

Efficient strategy of MCMC in
high-dimension and its application to
diffusion processes

Kengo Kamatani (Osaka Univ. and CREST, JST)

Mar 2015 at LeMans

1. New algorithm:

- **Markov chain Monte Carlo (MCMC)** produces a Markov chain X_0, \dots, X_{M-1} with a given invariant probability measure P . If it is ergodic, we have

$$M^{-1} \sum_{m=0}^{M-1} f(X_m) \rightarrow P(f) = \int f(x)P(dx).$$

We can approximate $P(f)$ by the empirical average.

- MCMC \ni RWM, Gibbs, MALA, Slice Sampler, HMC etc.
- Almost all MCMC satisfies **reversibility**, i.e., if $X_0 \sim P(dx)$,

$$\mathcal{L}(X_0, X_1, \dots, X_M) = \mathcal{L}(X_M, X_{M-1}, \dots, X_0).$$

1-a. RWM Algorithm: Let $P(dx) = p(x)dx$ be the target on \mathbb{R}^d .

1. Generate $x^* = x + w$ where $w \sim N_d(0, \sigma^2 I_d) = \Gamma_d$.
2. **Accept** x^* as the next state with probability $\alpha(x, x^*)$, and otherwise, **discard** x^* , where

$$\alpha(x, x^*) = \min \left\{ 1, \frac{p(x^*)}{p(x)} \right\}.$$

Proposal kernel [x to x^*] is reversible with respect to **the uniform distribution on \mathbb{R}^d** .

1-b. pCN Algorithm: Fix $\rho \in (0, 1)$. For $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ let $\|x\| = (\sum_{i=1}^d x_i^2)^{1/2}$.

1. Generate $x^* = \rho^{1/2}x + (1 - \rho)^{1/2}w$ where $w \sim N_d(0, I_d)$.

2. Accept x^* with probability $\alpha(x, x^*)$ where

$$\alpha(x, x^*) = \min \left\{ 1, \frac{p(x^*)\phi(x)}{p(x)\phi(x^*)} \right\}$$

where ϕ is the pdf of $N_d(0, I_d)$.

Proposal kernel [x to x^*] is reversible with respect to $N_d(0, I_d)$.

1-c. MpCN Algorithm (**New method**):

1. Generate $r \sim \text{Gamma}(d/2, \|x\|^2/2)$.
2. Generate $x^* = \rho^{1/2}x + (1 - \rho)^{1/2}r^{-1/2}w$ where $w \sim N_d(0, I_d)$.
3. Accept x^* with probability $\alpha(x, x^*)$ where

$$\alpha(x, x^*) = \min \left\{ 1, \frac{p(x^*) \|x\|^{-d}}{p(x) \|x^*\|^{-d}} \right\}.$$

Proposal kernel [x to x^*] is reversible with respect to $\|x\|^{-d} dx$.

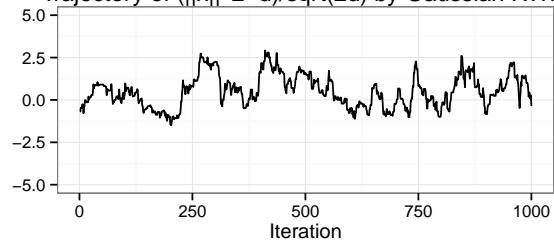
Note for application to Bayesian inference for complicated models

- $x = \theta$ and $P(dx) = P(d\theta|X^n) = p(\theta|X^n)d\theta$.
- For many advanced MCMC methods, we need to calculate $(\log p(x))' \approx$ (score function) in **each iteration** (ex. **10^6 times!**). Sometimes we also need to calculate $(\log p(x))''$.
- Previous three methods are nice in this point of view as long as the performance is nice.

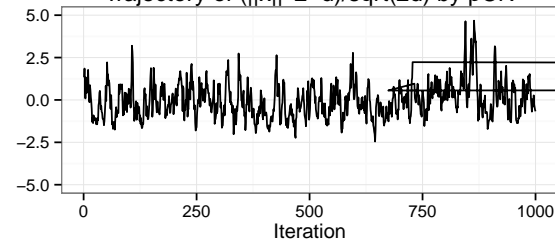
2. Application

2-a. Toy examples; $P(dx) = \text{standard normal distribution}$

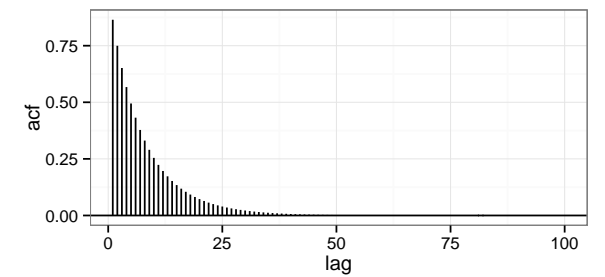
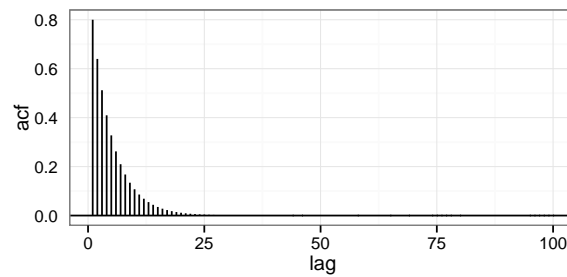
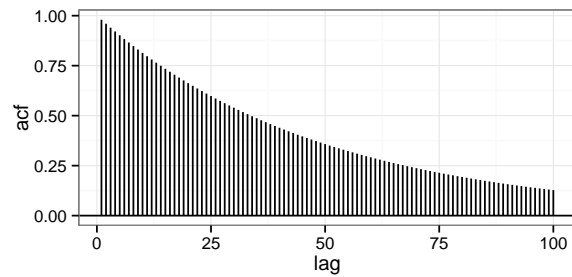
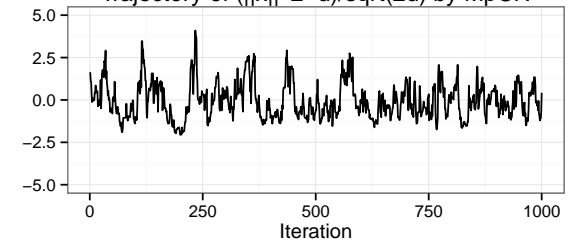
Trajectory of $(\|x\|^2 - d)/\sqrt{2d}$ by Gaussian RWM



