# Geometric Ergodicity through Variable Transformation in Metropolis Random Walk Markov Chain Monte Carlo

Charles J. Geyer

School of Statistics
University of Minnesota

April, 11, 2013

Co-author: Leif Johnson (Google) who did all the work

# Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is simple.

You have a probability or expectation that you cannot do exactly.

So simulate, but you cannot do that exactly either.

Set up Markov chain whose equilibrium distribution is the distribution you want to simulate (this is easy). Run it. Average over the simulations.

MCMC gives *not* independent, *not* indentically distributed "sample" from the distribution.

# Random Walk Metropolis

If the distribution you want to sample is *continuous* and *finite dimensional* and you can write down an *unnormalized* probability density function (don't need to know normalizing constant), you're done.

Write an R function that evaluates the log unnormalized density (LUD) function. R package `mcmc` does everything else.

No way for the user to screw up, except for coding the LUD function incorrectly.

# Random Walk Metropolis (cont.)

R package always runs Markov chain having the equilibrium distribution specified by the LUD function supplied *assuming there is such a distribution*.

If $h$ is LUD function, then $e^h$ is unnormalized density function, and the assumption is that the integral of $e^h$ is *finite*.

If $\int e^h = \infty$, then all bets are off. The distribution you think you want to sample does not exist.

MCMC does not do real analysis for you. Must do this bit (verifying $\int e^h < \infty$) yourself.

# Applications

Bayesian inference.

$$\text{likelihood} \times \text{unnormalized prior} = \text{unnormalized posterior}$$

Sampling any conditional distribution (Bayes is special case).

$$\text{unnormalized joint} = \text{unnormalized conditional}$$

Models specified by giving *unnormalized* density or mass functions. Also called Markov random field models. Widely used in spatial statistics and social networks.

# Random Walk Metropolis (cont.)

**Won't hurt if you don't get this.**

How random-walk Metropolis algorithm (RWMA) does one step of the Markov chain (at $X_n$, move to $X_{n+1}$) whose equilibrium distribution has LUD function $h$

Let $Z_n$ be independent draw from symmetric distribution centered at zero (e. g., mean-zero normal). Let $U_n$ be independent draw from Uniform$(0, 1)$. If

$$\log(U_n) < h(X_n + Z_n) - h(X_n)$$

set $X_{n+1} = X_n + Z_n$. Otherwise set $X_{n+1} = X_n$.

# Random Walk Metropolis (cont.)

Support of equilibrium distribution does not have to be whole space (just have LUD function return `-Inf` for points not in the support).

Initial state $X_1$ must be in support (LUD function returns finite value).

Operation of Metropolis algorithm assures chain stays in support when started in support.

# Random Walk Metropolis (cont.)

How to code a LUD function. Suppose state vector has three components, last of which must be positive. Start like this

```
ludfun <- function(theta) {
    stopifnot(length(theta) == 3)
    stopifnot(is.numeric(theta))
    stopifnot(is.finite(theta))
    boo <- theta[1]
    foo <- theta[2]
    moo <- theta[3]
    if (moo <= 0) return(-Inf)
    # now (boo, foo, moo) is in support
    # calculate LUD function and return (finite) value
}
```

# Foolproof (?!)

The referees for *Annals of Statistics* would not let us use the word "foolproof" for this algorithm and R package.

Indeed we have already mentioned two ways users can screw up

- Code up the LUD function incorrectly.
- The putative LUD function isn't ($\int e^h = \infty$).

But there is no other way the user can screw up (other than misreport the output), and we claim that this is *as foolproof as it is possible for an MCMC package to be*.

# Geometric Ergodicity

Remember that the Markov chain $X_1$, $X_2$, ... is a *not* independent and *not* identically distributed "sample" from the equilibrium distribution. The scare quotes remind you that this isn't your grandfather's notion of "random sample".

So if $X_1$, $X_2$, ... do not have the same distribution, how can they be a "sample" from the equilibrium distribution?

