Down-Up Metropolis-Hastings Algorithm for Multimodality

Hyungsuk Tak

Stat310

24 Nov 2015

Joint work with Xiao-Li Meng and David A. van Dyk

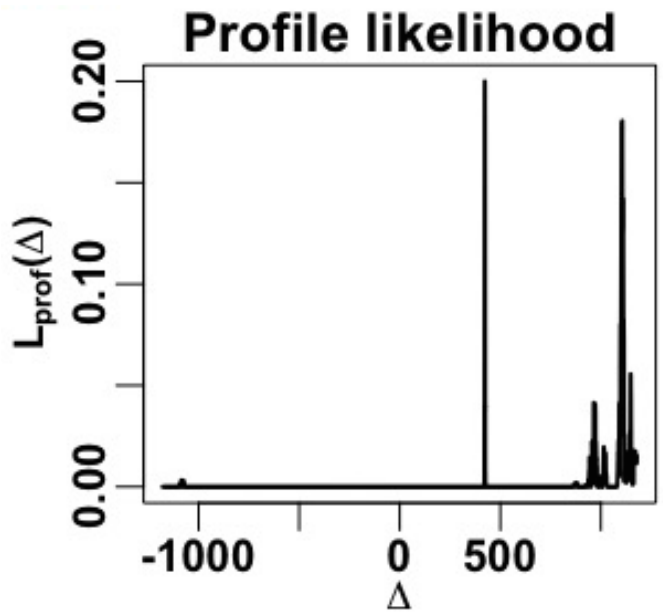
Outline

- Motivation & idea
- Down-Up Metropolis-Hastings (DUMH) algorithm
  - Mathematical specification
  - Auxiliary variable approach
  - Algorithmic specification
- Examples
  - A mixture of 20 bivariate Gaussian distributions (Kou et al., 2006)
  - Quasar Q0957+561
- Discussion

Motivation

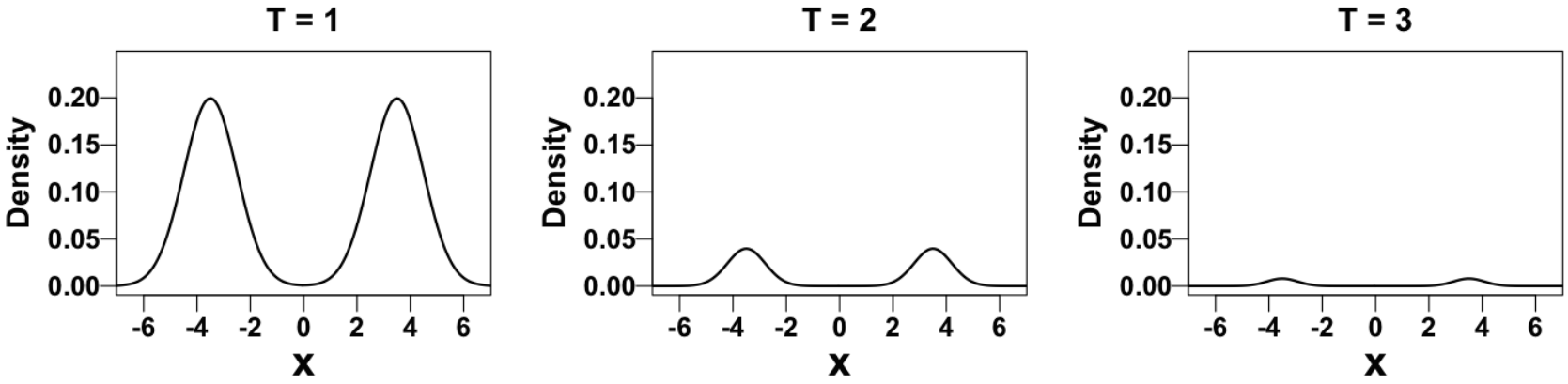
Time delay estimation between gravitationally-lensed quasar images to calculate the current expansion rate of the Universe.

- Full posterior density function: $\pi(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \text{Data})$.
- Profile likelihood: $L_{\text{prof}}(\Delta) \propto \pi(\Delta \mid \text{Data})$.
- A multimodal posterior distribution of the time delay (Δ in days) for Quasar Q0957+561



Motivation (cont.)

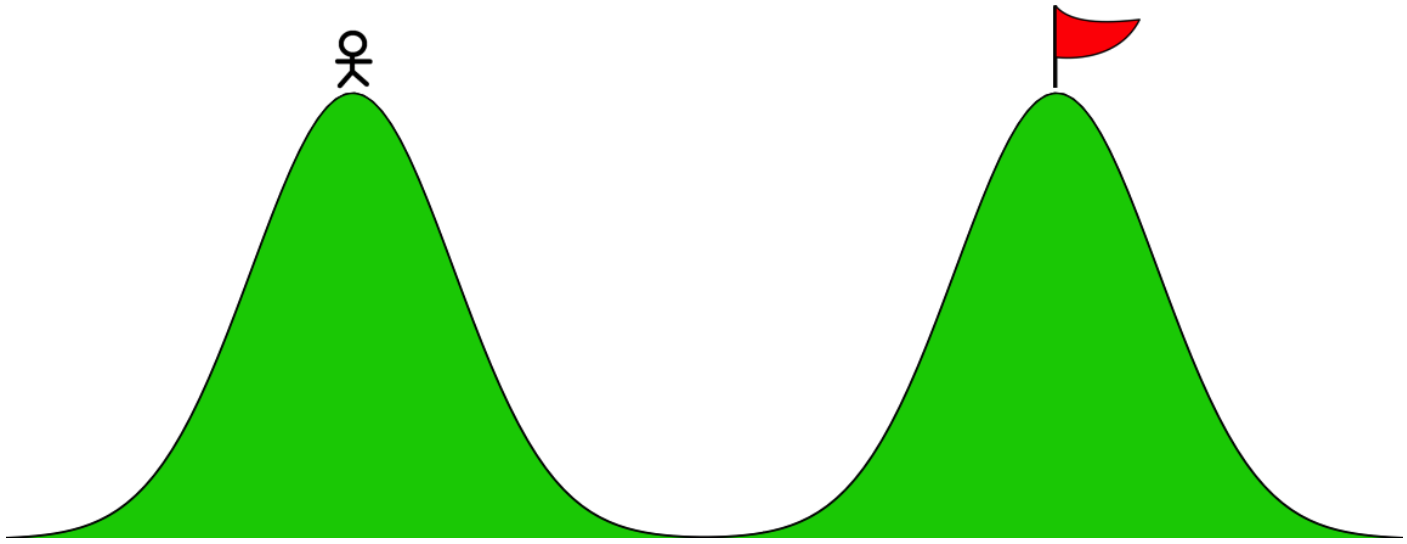
Popular remedy for multimodality is tempering, i.e., $\pi(x)^T$.