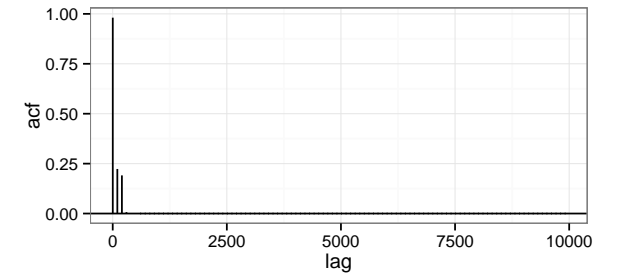
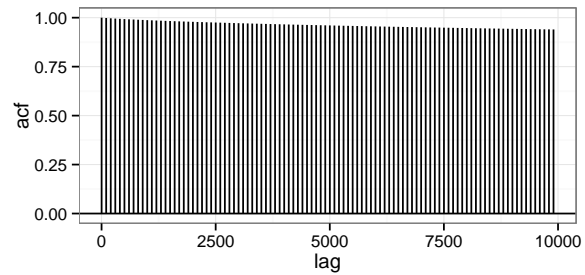
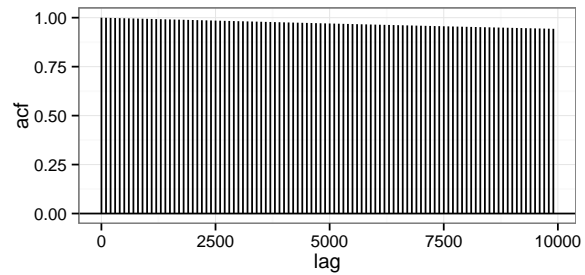
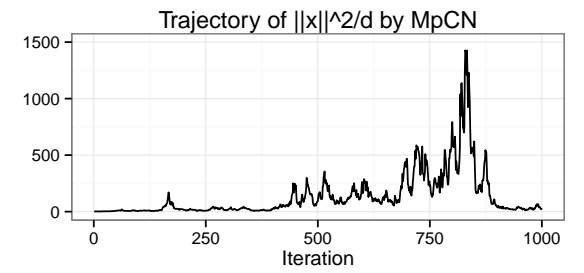
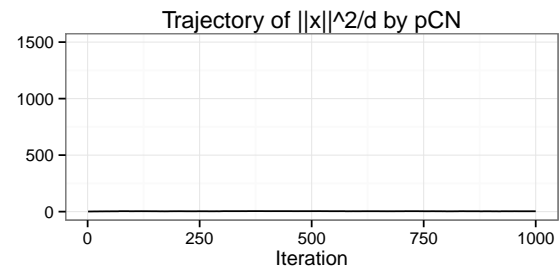
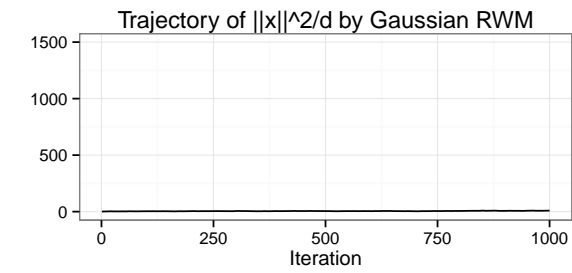
Trajectory of $(\|x\|^2 - d)/\sqrt{2d}$ by pCN



Trajectory of $(\|x\|^2 - d)/\sqrt{2d}$ by MpCN



2-a. Toy examples; t -distribution



2-b. Stochastic processes

Realistic examples. R with **Yuima** package. We consider some Bayesian parameter estimation for discretely observed stochastic processes.

Note

- **LA** (Likelihood analysis) is not available and we treat **QLA** (quasi-LA).
- QLA has been studied extensively. See Yoshida [9] and references therein.

Consider

$$dX_t = a(X_t, \theta)dt + b(X_t)dW_t; X_0 = 2, t \in [0, T]$$

where

$$a(x, \theta) = \theta_1 - \theta_2 x + 2 \sin(\theta_3 x), \quad b(x) = \frac{0.5 + x^2}{1 + 0.3x^2}.$$

$N = 5000, T = 250.$

$$P(d\theta|X^N) \propto \exp\left(-\frac{1}{2}\left(\sum_{n=1}^N \frac{(X_{nh} - X_{(n-1)h} - a(X_{(n-1)h}, \theta)h)^2}{hb(X_{(n-1)h})^2}\right)\right) P(d\theta)$$

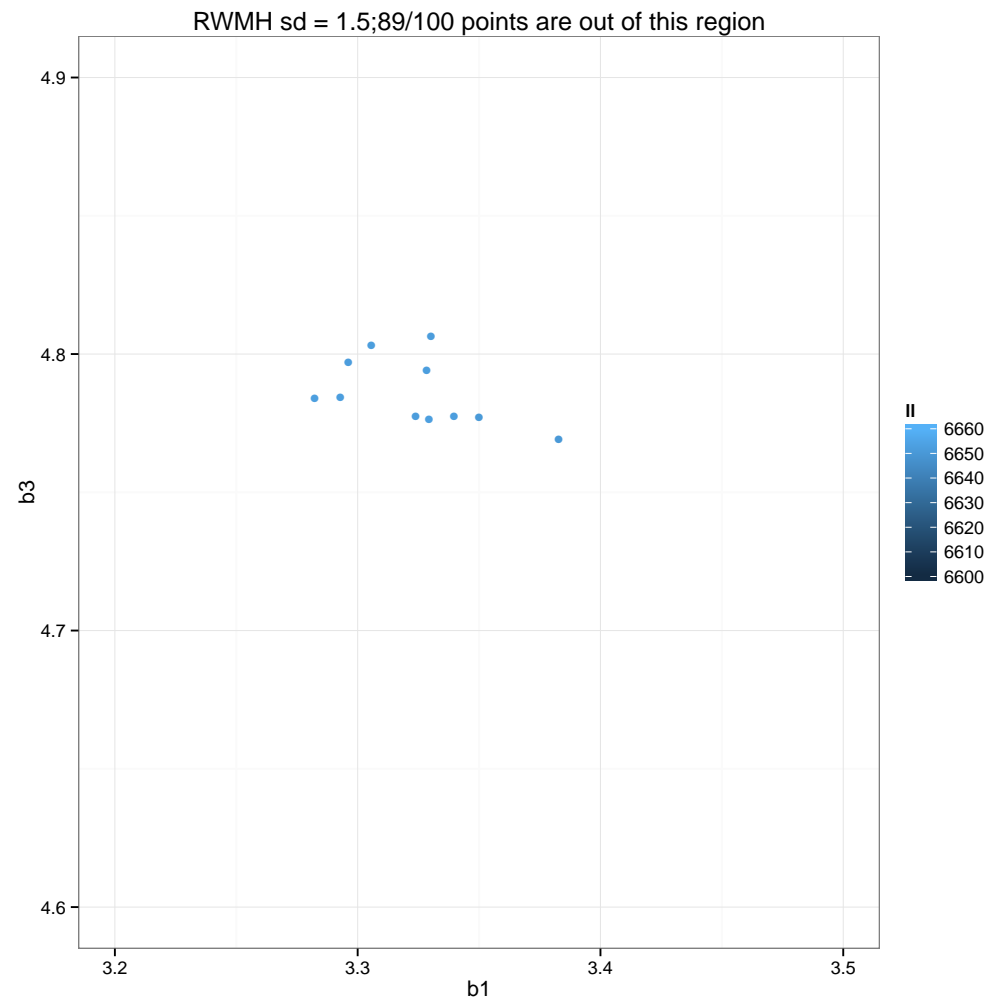
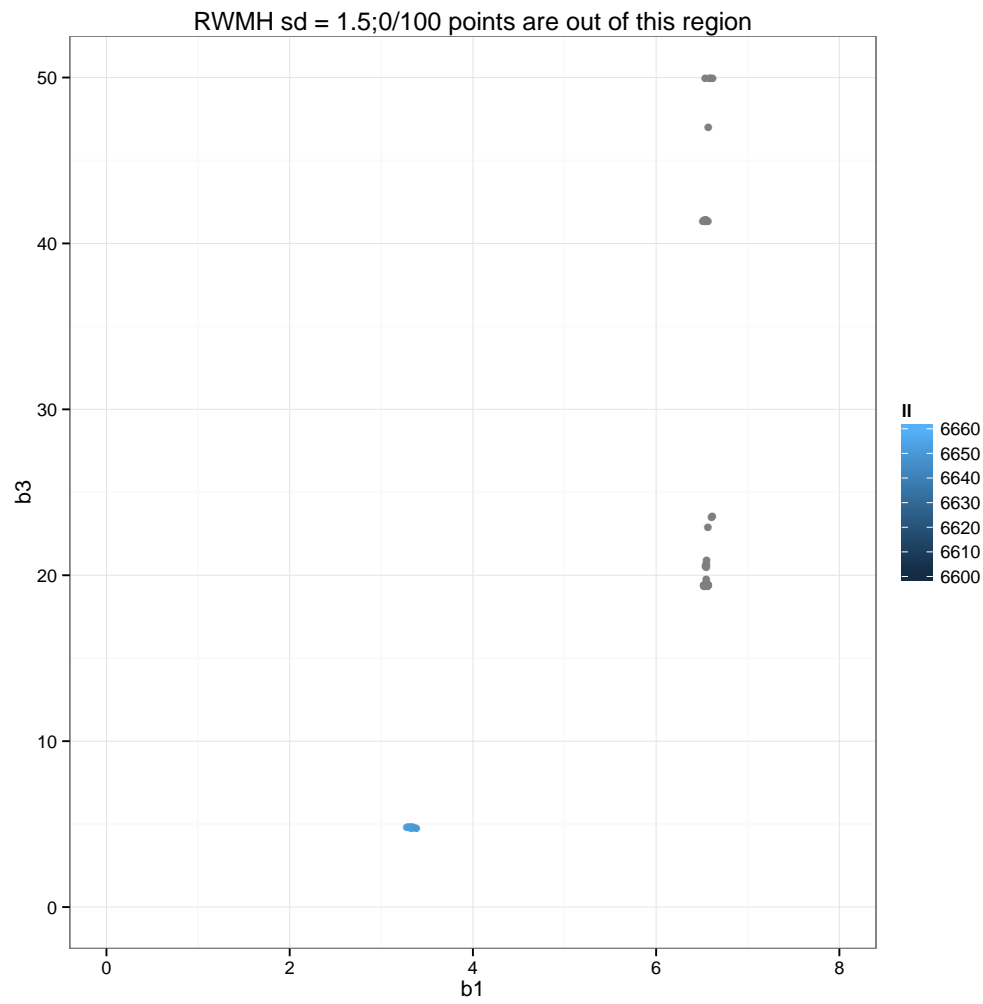
where $h = T/N$ ($Nh^3 = 0.625$). True is $\theta = (3, 7, 5)$.

