Markov Chain Monte Carlo

Michael Johannes and Nicholas Polson

Abstract This chapter provides an overview of Markov Chain Monte Carlo (MCMC) methods. MCMC methods provide samples from high-dimensional distributions that commonly arise in Bayesian inference problems. We review the theoretical underpinnings used to construct the algorithms, the Metropolis-Hastings algorithm, the Gibbs sampler, Markov Chain convergence, and provide a number of examples in financial econometrics.

1 Introduction

The Bayesian solution to any inference problem is a simple rule: compute the conditional distribution of unobserved variables given observed data. In financial time series settings, the observed data is asset prices, $y = (y_1, ..., y_T)$, and the unobservables are a parameter vector, θ, and latent variables, $x = (x_1, ..., x_T)$, and the inference problem is solved by $p(\theta, x|y)$, the posterior distribution. The latent variables are either unobserved persistent states such as expected returns or volatility or unobserved transient shocks such as price jump times or sizes.

Characterizing the posterior distribution, however, is often difficult. In most settings $p(\theta, x|y)$ is complicated and high-dimensional, implying that standard sampling methods either do not apply or are prohibitively expensive in terms of computing time. Markov Chain Monte Carlo (MCMC) methods provide a simulation based method for sampling from these high-dimensional

Michael Johannes
Graduate School of Business, Columbia University, 3022 Broadway, NY, 10027, e-mail: mj335@columbia.edu

Nicholas Polson
Graduate School of Business, University of Chicago, 5807 S. Woodlawn, Chicago IL 60637, e-mail: ngp@chicagogsb.edu

T.G. Anderson et al., *Handbook of Financial Time Series*,
DOI: 10.1007/978-3-540-71297-8_43, © Springer-Verlag Berlin Heidelberg 2009

distributions, and are particularly useful for analyzing financial time series models that commonly incorporate latent variables. These samples can be used for estimation, inference, and prediction.

MCMC algorithms generate a Markov chain, $\left\{\theta^{(g)}, x^{(g)}\right\}_{g=1}^{G}$, whose stationary distribution is $p\left(\theta, x|y\right)$. To do this, the first step is the Clifford-Hammersley (CH) theorem, which states that a high-dimensional joint distribution, $p\left(\theta, x|y\right)$, is completely characterized by a larger number of lower dimensional conditional distributions. Given this characterization, MCMC methods iteratively sample from these lower dimensional conditional distributions using standard sampling methods and the Metropolis-Hastings algorithm. Thus, the key to Bayesian inference is simulation rather than optimization.

The simulations are used to estimate integrals via Monte Carlo that naturally arise in Bayesian inference. Common examples include posterior moments of parameters, $E\left[\theta|y\right]$, or state variables, $E\left[x|y\right]$, or even expected utility. Monte Carlo estimates are given by

$$\begin{aligned}\widehat{E}\left(f\left(\theta, x\right)|y\right) &= G^{-1}\sum_{g=1}^{G} f\left(\theta^{(g)}, x^{(g)}\right)\\ &\approx \int f\left(\theta, x\right) p\left(\theta, x|y\right) d\theta dx = E\left(f\left(\theta, x\right)|y\right).\end{aligned}$$

The rest of the chapter is outlined as follows. In Section 2, we explain the components and theoretical foundations of MCMC algorithms. Section 3 provides a few examples from financial econometrics, and Section 4 provides a list of notable references.

2 Overview of MCMC Methods

To develop the foundations of MCMC in the simplest setting, we consider sampling from a bivariate posterior distribution $p\left(\theta_1, \theta_2|y\right)$, and suppress the dependence on the data for parsimony. For intuition, it is useful to think of θ_1 as traditional static parameters and θ_2 as latent variables.

2.1 Clifford–Hammersley theorem

The Clifford-Hammersley theorem (CH) proves that the joint distribution, $p\left(\theta_1, \theta_2\right)$, is completely determined by the conditional distributions, $p\left(\theta_1|\theta_2\right)$ and $p\left(\theta_2|\theta_1\right)$, under a positivity condition. The positivity condition requires that $p\left(\theta_1, \theta_2\right)$, $p\left(\theta_1\right)$ and $p\left(\theta_2\right)$ have positive mass for all points. These results are useful in practice because in most cases, $p\left(\theta_1,\theta_2\right)$ is only known up to proportionality and cannot be directly sampled. CH implies that the same information can be extracted from the lower-dimensional conditional distributions, breaking "curse of dimensionality" by transforming a higher dimensional problem, sampling from $p\left(\theta_1,\theta_2\right)$, into easier problems, sampling from $p\left(\theta_1|\theta_2\right)$ and $p\left(\theta_2|\theta_1\right)$.

