Statistics^2 and Bayesian Computation

Radu Craiu

Department of Statistical Sciences
University of Toronto

Fast and Curious MCMC
May 19, 2023

Computation $\overset{\heartsuit}{\rightarrow}$ Statistics

- Bayesian computation has, to a large extent, freed the statistical modeller.

Computation $\overset{\heartsuit}{\rightarrow}$ Statistics

- Bayesian computation has, to a large extent, freed the statistical modeller.
- Statistics $\overset{\heartsuit}{\longrightarrow}$ Computation?
- Monte Carlo as statistical model: A theory of statistical integration for Monte Carlo models ('03, Kong et al) & subsequent papers by Zhiqiang Tan
- Data Augmentaion - Hidden structures / statistical insight: Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency (Meng & Yu, '11)
- "Rao-Blackwellization" for MCMC (Robert and Roberts, '21)

Statistics $\overset{\heartsuit}{\rightarrow}$ Computation

- Antithetic variates $\leftarrow$ Design of experiments.
- (Randomized) Quasi-Monte Carlo $\leftarrow$ Stratified sampling
- Control variates $\leftarrow$ Estimation

Double Happiness: Coupled MCMC and Control Variates

joint with Xiao-Li Meng

Unbiased MCMC - Pierre Jacob et al. '20

- Assume interest in approximating $I = \mathrm{E}_\pi[h(X)]$ using $\hat{I} = \frac{1}{M}\sum_{t=B}^{M+B} h(X_t)$, where $\{X_t\}_{t\geq 0}$ are MCMC samples from some posterior π item $\hat{I}$ vulnerable to potential biases due to:
  - Insufficient burn-in B
  - Chain Initialization
- These biases can accumulate when one is approximating the expectation I repeatedly.
- For instance, with parallel computations

$$\hat{I} = \mathrm{E}[\mathrm{E}_\pi[h(X)|\mathcal{U}_j]]$$

where the inner expectation is the estimate obtained from the jth parallel process generated using random deviates $\mathcal{U}_j$, and the outer mean averages over all processes.

A Coupling-based Solution

- Consider two chains $\mathcal{X} = \{X_t, t \geq 0\}$ and $\mathcal{Y} = \{Y_t, t \geq 0\}$
- 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 $t \geq \tau$.

A Coupling-based Solution

- $H_k(\mathcal{X}, \mathcal{Y}) = h(X_k) + \sum_{j=k+1}^{\tau-1}[h(X_j) - h(Y_{j-1})]$ has (under mild conditions) the same mean as

$$\begin{aligned} I &= h(X_k) + \sum_{j=k+1}^{\infty}[h(X_j) - h(Y_{j-1})] \\ &= h(X_k) + \sum_{j=k+1}^{\infty}[h(X_j) - h(X_{j-1})] \end{aligned}$$

which is an unbiased estimator for $\mathrm{E}_\pi[h(X)]$ for any $k \geq 0$ (see Glynn and Rhee, 2014; Glynn, 2016; Jacob et al, 2020; Biswas et al, 2019)

A Coupling-based Solution

- Generalize to a general "lag" L, i.e. find τ such that $X_t = Y_{t-L}$ for all $t \geq \tau$
- $H_{k,L}(\mathcal{X}, Y) = h(X_k) + \sum_{j=1}^{J_{k,L}} \left[h(X_{k+jL}) - h(Y_{k+(j-1)L}) \right]$ is unbiased for I, where $J_{k,L} = \max \left\{ 0, \lceil \frac{\tau_L - L - k}{L} \rceil \right\}$.
- For our purpose it is useful to express $H_{k,L}$ in the equivalent form

