The random walk Metropolis - linking theory and practice through a case study.

Chris Sherlock

March 2009

Introduction

Much theory on creating efficient **random walk Metropolis (RWM)** algorithms.

Some applies to special cases, some more generally.

This talk:

1. Compares and contrasts a selection of RWM theory on scaling and shaping.
2. Uses this and other theory to suggest (often incremental) algorithmic improvements.
3. Examines algorithm performance on a non-trivial testing ground (the Markov Modulated Poisson Process).

The Random Walk Metropolis

The **RWM** algorithm explores a d-dimensional target density $\pi(\mathbf{x})$ by creating a Markov chain using a d-dimensional jump proposal density $\lambda^{-d}\ r(\mathbf{y}/\lambda)$ with $r(-\mathbf{y}) = r(\mathbf{y})$.

The Random Walk Metropolis

The **RWM** algorithm explores a d-dimensional target density $\pi(\mathbf{x})$ by creating a Markov chain using a d-dimensional jump proposal density $\lambda^{-d}\ r(\mathbf{y}/\lambda)$ with $r(-\mathbf{y}) = r(\mathbf{y})$.

From current position $\mathbf{X}_i$ propose a jump $\mathbf{Y}_i^*$.

Accept with probability $\alpha(\mathbf{x}_i, \mathbf{y}_i^*) = \min\left[1, \pi(\mathbf{x}_i + \mathbf{y}_i^*)/\pi(\mathbf{x}_i)\right]$

If accept $\mathbf{X}_{i+1} \leftarrow \mathbf{X} + \mathbf{Y}^*$ otherwise $\mathbf{X} \leftarrow \mathbf{X}$.

The Metropolis within Gibbs

The **MwG** algorithm explores a d-dimensional target density $\pi(\mathbf{x})$ by creating a Markov chain.

Jumps are proposed and accepted as for the RWM but have dimension $d^* < d$.

A **deterministic** MwG algorithm updates subsets of the components of $\mathbf{x}$ in some predetermined order.

A **random scan** MwG algorithm chooses at random the subset of components of $\mathbf{x}$ to be updated.

Integrated Autocorrelation Time

We wish to estimate $\mathbb{E}\left[f(\mathbf{X})\right]$ by $n^{-1}\sum_{1}^{n} f(\mathbf{x}^{(i)})$.

The MCMC sample is correlated and so the standard error of the estimate is $\mathrm{Var}\left[f(\mathbf{X})\right]/n_{eff}$ where $n_{eff} < n$ is the **effective sample size**.

Integrated Autocorrelation Time

We wish to estimate $\mathbb{E}\left[f(\mathbf{X})\right]$ by $n^{-1}\sum_1^n f(\mathbf{x}^{(i)})$.

The MCMC sample is correlated and so the standard error of the estimate is $\text{Var}\left[f(\mathbf{X})\right]/n_{eff}$ where $n_{eff} < n$ is the **effective sample size**.

For a stationary chain, let $\gamma_i = \text{Corr}\left[f(\mathbf{X}_k), f(\mathbf{X}_{k+i})\right]$.

The **integrated autocorrelation time** (ACT) is $\tau_f = 1 + 2\sum_1^\infty \gamma_i$, and $n_{eff} = n/\tau$.

We will use ACT to compare the output of the different MCMC algorithms.

Integrated Autocorrelation Time

We wish to estimate $\mathbb{E}\left[f(\mathbf{X})\right]$ by $n^{-1}\sum_1^n f(\mathbf{x}^{(i)})$.

The MCMC sample is correlated and so the standard error of the estimate is $\mathrm{Var}\left[f(\mathbf{X})\right]/n_{eff}$ where $n_{eff} < n$ is the **effective sample size**.

For a stationary chain, let $\gamma_i = \mathrm{Corr}\left[f(\mathbf{X}_k), f(\mathbf{X}_{k+i})\right]$.