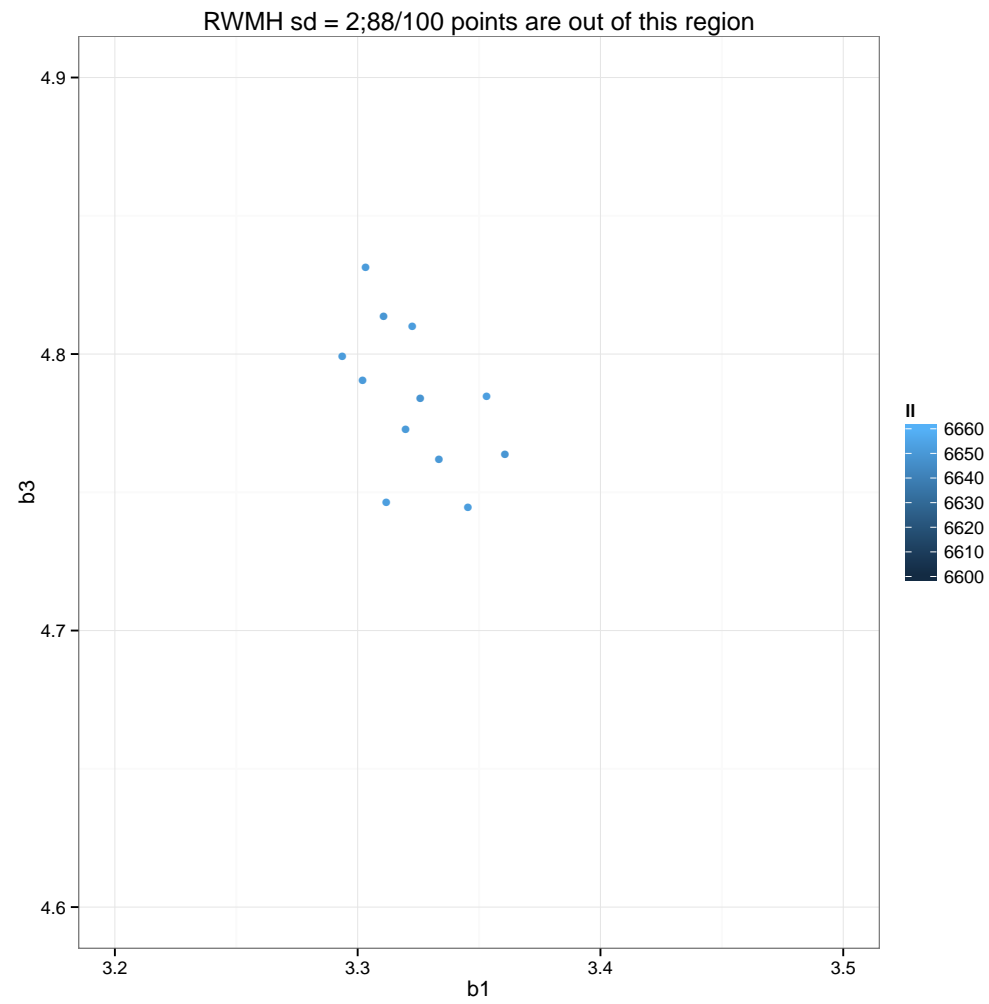
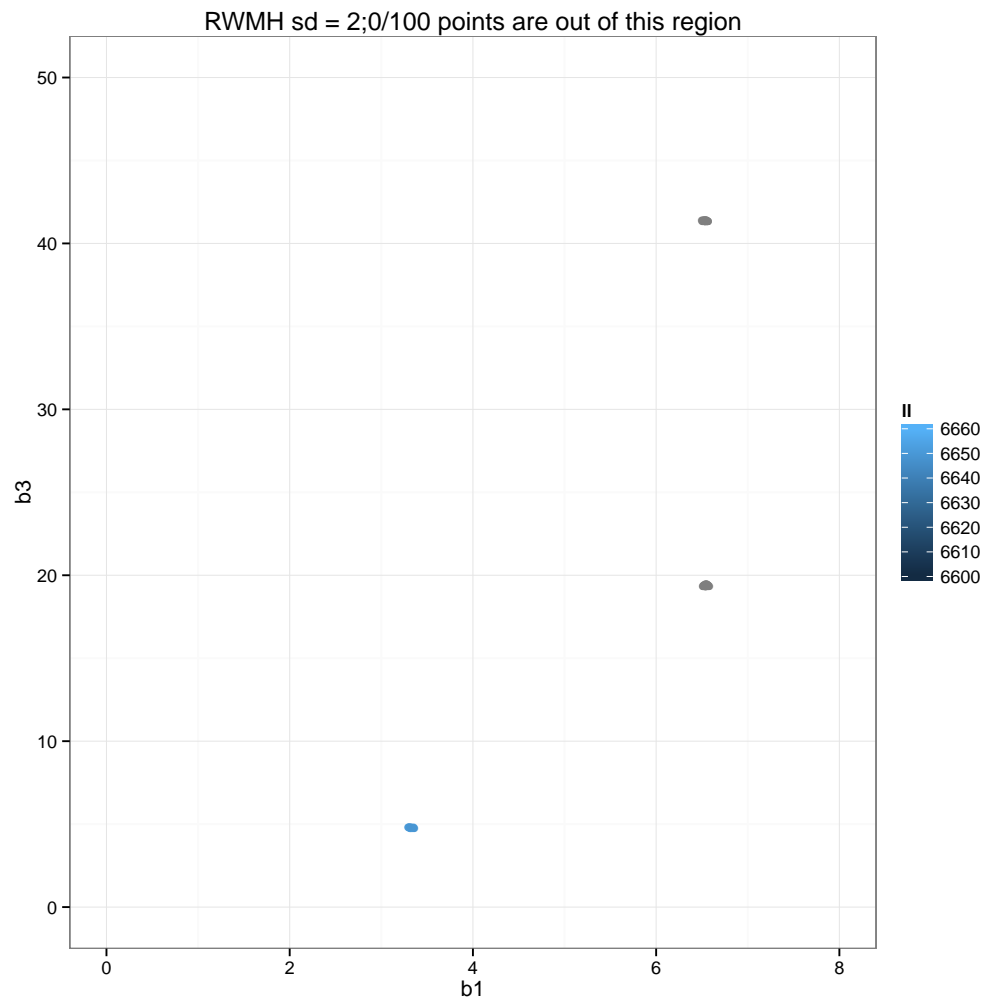
- Generate discrete observation X^N from the model for a true parameter.
- Run each MCMC for $M = 10^5$ iteration from 100 different starting points.

