The Universe of Sampling: Notes from a Statistical Odyssey

Radu Craiu

Department of Statistical Sciences
University of Toronto

Toronto, July 2026

Outline

- Josh's invite : "A review of MCMC" = Tall order
- Many existing excellent reviews
- Will focus instead on a few topics of interest
  - Unbiased MCMC via coupling
  - Noisy MCMC - with many impersonations
  - Sampling for models with intractable likelihoods
  - Cameos: control variates, adaptive MCMC, optimal coupling.

Sampling, what is it good for?

- How does my model perform in a new generative scenario?

Consider binary response $Y \in \{0, 1\}$ and covariates $x = (x_1, x_2, x_3) \in \mathbb{R}^3$

Of interest is simulating data that is "similar" to the observed, i.e. the number of cases and controls is the same

Goal: sample the covariates given an observed response, $x \mid Y = y$.

Suppose $x \sim \mathcal{N}_3(0, \Sigma)$, by Bayes' rule,

$$\pi(x \mid Y = y) \propto \underbrace{\Pr(Y = y \mid x)}_{\text{logistic}} \underbrace{\phi(x)}_{\text{Gaussian}}, \qquad \pi(x \mid Y = 1) \propto \frac{e^{-x^\top \Sigma^{-1} x/2}}{1 + e^{-(\beta_0 + \beta^\top x)}}.$$

Sampling, what is it good for?

- Compute the integral of a function, e.g. compute the marginal likelihood?

For a cluster with binary outcomes $y_1, \ldots, y_n$ and covariates x_i, consider the unobserved random effect $u \sim \mathcal{N}(0, \sigma^2)$:

$$\Pr(y_i = 1 \mid u) = \sigma(\beta_0 + \beta^\top x_i + u), \qquad \sigma(t) = \frac{1}{1 + e^{-t}}.$$

The marginal likelihood is obtained by integrating out the latent u:

$$L(\theta) = \int_{\mathbb{R}} \left[\prod_{i=1}^{n} \Pr(y_i \mid u, \theta) \right] \phi(u; 0, \sigma^2)\, du, \qquad \theta = (\beta_0, \beta, \sigma^2).$$

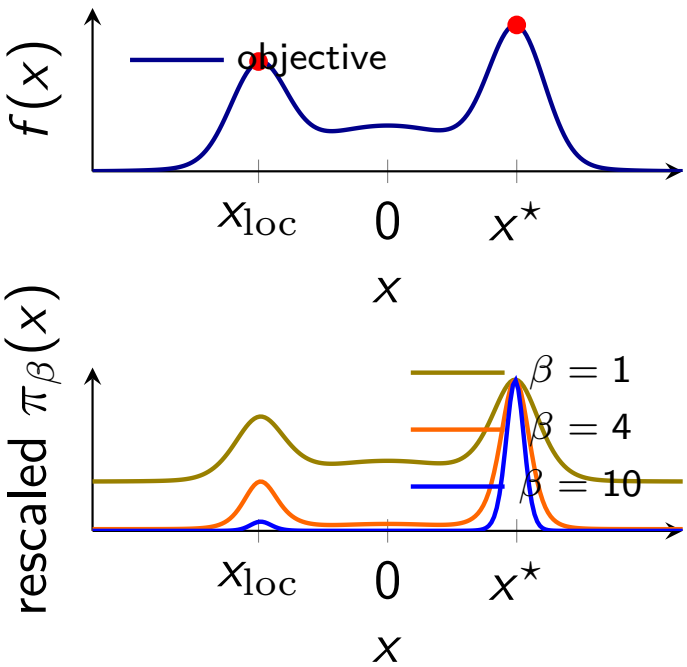
$\Rightarrow$ MCMC is needed: sample $u^{(1)}, \ldots, u^{(M)} \sim p(u \mid y, \theta)$ to form Monte Carlo estimates of $L(\theta)$ and its gradient — the basis of Monte Carlo EM and MCMC maximum likelihood (Geyer and Thompson, 1992).

Sampling for Optimization

- Find the global maximum of a function f, e.g. the MLE
- Maximizing f can be seen through a sampling lens: build a distribution that puts more mass where f is large,

$$\pi_\beta(x) \propto \exp\left(\beta f(x)\right).$$

- As β grows, samples from π_β concentrate near the maximizers of f — sampling becomes a stochastic search for optimization.



- Things work the other way around too. Sampling can be cast as an optimization problem - see variational approximation for Bayesian posteriors (Blei et al., 2017)

The sampling problem

- Of interest is $I = \int f(x)\pi(x)dx$ where $f(x)$ is a function and $\pi(x)$ is a density, $x \in \mathbb{R}^d$
- The (Vanilla) Monte Carlo solution relies on $x_1, \ldots, x_m$ iid from $\pi(x)$ to build $\widehat{I_m} = \frac{1}{m}\sum_{i=1}^{m} f(x_i)$.
- The accuracy and the cost of the approximation depends on the number of samples m, e.g. $\text{Var}(\widehat{I_m}) = O(m^{-1})$

Markov chain Monte Carlo

- Vanilla Monte Carlo is often not possible (π is too complicated, not fully known, not known at all, etc)
- Instead, we construct a Markov chain (discrete time, continuous or discrete state space) which has π as its stationary distribution

$$P(x, A) = \Pr(X_t \in A \mid X_{t-1} = x), \qquad x \in \mathcal{X},\ A \subseteq \mathcal{X},$$

so, started at $X_0 \sim \nu_0$, the t-th state has law

$$X_t \sim \nu_0 P^t, \qquad (\nu_0 P^t)(A) = \int_{\mathcal{X}} P^t(x, A)\, \nu_0(dx).$$