The **integrated autocorrelation time** (ACT) is
$\tau_f = 1 + 2\sum_1^\infty \gamma_i$, and $n_{eff} = n/\tau$.

We will use ACT to compare the output of the different MCMC algorithms.

Finite sample so use $\tau_f = 1 + 2\sum_1^{l-1}\hat{\gamma}_i$ where l is the first lag such that $\hat{\gamma}_l < 0.05$.

Squared jumping distances

Could measure theoretical efficiency in terms of **expected squared Euclidean jump distance**:

$$S^2_{d,Euc} := \mathbb{E}\left[||\mathbf{X}_{i+1} - \mathbf{X}_i||^2\right].$$

Maximising the Euclidean square jump distance for a component is equivalent to minimising the lag-1 autocorrelation for that component.

Squared jumping distances

Could measure theoretical efficiency in terms of **expected squared Euclidean jump distance**:

$$S^2_{d,Euc} := \mathbb{E}\left[||\mathbf{X}_{i+1} - \mathbf{X}_i||^2 \right].$$

Maximising the Euclidean square jump distance for a component is equivalent to minimising the lag-1 autocorrelation for that component.

For an *elliptical* target with contours along lines of constant $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ an alternative measure would be the **expected square jump distance**

$$S^2_d := \mathbb{E}\left[(\mathbf{X}_{i+1} - \mathbf{X}_i)' \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{X}_{i+1} - \mathbf{X}_i) \right].$$

Speed of a limiting diffusion

Consider a single component (e.g. the first) of the d-dimensional chain at iteration i and denote it $X_{1,i}^{(d)}$.

Speed of a limiting diffusion

Consider a single component (e.g. the first) of the d-dimensional chain at iteration i and denote it $X_{1,i}^{(d)}$.
Define a speeded up continuous time process which mimics the first component of the chain as

$$W_t^{(d)} := X_{1,[td]}^{(d)} \tag{1}$$

Speed of a limiting diffusion

Consider a single component (e.g. the first) of the d-dimensional chain at iteration i and denote it $X_{1,i}^{(d)}$.
Define a speeded up continuous time process which mimics the first component of the chain as

$$W_t^{(d)} := X_{1,[td]}^{(d)} \tag{1}$$

Under certain circumstances it is possible to show that the (weak) limit $\lim_{d\to\infty} W_t^{(d)}$ is a Langevin diffusion.

The *speed* of this diffusion is another measure of the algorithm's efficiency.

Optimal scaling (1)

Roberts and Rosenthal (2001) consider a target with independent components

$$\pi(\mathbf{x}) = \prod_{i=1}^{d} C_i \; f(C_i x_i),$$

where $\mathbb{E}\left[C_i\right] = 1$ and $\mathbb{E}\left[C_i^2\right] = b < \infty$. A Gaussian proposal is used: $\lambda\mathbf{Z}$ where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$.

It is shown that subject to moment conditions on f, and provided $\lambda = \mu/d^{1/2}$, for some fixed μ, then as $d \to \infty$, $C_1 W_t^{(d)}$ (from 1) does approach a Langevin diffusion.

Optimal scaling (1)

Roberts and Rosenthal (2001) consider a target with independent components

$$\pi(\mathbf{x}) = \prod_{i=1}^{d} C_i \; f(C_i x_i),$$

where $\mathbb{E}\left[C_i\right] = 1$ and $\mathbb{E}\left[C_i^2\right] = b < \infty$. A Gaussian proposal is used: $\lambda \mathbf{Z}$ where $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$.

It is shown that subject to moment conditions on f, and provided $\lambda = \mu / d^{1/2}$, for some fixed μ, then as $d \to \infty$, $C_1 W_t^{(d)}$ (from 1) does approach a Langevin diffusion.

The speed of this diffusion is $\mu^2 \overline{\alpha}_d \times C_1^2 / b$, where

$$\overline{\alpha}_d := 2\Phi\left(-\frac{1}{2}\mu I^{1/2}\right)$$

is the acceptance rate,and I is a measure of roughness.

