auto-correlations of $\{\mathbf{1}_{\{y_i^2 < \text{quantile}(\mathbf{y}^2, 0.5)\}}\}_{i=1}^n$,

(C7) Sum of the first 5 auto-correlations of $\{\mathbf{1}_{\{y_i^2 < \text{quantile}(\mathbf{y}^2, 0.9)\}}\}_{i=1}^n$.

Example: Stochastic Volatility cont..

Figure: First row compares SMC, ABC-RW and AABC-U samplers. Second row compares SMC, BSL-IS and ABSL-U. From left to right: θ_1, θ_2, θ_3 and θ_4.



ESS/CPU shows 400-800% improvements over competitive algorithms (SMC, ABC-RWM)

Example: Stochastic Volatility cont..

	Diff with SMC			Diff with true parameter			Efficiency	
Sampler	DIM	DIC	TV	$\sqrt{\text{Bias}^2}$	$\sqrt{\text{VAR}}$	$\sqrt{\text{MSE}}$	ESS	ESS/CPU
SMC	0.000	0.0000	0.000	0.221	0.201	0.299	468	0.267
ABC-RW	0.078	0.0126	0.205	0.248	0.198	0.317	24	0.069
ABC-IS	0.082	0.0151	0.306	0.232	0.221	0.320	26	0.071
AABC-U	0.069	0.0124	0.170	0.250	0.183	0.310	1303	1.617
AABC-L	0.069	0.0132	0.161	0.246	0.181	0.305	1256	1.546
BSL-RW	0.044	0.0116	0.122	0.225	0.181	0.289	123	0.037
BSL-IS	0.045	0.0103	0.125	0.226	0.177	0.287	285	0.084
ABSL-U	0.063	0.0133	0.228	0.225	0.181	0.289	832	0.735
ABSL-L	0.061	0.0140	0.230	0.236	0.183	0.299	757	0.671

Conclusions

- ABC and BSL are useful when the likelihood is not tractable.
- The computational burden can prohibit the full reach for these approximate methods.
- Currently exploring links between ABC and divide-and-conquer MCMC for large data.

All papers available at:
http://www.utstat.toronto.edu/craiu/Papers/index.html