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

1^{st} Happiness: Control Variates for Variance Reduction

- Let $\Delta_{k,j} = h(X_{k+jL}) - h(Y_{k+jL})$ and note that $E[\Delta_{k,j}] = 0$ for all $k, j \geq 0$.
- Then $C_\eta = \sum_{j \geq 1} \eta_j \Delta_{k,j}$ is a control variate for $H_{k,L}(\mathcal{X}, Y)$, where $\vec{\eta} \equiv \{\eta_j, j \geq 1\}$ is independent of $\{\mathcal{X}, \mathcal{Y}\}$, and $\sum_{j=1} \mathrm{E}_{\vec{\eta}} |\eta_j| < \infty$,
- Replace $H_{k,L}(\mathcal{X}, \mathcal{Y})$ with

$$H_{k,L}^{(\vec{\eta})}(\mathcal{X}, Y) = H_{k,L}(\mathcal{X}, Y) - \sum_{j \geq 1} \eta_j \Delta_{k,j}.$$

A Remarkable Result

- From
$$E[h(X_\pi) - h(X_k)] = E\left\{\sum_{j=1}^{J_{k,L}} \left[h(X_{k+jL}) - h(Y_{k+(j-1)L})\right]\right\}$$
$$\Rightarrow d_{\mathrm{TV}}(\pi_k, \pi) \leq \mathrm{E}[J_{k,L}]$$
- Instead of trying to minimize the variance of $H_{k,L}^{(\vec{\eta})}(\mathcal{X}, Y)$ we optimize $\vec{\eta}$ so that the resulting TV inequality is tighter!

2^{nd} Happiness: A Refined Bound

- We show:

$$\begin{aligned} d_{\mathrm{TV}}(\pi_k, \pi) &\leq 0.5 \left\{ \sum_{j \geq 1} \mathrm{E}|1_{\{j \leq J_{k,L}\}} - \eta_j| + \sum_{j \geq 1} \mathrm{E}|\eta_j - 1_{\{j \leq J_{k,L}-1\}}| \right. \\ &+ 0.5 \Pr(J_{k,L} > 0) \\ &= \sum_{j \geq 1} \mathrm{E}|1_{\{j \leq \tilde{J}_{k,L}\}} - \eta_j| + 0.5 \Pr(J_{k,L} > 0), \end{aligned}$$

  where $\tilde{J}_{k,L} = J_{k,L} - \xi$ and $\xi \sim \mathit{Bernoulli}(0.5)$

- Recall: for any given random variable V, $\min_{U \perp V} E|V - U| = E|V - m_V|$, where m_V is a median of V.

2^{nd} Happiness: A Refined Bound

- Let $\tilde{m}_{J_{k,L}}$ be the smallest integer median of $\tilde{J}_{k,L}$ and let $\eta_j = \mathbf{1}_{\{j<\tilde{m}_{k,L}\}}, \quad \text{for any } j.$
- In order for $\vec{\eta}$ to be independent from $\mathcal{X}, \mathcal{Y}$ we will use R pairs of coupled chains, independently run in parallel
- For each process use the estimate of $\tilde{m}_{J_{k,L}}$ obtained from the "other" $R-1$ processes.

2^{nd} Happiness: A Refined Bound

- We can show that this choice of $\vec{\eta}$ yields the bound

$$B_{k,L} \quad = \quad \mathrm{E}|J_{k,L} - m_{J_{k,L}}| + \Pr(J_{k,L} > 0) - S_{k,L}$$

  where $S_{k,L} = \max\{\Pr(J_{k,L} > m_{J_{k,L}}), \Pr(J_{k,L} < m_{J_{k,L}})\} \leq 0.5$ and $m_{J_{k,L}}$ is the smallest integer median of $J_{k,L}$.

- Always $B_{k,L} \leq E[J_{k,L}]\ \forall k, L$

- Whenever, $m_{J_{k,L}} = 0$, $B_{k,L} = E[J_{k,L}]$.

- Note that $B_{k,L}$ depends on the coupling time mean, its variance and the symmetry of its distribution.

Approximate Bayesian Computation with Friendly Neighbours

joint with Evgeny Levi

MCMC at the crossroads

- Large data and 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 $\boldsymbol{\theta} \in \mathbf{R}^d$.
- An MH proposal $\xi \sim q(\xi|\theta)$ is accepted with probability

$$\min\left\{1, \frac{\pi(\xi|\mathbf{y}_0)q(\theta|\xi)}{\pi(\theta|\mathbf{y}_0)q(\xi|\theta)}\right\}.$$

- 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)

Double Jeopardy: Intractable Likelihoods & Big Data

- 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."