- Plot empirical average for each 100 trials to approximate

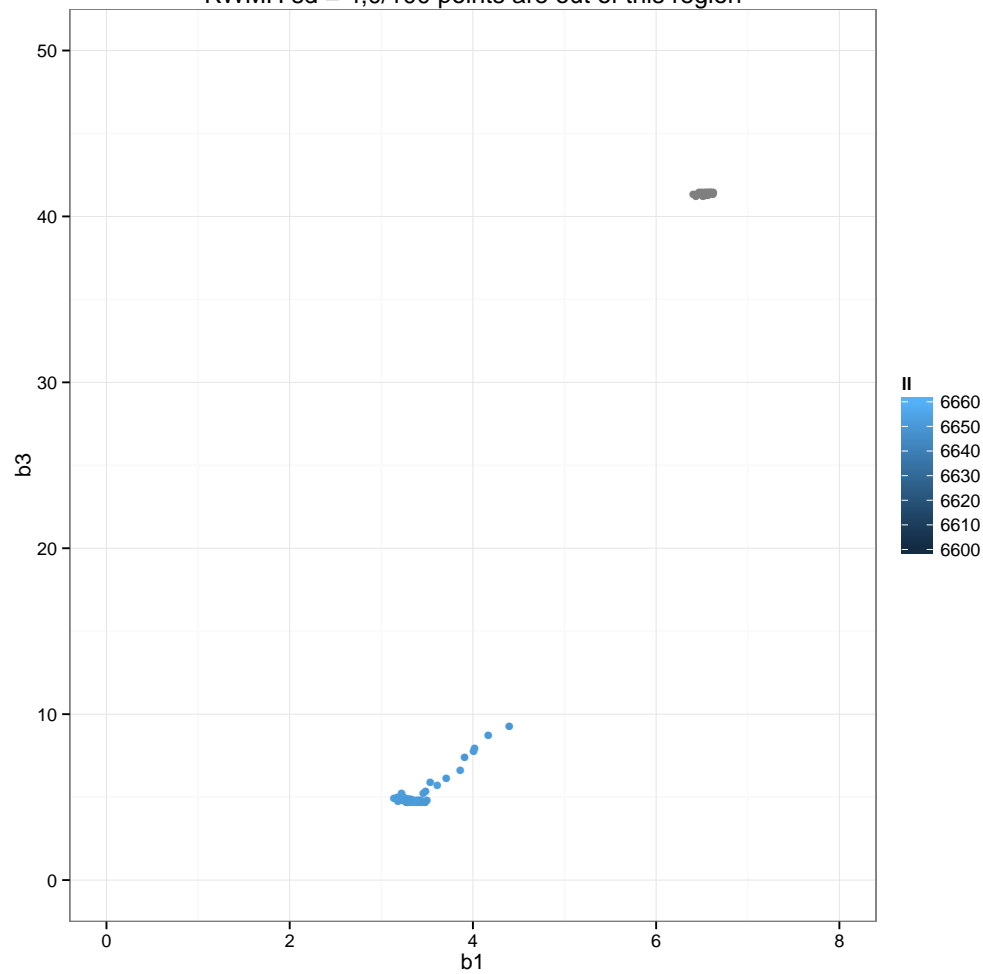
$$\int \theta P(d\theta | X^N).$$

- We compare RWM ($\sigma = 1.5, 2, 4$), pCN and MpCN.

