MCMC Methods for Continuous-Time Financial Econometrics

Michael Johannes and Nicholas Polson*

January 25, 2006

Abstract

This chapter develops Markov Chain Monte Carlo (MCMC) methods for Bayesian inference in continuous-time asset pricing models. The Bayesian solution to the inference problem is the distribution of parameters and latent variables conditional on observed data, and MCMC methods provide a tool for exploring these high-dimensional, complex distributions. We first provide a description of the foundations and mechanics of MCMC algorithms. This includes a discussion of the Clifford-Hammersley theorem, the Gibbs sampler, the Metropolis-Hastings algorithm, and theoretical convergence properties of MCMC algorithms. We next provide a tutorial on building MCMC algorithms for a range of continuous-time asset pricing models. We include detailed examples for equity price models, option pricing models, term structure models, and regime-switching models. Finally, we discuss the issue of sequential Bayesian inference, both for parameters and state variables.

*We would especially like to thank Chris Sims and the editors, Yacine Ait-Sahalia and Lars Hansen. We also thank Mark Broadie, Mike Chernov, Anne Gron, Paul Glasserman, and Eric Jacquier for their helpful comments. Johannes is at the Graduate School of Business, Columbia University, 3022 Broadway, NY, NY, 10027, mj335@columbia.edu. Polson is at the Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago IL 60637, ngp@gsb.uchicago.edu.

Contents

1 Introduction **4**

2 Overview of Bayesian Inference and MCMC **7**

2.1 MCMC Simulation and Estimation 8

2.2 Bayesian Inference 9

3 MCMC: Methods and Theory **12**

3.1 Clifford-Hammersley Theorem 12

3.2 Gibbs Sampling 14

3.3 Metropolis-Hastings 15

3.4 Convergence Theory 20

3.4.1 Convergence of Markov Chains 20

3.4.2 Convergence of MCMC algorithms 21

3.5 MCMC Algorithms: Issues and Practical Recommendations 26

4 Bayesian Inference and Asset Pricing Models **30**

4.1 States Variables and Prices 31

4.2 Time-discretization: computing $p(Y|X,\Theta)$ and $p(X|\Theta)$ 34

4.3 Parameter Distribution 37

5 Asset Pricing Applications **39**

5.1 Equity Asset Pricing Models 39

5.1.1 Geometric Brownian Motion 39

5.1.2 Black-Scholes 41

5.1.3 A Multivariate Version of Merton's Model 44

5.1.4 Time-Varying Equity Premium 50

5.1.5 Log-Stochastic Volatility Models 54

5.1.6 Alternative Stochastic Volatility Models 58

5.2 Term Structure Models 63

5.2.1 Vasicek's Model 64

5.2.2 Vasicek with Jumps 67

5.2.3 The CIR model 70

5.3 Regime Switching Models 72

6 Sequential Inference: Filtering **75**
6.1 The Particle Filter 76
6.1.1 Adapting the particle filter to continuous-time models 79
6.2 Practical Filtering 82

7 Conclusions and Future Directions **83**

8 References **85**

1 Introduction

Dynamic asset pricing theory uses arbitrage and equilibrium arguments to derive the functional relationship between asset prices and the fundamentals of the economy: state variables, structural parameters and market prices of risk. Continuous-time models are the centerpiece of this approach due to their analytical tractability. In many cases, these models lead to closed form solutions or easy to solve differential equations for objects of interest such as prices or optimal portfolio weights. The models are also appealing from an empirical perspective: through a judicious choice of the drift, diffusion, jump intensity and jump distribution, these models accommodate a wide range of dynamics for state variables and prices.

Empirical analysis of dynamic asset pricing models tackles the *inverse problem:* extracting information about latent state variables, structural parameters and market prices of risk from observed prices. The Bayesian solution to the inference problem is the distribution of the parameters, Θ, and state variables, X, conditional on observed prices, Y. This posterior distribution, $p(\Theta, X|Y)$, combines the information in the model and the observed prices and is the key to inference on parameters and state variables.

This chapter describes Markov Chain Monte Carlo (MCMC) methods for exploring the posterior distributions generated by continuous-time asset pricing models. MCMC samples from these high-dimensional, complex distributions by generating a *Markov Chain* over (Θ, X), $\left\{\Theta^{(g)}, X^{(g)}\right\}_{g=1}^{G}$, whose equilibrium distribution is $p(\Theta, X|Y)$. The *Monte Carlo* method uses these samples for numerical integration for parameter estimation, state estimation and model comparison.

Characterizing $p(\Theta, X|Y)$ in continuous-time asset pricing models is difficult for a variety of reasons. First, prices are observed discretely while the theoretical models specify that prices and state variables evolve continuously in time. Second, in many cases, the state variables are latent from the researcher's perspective. Third, $p(\Theta, X|Y)$ is typically of very high dimension and thus standard sampling methods commonly fail. Fourth, many continuous-time models of interest generate transition distributions for prices and state variables that are non-normal and non-standard, complicating standard estimation methods such as MLE or GMM. Finally, in term structure and option pricing models, parameters enter nonlinearly or even in a non-analytic form as the implicit solution to ordinary or partial differential equations. We show that MCMC methods tackle all of these issues.

To frame the issues involved, it is useful to consider the following example: Suppose on

$(\Omega, \mathcal{F}, \mathbb{P})$ an asset price, S_t, and its stochastic variance, V_t, jointly solve:

$$dS_t = S_t (r_t + \mu_t) dt + S_t \sqrt{V_t} dW_t^s (\mathbb{P}) + d\left(\sum_{j=1}^{N_t(\mathbb{P})} S_{\tau_j -} \left(e^{Z_j(\mathbb{P})} - 1 \right) \right) - \mu_t^{\mathbb{P}} S_t dt \quad (1)$$

$$dV_t = \kappa_v (\theta_v - V_t) dt + \sigma_v \sqrt{V_t} dW_t^v (\mathbb{P}) \quad (2)$$

where $W_t^s (\mathbb{P})$ and $W_t^v (\mathbb{P})$ are Brownian motions, $N_t (\mathbb{P})$ counts the number of jump times, τ_j, prior to time t, μ_t is the equity risk premium, $\mu_t^{\mathbb{P}} S_t$ is the jump compensator, $Z_j (\mathbb{P})$ are the jump sizes, and r_t is the spot interest rate. Researchers also often observe derivative prices, such as options. To price these derivatives, asset pricing theory asserts the existence of a probability measure, $\mathbb{Q}$, such that

$$dS_t = r_t S_t dt + S_t \sqrt{V_t} dW_t^s (\mathbb{Q}) + d\left(\sum_{j=1}^{N_t(\mathbb{Q})} S_{\tau_j -} \left(e^{Z_j(\mathbb{Q})} - 1 \right) \right) - \mu_t^{\mathbb{Q}} S_t dt$$

$$dV_t = \left[\kappa_v (\theta_v - V_t) + \lambda_v^{\mathbb{Q}} V_t \right] dt + \sigma_v \sqrt{V_t} dW_t^v (\mathbb{Q})$$

where all random variables are now defined on $(\Omega, \mathcal{F}, \mathbb{Q})$. Here $\lambda_v^{\mathbb{Q}}$ is the diffusive "price of volatility risk," and $\mu_t^{\mathbb{Q}}$ is the jump compensator. Under $\mathbb{Q}$, the price of a call option on S_t maturing at time T, struck at K, is

$$C_t = C (S_t, V_t, \Theta) = E^{\mathbb{Q}} \left[\exp\left(- \int_t^T r_s ds \right) (S_T - K)_+ | V_t, S_t, \Theta \right] \quad (3)$$

where $\Theta = \left(\Theta^{\mathbb{P}}, \Theta^{\mathbb{Q}} \right)$ are the structural and risk neutral parameters. The state variables, X, consist of the volatilities, the jump times and jump sizes.

The goal of empirical asset pricing is to learn about the risk neutral and objective parameters, the state variables, namely, volatility, jump times and jump sizes, and the model specification from the observed equity returns and option prices. In the case of the parameters, the marginal posterior distribution $p(\Theta|Y)$ characterizes the sample information about the objective and risk-neutral parameters and quantifies the estimation risk: the uncertainty inherent in estimating parameters. For the state variables, the marginal distribution, $p(X|Y)$, combines the model and data to provide a consistent approach for separating out the effects of jumps from stochastic volatility. This is important for empirical problems such as option pricing or portfolio applications which require volatility estimates. Classical methods are difficult to apply in this model as the parameters and volatility enter in a non-analytic manner in the option pricing formula, volatility, jump times and jump sizes are latent, and the transition density for observed prices is not known.

To design MCMC algorithms for exploring $p(\Theta, X|Y)$, we first follow Duffie (1996) and interpret asset pricing models as state space models. This interpretation is convenient for constructing MCMC algorithms as it highlights the modular nature of asset pricing models. The observation equation is the distribution of the observed asset prices conditional on the state variables and parameters while the evolution equation consists of the dynamics of state variables conditional on the parameters. In the example above, (1) and (3) form the observation equations and (2) is the evolution equation. Viewed in this manner, all asset pricing models take the general form of nonlinear, non-Gaussian state space models.

MCMC methods are particularly well-suited for continuous-time finance applications for several reasons.

1. Continuous-time asset models specify that prices and state variables solve parameterized stochastic differential equations (SDEs) which are built from Brownian motions, Poisson processes and other i.i.d. shocks whose distributions are easy to characterize. When discretized at any finite time-interval, the models take the form of familiar time series models with normal, discrete mixtures of normals or scale mixtures of normals error distributions. This implies that the standard tools of Bayesian inference directly apply to these models.

2. MCMC is a unified estimation procedure, simultaneously estimating both parameters and latent variables. MCMC directly computes the distribution of the latent variables and parameters given the observed data. This is a stark alternative the usual approach in the literature of applying approximate filters or noisy latent variable proxies. This allows the researcher, for example, to separate out the effects of jumps and stochastic volatility in models of interest rates or equity prices using discretely observed data.[1]

3. MCMC methods allow the researcher to quantify estimation and model risk. Estimation risk is the inherent uncertainty present in estimating parameters or state variables, while model risk is the uncertainty over model specification. Increasingly in practical problems, estimation risk is a serious issue whose impact must be quantified. In the case of option pricing and optimal portfolio problems, Merton (1980)

[1] Alternative approaches to separating out jumps and stochastic volatility rely on decreasing interval estimators. See, for example, Aït-Sahalia (2003), Barndorff-Nielson and Shephard (2002), and Andersen, Bollerslev, Diebold (2002).

argues that the "most important direction is to develop accurate variance estimation models which take into account of the errors in variance estimates" ($p.$ 355).

4. MCMC is based on conditional simulation, therefore avoiding any optimization or unconditional simulation. From a practical perspective, MCMC estimation is typically extremely fast in terms of computing time. This has many advantages, one of which is that it allows the researcher to perform simulation studies to study the algorithms accuracy for estimating parameters or state variables, a feature not shared by many other methods.

Armed with these tools for posterior simulation, we also analyze the problem of sequential Bayesian inference: iteratively computing the posterior distribution as additional data arrives. We discuss two approaches for sequential inference. The first, the particle filter, discretizes the filtering density and sequentially computes the posterior distribution by resampling the particles via MCMC or other simulation methods. This approach is particularly well suited for filtering state variables in continuous-time models. We also briefly discuss an alternative to the particle filter, the "practical" filter, which is a pure MCMC method for sequential estimation.

The rest of the chapter is outlined as follows. Section 2 provides a brief, non-technical overview of Bayesian inference and MCMC methods. Section 3 describes the mechanics of MCMC algorithms, provides an overview of the limiting properties of MCMC algorithms, and provides practical recommendations for implementing MCMC algorithms. Section 4 discusses the generic problem of Bayesian inference in continuous-time models. Section 5 provides a tutorial on MCMC methods, building algorithms for equity price, option price, term structure and regime switching models. Section 6 discusses filtering methods. Section 7 concludes and provides directions for future research.

2 Overview of Bayesian Inference and MCMC

This section provides a brief, nontechnical overview of MCMC and Bayesian methods. We first describe the mechanics of MCMC simulation and then we show how to use MCMC methods to compute objects of interest in Bayesian inference.

2.1 MCMC Simulation and Estimation

MCMC generates random samples from a given target distribution, in our case, the distribution of parameters and state variables given the observed prices, $p(\Theta, X|Y)$. One way to motivate the construction of MCMC algorithms is via a result commonly known as the Clifford-Hammersley theorem. The theorem states that a joint distribution can be characterized by its so-called complete conditional distributions. Specifically, the theorem implies that $p(X|\Theta, Y)$ and $p(\Theta|X, Y)$ completely characterize the joint distribution $p(\Theta, X|Y)$.

MCMC provides the recipe for combining the information in these distributions to generate samples from $p(\Theta, X|Y)$. Consider the following algorithm. Given two initial values, $\Theta^{(0)}$ and $X^{(0)}$, draw $X^{(1)} \sim p\left(X|\Theta^{(0)}, Y\right)$ and then $\Theta^{(1)} \sim p\left(\Theta|X^{(1)}, Y\right)$. Continuing in this fashion, the algorithm generates a sequence of random variables, $\left\{X^{(g)}, \Theta^{(g)}\right\}_{g=1}^{G}$. This sequence is not *i.i.d.*, but instead forms a *Markov Chain* with attractive properties: under a number of metrics and mild conditions, the distribution of the chain converges to $p(\Theta, X|Y)$, the target distribution.

The key to MCMC is that it is typically easier to characterize the complete conditional distributions, $p(\Theta|X, Y)$ and $p(X|\Theta, Y)$, then to directly analyze the higher-dimensional joint distribution, $p(\Theta, X|Y)$. In many models, the distribution of the state variables conditional on parameters and data, $p(X|\Theta, Y)$, can be computed using standard filtering and smoothing techniques. For example, in linear and Gaussian models, the Kalman filter generates samples from $p(X|\Theta, Y)$. Moreover, the distribution of the parameters given observed data and state variables, $p(\Theta|X, Y)$, is typically easy to simulate as it conditions on the latent states.

MCMC algorithms generically consist of two different steps. If the complete conditional distribution is known in closed form and can be directly sampled, the step in the MCMC algorithm is known as a "Gibbs" step. If all the conditionals can be directly sampled, the algorithm is referred to as a "*Gibbs sampler*." In many situations, one or more of the conditionals cannot be directly sampled and methods known as "*Metropolis-Hastings algorithms*" apply. These algorithms sample a candidate draw from a proposal density and then accept or reject the candidate draw based on an acceptance criterion. These algorithms generate random samples that form a Markov Chain with the appropriate equilibrium distribution. An algorithm can include only Gibbs steps, only Metropolis-Hastings steps or any combination of the two. This latter case, usually encountered in practice, generates a "hybrid" MCMC algorithm.

The samples $\left\{\Theta^{(g)}, X^{(g)}\right\}_{g=1}^{G}$ from the joint posterior can be used for parameter and state variable estimation using the *Monte Carlo* method. For a function $f(\Theta, X)$ satisfying technical regularity conditions, the Monte Carlo estimate of

$$E\left[f(\Theta, X)|Y\right] = \int f(\Theta, X)\, p(\Theta, X|Y)\, dX d\Theta$$

is given by $\frac{1}{G}\sum_{g=1}^{G} f\left(\Theta^{(g)}, X^{(g)}\right)$.

MCMC algorithms have attractive limiting behavior as $G \to \infty$. There are two types of convergence operating simultaneously. First, there is the convergence of the distribution of the Markov Chain to $p(\Theta, X|Y)$. Second, there is the convergence of the partial sums, $\frac{1}{G}\sum_{g=1}^{G} f\left(\Theta^{(g)}, X^{(g)}\right)$ to the conditional expectation $E\left[f(\Theta, X)|Y\right]$. The Ergodic Theorem for Markov Chains guarantees both types of convergence, and the conditions under which it holds can be generically verified for MCMC algorithms. In many cases, these limiting results can often be sharpened by deriving the rate of convergence of the Markov chain and geometric convergence rates are common. We discuss these issues in detail in Section 3.4.

2.2 Bayesian Inference

We now provide a brief, nontechnical overview of Bayesian inference. We refer the reader to Lindley (1972) or Bernardo and Smith (1995) for textbook treatments of Bayesian methods. The main advantage of Bayesian methods are the strong theoretical foundations of the Bayesian approach to inference and decision making. Bayesian inference provides a coherent approach for inference and is merely an implication of the laws of probability applied to the parameters and state variables. This approach is consistent with axiomatic decision theory. See, for example, the seminal work of Ramsey (1931), de Finetti (1931) and Savage (1954). We now discuss the key elements of Bayesian inference and decision-making problems.

The posterior distribution

The posterior distribution summarizes the information embedded in prices regarding latent state variables and parameters. Bayes rule factors the posterior distribution into is constituent components:

$$p(\Theta, X|Y) \propto p(Y|X, \Theta)\, p(X|\Theta)\, p(\Theta), \tag{4}$$

where $Y = \{Y_t\}_{t=1}^{T}$ are the observed prices, $X = \{X_t\}_{t=1}^{T}$ are the unobserved state variables, Θ are the parameters, $p(Y|X, \Theta)$ is the likelihood function, $p(X|\Theta)$ is the distribution of

the state variables, and $p(\Theta)$ is the distribution of the parameters, commonly called the prior. The parametric asset pricing model generates $p(Y|X,\Theta)$ and $p(X|\Theta)$ and $p(\Theta)$ summarizes any non-sample information about the parameters.

The Likelihood

There are two types of likelihood functions of interest. The distribution $p(Y|X,\Theta)$ is the full-information (or data-augmented) likelihood and conditions on the state variables and parameters. This is related to marginal likelihood function, $p(Y|\Theta)$, which integrates the latent variables from the augmented likelihood:

$$p(Y|\Theta) = \int p(Y,X|\Theta)\, dX = \int p(Y|X,\Theta)\, p(X|\Theta)\, dX.$$

In most models continuous-time asset pricing models, $p(Y|\Theta)$ is not available in closed form and simulation methods are required to perform likelihood-based inference. On the other hand, the full-information likelihood is usually known in closed form which is a key to MCMC estimation.

The Prior Distribution

The prior distribution, as an implication of Bayes rule, enters in the posterior distribution in (4). It is important to recognize that the importance of $p(\Theta)$ cannot be ignored: its presence in the posterior, like the presence of the likelihood, is merely an implication of the laws of probability. Additionally, this distribution serves important economic and statistical roles. The prior allows the researcher to incorporate nonsample information in a consistent manner. For example, the prior provides a consistent mechanism to impose important economic information such as positivity of certain parameters or beliefs over the degree of mispricing in a model. Statistically, the prior can impose stationarity, rule out near unit-root behavior, or separate mixture components, to name a few applications.

Expected Utility

The posterior distribution and Bayesian inference are only the first steps in a more complicated decision making process. When faced with a decision problem in the presence of uncertainty, a rational decision maker chooses an action, a, to maximize expected utility $E[U]$, where

$$E[U] = \int U(a,\Theta,X)\, p(\Theta,X|Y)\, d\Theta dX$$

where $U(a, \Theta, X)$ is the utility in state X, with parameter Θ, and for action a. When making decisions, a rational decision maker takes into account the uncertainty in the parameters and states by integrating out the uncertainty in these quantities and then maximizing expected utility by choosing the appropriate action. This shows the central role of the posterior distribution in decision making problems. For an overview of decision making in econometrics from a Bayesian perspective, see, for example, Chamberlain (2001).

Marginal Parameter Posterior

The information contained in the observed data regarding an individual parameter is summarized via the marginal posterior distribution

$$p(\Theta_i|Y) = \int p\left(\Theta_i, \Theta_{(-i)}, X|Y\right) dX d\Theta_{(-i)} \tag{5}$$

where Θ_i is the i^{th} element of the parameter vector and $\Theta_{(-i)}$ denotes the remaining parameters. The marginal posterior provides estimates (posterior means or medians) and characterizes estimation risk (posterior standard deviations, quantiles or credible sets).

State Estimation

State estimation is similar to parameter inference, but it is now important to focus on a number of different posteriors, depending on how much conditioning information is used. The following posterior distributions are all of interest:

$$\begin{aligned} \text{Smoothing} &: p\left(X_t|Y^T\right) \ t = 1, ..., T \\ \text{Filtering} &: p\left(X_t|Y^t\right) \ t = 1, ..., T \\ \text{Forecasting} &: p\left(X_{t+1}|Y^t\right) \ t = 1, ..., T. \end{aligned}$$

Here Y^t denotes the observed prices up to time t. The smoothing problem is a static problem, solved once using all of the data, the filtering and forecasting problems are inherently sequential.

The key to filtering latent states is once again Bayes rule which decomposes the filtering density into its components:

$$p\left(X_t|Y^t\right) \propto \int p(Y_t|X_t)\, p(X_t|X_{t-1})\, p\left(X_{t-1}|Y^{t-1}\right) dX_{t-1}.$$

Here $p(Y_t|X_t)$ is the likelihood, $p(X_t|X_{t-1})$ is the state evolution and $p(X_{t-1}|Y^{t-1})$ is the "prior" representing knowledge of the past states given prior price information. Simulation

based filtering methods such as the particle and practical filter provide computationally tractable approaches to approximate the filtering density, see Section 6.

Model Specification

The posterior distribution provides both formal and informal methods to evaluate model specification and to compare different models. Informally, the posterior can be used to analyze the in-sample fit. For example, the posterior can be used to test the normality of residuals or the independence of random variables, taking into account estimation risk. When there are a finite set of models under consideration, $\{\mathcal{M}_i\}_{i=1}^{M}$, we can compute the posterior odds of model i versus j. Formally, the posterior odds of $\mathcal{M}_i$ versus $\mathcal{M}_j$ is

$$\frac{p\left(\mathcal{M}_i|Y\right)}{p\left(\mathcal{M}_j|Y\right)} = \frac{p\left(Y|\mathcal{M}_i\right)}{p\left(Y|\mathcal{M}_j\right)} \frac{p\left(\mathcal{M}_i\right)}{p\left(\mathcal{M}_j\right)}.$$

Here, the ratio, $p\left(Y|\mathcal{M}_i\right)/p\left(Y|\mathcal{M}_j\right)$, is commonly referred to as the Bayes factor. If it is greater than one, the data favors model i over model j and vice versa. Formal Bayesian diagnostic tools such as Odds ratios or Bayes Factors can be computed using the output of MCMC algorithms, see, e.g., Kass and Raftery (1995) or Han and Carlin (2000) for reviews of the large literature analyzing this issue.

3 MCMC: Methods and Theory

In this section, we describe the mechanics of MCMC algorithms, their theoretical underpinnings and convergence properties. For a full textbook discussion, we recommend the book by Robert and Casella (1999) which contains numerous illustrations and a historical perspective.

3.1 Clifford-Hammersley Theorem

In many continuous-time asset pricing models, $p\left(\Theta, X|Y\right)$ is an extremely complicated, high-dimensional distribution and it is prohibitive to directly generate samples from this distribution. However, MCMC solves this problem by first breaking the joint distribution into its complete set of conditionals, which are of lower dimension and are easier to sample. It is in this manner that MCMC algorithms attacks the curse of dimensionality that plagues other methods.

The theoretical justification for breaking $p(\Theta, X|Y)$ into its complete conditional distributions is a remarkable theorem by Clifford and Hammersley.[2] The general version of the Clifford-Hammersley theorem (Hammersley and Clifford (1970) and Besag (1974)) provides conditions for when a set of conditional distributions characterizes a unique joint distribution. For example, in our setting, the theorem indicates that $p(\Theta|X, Y)$ and $p(X|\Theta, Y)$ uniquely determine $p(\Theta, X|Y)$.

This characterization of the joint posterior into two conditional posteriors may not be sufficient to break the curse of dimensionality, as may not be possible to directly sample from $p(\Theta|X, Y)$ and $p(X|\Theta, Y)$. If this case, another application of the Clifford-Hammersley theorem can be used to further simplify the problem. Consider $p(\Theta|X, Y)$ and assume that the $K-$dimensional vector Θ can be partitioned into $k \leq K$ components $\Theta = (\Theta_1, ..., \Theta_k)$ where each component could be uni- or multidimensional. Given the partition, the Clifford-Hammersley theorem implies that the following set of conditional distributions

$$\begin{aligned}
&\Theta_1|\Theta_2, \Theta_3, \ldots, \Theta_k, X, Y \\
&\Theta_2|\Theta_1, \Theta_3, \ldots, \Theta_k, X, Y \\
&\vdots \\
&\Theta_k|\Theta_2, \Theta_3, \ldots, \Theta_{k-1}, X, Y
\end{aligned}$$

uniquely determines $p(\Theta|X, Y)$. In the case of the state vector, the joint distribution $p(X|\Theta, Y)$ can be characterized by its own complete set of conditionals: $p\left(X_t|\Theta, X_{(-t)}, Y\right)$ for $t = 1, ..., T$ where $X_{(-t)}$ denotes the elements of X excluding X_t. In the extreme, the Clifford-Hammersley theorem implies that instead of drawing from a $T + K$ dimensional posterior, the same information is contained in $T + K$ *one* dimensional distributions.

The fact that complete conditionals fully characterize a joint distribution is not at all intuitively obvious. It is something unique to the problem of sampling from a joint distribution. A proof of the Clifford-Hammersley theorem based on the Besag formula (Besag (1974)) uses the insight that for any pair (Θ^0, X^0) of points, the joint density $p(\Theta, X|Y)$ is determined as

$$\frac{p(\Theta, X|Y)}{p(\Theta^0, X^0|Y)} = \frac{p(\Theta|X^0, Y)\, p(X|\Theta, Y)}{p(\Theta^0|X^0, Y)\, p(X^0|\Theta, Y)}$$

[2]Somewhat surprisingly, Clifford and Hammersley never published their results as they could not relax the positivity condition. For a discussion of the circumstances surrounding this, see the interesting discussion by Hammersley (1974) after the paper by Besag (1974).

as long as a *positivity* condition is satisfied. Thus, knowledge of $p(\Theta|X,Y)$ and $p(X|\Theta,Y)$, up to a constant of proportionality, is equivalent to knowledge of the joint distribution. The positivity condition in our case requires that for each point in the sample space, $p(\Theta,X|Y)$ and the marginal distributions have positive mass. Under very mild regularity conditions the positivity condition is always satisfied.