Intractable Likelihood

- The likelihood $L(\theta|\mathbf{y})$ is not computable but one can sample from $f(\mathbf{y}|\theta)$ for all θ's
- Approximate Bayesian Computation (ABC - Marin et al., Comp & Stat. 2012)
- Bayesian Synthetic Likelihood (BSL - Price et al, JCGS 2018) methods can be used.
- Indirect inference (Smith Jr, 1993; Gourieroux et al. 1993; Gallant and McCulloch, 2009)

Bayesian Synthetic Likelihood (BSL)

- Complex model: $f(\mathbf{y}|\boldsymbol{\theta})$ with intractable f
- Simpler model: $g(S(\mathbf{y})|\boldsymbol{\theta})$ approximates $f(S(\mathbf{y})|\boldsymbol{\theta})$
- g is Gaussian with parameters $\phi(\boldsymbol{\theta}) = (\mu_\theta, \Sigma_\theta)$
- The Synthetic Likelihood $\text{SL}(\theta|s_0) = \mathcal{N}(s_0; \mu_\theta, \Sigma_\theta)$, where $s_0 = S(\mathbf{y}_0)$.
- $\mu_\theta, \Sigma_\theta$ are estimated from m statistics $(s_1 := S(\mathbf{y}_1), \ldots, s_m := S(y_m))$ where $y_j \sim f(\mathbf{y}|\theta)$.
- The BSL posterior is $\pi(\theta|s_0) \propto p(\theta)\text{SL}(\theta|s_0)$

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 $\hat{\mu}_\theta$ and $\hat{\Sigma}_\theta$
- The synthetic likelihood is

$$\mathrm{SL}(\theta|s_0) = \mathcal{N}(S(\mathbf{y}_0); \hat{\mu}_\theta, \hat{\Sigma}_\theta), \tag{1}$$

  where $s_0 = S(\mathbf{y}_0)$

- A MH sampler requires $\mathrm{SL}(\theta|s_0)/\mathrm{SL}(\theta_t|s_0)$

BSL with precomputed proposals

- Precompute the proposals for the chain (parallelizable task)
- For $1 \leq h \leq H$ let $\xi_h \sim p(\xi)$ and m pseudo-data $\mathbf{w}_h^{(1)}, \ldots, \mathbf{w}_h^{(m)} \overset{iid}{\sim} f(\mathbf{w}|\xi_h)$;
- Set $s_h^{(k)} = S(\mathbf{w}_h^{(j)})$, $1 \leq j \leq m$, and $\mathcal{Z} = \{(\xi_h, s_h = [s_h^{(1)}, \ldots, s_h^{(m)}]) : 1 \leq h \leq H\}$
- Use $\mathcal{Z}$ for running MCMC-BSL.
- The number m of pseudo-data sets generated for each ξ is small so we do not have $\mathrm{SL}(\xi|s_0)$.

BSL with precomputed proposals

- When proposal is θ^*.

$$\begin{aligned}\hat{\mu}_{\boldsymbol{\theta}^*} &= \frac{\sum_{h=1}^{H}[W_h(\boldsymbol{\theta}^*)\frac{1}{m}\sum_{j=1}^{m} s_h^{(j)}]}{\sum_{h=1}^{H} W_h(\boldsymbol{\theta}^*)}, \\ \hat{\Sigma}_{\boldsymbol{\theta}^*} &= \frac{\sum_{h=1}^{H}[W_h(\boldsymbol{\theta}^*)\frac{1}{m}\sum_{j=1}^{m}(s_h^{(j)} - \hat{\mu}_{\boldsymbol{\theta}^*})(s_h^{(j)} - \hat{\mu}_{\boldsymbol{\theta}^*})^T]}{\sum_{h=1}^{H} W_h(\boldsymbol{\theta}^*)}.\end{aligned} \tag{2}$$

- $W_h(\boldsymbol{\theta}^*) = 1$ or $W_h(\boldsymbol{\theta}^*) = 1 - \|\xi_h - \boldsymbol{\theta}^*\|/\|\xi^* - \boldsymbol{\theta}^*\|$ and $\xi^* = \max_{\xi \in \mathcal{Z}} \|\xi - \boldsymbol{\theta}^*\|$, i.e. is the point in $\mathcal{Z}$ that is furthest away from $\boldsymbol{\theta}^*$.