Do have

$$\mathcal{L}(X_n) \to L$$

(the law of $X_n$ converges to the equilibrium distribution $L$).

The distribution of $X_n$ is never exactly $L$, but it gets closer and closer to $L$ as $n \to \infty$.

# Geometric Ergodicity (cont.)

**Won't hurt if you don't get this.**

In theory-speak, this invokes the *aperiodic ergodic theorem for Harris recurrent Markov chains*

$$\|\mathcal{L}(X_n) - L\| \to 0$$

where the norm here is total variation norm (stronger than convergence in distribution).

Geometric ergodicity says this happens exponentially fast: there exists $r > 1$ such that

$$\sum_{n=1}^{\infty} r^n \|\mathcal{L}(X_n) - L\| < \infty$$

# Geometric Ergodicity (cont.)

Geometric ergodicity is important because RWMA is *reversible* Markov chain and

**Theorem (Roberts and Rosenthal, 1997)**

*A reversible Markov chain has a central limit theorem for all square-integrable functionals if and only if it is geometrically ergodic.*

There are central limit theorems that don't use geometric ergodicity and apply to some but not all square-integrable functionals, but their conditions are very hard to verify. Geometric ergodicity is (fairly) easy to verify.

# The Big Issue (Finally!)

RWMA is not guaranteed to be geometrically ergodic.

Has been well studied (Mengersen and Tweedie, 1996; Roberts and Tweedie, 1996; Jarner and Hansen, 2000).

The main conclusion is that geometric ergodicity depends mainly on the tail behavior of the LUD function (light enough tails = geometrically ergodic).

The tail behavior of the proposal distribution (the distribution of the $Z_n$ in our description of RWMA) *does not matter* for geometric ergodicity.

# Jarner-Hansen Theory

**Superexponentially Light Tail Condition**

$$\limsup_{\|x\|\to\infty} \frac{x}{\|x\|} \cdot \nabla h(x) = -\infty$$

**Curvature Condition**

$$\limsup_{\|x\|\to\infty} \frac{x}{\|x\|} \cdot \frac{\nabla h(x)}{\|\nabla h(x)\|} < 0$$

Theorem (Jarner and Hansen, 2000)

*If LUD function $h$ is strictly positive and continuously differentiable and satisfies the Superexponentially Light Tail Condition and the Curvature Condition, then the RWMA is geometrically ergodic.*

# Jarner-Hansen Theory (cont.)

**Various Tail Conditions**

$$\limsup_{\|x\| \to \infty} \frac{x}{\|x\|} \cdot \nabla h(x) = L$$

| | |
|---|---|
| $L = -\infty$ | superexponentially light |
| $-\infty < L < 0$ | exponentially light |
| $L = 0$ | subexponentially light |
| $L > 0$ | there be dragons |

# Jarner-Hansen Theory (cont.)

How to always have superexponentially light tails if you are a Bayesian: put normal priors (which have $e^{-x^2}$ tails) on everything.

Example of merely exponentially (not superexponentially) light tails: discrete exponential family with natural parameters and conjugate priors (log-linear models for categorical data, logistic regression, Poisson regression).

Normal priors incorporate more prior information than in any finite amount of data, no matter how much! How to justify that?

Example of subexponentially light tails: Cauchy location model with flat prior.

# Johnson-Geyer Theory

If don't have superexponentially light tails, then we're screwed?

Not necessarily. Change-of-variable theorem to the rescue. If $X = g(Y)$

$$f_Y(y) = f_X[g(y)] \times |\det \nabla g(y)|$$

for LUD functions

$$h_Y(y) = h_X[g(y)] + \log|\det \nabla g(y)|$$

If $h_Y$ has superexponentially light tails and satisfies the curvature condition then RWMA for it is geometrically ergodic and

$$X_i = g(Y_i), \qquad n = 1, 2, \ldots$$

is also geometrically ergodic Markov chain (though not RWMA).