The CH theorem is based on the Besag formula: for any pairs (θ_1,θ_2) and $(\theta_1^{'},\theta_2^{'})$,

$$\frac{p\left(\theta_1,\theta_2\right)}{p\left(\theta_1^{'},\theta_2^{'}\right)}=\frac{p(\theta_1|\theta_2^{'})p(\theta_2|\theta_1)}{p\left(\theta_1^{'}|\theta_2^{'}\right)p\left(\theta_2^{'}|\theta_1\right)}. \tag{1}$$

The proof uses the fact that $p\left(\theta_1,\theta_2\right)=p\left(\theta_2|\theta_1\right)p\left(\theta_1\right)$, which, when applied to (θ_1,θ_2) and $(\theta_1^{'},\theta_2^{'})$, implies that

$$p\left(\theta_1\right)=\frac{p\left(\theta_1|\theta_2^{\prime}\right)p\left(\theta_2^{\prime}\right)}{p\left(\theta_2^{\prime}|\theta_1\right)}.$$

The general version of CH follows by analogy. Partitioning a vector as $\theta=(\theta_1,\theta_2,\theta_3,\ldots,\theta_K)$, then the general CH theorem states that

$$p\left(\theta_i|\theta_{-i}\right)\triangleq p\left(\theta_i|\theta_1,\theta_{1,\ldots,}\theta_{i-1},\theta_{i+1},...,\theta_K\right),$$

for $i=1,...,K$, completely characterizes the joint distribution $p\left(\theta_1,...,\theta_K\right)$.

An important case arises frequently in models with latent variables. Here, the posterior is defined over vectors of static fixed parameters, θ, and latent variables, x. In this case, CH implies that $p\left(\theta,x|y\right)$ is completely characterized by $p\left(\theta|x,y\right)$ and $p\left(x|\theta,y\right)$. The distribution $p\left(\theta|x,y\right)$ is the posterior distribution of the parameters, conditional on the observed data and the latent variables. Similarly, $p\left(x|\theta,y\right)$ is the smoothing distribution of the latent variables given the parameters.

2.2 Constructing Markov chains

To construct the Markov chains for MCMC with the appropriate limiting distribution, we use direct sampling methods for known distributions and otherwise use indirect sampling methods such as the Metropolis-Hastings algorithm. First, we describe the indirect methods and then explain the Gibbs sampler and hybrid algorithms, which combine aspects of Metropolis-Hastings and direct sampling methods.

2.2.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm provides a general approach for sampling from a given target density, $\pi(\theta)$. MH uses an accept-reject approach, drawing a candidate from a distribution $q(\theta)$ that is accepted or rejected based on an acceptance probability. Unlike traditional accept-reject algorithms, which repeatedly sample until acceptance, the MH algorithm samples only once at each iteration. If the candidate is rejected, the algorithm keeps the current value. In this original form of the algorithm, the entire vector θ is update at once. Below, modifications are discussed that update θ is blocks, using the intuition from CH.

Specifically, the MH algorithm repeats the following two steps G times: given $\theta^{(g)}$

$$\text{Step 1. Draw } \theta' \text{ from a proposal distribution, } q(\theta'|\theta^{(g)})$$
$$\text{Step 2. Accept } \theta' \text{ with probability } \alpha\left(\theta^{(g)}, \theta'\right),$$

where

$$\alpha\left(\theta^{(g)}, \theta'\right) = \min\left(\frac{\pi(\theta')}{\pi(\theta^{(g)})}\frac{q(\theta^{(g)}|\theta')}{q(\theta'|\theta^{(g)})}, 1\right).$$

To implement the accept-reject step, draw a uniform random variable, $U \sim U[0,1]$, and set $\theta^{(g+1)} = \theta'$ if $U < \alpha\left(\theta^{(g)}, \theta'\right)$, leaving $\theta^{(g)}$ unchanged $(\theta^{(g+1)} = \theta^{(g)})$ otherwise. It is important to note that the denominator in the acceptance probability cannot be zero, provided the algorithm is started from a $\pi-$positive point since q is always positive. The MH algorithm only requires that π can be evaluated up to proportionality.