- Parallel tempering (Geyer, 1991), simulated tempering (Marinari & Parisi, 1992), tempered transition (Neal, 1996), Equi-energy sampler (Kou et al., 2006), etc.
- Hard to tune several temperature-related tuning parameters off-line
- Computationally expensive
- Is there a simple and fast way to expedite jumps between modes?

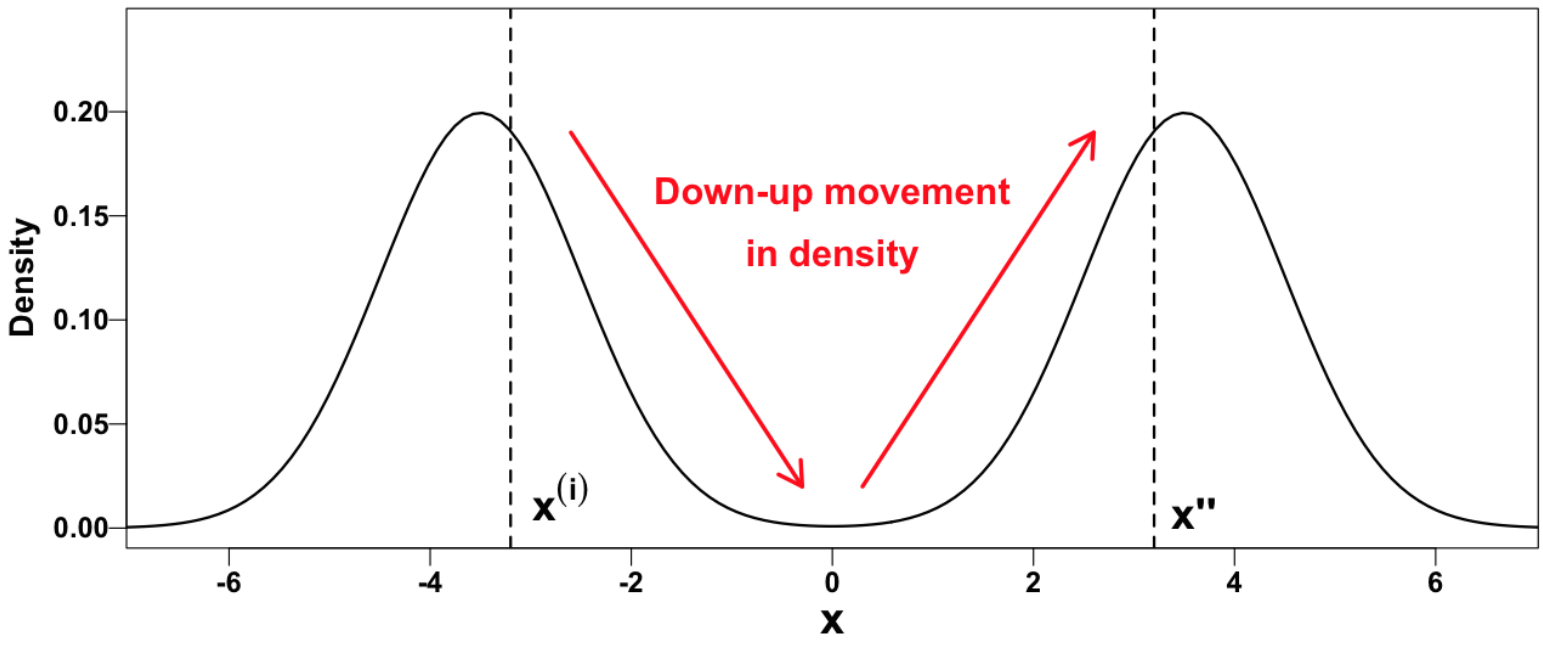
Idea

How would you go to the top of the other mountain?



Idea (cont.)

Let's make a down-up movement in density to generate a proposal **x"**



A down-up Metropolis-Hastings

Two-step Metropolis transitions: $x^{(i)}$: Current state
$\rightarrow x'$: Intermediate proposal
$\rightarrow x''$: Final proposal

- How do we make a downhill Metropolis transition from $x^{(i)}$ to x'?

$$\alpha^{\mathrm{D}}(x' \mid x^{(i)}) = \min\left\{1,\ \frac{\pi(x^{(i)})}{\pi(x')}\right\}$$

- How do we make an uphill Metropolis transition from x' to x''?

$$\alpha^{\mathrm{U}}(x'' \mid x') = \min\left\{1,\ \frac{\pi(x'')}{\pi(x')}\right\}$$

- For computational stability,

$$\alpha_{\epsilon}^{\mathrm{D}}(x' \mid x^{(i)}) = \min\left\{1,\ \frac{\pi(x^{(i)}) + \epsilon}{\pi(x') + \epsilon}\right\},\ \alpha_{\epsilon}^{\mathrm{U}}(x'' \mid x') = \min\left\{1,\ \frac{\pi(x'') + \epsilon}{\pi(x') + \epsilon}\right\}$$

A down-up Metropolis-Hastings (cont.)

