Accelerating MCMC by Interweaving Multiple Parameterisations

Yaming Yu

Department of Statistics, University of California, Irvine

Joint work with Xiao-Li Meng of Harvard University

Basic Setting

- **Model of interest:**

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

- **Introducing latent variable Z:**

$$p(Z, Y|\theta) = p(Z|Y, \theta)p(Y|\theta)$$

- **Gibbs sampling:**

$$Z|(\theta, Y) \longleftrightarrow \theta|(Z, Y)$$

- **Different choices of Z lead to different Gibbs samplers**

Parameterisation Matters: A Simple Example

- **One-way random effect model:**

$$Y = \theta + Z + \epsilon, \quad Z \sim N(0, V), \quad \epsilon \sim N(0, 1), \quad Z \perp \epsilon,$$

where Z is a latent variable.

- **Prior distribution:** $p(\theta) \propto 1$
- **Posterior distribution:** $\theta | Y \sim N\left(Y, 1 + V\right)$

- **Parameterise in terms of (θ, Z):**

$$Y = \theta + Z + \epsilon, \quad Z \sim N(0, V), \quad \epsilon \sim N(0, 1), \quad Z \perp \epsilon,$$

- **Ancillarity: $Z|\theta \sim N(0,\ V)$.**
- **Gibbs sampling (Scheme AA):**

$$Z|(\theta, Y) \sim N\left(\frac{V(Y-\theta)}{1+V},\ \frac{V}{1+V}\right),$$
$$\theta|(Z, Y) \sim N(Y - Z,\ 1).$$

- **Convergence rate $r_1 = V/(1+V)$: fast when V is small**

- **Parameterise in terms of θ and $W \equiv \theta + Z$:**

$$Y = W + \epsilon, \quad W|\theta \sim N(\theta, V), \quad \epsilon \sim N(0,1), \quad W \perp \epsilon.$$

- **Sufficiency: $Y|(\theta, W) \sim N(W, 1)$**
- **Gibbs sampling (Scheme SA):**

$$W|(\theta, Y) \sim N\left(\frac{\theta + VY}{1+V}, \ \frac{V}{1+V}\right),$$
$$\theta|(W, Y) \sim N\left(W, \ V\right).$$

- **Convergence rate $r_2 = 1/(1+V)$: fast when V is large**

How about using two Parameterisations?

$$\theta \rightleftarrows Z \qquad\qquad \theta \rightleftarrows W$$

Obvious ...

$$\theta \rightleftarrows Z \qquad \theta \rightleftarrows W$$

Obvious: Alternating Scheme

$$\begin{array}{ccccc} & & \theta & & \\ & \swarrow & & \nwarrow & \\ Z & & & & W \\ & \searrow & & \nearrow & \\ & & \theta & & \end{array}$$

- **Convergence rate** $r_1 r_2 = V/(1+V)^2 \leq 1/4$.
- **Robust! (to the value of** V**)**

Not so obvious ...

Obvious: Alternating Scheme

$$\begin{array}{ccccc} & & \theta & & \\ & \swarrow & & \nwarrow & \\ Z & & & & W \\ & \searrow & & \nearrow & \\ & & \theta & & \end{array}$$

Surprise: Interweaving Scheme

$$\begin{array}{ccccc} & & \theta & & \\ & \swarrow & & \nwarrow & \\ Z & & \longrightarrow & & W \end{array}$$

To achieve $Z \longrightarrow W$, draw $\theta | Z$ and then $W | (\theta, Z)$.

How Fast is Interweaving?