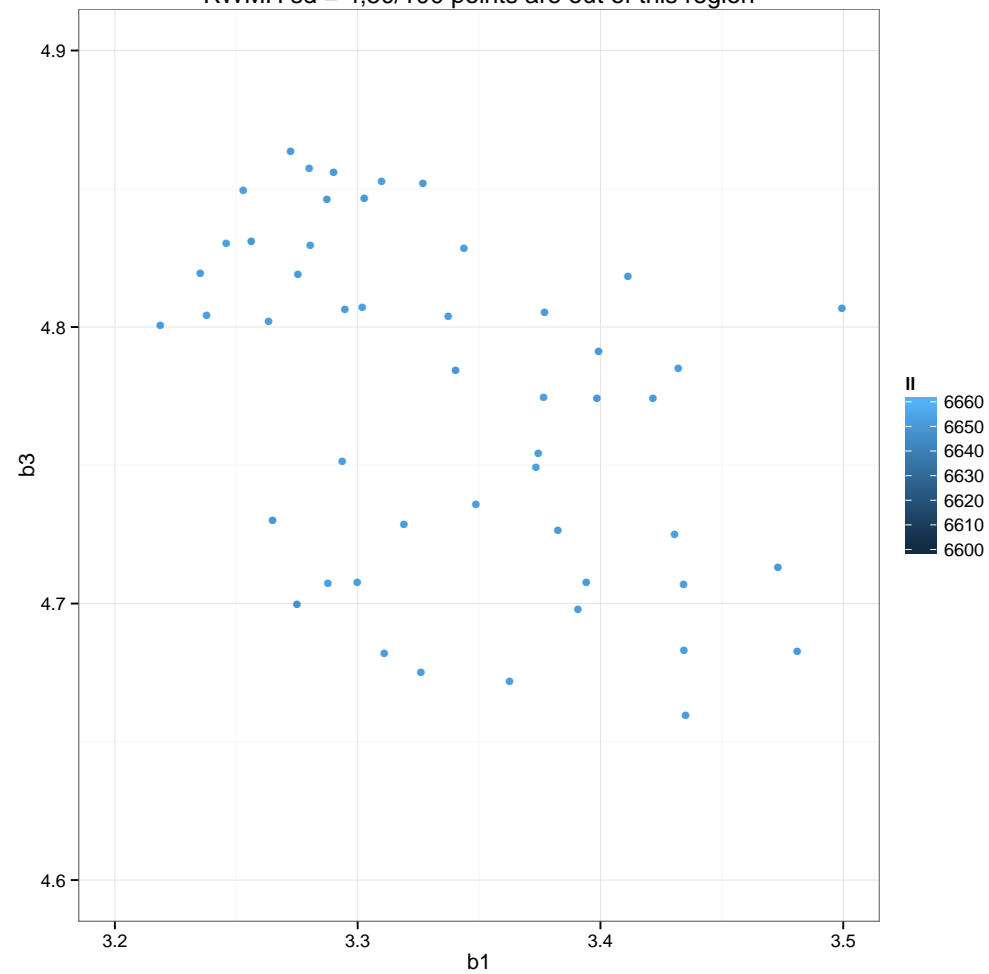


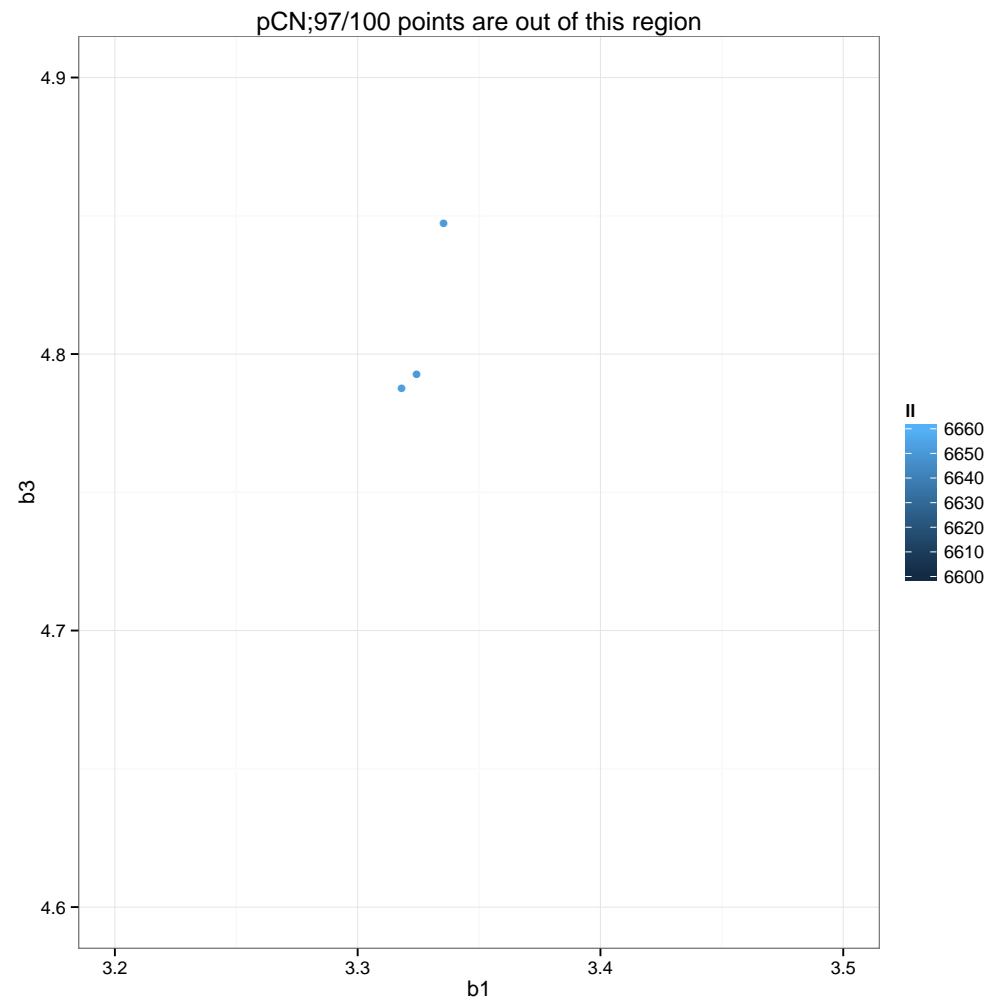
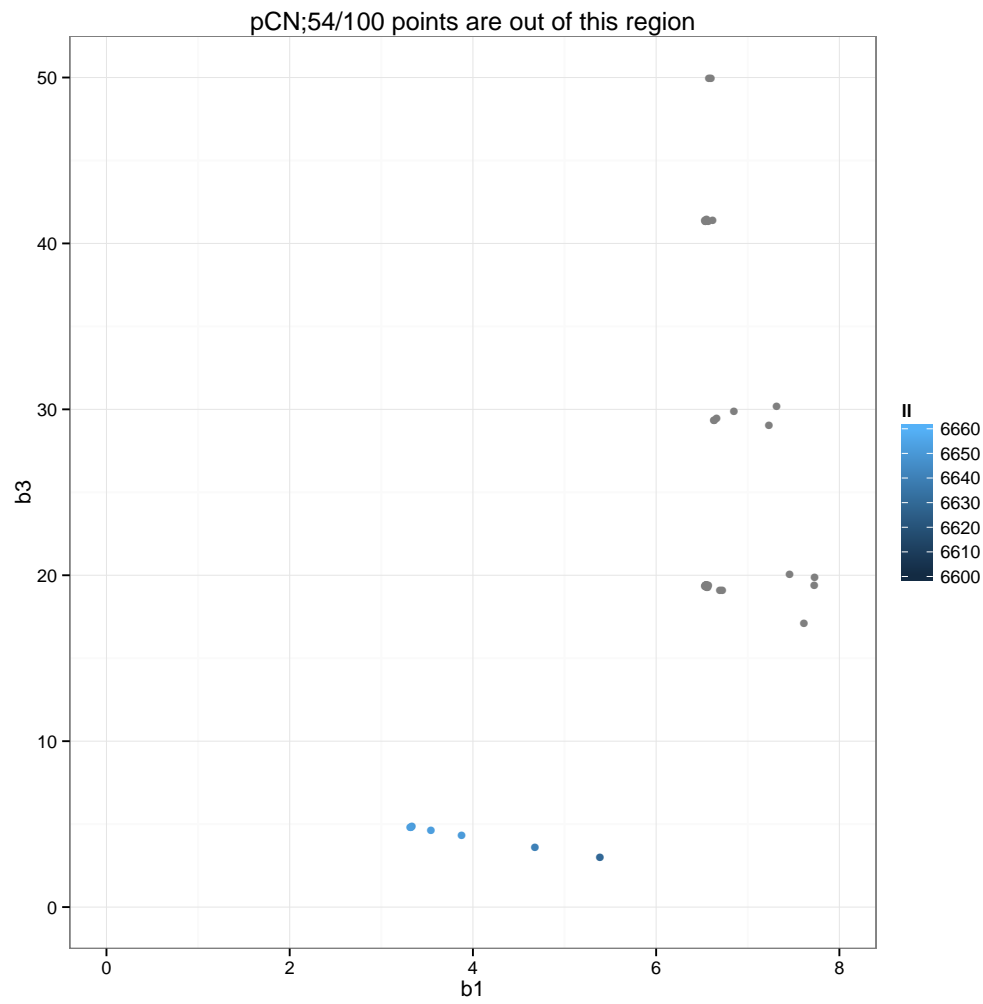