3.2 Gibbs Sampling

The simplest MCMC algorithm is called the Gibbs sampler, a label often attributed to the paper of Geman and Geman (1984), although there are clearly some logical predecessors.[3] When it is possible to directly sample iteratively from all of the complete conditionals, the resulting MCMC algorithm is a Gibbs sampler. For example, the following defines a Gibbs sampler: given $\left(\Theta^{(0)},X^{(0)}\right)$

$$\begin{aligned}&1.\text{ Draw }\Theta^{(1)}\sim p\left(\Theta|X^{(0)},Y\right)\\&2.\text{ Draw }X^{(1)}\sim p\left(X|\Theta^{(1)},Y\right).\end{aligned}$$

Continuing in this fashion, the Gibbs sampler generates a sequence of random variables, $\left\{\Theta^{(g)},X^{(g)}\right\}_{g=1}^{G}$, which, as we discuss later, converges to $p(\Theta,X|Y)$. Since the researcher controls G, the algorithm is run until it has converged, and then a sample is drawn from the limiting distribution.

If it is not possible generate direct draws from $p(\Theta|X,Y)$ and $p(X|\Theta,Y)$, these distributions can be further simplified via Clifford-Hammersley. For example, consider following Gibbs sampler: given $\left(\Theta^{(0)},X^{(0)}\right)$

$$\begin{aligned}&1.\text{ Draw }\Theta_1^{(1)}\sim p\left(\Theta_1|\Theta_2^{(0)},\Theta_3^{(0)},\Theta_r^{(0)},X^{(0)},Y\right)\\&2.\text{ Draw }\Theta_2^{(1)}\sim p\left(\Theta_2|\Theta_1^{(1)},\Theta_3^{(0)},\Theta_r^{(0)},X^{(0)},Y\right)\\&\vdots\\&r.\text{ Draw }\Theta_r^{(1)}\sim p\left(\Theta_r|\Theta_1^{(1)},\Theta_2^{(1)},...,\Theta_{r-1}^{(1)},X^{(0)},Y\right)\end{aligned}$$

and then draw the states $p(X|\Theta,Y)$. If the states cannot be drawn in a block, then a similar argument implies that we can factor $p(X|\Theta,Y)$ into a set of lower dimensional distributions.

[3] Robert and Casella (1999) provide a discussion of the predecessors at the end of Chapter 7.

The Gibbs sampler requires that one can conveniently draw from the complete set of conditional distributions. In many cases, implementing the Gibbs sampler requires drawing random variables from standard continuous distributions such as Normal, t, Beta or Gamma or discrete distributions such as Binomial, Multinomial or Dirichlet. The reference books by Devroye (1986) or Ripley (1992) provide algorithms for generating random variables from a wide class of recognizable distributions.

The Griddy Gibbs Sampler The Griddy Gibbs sampler is an approximation that can be applied to approximate the conditional distribution by a discrete set of points. Suppose that Θ is continuously distributed and univariate and that $p(\Theta|X,Y)$ can be evaluated on a point by point basis, but that the distribution $p(\Theta|X,Y)$ is nonstandard and direct draws are not possible. The Griddy Gibbs sample approximates the continuously distributed Θ with a discrete mass of $N-$points, $\{\Theta_j\}_{j=1}^N$. Given this approximation, Ritter and Tanner (1992) suggest the following algorithm:

1. Compute $p(\Theta_j|X,Y)$ for $j = 1, ..., N$ and set $w_j = p(\Theta_j|X,Y)$;
2. Normalize the weights to add to unity and use these weights to approximate the inverse CDF of $p(\Theta|X,Y)$;
3. Generate a sample from the approximate distribution by drawing a uniform on $[0,1]$ and inverting the CDF.

Ritter and Tanner (1991) discuss issues involved with the choice of grid of points and show that this algorithm can provide accurate characterization of the conditional distribution in certain cases. In general, the algorithm performs well when the discretization is performed on a small number of parameters. In high dimensional systems, the algorithm is not likely to perform extremely well.

3.3 Metropolis-Hastings

In some cases, one or more of the conditional distribution cannot be conveniently sampled, and thus the Gibbs sampler does not apply. For example, in models that are nonlinear in the parameters, parameter conditional distribution may be unrecognizable. In other cases, the distribution might be known, but there are not efficient algorithms for sampling from

it. In these cases, a very general approach known as the Metropolis-Hastings algorithms will often apply.

Consider the case where one of the parameter posterior conditionals, generically, $\pi(\Theta_i) \triangleq p(\Theta_i|\Theta_{(-i)}, X, Y)$, can be evaluated (as a function of Θ_i), but it is not possible to generate a sample from the distribution. For simplicity, consider the case of a single parameter and suppose we are trying to sample from a one-dimensional distribution, $\pi(\Theta)$. This is equivalent to suppressing the dependence of the other parameters and states in the conditional posterior, $p(\Theta_i|\Theta_{(-i)}, X, Y)$, and significantly reduces the notational demands.

To generate samples from $\pi(\Theta)$, a Metropolis-Hastings algorithm requires the researcher to specify a recognizable proposal or candidate density $q\left(\Theta^{(g+1)}|\Theta^{(g)}\right)$. In most cases this distribution will depend critically on the other parameters, the state variables and the previous draws for the parameter being drawn. As in Metropolis, et al. (1953), we only require that we can evaluate density ratio $\pi(\Theta^{(g+1)})/\pi(\Theta^{(g)})$ easily. This is a mild assumption which is satisfied in all of the continuous-time models that we consider.

The Metropolis-Hastings algorithm then samples iteratively similar to the Gibbs sampler method, but it first draws a candidate point that will be accepted or rejected based on the acceptance probability. The Metropolis-Hastings algorithm replaces a Gibbs sampler step with the following two stage procedure:

$$\text{Step 1} \quad : \quad \text{Draw } \Theta^{(g+1)} \text{ from the proposal density } q(\Theta^{(g+1)}|\Theta^{(g)}) \tag{6}$$

$$\text{Step 2} \quad : \quad \text{Accept } \Theta^{(g+1)} \text{ with probability } \alpha\left(\Theta^{(g)}, \Theta^{(g+1)}\right) \tag{7}$$

where

$$\alpha\left(\Theta^{(g)}, \Theta^{(g+1)}\right) = \min\left(\frac{\pi(\Theta^{(g+1)})/q(\Theta^{(g+1)}|\Theta^{(g)})}{\pi(\Theta^{(g)})/q(\Theta^{(g)}|\Theta^{(g+1)})}, 1\right). \tag{8}$$

Implementing Metropolis-Hastings requires only drawing from the proposal, drawing a uniform random variable and evaluating the acceptance criterion.[4] Intuitively, this algorithm "decomposes" the unrecognizable conditional distribution into two parts: a recognizable distribution to generate candidate points and an unrecognizable part from which the acceptance criteria arises. The acceptance criterion insures that the algorithm has the correct equilibrium distribution. Continuing in this manner, the algorithm generates samples $\left\{\Theta^{(g)}\right\}_{g=1}^{G}$ whose limiting distribution is $\pi(\Theta)$.

[4]Mechanically, the Metropolis-Hastings algorithm consists of the following steps: (1) draw a candidate $\widehat{\Theta}$ from $q\left(\Theta|\Theta^{(g)}\right)$, (2) draw $u \sim Uniform[0,1]$, (3) accept the draw, that is set $\Theta^{(g+1)} = \widehat{\Theta}$ if $u < \alpha\left(\Theta^{(g)}, \Theta^{(g+1)}\right)$, and (4) otherwise reject the draw, that is, set $\Theta^{(g+1)} = \Theta^{(g)}$.

The Metropolis-Hastings algorithm significantly extends the number of applications that can be analyzed as the complete conditionals conditional density need not be known in closed form. A number of points immediately emerge:

1. Gibbs sampling is a special case of Metropolis-Hastings, where $q(\Theta^{(g+1)}|\Theta^{(g)}) \propto \pi(\Theta^{(g+1)})$ and from (8) this implies that the acceptance probability is always one and the algorithm always moves. As Gibbs sampling is a special case of Metropolis, one can design algorithms consisting of Metropolis-Hastings or Gibbs steps as it is really only Metropolis. The case with both Metropolis and Gibbs steps is generally called a hybrid algorithm;

2. The Metropolis-Hastings algorithm allows the functional form of the density to be non-analytic, for example, which occurs when pricing functions require the solution of partial or ordinary differential equations. One only has to evaluate the true density at two given points;

3. There is an added advantage when there are constraints in the parameter space — one can just reject these draws. Alternatively, sampling can be done conditional on specific region, see, e.g. Gelfand, Smith and Lee (1992). This provides a convenient approach for analyzing parameter restrictions imposed by economic models.

Although theory places no restrictions on the proposal density, it is important to note that the choice of proposal density will greatly effect the performance of the algorithm. For example, if the proposal density has tails that are too thin relative to the target, the algorithm may converge slowly. In extreme case, the algorithm can get stuck in a region of the parameter space an may never converge. Later, we provide some practical recommendations based on the convergence rates of the algorithm.

There are two important special cases of the general Metropolis-Hastings algorithm which deserve special attention.

Independence Metropolis-Hastings

The general Metropolis-Hastings algorithm draws $\Theta^{(g+1)}$ from proposal density, $q(\Theta^{(g+1)}|\Theta^{(g)})$, which depends on the previous Markov state $\Theta^{(g)}$ (and, in general, other parameters and states also). An alternative is to draw the candidate $\Theta^{(g+1)}$ from a distribution independent of the previous state, $q(\Theta^{(g+1)}|\Theta^{(g)}) = q(\Theta^{(g+1)})$. This is known as an independence

Metropolis-Hastings algorithm:

$$\text{Step 1} \quad : \quad \text{Draw } \Theta^{(g+1)} \text{ from the proposal density } q(\Theta^{(g+1)}) \tag{9}$$

$$\text{Step 2} \quad : \quad \text{Accept } \Theta^{(g+1)} \text{ with probability } \alpha\left(\Theta^{(g)}, \Theta^{(g+1)}\right) \tag{10}$$

where

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

Even though the candidate draws, $\Theta^{(g+1)}$, are drawn independently of the previous state, the sequence $\left\{\Theta^{(g)}\right\}_{g=1}^{G}$ will be not be independent since the acceptance probability depends on previous draws. When using independence Metropolis, it is common to pick the proposal density to closely match certain properties of the target distribution.

Random-Walk Metropolis Random-walk Metropolis is the original algorithm considered by Metropolis, et al. (1953) and is the mirror image of the independence Metropolis-Hastings algorithm. It draws a candidate from the following random walk model, $\Theta^{(g+1)} = \Theta^{(g)} + \varepsilon_t$, where ε_t is an independent mean zero error term, typically taken to be a symmetric density function with fat tails, like a $t-$distribution. Note that the choice of the proposal density is generic, in the sense that it ignores the structural features of the target density.

Due to the symmetry in the proposal density, $q(\Theta^{(g+1)}|\Theta^{(g)}) = q(\Theta^{(g)}|\Theta^{(g+1)})$, the algorithm simplifies to

$$\text{Step 1} \quad : \quad \text{Draw } \Theta^{(g+1)} \text{ from the proposal density } q(\Theta^{(g+1)}|\Theta^{(g)}) \tag{11}$$

$$\text{Step 2} \quad : \quad \text{Accept } \Theta^{(g+1)} \text{ with probability } \alpha\left(\Theta^{(g)}, \Theta^{(g+1)}\right) \tag{12}$$

where

$$\alpha\left(\Theta^{(g)}, \Theta^{(g+1)}\right) = \min\left(\frac{\pi(\Theta^{(g+1)})}{\pi(\Theta^{(g)})}, 1\right).$$

In random walk Metropolis-Hastings algorithms, the researcher controls the variance of the error term and the algorithm must be tuned, by adjusting the variance of the error term, to obtain an acceptable level of accepted draws, generally in the range of 20-40%. We discuss this issue later.

Langevin Diffusion Metropolis An alternative to independence or random-walk Metropolis is based on the continuous-time Langevin diffusion. For a density $\pi(\Theta)$, the Langevin diffusion is the solution to the SDE

$$d\Theta_t = \frac{\sigma^2}{2}\nabla_\Theta \log \pi(\Theta_t)\, dt + \sigma dW_t,$$

where W_t is a Brownian motion with the same dimension as Θ. This algorithm uses the fact that the solution to this SDE, Θ_t, has a stationary density of $\pi(\Theta)$, under certain regularity conditions. To simulate this process, one can use an Euler time-discretization which implies that

$$\Theta^{(g+1)} = \Theta^{(g)} + \frac{\sigma^2}{2}\nabla_\Theta \log \pi\left(\Theta^{(g)}\right) + \sigma\left(W_{g+1} - W_g\right)$$

where we use superscripts instead of subscripts on the discretized process. As pointed out by Roberts and Tweedie (1996), the naive discretization may or may not share the same limiting properties (as g increases) as the continuous-time version.

Roberts and Tweedie (1996), following Besag (1994), suggest the following Metropolis correction to the naive discretization:

$$\text{Step 1} \quad : \quad \text{Draw } \Theta^{(g+1)} \sim N\left(\Theta^{(g)} + \frac{\sigma^2}{2}\nabla_\Theta \log \pi\left(\Theta^{(g)}\right), \sigma^2\right) \tag{13}$$

$$\text{Step 2} \quad : \quad \text{Accept } \Theta^{(g+1)} \text{ with probability } \alpha\left(\Theta^{(g+1)}, \Theta^{(g)}\right) \tag{14}$$

where

$$\alpha\left(\Theta^{(g)}, \Theta^{(g+1)}\right) = \min\left(\frac{\pi(\Theta^{(g+1)}) f\left(\Theta^{(g+1)}, \Theta^{(g)}\right)}{\pi(\Theta^{(g)}) f\left(\Theta^{(g)}, \Theta^{(g+1)}\right)}, 1\right)$$

where $f(x, y)$ is the multivariate normal kernel

$$f(x, y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\left\| y - x - \frac{\sigma^2}{2}\nabla_x \log\left(\pi(x)\right)\right\|^2\right)\right).$$

The Langevin diffusion proposal tilts the draws toward the mode, unlike random walk which moves symmetrically.

We now turn to the strong convergence theory underpinning MCMC.

3.4 Convergence Theory

Our MCMC algorithm generates sequence of draws for parameters, $\Theta^{(g)}$, and state variables, $X^{(g)}$. By construction, this sequence is Markov and the chain is characterized by its starting value, $\Theta^{(0)}$ and its conditional distribution or transition kernel $P\left(\Theta^{(g+1)}, \Theta^{(g)}\right)$, where, without any loss of generality, we abstract from the latent variables. One of the main advantages of MCMC is the attractive convergence properties that this sequence of random variables inherits from the general theory of Markov Chains.

3.4.1 Convergence of Markov Chains

Convergence properties of this sequence are based on the ergodic theory for Markov Chains. A useful reference text for Markov Chain theory is Meyn and Tweedie (1995) or Nummelin (1984). Tierney (1994) provides the general theory as applied to MCMC methods and Robert and Casella (1999) provide many additional references. We are interested in verifying that the chain produced by the MCMC algorithm converges and then identifying the unique equilibrium distribution of the chain as the correct joint distribution, the posterior. We now briefly review the basic theory of the convergence of Markov Chains.

A Markov chain is generally characterized by its $g - step$ transition probability,

$$P^{(g)}(x, A) = Prob\left[\Theta^{(g)} \in A | \Theta^{(0)} = x\right].$$

For a chain to have a unique equilibrium or stationary distribution, π, it must be irreducible and aperiodic. A Markov chain with invariant distribution π is irreducible if, for any initial state, it has positive probability of eventually entering any set which has $\pi-$positive probability. A chain is aperiodic if there are no portions of the state space that the chain visits at regularly spaced time intervals. If an irreducible and aperiodic chain has a proper invariant distribution, then π is unique and is also the equilibrium distribution of the chain. That is

$$\lim_{g \to \infty} \text{Prob}\left[\Theta^{(g)} \in A | \Theta^{(0)}\right] = \pi(A)$$

Given convergence, the obvious question is how fast does the chain converge? Here, the general theory of Markov chains also provides explicit convergence rates, see, e.g., Nummelin (1984) or Chapters 15 and 16 of Meyn and Tweedie (1995). The key condition to verify is a minorization condition for the transition kernel which leads in many cases to a convergence rate that is geometric.

While verifying geometric convergence is reassuring, there are well-known examples of geometrically ergodic Markov chains that do not converge in finite time (see the witches hat example in Polson (1991)). A stronger notion of convergence, polynomial time convergence, provides explicitly bounds on the actual convergence rate of the chain. Diaconis and Stroock (1991) show how the time-reversibility property can be used to characterize a bound known as the Poincare inequality for the convergence rate.

We now discuss the application of these general results to MCMC algorithms.

3.4.2 Convergence of MCMC algorithms

As the Gibbs sampler is a special case of the Metropolis-Hastings algorithm when the acceptance probability is unity, we can focus exclusively on the convergence of Metropolis-Hastings algorithms. In general, verifying the convergence of Markov chains is a difficult problem. Chains generated by Metropolis-Hastings algorithms, on the other hand, have special properties which allow convergence conditions to be verified in general, without reference to the specifics of a particular algorithm. We now review these conditions.

The easiest way to verify and find an invariant distribution is to check time-reversibility. Recall that for a Metropolis-Hastings algorithm, that the target distribution, π, is given and is proper being the posterior distribution. The easiest way of checking that π is an invariant distribution of the chain is to verify the detailed balance (time-reversibility) condition: a transition function P satisfies the detailed balance condition if there exists a function π such that

$$P(x,y)\pi(x) = P(y,x)\pi(y)$$

for any points x and y in the state space. Intuitively, this means that if the chain is stationary, it has the same probability of reaching x from y if started at y as it does of reaching y from x if started at x. This also implies that π is the invariant distribution since $\pi(y) = \int P(x,y)\,\pi(dx)$.

In the case of Gibbs sampling, verifying time-reversibility is an immediate implication of the Clifford-Hammersley theorem. The Gibbs sampler cycles through the one-dimensional conditional distributions. This generates the following transition density:

$$P(x,y) = \prod_{i=1}^{k} p(x_i|x_1,...,x_{i-1},y_{i+1},..,y_k).$$

By Clifford-Hammersley,

$$\frac{\pi(x)}{\pi(y)} = \prod_{i=1}^{k} \frac{p(x_i|x_1, ..., x_{i-1}, y_{i+1}, .., y_k)}{p(y_i|y_1, ..., y_{i-1}, x_{i+1}, .., x_k)}$$

which implies that

$$\frac{\pi(x)}{\pi(y)} = \frac{P(x,y)}{P(y,x)},$$

which is precisely the time-reversibility condition.

Checking time reversibility for Metropolis-Hastings algorithms is also straightforward. The transition function in the Metropolis-Hastings algorithm is

$$P(x,y) = \alpha(x,y)\, Q(x,y) + (1 - r(x))\, \delta_x(y) \tag{15}$$

where $r(x) = \int \alpha(x,y)\, Q(x,y)dy$ and $Q(x,y) = q(y|x)$. For the first term, the detailed balance condition holds because

$$\begin{aligned}
\alpha(x,y)\, Q(x,y)\pi(x) &= \min\left\{\frac{\pi(y)\, Q(y,x)}{\pi(x)\, Q(x,y)}, 1\right\} Q(x,y)\pi(x) \\
&= \min\left\{\pi(y)\, Q(y,x), Q(x,y)\pi(x)\right\} \\
&= \min\left\{1, \frac{Q(x,y)\pi(x)}{\pi(y)\, Q(y,x)}\right\} \pi(y)\, Q(y,x) \\
&= \alpha(y,x)\, Q(y,x)\pi(y)
\end{aligned}$$

and the derivation for the second term in (15) is similar. Thus Gibbs samplers and Metropolis-Hastings algorithms generate Markov Chains that are time-reversible and have the target distribution as an invariant distribution. Of course, the Gibbs sampler can also be viewed as a special case of Metropolis.

It is also straightforward to verify $\pi-$*irreducibility*, see Roberts and Polson (1994) for the Gibbs samplers and Roberts and Smith (1993) and Robert and Casella (1999) for Metropolis-Hastings algorithms. One sufficient condition is that $\pi(y) > 0$ implies that $Q(x,y) > 0$ (see, e.g., Mengersen and Tweedie (1996)). In the case of the Gibbs sampler, these conditions can be significantly relaxed to the assumption that x and y communicate, which effectively means that starting from x one can eventually reach state y. To verify aperiodicity, one can appeal to a theorem in Tierney (1994) which states that all $\pi-$irreducible Metropolis algorithms are Harris recurrent. Hence, there exists a unique stationary distribution to which the Markov chain generated by Metropolis-Hastings algorithms converges and hence the chain is ergodic.

Having discussed these results, it is important to note that we are rarely purely interested in convergence of the Markov chain. In practice, we are typically interested in sample averages of functionals along the chain. For example, to estimate the posterior mean for a given parameter, we are interested in the convergence of $\frac{1}{G}\sum_{g=1}^{G} f\left(\Theta^{(g)}\right)$. There are two subtle forms of convergence operating: first the distributional convergence of the chain, and second the convergence of the sample average. The following result provides both:

Proposition: (Ergodic Averaging) *Suppose $\Theta^{(g)}$ is an ergodic chain with stationary distribution π and suppose f is a real-valued function with $\int |f|\, d\pi < \infty$. Then for all $\Theta^{(g)}$ for any initial starting value $\Theta^{(g)}$*

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

almost surely.

In many cases, we go further with an ergodic central limit theorem:

Proposition: (Central Limit Theorem) *Suppose $\Theta^{(g)}$ is an ergodic chain with stationary distribution π and suppose that f is real-valued and $\int |f|\, d\pi < \infty$. Then there exists a real number $\sigma(f)$ such that*

$$\sqrt{G}\left(\frac{1}{G}\sum_{g=1}^{G} f\left(\Theta^{(g)}\right) - \int f(\Theta)\, d\pi\right)$$

converges in distribution to a mean zero normal distribution with variance $\sigma^2(f)$ for any starting value.

While re-assuring, these limiting theorems should be taken with a grain of salt. A warning regarding the use asymptotics of this type is given in Aldous (1989, *p. vii*),

> *The proper business of probabilists is calculating probabilities. Often exact calculations are tedious or impossible, so we resort to approximations. A limit theorem is an assertion of the form: the error in a certain approximation tends to 0 as (say) $G \to \infty$. Call such limit theorems naive if there is no explicit error bound in terms of G and the parameters of the underlying process. Such theorems are so prevalent in theoretical and applied probability that people seldom*

stop to ask their purpose.... It is hard to give any argument for the relevance of a proof of a naive limit theorem, except as a vague reassurance that your approximation is sensible, and a good heuristic argument seems equally reassuring.

One measure of speed of convergence, geometric convergence, implies that there exists a $\lambda < 1$ and a constant K such that

$$\left\| P^g\left(\cdot, \Theta^{(0)}\right) - \pi\left(\cdot\right) \right\| \leq K\lambda^{-G}$$

where $\|\|$ could denote any number of norms. Roberts and Polson (1994) prove that all Gibbs samplers are geometrically convergent under a minorization condition. For the Metropolis-Hastings algorithm, there are a number of results on the geometric convergence and the results rely on the tail behavior of the target and proposal density. Mengersen and Tweedie (1996) show that a sufficient condition for the geometric ergodicity of *independence* Metropolis-Hastings algorithms is that the tails of the proposal density dominate the tails of the target, which requires that the proposal density q is such that q/π is bounded over the entire support. Mengersen and Tweedie (1996) show that *random walk* algorithms converge at a geometric rate if the target density has geometric tails.

While geometric convergence is a major improvement on the central limit theorem, it can still give a false sense of security. A popular example of this is the witch's hat distribution (see Polson (1992) and Geyer (1992)). This distribution looks like a witch's hat: a broad flat brim with a sharp peak in the center. In this case, the Gibbs sampler is geometrically convergent, however, λ is so close to 1, that practically speaking, the algorithm never converges. The chance of moving from brim to peak is exponentially small and therefore in finite computing, one may never visit this region of the space. Another example of this sort of potentially degenerate behavior is given by an example from Elekes (1986).[5]

A stronger notion of convergence, polynomial convergence, is faster than geometric and guarantees convergence in finite or computing time. Diaconis and Stroock (1991), Frieze,

[5]Suppose that the state space is the unit ball in k-dimensions and consider any G points in this space. These G-points are the draws from the MCMC algorithm. The volume of the convex hull of these points is bounded by $G/2^k$. For example, suppose that $k = 50$ and one runs the MCMC algorithm for one billion draws, $G = 10^9$. The convex hull is $10^9/2^{50}$, which implies that any run of the Gibbs sampler will cover only an exponentially small portion of the state space.

Kannan and Polson (1994), Polson (1996) and Rosenthal (1995a, 1995b) provide polynomial convergence in a number of different cases. For example, Frieze, Kannan and Polson (1994) show that MCMC algorithms that draw from log-concave distributions generate polynomial convergent algorithms. While this assumption may seem restrictive, Polson (1996) shows that data augmentation can be used to convert a non-log-concave sampling problem into a log-concave problem. An example is representing a $t-$distribution as a scale mixture of normals with a latent variable indexing the scaling parameters. Thus careful data augmentation can significantly improve the convergence of the MCMC algorithm.

Second, in addition to the formal convergence theory, there is a large literature that studies the information content of sequence $\left\{\Theta^{(g)}\right\}_{g=1}^{G}$. Unlike importance sampling, MCMC algorithms generate dependent Monte Carlo simulation methodology and because of this, it is important to understand the nature of this dependency. On the one hand, while theory is clear that the chains converge, it is impossible to formally diagnose convergence from the realized output of the chain.[6] On the other hand, the output of the chain clearly has some informational content. Popular observed-chain based diagnostics include calculating parameter trace plots. The trace plots, plots of $\Theta_i^{(g)}$ versus g, show the history of the chain and are useful for diagnosing chains that get stuck in a region of the state space. It is also common to analyze the correlation structure of draws by computing the autocorrelation function (ACF). Again, one needs to take care when interpreting these ACF's, as algorithms that have low autocorrelation may never converge (the witch's hat distribution mentioned above). It is also easy to calculate Monte Carlo estimates for the standard errors of $\frac{1}{G}\sum_{g=1}^{G} f(\Theta^{(g)})$. The informational content of the chain regarding estimation of $E_\pi\left(f\left(\Theta\right)\right)$ is clearly summarized $\sigma^2\left(f\right)$. Geyer (1993), among others, show how to estimate the information using realizations of a provable convergent chain. This, in turn, allows the researcher to apply the Central Limit Theorem to assess the Monte Carlo errors inherent in MCMC estimation.