The output of the algorithm, $\left\{\theta^{(g)}\right\}_{g=1}^{\infty}$, is clearly a Markov chain. The key theoretical property is that the Markov chain, under mild regularity, has $\pi(\theta)$ as its limiting distribution. We discuss two important special cases that depend on the choice of q.

Independence MH

One special case draws a candidate *independently* of the previous state, $q(\theta'|\theta^{(g)}) = q(\theta')$. In this independence MH algorithm, the acceptance criterion simplifies to

$$\alpha\left(\theta^{(g)}, \theta'\right) = \min\left(\frac{\pi(\theta')}{\pi(\theta^{(g)})}\frac{q(\theta^{(g)})}{q(\theta')}, 1\right)$$

Even though θ' is drawn independently of the previous state, the sequence generated is not independent, since α depends on previous draws. The criterion implies a new draw is always accepted if target density ratio, $\pi(\theta')/\pi(\theta^{(g)})$, increases more than the proposal ratio, $q(\theta^{(g)})/q(\theta')$. When this is not satisfied, an balanced coin is flipped to decide whether or not to accept the proposal.

When using independence MH, it is common to pick the proposal density to closely match certain properties of the target distribution. One common criterion is to ensure that tails of the proposal density are thicker than the tails of the target density. By "blanketing" the target density, it is less likely that the Markov chain will get trapped in a low probability region of the state space.

Random-walk Metropolis

Random-walk (RW) Metropolis is the polar opposite of the independence MH algorithm. It draws a candidate from the following RW model,

$$\theta' = \theta^{(g)} + \sigma\varepsilon_{g+1},$$

where ε_t is an independent, mean zero, and symmetric error term, typically taken to be a normal or $t-$distribution, and σ is a scaling factor. The algorithm must be tuned via the choice of σ, the scaling factor. Symmetry implies that

$$q\left(\theta'|\theta^{(g)}\right) = q\left(\theta^{(g)}|\theta'\right),$$

with acceptance probability

$$\alpha\left(\theta^{(g)},\theta'\right) = \min\left(\pi(\theta')/\pi(\theta^{(g)}),1\right).$$

The RW algorithm, unlike the independence algorithm, learns about the density $\pi\left(\theta\right)$ via small symmetric steps, randomly "walks" around the support of π. If a candidate draw has a higher target density value than the current draw, $\pi(\theta') > \pi(\theta^{(g)})$, the draw is always accepted. If $\pi(\theta') < \pi(\theta^{(g)})$, then a unbalanced coin is flipped.

2.2.2 Gibbs sampling

The Gibbs sampler simulates multidimensional posterior distributions by iteratively sampling from the lower-dimensional conditional posteriors. The Gibbs sampler updates the chain one component at a time, instead of updating the entire vector. This requires either that the conditional posteriors distributions are discrete, are a recognizable distribution (e.g. normal) for which standard sampling algorithms apply, or that resampling methods, such as accept-reject, can be used.

In the case of $p\left(\theta_1,\theta_2\right)$, given current draws, $\left(\theta_1^{(g)},\theta_2^{(g)}\right)$, the Gibbs sampler consists of

$$\begin{aligned}&1. \text{ Draw } \theta_1^{(g+1)} \sim p\left(\theta_1|\theta_2^{(g)}\right)\\ &2. \text{ Draw } \theta_2^{(g+1)} \sim p\left(\theta_2|\theta_1^{(g+1)}\right),\end{aligned}$$

repeating G times. The draws generated by the Gibbs sampler form a Markov chain, as the distribution of $\theta^{(g+1)}$ conditional on $\theta^{(g)}$ is independent of past draws. Higher dimensional cases follow by analogy.

2.2.3 Hybrid chains

Given a partition of the vector θ via CH, a hybrid MCMC algorithm updates the chain one subset at a time, either by direct draws ('Gibbs steps') or via MH ('Metropolis step'). Thus, a hybrid algorithm combines the features of the MH algorithm and the Gibbs sampler, providing significant flexibility in designing MCMC algorithms for different models.