RWMH sd = 4; 0/100 points are out of this region

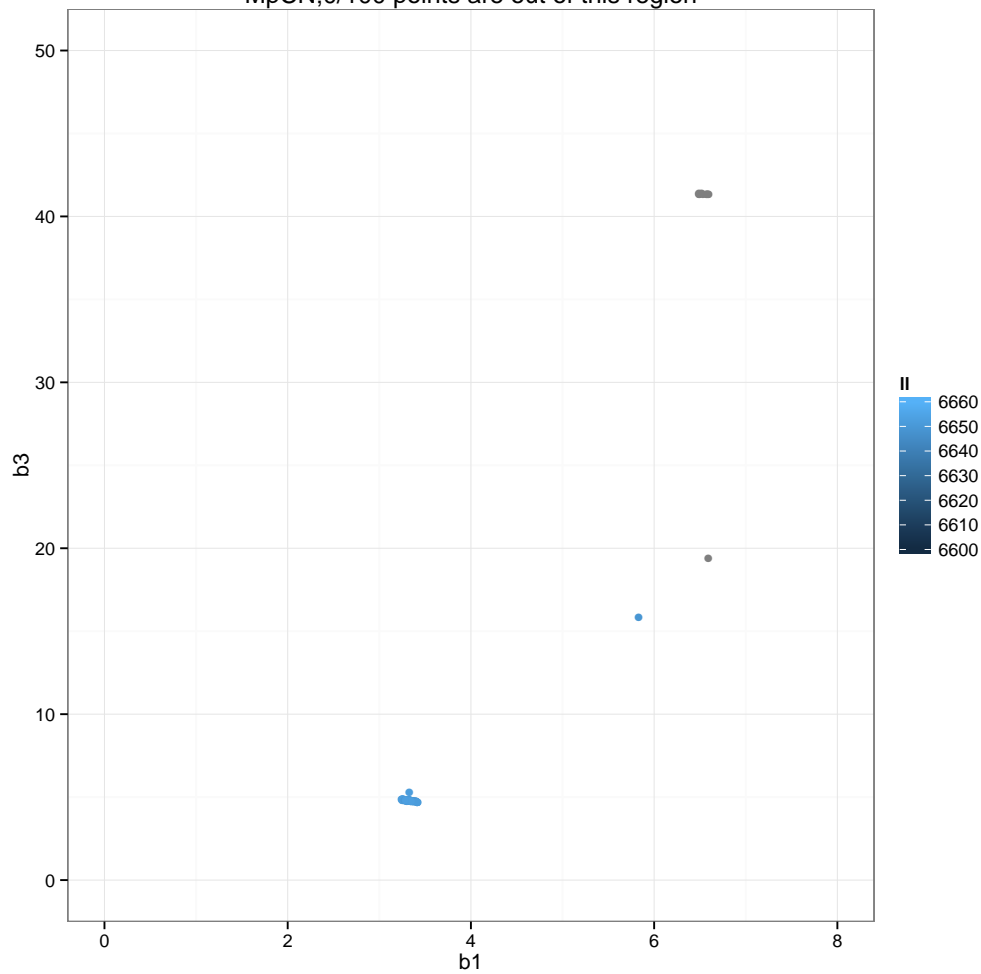


RWMH sd = 4; 50/100 points are out of this region

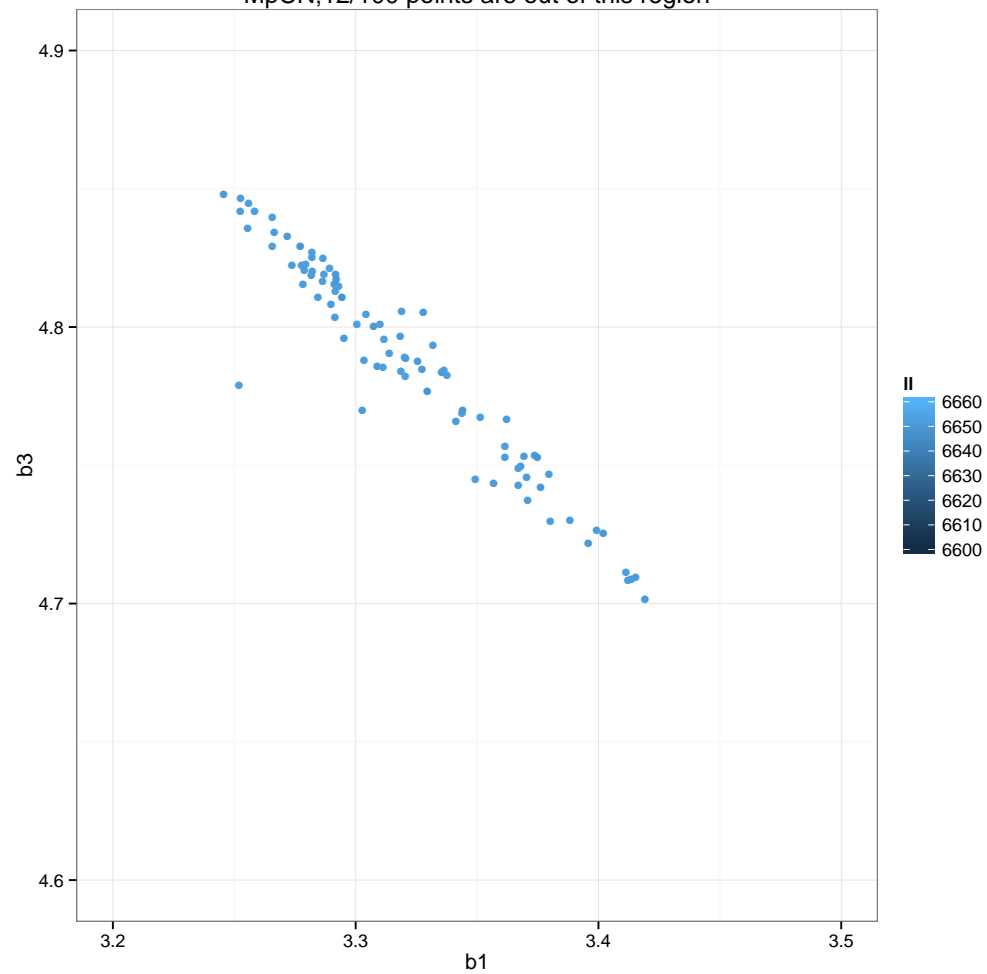




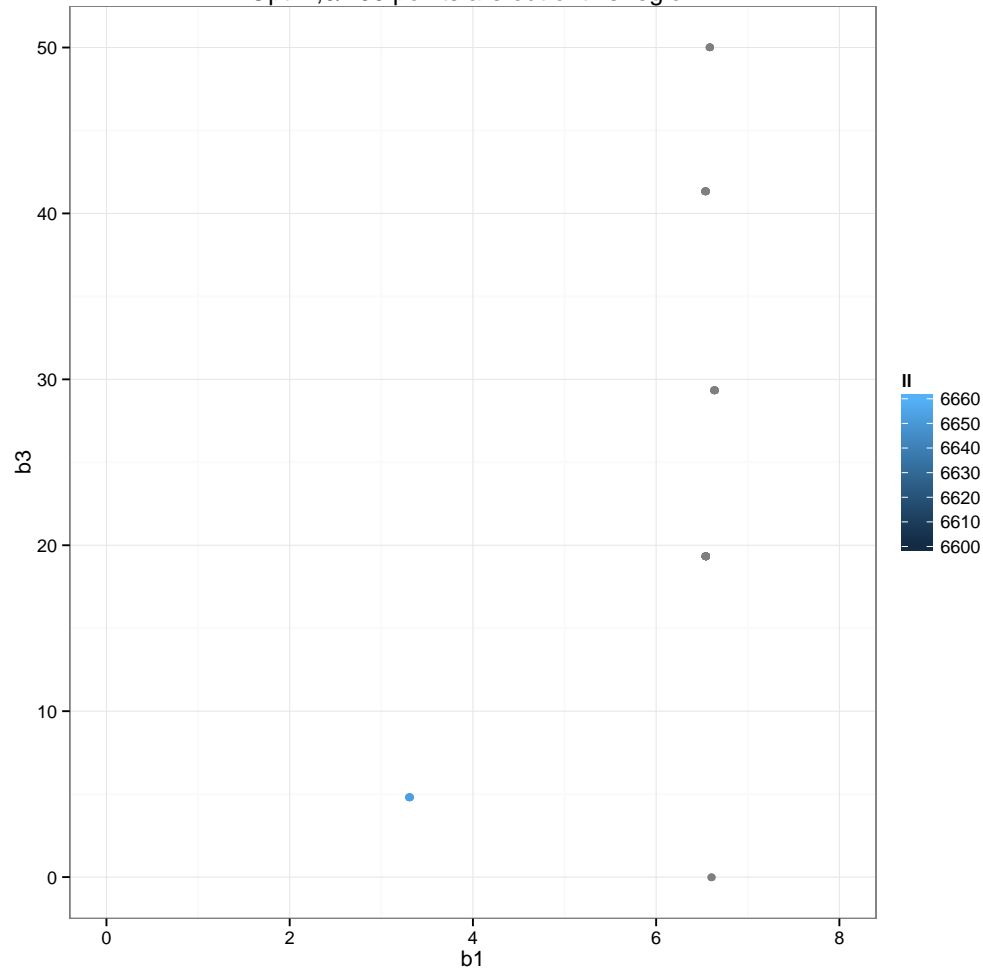
MpCN;0/100 points are out of this region



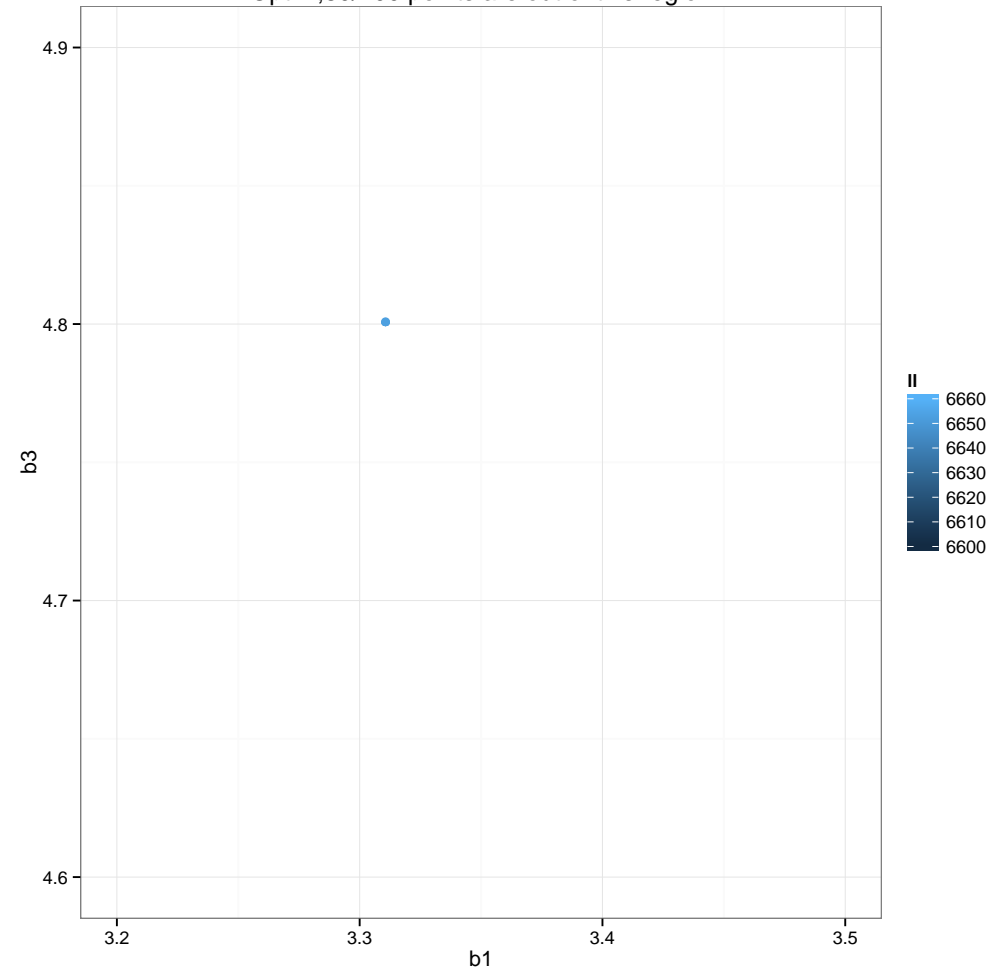
MpCN;12/100 points are out of this region



Optim;0/100 points are out of this region



Optim;89/100 points are out of this region



MpCN

- (Essentially) No tuning parameter.
- No derivative.
- Good performance.

3. Theoretical results

- We study (HDA) High-Dimensional asymptotics for MCMC.
- HDA is strong assumption \Rightarrow strong conclusion type framework.
- HDA was developed by Gelman et al. [5] and Roberts et al. [8] (RGG97).

- Introduce time scaling $t \mapsto [dt]$, where $[x]$ is the integer part of x , and consider

$$X_{[dt]}^d.$$

- Introduce projection $\pi_E(x) = (x_i)_{i \in E}$ for $E \subset \{1, \dots, d\}$ where $x = (x_1, \dots, x_d)$. Ex.

$$\pi_{\{3,5,10\}}(x) = (x_3, x_5, x_{10}) \text{ if } E = \{3, 5, 10\}$$

and consider

$$Y_t^d := \pi_{\{1\}}(X_{[dt]}^d).$$

- Introduce proposal scaling

$$\sigma^2 = l^2/d.$$

Theorem (GGR97). $Y^d \Rightarrow Y$ where

$$dY_t = h(l) \frac{(\log f)'(Y_t)}{2} dt + \sqrt{h(l)} dW_t$$