# Johnson-Geyer Theory (cont.)

Admittedly, very simple idea. Many people (including me) are embarrassed we didn't think of it years ago.

> *How extremely stupid not to have thought of that!*
> *— Thomas Henry Huxley about Darwin's theory of evolution by natural selection*

So we couldn't just say *that* and have an *Annals of Statistics* paper. Leif thought up and proved some theorems that make it as easy as possible to verify that the transformed LUD function satisfies the curvature and superexponentially light tail conditions.

# Isotropic Transformations

$$g(x) = \begin{cases} f(\|x\|)\frac{x}{\|x\|}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

These are used merely for mathematical convenience. Many others could be used. These simplify proofs.

The following conditions assure the LUD function for $Y$ is strictly positive and continuously differentiable (assuming the LUD function for $X$ was)

- $f$ is twice continuously differentiable with one-sided derivatives at zero.
- $f'(r) > 0$ for all $r$.
- $f''(0) = 0$.

# Isotropic Transformations (cont.)

Need $f$ (radial expansion function) that

- looks linear near zero
- pulls in the tails (increases faster than linear) for large $r$.

**Exponentially Light to Superexponentially Light**

$$f(r) = \begin{cases} r, & r < R \\ r + (r - R)^p, & r \geq R \end{cases}$$

**Subexponentially Light to Exponentially Light**

$$f(r) = \begin{cases} \frac{r^3 b^3 e}{6} + \frac{rbe}{2}, & r \leq \frac{1}{b} \\ e^{br} - \frac{e}{3}, & r > \frac{1}{b} \end{cases}$$

# Leif's Theorems

**Theorem**

*If a LUD function is exponentially light, then the first isotropic transformation takes us to a superexponentially light LUD function for the transformed variable.*

**Theorem**

*If a LUD function satisfies the curvature condition (in particular if its derivative is bounded), then the first isotropic transformation takes us to a LUD function for the transformed variable that satisfies the curvature condition.*

Hence we only need to check the curvature and exponentially light conditions on the original LUD function. Then variable transformation of the first kind (with polynomial radial expansion function) does the job.

# Leif's Theorems (cont.)

## Theorem

*If a LUD function $h$ is subexponentially light and for some $K$ greater than the dimension of the state space and $R < \infty$*

$$\frac{x}{\|x\|} \cdot \nabla h(x) \leq -\frac{K}{\|x\|}, \qquad \|x\| > R,$$

*then the second isotropic transformation takes us to an exponentially light LUD function.*
*If a LUD function $h$ is subexponentially light and for some $K$ greater than the dimension of the state space and $R < \infty$*

$$\|\nabla h(x)\| \leq -\frac{K}{\|x\|}, \qquad \|x\| > R,$$

*then the second isotropic transformation takes us to an exponentially light LUD function satisfying the conditions of the previous curvature condition theorem.*

# Leif's Theorems (cont.)

The last theorem is a bit annoying in that it doesn't produce superexponentially light tails in one step. So one has to apply both transformations (one after the other) to do the job. Also the two conditions in the theorem are not elegant, but the condition that $K$ is greater than the dimension of the state space is just enough to make $\int e^h < \infty$, so is not restrictive.

# Summary

RWMA with LUD $h$ is "foolproof" except that you have to

- prove $\int e^h < \infty$,
- correctly code up $h$ as R function,
- understand the tail behavior of $h$ so as to select the correct morphism (or none if superexponentially light already).

# My Favorite Application

Exponential family log likelihood

$$l_n(\theta) = n\bar{y}_n \cdot \theta - nc(\theta)$$

where

$$\nabla c(\theta) = E_\theta(Y)$$
$$\nabla^2 c(\theta) = \mathrm{var}_\theta(Y)$$

Log conjugate prior looks like log likelihood (Diaconis and Ylvisaker, 1979)

$$p(\theta) = \nu\eta \cdot \theta - \nu c(\theta)$$

where $\eta$ and $\nu$ are hyperparameters.