The following implementation procedure is typically used. Starting from a point $\Theta^{(0)}$, possibly at random, the general methodology is to discard a *burn-in* period of h initial iterations in order to reduce the influence of the choice of starting point. After the burn-in period the researcher makes an additional *estimation* period of G simulations, which results in one long chain of length G. When forming Monte Carlo averages every simulated point

[6] Peter Clifford had the following comments on detecting convergence of a chain purely from a simulated run of the chain in his discussion to Roberts and Smith (1993): "Can we really tell when a complicated Markov chain has reached equilibrium? Frankly, I doubt it" (p. 53).

in the chain after the burn-in period should be used. The estimation period G is chosen so as to make the Monte Carlo sampling error as small as desired. Standard errors are also easily computed. See Aldous (1987), Tierney (1994) and Polson (1996) for a theoretical discussion of the choice of (h, G) and the relationship between the estimation period G and Monte Carlo standard errors.

3.5 MCMC Algorithms: Issues and Practical Recommendations

This section provides a number of practical recommendations for building, testing and applying MCMC algorithms.

Building MCMC algorithms Due to the modular nature of MCMC algorithms, we recommend building the algorithms in a "bottom-up" fashion. That is, first program a simple version of the model and, after verifying that it works, add additional factors. For example, when estimating a stochastic volatility model with jumps, first implement a pure stochastic volatility model and a pure jump model, and then after both are working, combine them.

Moreover, there are always multiple ways of implementing a given model. For example, Robert and Casella (1999) provide examples where a Metropolis step may be preferred to Gibbs even when the conditional can directly be sampled. Therefore, we recommend trying multiple algorithms, assessing their accuracy and computational efficiency and then carefully choosing an algorithm along these two dimensions. Algorithms that appear to be "fast" in terms of computational properties may be very slow in terms of their theoretical convergence rates.

Polson (1996) shows that the introduction of additional latent state variables, known as data augmentation, can dramatically increase the rate of convergence. One must be careful, however, as the introduction of state variables can also degrade the provable convergence rate of algorithms.

Blocking When building algorithms, parameters or state variables that are correlated should, if possible, be drawn in blocks. As shown by Kong, Liu, and Wong (1994), drawing correlated parameters in blocks can improve the speed of convergence. Blocking plays a central role in models with latent states. States can be updated individually, commonly referred to as single-state updating, or in blocks. In many models, state variables are

persistent which implies that the correlation between neighbor states is typically high and the gains from drawing these states together in a block can be significant.

For example, in conditionally Gaussian models, the Kalman filtering recursions allow a block update of all states in one step, a highly efficient approach for updating states. As discussed in Carter and Kohn (1994) and Kim and Shephard (1994), models that involve discrete mixture of normal distributions have a structure that is amenable to updating the states in blocks. Unfortunately, it is difficult to generate generic algorithms for block-updating, as blocking schemes must use the specific stochastic structure of the model specification. We provide examples of these algorithms below.

Simulation studies Simulation studies, whereby artificial data sets are simulated and the efficiency and convergence of the algorithm can be checked, are always recommended. These studies provide a number of useful diagnostics. First, among other things, they provide insurance against programming errors, incorrect conditionals, poorly mixing Markov chains and improper priors. Second, they can also be used to compare MCMC against alternative estimation methodologies. For example, Andersen, Chung and Sorenson (1998) show that in a simple stochastic volatility, MCMC outperforms GMM, EMM, QMLE and simulated maximum likelihood in terms of root mean squared error.

Third, they characterize the impact of discretization error on the parameter estimates. Eraker, Johannes and Polson (2003) show that time-discretization of the double-jump model at a daily frequency does not induce any biases in the parameter estimates. Fourth, the simulation studies provide a guide for how long to run algorithms.

Provable Convergence Provable convergence rates are always important. Algorithms that are provably geometric convergent are preferred to those that are not. For example, care must be taken in using normal proposal densities when the target has fat tails, as the results in Mengersen and Tweedie (1996) imply that this algorithm will be "slow." When using independence or random-walk Metropolis, one should use fat tailed distributions such as a $t-$distribution.

Choosing proposal densities and tuning Metropolis algorithms In both independence and random-walk Metropolis algorithms, the is quite a bit of latitude that is available when choosing the proposal density. In the case of independence Metropolis, the functional form of the proposal density can be specified and in random walk and Langevin Metropolis,

the standard deviation of the shock distribution, also known as the scaling or tuning parameter, needs to be specified by the researcher. Theory only provides broad guidelines for how to specify these algorithms. For both independence and random-walk Metropolis, theory requires that the support of the distributions coincide and that, for "fast" convergence, the tails of the proposal density should dominate the tails of the target density.

Figure 1 provides three common pitfalls encountered when using Metropolis-Hastings algorithms. In each panel, the target density is shown in solid lines and the proposal density is shown as a dashed line. In each case, the proposal density is not properly chosen, scaled or tuned and the impact on the algorithm can be different depending on the case.

In the first case, the target density is $N(5,1)$ and the proposal density is $N(-5,1)$. In this case, it is clear that the algorithm, while converging nicely in theory as the normal distributions have the same tail behavior, will converge very slowly in computing time. Suppose that the current state is near the mode of the target. If a draw near the mode of the proposal is proposed, the algorithm will rarely accept this draw and the algorithm will not move. On the other hand, if the current state ever approaches, for example, the mode of the proposal density, it will continue to propose moves nearby which rarely will increase the acceptance probability. The case in the second panel is similar, except now the target has a much higher variance. In this case, the proposal will very often be accepted, however, the target distribution will not be efficiently explored because all of the proposals will be in a small range centered around zero. The third case is maybe most insidious. In this case, the two distributions have same mean and variance, but the target distribution is $t(1)$ and has extremely fat tails while the proposal is normal. This algorithm will likely have a high acceptance probability but the algorithm will never explore the tails of the target distribution. The algorithm appears to move around nicely, but theory indicates that convergence, in a formal sense, will be slow. Thus the researcher will receive a false sense of security as the algorithm appears to be behaving well.

How then should Metropolis proposals be chosen and tuned? We have a number of recommendations. First, as mentioned above, the researcher should be careful to insure that the Metropolis step is properly centered, scaled and has sufficiently fat tails. In most cases, a conditional posterior can be analytically or graphically explored and one should insure that the proposal has good properties. Two, we recommend simulation studies to insure that the algorithm is properly estimating parameters and states. This typically uncovers large errors when, for example, certain parameters easily get stuck either at the true value

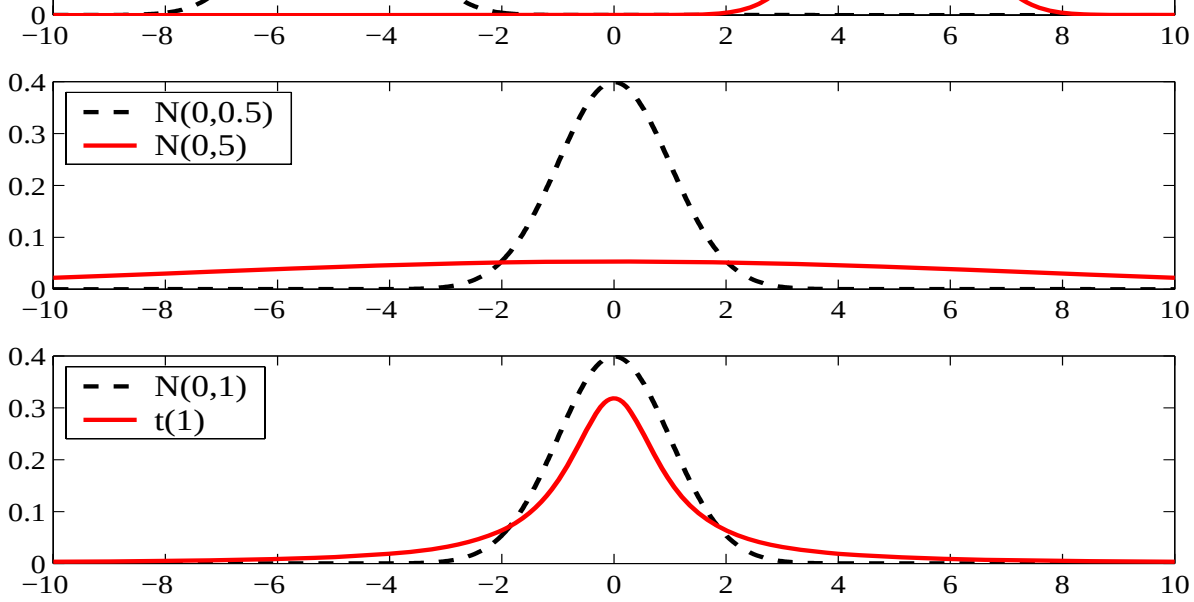

Figure 1: Examples of poorly chosen proposal densities. In each panel, the target density is shown as a solid line and the proposal as a dotted line.

or far away from the true value. Third, there are some asymptotic results for scaling random-walk and Langevin-diffusion Metropolis algorithms which provide the "optimal" asymptotic acceptance probability of random walk algorithms. Of course, optimal is relative to a specific criterion, but the results indicate that the acceptance probabilities should be in the range of 0.2-0.5. In our experience, these guidelines are reasonable in most cases.

Non-Informative priors One must be careful when using non-informative priors. Without care, conditional or joint posteriors can be improper, a violation of the Clifford-Hammersley Theorem. Hobert and Casella (1996) provide a number of general examples. For example, in a log-stochastic volatility, a "non-informative" prior on σ_v of $p(\sigma_v) \propto \sigma_v^{-1}$ results in a proper conditional posterior for σ_v but an improper joint posterior which leads

to a degenerate MCMC algorithm. In some cases, the propriety of the joint posterior cannot be checked analytically, and in this case, simulation studies can be reassuring. We recommend that proper priors, typically diffuse, always be used unless there is a very strong justification for doing otherwise.

Convergence Diagnostics and Starting Values We recommend carefully examining the parameter and state variable draws for a wide-range of starting values. For a given set of starting values, trace plots of a given parameter or state as a function of G provide important information. Trace plots are very useful for detecting poorly specified Markov Chains: chains that have difficultly moving from the initial condition, chains that get trapped in certain region of the state space, or chains that move slowly. We provide examples of trace plots below. Whenever the convergence of the MCMC algorithm is in question, careful simulation studies can provide reassurance that the MCMC algorithm is providing reliable inference.

Rao-Blackwellization In many cases, naive Monte Carlo estimates of the integrals can be improved using a technique known as Rao-Blackwellization. If there is an analytical form for the conditional density $p\left(\Theta_i|\Theta_{(-i)}, X, Y\right)$, then we can take advantage of the conditioning information to estimate the marginal posterior mean as

$$E\left(\Theta_i|Y\right) = E\left[E\left[\Theta_i|\Theta_{(-i)}, X, Y\right]|Y\right] \approx \frac{1}{G}\sum_{g=1}^{G} E\left[\Theta_i|\Theta_{(-i)}^{(g)}, X^{(g)}, Y\right].$$

Gelfand and Smith (1992) show that this estimator has a lower variance than the simple Monte Carlo estimate.

4 Bayesian Inference and Asset Pricing Models

The key to Bayesian inference is the posterior distribution which consists of three components, the likelihood function, $p\left(Y|X,\Theta\right)$, the state variable specification, $p\left(X|\Theta\right)$, and the prior distribution $p\left(\Theta\right)$. In this section, we discuss the connections between these components and the asset pricing models. Section 4.1 discusses the continuous-time specification for the state variables, or factors, and then how the asset pricing model generates the likelihood function. These distributions are abstractly given via the solution of stochastic

differential equations, and we use time-discretization methods which we discuss in Section 4.2 to characterize the likelihood and state dynamics. Finally, section 4.3 discusses the important role of the parameter distribution, commonly called the prior.

4.1 States Variables and Prices

Classical continuous-time asset pricing models such as the Cox, Ingersoll, and Ross (1985) model, begin with an exogenous specification of the underlying factors of the economy. In all of our examples, we assume that the underlying factors, labeled as F_t, arise as the exogenous solution to parameterized stochastic differential equations with jumps:

$$dF_t = \mu_f\left(F_t, \Theta^{\mathbb{P}}\right)dt + \sigma_f\left(F_t, \Theta^{\mathbb{P}}\right)dW_t^f\left(\mathbb{P}\right) + d\left(\sum_{j=1}^{N_t^f(\mathbb{P})} Z_j^f\left(\mathbb{P}\right)\right). \tag{16}$$

Here, $W_t^f\left(\mathbb{P}\right)$ is a vector of Brownian motions, $N_t^f\left(\mathbb{P}\right)$ is a counting process with stochastic intensity $\lambda_f\left(F_{t-}, \Theta^{\mathbb{P}}\right)$, $\Delta F_{\tau_j} = F_{\tau_j} - F_{\tau_j-} = Z_j^f\left(\mathbb{P}\right)$, $Z_j^f\left(\mathbb{P}\right) \sim \Pi_f\left(F_{\tau_j-}, \Theta^{\mathbb{P}}\right)$, and we assume the drift and diffusion are known parametric functions. For clarity, we are careful to denote the parameters that drive the objective dynamics of F_t by $\Theta^{\mathbb{P}}$. Throughout, we assume that characteristics have sufficient regularity for a well-defined solution to exist. While the factors are labeled F_t, we define the states as the variables that are latent from the perspective of the econometrician. Thus the "states" include jump times and jump sizes, in addition to F_t.

Common factors, or state variables, include stochastic volatility or a time-varying equity premium. This specification nests diffusions, jump-diffusions, finite-activity jump processes and regime-switching diffusions, where the drift and diffusion coefficients are functions of a continuous-time Markov chain. For many applications, the state variable specification is chosen for analytic tractability. For example, in many pure-diffusion models, the conditional density $p\left(F_t|F_{t-1}, \Theta^{\mathbb{P}}\right)$ can be computed in closed form or easily by numerical methods. Examples include Gaussian processes ($\mu_f\left(f, \Theta^{\mathbb{P}}\right) = \alpha_f + \beta_f f$, $\sigma_f\left(f, \Theta^{\mathbb{P}}\right) = \sigma_f$), the Feller "square-root" processes ($\mu_f\left(f, \Theta^{\mathbb{P}}\right) = \alpha_f^{\mathbb{P}} + \beta_f^{\mathbb{P}} f$, $\sigma_f\left(f, \Theta^{\mathbb{P}}\right) = \sigma_f^{\mathbb{P}}\sqrt{f}$), and more general affine processes (see, Duffie, Pan, and Singleton (2001) or Duffie, Filipovic, and Schachermayer (2003)). In these cases, the conditional density is either known in closed form or can be computed numerically using simple integration routines. Generally, the transition densities are not known in closed form and our MCMC approach relies on a time-discretization and data augmentation.

Given the state variables, arbitrage and equilibrium arguments provide the prices of other assets. We assume there are two types of prices. The first, denoted by a vector S_t are the prices whose dynamics we model. Common examples include equity prices, equity index values or exchange rates. The second case are derivatives such as option or bond prices, which can be viewed as derivatives on the short rate.

In the first case, we assume that S_t solves a parameterized SDE

$$dS_t = \mu_s\left(S_t, F_t, \Theta^{\mathbb{P}}\right) dt + \sigma_s\left(S_t, F_t, \Theta^{\mathbb{P}}\right) dW_t^s\left(\mathbb{P}\right) + d\left(\sum_{j=1}^{N_t^s(\mathbb{P})} Z_j^s\left(\mathbb{P}\right)\right), \tag{17}$$

where the objective measure dynamics are driven by the state variables, a vector of Brownian motion, $W_t^s\left(\mathbb{P}\right)$, a point process $N_t^s\left(\mathbb{P}\right)$ with stochastic intensity $\lambda^s\left(F_{t-}, \Theta^{\mathbb{P}}\right)$, and $S_{\tau_j} - S_{\tau_{j-}} = Z_j^s$ is a jump with $\mathcal{F}_{t-}$ distribution $\Pi^s\left(F_{t-}, \Theta^{\mathbb{P}}\right)$.

In the second case, the derivative prices, D_t are a function of the state variables and parameters, $D_t = D\left(S_t, F_t, \Theta\right)$ where $\Theta = \left(\Theta^{\mathbb{P}}, \Theta^{\mathbb{Q}}\right)$ contains risk premium parameters, $\Theta^{\mathbb{Q}}$. To price the derivatives, we assert the existence of an equivalent martingale measure, $\mathbb{Q}$,

$$dS_t = \mu_s\left(S_t, F_t, \Theta\right) dt + \sigma_s\left(S_t, F_t, \Theta^{\mathbb{P}}\right) dW_t^s(\mathbb{Q}) + d\left(\sum_{j=1}^{N_t^s(\mathbb{Q})} Z_j^s(\mathbb{Q})\right) \tag{18}$$

$$dF_t = \mu_f\left(F_t, \Theta\right) dt + \sigma_f\left(F_t, \Theta^{\mathbb{P}}\right) dW_t^f(\mathbb{Q}) + d\left(\sum_{j=1}^{N_t^f(\mathbb{Q})} Z_j^f(\mathbb{Q})\right). \tag{19}$$

where, it important to note that the drift now depends potentially on both $\Theta^{\mathbb{P}}$ and $\Theta^{\mathbb{Q}}$ (we assume for simplicity that the functional form of the drift does not change under $\mathbb{Q}$), $W_t^s(\mathbb{Q})$ and $W_t^f(\mathbb{Q})$ are Brownian motions under $\mathbb{Q}$, $N_t^f(\mathbb{Q})$ and $N_t^s(\mathbb{Q})$ are point process with stochastic intensities $\left\{\lambda^i\left(F_{t-}, S_{t-}, \Theta^{\mathbb{Q}}\right)\right\}_{i=s,f}$ and $\left(Z_j^f, Z_j^s\right)$ have joint distribution $\Pi\left(F_{t-}, S_{t-}, \Theta^{\mathbb{Q}}\right)$. Due to the absolute continuity of the changes in measure, the diffusion coefficients depend only on $\Theta^{\mathbb{P}}$. The likelihood ratio generating the change of measure for jump-diffusions is given in Aase (1988) or the review paper by Runggaldier (2003).

We only assume that this pricing function, $D\left(s, x, \Theta\right)$, can be computed numerically and do not require it to be analytically known. This implies that our methodology covers the important cases of multi-factor term structure and option pricing models. In multi-factor term structure models, the short rate process, r_t, is assumed to be a function of a

set of state variables, $r_t = r(F_t)$, and bond prices are given by

$$D_t = D(F_t, \Theta) = E_t^{\mathbb{Q}}\left[e^{-\int_t^T r(F_s)ds}\right]$$

where $\mathbb{Q}$ is an equivalent martingale measure f can be computed either analytically or as the solution to ordinary or partial differential equation. In models of option prices, the mapping is given via

$$D_t = f(S_t, F_t, \Theta) = E_t^{\mathbb{Q}}\left[e^{-\int_t^T r(F_s)ds}(S_T - K)_+\right]$$

and F_t could be, for example, stochastic volatility.

Derivative prices raise an important issue: the observation equation is technically a degenerate distribution as the prices are known conditional on state variables and parameters. In this case, if the parameters are known, certain state variables can often be inverted from observed prices, if the parameters were known. An common example of this is Black-Scholes implied volatility. In practice there are typically more prices observed than parameters which introduces a stochastic singularity: the model is incapable of simultaneously fitting all of the prices. This over-identification provides a rich source of information for testing. To circumvent the stochastic singularities, researchers commonly assume there exists a pricing error, ε_t. In the case of an additive pricing error,[7]

$$D_t = D(S_t, F_t, \Theta) + \varepsilon_t$$

where $\varepsilon_t \sim N(0, \Sigma_\varepsilon)$. This implies that prices are not fully revealing of state variables or parameters.

There are a number of motivations for introducing pricing errors. First, there is often a genuine concern with noisy price observations generated by bid-ask spreads. For example, consider an at-the-money equity index option. For the S&P 100 or 500 the contract typically has a bid-ask spread of around 5-10% of the value of the option. In fixed income, zero yields are often measured with error as they are obtained by interpolation par bond yields. The pricing error breaks the stochastic singularity that arises when there are more observed asset prices than state variables. Second, even if the econometrician does not believe the prices are observed with error, the addition of an extremely small pricing error can be viewed as a tool to simplify econometric analysis. Third, our models are clearly abstractions from

[7]It some cases, it might be more appropriate to use a multiplicative pricing error $Y_t = f(X_t, \Theta)e^{\varepsilon_t}$, which can, for example, guarantee positive prices.

reality and will never hold perfectly. Pricing errors accommodate this misspecification in a tractable manner. These pricing errors provide a useful model diagnostic, and MCMC are useful for investigating the small sample behavior of the pricing errors.

4.2 Time-discretization: computing $p(Y|X,\Theta)$ and $p(X|\Theta)$

In this subsection, we describe how the time-discretization of the stochastic differential equations can be used to compute the likelihood function and state dynamics. The researcher typically observed a panel of prices, Y, where $Y=(S,D)$ and $S=(S_1,...,S_T)$ and $D=(D_1,...,D_T)$. We assume the prices are observed at equally spaced, discrete intervals. For simplicity, we normalize the observation interval to unity. This generates the following continuous-time state space model for the derivative prices,

$$D_t = D\left(S_t, F_t, \Theta\right) + \varepsilon_t, \tag{20}$$

and the prices and factors

$$S_{t+1} = S_t + \int_t^{t+1} \mu_s\left(S_u, F_u, \Theta^{\mathbb{P}}\right) du + \int_t^{t+1} \sigma_s\left(S_u, F_u, \Theta^{\mathbb{P}}\right) dW_u^s\left(\mathbb{P}\right) + \sum_{j=N_t^s(\mathbb{P})+1}^{N_{t+1}^s(\mathbb{P})} Z_j^s\left(\mathbb{P}\right) \tag{21}$$

$$F_{t+1} = F_t + \int_t^{t+1} \mu_f\left(F_u, \Theta^{\mathbb{P}}\right) du + \int_t^{t+1} \sigma_f\left(F_u, \Theta^{\mathbb{P}}\right) dW_u^f\left(\mathbb{P}\right) + \sum_{j=N_t^f(\mathbb{P})+1}^{N_{t+1}^f(\mathbb{P})} Z_j^f\left(\mathbb{P}\right). \tag{22}$$

Equations (20) and (21) are the observation equations and (22) are the evolution equation.

In continuous-time, these models take the form of a very complicated state space model. Even if ε_t is normally distributed, $D\left(S_t, X_t, \Theta\right)$ is often non-analytic which generates Gaussian, but non-linear and non-analytic observation equation. Similarly, in (21) the error distribution is generated by

$$\int_t^{t+1} \sigma_s\left(S_u, F_u, \Theta^{\mathbb{P}}\right) dW_u^s\left(\mathbb{P}\right) + \sum_{j=N_t^s(\mathbb{P})+1}^{N_{t+1}^s(\mathbb{P})} Z_j^s\left(\mathbb{P}\right).$$

Together, the model is clearly a non-linear, non-Gaussian state space model.

At this stage, it is important to recognize the objects of interest. From the perspective of the econometrician, the jump times, jump sizes and F_t are latent, although it is typically assumed that the agents in the economy pricing the assets observe these variables. While

the variables F_t solve the stochastic differential equation, and thus, would commonly be referred to as the states, we include in our state vector the jump times, jump sizes, and spot factors, F_t, as these are all objects of interest in asset pricing models. Pricing models commonly integrate out of the the jump times and sizes, and condition solely on F_t and other prices.

At this level, it is clearly not possible to compute either the likelihood, $p(Y|X,\Theta)$, or the latent state distribution, $p(X|\Theta)$. To compute these quantities, we time-discretize the stochastic differential equations which then allows us to compute the likelihood and state variable evolution. To start, assume that the time-discretization interval matches the observed frequency. This generates the following time-discretized state space model:

$$\begin{aligned} D_{t+1} &= D(S_{t+1}, F_{t+1}, \Theta) + \varepsilon_{t+1} && (23)\\ S_{t+1} &= S_t + \mu_s\left(S_t, F_t, \Theta^{\mathbb{P}}\right) + \sigma_s\left(S_t, F_t, \Theta^{\mathbb{P}}\right)\varepsilon^s_{t+1} + Z^s_{t+1}J^s_{t+1} && (24)\\ F_{t+1} &= F_t + \mu_f\left(F_t, \Theta^{\mathbb{P}}\right) + \sigma_f\left(F_t, \Theta^{\mathbb{P}}\right)\varepsilon^f_{t+1} + Z^f_{t+1}J^f_{t+1}, && (25) \end{aligned}$$

where $\varepsilon_t \sim N(0, \sigma_D^2)$, $\varepsilon^f_{t+1}, \varepsilon^s_{t+1} \sim \mathcal{N}(0, I)$, $Z^f_{t+1} \sim \Pi^f\left(F_t, \Theta^{\mathbb{P}}\right)$, $Z^s_{t+1} \sim \Pi^s\left(F_t, \Theta^{\mathbb{P}}\right)$, $J^f_{t+1} \sim Ber\left[\lambda^f\left(X_t, \Theta^{\mathbb{P}}\right)\right]$, and $J^f_{t+1} \sim Ber\left[\lambda^s\left(F_t, \Theta^{\mathbb{P}}\right)\right]$.

Since the shocks in the model, the Brownian increments, jump times and jump sizes, are conditionally independent, once the model is discretized, it is easy to characterize the likelihood and the state dynamics. We define the latent state vector as $X_t = \left(F_{t-1}, Z^s_t, Z^f_t, J^s_t, J^f_t\right)$, which implies that

$$p(Y|X,\Theta) = \prod_{t=1}^{T} p\left(S_t|S_{t-1}, X_t, \Theta^{\mathbb{P}}\right) p\left(D_t|S_t, F_t, \Theta\right)$$

where $p\left(S_t|S_{t-1}, X_t, \Theta^{\mathbb{P}}\right)$ and $p(D_t|S_t, F_t, \Theta)$ are multivariate normal distribution. This shows the simple structural form of the model given the time-discretization. It is important to note that $\Theta^{\mathbb{Q}}$ will only appear in the second term, $p(D_t|S_t, X_t, \Theta)$. For example, in the option pricing or term structure examples, the parameters determining the risk neutral behavior of stock prices, stochastic volatility or interest rates, only appear in the option price or bond yields. The price and state evolutions, as they are observed under $\mathbb{P}$, provide no information regarding the risk-neutral behavior.