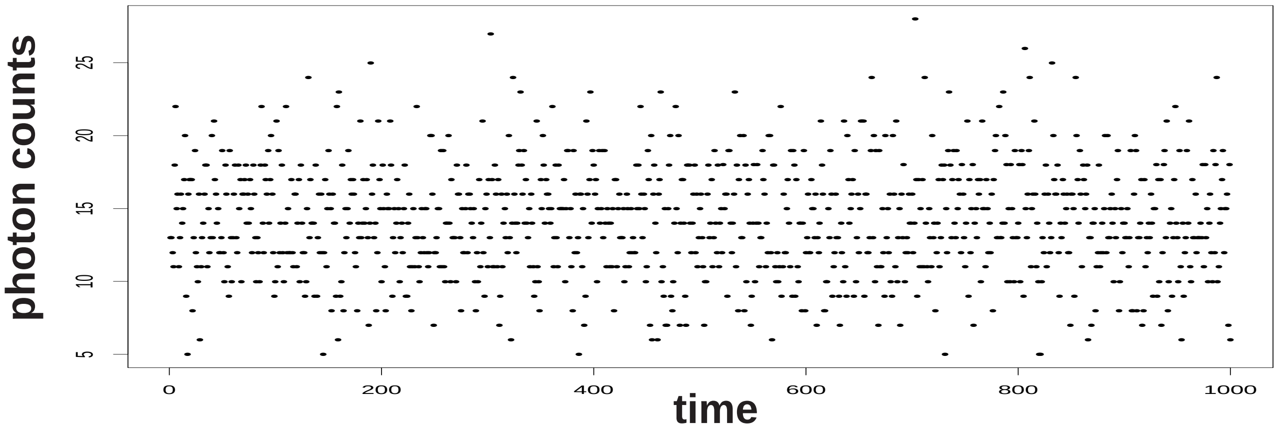
- **SA:** $(A)\ W|\theta \sim N\left(\frac{\theta+VY}{1+V},\ \frac{V}{1+V}\right), \quad (B)\ \theta|W \sim N(W,\ V)$.
- **AA:** $(A')\ Z|\theta \sim N\left(\frac{V(Y-\theta)}{1+V},\ \frac{V}{1+V}\right), \quad (B')\ \theta|Z \sim N(Y-Z,\ 1).$
- **Interweaving: Replace (A') by $Z = W - \theta$**
  - **Perform (A) to obtain $W^{(t+1)}$ and then (B) to obtain an *intermediate***
    $\theta^{(t+0.5)} = W^{(t+1)} + \sqrt{V}\delta_1, \qquad \delta_1 \sim N(0,1)$
  - **Let $Z^{(t+1)} = W^{(t+1)} - \theta^{(t+0.5)} = -\sqrt{V}\delta_1$ and then perform (B'):**
    $\theta^{(t+1)} = Y - Z^{(t+1)} + \delta_2, \qquad \delta_2 \sim N(0,1)$
  - **But then $r_{ASIS} = 0$ because ASIS converges in a SINGLE iteration:**
    $\theta^{(t+1)} = Y + \sqrt{V}\delta_1 + \delta_2 \sim N(Y, 1+V) \longleftarrow$ **target distribution!**

For More Details...

- **Conditional augmentation and marginal augmentation (Meng and van Dyk, 1997, 1999; van Dyk and Meng, 2001);**
- **PX-DA (Liu and Wu, 1999);**
- **Cerntering, non-centering and partially non-centering (e.g., Gelfand, Sahu, and Carlin, 1995, 1996; Papaspiliopoulos, Roberts and Sköld, 2003, 2007)**
- **Sandwich Algorithms (Hobert and Marchev, 2008)**
- **Ancillary-sufficient Interweaving Strategy (ASIS; Yu and Meng, 2011)**
- **...**

Chandra X-ray: The problem motivated us ...

- **Poisson variation of the counts given the intensity.**
- **Variation of the intensity itself, due to X-ray flare, Binary systems, Gradual cooling, etc.**
- **The neutron star/quark star candidate RX J1856.5-3754 observed by Chandra (exposure time 55476 seconds, divided into 1000 bins).**



A Parameter-Driven Poisson Time Series Model

$\theta = (\beta_0 \quad \beta_1 \quad \rho \quad \delta)$

baseline — trend — autocorr. — residual s.d.