where

$$h(l) = 2l^2 \Phi \left(-\frac{l\sqrt{I}}{2} \right),$$
$$I = \int \{(\log f)'(x)\}^2 f(x) dx.$$

Interpretation of RGG97's

- The rate of convergence is d . Thus the number of iteration should be proportional to d .
- For the limit process Y , the convergence rate is determined by $h(l)$.
- The function $h(l)$ is maximised if the average acceptance probability is approximately 0.23.

The result gives a criterion for constructing a good RWM.

After the seminal paper RGG97, there are many studies for the generalization of the result.

Generalization of P_d Non i.i.d. Bédard [2], Perturbation of Gaussian Beskos et al. [4], etc.

Better convergence rate Metropolis adjusted Langevin algorithm (MaLa, $d^{1/3}$ Roberts and Rosenthal [7]), Hybrid Monte Carlo ($d^{1/4}$ Beskos et al. [3]), Metropolis-Coupled MCMC Atchadé et al. [1], etc.

Our plan:

- (perturbation of) Gaussian = ideal situation. **Heavy-tail \approx realistic, non-ideal situation.** We want to know the rate of convergence (time scaling). It is d for RWM for Gaussian case.
- We want to construct MCMC, which works well for a difficult target distribution.
- We only consider a special class of heavy-tailed distribution. By this, we can apply **Stein's techniques** and **Malliavin calculus**.