Proposal density function of DUMH

$$q^{\mathrm{DU}}(x'' \mid x^{(i)}) = \int q^{\mathrm{D}}(x' \mid x^{(i)}) q^{\mathrm{U}}(x'' \mid x') dx',$$

where the downhill kernel density is

$$q^{\mathrm{D}}(x' \mid x^{(i)}) = N(x' \mid x^{(i)}, \sigma^2) \alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)}) + r^{\mathrm{D}}(x^{(i)}) \delta_{x^{(i)}}(x'),$$
$$r^{\mathrm{D}}(x^{(i)}) = 1 - \int N(x' \mid x^{(i)}, \sigma^2) \alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)}) dx',$$

and similarly the uphill kernel density is

$$q^{\mathrm{U}}(x'' \mid x') = N(x'' \mid x', \sigma^2) \alpha_\epsilon^{\mathrm{U}}(x'' \mid x') + r^{\mathrm{U}}(x') \delta_{x'}(x''),$$
$$r^{\mathrm{U}}(x') = 1 - \int N(x'' \mid x', \sigma^2) \alpha_\epsilon^{\mathrm{U}}(x'' \mid x') dx'',$$

A down-up Metropolis-Hastings (cont.)

Four possibilities of the final proposal:

- Down-Up: Both accepted
- Down- : Downhill accepted & uphill rejected (not helpful for jump)
- -Up: Downhill rejected & uphill accepted (not helpful for jump)
- : Both rejected ($x^{(i)} = x''$) (waste of iterations)

We only need "Down-up" movement to boost jumps between modes!

A down-up Metropolis-Hastings (cont.)

Exclude the last three possibilities via forced Metropolis transitions.

$$q^{\mathrm{D}}(x' \mid x^{(i)}) = N(x' \mid x^{(i)}, \sigma^2)\alpha^{\mathrm{D}}(x' \mid x^{(i)}) + r^{\mathrm{D}}(x^{(i)})\delta_{x^{(i)}}(x'),$$

$$r^{\mathrm{D}}(x^{(i)}) = 1 - \int N(x' \mid x^{(i)}, \sigma^2)\alpha^{\mathrm{D}}(x' \mid x^{(i)})dx',$$

$$\rightarrow q^{\mathrm{D}}(x' \mid x^{(i)}) = \frac{N(x' \mid x^{(i)}, \sigma^2)\alpha^{\mathrm{D}}(x' \mid x^{(i)})}{1 - r^{\mathrm{D}}(x^{(i)})}$$

- How do we generate $x' \sim q^{\mathrm{D}}(x' \mid x^{(i)})$?

  Repeatedly generate $x' \sim q^{\mathrm{D}}(x' \mid x^{(i)})$ until it is accepted!
- Intuition: Flip a coin twice. Restrict the sample space to {HT, HH}. How do we generate {HT} or {HH}? Flip the coin twice until either {HT} or {HH} appears, ignoring {TT} and {TH}.

A down-up Metropolis-Hastings (cont.)

Accept x'' with an MH acceptance probability

$$\alpha^{\mathrm{DU}}(x'' \mid x^{(i)}) = \min\left\{1,\ \frac{\pi(x'')q^{\mathrm{DU}}(x^{(i)} \mid x'')}{\pi(x^{(i)})q^{\mathrm{DU}}(x'' \mid x^{(i)})}\right\}$$
$$= \min\left\{1,\ \frac{\pi(x'')(1 - r^{\mathrm{D}}(x^{(i)}))}{\pi(x^{(i)})(1 - r^{\mathrm{D}}(x''))}\right\}.$$

The second equality holds because

$$\alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)}) = \min\left\{1,\ \frac{\pi(x^{(i)}) + \epsilon}{\pi(x') + \epsilon}\right\} = \alpha_\epsilon^{\mathrm{U}}(x^{(i)} \mid x')$$

and thus

$$q^{\mathrm{DU}}(x'' \mid x^{(i)}) = \int \frac{N(x' \mid x^{(i)}, \sigma^2)\alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)})}{1 - r^{\mathrm{D}}(x^{(i)})} \frac{N(x'' \mid x', \sigma^2)\alpha_\epsilon^{\mathrm{U}}(x'' \mid x')}{1 - r^{\mathrm{U}}(x')} dx'$$

$$q^{\mathrm{DU}}(x'' \mid x^{(i)})\{1 - r^{\mathrm{D}}(x^{(i)})\} = \int N(x' \mid x^{(i)}, \sigma^2)\alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)}) \frac{N(x'' \mid x', \sigma^2)\alpha_\epsilon^{\mathrm{U}}(x'' \mid x')}{1 - r^{\mathrm{U}}(x')} dx'$$

$$= \int N(x^{(i)} \mid x', \sigma^2)\alpha_\epsilon^{\mathrm{U}}(x^{(i)} \mid x') \frac{N(x' \mid x'', \sigma^2)\alpha_\epsilon^{\mathrm{D}}(x' \mid x'')}{1 - r^{\mathrm{U}}(x')} dx'$$

$$= \int N(x' \mid x'', \sigma^2)\alpha_\epsilon^{\mathrm{D}}(x' \mid x'') \frac{N(x^{(i)} \mid x', \sigma^2)\alpha_\epsilon^{\mathrm{U}}(x^{(i)} \mid x')}{1 - r^{\mathrm{U}}(x')} dx'$$

$$= q^{\mathrm{DU}}(x^{(i)} \mid x'')\{1 - r^{\mathrm{D}}(x'')\}.$$

A down-up Metropolis-Hastings (cont.)