To see the mechanics, consider the two-dimensional example. First, assume that the distribution $p(\theta_2|\theta_1)$ is recognizable and can be directly sampled. Second, suppose that $p(\theta_1|\theta_2)$ can only be evaluated and not directly sampled. Thus we use a Metropolis step to update θ_1 given θ_2. For the MH step, the candidate is drawn from $q\left(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)}\right)$, which indicates that the step can depend on the past draw for θ_1. We denote the Metropolis step as $MH\left[q\left(\theta_1|\theta_1^{(g)}, \theta_2^{(g)}\right)\right]$, which implies that we draw $\theta_1^{(g+1)} \sim q\left(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)}\right)$ and then accept/reject based on

$$\alpha\left(\theta_1^{(g)}, \theta_1'\right) = \min\left(\frac{p\left(\theta_1'|\theta_2^{(g)}\right)}{p\left(\theta_1^{(g)}|\theta_2^{(g)}\right)} \frac{q\left(\theta_1^{(g)}|\theta_1', \theta_2^{(g)}\right)}{q\left(\theta_1'|\theta_1^{(g)}, \theta_2^{(g)}\right)}, 1\right).$$

The general hybrid algorithm is as follows. Given $\theta_1^{(g)}$ and $\theta_2^{(g)}$, for $g = 1, ..., G$,

$$\begin{aligned}&1. \text{ Draw } \theta_1^{(g+1)} \sim MH\left[q\left(\theta_1|\theta_1^{(g)}, \theta_2^{(g)}\right)\right]\\ &2. \text{ Draw } \theta_2^{(g+1)} \sim p\left(\theta_2|\theta_1^{(g+1)}\right).\end{aligned}$$

In higher dimensional cases, a hybrid algorithm consists of any combination of Gibbs and Metropolis steps. Hybrid algorithms significantly increase the applicability of MCMC methods, as the only requirement is that the model generates posterior conditionals that can either be sampled or evaluated.

2.3 Convergence theory

To understand why MCMC algorithms work, we briefly discuss convergence of the underlying Markov chain for the case of the Gibbs sampler. The arguments for convergence of MH or hybrid algorithms are similar.

The Markov transition kernel from state θ to state θ' is $\mathbb{P}(\theta, \theta') = p(\theta_1'|\theta_2)\, p(\theta_2'|\theta_1')$,and by definition, $\int \mathbb{P}(\theta, \theta')\, d\theta' = 1$. The densities $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ will typically have either discrete or continuous support, and in nearly all cases the chain can reach any point or set in the state space in one step. To establish convergence, we first identify the limiting distribution. A stationary probability distribution, π, satisfies the integral equation

$$\pi(\theta') = \int \mathbb{P}(\theta, \theta')\, \pi(\theta)\, d\theta.$$

If the chain converges, then π is also called the limiting distribution. It is easy to verify that the stationary distribution of the Markov chain generated by the Gibbs sampler is the posterior distribution, $\pi(\theta) = p(\theta_1, \theta_2)$:

$$\begin{aligned}\int \mathbb{P}(\theta, \theta')\, p(\theta)\, d\theta &= p(\theta_2'|\theta_1') \int_{\theta_2} \int_{\theta_1} p(\theta_1'|\theta_2)\, p(\theta_1, \theta_2)\, d\theta_1 d\theta_2 \\ &= p(\theta_2'|\theta_1') \int_{\theta_2} p(\theta_1'|\theta_2)\, p(\theta_2)\, d\theta_2 \\ &= p(\theta_2'|\theta_1')\, p(\theta_1') = p(\theta_1', \theta_2') = \pi(\theta').\end{aligned}$$

To establish convergence to the limiting distribution, the chain must satisfy certain regularity conditions on how it traverses the state space. Starting from an initial π-positive point, the Markov chain in Gibbs samplers can typically reach any set in the state space in one step, implying that states communicate and the chain is irreducible. This does not imply that a chain starting from a given point, will return to that point or visit nearby states *frequently.* Well-behaved chains are not only irreducible, but stable, in the sense that they make many return visits to states. Chains that visit states or sets frequently are recurrent. Under very mild conditions, the Gibbs sampler generates an irreducible and recurrent chain. In most cases, a measure theoretical condition called Harris recurrence is also satisfied, which implies that the chains converge for any starting values.