3-b. Setting

- P_d is a scale mixture of the Gaussian distribution;

$$P_d = \mathcal{L}(X_0^d), \text{ where } X_0^d|Y \sim N_d(0, I_d Y), Y \sim Q(dy).$$

- The class of $P_d \ni N_d(0, I_d)$, Student t -distribution and the stable distribution.

If P_d is heavy-tailed, the rate of convergence is difficult to define.

- (Usual) Consistency

$\xi^d = (\xi_m^d)_{m \in \mathbb{N}_0}$ ($d \in \mathbb{N}$) is consistent if

$$\frac{1}{M} \sum_{m=0}^{M-1} f(\xi_m^d) - \int f(x) \Pi_d(dx) = o_{\mathbb{P}}(1) \quad (M, d \rightarrow \infty)$$

for any bounded continuous function f (K. 2014 [6]).

- Since the dimension grows as $d \rightarrow \infty$, it is not suitable the current study. We make a generalization of this definition.

- For any bounded continuous function $f : \mathbb{R}^k \rightarrow \mathbb{R}$ and for any sequence $E_d^k \subset \{1, \dots, d\}$ s.t. $\#E_d^k = k$,

$$\frac{1}{M_d} \sum_{m=0}^{M_d-1} f \circ \pi_{E_d^k}(X_m^d) - \int f \circ \pi_{E_d^k}(x) P_d(dx) = o_{\mathbb{P}}(1)$$

for any $M_d \rightarrow \infty$ then we call that $(X^d)_d$ is consistent.

- If above satisfies all M_d such that $M_d/T_d \rightarrow \infty$, then we call that T_d is the **convergence rate**.
- This is just a formalisation of the rate of convergence used in HDA community.

3-c. Gaussian case; $P_d = N_d(0, I_d)$

- Let $\mu_k(\sigma) = \mathbb{E}[|\xi|^k \exp(-\xi^+)]$ for $\xi \sim N(\frac{\sigma^2}{2}, \sigma^2)$, and $\xi^+ = \max\{0, \xi\}$.

- Let

$$r_d(x) = \sqrt{d} \left(\frac{\|x\|^2}{d} - 1 \right) \quad (x \in \mathbb{R}^d).$$

Proposition (Gaussian RWM). Consider the Gaussian RWM and set $\sigma^2 = l^2/d$. Set $Y_t^d = r_d(X_{[dt]}^d)$. Then $Y^d \Rightarrow Y$ where

$$dY_t = -\frac{\sigma(l)^2}{4}Y_t dt + \sigma(l)dW_t; Y_0 \sim N(0, 2).$$

where $\sigma(l)^2 = 4\mu_2(l)$. By this, the Gaussian RWM is weakly consistent with the rate d .

Theorem (Optimality). The above RWM attains the optimal rate among *all* the RWM algorithms.

Proposition. Both *pCN* and *MpCN* algorithms have the rate **1**.

The key of the proof is reversibility.

$$\begin{aligned}
\mathbb{P}\left(\left|\|X_1^d\|^2 - \|X_0^d\|^2\right| > \epsilon\right) &= 2\mathbb{P}\left(\|X_1^d\|^2 - \|X_0^d\|^2 < -\epsilon\right) \\
&\leq 2\mathbb{P}\left(\|X_0^d + W_1^d\|^2 - \|X_0^d\|^2 < -\epsilon\right) \\
&= 2\mathbb{P}\left(2Z^d < -\epsilon\right)
\end{aligned}$$

where

$$Z^d := \frac{\|X_0^d + W_1^d\|^2 - \|X_0^d\|^2}{2} = \langle X_0^d, W_1^d \rangle + \frac{\|W_1^d\|^2}{2}.$$

We have

$$Z^d = \frac{1}{2} \left(\|W_1^d\| + \left\langle X_0^d, \frac{W_1^d}{\|W_1^d\|} \right\rangle \right)^2 - \left\langle X_0^d, \frac{W_1^d}{\|W_1^d\|} \right\rangle^2 \geq - \left\langle X_0^d, \frac{W_1^d}{\|W_1^d\|} \right\rangle^2.$$

3-d. Heavy-tail case

Set

$$r_d(x) = \frac{\|x\|^2}{d} \quad (x \in \mathbb{R}^d).$$

Proposition. Let $\Gamma_d = N_d(0, l^2 I_d/d)$ and set $Y_t^d = r_d(X_{[d^2 t]}^d)$.

Then $Y^d \Rightarrow Y$ where

$$dY_t = a(Y_t)dt + \sqrt{b(Y_t)}dW_t; Y_0 \sim Q$$

where

$$a(y) = 2(y + (\log q)'(y)y^2)\mu_2(l/\sqrt{y}) + l^2\mu_1(l/\sqrt{y}), \quad b(y) = 4y^2\mu_2(l/\sqrt{y}).$$

In particular, the Gaussian RWM has the rate d^2 .

Theorem. *The above RWM attains the optimal rate for the weak consistency. Thus d^2 is the optimal rate of RWM.*

Proposition. *In this case, pCN does not have any polynomial rate and MpCN has the rate d .*

Summary

	Light-tail	Heavy-tail
RMW	d	d^2
pCN	1	∞
MpCN	1	d

Summary

- We propose a new MCMC algorithm, MpCN algorithm.
- It works well for both toy models, and stochastic process examples.
- High-dimensional asymptotic theory was provided.

- [1] Yves F. Atchadé, Gareth O. Roberts, and Jeffrey S. Rosenthal. Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Statistics and Computing*, 21(4):555–568, October 2011. ISSN 0960-3174. doi: 10.1007/s11222-010-9192-1. URL <http://dx.doi.org/10.1007/s11222-010-9192-1>.

- [2] Mylène Bédard. Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.*, 17(4):1222–1244, 2007. ISSN 1050-5164. doi: 10.1214/105051607000000096. URL <http://dx.doi.org/10.1214/105051607000000096>.

- [3] A. Beskos, N. Pillai, G.O. Roberts, J.-M. Sanz-Serna, and A.M. Stuart. Optimal tuning of hybrid monte-carlo. to appear, 2013.

- [4] Alexandros Beskos, Gareth Roberts, and Andrew Stuart. Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.*, 19(3):863–898, 2009. ISSN 1050-5164. doi: 10.1214/08-AAP563. URL <http://dx.doi.org/10.1214/08-AAP563>.

- [5] A. Gelman, G. O. Roberts, and W. R. Gilks. Efficient Metropolis jumping rules. In *Bayesian statistics, 5 (Alicante, 1994)*, Oxford Sci. Publ., pages 599–607. Oxford Univ. Press, New York, 1996.
- [6] Kengo Kamatani. Local consistency of Markov chain Monte Carlo methods. *Ann. Inst. Statist. Math.*, 66(1):63–74, 2014. ISSN 0020-3157. doi: 10.1007/s10463-013-0403-3.
- [7] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):255–268, 1998. ISSN 1467-9868. doi: 10.1111/1467-9868.00123.
- [8] Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997. ISSN 1050-5164. doi: 10.1214/aoap/1034625254.

- [9] Nakahiro Yoshida. Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Ann. Inst. Statist. Math.*, 63(3):431–479, 2011. ISSN 0020-3157. doi: 10.1007/s10463-009-0263-z.