Useless due to the intractable integrations in the acceptance probability

$$\alpha^{\mathrm{DU}}(x'' \mid x^{(i)}) = \min\left\{1,\ \frac{\pi(x'')(1 - r^{\mathrm{D}}(x^{(i)}))}{\pi(x^{(i)})(1 - r^{\mathrm{D}}(x''))}\right\}$$
$$= \min\left\{1,\ \frac{\pi(x'')\int N(x' \mid x^{(i)}, \sigma^2)\alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)})dx'}{\pi(x^{(i)})\int N(x' \mid x'', \sigma^2)\alpha_\epsilon^{\mathrm{U}}(x' \mid x'')dx'}\right\}.$$

Is there any way to cancel out this ratio?

Auxiliary variable approach

If we explore a larger space, then there can be a way to cancel the ratio! (Møller et al., 2006)

An auxiliary variable A such that $\pi^{\mathrm{C}}(A \mid x)$ is well-defined.

- Joint target density: $\pi^{\mathrm{J}}(A, x)$
- Joint proposal density:

$$q^{\mathrm{J}}(A'', x'' \mid A^{(i)}, x^{(i)}) = q_1(x'' \mid A^{(i)}, x^{(i)}) q_2(A'' \mid x'', A^{(i)}, x^{(i)})$$

- Joint acceptance probability:

$$\alpha^{\mathrm{J}}(A'', x'' \mid A^{(i)}, x^{(i)}) = \min\left[1, \frac{\pi^{\mathrm{J}}(A'', x'') q^{\mathrm{J}}(A^{(i)}, x^{(i)} \mid A'', x'')}{\pi^{\mathrm{J}}(A^{(i)}, x^{(i)}) q^{\mathrm{J}}(A'', x'' \mid A^{(i)}, x^{(i)})}\right]$$

$$= \min\left[1, \frac{\pi(x'') \pi^{\mathrm{C}}(A'' \mid x'') q_1(x^{(i)} \mid A'', x'') q_2(A^{(i)} \mid x^{(i)}, A'', x'')}{\pi(x^{(i)}) \pi^{\mathrm{C}}(A^{(i)} \mid x^{(i)}) q_1(x'' \mid A^{(i)}, x^{(i)}) q_2(A'' \mid x'', A^{(i)}, x^{(i)})}\right]$$

Auxiliary variable approach (cont.)

Let's choose π^{C}, q_1, and q_2 to cancel out the intractable ratio.

$$q_1(x'' \mid A^{(i)}, x^{(i)}) = q^{\mathrm{DU}}(x'' \mid x^{(i)})$$
$$q_2(A'' \mid x'', A^{(i)}, x^{(i)}) = q^{\mathrm{D}}(A'' \mid x'')$$
$$\pi^{\mathrm{C}}(A'' \mid x'') = N(A'' \mid x'', \sigma^2)$$

Then,

$$\alpha^{\mathrm{J}}(A'', x'' \mid A^{(i)}, x^{(i)}) = \min\left[1,\ \frac{\pi(x'')N(A'' \mid x'', \sigma^2)q^{\mathrm{DU}}(x^{(i)} \mid x'')q^{\mathrm{D}}(A^{(i)} \mid x^{(i)})}{\pi(x^{(i)})N(A^{(i)} \mid x^{(i)}, \sigma^2)q^{\mathrm{DU}}(x'' \mid x^{(i)})q^{\mathrm{D}}(A'' \mid x'')}\right]$$

$$= \min\left[1,\ \frac{\pi(x'')N(A'' \mid x'', \sigma^2)\{1 - r^{\mathrm{D}}(x^{(i)})\}\frac{N(A^{(i)}|x^{(i)},\sigma^2)\alpha_\epsilon^{\mathrm{D}}(A^{(i)}|x^{(i)})}{1-r^{\mathrm{D}}(x^{(i)})}}{\pi(x^{(i)})N(A^{(i)} \mid x^{(i)}, \sigma^2)\{1 - r^{\mathrm{D}}(x'')\}\frac{N(A''|x'',\sigma^2)\alpha_\epsilon^{\mathrm{D}}(A''|x'')}{1-r^{\mathrm{D}}(x'')}}\right]$$

$$= \min\left[1,\ \frac{\pi(x'')\alpha_\epsilon^{\mathrm{D}}(A^{(i)} \mid x^{(i)})}{\pi(x^{(i)})\alpha_\epsilon^{\mathrm{D}}(A'' \mid x'')}\right] = \min\left[1,\ \frac{\pi(x'')\min\{1, \frac{\pi(x^{(i)})+\epsilon}{\pi(A^{(i)})+\epsilon}\}}{\pi(x^{(i)})\min\{1, \frac{\pi(x'')+\epsilon}{\pi(A'')+\epsilon}\}}\right].$$

Algorithmic specification

The DUMH algorithm is composed of eight steps for each iteration.

Set initial values $A^{(0)}$ and $x^{(0)}$. For $i = 0, 1, \ldots, n-1$,

Step 1: Sample $x' \sim N(x' \mid x^{(i)}, \sigma^2)$ and $u_1 \sim \text{Unif}(0, 1)$.

Step 2: Repeat Step 1 until $u_1 < \alpha_\epsilon^{\text{D}}(x' \mid x^{(i)})$ for forced downhill move.