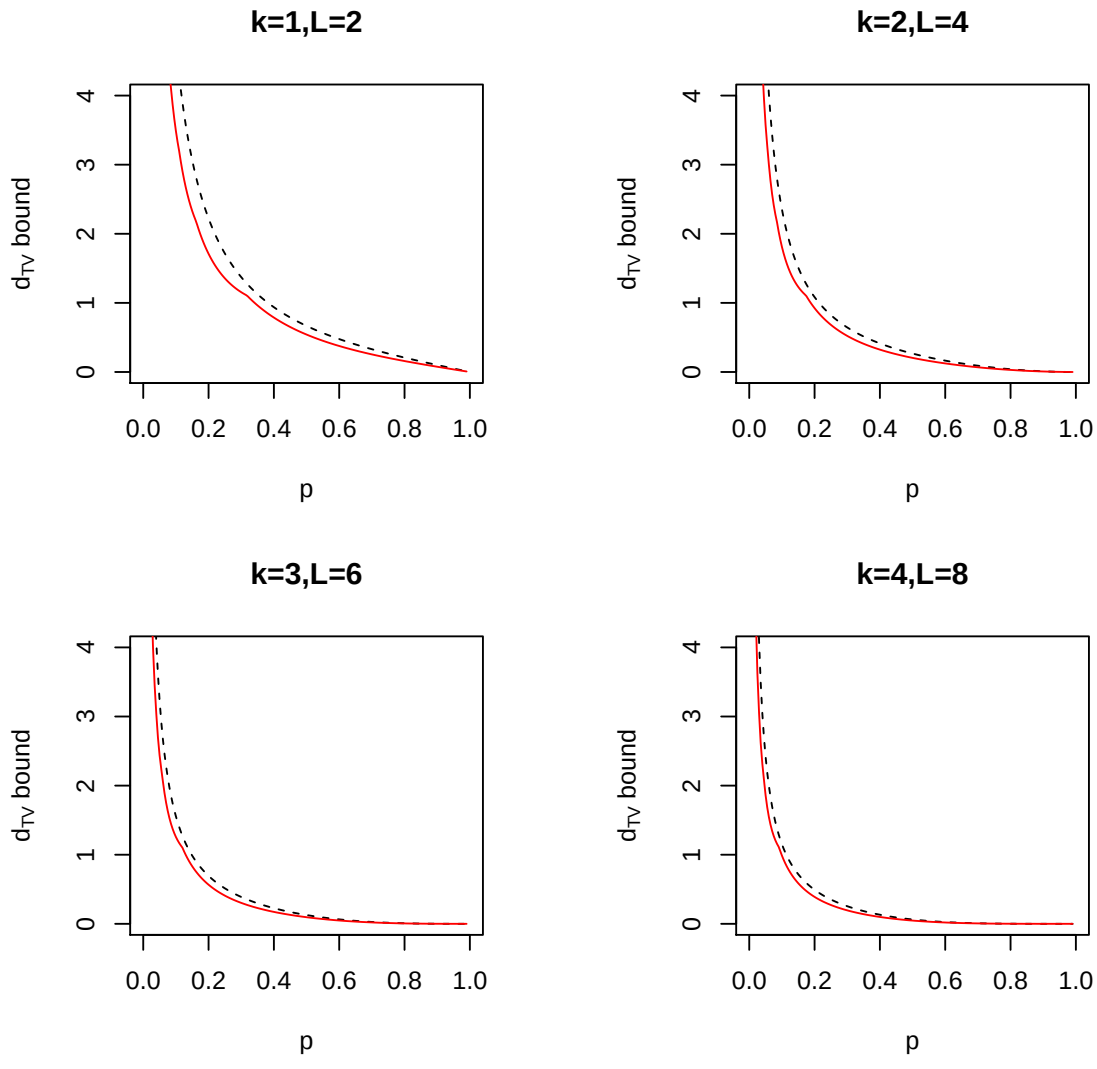
ABC with precomputed proposals

- The target of ABC-MCMC is

$$\pi_\epsilon(\theta|s_0) \propto p(\theta)\Pr(d(S(\mathbf{y}), s_0) \leq \epsilon|\theta)$$

- At each step one needs to estimate $\Pr(d(S(\mathbf{y}), s_0) \leq \epsilon|\theta)$.
- An unbiased estimator requires generating pseudo-data for each proposal θ^*
- Details in Levi & C (2022).

TV bound: Geometric case



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.
- $\beta \sim N(0, 10\mathbf{I}_d)$, $\Pr(Y_i = 1|x_i) = [1 + \exp(-x_i^T \beta)]^{-1}$