- Ideally, after a burn-in period the states visited by the chain are dependent samples approximately distributed with π

$$X_{t-1} \sim \pi \implies X_t \sim \pi, \qquad \text{i.e.} \qquad \pi(A) = \int_{\mathcal{X}} P(x, A)\, \pi(dx)\ \forall A \quad (\pi P = \pi),$$

Markov chain Monte Carlo

- We trade independence for feasibility. In the process we acquire wider applicability and several headaches
- Some headaches include:
  - serial dependence between draws can be high so we explore poorly/slowly the sample space
  - sometimes hard to determine when the chain is in (or close to) the stationary regime
  - compute the variance of the MCMC estimators, i.e. $\mathrm{Var}(\widehat{\mathrm{I}})$.
  - improve the design, tune the parameters, etc

One slide on Metropolis-Hastings

- At state $X_t = x$, propose $Y \sim q(\cdot \mid x)$ and accept it with probability

$$\alpha(x, Y) = \min\left\{1, \frac{\pi(Y)q(x \mid Y)}{\pi(x)q(Y \mid x)}\right\}.$$

- If accepted, set $X_{t+1} = Y$; otherwise set $X_{t+1} = x$.
- For $\widehat{\mathrm{I}}_M = M^{-1}\sum_{t=1}^{M} f(X_t)$,

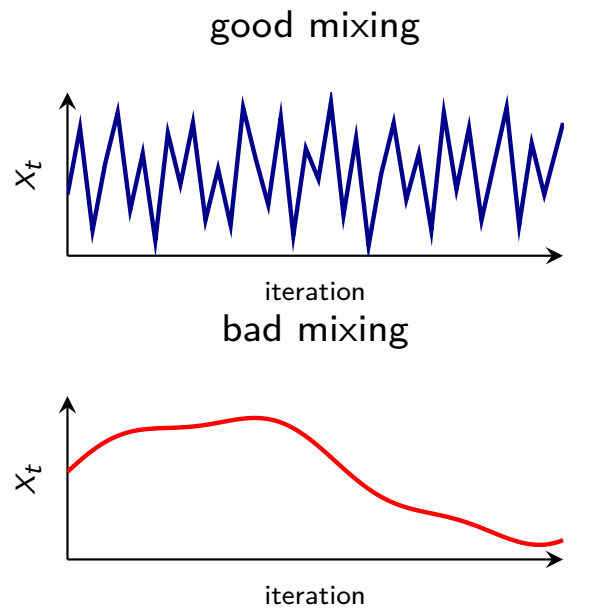
$$\mathrm{Var}(\widehat{\mathrm{I}}_M) \approx \frac{\sigma_f^2}{M}\left(1 + 2\sum_{k\geq 1}\rho_k\right),$$

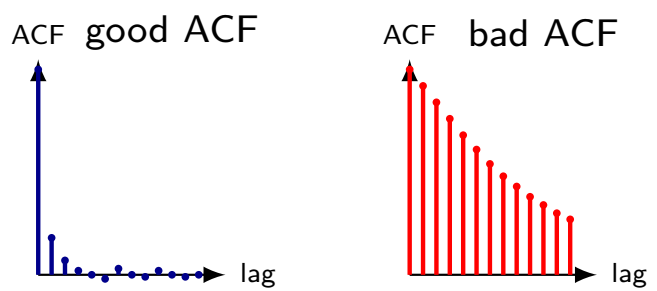
where $\sigma_f^2 = \mathrm{Var}_\pi\{f(X)\}$ and ρ_k is the lag-k autocorrelation of $f(X_t)$.

- The effective sample size is

$$ESS = \frac{M}{1 + 2\sum_{k\geq 1}\rho_k}.$$

Thus $\mathrm{Var}(\widehat{\mathrm{I}}_M) \approx \sigma_f^2/ESS$.

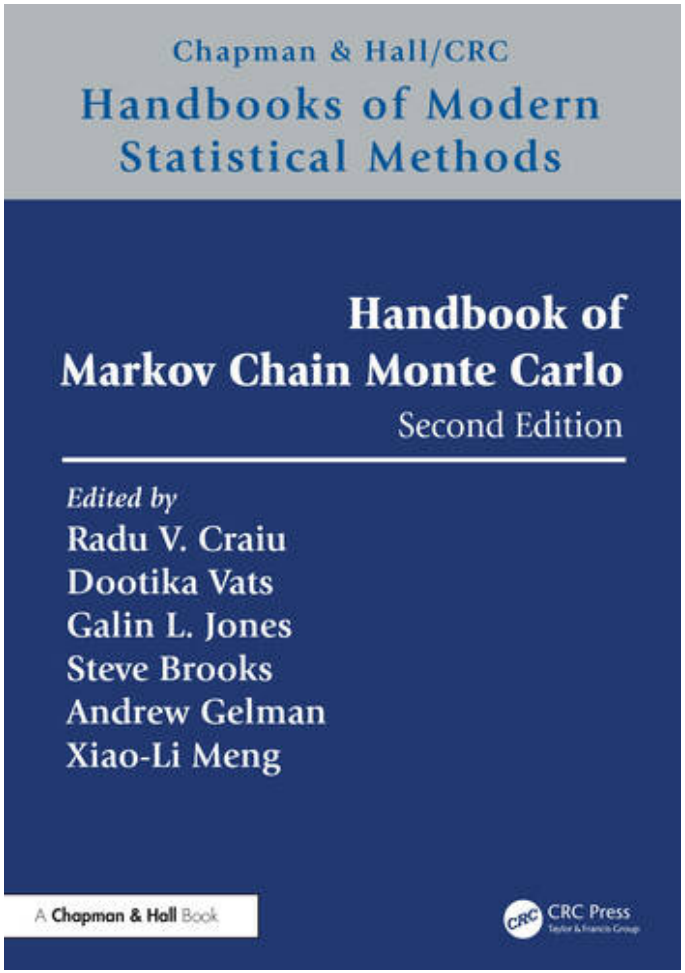




Slow exploration means large positive autocorrelations, smaller M_{eff}, and larger Monte Carlo variance.