Optimal scaling (2)

Bedard (2007) considers targets with independent components and a triangular sequence of inverse scale coefficients $c_{i,d}$, and shows a similar result provided

$$\frac{\max_i c_{i,d}^2}{\sum_{i=1}^d c_{i,d}^2} \to 0. \tag{2}$$

Optimal scaling (3)

Sherlock and Roberts (2009) consider sequences of elliptically symmetric targets $\mathbf{X}^{(d)}$ explored by a spherically symmetric proposal $\lambda \mathbf{Z}^{(d)}$ and use ESJD as a measure of efficiency.

For many spherically symmetric distributions, as $d \to \infty$ all of the mass converges to a particular radius. It is shown than if $\lambda = \mu / d^{1/2} \times k_x^{(d)} / k_z^{(d)}$, and

$$\frac{\left|\mathbf{X}^{(d)}\right|}{k_x^{(d)}} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\left|\mathbf{Z}^{(d)}\right|}{k_z^{(d)}} \xrightarrow{m.s.} 1,$$

and the inverse scale parameters of the axes of the elliptical target satisfy 2 then

Optimal scaling (3)

Sherlock and Roberts (2009) consider sequences of elliptically symmetric targets $\mathbf{X}^{(d)}$ explored by a spherically symmetric proposal $\lambda \mathbf{Z}^{(d)}$ and use ESJD as a measure of efficiency.

For many spherically symmetric distributions, as $d \to \infty$ all of the mass converges to a particular radius. It is shown than if $\lambda = \mu / d^{1/2} \times k_x^{(d)} / k_z^{(d)}$, and

$$\frac{\left|\mathbf{X}^{(d)}\right|}{k_x^{(d)}} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\left|\mathbf{Z}^{(d)}\right|}{k_z^{(d)}} \xrightarrow{m.s.} 1,$$

and the inverse scale parameters of the axes of the elliptical target satisfy 2 then

$$\frac{d}{k_x^{(d)^2}} S_d^2(\mu) \to \mu^2 \ \overline{\alpha}_d \quad \text{with} \quad \overline{\alpha}_d(\mu) := 2\Phi\left(-\frac{1}{2}\mu\right).$$

Optimal scaling (4)

Optimising the efficiency measure w.r.t. μ and substituting gives

$$\lambda_d = \frac{2.38}{d^{1/2}\ I^{1/2}} \text{ (R and R)} \quad \text{and} \quad \lambda_d = \frac{2.38\ k_x^{(d)}}{d^{1/2}\ k_z^{(d)}} \text{ (S and R).}$$

Optimal scaling (4)

Optimising the efficiency measure w.r.t. μ and substituting gives

$$\lambda_d = \frac{2.38}{d^{1/2}\ I^{1/2}} \text{ (R and R)} \quad \text{and} \quad \lambda_d = \frac{2.38\ k_x^{(d)}}{d^{1/2}\ k_z^{(d)}} \text{ (S and R).}$$

Both lead to an optimal acceptance rate of 0.234.

Optimal scaling (4)

Optimising the efficiency measure w.r.t. μ and substituting gives

$$\lambda_d = \frac{2.38}{d^{1/2}\ I^{1/2}} \text{ (R and R)} \quad \text{and} \quad \lambda_d = \frac{2.38\ k_x^{(d)}}{d^{1/2}\ k_z^{(d)}} \text{ (S and R).}$$

Both lead to an optimal acceptance rate of 0.234.

Algorithm 1: proposal $\mathbf{Y} \sim N(\mathbf{0}, \lambda^2 I)$ with λ chosen so that the acceptance rate is between 0.2 and 0.3.

Optimal scaling (5)

NB The limiting optimal acceptance rate need not be 0.234 - e.g. Bedard (2008), Sherlock and Roberts (2009).

Optimal scaling for MwG (1)

Neal and Roberts (2006) consider the behaviour of the random scan MwG algorithm on a target with iid components.