The state dynamics are given by $p(X|\Theta)$ which is given by: similarly straightforward.

$$p(X|\Theta) = \prod_{t=1}^{T} p\left(X_t|X_{t-1}, \Theta^{\mathbb{P}}\right)$$

where

$$p\left(X_{t+1}|X_t,\Theta^{\mathbb{P}}\right)=p\left(F_t|F_{t-1},Z_t^f,J_t^f\right)p\left(Z_{t+1}^f|F_t\right)p\left(J_{t+1}^f|F_t\right)p\left(Z_{t+1}^s|F_t\right)p\left(J_{t+1}^s|F_t\right).$$

The time-discretization plays an integral part: it provides a methods to analytically compute the likelihood and the state variable evolution.

The previous approximation normalized the discretization interval to unity, but the accuracy of this approximation depends on the length of the interval between observations and the characteristics of the process. In the diffusion case, if the drift and diffusion are constant, the time-discretization is exact. If the time interval between observations is small (e.g., daily) and the drift and diffusion coefficients are smooth functions of the states, the approximation error via the time-discretization is also likely small. For example, in a term structure model, Stanton (1997) finds that approximation errors in the conditional moments of the process of certain diffusions in negligible for time intervals up to a month, while Eraker, Johannes and Polson (2003) find that time-discretizations of equity prices models with jumps do not introduce any biases in the parameter estimates. As noted by Pritsker (1997, 1998) and Johannes (2004), the sampling variation (due to finite samples) typically dwarfs any discretization bias when data is sampled at reasonably high frequencies such as daily.

In other cases, the simple time-discretization may not be accurate. This can occur when the sampling interval is long (weeks or months) or the drift and diffusion coefficients are highly variable functions of the states. In this case, the solution to the difference equation in the time-discretization (**??**) is substantively different than the true solution to SDE in (22). When this occurs, it is straightforward to use the Euler scheme to obtain a more accurate characterization of the solution. The idea is to simulate additional states between times t and $t+1$ at intervals $1/M$ for $M>1$:

$$S_{t_{j+1}} = S_{t_j}+\mu_s\left(S_{t_j},F_{t_j},\Theta^{\mathbb{P}}\right)/M+\sigma_s\left(S_{t_j},F_{t_j},\Theta^{\mathbb{P}}\right)\varepsilon^s_{t_{j+1}}+Z^s_{t_{j+1}}J^s_{t_{j+1}} \quad (26)$$

$$F_{t_{j+1}} = F_{t_j}+\mu_f\left(F_{t_j},\Theta^{\mathbb{P}}\right)/M+\sigma_f\left(F_{t_j},\Theta^{\mathbb{P}}\right)\varepsilon^f_{t_{j+1}}+Z^f_{t_{j+1}}J^f_{t_{j+1}} \quad (27)$$

where $t_j=t+\frac{j}{M}$, $\varepsilon^f_{t_{j+1}},\varepsilon^s_{t_{j+1}}\sim\mathcal{N}\left(0,M^{-1}\right)$, $Z^f_{t_j+1}\sim\Pi^f\left(f_{t_j},\Theta^{\mathbb{P}}\right)$, $Z^s_{t_j+1}\sim\Pi^s\left(F_{t_j},\Theta^{\mathbb{P}}\right)$, $J^f_{t_j+1}\sim Ber\left[\lambda^s\left(F_{t_j},\Theta^{\mathbb{P}}\right)M^{-1}\right]$, and $J^f_{t_j+1}\sim Ber\left[\lambda^f\left(F_{t_j},\Theta^{\mathbb{P}}\right)M^{-1}\right]$.

With the additional simulations, we can augment the original state vector with the intermediate jump times, jump sizes and f_{t_j}'s to obtain a conditionally normal distribution. Jones (1998), Eraker (2001), Elerian, Shephard and Chib (2001), and Chib, Pitt

and Shephard (2003) examine various approaches for using time discretizations to estimate continuous-time models using MCMC methods. We refer the interested reader to these papers for further details and examples.

4.3 Parameter Distribution

The final component of the joint posterior distribution is the prior distribution of the parameters, $p(\Theta)$. This represent non-sample information regarding the parameters and one typically chooses a parameterized distribution. This implies that the researcher must choose both a distribution for the prior and the so-called hyperparameters that index the distribution. Through both the choice of distribution and hyperparameters, the researcher can introduce non-sample information or, alternatively, choose to impose little information. In the latter case, an "uninformative" or diffuse prior is one that provides little or no information regarding the location of the parameters.

When possible we recommend using standard conjugate prior distributions, see, for example Raiffa and Schlaifer (1961) or DeGroot (1970). They provide a convenient way of finding closed-form, easy to simulate, conditional posteriors. A conjugate prior is a distribution for which the conditional posterior is the same distribution with different parameters. For example, suppose and random variable Y is normally distributed, $Y_t|\mu,\sigma^2 \sim N(\mu,\sigma^2)$. Assuming a normal prior on μ, $\mu \sim N(a,A)$, the conditional posterior distribution for the mean, $p(\mu|\sigma^2,Y)$, is also normally distributed, $N(a^*,A^*)$, where the starred parameters depend on the data, sample size and the hyperparameters a and A. In this case, the posterior mean is a weighted combination of the prior mean and the sample information with the weights determined by the relative variances. Choosing A to be very large generates what is commonly referred to as an uninformative prior. Of course, depending on the parameter of interest, no prior can be truly uninformative (Poincare (1901)). For the variance parameter, the inverted gamma distribution is also conjugate. Bernardo and Smith (1995) provide a detailed discussion and list of conjugate priors.

In some cases, researchers may specify a flat prior, which is completely uninformative. For example, in a geometric Brownian motion model of returns, $Y_t \sim N(\mu,\sigma^2)$, it is common to assume a flat prior distribution for the mean by setting $p(\mu,\sigma^2) \propto \sigma^{-1}$. While a flat prior distribution may represent lack of knowledge, it may also also lead to serious computational problems as a flat prior does not integrate to one. To see this, note that the

parameter posterior is given by

$$p(\Theta|Y) \propto p(Y|\Theta)\, p(\Theta).$$

For inference, this distribution must be proper, that is $\int_\Theta p(\Theta|Y)\, d\Theta = 1$. In many cases, flat priors can lead to an improper posterior. This is more problematic in state space models where the marginal likelihood, $p(Y|\Theta)$, is unavailable in closed form and where one cannot always check that the propriety of the posterior. Additionally, joint posterior propriety is a necessary condition for MCMC algorithms to converge as we discuss later. This implies that another motivation for using diffuse proper priors is as a computational tool for implementation via MCMC.

There are often statistical and economic motivations for using informative priors. For example, in many mixture models, priors must at least partially informative to overcome degeneracies in the likelihood. Take, for example, Merton's (1976) jump diffusion model for log-returns $Y_t = \log(S_{t+\Delta}/S_t)$. In this case, returns are given by

$$Y_t = \mu + \sigma(W_{t+\Delta} - W_t) + \sum_{j=N_t}^{N_{t+\Delta}} Z_j$$

and the jump sizes are normally distributed with mean μ_J and variance σ_J^2. As shown by Lindgren (1978), Kiefer (1978) and Honore (1997), the maximum likelihood estimator is not defined as the likelihood takes infinite values from some parameters. This problem does not arise when using an informative prior, as the prior will typically preclude these degeneracies.

Informative priors can also be used to impose stationarity on the state variables. Models of interest rates and stochastic volatility often indicate near-unit-root behavior. In the stochastic volatility model discussed earlier, a very small κ_v introduces near-unit root behavior. For practical applications such as option pricing or portfolio formation, one often wants to impose mean-reversion to guarantee stationarity. This enters via the prior on the speed of mean reversion that imposes that κ_v are positive and are bounded away from zero.

For regime-switching models, the prior distribution $p(\Theta)$ can be used to solve a number of identification problems. First, the labeling problem of identifying the states. The most common way of avoiding this problem is to impose a prior that orders the mean and variance parameters. One practical advantage of MCMC methods are that they can easily handle truncated and ordered parameter spaces, and hence provide a natural approach for regime switching models.

It is increasingly common in many applications to impose economically motivated priors. For example, Pastor and Stambaugh (2000) use the prior to represent an investor's degree of belief over a multi-factor model of equity returns. In other cases, an economically motivated prior might impose that risk premium are positive, for example.

In practice, researchers often perform sensitivity analysis to gauge the impact of certain prior parameters on the parameter posterior. Occasionally, the posterior for certain may depend critically on the choice. As the posterior is just the product of the likelihood and the prior, this only indicates that the likelihood does not provide any information regarding the location of these parameters. One extreme occurs when the parameters are not identified by the likelihood and the posterior is equal to the prior.

5 Asset Pricing Applications

In this section, we describe a number of asset pricing models and the associated MCMC algorithms for estimating the parameters and latent states. We first consider equity models where we assume that equity returns and option prices are observed. We consider the Black-Scholes-Merton model, time-varying equity premium models, stochastic volatility models and multi-variate models with jumps. Next, we consider models of interest rates, and consider Gaussian, square-root and multi-factor models. Finally, we discuss general estimation of regime-switching models.

5.1 Equity Asset Pricing Models

5.1.1 Geometric Brownian Motion

The simplest possible asset pricing model is the geometric Brownian specification for an asset price. Here, the price, S_t, solves the familiar SDE

$$dS_t = \left(\mu + \frac{1}{2}\sigma^2\right) S_t dt + \sigma S_t dW_t \left(\mathbb{P}\right)$$

where μ is the continuously-compounded expected return and σ is the volatility. Prices are always recorded at discrete-spaced time intervals, and, for simplicity, we assume they are equally space. This model has a closed-form solution for continuously-compounded returns:

$$Y_t = \log\left(S_t/S_{t-1}\right) = \mu + \sigma\varepsilon_t,$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$. The model generates a conditional likelihood likelihood for the vector of continuously-compounded returns of

$$p\left(Y|\mu,\sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^T \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(Y_t-\mu\right)^2\right),$$

where $Y = (S_0, ..., S_T)$. There are no latent variables in this model which implies that the posterior is $p(\Theta|Y) = p(\mu,\sigma^2|Y)$. Under standard prior assumptions, the posterior distribution, $p(\Theta|Y) = p(\mu,\sigma^2|Y)$, is known in closed form. However, to develop intuition, we describe an MCMC approach for sampling from $p(\mu,\sigma^2|Y)$.

The first step is an application of Clifford-Hammersley theorem which implies that $p(\mu|\sigma^2,Y)$ and $p(\sigma^2|\mu,Y)$ are the complete conditionals. Assuming independent priors on μ and σ^2,[8] Bayes rule implies that

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

where $p(\mu)$ and $p(\sigma^2)$ are the priors. Assuming a normal prior for μ, $p(\mu) \sim \mathcal{N}$, and an inverted gamma prior for σ^2, $p(\sigma^2) \sim \mathcal{IG}$,[9] the posteriors are conjugate, which means that $p(\mu|\sigma^2,Y)$ is also normal and $p(\sigma^2|\mu,Y)$ is also inverse Gamma. The MCMC algorithm consists of the following steps: given $\mu^{(g)}$ and $(\sigma^2)^{(g)}$

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

where the arguments of the normal and inverted Gamma distributions are easy to derive and are omitted for notational simplicity. Both of these distributions can be directly sampled, thus the MCMC algorithm is a Gibbs sampler. Iterating, this algorithm produces a sample $\left\{\mu^{(g)},(\sigma^2)^{(g)}\right\}_{g=1}^{G}$ from the posterior $p(\mu,\sigma^2|Y)$.

[8] Alternatively, one could use dependent conditional conjugate priors such as $p\left(\mu,\sigma^2\right) = p\left(\mu|\sigma^2\right)p\left(\sigma^2\right)$. In this model, the $p\left(\mu|\sigma^2\right)$ is normal and $p\left(\sigma^2\right)$ is inverted Gamma. Later, we discuss the multivariate version of this, the normal-inverted Wishart prior which leads to a joint posterior for μ and σ^2 which can be directly sampled.

[9] The inverted Gamma distribution is a common prior for a variance parameter. The inverted Gamma distribution, $\mathcal{IG}(\alpha,\beta)$, has support on the positive real line and the density is given by $f(x|\alpha,\beta) = \beta^\alpha x^{\alpha-1}e^{-\beta/x}/\Gamma(\alpha)$.

This simple example previews the general approach to MCMC estimation:

Step 1	:	Write out the price dynamics and state evolution in state space form;
Step 2	:	characterize the joint distribution by its complete conditionals
Step 3	:	use standard random sampling methods to generate draws from joint posterior

5.1.2 Black-Scholes

In many cases, option prices are also observed. If the underlying follows a geometric Brownian motion, the Black-Scholes (1973) formula implies that the price of a call option struck at K is given by

$$C_t = BS(\sigma, S_t) = S_t N(d_1) - e^{r(T-t)} K N\left(d_1 - \sigma\sqrt{T-t}\right)$$

where

$$d_1 = \frac{\log(S_t/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}},$$

and we assume the continuously-compounded interest rate is known. The addition of option prices generates only minor alterations to the MCMC algorithm of the previous section. Our analysis follows and is a special case of Polson and Stroud (2002) and Eraker (2003) who allow for stochastic volatility, jumps in returns and jumps in volatility. Jacquier and Jarrow (2000) also provide a similar analysis.

The first thing to notice about the model is the stochastic singularity: if a single option price is observed without error, volatility can be inverted from the price. To break this singularity, we assume that option prices are observed with a normally distributed error. This implies the following state space model is

$$\begin{aligned} \log(S_t/S_{t-1}) &= \mu + \sigma\varepsilon_t \\ C_t &= BS(\sigma, S_t) + \varepsilon_t^c \end{aligned}$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ and $\varepsilon_t^c \sim \mathcal{N}(0, \sigma_c^2)$. This state space model is conditionally normal, but nonlinear in the parameters as $BS(\sigma, S_t)$ is not known analytically.

The joint likelihood function is the product of the equity return likelihood, $p\left(S|\mu,\sigma^2\right)$, and the option likelihood, $p\left(C|,S,\mu,\sigma^2,\sigma_c^2\right)$:

$$p\left(S,C|\mu,\sigma^2,\sigma_c^2\right)=\prod_{t=1}^{T}p\left(C_t|S_t,\sigma^2,\sigma_c^2\right)p\left(\log\left(S_t/S_{t-1}\right)|\mu,\sigma^2\right).$$

Here $S=\left(S_1,...,S_T\right)$, and $C=\left(C_1,..,C_T\right)$ are the vector with the underlying and option prices. The equity return portion of the likelihood is the same as in previous section. The option price component of the likelihood is given by

$$p\left(C_t|S_t,\sigma^2,\sigma_c^2\right)\propto\exp\left(-\frac{1}{2\sigma_c^2}\left(C_t-BS\left(\sigma,S_t\right)\right)^2\right)$$

Notice that the distribution of the option prices conditional on S_t,σ^2 and σ_c^2 is independent of μ and the distribution of the stock returns is independent of σ_c.

The MCMC algorithm samples from the joint posterior, $p\left(\mu,\sigma^2,\sigma_c^2|S,C\right)$. The complete conditionals are $p\left(\mu|\sigma^2,S\right)$, $p\left(\sigma^2|\mu,\sigma_c^2,S,C\right)$ and $p\left(\sigma_c^2|\sigma^2,S,C\right)$. Assuming the independent priors, $p\left(\mu\right)\sim\mathcal{N}$, $p\left(\sigma^2\right)\sim\mathcal{IG}$ and $p\left(\sigma_c^2\right)\sim\mathcal{IG}$, the conditional posteriors for μ and σ_c^2 are conjugate. Due to the option pricing formula, $p\left(\sigma^2|\mu,\sigma_c^2,S,C\right)$ is not a known distribution and the Metropolis-Hastings algorithm will be used to update σ^2. The MCMC algorithm cycles through the conditionals:

$$\begin{aligned}
\mu^{(g+1)} &\sim p\left(\mu|\left(\sigma^2\right)^{(g)},S\right)\sim\mathcal{N}\\
\left(\sigma_c^2\right)^{(g+1)} &\sim p\left(\sigma_c^2|\left(\sigma^2\right)^{(g)},S,C\right)\sim\mathcal{IG}\\
\left(\sigma^2\right)^{(g+1)} &\sim p\left(\sigma^2|\mu^{(g+1)},\left(\sigma_c^2\right)^{(g+1)},S,C\right):\text{Metropolis}
\end{aligned}$$

There are a number of alternatives for the Metropolis step. By Bayes rule, the conditional posterior for σ is

$$\pi\left(\sigma^2\right)\triangleq p\left(\sigma^2|\mu,C,S\right)\propto p\left(C|\sigma^2,S\right)p\left(S|\mu,\sigma^2\right)p\left(\sigma^2\right)$$

which clearly shows how both the returns and the option prices contain information about σ^2. Since $BS\left(\sigma,S_t\right)$ is given as an integral, it is not possible to sample directly from $p\left(\sigma^2|\mu,\sigma_c^2,S,Y\right)$ as

$$p\left(C_t|S_t,\sigma^2,\sigma_c^2\right)\propto\exp\left(-\frac{1}{2\sigma_c^2}\left(C_t-BS\left(\sigma,S_t\right)\right)^2\right)$$

is not known in closed form as a function of σ. One approach is to use independence Metropolis. In this case, the algorithm proposes using the data from the returns, $p\left(\sigma^2|S,\mu\right)$, and then accepts/rejects based on the information contained in the option prices. Specifically, consider a proposal of the form

$$q\left(\sigma^2\right) = p\left(\sigma^2|\mu,S\right) \propto p\left(S|\mu,\sigma^2\right) p\left(\sigma^2\right) \sim \mathcal{IG}.$$

The Metropolis algorithm is

$$\text{Step 1} \quad : \quad \text{Draw} \ \left(\sigma^2\right)^{(g+1)} \ \text{from} \ q\left(\sigma^2\right) \sim \mathcal{IG} \tag{28}$$

$$\text{Step 2} \quad : \quad \text{Accept} \ \left(\sigma^2\right)^{(g+1)} \ \text{with probability} \ \alpha\left(\left(\sigma^2\right)^{(g+1)}, \left(\sigma^2\right)^{(g)}\right) \tag{29}$$

where

$$\alpha\left(\left(\sigma^2\right)^{(g)}, \left(\sigma^2\right)^{(g+1)}\right) = \min\left(\frac{p\left(C|\left(\sigma^2\right)^{(g+1)}, S\right)}{p\left(C|\left(\sigma^2\right)^{(g)}, S\right)}, 1\right).$$

As the Black-Scholes price is always bounded by the underlying price, $BS\left(\sigma, S_t\right) \leq S_t$, so the tail behavior of $\pi\left(\sigma^2\right)$ is determined by the likelihood component and the algorithm is geometrically convergent.

In practice, this Metropolis algorithm is susceptible to two common problems mentioned earlier. First, option prices often embed volatility or jump risk premiums which implies that Black-Scholes implied volatility is generally higher than historical volatility. In fact, for the period from 1997-2002, Black-Scholes implied volatility as measured by the VIX index has averaged about 30%, while the underlying volatility is about 18%. In this case, the target may be located to the right of the proposal, as historical volatility is lower than implied volatility. This would result in very low acceptance probabilities. Of course, this is not a problem with MCMC per se, but is due to a misspecified model. Second, option prices are likely to be more informative about volatility than historical returns. This implies that the target will have lower variance than the proposal. The proposal will generate large moves which will often be rejected by the Metropolis algorithm, again, potentially generating a slowly moving chain. An alternative to the independence algorithm is to use a random-walk Metropolis algorithm with a fat-tailed innovation such as a $t-$distribution. The variance can be tuned to insure that the acceptance rates are sufficiently high. In many cases this is an attractive alternative. As mentioned in Section 2.5, it is important to implement multiple algorithms and choose one appropriate for the problem at hand.

5.1.3 A Multivariate Version of Merton's Model

Consider an extension of the geometric model: a multivariate version of Merton's (1976) jump-diffusion model. Here a $K-$vector of asset prices solves

$$dS_t = \mu S_t dt + \sigma S_t dW_t(\mathbb{P}) + d\left(\sum_{j=1}^{N_t(\mathbb{P})} S_{\tau_{j-}}(e^{Z_j(\mathbb{P})} - 1)\right)$$

where $W_t(\mathbb{P})$ is a vector standard Brownian motion, $\sigma\sigma' = \Sigma \in R^K \times R^K$ is the diffusion matrix, $N_t(\mathbb{P})$ is a Poisson process with constant intensity λ and the jump sizes, $Z_j \in R^K$ are multivariate normal with mean μ_J and variance-covariance matrix Σ_J. This model assumes the prices have common jumps with correlated sizes, although it is easy to add an additional jump process to characterize idiosyncratic jumps.

Solving this stochastic differential equation, continuously compounded equity returns over a daily interval ($\Delta = 1$) are

$$\log(S_t/S_{t-1}) = \mu + \sigma(W_{t+1}(\mathbb{P}) - W_t(\mathbb{P})) + \sum_{j=N_t+1(\mathbb{P})}^{N_{t+1}(\mathbb{P})} Z_j(\mathbb{P})$$

where, again, we have redefined the drift vector to account for the variance correction. We time discretize the jump component assuming that at most a single jump can occur over each time interval:

$$Y_t \equiv \log(S_t/S_{t-1}) = \mu + \sigma\varepsilon_t + J_t Z_t$$

where $\mathbb{P}[J_t = 1] = \lambda \in (0,1)$ and the jump sizes retain their structure. Johannes, Kumar and Polson (1999) document that, in the univariate case, the effect of time-discretization in the Poisson arrivals is minimal, as jumps are rare events. To see why, suppose the jump intensity (scaled to daily units) is $\lambda = 0.05$. Since

$$Prob[N_{t+1}(\mathbb{P}) - N_t(\mathbb{P}) = j] = \frac{e^{-\lambda}(\lambda)^j}{j!}$$

the probability of two or more jumps occurring over a daily interval is approximately 0.0012 or 1/10th of one percent, which is why the discretization bias is likely to be negligible.

The MCMC algorithm samples from $p(\Theta, X|Y)$, where $\Theta = (\mu, \Sigma, \lambda, \mu_J, \Sigma_J)$ and $X = (J, Z)$, where J, Z, and Y are vectors of jump times, jump sizes and observed prices. Returns are independent through time, which implies that the full-information likelihood is a product of multivariate normals,

$$p(Y|\Theta, J, Z) = \prod_{t=1}^{T} p(Y_t|\Theta, J_t, Z_t)$$

where

$$p\left(Y_t|J_t,Z_t,\Theta\right)\propto\left|\Sigma\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(Y_t-\mu-Z_tJ_t\right)'\Sigma^{-1}\left(Y_t-\mu-Z_tJ_t\right)\right\}.$$

In contrast, the the observed likelihood, $p\left(Y_t|\Theta\right)$, integrates out Z_t and J_t and is mixture of multivariate normal distributions. Discrete mixture distributions introduce a number of problems. For example, in the univariate case, it is well known that the observed likelihood has degeneracies as certain parameter values lead an infinite likelihood. Multi-variate mixtures are even more complicated and direct maximum likelihood is rarely attempted.

For the parameters, Clifford-Hammersley implies that $p\left(\mu,\Sigma|J,Z,Y\right)$, $p\left(\mu_J,\Sigma_J|J,Z\right)$, and $p\left(\lambda|J\right)$ characterize $p\left(\Theta|X,Y\right)$. For the states, $p\left(Z_t|\Theta,J_t,Y_t\right)$ and $p\left(J_t|\Theta,Z_t,Y_t\right)$ for $t=1,...,T$ characterize $p\left(J,Z|\Theta,Y\right)$. Assuming standard conjugate prior distributions for the parameters, $\mu\sim\mathcal{N}$,$\Sigma\sim\mathcal{IW}$, $\mu_J\sim\mathcal{N},\Sigma_J\sim\mathcal{IW}$, and $\lambda\sim\mathcal{B}$, where $\mathcal{IW}$ is an inverted Wishart (multivariate inverted gamma) and $\mathcal{B}$ is the beta distribution,[10] all of the conditional parameter posteriors are conjugate. We make one adjustment allowing for certain parameters to have conditional priors. We assume that Σ and Σ_J have $\mathcal{IW}$ priors, but that $\mu|\Sigma\sim\mathcal{N}\left(a,b\Sigma\right)$ and $\mu_J|\Sigma_J\sim\mathcal{N}\left(a_J,b_J\Sigma_J\right)$. This allows us to draw $\left(\mu,\Sigma\right)$ and $\left(\mu_J,\Sigma_J\right)$ in blocks. Since these parameters are likely to be correlated, this will improve the efficiency of the MCMC algorithm. We now derive the conditional posteriors for λ, J_t, and Z_t.