Hydration break

An advertisement



See also: `https://github.com/dvats/HandbookOfMCMC`

Some topics of interest

- There are several reviews of basic MCMC available, including Andrieu et al. (2003); Craiu and Rosenthal (2014), and Chapter 1 in the Handbook.
- Hamiltonian Monte Carlo (HMC) has emerged as a powerful tool for sampling densities in high dimensions (many examples and reviews available, see also Chapter 2 in Handbook)
- We will discuss:
  - Coupling and unbiasedness
  - Sampling from intractable densities
  - Noisy and approximate MCMC
  - Recent AI-based samplers for intractable distributions

Unbiased MCMC

Unbiased MCMC

- Assume interest in approximating $\mathsf{I} = \mathbb{E}_\pi[f(X)]$ using $\widehat{\mathsf{I}} = \frac{1}{M}\sum_{t=B}^{M+B} f(X_t)$, where $\{X_t\}_{t\geq 0}$ are MCMC samples from some posterior π
- $\widehat{\mathsf{I}}$ is vulnerable to potential biases due to:
  - Insufficient burn-in B
  - Poor chain Initialization
- These biases can accumulate when one is approximating the expectation I repeatedly.
- Naghib et al. (2019) develop a framework for telescope scheduling and their approach can be generalized to other applications that rely on sequential decision-making under uncertainty.

Recursive expectations amplify the bias

- In sequential decision making (e.g. telescope scheduling) the value of a state solves a backward recursion: for $k = N-1, \ldots, 0$,

$$V_k(s) = \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \beta\, \mathbb{E}_\pi \left[h(\theta, a, V_{k+1}) \right] \right\},$$

  with intractable posterior $\pi(\theta) = p(\theta \mid D)$ and

$$h(\theta, a, V_{k+1}) = \sum_{s'} P_\theta(s' \mid s, a)\, V_{k+1}(s').$$

- Each expectation is approximated with MCMC draws $\theta_1, \ldots, \theta_L$:

$$\widehat{h}_L = \frac{1}{L} \sum_{\ell=1}^{L} h(\theta_\ell, a, V_{k+1}), \qquad \mathbb{E}[\widehat{h}_L] \neq \mathbb{E}_\pi[h] \ \text{(finite burn-in)}.$$

- The recursion then applies a biased operator $\widetilde{T} \neq T$, and the biased $\widehat{V}_{k+1}$ feeds into stage k: the per-stage error is compounded through all N steps.

$\Rightarrow$ We need each conditional expectation to be unbiased, so the recursion injects no error.

A Coupling-based Solution

- Consider two chains $\mathcal{X} = \{X_t, t \geq 0\}$ and $\mathcal{Y} = \{Y_t, t \geq 0\}$ are coupled in such a way that:
  - They have the same initial distribution and transition kernel
  - With probability one there exists a finite stopping time τ such that $X_t = Y_{t-1}$ for all $t \geq \tau$.
- Then $H_k(\mathcal{X}, \mathcal{Y}) = f(X_k) + \sum_{j=k+1}^{\tau-1}[f(X_j) - f(Y_{j-1})]$ has mean

$$\mathbb{E}\left[H_k(\mathcal{X}, \mathcal{Y})\right] = \mathbb{E}\left\{f(X_k) + \sum_{j=k+1}^{\infty}\left[f(X_j) - f(Y_{j-1})\right]\right\}$$

$$= \mathbb{E}\left\{f(X_k) + \sum_{j=k+1}^{\infty}\left[f(X_j) - f(X_{j-1})\right]\right\} = I$$

- So $H_k(\mathcal{X}, \mathcal{Y})$ is unbiased for I for any $k \geq 0$ (see Chapter 17 in the Handbook)

From Unbiasedness to Convergence criteria

- Generalize to a general "lag" L, i.e. find τ such that $X_t = Y_{t-L}$ for all $t \geq \tau$
- $H_{k,L}(\mathcal{X}, Y) = f(X_k) + \sum_{j=1}^{J_{k,L}} \left[f(X_{k+jL}) - f(Y_{k+(j-1)L}) \right]$ is unbiased for I, where $J_{k,L} = \max\left\{0, \lceil \frac{\tau_L - L - k}{L} \rceil \right\}$.
- Biswas et al. (2019) have shown:

$$d_{\mathrm{TV}}(\pi_k, \pi) \leq \mathbb{E}[J_{k,L}]$$

Variance reduction via control variates

- For our purpose it is useful to express $H_{k,L}$ in the equivalent form

$$H_{k,L}(\mathcal{X}, Y) = f(X_{k+LJ_{k,L}}) + \sum_{j=0}^{J_{k,L}-1} \left[f(X_{k+jL}) - f(Y_{k+jL}) \right].$$