Y_{obs} : counts observed

Y_{mis} : depends on the augmentation scheme

$$Y_t|(\xi_t, \beta) \overset{ind}{\sim} Pois\left(d_t e^{\beta_0+\beta_1 t+\xi_t}\right);$$

$$\xi_t|(\xi_{<t}, \beta, \rho, \delta) \sim N(\rho\xi_{t-1},\ \delta^2).$$

- Y_t: counts in bin t, $\quad t = 1, \ldots, T$;
- d_t: width (e.g., in seconds) of bin t;
- $\xi = \{\xi_t\}$ is a stationary AR(1) process;
  $\xi_t \sim N(0,\ \tau^2)$, where $\tau^2 = \delta^2/(1-\rho^2)$.

The Standard Gibbs Sampler

$$Y_t \sim Pois\left(d_t e^{\beta_0+\beta_1 X_t+\xi_t}\right), \quad \xi_t|\xi_{<t} \sim N(\rho\xi_{t-1},\ \delta^2); \quad p(\beta,\rho,\tau) \propto 1$$

1. $\xi|(\beta,\ \rho,\ \delta)$: **draw ξ, the missing data.**

   Difficult to update all ξ's simultaneously, so update $\xi_t|(\xi_{t-1},\ \xi_{t+1})$ in turn.

2. $\beta|(\xi,\ \rho,\ \delta)$, **or** $\beta|\xi$

   Equivalent to posterior sampling of a Poisson GLM. Need an M–H move.

3. $(\rho,\ \delta)|(\xi,\ \beta)$, **or** $(\rho,\ \delta)|\xi$

   Equivalent to Bayesian fitting of an AR(1) model:

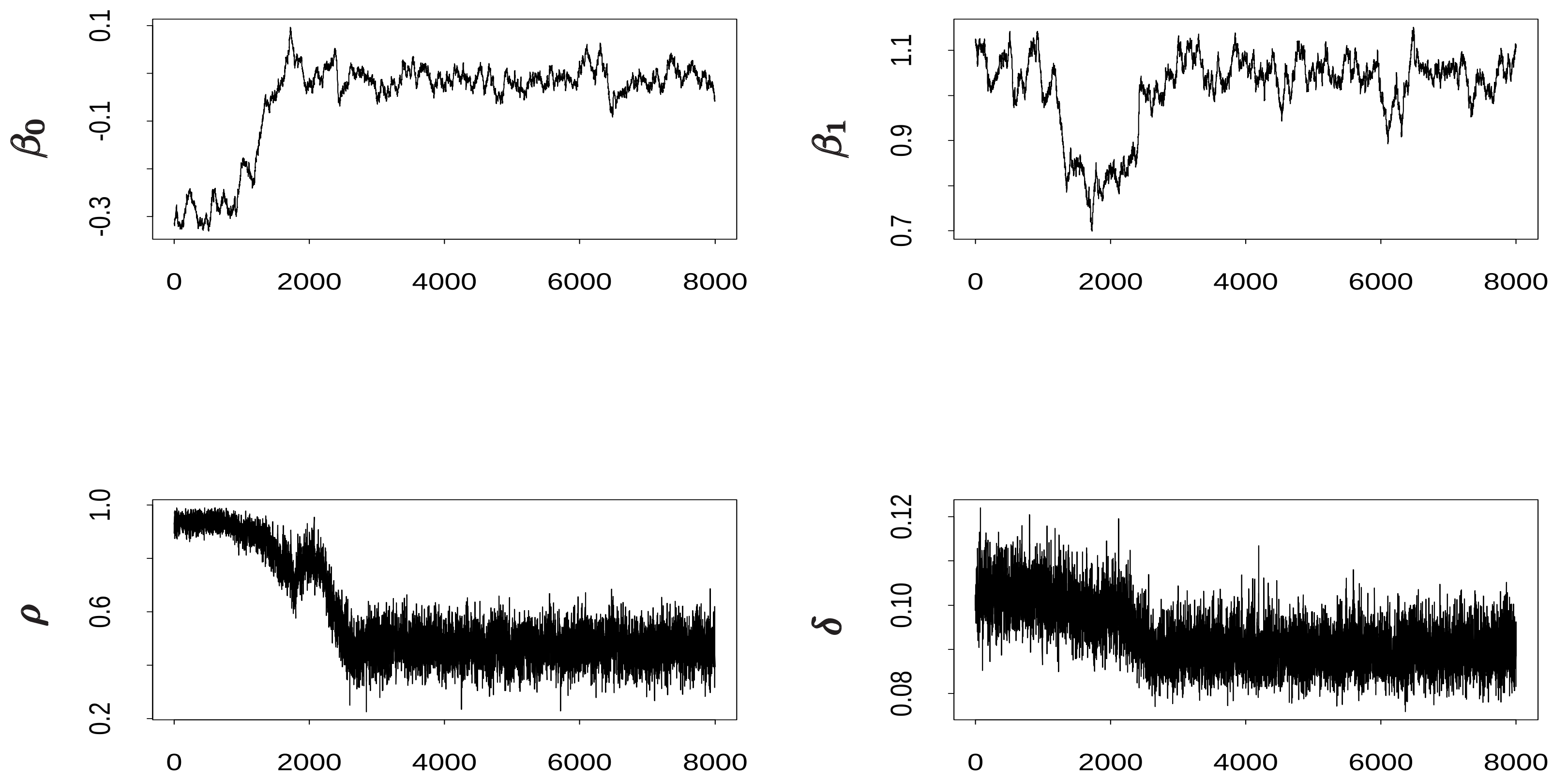
$$\xi_t = \rho\xi_{t-1} + N(0,\ \delta^2), \quad t = 1, \ldots, T.$$