The posterior for λ is conjugate and is given by Bayes rule as

$$p\left(\lambda|J\right)\propto p\left(J|\lambda\right)p\left(\lambda\right)\propto\left[\lambda^{\sum_{t=1}^{T}J_t}\left(1-\lambda\right)^{T-\sum_{t=1}^{T}J_t}\right]\lambda^{\alpha-1}\left(1-\lambda\right)^{\beta-1}\propto\mathcal{B}\left(\alpha^*,\beta^*\right)$$

where $p\left(\lambda\right)=\mathcal{B}\left(\alpha,\beta\right)$, $\alpha^*=\sum_{t=1}^{T}J_t+\alpha$ and $\beta^*=T-\sum_{t=1}^{T}J_t+\beta$. The conditional

[10] An $n\times n$ matrix Σ has an inverse Wishart distribution, denoted $\mathcal{W}^{-1}\left(a,A\right)$, with scaler parameter $a>0$ and matrix parameter A positive definite, its density is given by:

$$p\left(\Sigma|a,b\right)=\left|A\right|^{\frac{\left(a-n-1\right)}{2}}\left|\Sigma\right|^{-\frac{a}{2}}\exp\left(-\frac{1}{2}tr\left(\Sigma^{-1}B\right)\right).$$

The beta distribution, $\mathcal{B}\left(\alpha,\beta\right)$ for $\alpha,\beta>0$, has support over the unit interval and its density is given by

$$p\left(x|\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)\beta^{\alpha}}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}x^{\alpha-1}\left(1-x\right)^{\beta-1}.$$

posterior for Z_t is normal:

$$\begin{aligned} p\left(Z_t|Y_t, J_t, \Theta\right) &\propto \exp\left(-\frac{1}{2}\left[r_t'\Sigma^{-1}r_t + \left(Z_t-\mu_Z\right)'\Sigma_J^{-1}\left(Z_t-\mu_Z\right)\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left(Z_t-m_t\right)'V_t^{-1}\left(Z_t-m_t\right)\right) \end{aligned}$$

where $r_t = Y_t - \mu - Z_t J_t$ and

$$\begin{aligned} V_t &= \left(J_t\Sigma^{-1} + \Sigma_J^{-1}\right)^{-1} \\ m_t &= \Sigma_J^{-1}\left(J_t\Sigma^{-1}\left(Y_t-\mu\right) + \Sigma_J^{-1}\mu_Z\right). \end{aligned}$$

For the jump times, the conditional posterior is Bernoulli since J_t can only take two values. The Bernoulli probability is

$$\begin{aligned} p\left(J_t=1|\Theta, Z_t, Y_t\right) &\propto p\left(Y_t|J_t=1,\Theta,Z_t\right)p\left(J_t=1|\Theta\right) \\ &\propto \lambda\exp\left(-\frac{1}{2}\left[\left(Y_t-\mu-Z_t\right)'\Sigma^{-1}\left(Y_t-\mu-Z_t\right)\right]\right). \end{aligned}$$

Computing $p\left(J_t=0|\Theta, Z_t, Y_t\right)$ then provides the Bernoulli probability. This completes the specification of our MCMC algorithm. The arguments in Rosenthal (1995a, b) show that the algorithm is in fact polynomial time convergent, and thus, converges quickly.

As all of the conditional posteriors can be directly sampled, the MCMC algorithm is a Gibbs sampler iteratively drawing from

$$\begin{aligned} p\left(\mu,\Sigma|J,Z,Y\right) &\sim \mathcal{N}/\mathcal{IW} \\ p\left(\mu_J,\Sigma_J|J,Z\right) &\sim \mathcal{N}/\mathcal{IW} \\ p\left(\lambda|J\right) &\propto \mathcal{B} \\ p\left(Z_t|\Theta,J_t,Y_t\right) &\sim \mathcal{N} \\ p\left(J_t|\Theta,Z_t,Y_t\right) &\sim \textit{Binomial} \end{aligned}$$

where $\mathcal{N}/\mathcal{IW}$ is the normal-inverted Wishart joint distribution. Sampling from this distribution requires two steps, but is standard (see, for example, Bernardo and Smith (1995)).

To illustrate the methodology, consider a bivariate jump-diffusion model for S&P 500 and Nasdaq 100 equity index returns from 1986-2000. The model is a lower-dimensional version of the those considered in Duffie and Pan (1999) and is given by

$$\begin{pmatrix} Y_t^1 \\ Y_t^2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}^{1/2} \begin{pmatrix} \varepsilon_t^1 \\ \varepsilon_t^2 \end{pmatrix} + J_t \begin{pmatrix} Z_t^1 \\ Z_t^2 \end{pmatrix}$$

Table 1: Parameter estimates for the bi-variate jump-diffusion model for daily S&P 500 and Nasdaq 100 returns from 1986-2000.

	Prior		Posterior		
	Mean	Std	Mean	Std	(5,95)% Credible Set
μ_1	0	$\sqrt{1000}$	0.1417	0.0229	0.1065, 0.1797
μ_2	0	$\sqrt{1000}$	0.0839	0.0148	0.0589, 0.1082
σ_1	1.7770	0.9155	1.2073	0.0191	1.1778, 1.2396
σ_2	1.7770	0.9155	0.7236	0.0369	0.6903, 0.7599
ρ	0	0.1713	0.6509	0.0115	0.6317, 0.6690
λ	0.0476	0.0147	0.0799	0.0081	0.0663, 0.0933
$\mu_{1,J}$	0	$\sqrt{1000}$	-0.5747	0.2131	-0.9320, -0.2351
$\mu_{2,J}$	0	$\sqrt{1000}$	-0.3460	0.1765	-0.6537, -0.0648
$\sigma_{1,J}$	2.1113	1.1715	2.9666	0.1647	2.7073, 3.2435
$\sigma_{2,J}$	2.1113	1.1715	2.5873	0.1458	2.3540, 2.8233
ρ_J	0	0.1519	0.5190	0.0490	0.4360, 0.5986

where $\sigma\sigma' = \Sigma$, $Z_t = [Z_t^1, Z_t^2]' \sim N(\mu_J, \Sigma_J)$ and the jump arrivals, common to both returns, have constant intensity λ.

We run the Gibbs sampler for 1250 iterations and discard the first 250 as a burn-in period, using the last 1000 draws to summarize the posterior distribution. Table 1 provides the prior mean and standard deviation and the posterior mean, standard deviation and a $(5, 95)\%$ credible set. The prior on λ is informative, in the sense that it specifies that jumps are rare events. Our prior represents our belief that the variance of jump sizes is larger than the daily diffusive variance. For all parameters, the data is very informative as the posterior standard deviation is much smaller than the prior indicating that that the parameters are easily learned from the data. This should not be a surprise as returns in the model are i.i.d.. Figure 2 provides parameter trace plots and shows how, after burn-in, the Gibbs sampler moves around the posterior distribution.

Figure 3 provides Monte Carlo estimates of the jump sizes in returns ($Z_t J_t$). Since the model has constant volatility, there are periods when jumps are clustered which is clearly capturing time-variation in volatility that the model does not have built in. We address this issue later by introducing time-varying and stochastic volatility.

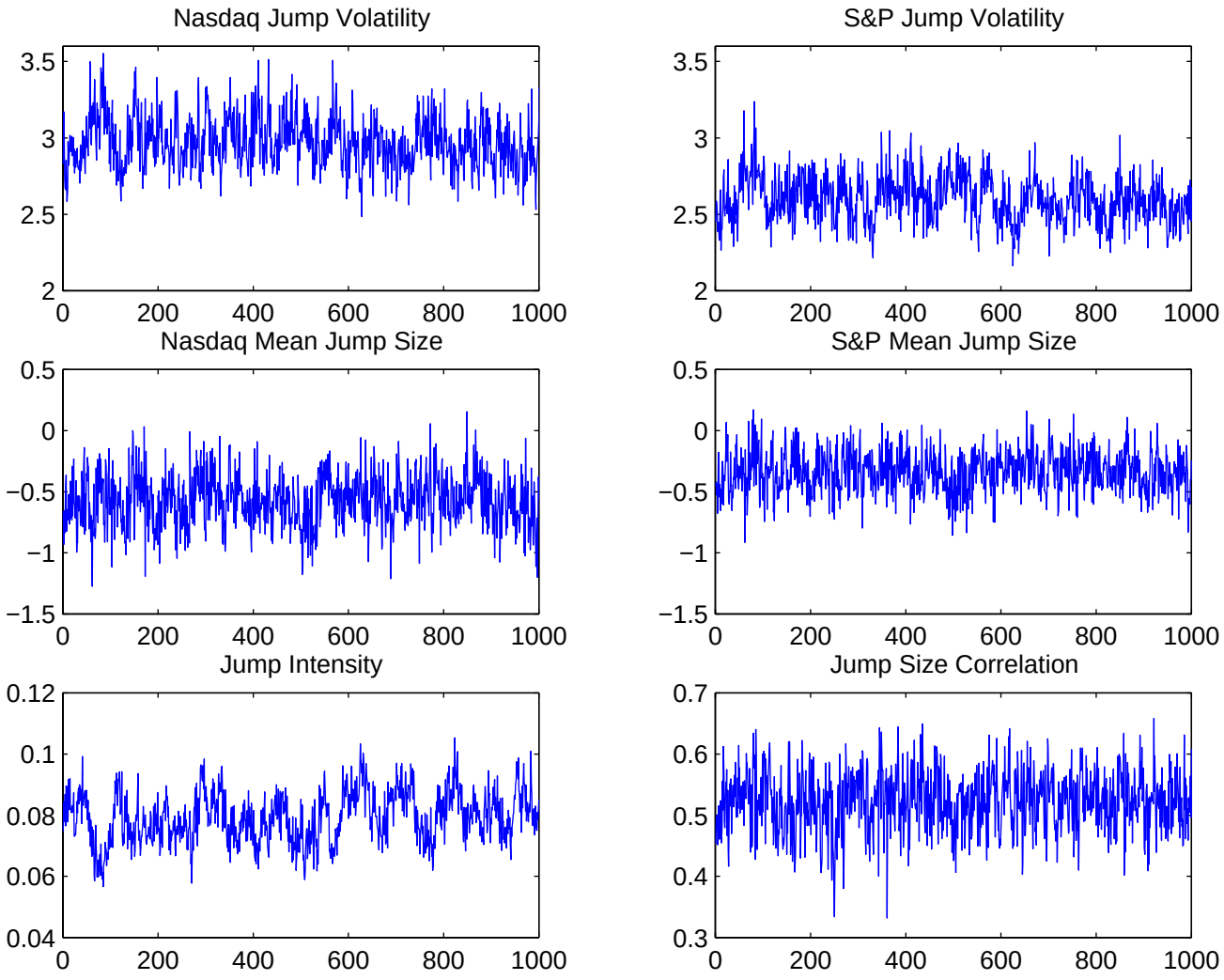

Figure 2: Parameter trace plots for the jump parameters. Each panel shows $\left\{\Theta^{(g)}\right\}_{g=1}^{G}$ for the individual parameters.

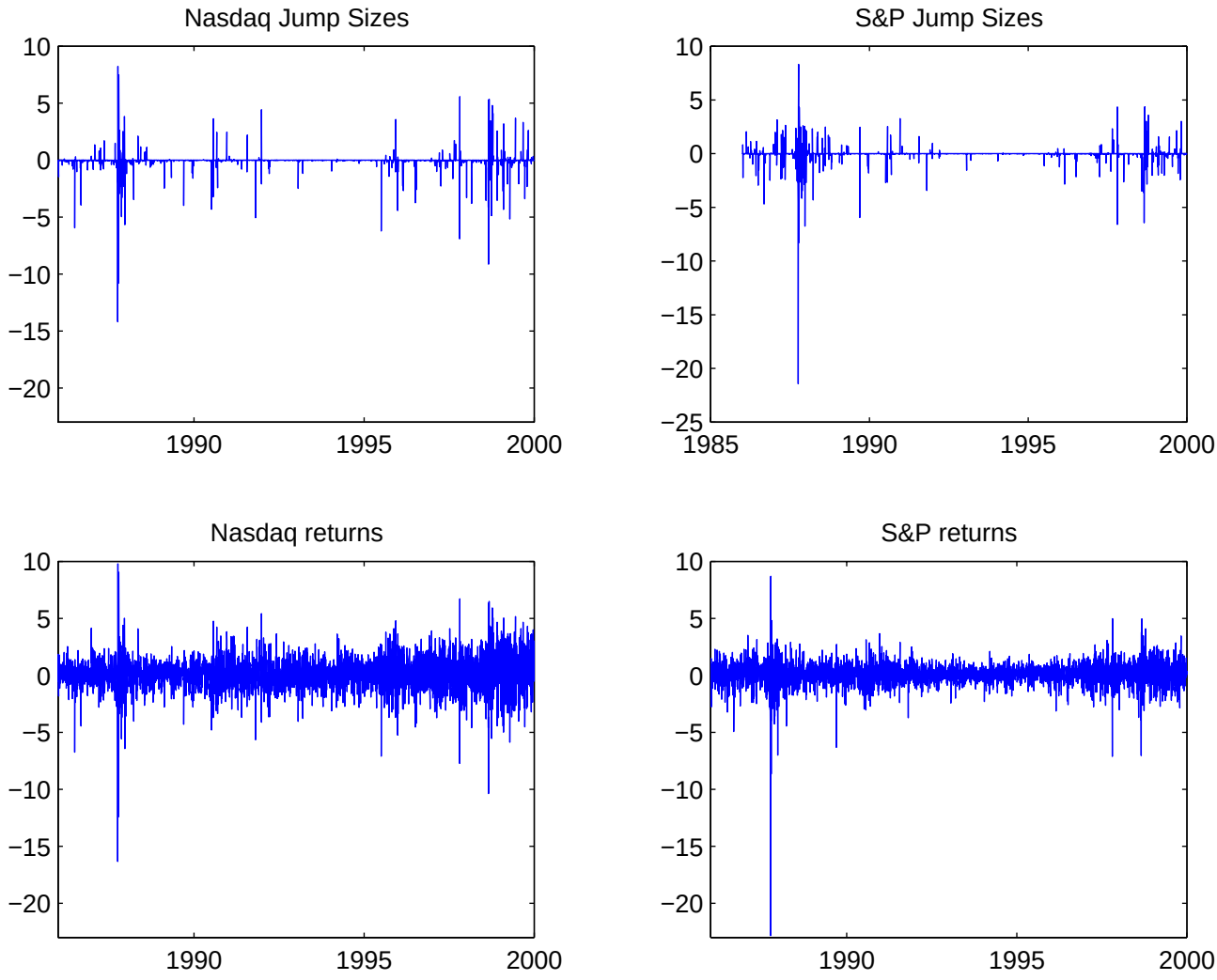

Figure 3: Estimated jump sizes in returns for the Nasdaq and S&P 500 and actual returns over the same period.

5.1.4 Time-Varying Equity Premium

The Black-Scholes model assumes that μ and σ are constant. Extensions of the Black-Scholes model allow these parameters to vary over time. In the case of the expected return, a straightforward extension posits that the equity premium, μ_t, is time-varying:

$$\frac{dS_t}{S_t} = \left[r_t + \mu_t + \frac{1}{2}\sigma^2 \right] dt + \sigma dW_t^s \left(\mathbb{P} \right)$$

where r_t is the spot rate, the equity premium solves

$$d\mu_t = \kappa_\mu \left(\theta_\mu - \mu_t \right) dt + \sigma_\mu dW_t^\mu \left(\mathbb{P} \right)$$

and the Brownian motions could be correlated. The mean-reverting specification for expected returns is popular in the portfolio choice literature and was introduced by Merton (1971) and recently used by Kim and Omberg (1996), Liu (1999), and Wachter (2000). Brandt and Kang (2000) and Johannes, Polson and Stroud (2001) provide empirical analyses of this model.

Solving the SDE, the increments of μ_t are

$$\mu_t = \mu_{t-1} e^{-\kappa_\mu} + \theta_\mu \left(1 - e^{-\kappa_\mu} \right) + \sigma_\mu \int_{t-1}^{t} e^{-\kappa_\mu (t-s)} dW_s^\mu \left(\mathbb{P} \right)$$

and the model is a discrete-time AR(1):

$$\mu_t = \alpha_\mu + \beta_\mu \mu_{t-1} + \sigma_\mu \varepsilon_t^\mu$$

where $\alpha_\mu = \theta_\mu \left(1 - e^{-\kappa_\mu} \right)$, $\beta_\mu = e^{-\kappa_\mu}$ and we have redefined σ_μ. The state space form is

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

where $Y_t = \log \left(\frac{S_t}{S_{t-1}} \right) - \int_{t-1}^{t} r_s ds$ are excess returns. The parameters are $\Theta = \left\{ \alpha_\mu, \beta_\mu, \sigma_\mu, \sigma \right\}$, the state variables are $X = \mu = \left\{ \mu_t \right\}_{t=1}^{T}$, and the posterior distribution is $p\left(\Theta, \mu | Y \right)$. Clifford-Hammersley implies that $p\left(\alpha_\mu, \beta_\mu, \sigma_\mu^2, \sigma^2 | \mu, Y \right)$ and $p\left(\mu | \alpha_\mu, \beta_\mu, \sigma_\mu^2, \sigma^2, Y \right)$ are complete conditionals. Assuming normal-inverted Wishart priors conditional priors for the parameters, the parameters can be updated as a single block.

Drawing from $p(\mu|Y,\alpha,\beta,\sigma_v^2)$, a $T-$dimensional distribution, might appear to be difficult. However, it is possible to use the Kalman filter to obtain this density via the forward-filtering backward sampling (FFBS) algorithm described in Carter and Kohn (1993). This generates the following Gibbs sampler:

$$\begin{aligned} p\left(\alpha_\mu,\beta_\mu,\sigma_\mu^2,\sigma^2|\mu,Y\right) &\sim \mathcal{N}/\mathcal{IG} \\ p\left(\mu|\alpha_\mu,\beta_\mu,\sigma_\mu^2,\sigma^2,Y\right) &: FFBS. \end{aligned}$$

The mechanics of the FFBS algorithm are quite simple. Consider the following decomposition of the joint expected returns posterior:

$$p(\mu|Y,\Theta) \propto p(\mu_T|Y,\Theta)\prod_{t=0}^{T-1} p\left(\mu_t|\mu_{t+1},Y^t,\Theta\right)$$

where $Y^t=[Y_1,...,Y_t]$. To simulate from $p(\mu|Y,\Theta)$, consider the following procedure:

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

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

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

The first step is the usual Kalman filtering algorithm (forward filtering) and then the last two steps move backward to unwind the conditioning information (backward sampling). Anderson and Moore (1979, p. 105) and Carter and Kohn (1994, Appendix 1) show that the samples $(\widehat{\mu}_1,...,\widehat{\mu}_T)$ are a direct block draw from $p(\mu|Y,\Theta)$. It is important to recognize that the Kalman filter is just one part of the MCMC algorithm, and the other step (parameter updating) indicates that the algorithm accounts for parameter uncertainty.

To get a sense of some empirical results, we estimate the model above using S&P 500 returns and Nasdaq 100 returns from 1973 to 2000 and 1987 to 2000, respectively, and report summary statistics of $p(\mu_t|Y)$ over the common period 1987 to 2000. Figure 3 provides posterior estimates of μ_t, $E[\mu_t|Y]$, and 95 percent confidence bounds. Note that the MCMC algorithm provides the entire distribution of the states taking into account estimation risk. Not surprisingly, the risk present in estimating μ_t is quite large. In fact, for many periods of time, the confidence band includes zero, which shows that there is often not a statistically significant equity premium, although the point estimate of the equity premium $E[\mu_t|Y]$ is positive.

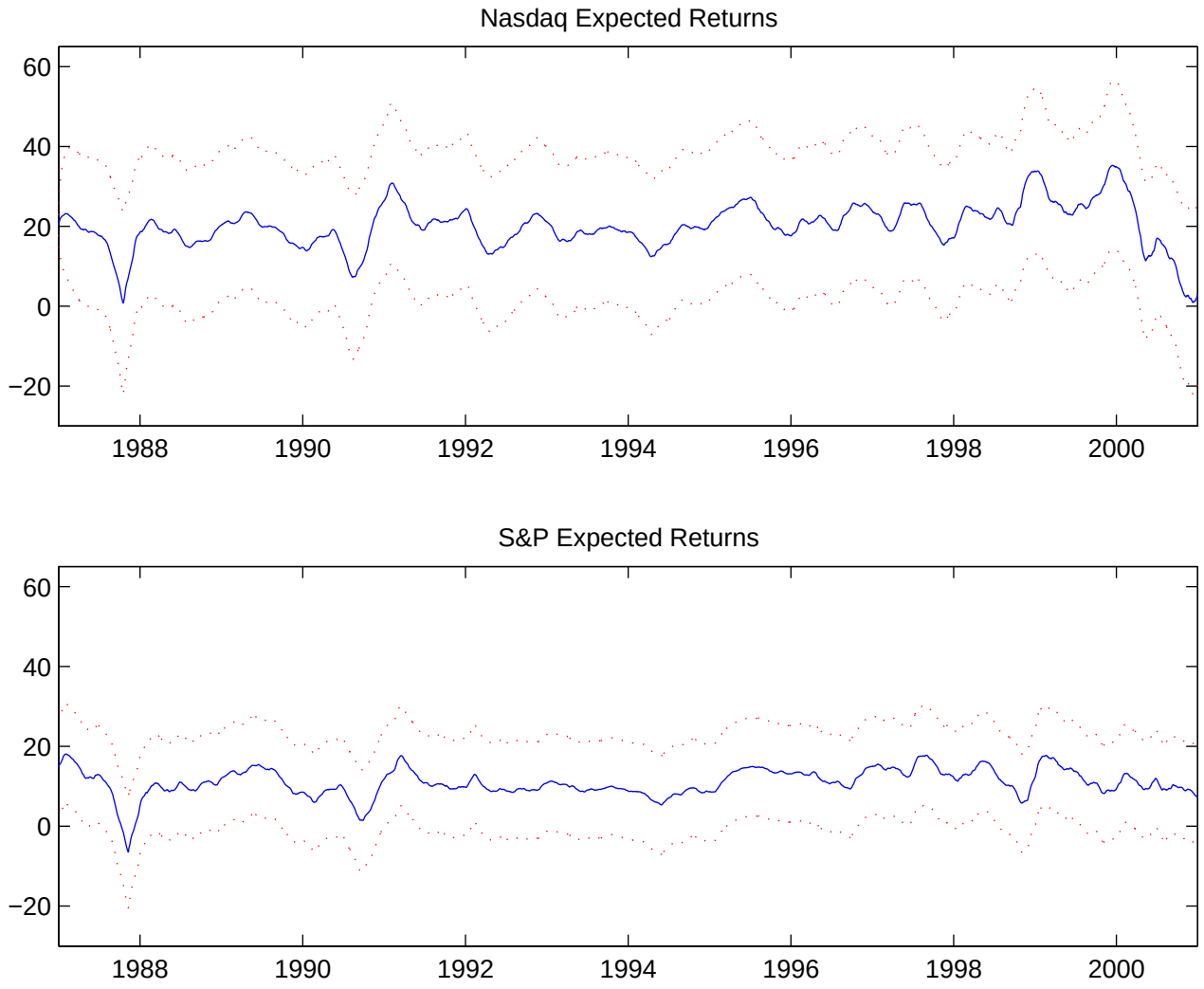

Figure 4: Smoothed expected return paths (with confidence bands) for the S&P 500 and Nasdaq 100 from 1987-2001.

Time-varying parameter models The algorithm above, applies, in a slightly modified form to more general time-varying parameter models. Consider the more general setting of Xia (2001):

$$\frac{dS_t}{S_t} = \left[\alpha_t + \beta_t' X_t + \frac{1}{2}\sigma^2\right] dt + \sigma dW_t^s \left(\mathbb{P}\right)$$

where the equity premium is $\mu_t = \alpha_t + \beta_t' X_t$, X_t is a vector of predictor variables, and β_t is a vector of time-varying coefficients. Xia (2001) assumes they jointly solve

$$\begin{aligned} d\beta_t &= \kappa_\beta \left(\theta_\beta - \beta_t\right) dt + \sigma_\beta dW_t^\beta \left(\mathbb{P}\right) \\ dX_t &= \kappa_x \left(\theta_x - X_t\right) dt + \sigma_x dW_t^x \left(\mathbb{P}\right) \end{aligned}$$

where all Brownian motions can be correlated. In discrete-time, this model takes the form of a linear, Gaussian state-space model. The conditional posteriors for the parameters are all standard, while the conditional posterior for the latent variable, β_t, can be obtained via the FFBS algorithm mentioned above.

Merton's Model of Defaultable Debt The methods in the previous section can be easily adapted to handle Merton's (1974) model of defaultable debt. In this model, a firm has assets with a market value of A_t and has outstanding bond obligations equal to a zero coupon bond expiring at time T with par value B. Equity holders, as residual claimants, receive any excess value over that which is given to the bond holders, that is, at time T the equity holders receive $(A_T - B)_+$. In this case, standard arguments imply that the value of equity, S_t, is given by $S_t = E^{\mathbb{Q}}\left[e^{-r(T-t)}\left(A_T - B\right)_+ | A_t\right]$.

Given this, the state space representation for structural models of default implies that

$$\begin{aligned} S_t &= E_t^{\mathbb{Q}}\left[e^{-r(T-t)}\left(A_T - B\right)_+ | A_t\right] \\ dA_t &= \mu A_t dt + \sigma A_t dW_t. \end{aligned}$$

In the case of geometric Brownian motion for the firm value, the equity price is given by the Black-Scholes formula. It is also important to remember that, from the econometrician's perspective, the firm value, A_t, is an unobserved state variable and estimating it is one of the primary objectives. A time-discretization of this model leads to a combination of the Black-Scholes option pricing model and a time-varying parameter model.

5.1.5 Log-Stochastic Volatility Models

Stochastic volatility models are one of the most successful applications of MCMC methods. A wide range of stochastic volatility models appear in the literature, and they all present difficulties for estimation as they are non-Gaussian and nonlinear state space models. We consider a number of different stochastic volatility models: the log-stochastic volatility model, a model incorporating the leverage effect, Heston's (1993) square-root stochastic volatility model, and the double-jump model of Duffie, Pan, and Singleton (2000).

Log-volatility The log-stochastic volatility is arguably the most popular specification for modeling stochastic volatility. In this model, volatility solves a continuous-time AR(1) in logs:

$$\begin{aligned} d\log(S_t) &= \mu_t dt + \sqrt{V_t} dW_t^s \\ d\log(V_{t+1}) &= \kappa_v(\theta_v - \log(V_t))\, dt + \sigma_v dW_t^v. \end{aligned}$$

where, for simplicity, we assume that the Brownian motions are independent, although this assumption was relaxed in Jacquier, Polson and Rossi (2004). To abstract from conditional mean dynamics, set $\mu_t = 0$. An Euler time discretization implies that

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

where Y_t are the continuously compounded returns, $\alpha_v = \kappa_v\theta_v$ and $\beta_v = 1 - \kappa_v$. This reparameterization allows us to use standard conjugate updating theory for the parameters. Define the parameter and state vectors as $\Theta = \{\alpha_v, \beta_v, \sigma_v^2\}$ and $X = V = \{V_t\}_{t=1}^T$.