- 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, $\beta \in \mathbf{R}^{49}$ 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

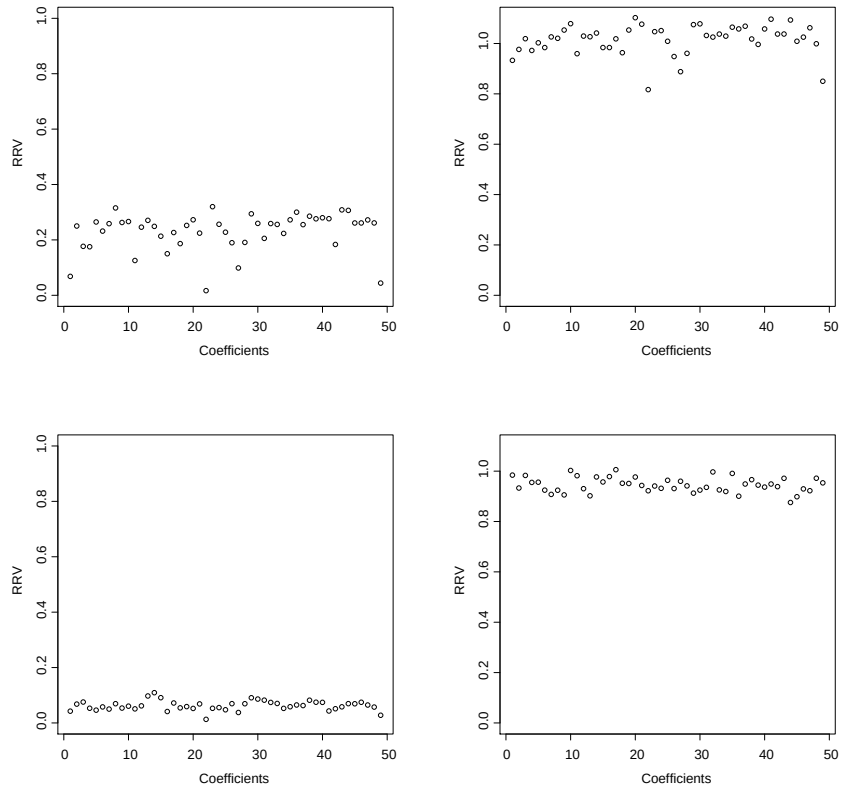


Figure: German Credit Data. Relative (RRV) reduction in variance for the 49 regression coefficients. *Top panels*: the lag is $L = 5$. *Bottom panels*: the lag is $L = 20$. *Left panels*: RRV is obtained using $k = 5$. *Right panels*: RRV is obtained from the average estimators with $k = 5$ and $r = 30$.

Conclusions

- Computation $\overset{\heartsuit}{\rightarrow}$ Statistics.
- Is it time for more Statistics $\overset{\heartsuit}{\rightarrow}$ Computation?