In this case, the ergodic theorem holds: for a sufficiently integrable function f and for all starting points θ,

$$\lim_{G \to \infty} \frac{1}{G} \sum_{g=1}^{G} f\left(\theta^{(g)}\right) = \int f(\theta)\, \pi(\theta)\, d\theta = E[f(\theta)]$$

almost surely. Notice the two subtle modes of convergence: there is the convergence of the Markov chain to its stationary distribution, and Monte Carlo convergence, which is the convergence of the partial sums to the integral.

In practice, a chain is typically run for an initial length, often called the burn-in, to remove any dependence on the initial conditions. Once the chain has converged, then a secondary sample of size G is created for Monte Carlo inference.

3 Financial Time Series Examples

While there are many examples of MCMC methods analyzing financial time series models, we focus on just three prominent examples, providing references at the end for other applications.

3.1 Geometric Brownian motion

The geometric Brownian motion is the simplest model,

$$y_t = \mu + \sigma \varepsilon_t,$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ and y_t are continuously compounded returns. The likelihood function is $p\left(y|\mu,\sigma^2\right)$, and $p\left(\mu,\sigma^2|y\right)$ is the joint posterior. We assume independent conjugate priors, $p(\mu) \sim \mathcal{N}(a,A)$ and $p\left(\sigma^2\right) \sim \mathcal{IG}\left(\frac{b}{2},\frac{B}{2}\right)$, where $\mathcal{IG}$ denotes the inverse Gamma distribution, and a, A, b, and B are hyperparameters.

CH implies that $p\left(\mu|\sigma^2,y\right)$ and $p\left(\sigma^2|\mu,y\right)$ are the complete conditionals, which are given by Bayes rule as

$$\begin{aligned} p\left(\mu|\sigma^2,y\right) &\propto p\left(y|\mu,\sigma^2\right)p(\mu) \\ p\left(\sigma^2|\mu,y\right) &\propto p\left(y|\mu,\sigma^2\right)p\left(\sigma^2\right). \end{aligned}$$

Straightforward algebra implies that

$$p\left(\mu|y,\sigma^2\right) \sim \mathcal{N}\left(a^T,A^T\right) \text{ and } p\left(\sigma^2|y,\mu\right) \sim \mathcal{IG}\left(\frac{b^T}{2},\frac{B^T}{2}\right),$$

where

$$\begin{aligned} a^T &= A^T\left(\frac{\overline{y}}{\sigma^2/T}+\frac{a}{A}\right),\ A^T = \left(\frac{1}{\sigma^2/T}+\frac{1}{A}\right)^{-1} \\ b^T &= b+T \text{ and } B^T = B+\sum\nolimits_{t=1}^{T}\left(y_t-\mu\right)^2, \end{aligned}$$

where $T^{-1}\sum_{t=1}^{T} y_t = \overline{y}$.

The fact that the conditional posterior is the same distribution (with different parameters) as the prior distribution is a property of the prior known as conjugacy.

Since both distributions are standard distributions, the MCMC algorithm is a two-step Gibbs sampler. Given current draws, $\left(\mu^{(g)}, \left(\sigma^2\right)^{(g)}\right)$, the algorithm iteratively simulates

$$\begin{aligned} &1.\text{ Draw } \mu^{(g+1)} \sim p\left(\mu | \left(\sigma^2\right)^{(g)}, y\right) \sim \mathcal{N} \\ &2.\text{ Draw } \left(\sigma^2\right)^{(g+1)} \sim p\left(\sigma^2 | \mu^{(g+1)}, y\right) \sim \mathcal{IG}. \end{aligned}$$

This example is meant to develop intuition. In most cases, one would chose a dependent prior of the form

$$p\left(\mu, \sigma^2\right) \propto p\left(\mu|\sigma^2\right) p\left(\sigma^2\right),$$

where $p\left(\mu|\sigma^2\right) \sim \mathcal{N}$ and $p\left(\sigma^2\right) \sim \mathcal{IG}$. This is known as the $\mathcal{NIG}$ is the normal-inverse gamma joint prior. In this case, MCMC is not required as one can draw directly from $p\left(\mu, \sigma^2|y\right)$.