Step 3: Sample $x'' \sim N(x'' \mid x', \sigma^2)$ and $u_2 \sim \text{Unif}(0, 1)$.

Step 4: Repeat Step 3 until $u_2 < \alpha_\epsilon^{\text{U}}(x'' \mid x')$ for forced uphill move.

Step 5: Sample $A'' \sim N(A'' \mid x'', \sigma^2)$ and $u_3 \sim \text{Unif}(0, 1)$.

Step 6: Repeat Step 5 until $u_3 < \alpha_\epsilon^{\text{D}}(A'' \mid x'')$ for forced downhill move.

Step 7: Sample $u_4 \sim \text{Unif}(0, 1)$.

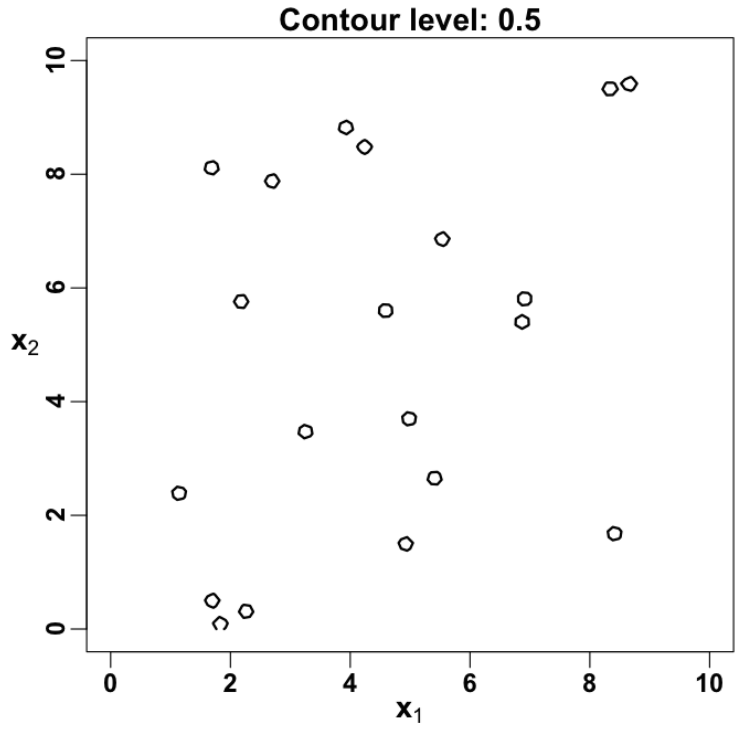
Step 8: Accept (A'', x'') as $(A^{(i+1)}, x^{(i+1)})$ if $u_4 < \alpha^{\text{J}}(A'', x'' \mid A^{(i)}, x^{(i)})$,
or otherwise set $(A^{(i+1)}, x^{(i+1)})$ to $(A^{(i)}, x^{(i)})$.

Example 1

A mixture of 20 bivariate Gaussian distributions (Kou et al., 2006)

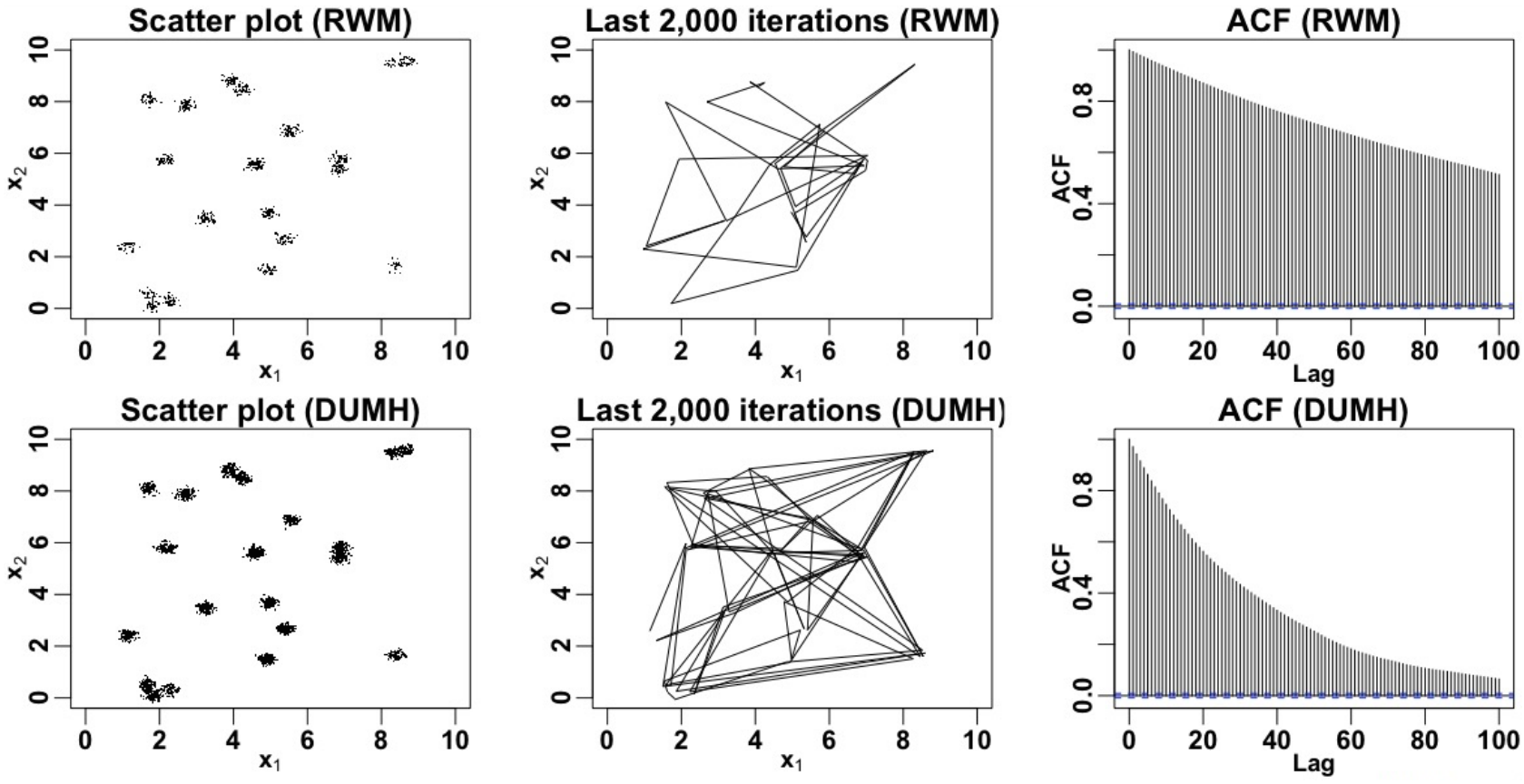
$$\pi(x) \propto \frac{1}{20} \sum_{i=1}^{20} \frac{1}{2\pi\sigma_i^2} \exp\left(-\frac{1}{2\sigma_i^2} (x - \mu_i)^\top (x - \mu_i) \right),$$