- Craiu and Meng (2022) noted that $\Delta_{k,j} := f(X_{k+jL}) - f(Y_{k+jL})$ has mean zero so that

$$\tilde{H}_{k,L}(\mathcal{X}, Y) = f(X_{k+LJ_{k,L}}) + \sum_{j=0}^{J_{k,L}-1} \left[f(X_{k+jL}) - f(Y_{k+jL}) \right] - C_\eta$$

  is still unbiased when $C_\eta = \sum_{j \geq 1} \eta_j \Delta_{k,j}$

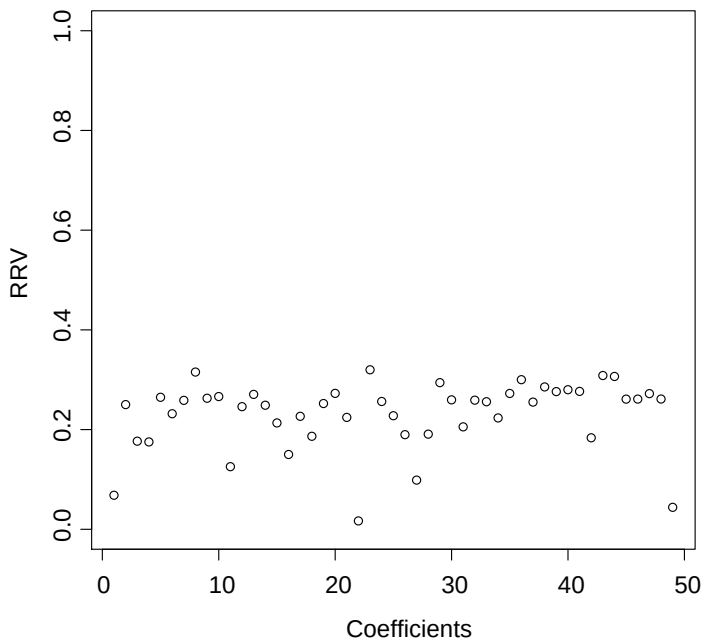
- This leads to tighter bounds on the TV distance.

Control Variates: German Credit Data

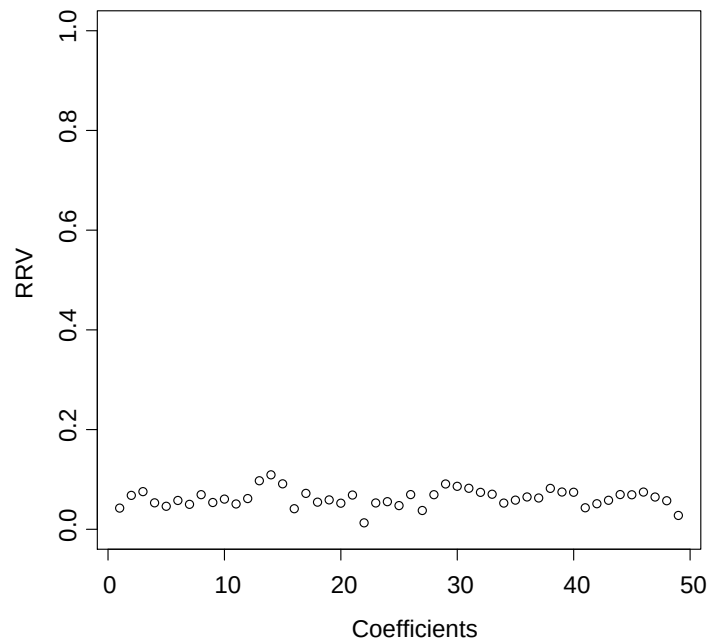
- Bayesian logistic regression model for the German Credit data.
- Data consist of $n = 1000$ binary responses and $d = 49$ covariates.
- The relative reduction in variance (RRV) computed as $\text{RRV} = \frac{\text{var}_{MCCV}(\hat{\beta})}{\text{var}_{MC}(\hat{\beta})}$ where $\hat{\beta}$ is the posterior mean of the regression coefficients, $\beta \in \mathbb{R}^{49}$ and Var_{MC}, Var_{MCCV} denote the estimated Monte Carlo variances of $\hat{\beta}$ obtained without and, respectively, with control variates.

Control Variates: German Credit Data

$L = 5$



$L = 20$



Noisy MCMC

Noisy MCMC

- Idea: replace the transition kernel of an ideal but unrealistic MCMC algorithm with a practical one.
- Consequence: The stationary distribution of the modified chain is no longer π (the one we want) but $\tilde{\pi}$.
- Rule of the Game: Make sure that the approximating kernel doesn't lead to a $\tilde{\pi}$ that is very different from π.

A success story: Pseudo-Marginal Metropolis-Hastings

- Assume there exists a computable $\widehat{L}(\theta, U) \geq 0$, such that

$$\mathbb{E}\{\widehat{L}(\theta, U)\} = L(\theta),$$

where U denotes the random numbers or auxiliary variables used in the estimate.

- Run MH on the extended state (θ, U) with proposal

$$\theta^\star \sim q(\cdot \mid \theta), \qquad U^\star \sim m(\cdot \mid \theta^\star),$$

and accept with probability $\min\{1, A\}$, where

$$A = \frac{\widehat{L}(\theta_{new}, U_{new}) p(\theta_{new}) q(\theta_{old} \mid \theta_{new})}{\widehat{L}(\theta_{old}, U) p(\theta_{old}) q(\theta_{new} \mid \theta_{old})}.$$

- The marginal stationary distribution of θ is the exact posterior $\pi(\theta \mid y) \propto L(\theta) p(\theta)$, despite using a random likelihood estimate (Andrieu and Roberts, 2009).