Jacquier, Polson and Rossi (1994) were the first to use MCMC methods to analyze this model, and since then, there have a been a number of important alternative MCMC algorithms proposed. The Clifford-Hammersley theorem implies that $p(\Theta, V|Y)$ is completely characterized by $p(\alpha_v, \beta_v|\sigma_v, V, Y)$, $p(\sigma_v^2|\alpha_v, \beta_v, V, Y)$ and $p(V|\alpha_v, \beta_v, \sigma_v^2, Y)$. JPR (1994) assume conjugate priors for the parameters, $\alpha_v, \beta_v \sim \mathcal{N}$ and $\sigma_v^2 \sim \mathcal{IG}$, which implies that

$$p(\alpha_v, \beta_v|\sigma_v, V, Y) \propto \prod_{t=1}^{T} p(V_t|V_{t-1}, \alpha_v, \beta_v, \sigma_v)\, p(\alpha_v, \beta_v) \propto \mathcal{N}$$

and for σ_v, we have that

$$p\left(\sigma_v^2|\alpha_v, \beta_v, V, Y\right) \propto \prod_{t=1}^{T} p(V_t|V_{t-1}, \alpha_v, \beta_v, \sigma_v)\, p\left(\sigma_v^2\right) \propto \mathcal{IG}.$$

The only difficult step arises in updating the volatility states. The full joint posterior for volatility is

$$p\left(V|\Theta,Y\right)\propto p\left(Y|\Theta,V\right)p\left(V|\Theta\right)\propto\prod_{t=1}^{T}p\left(Y_t|V_t,\Theta\right)p\left(V_t|V_{t-1,}\Theta\right)$$

since $p\left(Y|\Theta,V\right)=\prod_{t=1}^{T}p\left(Y_t|V_t,\Theta\right)$ and $p\left(V|\Theta\right)\propto\prod_{t=1}^{T}p\left(V_t|V_{t-1,}\Theta\right)$, by the conditional independence and Markov property, respectively. Now the complete conditional are given by $p\left(V_t|V_{(-t)},\Theta,Y\right)=p\left(V_t|V_{t-1},V_{t+1},\Theta,Y\right)$ by the Markov property: conditional on Y, V_t is only influenced by its local neighbors. By Bayes rule, the complete conditional is given by:

$$\begin{aligned}p\left(V_t|V_{t-1},V_{t+1},\Theta,Y\right) &\propto p\left(V_{t-1},V_t,V_{t+1}|\Theta,Y\right)\\ &\propto p\left(Y_t|V_t,\Theta\right)p\left(V_{t-1},V_t,V_{t+1}|\Theta\right)\\ &\propto p\left(Y_t|V_t,\Theta\right)p\left(V_t|V_{t-1},\Theta\right)p\left(V_{t+1}|V_t,\Theta\right),\end{aligned}$$

where all terms not directly involving V_t are removed. As a function of V_t, the conditional variance posterior is quite complicated:

$$p\left(V_t|V_{t-1},V_{t+1},\Theta,Y\right)\propto V_t^{-\frac{1}{2}}\exp\left(-\frac{Y_t^2}{2V_t}\right)\exp\left(-\frac{e_t^2}{2\sigma_v^2}\right)V_t^{-1}\exp\left(-\frac{e_{t+1}^2}{2\sigma_v^2}\right),$$

where $e_t=\log\left(V_t\right)-\alpha_v-\beta_v\log\left(V_{t-1}\right)$. Note that V_t enters in four different places. As this distribution is not recognizable, Metropolis-Hastings is required to sample from it.

We first consider a "single state" Metropolis updating scheme as the joint volatility posterior, $p\left(V|\Theta,Y\right)$, cannot directly drawn from without approximations. The MCMC algorithm therefore consists of the following steps:

$$\begin{aligned}p\left(\alpha_v,\beta_v|\sigma_v,V,Y\right) &\sim \mathcal{N}\\ p\left(\sigma_v^2|\alpha_v,\beta_v,V,Y\right) &\sim \mathcal{IG}\\ p\left(V_t|V_{t-1},V_{t+1},\Theta,Y\right) &: \mathit{Metropolis}\end{aligned}$$

Jacquier, Polson, and Rossi (1994) use an independence Metropolis-Hastings algorithm to update the states. This is preferable to a random-walk algorithm because we can closely approximate conditional distribution, $p\left(V_t|V_{t-1},V_{t+1},\Theta,Y\right)$, especially in the tails. The proposal density is a Gamma proposal density motivated by the observation that the first

term in the posterior is an inverse Gamma and the second log-normal term can be approximated (particularly in the tails) by a suitable chosen inverse Gamma. If we refer to the proposal density as $q(V_t)$ and the true conditional density as $\pi(V_t) \triangleq p(V_t|V_{t-1}, V_{t+1}, \Theta, Y)$, this implies the Metropolis-Hastings step is given by:

$$\begin{aligned}&\text{Step 1. Draw } V_t^{(g+1)} \text{ from } q(V_t)\\ &\text{Step 2. Accept } V_t^{(g+1)} \text{ with probability } \alpha\left(V_t^{(g+1)}, V_t^{(g)}\right)\end{aligned}$$

where

$$\alpha\left(V_t^{(g)}, V_t^{(g+1)}\right) = \min\left(\frac{\pi\left(V_t^{(g+1)}\right) q\left(V_t^{(g)}\right)}{\pi\left(V_t^{(g)}\right) q\left(V_t^{(g+1)}\right)}\right).$$

As shown by Jacquier, Polson, and Rossi (1994) using simulations, this algorithm provides accurate inference. Given that the gamma distribution bounds the tails of the true conditional density, the algorithm is geometrically convergent.

Figure 2 provides posterior means, $E(V_t|Y)$, of the latent volatility states with (5,95)% credible sets for the S&P 500 and Nasdaq 100. These are smoothed volatility estimates, as opposed to filtered volatility estimates, and account for estimation risk as they integrate out parameter uncertainty.

Correlated Shocks: The Leverage Effect One common extension of the model presented above relax the assumption that the shocks in volatility and prices are independent and instead assume that $corr(W_t^v, W_t^s) = \rho$. This "leverage" effect of Black (1976) has been shown to be an important component of returns. For equity returns, this parameters is negative which indicates that negative returns signal increases in volatility. For MCMC estimation, the leverage effect adds two complications. First, updating volatility is slightly more complicated as equity returns now are directionally correlated with changes in volatility. Second, there is an additional parameter that is present.

Jacquier, Polson and Rossi (2004) consider the log-volatility model and show how to incorporate the leverage effects into the model. In discrete-time, they write the model as:

$$\begin{aligned} Y_t &= \sqrt{V_{t-1}}\varepsilon_t^s \\ \log(V_t) &= \alpha_v + \beta_v \log(V_{t-1}) + \sigma_v\left[\rho\varepsilon_t^s + \sqrt{1-\rho^2}\varepsilon_t^v\right]\end{aligned}$$

where ε_t^s and ε_t^v are uncorrelated. Jacquier, Polson and Rossi (2004) reparameterize the model by defining $\phi_v = \sigma_v\rho$ and $\omega_v = \sigma_v^2(1-\rho^2)$. They assume $\alpha_v, \beta_v \sim \mathcal{N}$, $\phi_v \sim \mathcal{N}$ and

$\omega_v \sim \mathcal{IG}$. This generates the following MCMC algorithm:

$$\begin{aligned} p\left(\alpha_v, \beta_v | \sigma_v, V, Y\right) &\sim \mathcal{N} \\ p\left(\phi_v, \omega_v | \alpha_v, \beta_v, V, Y\right) &\sim \mathcal{N}\text{-}\mathcal{IG} \\ p\left(V_t | V_{t-1}, V_{t+1}, \Theta, Y\right) &: \quad Metropolis. \end{aligned}$$

We refer the reader to JPR (2004) for the details of each of these steps and for extensions to multivariate stochastic volatility models.

Blocking Volatility States Since the volatility states are correlated, one would ideally like to update them in a block. Unfortunately, direct block updating is extremely difficult and therefore a number of authors have considered an approximation to the model which can then be used to update volatility in a block. One alternative is to approximate the model, in hopes that the approximating model will be negligibly different and will allow black updating. Kim, Shephard and Chib (1998) first square returns and add a small constant, to avoid taking the logarithm of zero. In log-form, the state space model is

$$\begin{aligned} \log\left(\left(Y_{t+1}^2 + c\right)\right) &= \log\left(V_t\right) + \log\left(\left(\varepsilon_{t+1}^s\right)^2\right) \\ \log(V_{t+1}) &= \alpha_v + \beta_v \log(V_t) + \sigma_v \varepsilon_{t+1}^v. \end{aligned}$$

If we re-label the log-volatilities as $\widetilde{V}_t = \log\left(V_t\right)$, we see the model now takes the form of a non-normal, but linear state space model:

$$\begin{aligned} \log\left(\left(Y_{t+1}^2 + c\right)\right) &= \widetilde{V}_t + \log\left(\left(\varepsilon_{t+1}^s\right)^2\right) \\ \widetilde{V}_{t+1} &= \alpha_v + \beta_v \widetilde{V}_t + \sigma_v \varepsilon_{t+1}^v. \end{aligned}$$

This simplifies the model along one line by removing the non-linearity, but introduces non-normalities.

Kim, Shephard and Chib (1998) argue that $x_{t+1} = \log\left(\left(\varepsilon_{t+1}^s\right)^2\right)$ can be easily approximated by a 7-component mixture of normals:

$$p\left[x_t\right] \approx \sum_{j=1}^{7} q_j \phi\left(x_t | \mu_j, \sigma_j^2\right)$$

where $\phi\left(x|\mu, \sigma^2\right)$ is a normal distribution with mean μ and variance σ^2. The constants q_j, μ_j and σ_j are chosen to approximate the distribution of $\log\left(\left(\varepsilon_{t+1}^s\right)^2\right)$. Formally, this

requires the addition of a latent state variable, s_t, such that

$$\begin{aligned} x_t|s_t &= j \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right) \\ \text{Prob}\left[s_t = j\right] &= q_j. \end{aligned}$$

In this transformed model, the posterior is $p\left(\alpha_v, \beta_v, \sigma_v^2, \widetilde{V}, s|Y\right)$. The MCMC algorithm is now

$$\begin{aligned} p\left(\alpha_v, \beta_v|\sigma_v, s, \widetilde{V}, Y\right) &\sim \mathcal{N} \\ p\left(\sigma_v^2|\alpha_v, \beta_v, s, \widetilde{V}, Y\right) &\sim \mathcal{IG} \\ p\left(\widetilde{V}|\alpha_v, \beta_v, \sigma_v, s, Y\right) &: FFBS \\ p\left(s_t|\alpha_v, \beta_v, \sigma_v, Y, \widetilde{V}\right) &\sim Multinomial. \end{aligned}$$

The key advantage is that, conditional on the indicators, the model is a linear, normal state space model and the Kalman recursions deliver $p\left(\widetilde{V}|\alpha_v, \beta_v, \sigma_v, s, Y\right)$. Further details of the algorithm and the exact conditional posteriors are given in Kim, Shephard and Chib (1998).

The algorithm generates a Markov Chain with very low autocorrelations. Due to the low autocorrelations, the algorithm is often referred to as a rapidly converging algorithm. This is certainly true of the approximated model, if the shocks to the volatility equation have the postulated mixture representation. If data is simulated from the true distribution, the algorithm could provide inaccurate inference as the state evolution is misspecified. However, this affect appears to be quite small for many financial time series. There are other potential problems with this algorithm. It suffers from the inlier problem when $\varepsilon_{t+1}^s \approx 0$, and it cannot handle models with a leverage effect ($corr(\varepsilon_{t+1}^s, \varepsilon_{t+1}^v) \neq 0$). Moreover, it drastically increases the state space by introducing indicator variables. Rosenthal (1995) discusses potential convergence problems with Markov Chains over discrete state spaces. The algorithm is difficult to extend to other interesting cases, such as the square-root stochastic volatility model, although this is an area of research that certainly deserves further attention.

5.1.6 Alternative Stochastic Volatility Models

Although the log-volatility model is common for many applications, it has a number of potential shortcomings. First, the model falls outside the affine class which implies it is

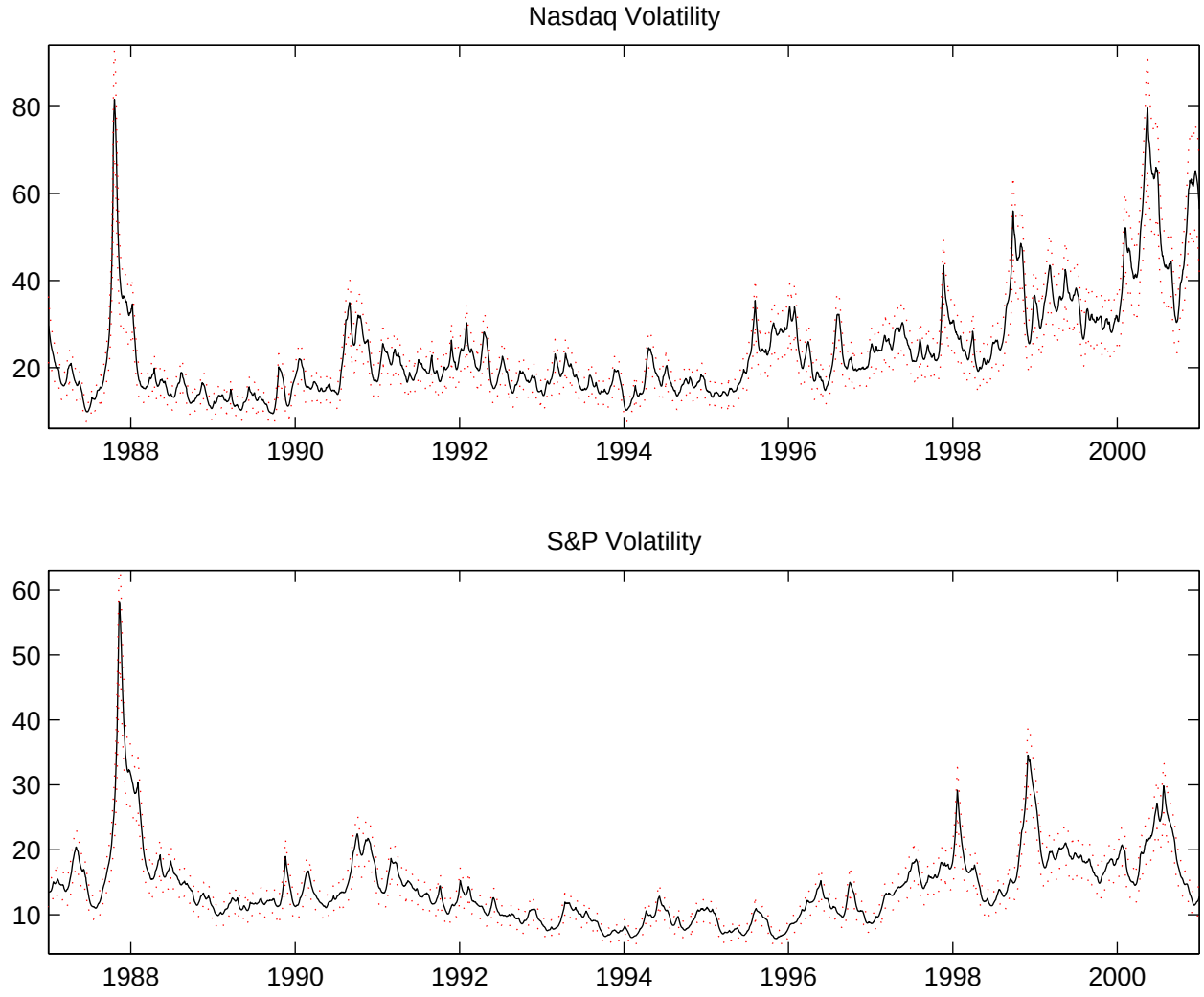

Figure 5: Smoothed volatility paths (with posterior confidence bands) for the S&P 500 and Nasdaq 100 from 1987-2001.

numerically costly to compute option prices or portfolio weights in applications as partial differential equation must be solved. Second, the volatility of volatility is constant, a potentially counterfactual implication. To address these concerns, a number of alternative models have been introduced into the literature.

Heston's Square-Root Volatility Model In the option pricing and portfolio allocation literature, it is common to use an "affine" model specification for volatility. Heston (1993) introduced the square-root stochastic volatility model

$$dS_t = S_t\left(r_t + \eta_v V_t + \frac{1}{2}V_t\right)dt + S_t\sqrt{V_t}dW_t^s\left(\mathbb{P}\right) \tag{30}$$

$$dV_t = \kappa_v\left(\theta_v - V_t\right)dt + \sigma_v\sqrt{V_t}dW_t^v\left(\mathbb{P}\right) \tag{31}$$

where the Brownian motions have constant correlation ρ. Discretizing the SDE, we have that

$$\begin{aligned} Y_t &= \eta_v V_{t-1} + \sqrt{V_{t-1}}\varepsilon_t^s \\ V_t &= \alpha_v + \beta_v V_{t-1} + \sigma_v\sqrt{V_{t-1}}\varepsilon_t^v \end{aligned}$$

where Y_t are excess returns, and we have re-defined the parameters in the drift process of the volatility process.

Before continuing with the algorithm, it is important to note that there is a clear problem with the time-discretization of the square-root model. Provided the Feller condition holds, the SDE has a positive solution, that is, starting from an initially positive volatility, the solution remains positive for all t. The time-discretized model does not share this property as the shocks to V_t are normally distributed, which implies that a simulated V_t could be negative. There are three ways to deal with this. First, the original SDE could be transformed by Ito's lemma into logarithms, and the solution simulated in logs.[11] This makes parameter updating more difficult as the volatility appears in both the drift and diffusion. Second, one can ignore the problem and hope that it does not affect the results. For certain time series, such as U.S. equity indices, volatility tends to be rather high and the problem is likely very small as $p\left(V_t|V_{t-1}\right)$ has very little support below zero, especially

[11] By Ito's lemma, we for $log\left(V_t\right) = h_t$, $dh_t = e^{h_t}\left[k_v\left(\theta_v - e^{h_t}\right) - \frac{1}{2}\sigma_v^2\right]dt + e^{h_t/2}dW_t^v$. Simulating this process in discrete-time in logarithms guarantees that $V_t = \exp\left(h_t\right) > 0$.

when the discretization interval is daily.[12] For other time series, such as exchange rates where volatility is very low, ignoring the problem may not as innocuous. Third, one could fill in missing data points to reduce the impact of discretization.

We assume normal independent priors for η_v and (α_v, β_v), an inverted Gamma prior for σ_v^2, and a uniform prior for ρ. Eraker, Johannes, and Polson (2003) examine this model using MCMC methods, as well as extensions that include jumps in returns and jump in volatility. The Clifford-Hammersley theorem implies that $p(\alpha_v, \beta_v|\sigma_v, \rho, V, Y)$, $p(\sigma_v^2|\alpha_v, \beta_v, \rho, V, Y)$, $p(\rho|\alpha_v, \beta_v, \sigma_v^2, V, Y)$ and $p(V|\alpha_v, \beta_v, \sigma_v^2, Y)$ are the complete conditionals. The MCMC algorithm is given by:

$$\begin{aligned}
p(\eta_v|\alpha_v, \beta_v, \sigma_v, \rho, V, Y) &\sim \mathcal{N} \\
p(\alpha_v, \beta_v|\sigma_v, \rho, V, Y) &\sim \mathcal{N} \\
p\left(\sigma_v^2|\alpha_v, \beta_v, \rho, V, Y\right) &\sim \mathcal{IG} \\
p\left(\rho|\alpha_v, \beta_v, \sigma_v^2, V, Y\right) &: \quad Metropolis \\
p(V_t|V_{t-1}, V_{t+1}, \Theta, Y) &: \quad Metropolis.
\end{aligned}$$

The parameter posteriors are similar to those in the previous sections and are omitted. Eraker, Johannes, and Polson (2003) use a random walk Metropolis-Hastings algorithm for both the correlation parameter and the volatility states. The functional form of $p(\rho|\alpha_v, \beta_v, \sigma_v^2, V, Y)$ and $p(V_t|V_{t-1}, V_{t+1}, \Theta, Y)$ are particularly complicated. In both cases, the state dependence in the variance of V_t creates complications and it is difficult to find a good proposal for independence Metropolis. Eraker, Johannes, and Polson (2003) provide a simulation study to to show the efficacy of their algorithm to estimate the parameters of the underlying continuous-time process. It would be particularly useful if a block-updating scheme were available for updating the volatilities.

Stochastic volatility with jumps in returns and volatility As an example, consider the popular double-jump model of Duffie, Pan, and Singleton (2000) which assumes the

[12] For example, using the estimates in Eraker, Johannes, and Polson (2003), we have that

$$\begin{aligned}
V_{t+1}|V_t &\sim N\left(V_t + \kappa_v(\theta_v - V_t), \sigma^2 V_t\right) \\
&\sim N\left(V_t + 0.02(0.9 - V_t), (0.14)^2 V_t\right).
\end{aligned}$$

If daily volatility is 1%, $V_t = 1$ (roughly 15 percent annualized) it requires a 50 standard deviation move to make volatility go negative.

equity price, S_t, and its stochastic variance, V_t, jointly solve

$$dS_t = S_t\left(r_t + \eta_v V_t\right)dt + S_t\sqrt{V_t}dW_t^s\left(\mathbb{P}\right) + d\left(\sum_{j=1}^{N_t(\mathbb{P})} S_{\tau_j -}\left(e^{Z_j^s(\mathbb{P})} - 1\right)\right) \quad (32)$$

$$dV_t = \kappa_v\left(\theta_v - V_t\right)dt + \sigma_v\sqrt{V_t}dW_t^v\left(\mathbb{P}\right) + d\left(\sum_{j=1}^{N_t(\mathbb{P})} Z_j^v\left(\mathbb{P}\right)\right) \quad (33)$$

where $W_t^s\left(\mathbb{P}\right)$ and $W_t^v\left(\mathbb{P}\right)$ are correlated (ρ) Brownian motions, $N_t\left(\mathbb{P}\right) \sim Poisson\left(\lambda\right)$, τ_j are the jump times, $Z_j^s\left(\mathbb{P}\right)|Z_j^v \sim N\left(\mu_s + \rho_s Z_j^v, \sigma_s^2\right)$ are the return jumps, $Z_j^v\left(\mathbb{P}\right) \sim \exp\left(\mu_v\right)$ are the variance jumps, and r_t is the spot interest rate. This model plays a prominent role given its importance for practical applications such as option pricing and portfolio analysis. Heston (1993) introduced the square-root stochastic volatility specification and Bates (1996, 2001), Pan (2001) and Duffie, Pan, and Singleton (2000) introduced generalizations with jumps in returns and volatility. Eraker, Johannes and Polson (2003) estimate stochastic volatility models with jumps in returns and volatility using MCMC methods. Eraker (2002) extends Eraker, Johannes, and Polson (2003) to incorporate option prices. Liu, Longstaff and Pan (2001) analyze the portfolio implications of models with jumps in stock prices and in volatility.

A time discretization of this model

$$\begin{aligned} Y_t &= \mu + \eta_v V_{t-1} + \sqrt{V_{t-1}}\varepsilon_t^s + J_t Z_t^s \\ V_t &= \alpha_v + \beta_v V_{t-1} + \sigma_v\sqrt{V_{t-1}}\varepsilon_t^v + J_t Z_t^v. \end{aligned}$$

Given the time discretization, Clifford-Hammersley implies we can factor the parameters and states into the following groups, $[(\mu, \eta_v), (\alpha_v, \beta_v), \sigma_v^2, \rho, \lambda, \mu_v, (\mu_s, \rho_s), \sigma_s^2, J, Z^s, Z^v, V]$. We assume normal independent priors for (μ, η_v), (α_v, β_v), and (μ_s, ρ_s), inverted Gamma priors for σ_v^2 and σ_s^2, a Beta prior for λ, an Gamma prior for μ_v, and a uniform prior for ρ.

Although the model has a number of parameters, deriving the conditional posteriors for many of them is straightforward given the results in the previous section. This is due to the modular nature of MCMC algorithms. For example, the conditional posteriors for the "diffusive" parameters are the same as in the previous section, with an adjusted return and volatility series. Conditional on jump times and sizes, we can define the jump-adjusted returns and volatilities to get

$$\begin{aligned} \widetilde{r}_t &= Y_t - J_t Z_t^s = \mu + \eta_v V_{t-1} + \sqrt{V_{t-1}}\varepsilon_t^s \\ \widetilde{V}_t &= V_t - J_t Z_t^v = \alpha_v + \beta_v V_{t-1} + \sigma_v\sqrt{V_{t-1}}\varepsilon_t^v \end{aligned}$$

which implies the functional forms conditional posteriors for $(\mu, \eta_v), (\alpha_v, \beta_v), \sigma_v^2$, and ρ are the same as in the previous section. Drawing λ is the same as in previous section. Conditional on the jump sizes, the parameters of the jump distributions are conjugate.

The MCMC algorithm draws from the conditional parameter posteriors

$$\begin{aligned}
p(\mu, \eta_v | ..., J, Z, V, Y) &\sim \mathcal{N} \\
p(\alpha_v, \beta_v | ..., J, Z, V, Y) &\sim \mathcal{N} \\
p\left(\sigma_v^2 | ..., J, Z, V, Y\right) &\sim \mathcal{IG} \\
p(\lambda | J) &\sim \mathcal{B} \\
p(\mu_s, \rho_s | ..., J, Z^s, Z^v) &\sim \mathcal{N} \\
p\left(\sigma_s^2 | ..., J, Z^s, Z^v\right) &\sim \mathcal{IG} \\
p(\mu_v | ..., J, Z, V, Y) &\sim \mathcal{G} \\
p\left(\rho | \alpha_v, \beta_v, \sigma_v^2, V, Y\right) &: \quad Metropolis
\end{aligned}$$

and the conditional state variable posteriors

$$\begin{aligned}
p(Z_t^v | ..., Z_t^s, J_t, V_t, V_{t-1}) &\sim \mathcal{TN} \\
p(Z_t^s | ..., Z_t^v, J_t, Y_t, V_t, V_{t-1}) &\sim \mathcal{N} \\
p(J_t = 1 | ..., Z_t^s, Z_t^v, Y_t, V_t, V_{t-1}) &\sim Bernoulli \\
p(V_t | V_{t-1}, V_{t+1}, \Theta, Y) &: \quad Metropolis.
\end{aligned}$$

For both ρ and the volatilities, Eraker, Johannes, and Polson (2003) use a random walk Metropolis algorithm, properly tuned to deliver acceptance rates in the 30-60% range. Eraker, Johannes, and Polson (2003) provide simulation evidence to document the algorithm's ability to estimate the parameters of interest.

