Bayesian Computation Strategies for Big Data and Intractable Models

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with Evgeny Levi (Toronto)

Babeș-Bolyai, Cluj, February 8, 2023

A Bayesian's Best Friend: MCMC

- Consider observed data $\mathbf{y}_0 \in \mathcal{Y}$, likelihood function $L(\boldsymbol{\theta}|\mathbf{y}_0)$ (or sampling distribution $f(\mathbf{y}_0|\boldsymbol{\theta})$), prior $p(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^d$.
- Inference is based on $\pi(\boldsymbol{\theta}|\mathbf{y}_0) = \frac{f(\mathbf{y}_0|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f(\mathbf{y}_0|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{f(\mathbf{y}_0|\boldsymbol{\theta})p(\boldsymbol{\theta})}{m(\mathbf{y}_0)}$.
- Denominator $m(\mathbf{y}_0)$ is usually intractable so to estimate $I = \int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y}_0)d\boldsymbol{\theta}$ for a given h we need to sample from π.
- MCMC methods rely on constructing and running a Harris recurrent and irreducible Markov chain with state space Θ and stationary distribution $\pi(\boldsymbol{\theta}|\mathbf{y}_0)$.
- MCMC developments have allowed Bayesian statisticians to think freely about a statistical model for 30+ years.

MCMC at the crossroads

- 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)
- Large data and/or intractable likelihoods have brought Bayesian computation at a crossroads.

Massive data set

- $L(\theta|\mathcal{D})$ is computable, but data is massive.
- Possible remedies:
  - 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)

Divide and conquer

- 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.
- Challenge: provide theoretical guarantees or assess approximation errors beyond the Gaussian case.

Example: Consensus Monte Carlo (Scott et al., 2016)

- The batch-specific posterior is defined as $\pi_{k,CMC} \propto [p(\boldsymbol{\theta})]^{\frac{1}{K}} f(\mathbf{y}_0^{(k)}|\boldsymbol{\theta})$ so that

$$\pi(\boldsymbol{\theta}|\mathbf{y}_0) \propto \prod_{k=1}^{K} \pi_{k,CMC}(\boldsymbol{\theta}|\mathbf{y}_0^{(k)}).$$

- MCMC samples $\boldsymbol{\theta}_i^{(h)} \sim \pi_{h,CMC}$ for $i = 1, \ldots, M$ and are combined using a weighted average

$$\boldsymbol{\theta}_i = \frac{\sum_{k=1}^{K} w_k \boldsymbol{\theta}_i^{(k)}}{\sum_{k=1}^{K} w_k}$$

  where $w_k = Var(\boldsymbol{\theta}|\mathbf{y}_0(k))$.
- Theory works if the posteriors are Gaussian.
- Entezari, C. and Rosenthal (2018) $\rightarrow$ CMC for BART.

ABC and BSL

- 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

- Groundwater studies (Cui et al. 2018)

  "A critical issue that limits the application of Bayesian inference is the difficulty to define an explicit likelihood function for complex and non-linear groundwater models"

- Hurrican surge (Plumlee et al., 2021)

  "Storm surge is simulated by solving a set of partial differential equations known as the shallow water equations to yield water elevation and velocity in space and time [...]. A mesh of nodes, which are points in geographic space, is constructed to capture the shape of the seafloor and overland topography. The partial differential equations are then solved on the mesh and integrated forward in time over several days for a single storm simulation."

- Most of methods that address the challenge of large data cannot be used directly for intractable models.

A remarkable algorithm- ABC

- ABC:
  - Sample $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and $\mathbf{y} \sim f(\mathbf{y}|\boldsymbol{\theta})$;
  - Compute distance in terms of a user-defined statistics $\mathbf{S}$

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

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

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

- If $\epsilon = 0$ and $\mathbf{S}$ is a sufficient statistics then $\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}_0) = \pi(\boldsymbol{\theta}|\mathbf{y}_0)$ so ABC is exact.

Zooming in on the target

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

$$\pi_\epsilon(\boldsymbol{\theta}|\mathbf{y}_0) = \int_{\mathcal{Y}} \pi_\epsilon(\boldsymbol{\theta}, \mathbf{y}|\mathbf{y}_0)d\mathbf{y} \propto$$

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

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

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

  (L) The *linear* kNN with $W_{Nn}(\zeta^*) = W(\|\tilde{\zeta}_n - \zeta^*\|) = 1 - \|\tilde{\zeta}_n - \zeta^*\| / \|\tilde{\zeta}_K - \zeta^*\|$ for $n \leq K$ 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 (Gourieroux et al. 1993; Smith Jr 1993)
- 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 $\mathbf{y}_1, \ldots, \mathbf{y}_K$ 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 $\mathbf{S}(\mathbf{y})$ given θ is Gaussian with parameters $\phi(\boldsymbol{\theta}) = (\mu_\theta, \Sigma_\theta)$
- The Synthetic Likelihood (SL) procedure assigns to each θ the likelihood $\text{SL}(\theta|s_0) = \mathcal{N}(s_0; \mu_\theta, \Sigma_\theta)$, where $s_0 = \mathbf{S}(\mathbf{y}_0)$.
- The BSL posterior is $\pi(\theta|s_0) \propto p(\theta)\text{SL}(\boldsymbol{\theta}|s_0)$.
- To estimate μ_θ and Σ_θ we can use m realizations of $\mathbf{S}$, $\mathbf{S}(\mathbf{y}_1), \cdots, \mathbf{S}(\mathbf{y}_m)$, where $\mathbf{y}_1, \ldots, \mathbf{y}_m \overset{iid}{\sim} f(y|\boldsymbol{\theta})$