3.2 Time-varying expected returns

Next, consider a model with time-varying expected returns,

$$\begin{aligned} y_t &= \mu_t + \sigma\varepsilon_t \\ \mu_t &= \alpha_\mu + \beta_\mu \mu_{t-1} + \sigma_\mu \varepsilon_t. \end{aligned}$$

The parameter vector is $\theta = \left(\sigma^2, \alpha_\mu, \beta_\mu, \sigma_\mu^2\right)$ and the state variables are $\mu = (\mu_1, ..., \mu_T)$. We assume standard conjugate priors, $\sigma^2 \sim \mathcal{IG}$ and $(\alpha_\mu, \beta_\mu, \sigma_\mu) \sim \mathcal{NIG}$, suppressing the parameter of these distributions. CH implies that $p\left(\sigma^2|\alpha_\mu, \beta_\mu, \sigma_\mu^2, \mu, y\right)$, $p\left(\alpha_\mu, \beta_\mu, \sigma_\mu^2|\sigma^2, \mu, y\right)$, and $p\left(\mu|\sigma^2, \alpha_\mu, \beta_\mu, \sigma_\mu^2, y\right)$ are the complete conditionals.

The Gibbs sampler for this model is given by:

$$\begin{aligned} &1.\ \left(\sigma^2\right)^{(g+1)} \sim p\left(\sigma^2 | \alpha_\mu^{(g)}, \beta_\mu^{(g)}, \left(\sigma_\mu^2\right)^{(g)}, \mu^{(g)}, y\right) \sim \mathcal{IG} \\ &2.\ \left(\alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}, \left(\sigma_\mu^2\right)^{(g+1)}\right) \sim p\left(\alpha_\mu, \beta_\mu, \sigma_\mu^2 | \left(\sigma^2\right)^{(g+1)}, \mu^{(g)}, y\right) \sim \mathcal{NIG} \\ &3.\ \mu^{(g+1)} \sim p\left(\mu | \left(\sigma^2\right)^{(g+1)}, \alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}, \left(\sigma_\mu^2\right)^{(g+1)}, y\right) \sim FFBS, \end{aligned}$$

where the third step refers to the forward-filtering, backward sampling algorithm. This algorithm applies in conditionally Gaussian state space models, and requires three steps:

Step 1. Run the Kalman filter forward for $t = 1, ..., T$ to get the moments of $p\left(\mu_t|\theta, y^t\right)$

Step 2. Sample the last state from $\widehat{\mu}_T \sim p\left(\mu_T|\theta, y^T\right)$

Step 3. Sample backward through time: $\widehat{\mu}_t \sim p\left(\mu_t|\widehat{\mu}_{t+1}, \theta, y^t\right)$.

where $y^t = (y_1, ..., y_t)$. The FFBS algorithm provides a direct draw of the vector μ from its conditional distribution, which is more efficient than sampling the expected returns, μ_t, one state at a time.

The output of the algorithm can be used for Monte Carlo integration. For example, the smoothed estimate of the latent state at time t is given by

$$\frac{1}{G}\sum_{g=1}^{G}\mu_t^{(g)} \approx \int \mu_t p\left(\mu_t|y\right) d\mu_t = E\left(\mu_t|y\right).$$

3.3 Stochastic volatility models

A popular discrete-time stochastic volatility model is given by

$$\begin{aligned} y_t &= \sqrt{V_{t-1}}\varepsilon_t \\ \log\left(V_t\right) &= \alpha_v + \beta_v \log\left(V_{t-1}\right) + \sigma_v \varepsilon_t^v, \end{aligned}$$

where, for simplicity, we assume the errors are uncorrelated. Again, a $\mathcal{NIG}$ prior for $\left(\alpha_v, \beta_v, \sigma_v^2\right)$ is conjugate for the parameters, conditional on the volatilities.

The only difficulty in this model is sampling from $p\left(V|\alpha_v, \beta_v, \sigma_v^2, y\right)$. This distribution is not a recognizable distribution, and due to its high dimension, a direct application of MH is not recommended. The simplest approach is to use the CH theorem to break the T-dimensional distribution $p\left(V|\alpha_v, \beta_v, \sigma_v^2, y\right)$ into T 1-dimensional distributions,

$$p\left(V_t|V_{t-1}, V_{t+1}, \theta, y_{t+1}\right) \propto p\left(y_{t+1}|V_t\right) p\left(V_{t+1}|V_t, \theta\right) p\left(V_t, |V_{t-1}, \theta\right),$$

for $t = 1, ..., T$. This distribution is again not recognizable, but it is easy to develop proposal distributions that closely approximate the distribution using independence MH, although the random-walk algorithm also applies and works well in practice. This is typically referred to as a single-state volatility updating step.