Optimal scaling for MwG (1)

Neal and Roberts (2006) consider the behaviour of the random scan MwG algorithm on a target with iid components.

The optimal scale parameter is larger than for a full update (since the dimension of the update is smaller) but the limiting optimal acceptance rate is still 0.234.

Optimal scaling for MwG (2)

Sherlock (2006) considers a deterministic MwG algorithm on a sequence of elliptical targets (subject to 2) with updates proposed from a spherical distribution, but allowing different scalings for different sub-blocks of principal components of the ellipse.

For equal-sized sub-blocks the limiting relative efficiency (compared the optimal RWM with a single spherical proposal) is shown to be

$$r_{MwG/RWM} = \frac{\frac{1}{k}\sum \overline{c^2}_i}{\left(\frac{1}{k}\sum \overline{c^2}_i^{-1}\right)^{-1}}$$

where $\overline{c^2}_i$ is the mean of the squares of the inverse scale parameters for the i^{th} sub-block.

Optimal scaling for MwG (3)

An optimally tuned MwG algorithm (for orthogonal sub-blocks) will be more efficient than a single block update.

Algorithm 2: MwG with proposed jumps $Y_i \sim N(0, \lambda_i^2)$ optimised along each component ($\alpha \approx 0.4$).

Optimal shaping (1)

Sherlock (2006) considers elliptical targets explored either using an optimally tuned spherical proposal or and optimally tuned elliptical proposal of the same shape and orientation as the target.

Optimal shaping (1)

Sherlock (2006) considers elliptical targets explored either using an optimally tuned spherical proposal or and optimally tuned elliptical proposal of the same shape and orientation as the target.

For a sequence where the target with dimension d has elliptical axes with inverse scale parameters $c_{d,1}, \ldots, c_{d,d}$, the limiting ratio of expected squared Euclidean jump distances is

$$r_{sph/ell} = \frac{\lim_{d\to\infty} \left(\frac{1}{d} \sum_{i=1}^{d} c_{d,i}^{-2} \right)^{-1}}{\lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} c_{d,i}^{2}}.$$

Optimal shaping (2)

Roberts and Rosenthal (2001) examine targets of the form

$$\prod C_i f(C_i x_i)$$

and compare the efficiencies of the limiting Langevin diffusions for spherical Gaussian proposals and Gaussian proposals with inverse scale parameter C_i for the i^{th} component.

Optimal shaping (2)

Roberts and Rosenthal (2001) examine targets of the form

$$\prod C_i f(C_i x_i)$$

and compare the efficiencies of the limiting Langevin diffusions for spherical Gaussian proposals and Gaussian proposals with inverse scale parameter C_i for the i^{th} component.

The limiting efficiency was found to be

$$r_{id/iid} = \frac{\mathbb{E}\left[C^2\right]}{\mathbb{E}\left[C\right]^2}$$

Optimal shaping (3)

We should therefore explore the target using a proposal with a similar shape and orientation to the target.

Algorithm 3: use 1000 iterations from Algorithm 1 to estimate the covariance matrix $\hat{\boldsymbol{\Sigma}}$ then propose from $N(\mathbf{0}, \lambda\hat{\boldsymbol{\Sigma}})$ with λ chosen to give an acceptance rate between 0.2 and 0.3.

Exploring heavy tails

There is evidence (e.g. Roberts, 2003) to suggest that a heavy tailed proposal should better explore a heavy tailed target.

Algorithm 4 proposes from a Cauchy distribution with modal hessian $\hat{\boldsymbol{\Sigma}}^{-1}$, and scaling chosen so as to minimise the ACT.

Exploring heavy tails

There is evidence (e.g. Roberts, 2003) to suggest that a heavy tailed proposal should better explore a heavy tailed target.

Algorithm 4 proposes from a Cauchy distribution with modal hessian $\hat{\boldsymbol{\Sigma}}^{-1}$, and scaling chosen so as to minimise the ACT.