Example: von Mises regression for directional data

- Directional/angular data (wind directions, orientations, phase angles) live on the circle $(-\pi, \pi]$. A von Mises regression links each angle to covariates z_i:

  $$X_i \mid \mu_i, \kappa \sim \mathrm{vM}(\mu_i, \kappa), \quad \mu_i = 2\arctan(\beta^\top z_i), \quad \beta \sim \mathcal{N}(0, 10^2 I), \ \log(\kappa) \sim \mathcal{N}(0, 10^2).$$

- von Mises density (mean direction μ, concentration κ; I_0 = modified Bessel function of order 0):

  $$f(x \mid \mu, \kappa) = \frac{\exp(\kappa \cos(x - \mu))}{2\pi I_0(\kappa)}, \qquad -\pi < x < \pi.$$

- With $\theta = (\beta, \log(\kappa))^\top$ the per-observation log-likelihood is

  $$\log(f(X_i \mid \theta)) = \kappa \cos\Big(X_i - 2\arctan(\beta^\top z_i)\Big) - \log(2\pi) - \log(I_0(\kappa)).$$

- Setup: $n = 10{,}000$ angles, 10 covariates — a tall-data target where recomputing $\sum_{i=1}^{n} \log(f(X_i \mid \theta))$ at every step is the bottleneck (Quiroz et al., 2019).

How the control variates are obtained and used

- Draw a subsample S, $|S| = k \ll n$, and estimate the log-likelihood (a sum) using one control variate $q(X_i \mid \theta)$ per point:

$$\widehat{\ell}(\theta) = q(\theta) + \frac{n}{|S|} \sum_{i \in S} \Big(\log\left(f(X_i \mid \theta)\right) - q(X_i \mid \theta) \Big), \qquad q(\theta) = \sum_{i=1}^{n} q(X_i \mid \theta),$$

  which is unbiased for $\sum_i \log\left(f(X_i \mid \theta)\right)$ for any choice of q.

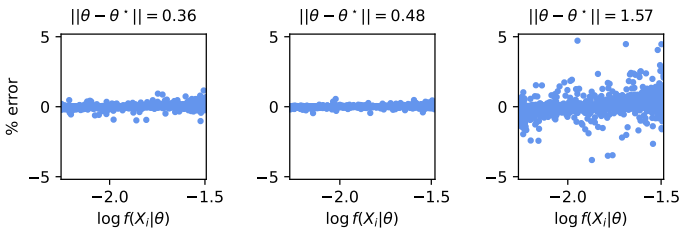
- How q is obtained: take the second-order Taylor expansion of $\log\left(f(X_i \mid \theta)\right)$ about the posterior mode $\theta^\star$ (Bardenet et al., 2017). Then $q(\theta) = \sum_i q(X_i \mid \theta)$ is a fixed low-order polynomial in θ whose coefficients are precomputed data summaries (e.g. sums of z_i, $z_i z_i^\top$), so $q(\theta)$ costs $O(1)$ per step and the estimator costs $O(|S|)$, not $O(n)$.

- Why it works: the estimator variance is $\sigma^2_{\widehat{\ell}} = \dfrac{n^2}{|S|}\, \sigma_d^2$, where σ_d^2 is the population variance of the residuals $d_i(\theta) = \log\left(f(X_i \mid \theta)\right) - q(X_i \mid \theta)$. A good q homogenizes the d_i, shrinking σ_d^2 and hence $\sigma^2_{\widehat{\ell}}$.

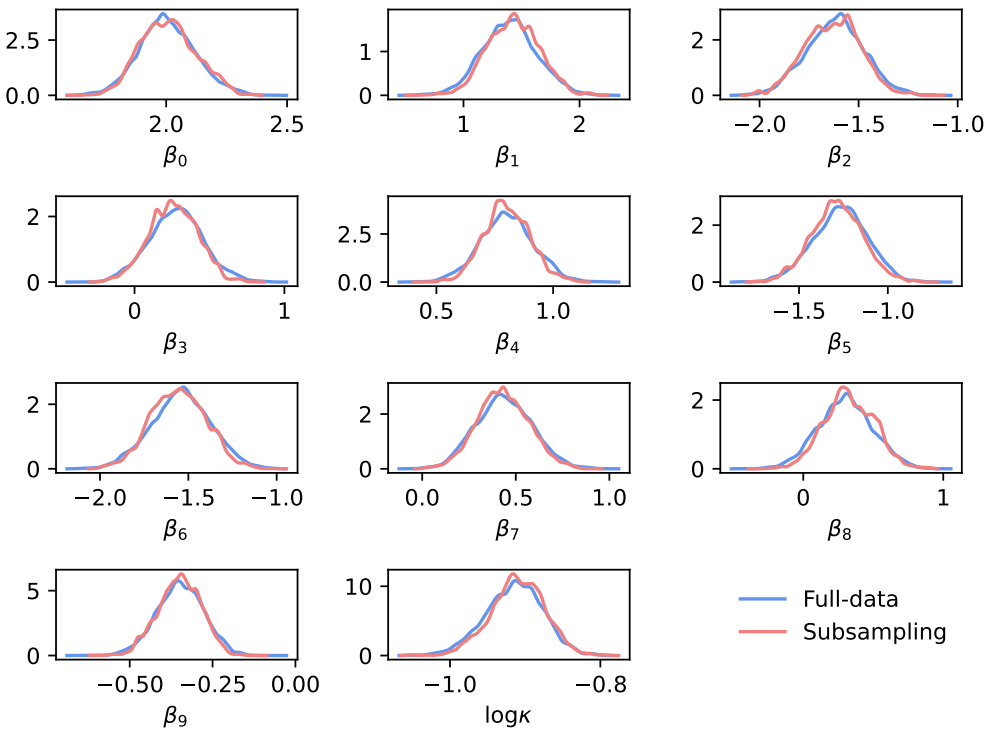
- How $\widehat{\ell}$ is used: feed the bias-corrected likelihood $\exp\left(\widehat{\ell} - \frac{1}{2}\widehat{\sigma}^2_{\widehat{\ell}}\right)$ into pseudo-marginal MH (Ceperley and Dewing, 1999); tune $|S|$ so $\sigma^2_{\widehat{\ell}} \approx 1$ to keep the chain from sticking.