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

We're HIRING: Assistant Prof. in Computational Statistics at U of Toronto! Deadline: Middle of November

Numerical Experiments: Ricker's Model

- A particular instance of hidden Markov model:

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

  where $\mathit{Pois}(\lambda)$ 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 $\mathbf{y}$, $\bar{y}$,

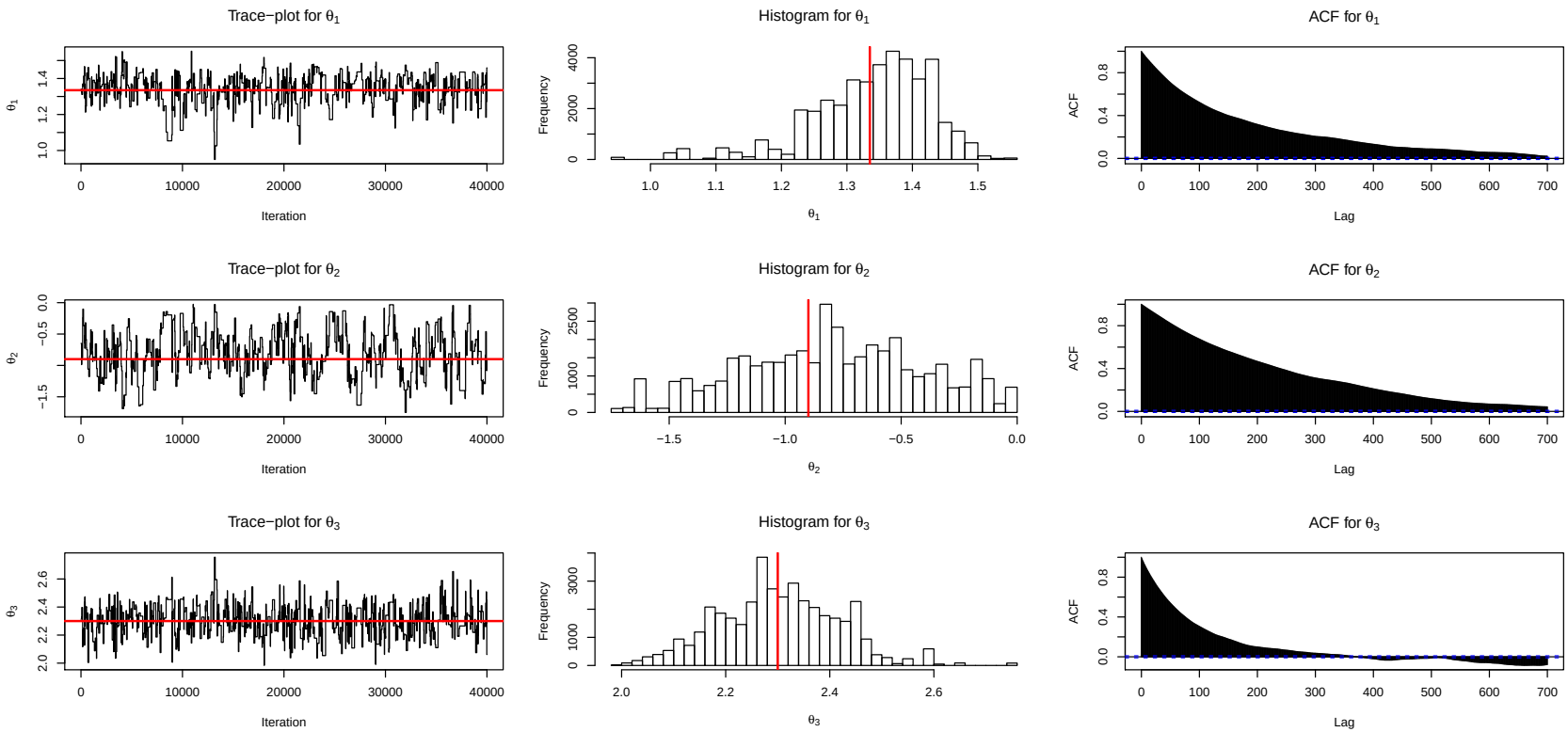
(C3:C7) Sample auto-correlations at lags 1 through 5,