Thus, the hybrid MCMC algorithm for estimating the stochastic volatility requires the following steps: given

1. $\left(\alpha_v^{(g+1)}, \beta_v^{(g+1)}, \left(\sigma_v^2\right)^{(g+1)}\right) \sim p\left(\alpha_v, \beta_v, \sigma_v^2 | V^{(g)}, y\right) \sim \mathcal{NIG}$
2. $V_t^{(g+1)} \sim MH\left[q\left(V_t | V_{t-1}^{(g)}, V_t^{(g)}, V_{t+1}^{(g)}, \theta^{(g+1)}\right)\right]$ for $t = 1, ..., T$.

When implementing this model, care needs to be taken with the Metropolis step. It is common to try alternative proposal distribution and perform simulation studies to ensure the algorithm is working properly.

4 Further Reading

For a textbook discussion of the Bayesian approach to inference, we recommend the books by Raiffa and Schlaifer (1961), Bernardo and Smith (1995), Robert (2001), or O'Hagan (2004). Robert and Casella (2005) or Gamerman and Lopes (2006) provide excellent textbook treatments of MCMC methods. They provide with details regarding the algorithms (e.g., tuning MH algorithms) and numerous examples.

It is impossible to cite all of the important papers developing MCMC theory and building MCMC algorithms in different applications. We here provide the briefest possible list, with an emphasis on the initial MCMC approaches for various different models. The extensions to these foundational papers are numerous.

One important precursor to MCMC methods in Bayesian statistics is Tanner and Wong (1987), who introduced algorithms using data augmentation. Gelfand and Smith (1990) provided the first MCMC applications in Bayesian statistics. Smith and Roberts (1993) and Besag, Green, Higdon, and Mengersen (1995) provide overviews of MCMC methods.

Regarding the underlying theory of MCMC algorithms, The Clifford-Hammersley theorem was originally shown in Hammersley and Clifford (1970) and the Besag formula is in Besag (1974). The original Metropolis random-walk algorithm is given in Metropolis et al. (1953), and the independence version in Hastings (1973). Geman and Geman (1984) introduced the Gibbs sampler for sampling posterior distributions and proved convergence properties. Tierney (1994) provides a wide range of theoretical convergence results for MCMC algorithms, providing verifiable conditions for various forms of convergence and discussing hybrid algorithms. Chib and Greenberg (1995) provide an overview of the Metropolis-Hastings algorithm.

With regard to specific models, there are a number of important foundational references. For simplicity, we list them in chronological order. Carlin and Polson (1991) developed MCMC algorithms for models with scale mixture of normal distribution errors, which includes the t, double exponential, logistic, and exponential power family error distributions. Carlin and Polson (1992) and develop MCMC algorithms for discrete regression and categorical observations and for the probit model, see Albert and Chib (1993). Carlin, Gelfand, and Smith (1992) and Chib (1998) developed algorithms for time series models with change-points. Diebold and Robert (1994) analyzed finite mixture models with MCMC methods. Carlin, Polson, and Stoffer (1992) develop MCMC methods for nonlinear and non-normal state space models, and Carter and Kohn (1994, 1996) developed the FFBS algorithm for estimation in a range of non-normal state space models. McCulloch and Tsay (1993) analyze Markov switching models.

MCMC methods have been broadly applied in stochastic volatility models. Jacquier, Polson, and Rossi (1994) first developed MCMC algorithms for the log-stochastic volatility models, with Jacquier, Polson, and Rossi (2004) providing extensions to correlated and non-normal error distributions. Eraker, Johannes, and Polson (2003) analyzed time series models with jumps in prices and volatility. Jones (1998), Eraker (2001) and Elerian, Shephard, and Chib (2001) develop approaches for MCMC analysis of continuous-time models by simulating additional high-frequency data points between observations. Also, see the chapter by Chib, Omori, and Asai in this handbook for further references for multivariate problems. For a more extensive review of MCMC methods for financial econometrics, see Johannes and Polson (2005).