# My Favorite Application (cont.)

Log unnormalized posterior (LUD function)

$$h(\theta) = (n\bar{y}_n + \nu\eta) \cdot \theta - (n + \nu)c(\theta)$$

**Theorem (Diaconis and Ylvisaker (1979))**

*$\int e^h < \infty$ if and only if $h$ has a unique mode, which happens if and only if*

- $n + \nu > 0$
- $(n\bar{y}_n + \nu\eta)/(n + \nu)$ *is possible value of* $E_\theta(Y)$

Proper posterior always guaranteed by proper prior, which happens if and only if $\nu > 0$ and $\eta$ is possible value of $E_\theta(Y)$.

# My Favorite Application (cont.)

The following all shown in Johnson and Geyer (2012).

If natural statistic vector $Y$ is bounded in any direction, then $h$ does not have superexponentially light tails.

If $\int e^h < \infty$, the $h$ does have exponentially light tails and satisfies the curvature condition.

Hence isotropic transformation of the first kind produces geometrically ergodic transformed chain.

Applications: log-linear models for categorical data, logistic regression, Poisson regression (log link).

# Leif's Favorite Application

Univariate $t$ distribution (RWMA not geometrically ergodic)

How to code (from vignette in R package `mcmc`)

```
lud <- function(x) dt(x, df=3, log=TRUE)
out <- morph.metrop(lud, 0, blen=100, nbatch=100,
     morph=morph(b=1))
out <- morph.metrop(out, scale=4)
```

The later to adjust the scale to get about 20% acceptance rate.

```
t.test(out$batch)
```

gives confidence interval for "unknown" true posterior mode (which we actually know is zero because this is a toy problem but we wouldn't know in a real application).

# Leif's Favorite Application (cont.)

If we use unmorphed RWMA, it doesn't work.

With same LUD function (`lud`) as before.

```
out <- metrop(lud, 0, blen=100, nbatch=100)
out <- metrop(out, scale=6)
out <- metrop(out, blen=1e6, nbatch=1e3)
```