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

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

- Acceptance probability requires repeated estimation of (2)

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

A different POV: Precomputation

- Given a proposal q, precompute $\mathcal{Z} = \{(\xi_h, s_h = (s_h^{(1)}, \ldots, s_h^{(m)})^T) : 1 \leq h \leq H\}$ where $\xi_h \sim q$, $\mathbf{w}_h^{(1)}, \ldots, \mathbf{w}_h^{(m)} \overset{iid}{\sim} f(\mathbf{w}|\xi_h)$ and set $s_h^{(j)} = S(\mathbf{w}_h^{(j)})$ for all $1 \leq j \leq m$.
- Given a proposal $\boldsymbol{\theta}^*$ at t-th iteration

$$\tilde{\mu}(\boldsymbol{\theta}^*) = \frac{\sum_{h=1}^{H}[W_h(\boldsymbol{\theta}^*)\frac{1}{m}\sum_{j=1}^{m} s_h^{(j)}]}{\sum_{h=1}^{H} W_h(\boldsymbol{\theta}^*)},$$

$$\tilde{\Sigma}(\boldsymbol{\theta}^*) = \frac{\sum_{h=1}^{H}[W_h(\boldsymbol{\theta}^*)\frac{1}{m}\sum_{j=1}^{m}(s_h^{(j)} - \hat{\mu}_{\boldsymbol{\theta}^*})(s_h^{(j)} - \hat{\mu}_{\boldsymbol{\theta}^*})^T]}{\sum_{h=1}^{H} W_h(\boldsymbol{\theta}^*)}. \tag{3}$$

- We use $m = 1$.

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 computed exactly.
- The stationary distribution of a chain with kernel $P(\cdot, \cdot)$ is 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: 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 $\mathbf{y}$, $\bar{y}$,

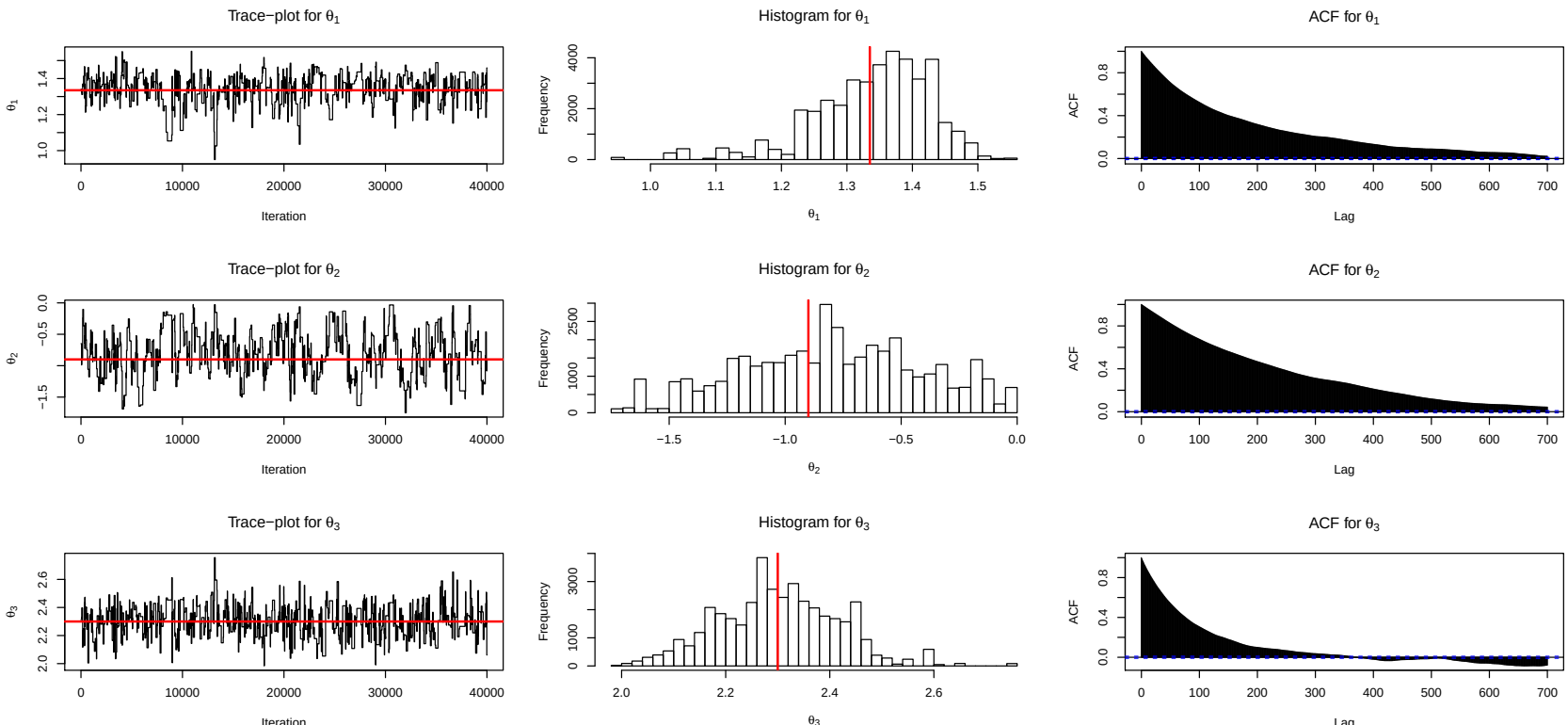
(C3:C7) Sample auto-correlations at lags 1 through 5,