An alternative strategy is to transform the target to one with lighter tails. Dellaportas and Roberts (2003) use a random walk on the posterior for the log of each parameter: the **multiplicative random walk**.

Algorithm 5 uses a Gaussian proposal on a transformed parameter set $\{\log \theta_1, \ldots, \log \theta_4\}$, with shape matrix estimated as for Algorithm 3 (but on the log parameters!).

Adaptive MCMC (1)

Rather than estimating $\mathbf{\Sigma}$ and λ from finite tuning runs, we could let a single algorithm learn from its own output.

It is important that changes to the MCMC kernel become vanishingly small as iteration $i \to \infty$ (e.g. Roberts and Rosenthal, 2007).

Adaptive MCMC (1)

Rather than estimating $\boldsymbol{\Sigma}$ and λ from finite tuning runs, we could let a single algorithm learn from its own output.

It is important that changes to the MCMC kernel become vanishingly small as iteration $i \to \infty$ (e.g. Roberts and Rosenthal, 2007).

Algorithm 6 uses a random walk on the posterior for the log parameters. The jump proposal is

$$\mathbf{Y} \sim \begin{cases} N\left(\mathbf{0}, m^2\hat{\boldsymbol{\Sigma}}_n\right) & w.p. \quad 1-\delta \\ N\left(\mathbf{0}, \frac{1}{d}\lambda_0^2\mathbf{I}\right) & w.p. \quad \delta. \end{cases}$$

Here $\delta = 0.05$, $d = 4$, and $\hat{\boldsymbol{\Sigma}}_n$ is estimated from the logarithms of the posterior sample to date.

Adaptive MCMC (2)

$$\mathbf{Y} \sim \begin{cases} N\left(\mathbf{0}, m^2\hat{\boldsymbol{\Sigma}}_n\right) & w.p. \quad 1-\delta \\ N\left(\mathbf{0}, \frac{1}{d}\lambda_0^2\mathbf{I}\right) & w.p. \quad \delta. \end{cases}$$

A few minutes were spent tuning the block multiplicative random walk with proposal variance $\frac{1}{4}\lambda_0^2\mathbf{I}$ to give at least a reasonable value for λ_0 (acceptance rate ≈ 0.3), although this is not stricly necessary.

Adaptive MCMC (2)

$$\mathbf{Y} \sim \begin{cases} N\left(\mathbf{0}, m^2\hat{\mathbf{\Sigma}}_n\right) & w.p. \quad 1-\delta \\ N\left(\mathbf{0}, \frac{1}{d}\lambda_0^2\mathbf{I}\right) & w.p. \qquad \delta. \end{cases}$$

A few minutes were spent tuning the block multiplicative random walk with proposal variance $\frac{1}{4}\lambda_0^2\mathbf{I}$ to give at least a reasonable value for λ_0 (acceptance rate ≈ 0.3), although this is not stricly necessary.

m was updated as follows: if the proposal was rejected then $m < -m - \Delta/i^{1/2}$, otherwise $m < -m + 2.3\Delta/i^{1/2}$. This leads to an equilibrium acceptance rate of $1/3.3$ (Δ is some small fixed quantity).

The MMPP

A Markov modulated Poisson process (MMPP) is a Poisson process, the intensity of which, $\lambda(X_t)$, depends on the state of a continuous time discrete space Markov chain X_t.

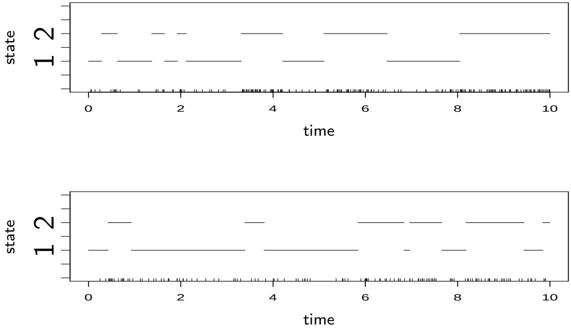

Figure: Two 2-state cts time MCs simulated from generator $\mathbf{Q}$ with $q_{12} = q_{21} = 1$; rug plots show events from MMPPs simulated from these chains, with intensity $\psi = (10, 30)$ (upper) and $\psi = (10, 17)$ (lower).