References

Albert, J. and Chib, S. (1993): Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88**, 669–679.

Bernardo, J. and Smith, A. (1995): *Bayesian Theory* Wiley, New York.

Besag, J. (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* **36**, 192–326.

Besag, J. Green, E., Higdon, D. and Mengersen, K. (1995): Bayesian computuation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Carlin, B. and Polson, N. (1991): Inference for Nonconjugate Bayesian Models using the Gibbs sampler. *Canadian Journal of Statistics* **19**, 399–405.

Carlin, B. and Polson, N. (1992): Monte Carlo Bayesian Methods for Discrete Regression Models and Categorical Time Series. *In: Bernardo, J.M. et al (Eds.): Bayesian Statistics* **4**, 577–586. Oxford University Press, Oxford.

Carlin, B., Polson, N. and Stoffer, D. (1992): A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association* **87**, 493–500.

Carlin, B., Gelfand, A. and Smith, A. (1992): Hierarchical Bayesian analysis of change point process. *Applied Statistics, Series C* **41**, 389–405.

Carter, C., and Kohn, R. (1994): On Gibbs Sampling for State Space Models. *Biometrika* **81**, 541–553.

Carter, C. and Kohn,R. (1996): Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika* **83**, 589–601.

Chib, S. (1998): Estimation and Comparison of Multiple Change Point Models. *Journal of Econometrics* **86**, 221–241.

Chib, S. and Greenberg, E. (1995): Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49**, 327–335.

Diebolt, J. and Robert, C. (1994): Estimation of finite mixture distributions by Bayesian sampling. *Journal of the Royal Statistical Society Series B* **56**, 363–375.

Elerian, O., Shephard, N. and Chib, S. (2001): Likelihood inference for discretely observed non-linear diffusions. *Econometrica* **69**, 959–993.

Eraker, B. (2001): MCMC Analysis of Diffusion Models with Applications to Finance. *Journal of Business and Economic Statistics* **19**, 177–191.

Eraker, B., Johannes, M. and Polson, N. (2003): The Impact of Jumps in Equity Index Volatility and Returns. *Journal of Finance* **58**, 1269–1300.

Gamerman, D. and Lopes, H. (2006): *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* Boca Raton, Chapman & Hall/CRC.

Gelfand, A. and Smith, A. (1990): Sampling Based approaches to calculating Marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Geman, S. and Geman, D. (1984): Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hammersley, J. and Clifford, P. (1970): *Markov fields on finite graphs and lattices. Unpublished Manuscipt.*

Hastings, W. K. (1970): Monte Carlo sampling Methods using Markov Chains and their Applications. *Biometrika* **57**, 97–109.

Jacquier, E., Polson, N. and Rossi, P. (1994): Bayesian analysis of Stochastic Volatility Models (with discussion). *Journal of Business and Economic Statistics* **12**, 371–417.

Jacquier, E., Polson, N. and Rossi, P. (2004): Bayesian Inference for SV models with Correlated Errors. *Journal of Econometrics* forthcoming.

Johannes, M. and Polson, N. (2005): MCMC methods for Financial Econometrics. *In: Ait-Sahalia, Y. and Hansen, L. (Eds.): Handbook of Financial Econometrics* forthcoming.

Jones, C. (1998): Bayesian Estimation of Continuous-Time Finance Models. *Working paper.*

McCulloch, R. and Tsay, R. (1993): Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series. *Journal of the American Statistical Association* **88**, 968–978.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953): Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21**, 1087–1091.

O'Hagan, A. and Forster, J. (2004): *Kendall's Advanced Theory of Statistics: Bayesian Inference.* **2**, Hodder-Arnold.

Raiffa, H. and Schlaifer, R. (1961): *Applied Statistical Decision Theory* Harvard University, Boston, MA.

Robert, C. (2001): *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, Springer-Verlag, New York.

Robert, C. and Casella, G. (2005): *Monte Carlo Statistical Methods* New York, Springer.

Smith, A. and Gareth, R. (1993): Bayesian Computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Association Series B* **55**, 3–23.

Tanner, M. and Wong, W. (1987): The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Tierney, L. (1994): Markov Chains for exploring Posterior Distributions (with discussion). *Annals of Statistics* **22**, 1701–1786.