Performance of the Standard Gibbs Sampler

A simulation:

- **Counts are generated according to the correct model**
  - $T = 200$, $d_t = 5000$, **and** $X_t = t/T$.
  - **Parameter values:** $(\beta_0, \beta_1, \rho, \delta) = (0, 1, 0.5, 0.1)$.
- **Counts are in the order of thousands.**

It’s not working!



Component-wise ASIS

- $Y_t \sim Pois\left(d_t e^{\beta_0+\beta_1 X_t+\xi_t}\right); \quad \xi_t|\xi_{<t} \sim N(\rho\xi_{t-1},\ \delta^2).$
- $\eta_t = \xi_t + \beta_0 + \beta_1 X_t$
  - η is the **conditional** sufficient augmentation (CSA) for β;
  - ξ is the **conditional** ancillary augmentation (CAA) for β.

  Step 2': $\beta|(\eta,\ \rho,\ \delta)$. (**In addition** to Steps 1–3.)
- The Poisson likelihood does not play a role; only need linear regression.
- To keep track of ξ, set $\xi_t^{new} = \eta_t - \beta_0^{new} - \beta_1^{new} X_t$ right after Step 2'.

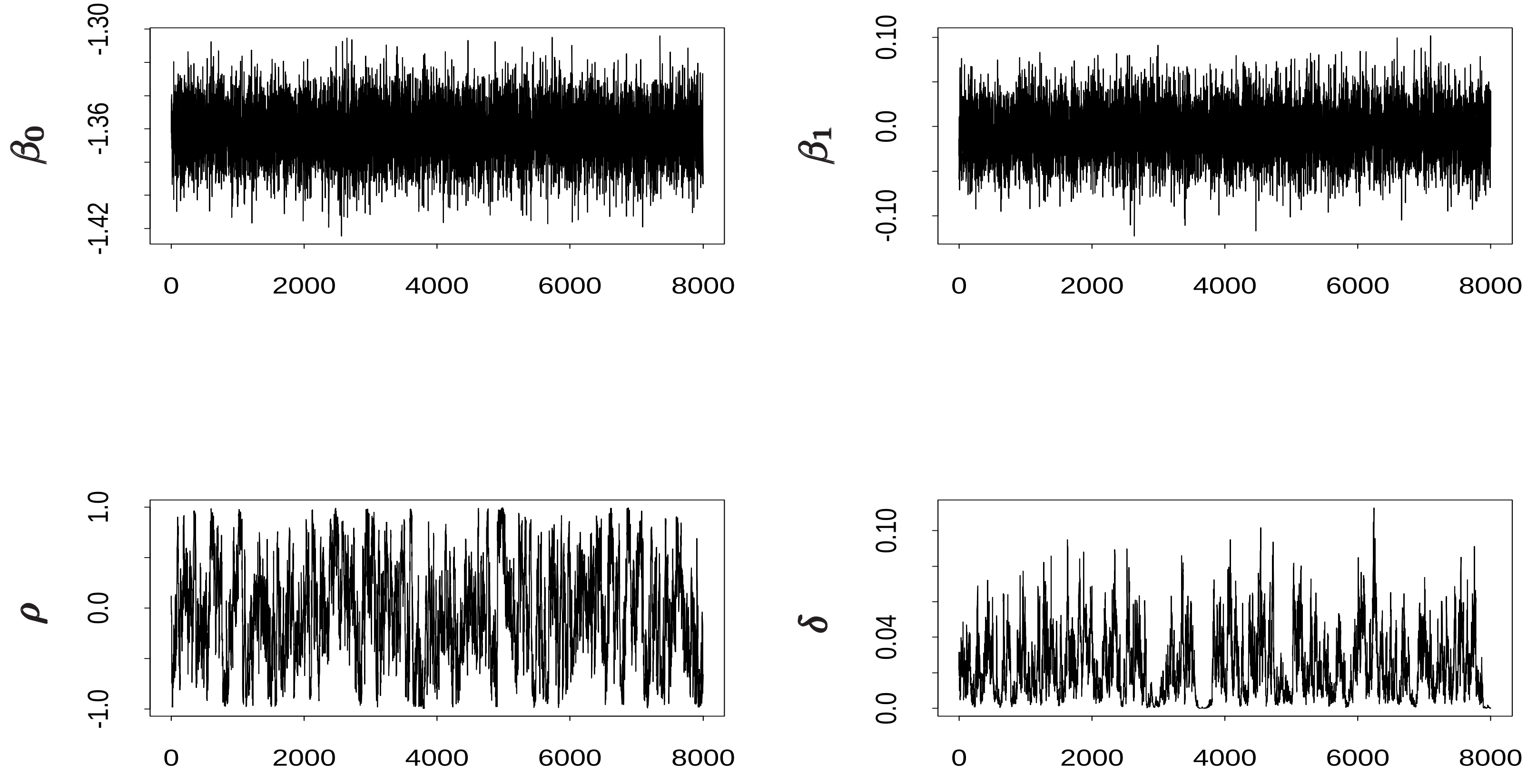
The flexibility of CIS

- **Step 1: draw** $\xi|(\beta,\ \rho,\ \delta)$**.**
- **Steps 2–2' and 3–3": draw**

	CSA	CAA
$\beta\|$	$\eta:\ \eta_t = \xi_t + \beta_0 + \beta_1 X_t$	ξ *
$\rho\|$	ξ *	$\zeta:\ \zeta_t = \xi_t - \rho\xi_{t-1}$
$\delta\|$	ξ *	$\kappa:\ \kappa_t = \xi_t/\delta$

  - ***: Steps performed by the standard Gibbs sampler.**
  - $\rho|(\xi,\ \beta,\ \delta)$ **and** $\delta|(\xi,\ \beta,\ \rho)$ **are combined into** $(\rho,\ \delta)|(\xi,\ \beta)$**.**

It is vastly improved!



Posterior Summary: Intensity Does Not Change