The MMPP Test Data

Simulated test data was from 100 secs of MMPPs with $q_{12} = q_{21} = 1$ and either $\boldsymbol{\psi} = (10, 30)$ (D1 - 3 replicates) or $\boldsymbol{\psi} = (10, 17)$ (D2 - 3 replicates).

The MMPP Test Data

Simulated test data was from 100 secs of MMPPs with $q_{12} = q_{21} = 1$ and either $\boldsymbol{\psi} = (10, 30)$ (D1 - 3 replicates) or $\boldsymbol{\psi} = (10, 17)$ (D2 - 3 replicates).

D1 - more events + easier to distinguish the state of the underlying chain ⇒ lighter tails + parameters $(\psi_1, \psi_2, q_{12}, q_{21})$ closer to orthogonal.

The MMPP Test Data

Simulated test data was from 100 secs of MMPPs with $q_{12} = q_{21} = 1$ and either $\boldsymbol{\psi} = (10, 30)$ (D1 - 3 replicates) or $\boldsymbol{\psi} = (10, 17)$ (D2 - 3 replicates).

D1 - more events + easier to distinguish the state of the underlying chain ⇒ lighter tails + parameters $(\psi_1, \psi_2, q_{12}, q_{21})$ closer to orthogonal.

D2 - fewer events + harder to distinguish the state of the underlying chain ⇒ heavier tails + parameters far from orthogonal.

Using problem specific information

When $\psi_1 \approx \psi_2$ can Taylor expand likelihood in $\boldsymbol{\psi}$ about $\overline{\psi}\mathbf{1}$.

Leads to a new reparamterisation with new parameters approximately orthogonal (when $\psi_2 \approx \psi_1$).

Algorithm 7: MwG updates on the new parameters, multiplicative where possible (3/4).

Analysis

Priors: Exponential, with mean the known “true” parameter value.

Runs of 10 000 iterations (+ burn in of 1000)

Accuracy? Compared with 100 000 iterations of a Gibbs sampler (Sherlock and Fearnhead, 2006). All *OK*.

Efficiency: ACT (mutiplied by 4 for MwG).

ACT Results (1)

All algorithms performed better on D1 than D2 because D1 has lighter tails.

ACT Results (1)

All algorithms performed better on D1 than D2 because D1 has lighter tails.

Alg2 (MwG, $N(0, \lambda_i^2 \mathbf{I})$) 2-3 times better than **Alg1** for D1 but only 1.5 times better than Alg1 for D2, as parameters closer to orthogonal for D1.

ACT Results (1)

All algorithms performed better on D1 than D2 because D1 has lighter tails.

Alg2 (MwG, $N(0, \lambda_i^2 \mathbf{I})$) 2-3 times better than **Alg1** for D1 but only 1.5 times better than Alg1 for D2, as parameters closer to orthogonal for D1.

Alg3 ($N(\mathbf{0}, \hat{\mathbf{\Sigma}})$) 4-6 times better than Alg1 ($N(\mathbf{0}, \lambda^2 \mathbf{I})$).

ACT Results (1)

All algorithms performed better on D1 than D2 because D1 has lighter tails.

Alg2 (MwG, $N(0, \lambda_i^2 \mathbf{I})$) 2-3 times better than **Alg1** for D1 but only 1.5 times better than Alg1 for D2, as parameters closer to orthogonal for D1.

Alg3 ($N(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$) 4-6 times better than Alg1 ($N(\mathbf{0}, \lambda^2 \mathbf{I})$).

Improvements in Alg2 and Alg3 best for ψ as Alg1 limited by variance of q.