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

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

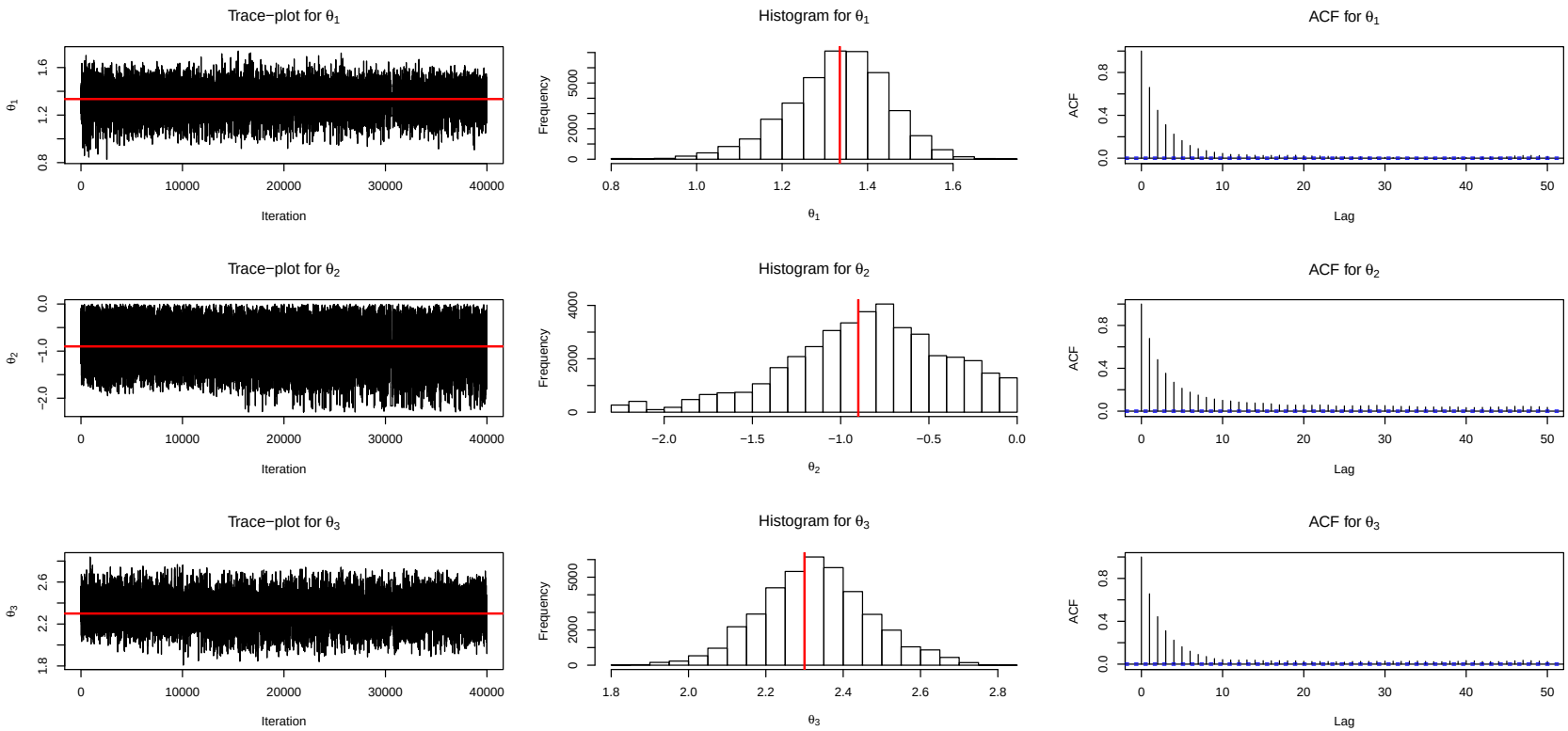
Numerical Experiments: Ricker's Model - ABC/RWM

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 - 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

- Our methods show good results even if $q(\xi|\theta) = \mathcal{N}(\theta, \Sigma)$ but theory is not fully developed.
- Ideally we want to combine with adaptive MCMC.
- The computational burden can prohibit the full reach for these approximate methods so more solutions are needed.

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