Accurate control variates: subsampling succeeds

Second-order Taylor control variates around $\theta^\star$: the residual error stays small even far from the mode, so $\sigma^2_{\widehat{\ell}} \approx 1$ (here $|S| = 100$) and the subsampling posterior matches the full-data posterior.



% error of q vs. $\log\left(f(X_i \mid \theta)\right)$ — small

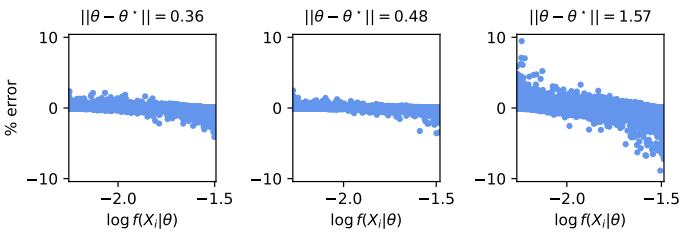


marginal posteriors: full-data vs. subsampling coincide

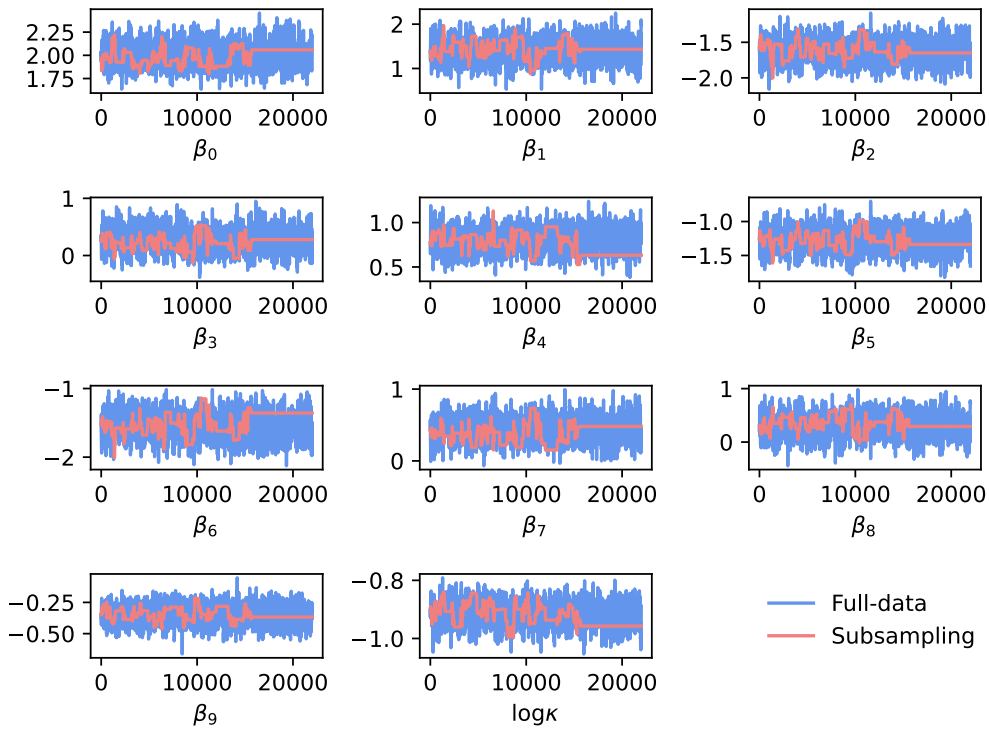
Relative efficiency 45–68 vs. full-data MCMC. Figures from Chapter 20 of the Handbook; method of Quiroz et al. (2019).

Inaccurate control variates: the chain gets stuck

First-order Taylor control variates: the residual error grows with $\|\theta - \theta^\star\|$, so $\sigma^2_{\widehat{\ell}}$ is large (mean ≈ 11) and the pseudo-marginal chain sticks — at the same subsample size.



% error of q — systematic and larger



traces: subsampling (red) is sticky vs. full-data (blue)

The accuracy of the control variate is important for the subsampling to work. Figures from Chapter 20 of the Handbook.

Generative models: ABC

- 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"

- Hurricane 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.
- Having a hard computation problem is like a toothache....

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

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

- 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. (2017) make the case of using $J = 1$.
- Levi and Craiu (2022) 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

- Idea: If the parameter set is discrete a proposed ζ^* will appear in the history along with the corresponding distances, δ.
- This would give us "computation-free" replications!
- We don't have to restrict only to samples that were accepted.
- When the parameter space is continuous instead of looking for ties, we look for past proposals that are not too far from ζ^*, and their corresponding distances

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 $W : \mathbb{R} \to [0, \infty)$ 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, consistency may reduce variability.
- When K is large, consistency may reduce costs.