ACT Results (2)

Alg4 (Cauchy, $\hat{\boldsymbol{\Sigma}}$) performs ≈ 1.5 times *worse* than Alg3 (Normal, $\hat{\boldsymbol{\Sigma}}$) for *both* algorithms!

More negative proposals? $\hat{\boldsymbol{\Sigma}}$ not representative away from the modes?

ACT Results (2)

Alg4 (Cauchy, $\hat{\boldsymbol{\Sigma}}$) performs ≈ 1.5 times *worse* than Alg3 (Normal, $\hat{\boldsymbol{\Sigma}}$) for *both* algorithms!

More negative proposals? $\hat{\boldsymbol{\Sigma}}$ not representative away from the modes?

Alg5 (Multiplicative, Normal, $\hat{\boldsymbol{\Sigma}}_*$) performs the same as Alg3 for D1 and $\approx 1.5 - 2$ times better than Alg3 for D2.
Heavier tails.

ACT Results (2)

Alg4 (Cauchy, $\hat{\mathbf{\Sigma}}$) performs ≈ 1.5 times *worse* than Alg3 (Normal, $\hat{\mathbf{\Sigma}}$) for *both* algorithms!

More negative proposals? $\hat{\mathbf{\Sigma}}$ not representative away from the modes?

Alg5 (Multiplicative, Normal, $\hat{\mathbf{\Sigma}}_*$) performs the same as Alg3 for D1 and $\approx 1.5 - 2$ times better than Alg3 for D2.
Heavier tails.

Alg6 (Adap, mult; Normal, $\hat{\mathbf{\Sigma}}_*$) performs the same as Alg5 for D1 and 1-1.5 times better than Alg5 for D2.
Takes > 1000 iterations to estimate $\hat{\mathbf{\Sigma}}$?

ACT Results (3)

Alg7 (Reparameterisation; MwG, mult. where possible, Normal) performs ≈ 2 times worse than Alg6 (Adap, mult; Normal, $\hat{\boldsymbol{\Sigma}}_*$) for D1 *but* performance is very similar to Alg6 for D2.

Alg7 was designed for cases such as D2.

Summary

- Two different approaches to optimising RWM.
- Apply to different distributions (independent components / elliptical contours).
- Use different measures (diffusion speed / ESJD)
- Suggest similar methods for producing efficient algorithms.

Summary

- Two different approaches to optimising RWM.
- Apply to different distributions (independent components / elliptical contours).
- Use different measures (diffusion speed / ESJD)
- Suggest similar methods for producing efficient algorithms.
- Algorithms perform as might be expected, except for the Cauchy proposal - worse.
- On the heavier tailed data set, the adaptive algorithm performs as well as the algorithm which uses problem specific knowledge.

References

Bedard, M. (2007). Weak convergence of Metropolis algorithms for non-iid target distributions. *Ann. Appl. Probab.* **17**(4), 1222-1244.

Bedard, M. (2008). Optimal acceptance rates for Metropolis algorithms: moving beyond 0.234. *Stochastic Process. Appl.* **118**(12), 2198-2222.

Dellaportas, P. and Roberts, G.O. An introduction to MCMC. In number 173 in Lecture Notes in Statistics, Springer, Berlin, 1-41.

Fearnhead, P. and Sherlock, C. (2006). An exact Gibbs sampler for the Markov Modulated Poisson Process. *J.R.Stat.Soc.Ser.B Stat. Methodol.* **68**(5), 767-784.

Neal, P. and Roberts, G. (2006). Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Probab.* **16**, 475-515.

Roberts, G.O. (2003). Linking theory and practice of MCMC. In volume 27 of *Oxford Statist. Sci. Ser.*, Oxford Univ. Press, Oxford.

Roberts, G.O., and Rosenthal, J.S. (2007). Optimal scaling of various Metropolis-Hastings algorithms. *Statistical Science.* **16**, 351-367.

Sherlock, C. (2006). Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis. PhD thesis, Lancaster.

Sherlock, C. and Roberts, G.O. (2009). Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, to appear.