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

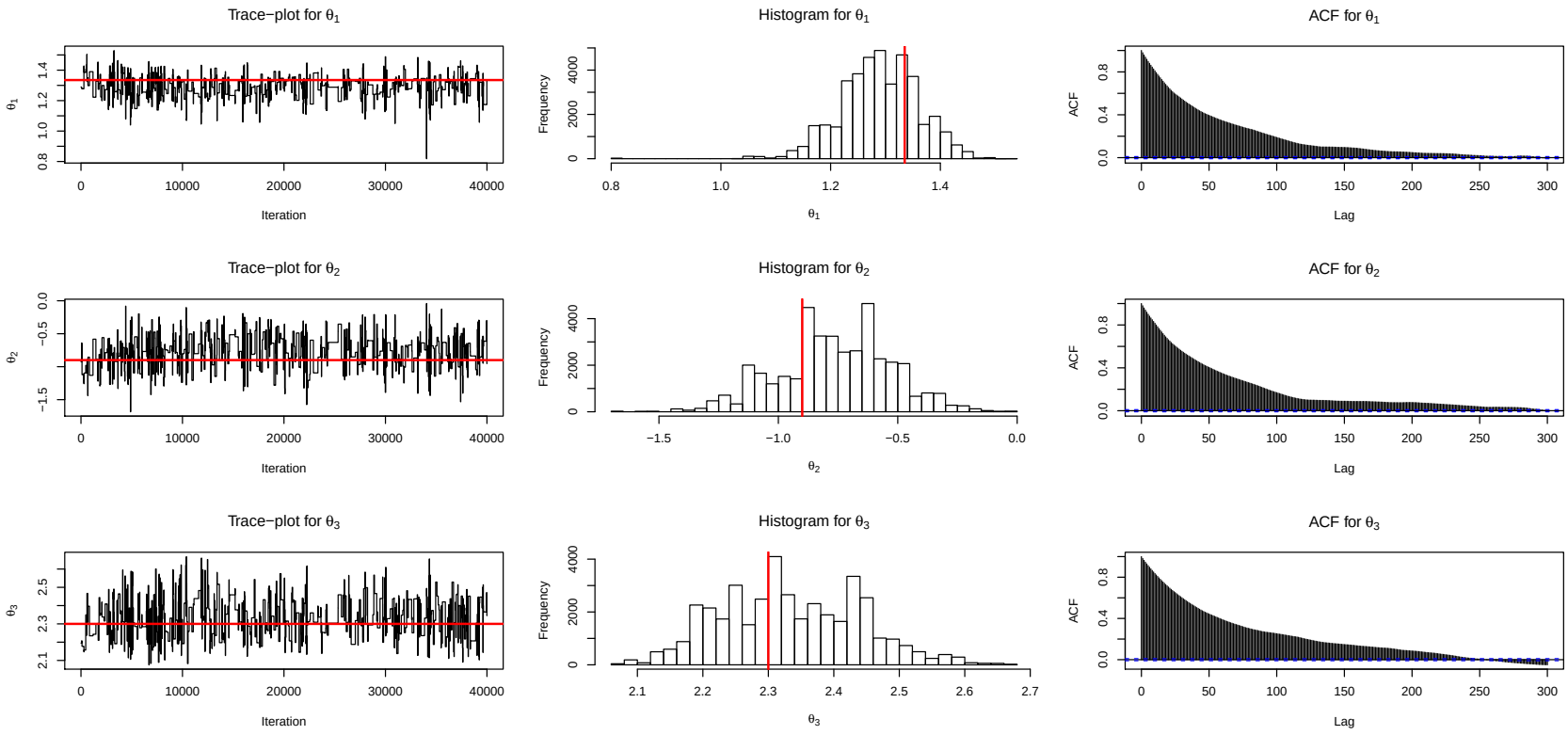
Numerical Experiments: Ricker's Model - ABC/RWM

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.