where $x = (x_1, x_2)^\top$, $\sigma_i = 0.1$, and the 20 mean vectors, $(\mu_1, \mu_2, \ldots, \mu_{20})^\top$, are given in Kou et al. (2006).



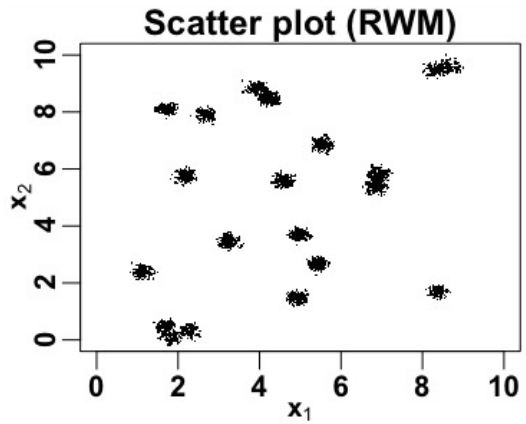
Example 1 (cont.): RWM vs DUMH

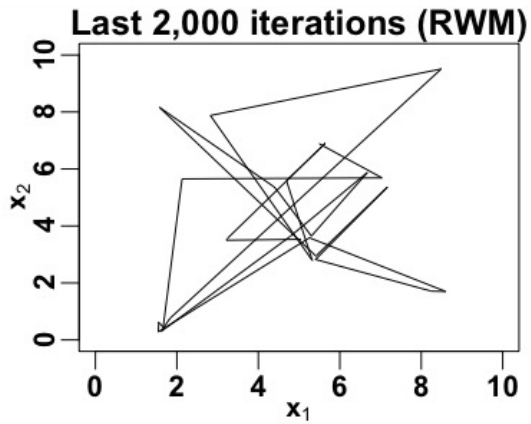
50,000 iterations after discarding the first 50,000 iterations.

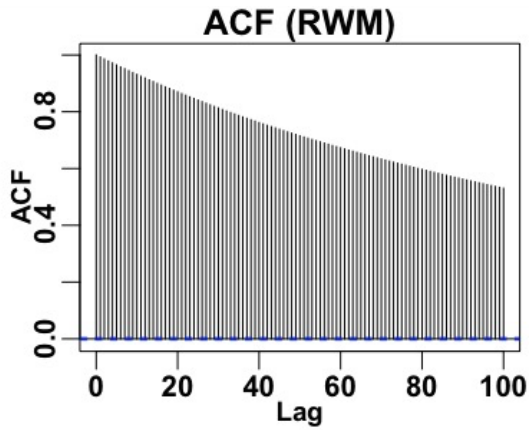


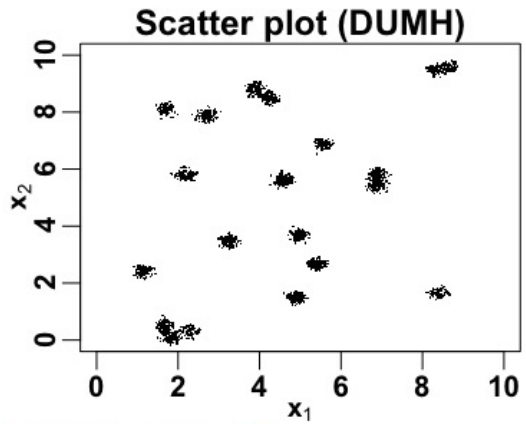
	σ	Accept. Rate	Sample size	CPU time (sec.)
RWM	4	0.015	50,000	535
DUMH	4	0.045	50,000	2,758