Total run length for last run

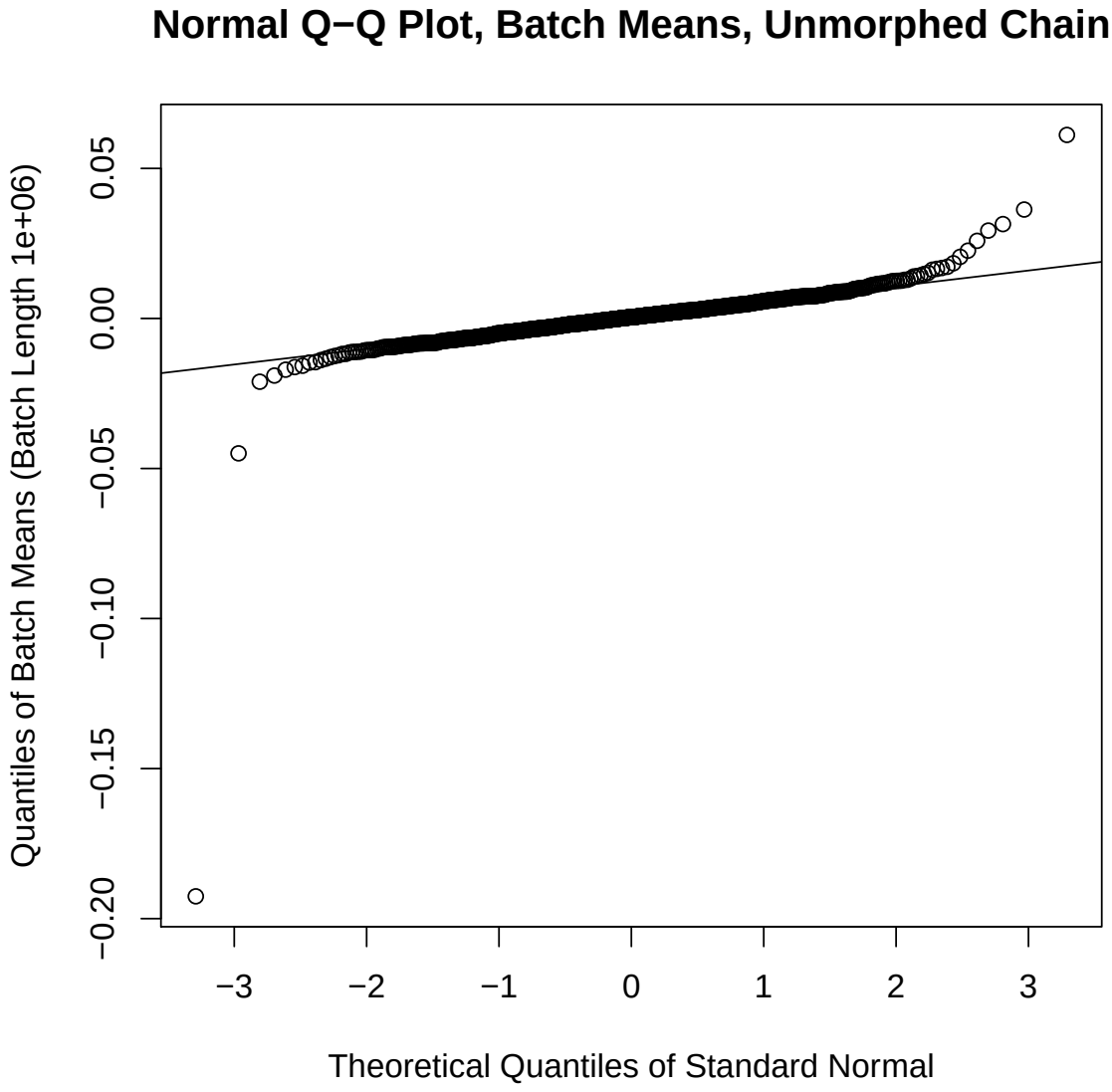
$$\text{batch length } (\texttt{blen}) \times \text{number of batches } (\texttt{nbatch}) = 10^9$$

# Leif's Favorite Application (cont.)

Batch means should be normally distributed if batch length $b = 10^6$ is long enough to be "large" in the Markov chain central limit theorem (MCCLT)

$$\sqrt{b}(\overline{X}_b - \mu) \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma^2)$$