5.2 Term Structure Models

One of the great successes of continuous-time asset pricing are term structure models. These models start with a specification for the instantaneous spot rate, r_t, under both the risk-neutral and objective measures, from which bond prices are computed. These models can be justified both on equilibrium (Cox, Ingersoll, and Ross (1985)) and arbitrage grounds and provide an excellent framework for understanding the cross-section and dynamics of bond prices.

Term structure models pose a number of difficult problems for estimation. In general, the parameters enter the state space model in a highly nonlinear fashion, often non-analytically. For example, in general affine models, the parameters appear in ODE's that can only be solved numerically. Second, most models specify a low-dimensional state vector that drives all bond prices. When the number of observed yields or bond prices is greater than the number of state variables, there is a stochastic singularity as the observed yields never conform exactly to the specified model.

We discuss a number of common models, although our discussion is by no means complete. We refer the reader to papers by Lamoureaux and Witte (2002), Polson and Stroud (2002), and Bester (2003) for multifactor implementations using MCMC methods.

5.2.1 Vasicek's Model

The first term structure model we consider is the univariate, Gaussian model of Vasicek (1977) which assumes that r_t solves a continuous-time AR(1) on $(\Omega, \mathcal{F}, \mathbb{P})$:

$$dr_t = \left(a_r^{\mathbb{P}} - b_r^{\mathbb{P}} r_t\right) dt + \sigma_r dW_t^r \left(\mathbb{P}\right),$$

where $W_t^r \left(\mathbb{P}\right)$ is a standard Brownian motion.[13] Assuming a general, "essentially affine" risk premium specification, (see the review paper of Dai and Singleton (2003) for details), the spot rate evolves under the equivalent martingale measure $\mathbb{Q}$ via

$$dr_t = \left(a_r^{\mathbb{Q}} - b_r^{\mathbb{Q}} r_t\right) dt + \sigma_r dW_t^r \left(\mathbb{Q}\right)$$

where $W_t^r \left(\mathbb{Q}\right)$ is a standard Brownian motion on $(\Omega, \mathcal{F}, \mathbb{P})$. We label $\Theta^{\mathbb{Q}} = \left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}\right)$ and $\Theta^{\mathbb{P}} = \left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, \sigma_r\right)$. To avoid any confusion, we explicitly label $\mathbb{P}$ and $\mathbb{Q}$ measure parameters. Given the risk premium specifications, the price of a zero coupon, default-free bond maturing at time τ is

$$P\left(r_t, \tau\right) = E_t^{\mathbb{Q}}\left[e^{-\int_t^{t+\tau} r_s ds}\right] = \exp\left(\beta\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \tau\right) + \beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right) r_t\right)$$

where the loading functions are known in closed form:

$$\begin{aligned}
\beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right) &= \frac{1}{b_r^{\mathbb{Q}}}\left(e^{-b_r^{\mathbb{Q}}\tau} - 1\right) \\
\beta\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \tau\right) &= \frac{1}{2}\left[\frac{\sigma_r^2}{\left(b_r^{\mathbb{Q}}\right)^2} + -\frac{a_r^{\mathbb{Q}}}{b_r^{\mathbb{Q}}}\right]\left(\tau - \beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right)\right) - \frac{\sigma_r^2}{4b_r^{\mathbb{Q}}}\beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right)^2.
\end{aligned}$$

[13] We note that MCMC easily can accomodate a model of the form in Duarte (2002) which has a nonlinear drift under $\mathbb{P}$ but a linear drift under $\mathbb{Q}$.

We assume that a there exist a panel of zero coupon, continuously-compounded yields $Y_t = [Y_{t,\tau_1}, ..., Y_{t,\tau_k}]$ where $Y_{t,\tau} = -\log\left(P\left(\Theta, r_t, \tau\right)\right)$ and the maturities are $\tau = \tau_1, ..., \tau_n$.[14] In this model, if the parameters are known, the spot rate is observable from a single yield. If four yields are observed, the yields can be inverted to compute $a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r$ and r_t without error, in much the same way volatility is often "implied" from option prices in the Black-Scholes model.

To break this stochastic singularity, it is common to add an additive pricing error:[15]

$$\begin{aligned} Y_t &= \beta\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \tau\right) + \beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right) r_t + \varepsilon_t \\ r_{t+1} &= r_t + a_r^{\mathbb{P}} + b_r^{\mathbb{P}} r_t + \sigma_r \varepsilon_{t+1}^r \end{aligned}$$

where, for notational simplicity, we relabel $Y_{t,\tau}$ as the log-bond prices, $\varepsilon_t^r \sim \mathcal{N}(0,1)$ is standard normal, and $\varepsilon_t \sim \mathcal{N}(0, \Sigma_\varepsilon)$ is the vector of pricing errors. Since the spot rate evolution is Gaussian, an alternative is to use the exact transitions for the spot rate:

$$r_{t+1} = r_t e^{-b_r^{\mathbb{P}}} + \left(1 - e^{-b_r^{\mathbb{P}}\Delta}\right) \frac{a_r^{\mathbb{P}}}{b_r^{\mathbb{P}}} + \int_t^{t+1} e^{-b_r^{\mathbb{P}}(t-s)} \sigma_r dW_t^r\left(\mathbb{P}\right).$$

For parameters typically estimated from data and for common time-intervals such as daily or weekly, the discretization bias is negligible.

Before we discuss MCMC estimation, we provide a brief and informal discussion of parameter identification. Given our risk premium assumptions, it is clear that $a_r^{\mathbb{Q}}$ and $b_r^{\mathbb{Q}}$ are identified solely from the cross-section of bond prices, $a_r^{\mathbb{P}}$ and $b_r^{\mathbb{P}}$ are identified solely by the dynamics of the short rate, and σ_r is identified jointly from the cross-section of bond prices and the dynamics of the short rate. The average slope and curvature of the yield curve determine the risk premium parameters, as they are assumed to be constant over time. The parameter $a_r^{\mathbb{Q}}$ enters linearly into $\beta\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \tau\right)$, and thus it plays the role of a constant regression parameter. It controls the average long run level of the yield curve. Since $\left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}\right)$ do not appear in the bond yield expressions, they enter only as regression parameters in the state evolution. While easy to estimate in principle, interest rates are very persistent which implies that long time series will be required to accurately estimate the drift.

[14] If discretely compounded bond yields are observed, they can be converted from the discount basis to continuously compounded rates. If par rates or swap rates are observed, it is common to bootstrap these rates using interpolation, if necessary, to obtain zero coupon bond prices.

[15] An alternative justification, which is also plausible, is that the model is misspecified and the pricing error captures some of this misspecification.

Finally, σ_r enters both in the yields and the dynamics. In principle, either the dynamics of the short rate or the cross-section should identify this parameter as it enters linearly in the bond yields or as a variance parameter in the regression. However, recent research indicates that yield-based information regarding volatility is not necessarily consistent with information based on the dynamics of the spot rate, a time-invariant version of the so-called unspanned volatility puzzle (see, Collin-Dufresne and Goldstein (2002) or Collin-Dufresne, Goldstein and Jones (2003)). This implies that it may be difficult to reconcile the information regarding spot rate volatility from yields and the dynamics of spot rates. Again, as in the case of Black-Scholes implied volatility, this is not a problem with the model or an estimation scheme per se, rather it is indicative of a sort of misspecification encountered when applying these models to real data.

For the objective measure parameters, we choose standard conjugate priors, $\left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}\right) \sim \mathcal{N}$ and $\sigma_r \sim \mathcal{IG}$. One might also want to impose stationarity, that is, $b_r^{\mathbb{P}} > 0$, which could be imposed by using a truncated prior or just by removing any draws in the MCMC algorithm for which $b_r^{\mathbb{P}} < 0$. We assume that $\Sigma_\varepsilon \sim \mathcal{IW}$ and that $\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}\right) \sim \mathcal{N}$. The posterior distribution is $p\left(\Theta, r|Y\right)$ where $\Theta = \left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \Sigma_\varepsilon\right)$, $r = (r_1, ..., r_T)$, and $Y = (Y_1, ..., Y_T)$. The MCMC algorithm consists of the following steps:

$$\begin{aligned}
p\left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}|\sigma_r, r\right) &\sim \mathcal{N} \\
p\left(a_r^{\mathbb{Q}}|\sigma_r, \Sigma_\varepsilon, r, Y\right) &\sim \mathcal{N} \\
p\left(\Sigma_\varepsilon|a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, r, Y\right) &\sim \mathcal{IW} \\
p\left(b_r^{\mathbb{Q}}|a_r^{\mathbb{Q}}, \sigma_r, \Sigma_\varepsilon, r, Y\right) &: \; RW\ Metropolis \\
p\left(\sigma_r^2|a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \Sigma_\varepsilon, r, Y\right) &: \; Metropolis \\
p\left(r|a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \Sigma_\varepsilon, Y\right) &: \; FFBS.
\end{aligned}$$

The updates for $a_r^{\mathbb{P}}, b_r^{\mathbb{P}}$, $a_r^{\mathbb{Q}}$, and Σ_ε are conjugate and the spot rates can be updated in a single block using the FFBS algorithm developed earlier in Section 5.1.4. It is not possible to directly draw interest rate volatility and the risk-neutral speed of mean reversion as the conditional posterior distributions are not standard due to the complicated manner in which these parameters enter into the loading functions. Since $b_r^{\mathbb{Q}}$ only appears in the yield equation, it can be difficult to generate a reasonable proposal for independence Metropolis, and thus we recommend a fat-tailed random walk Metropolis step for $b_r^{\mathbb{Q}}$. The Griddy Gibbs sampler would be also be appropriate. For σ_r, the conditional posterior is given as

$$p\left(\sigma_r^2|a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \Sigma_\varepsilon, r, Y\right) \propto p\left(r|a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, \sigma_r\right) p\left(Y|a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \Sigma_\varepsilon, r\right) p\left(\sigma_r^2\right)$$

which is not a recognizable distribution. The Griddy Gibbs sampler, random walk Metropolis or independence Metropolis are all possible for updating $\sigma_r^{\mathbb{P}}$. For independence Metropolis since, as a function of σ_r, $p\left(Y|a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \Sigma_\varepsilon, r\right)$ is also not a recognizable, one could propose from $p\left(r|a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, \sigma_r^{\mathbb{P}}\right) p\left(\sigma_r^{\mathbb{P}}\right) \sim \mathcal{IG}$ and accept/reject based on the yields. If the information regarding volatility is consistent between the spot rate evolution and yields, this approach will work well. As in all cases when Metropolis is applied, we recommend trying multiple algorithms and choosing the one that has both good theoretical and empirical convergence properties.

5.2.2 Vasicek with Jumps

A number of authors argue that interest rates contain a jump component, in addition to the usual diffusion components. These jumps are often generated by news about the macroeconomy and the jumps arrive at either deterministic or random times. For simplicity, we focus on the latter case, although Piazzesi (2003) addresses the former case. Jumps occurring at predictable times are in fact easier to deal with as there is no timing uncertainty, and in discrete-time, the jump component consists of a random size multiplied by a dummy variable indicating the announcement date.

Consider an extension of Vasicek's (1977) model to incorporate jumps in the short rate:

$$dr_t = \left(a_r^{\mathbb{P}} - b_r^{\mathbb{P}} r_t\right) dt + \sigma_r dW_t^r\left(\mathbb{P}\right) + d\left(\sum_{j=1}^{N_t^{\mathbb{P}}} Z_j^{\mathbb{P}}\right)$$

where $N_t^{\mathbb{P}}$ is a Poisson process with intensity $\lambda_r^{\mathbb{P}}$ and $Z_j^{\mathbb{P}} \sim N\left(\mu_J^{\mathbb{P}}, \left(\sigma_J^{\mathbb{P}}\right)^2\right)$. Assuming essentially affine risk premiums for the diffusive risks and constant risk premiums for the risks associated with the timing and sizes of the jumps, the risk neutral evolution of r_t is

$$dr_t = \left(a_r^{\mathbb{Q}} - b_r^{\mathbb{Q}} r_t\right) dt + \sigma_r dW_t^r(\mathbb{Q}) + d\left(\sum_{j=1}^{N_t^{\mathbb{Q}}} Z_j^{\mathbb{Q}}\right)$$

where $W_t^r(\mathbb{Q})$ is a standard Brownian motion, $N_t^{\mathbb{Q}}$ has intensity $\lambda^{\mathbb{Q}}$, and $Z_j^{\mathbb{Q}} \sim N\left(\mu_J^{\mathbb{Q}}, \left(\sigma_J^{\mathbb{Q}}\right)^2\right)$, all of which are defined on $\mathbb{Q}$.[16] Jumps affect all of the risk-neutral moments and provide three additional parameters for matching term structure shapes.

[16] These are, of course, very restrictive assumptions on the market prices of risk, especially for the jump components. Provided we allow for affine dynamics under $\mathbb{Q}$, we could specify any dynamics under $\mathbb{P}$. This could include nonlinear drifts, state dependent jump intensities or state dependent jump distributions.

This model delivers exponential affine bond prices

$$P\left(r_t,\tau\right)=E_t^{\mathbb{Q}}\left[e^{-\int_t^{t+\tau}r_s ds}|r_t\right]=\exp\left(\beta\left(\Theta^{\mathbb{Q}},\sigma_r,\tau\right)+\beta^r\left(b_r^{\mathbb{Q}},\tau\right)r_t\right)$$

where $\Theta^{\mathbb{Q}}=\left(a_r^{\mathbb{Q}},b_r^{\mathbb{Q}},\lambda^{\mathbb{Q}},\mu_J^{\mathbb{Q}},\sigma_J^{\mathbb{Q}}\right)$ and the loading functions solve the system of ODEs:

$$\begin{aligned}\frac{\beta^r\left(b_r^{\mathbb{Q}},\tau\right)}{d\tau} &= 1+\beta^r\left(b_r^{\mathbb{Q}},\tau\right)b_r^{\mathbb{Q}}\\ \frac{d\beta\left(\Theta^{\mathbb{Q}},\sigma_r,\tau\right)}{d\tau} &= \beta^r a_r^{\mathbb{Q}}+\frac{1}{2}\left(\sigma_r\beta^r\right)^2+\lambda^{\mathbb{Q}}\left[\exp\left(\beta^r\mu_J^{\mathbb{Q}}+\frac{1}{2}\left(\beta^r\sigma_J^{\mathbb{Q}}\right)^2\right)-1\right]\end{aligned}$$

subject to $\beta^r(b_r^{\mathbb{Q}},0)=\beta\left(\Theta^{\mathbb{Q}},\sigma_r,0\right)=0$. We have suppressed the dependence of $\beta^r\left(b_r^{\mathbb{Q}},\tau\right)$ on $b_r^{\mathbb{Q}}$ and τ for notational simplicity on the right hand-side of the second ODE. It is not possible to analytically solve these ordinary differential equations, although it is straightforward to solve them numerically.

Time-discretizing the model gives:

$$\begin{aligned}Y_{t,\tau} &= \beta\left(\Theta^{\mathbb{Q}},\sigma_r,\tau\right)+\beta^r\left(b_r^{\mathbb{Q}},\tau\right)r_t+\varepsilon_t\\ r_{t+1} &= r_t+a_r^{\mathbb{P}}-b_r^{\mathbb{P}}r_t+\sigma_r\varepsilon_{t+1}^r+Z_{t+1}^{\mathbb{P}}J_{t+1}^{\mathbb{P}}\end{aligned}$$

$\varepsilon_t^r\sim\mathcal{N}\left(0,1\right)$, $\varepsilon_t\sim\mathcal{N}\left(0,\Sigma_\varepsilon\right)$ is the vector of pricing errors, $Z_t^{\mathbb{P}}\sim N\left(\mu_J^{\mathbb{P}},\left(\sigma_J^{\mathbb{P}}\right)^2\right)$, and $J_t=1$ with $\mathbb{P}-$probability $\lambda^{\mathbb{P}}$. With jumps, it is important to explicitly address the issue of potential biases arising from the time-discretization. Empirical estimates indicate that jumps in interest rates occur more often than jumps in equity returns (Johannes (2003)). This implies that a simple Bernoulli approximation to the Poisson process could be inaccurate and stresses the importance of using high-frequency data. For the U.S. Treasury market, daily data is available for the past twenty years and reliable daily LIBOR/swap data is available since 1990.

The posterior distribution is defined over the objective measure jump parameters, $\Theta_J^{\mathbb{P}}=\left(\lambda_r^{\mathbb{P}},\mu_J^{\mathbb{P}},\sigma_J^{\mathbb{P}}\right)$, objective measure diffusive parameters, $\Theta_D^{\mathbb{P}}=\left(a_r^{\mathbb{P}},b_r^{\mathbb{P}},\sigma_r\right)$, the risk-neutral parameters, $\Theta^{\mathbb{Q}}$, and the latent state variables, (r,Z,J) where r, Z, and J are vectors containing the spot rates, jump sizes and jump times. For $\Theta^{\mathbb{P}}$ we assume conjugate priors: $a_r^{\mathbb{P}}$, $b_r^{\mathbb{P}}$, and $\mu_J^{\mathbb{P}}$ are normal, $\sigma_J^{\mathbb{P}}$ and σ_r are inverted Gamma and $\lambda_r^{\mathbb{P}}$ is Beta. For simplicity, we assume the same functional forms for the corresponding risk-neutral parameters. Clifford-

Hammersley implies that

$$
\begin{aligned}
&p\left(\Sigma_\varepsilon, \Theta_D^{\mathbb{P}} | \Theta_J^{\mathbb{P}}, \Theta^{\mathbb{Q}}, r, Z, J, Y\right) \\
&p\left(\Theta_J^{\mathbb{P}} | \Theta_D^{\mathbb{P}}, \Theta^{\mathbb{Q}}, r, Z, J, Y\right) \\
&p\left(\Theta^{\mathbb{Q}} | \Theta_D^{\mathbb{P}}, \Theta_J^{\mathbb{P}}, r, Z, J, Y\right) \\
&p\left(r | \Theta_D^{\mathbb{P}}, \Theta_J^{\mathbb{P}}, \Theta^{\mathbb{Q}}, Z, J, Y\right) \\
&p\left(Z | \Theta_D^{\mathbb{P}}, \Theta_J^{\mathbb{P}}, \Theta^{\mathbb{Q}}, r, J, Y\right) \\
&p\left(J | \Theta_D^{\mathbb{P}}, \Theta_J^{\mathbb{P}}, \Theta^{\mathbb{Q}}, r, Z, Y\right)
\end{aligned}
$$

are complete conditionals. We discuss each of these in turn.

The first portion of the MCMC algorithm updates $\Sigma_\varepsilon, \Theta_D^{\mathbb{P}}$ and $\Theta_J^{\mathbb{P}}$. We factor these distributions further via Clifford-Hammersley and sequentially draw $\left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}\right)$, σ_r, Σ_ε, $\lambda_r^{\mathbb{P}}$, and $\left(\mu_J^{\mathbb{P}}, \sigma_J^{\mathbb{P}}\right)$. Since $a_r^{\mathbb{P}}, b_r^{\mathbb{P}}$, and $\Theta_J^{\mathbb{P}}$ do not appear in the ODE's, their conditional posteriors simplify since, conditional on r, the parameter posteriors are independent of Y. To update $\Theta_D^{\mathbb{P}}$ and $\Theta_J^{\mathbb{P}}$, we sequentially draw

$$
\begin{aligned}
p\left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}} | \sigma_r, r, Z, J\right) &\sim \mathcal{N} \\
p\left(\lambda_r^{\mathbb{P}} | J\right) &\sim \mathcal{B} \\
p\left(\Sigma_\varepsilon | \sigma_r, \Theta^{\mathbb{Q}}, r, Z, J, Y\right) &\sim \mathcal{IW} \\
p\left(\mu_J^{\mathbb{P}}, \sigma_J^{\mathbb{P}} | \Theta_D^{\mathbb{P}}, Z, J\right) &\sim \mathcal{N} - \mathcal{IW} \\
p\left(\sigma_r | a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, \Theta^{\mathbb{Q}}, \Sigma_\varepsilon, r, Z, J, Y\right) &\sim \textit{Metropolis}.
\end{aligned}
$$

The updates for $a_r^{\mathbb{P}}$, $b_r^{\mathbb{P}}$ are standard as, conditional on the jump times and sizes, they enter as regression parameters:

$$
r_{t+1} - r_t - Z_{t+1}^{\mathbb{P}} J_{t+1}^{\mathbb{P}} = a_r^{\mathbb{P}} - b_r^{\mathbb{P}} r_t + \sigma_r \varepsilon_{t+1}^r.
$$

The conditional posteriors for the jump intensity and parameters of the jump size distribution are similar to those in Section 5.1.3. The conditional posterior for σ_r is the same as in the previous section, as this parameter appears in the bond prices and the structural evolution of spot rates.

Next, we update the risk premium parameters. $a_r^{\mathbb{Q}}$ and $\lambda^{\mathbb{Q}}$ enter linearly and, at least for, $a_r^{\mathbb{Q}}$, a normal prior generates a normal conditional posterior. For $\lambda^{\mathbb{Q}}$, we need to impose positivity, and thus we assume a Beta prior. This has the flexibility to impose that jumps

are rare under $\mathbb{Q}$. For $\lambda^{\mathbb{Q}}$ and the other risk premium these parameters, we use a random walk Metropolis algorithm. This implies the following steps:

$$\begin{aligned} p\left(a_r^{\mathbb{Q}}|\sigma_r, b_r^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \mu_J^{\mathbb{Q}}, \sigma_J^{\mathbb{Q}}, \Sigma_\varepsilon, r, Y\right) &\sim \mathcal{N} \\ p\left(b_r^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \mu_J^{\mathbb{Q}}, \sigma_J^{\mathbb{Q}}|a_r^{\mathbb{Q}}, \sigma_r, \Sigma_\varepsilon, r, Y\right) &: \quad RW\ Metropolis. \end{aligned}$$

The final stage updates the state variables. In this model, we are able to draw each vector of state variables in blocks

$$\begin{aligned} p\left(r|\Theta_D^{\mathbb{P}}, \Theta_J^{\mathbb{P}}, \Theta^{\mathbb{Q}}, \Sigma_\varepsilon, Z, J, Y\right) &: \quad FFBS \\ p\left(Z|\Theta_D^{\mathbb{P}}, \Theta_J^{\mathbb{P}}, r, J\right) &\sim \mathcal{N} \\ p\left(J|\Theta_D^{\mathbb{P}}, \Theta_J^{\mathbb{P}}, r, Z, \right) &\sim Bernoulli. \end{aligned}$$

The FFBS update for spot rates follows directly from the fact that, conditional parameters, jump times and jump sizes, the model is a linear, Gaussian state space model. The updates from J and Z are, conditional on r, the same as in Section 5.1.3.

The jump-diffusion model can be easily generalized to multiple dimensions. Assume that $r_t = \alpha + \alpha_x' X_t$ where the $N-$vector of state variables, X_t, solves

$$dX_t = \left(A_x^{\mathbb{P}} - B_x^{\mathbb{P}} X_t\right) dt + \sigma_x dW_t^r\left(\mathbb{P}\right) + d\left(\sum_{j=1}^{N_t^{\mathbb{P}}} Z_j^{\mathbb{P}}\right),$$

where $A_x^{\mathbb{P}} \in R^N$, $\sigma_x, B_x^{\mathbb{P}} \in R^{N\times N}$, $\Sigma_x^{\mathbb{P}} = \sigma_x'\sigma_x$, $W_t^r\left(\mathbb{P}\right)$ is a $N-$dimensional standard Brownian motion, and $Z_j^{\mathbb{P}} \in R^N \sim N\left(\mu_J^{\mathbb{P}}, \Sigma_J^{\mathbb{P}}\right)$, and for simplicity we assume that the jump arrivals are coincident across the state variables. In the case of no jumps, this models takes the form of a multi-factor Gaussian model.

It is common to assume there are three states, which are typically identified as the short rate, the slope of the curve, and the curvature. This would allow for correlated jumps and generates some potentially interesting issues. For example, while most would agree that the short rate jumps (as it is related to monetary policy), but if one finds that the slope factor also jumps, does this imply that the FED influences the long end of the yield curve? These and other issues can be addressed in a multivariate jump model.

5.2.3 The CIR model

Gaussian models have three potentially problematic assumptions: (1) interest rate volatility is constant, (2) interest rate increments are normally distributed, and (3) the spot rate can

be negative. As r_t is typically assumed to be a nominal rate, this is an unattractive feature. The classic square-root model of Cox, Ingersoll and Ross (1985) corrects these shortcomings. CIR assume the spot rate follows a Feller (1951) square root process

$$dr_t = \left(a_r^{\mathbb{P}} - b_r^{\mathbb{P}} r_t\right) dt + \sigma_r \sqrt{r_t} dW_t^r(\mathbb{P})$$

where W_t^r is a Brownian motion under the objective measure, $\mathbb{P}$. As r_t falls to zero, $\sqrt{r_t}$ falls to zero, effectively turning off the randomness in the model. If $b_r^{\mathbb{P}} > 0$ and together the parameters satisfy the Feller condition, the drift will pull r_t up from low rates. Under regularity, this model generates a form of time-varying volatility, (slightly) non-normal increments and positive interest rates.

Assuming essentially affine risk premiums, the evolution under $\mathbb{Q}$ is

$$dr_t = \left(a_r^{\mathbb{Q}} - b_r^{\mathbb{Q}} r_t\right) dt + \sigma_r \sqrt{r_t} dW_t^r(\mathbb{Q})$$

and the price of a zero coupon bond maturing at time τ_i is

$$P\left(r_t, \tau\right) = \exp\left(\beta\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \tau\right) + \beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right) r_t\right),$$

where again we label $\Theta^{\mathbb{Q}} = \left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}\right)$ and $\Theta^{\mathbb{P}} = \left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, \sigma_r\right)$. The loading functions are given by:

$$\begin{aligned}
\beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right) &= \frac{2\left(1 - \exp\left(\gamma\tau\right)\right)}{\left(\gamma + b_r^{\mathbb{Q}}\right)\left(\exp\left(\gamma\tau\right) - 1\right) + 2\gamma} \\
\beta\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \tau\right) &= \frac{a_r^{\mathbb{Q}}}{\sigma_r^2}\left[2\ln\left(\frac{2\gamma}{\left(b_r^{\mathbb{Q}} + \gamma\right)\left(\exp\left(\gamma\tau\right) - 1\right) + 2\gamma}\right) + \left(b_r^{\mathbb{Q}} + \gamma\right)\tau\right]
\end{aligned}$$

where $\gamma = \left[\left(b_r^{\mathbb{Q}}\right)^2 + 2\sigma_r^2\right]^{1/2}$.