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 π_ϵ
- 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

Under some mild regularity conditions, then

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

where ν be any probability measure on $(\Theta, \mathcal{F}_0)$.

Numerical Experiments: Ricker's Model

- A particular instance of hidden Markov model:

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

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

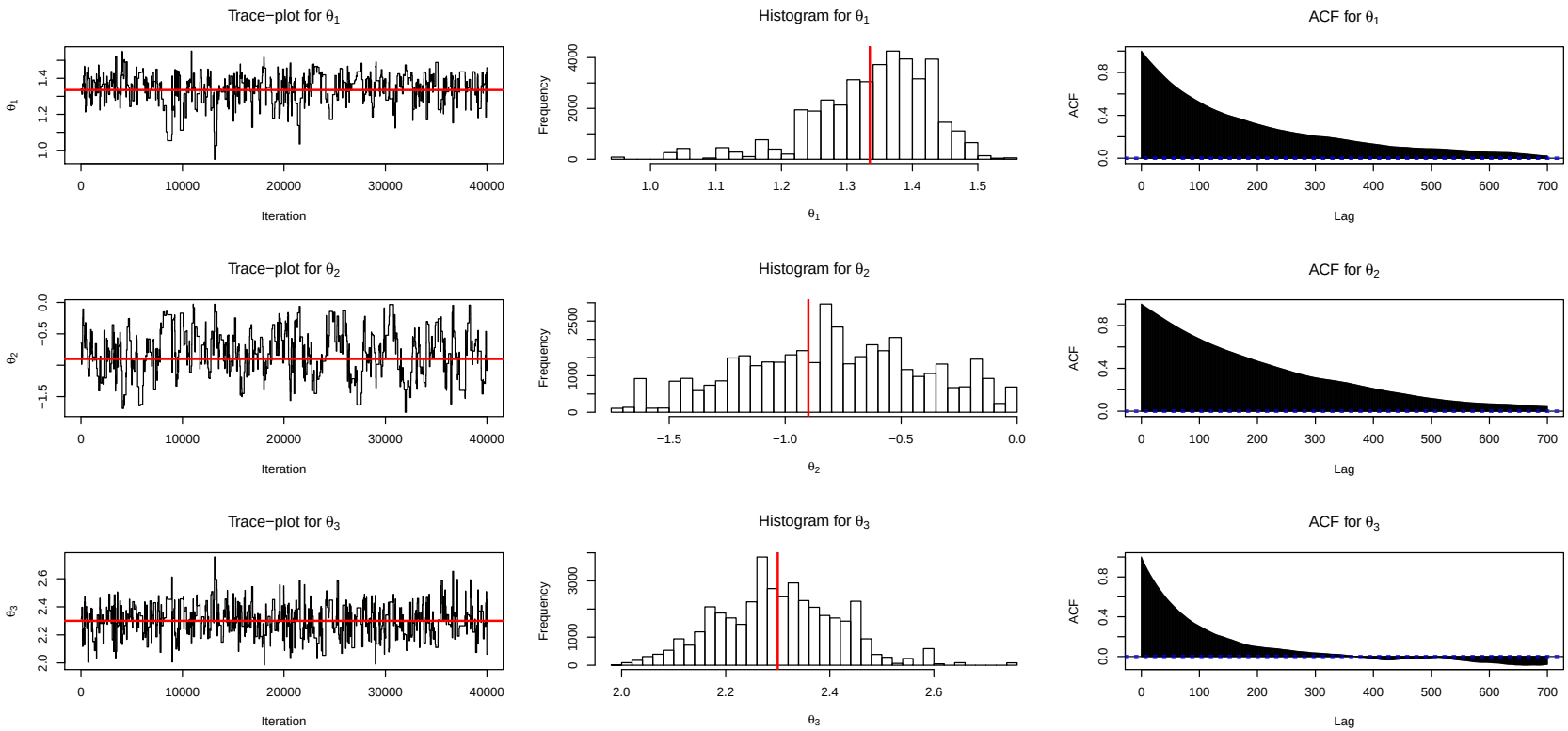
(C3:C7) Sample auto-correlations at lags 1 through 5,

(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
$y_i^{0.3} = \beta_0 + \beta_1 y_{i-1}^{0.3} + \beta_2 y_{i-1}^{0.6} + \epsilon_i,\ i = 2, \ldots, n.$