# Leif's Favorite Application (cont.)



# Leif's Favorite Application (cont.)

The generalized MCCLT (GMCCLT) says for $1 < \alpha \leq 2$

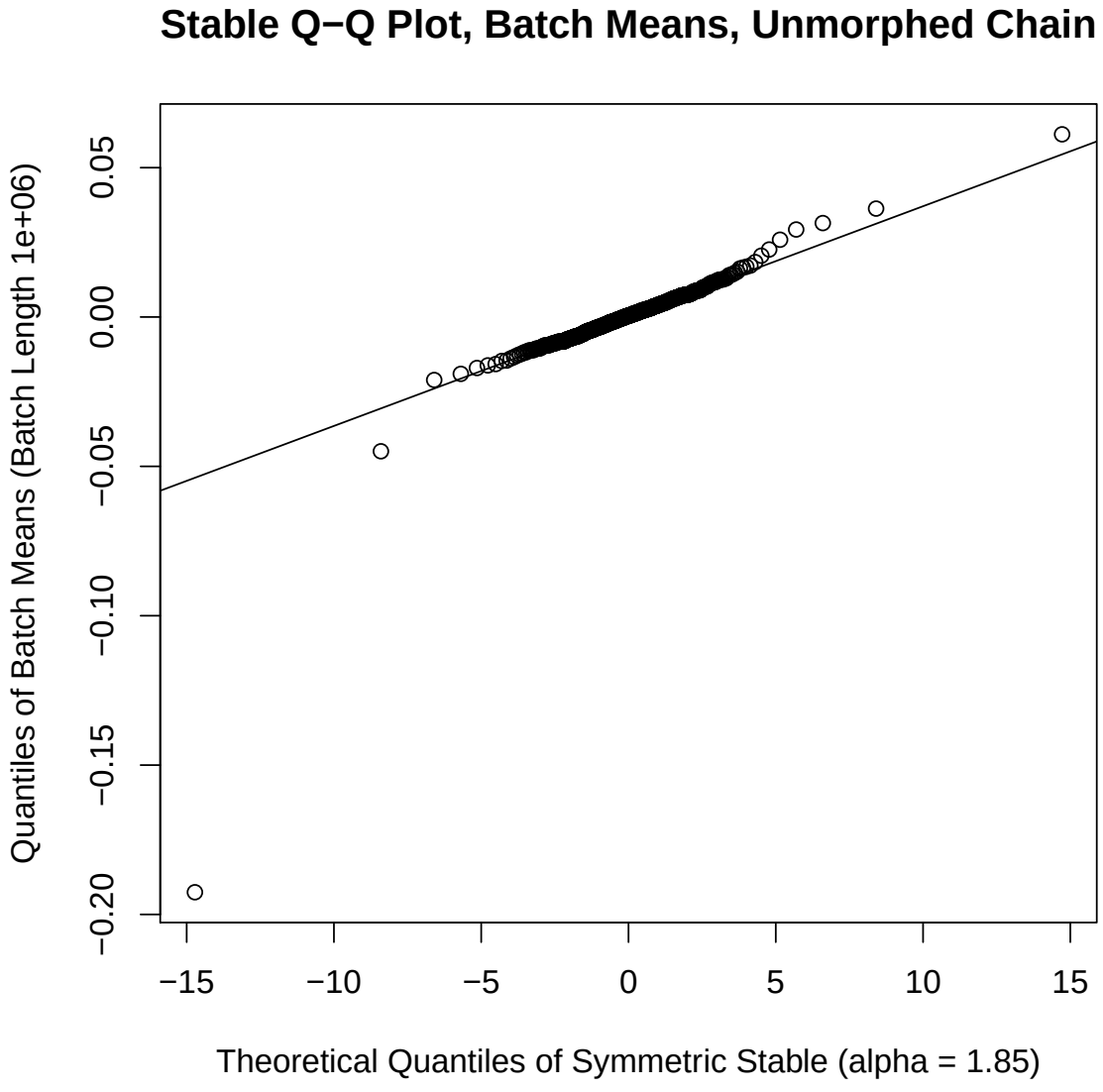
$$b^{1-1/\alpha} s(b)(\overline{X}_b - \mu) \xrightarrow{\mathcal{D}} \text{Stable}(\alpha, \beta, \gamma, 0)$$

where $s$ is a slowly varying function.