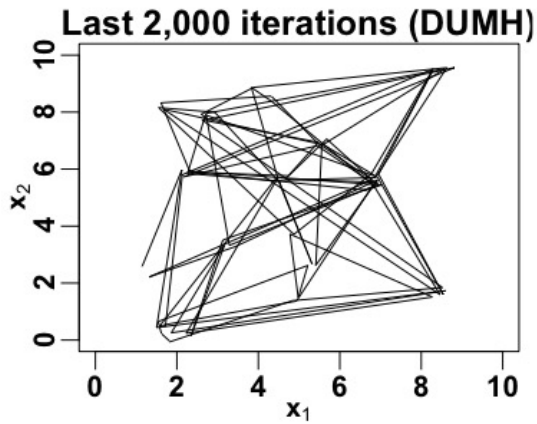
Example 1 (cont.): RWM vs DUMH

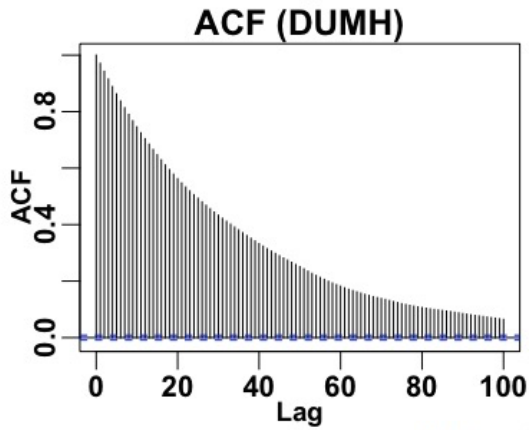








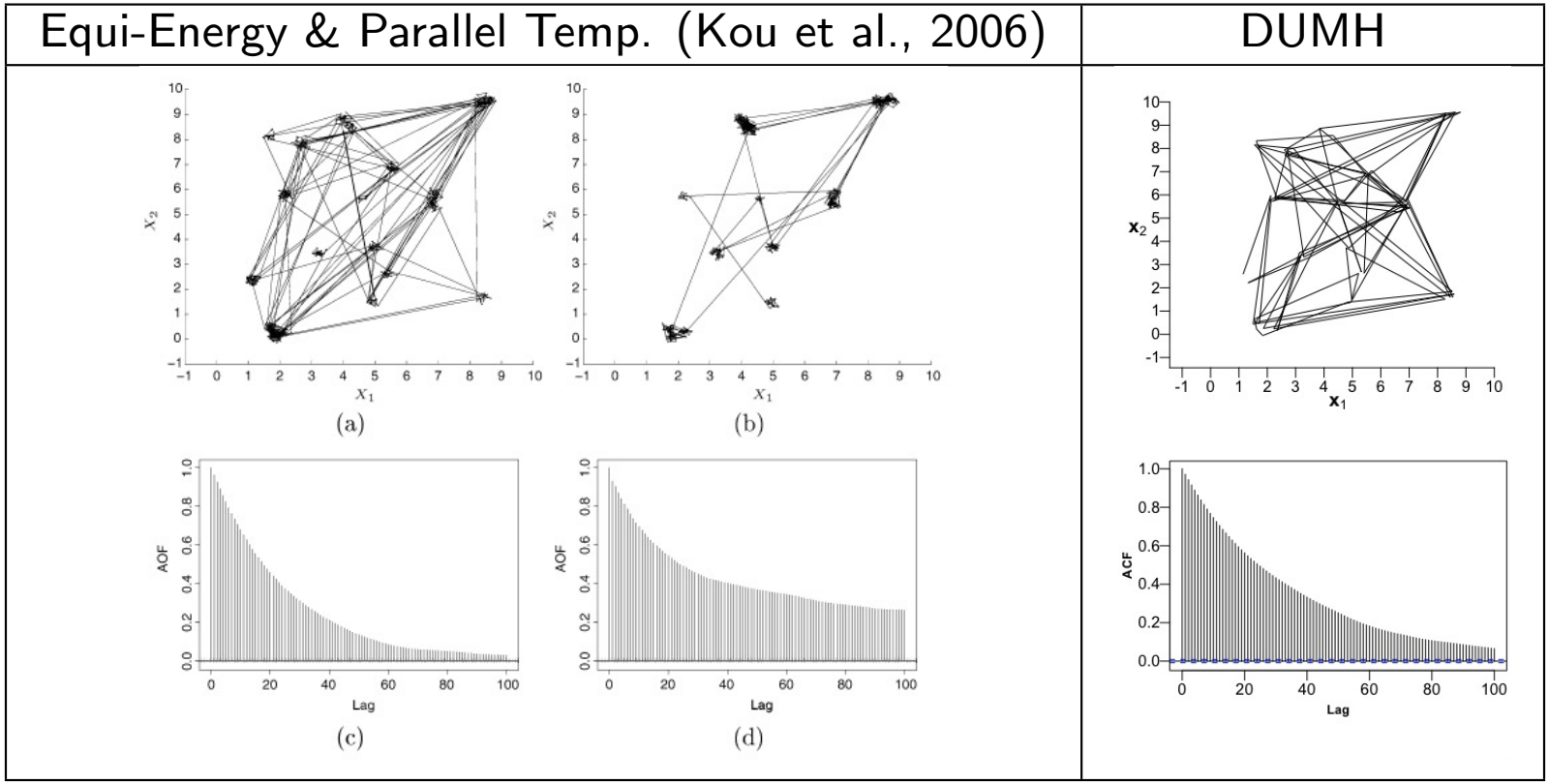




	σ	Accept. Rate	Sample size	CPU time (sec.)
RWM	4	0.013	257,757	2,663
DUMH	4	0.046	50,000	2,758

"Do (DU) MH instead of the naive RWM!"

Example 1 (cont.): EE vs PT vs DUMH



	$E(X_1)$	$E(X_2)$	$E(X_1^2)$	$E(X_2^2)$
True value	4.478	4.905	25.605	33.920
DUMH	4.5014 (0.095)	4.8847 (0.141)	25.5858 (0.977)	33.6423 (1.371)
EE	4.5019 (0.107)	4.9439 (0.139)	25.9241 (1.098)	34.4763 (1.373)
PT	4.4185 (0.170)	4.8790 (0.283)	24.9856 (1.713)	33.5966 (2.867)
$\frac{\text{MSE(EE)}}{\text{MSE(DUMH)}}$	1.26	1.03	1.37	1.12
$\frac{\text{MSE(PT)}}{\text{MSE(DUMH)}}$	3.39	3.98	3.47	4.25

Better than the EE and PT in terms of accuracy! (Do MH!)