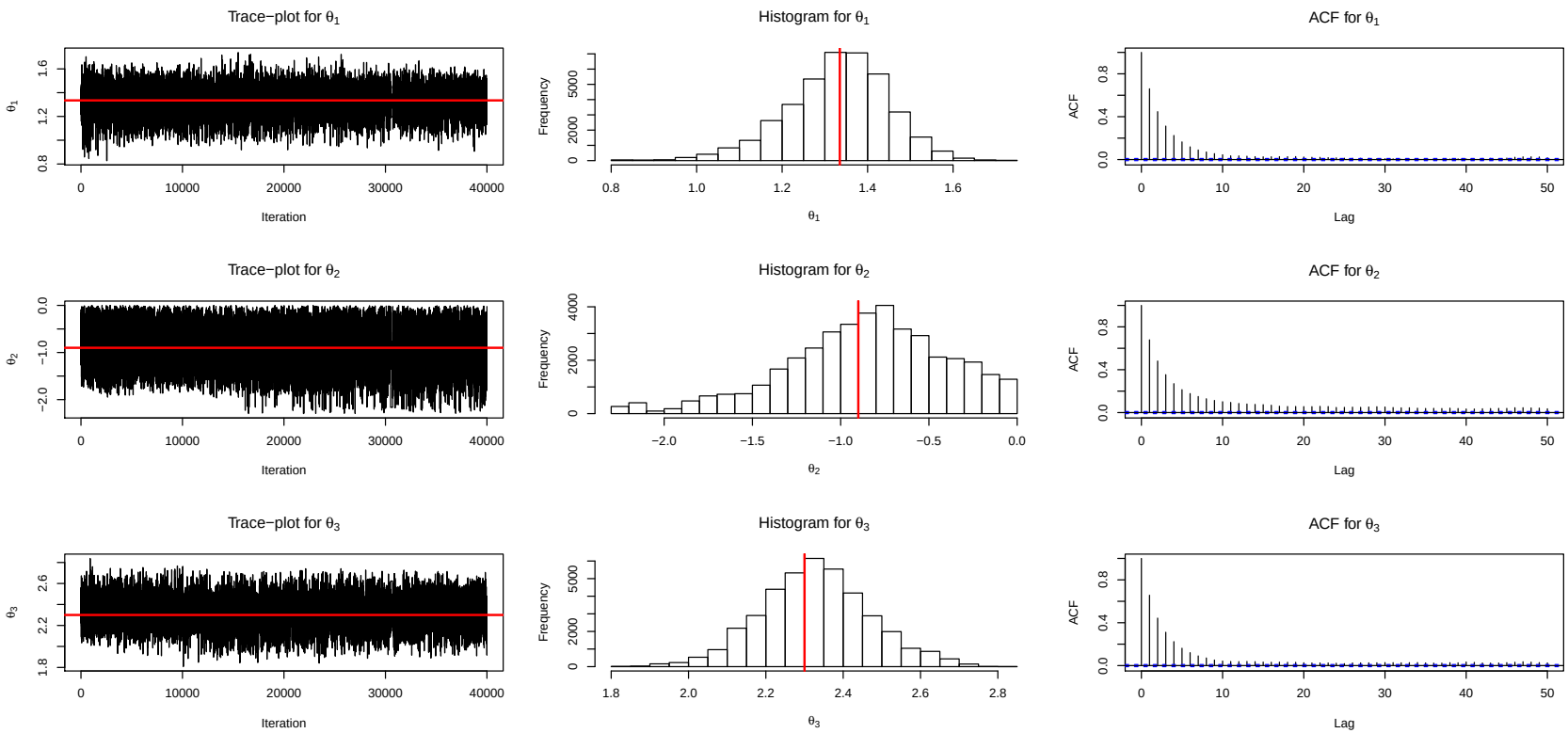
Numerical Experiments: Ricker's Model - ABC/RWM

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 - 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

Final Comments

- Modern MCMC is less about one universal sampler and more about controlled compromises.
- Approximate MCMC methods trade exact likelihoods for simulation, summaries, or synthetic models.
- Coupling methods trade extra simulation for bias removal, diagnostics, and computable error bounds.

References

Andrieu, C., de Freitas, N., Doucet, A. and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning* **50** 5–43.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* **37** 697–725.

Bardenet, R., Doucet, A. and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research* **18** 1–43.

Biswas, N., Jacob, P. E. and Vanetti, P. (2019). Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association* **112** 859–877.

Bornn, L., Pillai, N. S., Smith, A. and Woodard, D. (2017). The use of a single pseudo-sample in approximate Bayesian computation. *Statistics and Computing* **27** 583–590.

Ceperley, D. M. and Dewing, M. (1999). The penalty method for random walks with uncertain energies. *The Journal of Chemical Physics* **110** 9812–9820.

Craiu, R. V. and Meng, X.-L. (2022). Double happiness: Enhancing the coupled gains of L-lag coupling via control variates. *Statistica Sinica* **32** 1745–1766.

Craiu, R. V. and Rosenthal, J. S. (2014). Bayesian computation via markov chain monte carlo. *Annual Review of Statistics and Its Application* **1** 179–201.

Cui, T., Peeters, L., Pagendam, D., Gilfedder, M. et al. (2018). Emulator-enabled approximate Bayesian computation (ABC) and uncertainty analysis for computationally expensive groundwater models. *Journal of Hydrology* **564** 191–207.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B* **54** 657–699.

Lee, A., Andrieu, C. and Doucet, A. (2012). Automatic tuning of pseudo-marginal MCMC-ABC kernels. *Journal of the Royal Statistical Society, Series B* **74** 419–474. Discussion of Fearnhead and Prangle.

Levi, E. and Craiu, R. V. (2022). Finding our way in the dark: Approximate MCMC for approximate Bayesian methods. *Bayesian Analysis* **17** 193 –221.

Naghib, E., Yoachim, P., Vanderbei, R. J., Connolly, A. J. and Jones, R. L. (2019). A framework for telescope schedulers: with applications to the large synoptic survey telescope. *The Astronomical Journal* **157** 151.

Plumlee, M., Asher, T. G., Chang, W. and Bilskie, M. V. (2021). High-fidelity hurricane surge forecasting using emulation and sequential experiments. *The Annals of Applied Statistics* **15** 460–480.

Quiroz, M., Kohn, R., Villani, M. and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association* **114** 831–843.

Wilkinson, R. D. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology* **12** 129–141.