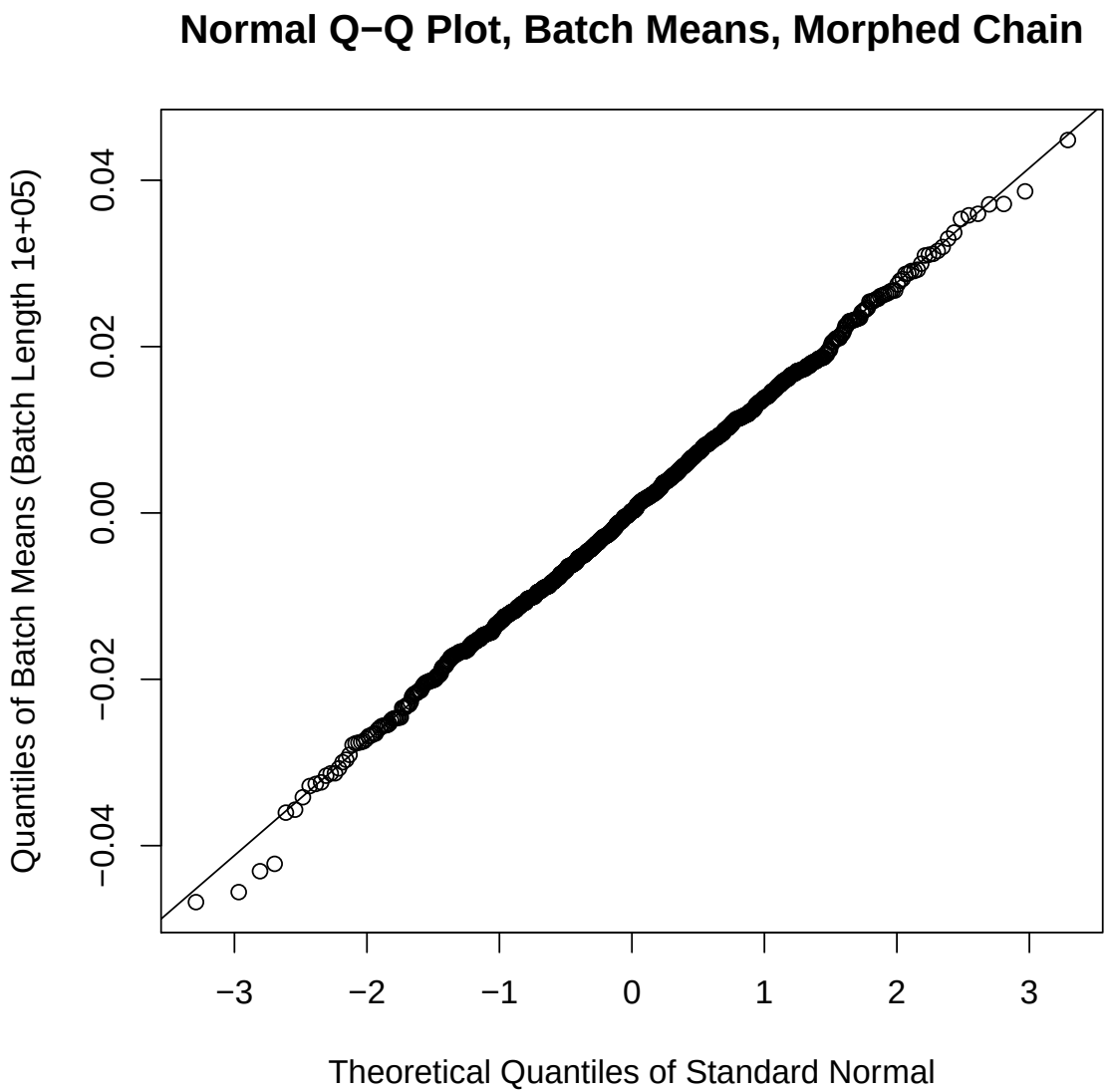
The ordinary MCCLT is $\alpha = 2$ special case ($\alpha$ is index, $\beta$ is skewness, $\gamma$ is scale). Here we expect $\beta = 0$ by symmetry of equilibrium distribution.

Unknown (to me) if GMCCLT actually holds in this case or what $\alpha$ would be if it does hold.

# Leif's Favorite Application (cont.)



# Appendix I


Normal Q−Q Plot, Batch Means, Morphed Chain
Quantiles of Batch Means (Batch Length 1e+05)
0.04
0.02
0.00
−0.02
−0.04
−3
−2
−1
0
1
2
3
Theoretical Quantiles of Standard Normal

# Appendix II

Details of bounding the tail behavior for exponential families.

$\int e^h < \infty$ if and only if $h$ has a unique mode $\tilde{\theta}_n$.