Example 2: Quasar Q0957+561

MH within Gibbs sampler for $p(\mathbf{X}(\mathbf{t}^{\boldsymbol{\Delta}}), \Delta, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \text{Data})$.

Step 1: Sample $(\mathbf{X}^{(l)}(\mathbf{t}^{\boldsymbol{\Delta}^{(l)}}), \Delta^{(l)}) \sim p(\mathbf{X}(\mathbf{t}^{\boldsymbol{\Delta}}), \Delta \mid \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)})$

$= p(\mathbf{X}(\mathbf{t}^{\boldsymbol{\Delta}}) \mid \Delta, \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}) \times p(\Delta \mid \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)})$ by M-H

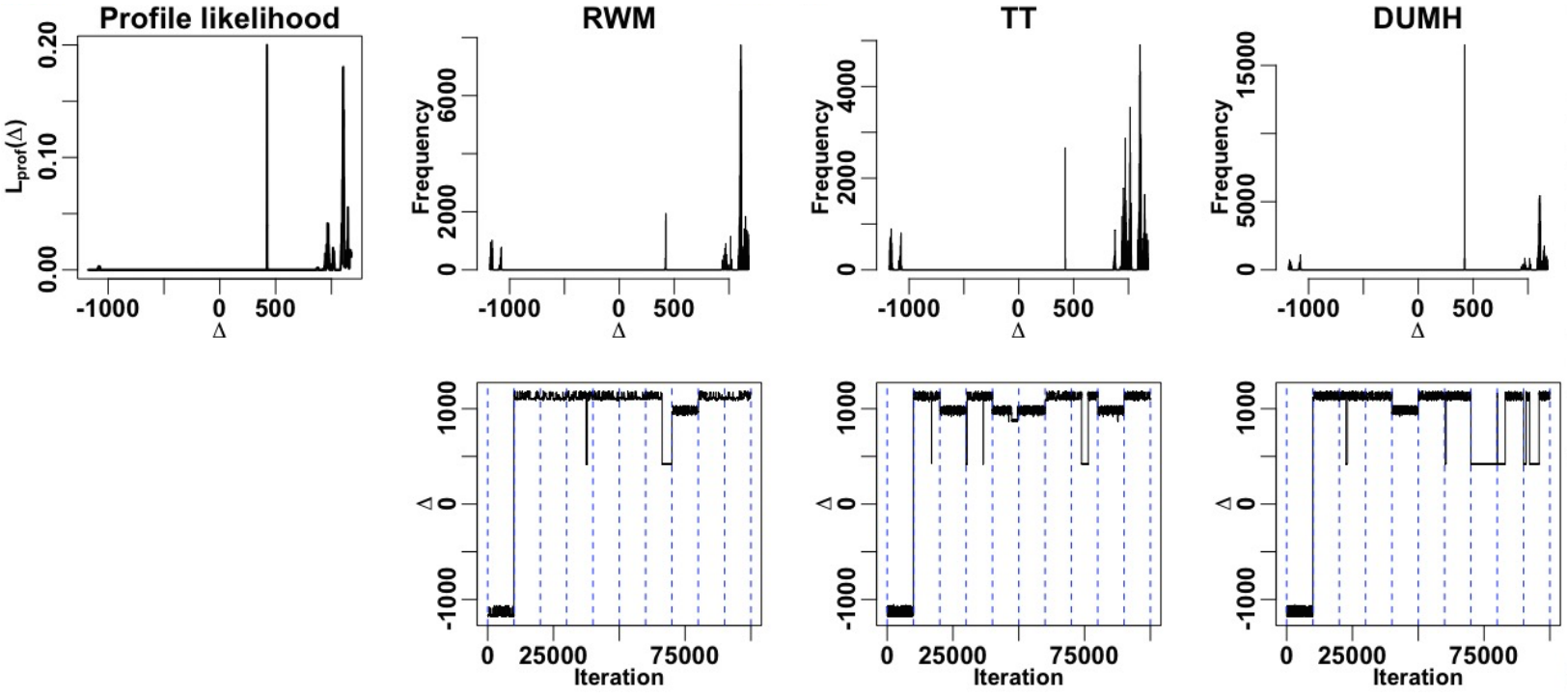
Step 2: Sample $\boldsymbol{\beta}^{(l)} \sim p(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(l-1)}, \mathbf{X}^{(l)}(\mathbf{t}^{\boldsymbol{\Delta}^{(l)}}), \Delta^{(l)})$

Step 3: Sample $\boldsymbol{\theta}^{(l)} \sim p(\boldsymbol{\theta} \mid \mathbf{X}^{(l)}(\mathbf{t}^{\boldsymbol{\Delta}^{(l)}}), \Delta^{(l)}, \boldsymbol{\beta}^{(l)})$

RWM, tempered transition (Neal, 1996), and DUMH applied to $p(\Delta \mid \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)})$.

Example 2: Quasar Q0957+561

Each chain (among 10 chains) has 10,000 samples after discarding the first 5,000.



	σ	Accept. Rate	Sample size	CPU time/chain
RWM	600	0.018	100,000	48
TT	300	0.041	100,000	903
DUMH	1100	0.053	100,000	493

Discussion

How do we fairly evaluate various MCMC samplers for multimodality?

- How do we measure the time for the off-line tuning work involved in all the temperature-based samplers?
- How do we theoretically evaluate the MCMC samplers?