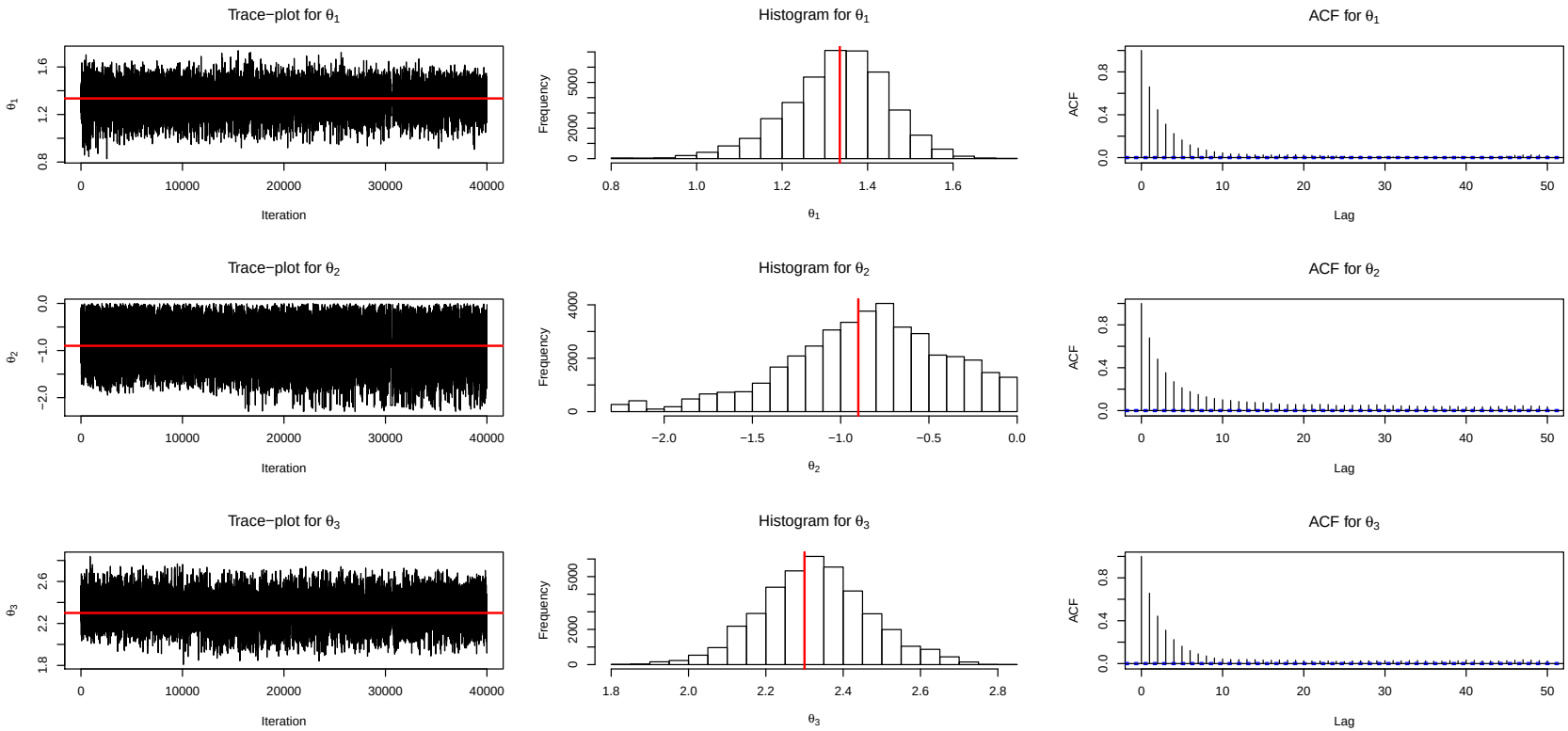
Numerical Experiments: Ricker's Model - BSL

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