By strict concavity $\nabla h$ is strictly multivariate monotone function

$$\left[\nabla h(\theta_1) - \nabla h(\theta_2)\right] \cdot (\theta_1 - \theta_2) < 0, \qquad \text{whenever } \theta_1 \neq \theta_2$$

and

$$\nabla h(\theta) \cdot (\theta - \tilde{\theta}_n) < 0, \qquad \text{whenever } \theta \neq \tilde{\theta}_n$$

and

$$\nabla h(\theta) \cdot \frac{\theta - \tilde{\theta}_n}{\|\theta - \tilde{\theta}_n\|} < 0, \qquad \text{whenever } \theta \neq \tilde{\theta}_n$$

Since left hand side is continuous function of $\theta$, its supremum over any compact set is negative. Let $-\varepsilon$ be the supremum over the boundary of the ball of radius one centered at $\tilde{\theta}_n$.

# Appendix II (cont.)

For any $\theta_1$ in the exterior of the ball of radius one centered at $\tilde{\theta}_n$, define

$$\theta_2 = \frac{\theta_1 - \tilde{\theta}_n}{\|\theta_1 - \tilde{\theta}_n\|}$$

and

$$\nabla h(\theta_1) \cdot \frac{\theta_1 - \tilde{\theta}_n}{\|\theta_1 - \tilde{\theta}_n\|} = \left[\nabla h(\theta_1) - \nabla h(\theta_2)\right] \cdot \frac{\theta_1 - \tilde{\theta}_n}{\|\theta_1 - \tilde{\theta}_n\|} + \nabla h(\theta_2) \cdot \frac{\theta_1 - \tilde{\theta}_n}{\|\theta_1 - \tilde{\theta}_n\|}$$

The first term on the right hand side is negative because of multivariate monotonicity and the fact that $\theta_1 - \tilde{\theta}_n$ is parallel to $\theta_1 - \theta_2$ and the second term is less than or equal to $-\varepsilon$.

# Appendix II (cont.)

We have now established

$$\nabla h(\theta_1) \cdot \frac{\theta_1 - \tilde{\theta}_n}{\|\theta_1 - \tilde{\theta}_n\|} \leq -\varepsilon, \qquad \text{whenever } \theta_1 \in E$$

where $E$ is the exterior of the ball of radius one centered at $\tilde{\theta}_n$.

Taking the limit superior as $\theta_1 \to \infty$ establishes that $h$ has exponentially light tail behavior.

# Appendix II (cont.)

For an exponential family with bounded sufficient statistic (logistic regression, log-linear models for categorical data), the curvature condition is trivial (Leif's theorem has boundedness as one condition).

For other exponential families (Poisson regression with log link), the curvature condition takes some work. Multivariate monotonicity also implies

$$\frac{\nabla h(\theta)}{\|\nabla h(\theta)\|} \cdot \frac{\theta - \tilde{\theta}_n}{\|\theta - \tilde{\theta}_n\|} < 0, \qquad \text{whenever } \theta \neq \tilde{\theta}_n$$

and the proof starting here is almost the same as the one for exponentially light.

# Appendix III

Johnson, L. T. and Geyer, C. J. (2012)
Variable transformation to obtain geometric ergodicity
in the random-walk Metropolis algorithm.
*Annals of Statistics*, **40**, 3050–3076.

Jarner, S. F. and Hansen, E. (2000).
Geometric ergodicity of Metropolis algorithms.
*Stochastic Processes and their Applications*, **85**, 341–361.

Roberts, G. O. and Rosenthal, J. S. (1997).
Geometric ergodicity and hybrid Markov chains.
*Electronic Journal of Probability*, **2**, 13–25.

# Appendix III (cont.)

Ibragimov, I. A. and Linnik, Yu. V. (1971).
*Independent and Stationary Sequences of Random Variables*.
Walters-Nordhoff, Gronigen.
(Generalized CLT for stationary stochastic processes
is Theorem 18.1.1.)

Tierney, L. (1994).
Markov chains for exploring posterior distributions
(with discussion).
*Annals of Statistics*, **22**, 1701–1762.
(Named RMWA, showed Harris ergodicity).