Given the usual observed panel of yields, and assuming a time-discretization[17] of the interest rate increments, the state space is given by:

$$\begin{aligned}
Y_{t,\tau} &= \beta\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \tau\right) + \beta^r\left(b_r^{\mathbb{Q}}, \sigma_r, \tau\right) r_t + \varepsilon_t \\
r_{t+1} &= r_t + a_r^{\mathbb{P}} + b_r^{\mathbb{P}} r_t + \sigma_r \sqrt{r_t} \varepsilon_{t+1}^r.
\end{aligned}$$

[17] As in the Vasicek model, the exact transitions of the of the interest rate are known and are given by

$$p\left(r_{t+1} | r_t\right) \propto e^{-u-v}\left(\frac{u}{v}\right)^{\frac{q}{2}} I_q\left(2\left(uv\right)^{1/2}\right)$$

where $u = cr_t e^{-b_r}$, $v = cr_{t+1}$ and $c = \frac{2b_r}{\sigma_r^2\left(1 - e^{-b_r}\right)}$. Lamoureux and Witte (2001) discretize the state space and implement a "Griddy" Gibbs sampler. An attractive alternative to this would be to use a Metropolis algorithm to update the states.

The state space is still linear and conditionally Gaussian in the states, but the spot rate evolution has conditional heteroskedasticity.

The posterior distribution is given by $p\left(\Theta^{\mathbb{P}}, \Theta^{\mathbb{Q}}, r|Y\right)$ and the parameter component of the MCMC algorithm we consider is similar to the one in the previous section. For priors, we can choose, for example, $\left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}\right) \sim \mathcal{N}$, $\sigma_r \sim \mathcal{IG}$, $\Sigma_\varepsilon \sim \mathcal{IW}$, $\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}\right) \sim \mathcal{N}$. The MCMC algorithm consists of the following steps:

$$\begin{aligned}
p\left(a_r^{\mathbb{P}}, b_r^{\mathbb{P}}|\sigma_r, r\right) &\sim \mathcal{N} \\
p\left(\Sigma_\varepsilon|a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, r, Y\right) &\sim \mathcal{IW} \\
p\left(a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}|\sigma_r, \Sigma_\varepsilon, r, Y\right) &: \; RW\ Metropolis \\
p\left(\sigma_r^2|a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \Sigma_\varepsilon, r, Y\right) &: \; Metropolis \\
p\left(r|a_r^{\mathbb{P}}, b_r^{\mathbb{P}}, a_r^{\mathbb{Q}}, b_r^{\mathbb{Q}}, \sigma_r, \Sigma_\varepsilon, Y\right) &: \; Metropolis.
\end{aligned}$$

All of these are familiar from the previous sections, with the exception of r. Since the spot rates appear in the conditional variance of the spot rate evolution, the Kalman filter and thus the FFBS algorithm does not apply. To update the spot rates, independence or random walk is required.

It is straightforward to extend this algorithm to multi-factor square-root models. Lamoureaux and Whitte (2002) consider a two-factor square-root model and use an alternative approach based on the Griddy-Gibbs sampler for all of the parameters and the state variables. This avoids discretization bias, but is extremely computationally demanding. Polson, Stroud, and Muller (2001) analyze a square-root stochastic volatility model using Treasury rates. Bester (2003) analyzes multi-factor affine and string models using MCMC.

5.3 Regime Switching Models

We first considered the Black-Scholes model, a model with a constant expected return and volatility. In the sections that followed, we considered models that relaxed this constant parameter specification, allowing the expected return and volatility to vary over time. In those models, expected returns or volatility were modeled as diffusions or jump-diffusions, where the jump component was i.i.d. In this section, we consider an alternative: the drift and diffusion are driven by a continuous-time, discrete state Markov Chain. The models are commonly called regime-switching models, Markov switching models or Markov modulated diffusions.

The general form of the model is

$$dS_t = \mu(\Theta, X_t, S_t)\, dt + \sigma(\Theta, X_t, S_t)\, dW_t$$

where X_t takes values in a discrete space $X_t = x_1, ..., x_k$ with transition matrix $P_{ij}(t)$. $\Theta = (\Theta_1, ..., \Theta_J)$. Intuitively, if the process is in state i, the process solves

$$dS_t = \mu(\Theta, i, S_t)\, dt + \sigma(\Theta, i, S_t)\, dW_t.$$

Common specifications assume the drift and diffusion coefficients are parametric functions and the parameters switch over time. In this case, it is common to write the model as

$$dS_t = \mu(\Theta_{X_t}, S_t)\, dt + \sigma(\Theta_{X_t}, S_t)\, dW_t.$$

Term structure models with regime-switches are analyzed in Lee and Naik (1994), Landen (2000), Dai and Singleton (2002), Ang and Bekaert (2000), and Gray (1996). For example, an regime-switching extension of the Vasicek model assumes that the long run mean and the volatility can switch over time:

$$dr_t = \kappa_r(\theta_{X_t} - r_t)\, dt + \sigma_{X_t} dB_t$$

There has been an enormous amount of theoretical and practical work on regime-switching models using MCMC methods. For example, see the monograph by Kim and Nelson (2002) and the earlier papers by Carlin and Polson (1992) and Chib (1996, 1998). We provide a general algorithm, based on Scott (2002) who adapts the FFBS algorithm to the case of regime-switching models. Time discretized, we consider the following model:

$$S_t = \mu(\Theta_{X_t}, S_{t-1}) + \sigma(\Theta_{X_t}, S_{t-1})\, \varepsilon_t.$$

Note that we use the standard notation from discrete-time models where the time index on the Markov state is equal to the current observation. The discrete-time transition probabilities are

$$P_{ij} = P(X_t = i | X_{t-1} = j)$$

and we assume, apriori, that the transition functions are time and state invariant. The joint likelihood is given by

$$p(S|X, \Theta) = \prod_{t=1}^{T} p(S_t | S_{t-1}, X_{t-1}, \Theta)$$

where $p\left(S_t|S_{t-1}, X_{t-1}, \Theta\right) = N\left(\mu\left(\Theta_{X_{t-1}}, S_{t-1}\right), \sigma^2\left(\Theta_{X_{t-1}}, S_{t-1}\right)\right)$.

Clifford-Hammersley implies that the complete conditionals are given by $p\left(\Theta|X, S, P\right)$, $p\left(P|X, S, \Theta\right)$, and $p\left(X|P, S, \Theta\right)$. We do not directly address the first step. Conditional on the states and the transition probabilities, updating the parameters is straightforward. Conditional on the states, the transition matrix has a Dirchlet distribution, and updating this is also straightforward. To update the states, define the following quantities

$$\begin{aligned}\pi^t\left(X_t = i|\Theta\right) &= p\left(X_t = i|\Theta, S_{1:t}\right) \\ \widetilde{\pi}^t\left(X_t = i|\Theta\right) &= p\left(X_t = i|\Theta, S_{1:T}\right).\end{aligned}$$

The first distribution is the filtering distribution of the states and the second the smoothing distribution of the states. The updating algorithm is a discrete probability modification of the FFBS algorithm. We first forward filter the states, given the forward filtering distribution, we backward sample. The forward matrices are given by: $P^1, ..., P^T$, where $P^t = \left(P^t_{ij}\right)$ is $P^t_{ij} = p\left(X_{t-1} = i, X_t = j|\Theta, S_{1:t}\right)$. To compute the forward matrix, note the recursive structure of the filtering density:

$$\begin{aligned}P^t_{ij} &\propto p\left(S_t, X_{t-1} = i, X_t = j|\Theta, S_{1:t-1}\right) \\ &\propto p\left(S_t|X_t = j, \Theta\right) p\left(X_t = j|X_{t-1} = i, \Theta\right) \pi^{t-1}\left(X_{t-1} = i|\Theta\right).\end{aligned}$$

This provides a recursive solution for the forward matrices, P^t_{ij} for $t = 1, ..., T$. This is similar to the Kalman filter in Gaussian state space models. Next, we iterate backward in time by finding

$$\widetilde{P}^t_{ij} = p\left(X_{t-1} = i, X_t = j|\Theta, S_{1:T}\right).$$

The formula for the backward matrices is:

$$\widetilde{P}^t_{ij} \propto P^t_{ij} \frac{\widetilde{\pi}^t\left(X_t = j|\Theta\right)}{\pi^t\left(X_t = j|\Theta\right)}$$

which again is computed backward in time for $t = T, .., 1$.

An important component of regime switching models is the prior distribution. Regime switching models (and most mixture models) are not formally identified. For example, in all regime switching models, there is a labeling problem: there is no unique way to identify the states. A common approach to overcome this identification issue is to order the parameters.

6 Sequential Inference: Filtering

We now turn to the issue of filtering: estimating latent state based on contemporaneously available data. Throughout, we assume the parameters are known and we discuss various approaches to relaxing this assumption. Of primary interest are the filtering distributions, $p(X_t|Y^t)$, as a function of t, and the forecasting distribution, $p(X_{t+1}|Y^t)$. These are closely related to the likelihood $p(Y_t|X_t)$ and the state transition, $p(X_{t+1}|X_t)$.

We assume that prices are observed at a fixed observation frequency, for simplicity normalized to unit length. Thus the researcher observes $S_{1:t} = (S_1, ..., S_t)$. From this data, the goal is to estimate the latent variables, which we denote X_t. To understand the issues involved in the filtering problem, we write the state space model in its integrated form:

$$S_{t+1} = S_t + \int_t^{t+1} \mu^s(S_s, F_s)\, ds + \int_t^{t+1} \sigma^s(S_{s-}, F_{s-})\, dW_s^p + \sum_{t<\tau_n\leq t+1} Z_n^s \tag{34}$$

$$F_{t+1} = F_t + \int_t^{t+1} \mu^f(F_s)\, ds + \int_t^{t+1} \sigma^f(F_{s-})\, dW_s^x + \sum_{t<\tau_n\leq t+1} Z_n^f. \tag{35}$$

Our goal is to solve the filtering problem, the computation of the sequence of conditional densities, $p(X_t|S_{1:t})$, for $t = 1, ...T$. Again, X_t contains the factor states, jump times and jump sizes. This density can be represented via Bayes rule as:

$$p(X_{t+1}|S_{1:t+1}) = \frac{p(S_{t+1}|X_{t+1}, S_t)\, p(X_{t+1}|S_{1:t})}{p(S_{t+1}|S_{1:t})} \tag{36}$$

where we label $p(S_{t+1}|X_{t+1}, S_t)$ as the likelihood (the distribution of observed prices conditional on latent states and past prices), $p(X_{t+1}|S_{1:t})$, the predictive distribution of latent states, and $p(S_{t+1}|S_{1:t})$, the predictive distribution of prices. If all of these densities could be computed, then the filtering problem could be solved. Unfortunately, computing these densities is difficult as none are known analytically, and they are difficult to compute with brute force. For example,

$$p(X_{t+1}|S_{1:t}) = \int p(X_{t+1}|X_t)\, p(X_t|S_{1:t})\, dX_t$$

is a high dimension integration problem that cannot be solved analytically.

The problem of filtering continuous-time models with discrete-time observations must address two separate issues. First, suppose that we could directly evaluate the likelihood $p(S_{t+1}|X_{t+1}, S_t)$ and the state evolution $p(X_{t+1}|X_t)$. In this case, the filtering problem

becomes an issue of how to update from one period's filtering density, $p(X_t|S_{1:t})$, to the next period's, $p(X_{t+1}|S_{1:t+1})$. This problem is essentially a problem of estimating the integral, $\int p(X_{t+1}|X_t)\, p(X_t|S_{1:t})\, dX_t$ and is solved by a discretization of the filtering density or by directly using Monte Carlo methods. Second, it is not possible to numerically evaluate the likelihood, $p(S_{t+1}|X_{t+1}, S_t)$, and how to simulate the state evolution $p(X_{t+1}|X_t)$ when the prices and latent variables arise from continuous-time models. To solve this step, one typically time-discretizes the model and simulates the state variables. Both of these steps can be computationally intensive, so it is important to develop methods that are both accurate and computationally efficient.

We now describe two Monte Carlo approaches for filtering and sequential parameter estimation: the particle filter and the practical filter.

6.1 The Particle Filter

A state space model is built from the likelihood function, $p(S_{t+1}|X_{t+1}, S_t)$, and the state evolution, $p(X_{t+1}|X_t)$. The particle filter requires only two assumptions:

(A1) That the state evolution can be exactly simulated, that is, one can obtain a random draw from the distribution $p(X_{t+1}|X_t)$,

(A2) That the likelihood can be exactly evaluated as a function of X_t and S_t. That is, given (X_{t+1}, S_t), one can functionally evaluate $(X_{t+1}, S_t) \mapsto p(S_{t+1}|X_{t+1}, S_t)$.

Under these assumptions, the particle filter delivers an estimate of the true filtering density, $p(X_t|S_{1:t})$, which converges (as the number of particles increases) to the true filtering density. We now describe the mechanics of the particle filtering algorithm. The particle filter was developed in Gordon, Salmond, and Smith (1993) or Kitagawa (1994, 1996). For an overview of particle filtering methods, see edited volume by Doucet, deFreitas and Gordon (2001). Pitt and Shephard (2000) also provide a short overview, describe some common problems when applying the particle filter, and offer an extension that is important for practical applications.

The particle filter approximates the filtering density by a discrete probability distribution. The distribution $p(X_t|S_{1:t})$ is approximated by a set of particles, $\left\{X_t^{(i)}\right\}_{i=1}^{N}$ with

probability weights $\left\{\pi_t^{(i)}\right\}_{i=1}^N$, namely

$$p^N\left(X_t|S^t\right) = \sum_{i=1}^N \delta_{X_t^{(i)}} \pi_t^i.$$

Here p^N refers to an estimated density with N particles and δ is the Dirac function. Once the distribution is discretized, integrals become sums and estimates of the filtering and predictive densities are

$$p^N\left(X_{t+1}|S^t\right) = \sum_{i=1}^N p\left(X_{t+1}|X_t^{(i)}\right)\pi_t^i \approx \int p\left(X_{t+1}|X_t\right) p\left(X_t|S^t\right) dX_t.$$

When combined with the conditional likelihood, the filtering density at time $t+1$ is defined via the recursion:

$$p^N\left(X_{t+1}|S^{t+1}\right) \propto p\left(S_{t+1}|X_{t+1}, S_t\right) \sum_{i=1}^N p\left(X_{t+1}|X_t^{(i)}\right)\pi_t^i.$$

This recursion is just Bayes rule and show how to mechanically translate a particle representation of $p^N\left(X_t|S_{1:t}\right)$ into $p^N\left(X_{t+1}|S^{t+1}\right)$. However, the key to the particle filtering algorithm is a mechanism to generate samples from $p^N\left(X_{t+1}|S_{1:t+1}\right)$ in an efficient and accurate manner. That is, the algorithm requires a method to propagate the old particles and probabilities $\left\{X_t^{(i)}, \pi_t^{(i)}\right\}_{i=1}^N$ into new particles and probabilities $\left\{X_t^{(i)}, \pi_t^{(i)}\right\}_{i=1}^N$. Naive methods of generating samples often lead to degeneracies: the algorithm places all the probabilities on a few particles and gets stuck.

There are a number of different approaches that can be used to update or propagate the particles: the weighted bootstrap (also known as the sampling/importance sampling (SIR) algorithm), rejection sampling and MCMC. Sampling resampling takes X_t^i and simulates the state forward using the transition kernel, $p\left(X_{t+1}|X_t^i\right)$, to get X_{t+1}^i, and the reweights these samples according to $p\left(Y_{t+1}|X_{t+1}^i\right)$. Rejection sampling takes a draw from $\sum_{i=1}^N \pi_t^i p\left(X_{t+1}|X_t^i\right)$ and then accepts it with probability proportional to $p\left(Y_{t+1}|X_{t+1}^i\right)$. This requires that $p\left(Y_{t+1}|X_{t+1}\right)$ is bounded as a function of X_{t+1}. Finally, one can also use the independence Metropolis algorithm. In this case, one proposes from $\sum_{i=1}^N \pi_t^i p\left(X_{t+1}|X_t^i\right)$ and then accepts/rejects based on $p\left(Y_{t+1}|X_{t+1}\right)$. Doucet, deFreitas and Gordon (2001) provide detailed descriptions of these approaches and examples of how the different approaches perform in different settings. We describe the weighted bootstrap/SIR algorithm (Smith

and Gelfand (1992), Gordon, Salmond and Smith (1993)) for its generality and simplicity. A variant known as the auxiliary particle filter (Pitt and Shephard (1999)) often provides large efficiency gains and is popular for applications.

To understand the mechanics of the weighted bootstrap, we can view $p^N\left(X_{t+1}|S_{1:t}\right)$ as the prior and $p\left(S_{t+1}|X_{t+1},S_t\right)$ as the likelihood, and by Bayes rule the updated distribution is given by

$$p^N\left(X_{t+1}|S^{t+1}\right) \propto \underbrace{p\left(S_{t+1}|X_{t+1},S_t\right)}_{Likelihood}\underbrace{\sum_{i=1}^N p\left(X_{t+1}|X_t^{(i)}\right)\pi_t^i}_{\text{Prior}}.$$

Generating the next state is straightforward as we assumed the state transition can be exactly sampled. The first step is to simulate $X_{t+1}^{(i)}$ from the latent state evolution by drawing from the distribution $p\left(X_{t+1}|X_t^{(i)}\right)$. Given the updated states, $\left\{X_{t+1}^{(i)}\right\}_{i=1}^N$, the weighted bootstrap then re-samples these states $\left\{X_{t+1}^{(i)}\right\}_{i=1}^N$ with weights $\pi_{t+1}^i \propto p\left(S_{t+1}|X_{t+1}^{(i)},S_t\right)$. For clarity, we state the algorithm in steps:

1. Given $\left\{X_t^{(i)},\pi_t^{(i)}\right\}_{i=1}^N$, simulate the latent state vector forward using the transition equation. That is, draw $X_{t+1}^{(i)} \sim p\left(X_{t+1}|X_t^{(i)}\right)$.

2. Evaluate the likelihood function at the new state, $p\left(S_{t+1}|X_{t+1}^{(i)}\right)$, and set

$$\pi_{t+1}^{(i)} = \frac{p\left(S_{t+1}|X_{t+1}^{(i)}\right)}{\sum_{i=1}^N p\left(S_{t+1}|X_{t+1}^{(i)}\right)}.$$

3. Finally, resample N particles with replacement from the multinomial distribution of $\left\{X_{t+1}^{(i)}\right\}_{i=1}^N$ with probabilities $\left\{\pi_{t+1}^{(i)}\right\}_{i=1}^N$.

Gordon, Salmond, and Smith (1993), using an argument from Smith and Gelfand (1992), show that these resampled draws, $\left\{X_{t+1}^{(i)}\right\}_{i=1}^N$ with probabilities $\left\{\pi_{t+1}^{(i)}\right\}_{i=1}^N$ provide a sample from the approximated filtering density, $p^N\left(X_{t+1}|S_{1:t+1}\right)$. An advantage of the weighted bootstrap is that by sampling with replacement according to the probabilities $\pi_{t+1}^{(i)}$, the procedure selectively and over time eliminates states with very low probability by disproportionately resampling states with higher probability. Thus the algorithm propagates high likelihood states forward while discarding low likelihood states.

The particle filter has other advantages which include its computational efficiency, ease of implementation and modular nature. Moreover, there are a number of approaches which exist to improve on the performance of the naive particle filter. One approach, the auxiliary particle filter, is a straightforward extension of the particle filter and is described in Pitt and Shephard (1999). The auxiliary particle filter "peaks" forward via an initial resampling step and then propagates these higher likelihood samples forward with the particle filter. We use this approach to improve the algorithm's performance.

There are a number of applications of particle filtering using discrete and continuous-time finance models. Johannes, Polson and Stroud (2002) provide a general particle filtering approach for multivariate jump−diffusion models. The problem with continuous-time models is that $p(X_{t+1}|X_t)$ can rarely be directly sampled. Due to this, Johannes, Polson and Stroud (2002) follow Pedersen (1995), Elerian, Shephard and Chib (2001) and Eraker (2001), and simulates a number of additional time steps in between the observed data points. They show that particle filtering can handle combinations of factors like jumps in returns, stochastic volatility and jumps in volatility. Particle filtering has also been applied in a number of discrete-time models (Chib, Nardari and Shephard (2001) and Kim and Shephard (1998)) and also in diffusion models to construct the likelihood function (Durham and Gallant (2001), Pitt (2002)). Particle filtering can also, in certain cases, be extended to deal with the issue of sequential parameter learning, see, for example, Storvik (2002). However, particle filters in the presence of parameter uncertainty often degenerate.

6.1.1 Adapting the particle filter to continuous-time models

The previous section shows that the particle filter effectively imposes only two requirements on the state space model: (1) simulate the latent state variables forward and (2) evaluate the likelihood function as a function of the latent states and observables In continuous-time models, both of these are generally impossible to do without approximation. Johannes, Polson and Stroud (2003) describe how to do this in jump-diffusion models. We outline the basics of the algorithm here.

To implement the particle filter, we need to simulate the state variables forward and evaluate the likelihood function. To do this, we use time-discretized solutions to the stochastic differential equations. Assuming that prices are observed at times t and $t+1$, we simulate an additional $M-1$ points in between those observations via the Euler-type

discretization

$$S_{t+\frac{j+1}{M}} = S_{t+\frac{j}{M}} + \mu^s\left(S_{t+\frac{j}{M}}, F_{t+\frac{j}{M}}\right)M^{-1} + \sigma^s\left(S_{t+\frac{j}{M}}, F_{t+\frac{j}{M}}\right)\varepsilon^s_{t+\frac{j+1}{M}} + Z^s_{t+\frac{j+1}{M}}J^s_{t+\frac{j+1}{M}}$$
$$F^{(i)}_{t+\frac{j}{M}} = F^{(i)}_{t+\frac{j-1}{M}} + \mu^f\left(F_{t+\frac{j}{M}}\right)M^{-1} + \sigma^f\left(F_{t+\frac{j}{M}}\right)\varepsilon^f_{t+\frac{j+1}{M}} + Z^f_{t+\frac{j+1}{M}}J^f_{t+\frac{j+1}{M}}.$$

where $j = 1, ..., M-1$, ε^s_t and ε^x_t are mean zero, jointly normally distributed (potentially correlated) with common variance M^{-1}, J^s_t and J^f_t are Bernoulli random variables with respective intensities $\lambda^s\left(S_{t-M^{-1}}, F_{t_{t-M^{-1}}}\right)M^{-1}$ and $\lambda^f\left(F_{t-M^{-1}}\right)M^{-1}$. The jump size distribution remains unchanged.

While we use an Euler-type discretization for the jump-diffusion, there are other discretization schemes for the diffusion and jump components. For the diffusion components, higher-order discretization schemes are given in Kloeden and Platen (2003) and can all be used in our particle filtering approach. Typically these schemes require additional differentiability assumptions which commonly hold in applications. For the jump times, we use a Bernoulli discretization. If the jumps are Poisson (constant arrival intensity), the inter-arrival times can be exactly simulated which implies there is no discretization bias in the jump component. When the jump intensity is state dependent, the Bernoulli approximation is straightforward, but there are other algorithms available (see, e.g., Glasserman and Merener (2003)).

Given the time discretization, we define the following quantities: $F^M_{t+1} = \left(F_t, ..., F_{t+\frac{M-1}{M}}\right)$, $S^M_{t+1} = \left(S_{t+\frac{1}{M}}, ..., S_{t+\frac{M-1}{M}}\right)$, $Z^{k,M}_{t+1} = \left(Z^k_{t+\frac{1}{M}}, ..., Z^k_{t+1}\right)$, and $J^{k,M}_{t+1} = \left(J^k_{t+\frac{1}{M}}, ..., J^k_{t+1}\right)$ where $k = s, f$. The entire matrix of latent variables is

$$X^M_{t+1} = \left(F^M_{t+1}, S^M_{t+1}, Z^{s,M}_{t+1}, Z^{f,M}_{t+1}, J^{s,M}_{t+1}, J^{f,M}_{t+1}\right).$$

Not that the latent variables include augmented prices between observations in S^M_t and that the states are simulated up to one-discretization interval before the next observation. It is on this quantity that we define the particle filter.

Given the time-discretization, we can now modify the particle filter to handle the case of jump diffusions. Conditional on $\left\{\left(X^M_t\right)^{(i)}\right\}_{i=1}^N$, the first stage involves simulating the state variables forward. To do this, first generate jump times, jump sizes and discretized Brownian increments. All of these draws are straightforward as they are i.i.d. draws. Feed the jump times, jump sizes and Brownian increments through the Euler scheme to generate $\left(F^M_{t+1}\right)^{(i)}$ and $\left(S^M_{t+1}\right)^{(i)}$. This provides the propagated state vector, $\left\{\left(X^M_{t+1}\right)^{(i)}\right\}_{i=1}^N$. Finally, these updated states are resampled with the appropriate probabilities.

We now provide the details of the algorithm. For simplicity, we first describe it in the case of no derivative prices and then discuss how to adapt the procedure to deal with derivatives.

1. Given $\left\{\left(X_{t+1}^{M}\right)^{(i)}, \pi_{t+1}^{(i)}\right\}_{i=1}^{N}$, simulate the latent state vector and latent prices forward. This requires the following steps. For $j = 1, ..., M-1$, conditional on $S_{t+(j-1)M^{-1}}^{(i)}$ and $F_{t+(j-1)M^{-1}}^{(i)}$:

   (a) sample Brownian increments

$$\left(\left(\varepsilon_{t+jM^{-1}}^{s}\right)^{(i)}, \left(\varepsilon_{t+jM^{-1}}^{f}\right)^{(i)}\right) \sim N\left(0, M^{-1}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

   (b) sample jump sizes from their conditional distributions

$$\left(Z_{t+jM^{-1}}^{s}\right)^{(i)} \sim \Pi^{s}\left(S_{t+\frac{(j-1)}{M}}^{(i)}, X_{t+\frac{(j-1)}{M}}^{(i)}\right)$$
$$\left(Z_{t+jM^{-1}}^{f}\right)^{(i)} \sim \Pi^{f}\left(F_{t+\frac{(j-1)}{M}}^{(i)}\right);$$

   sample jump times

$$J_{t+\frac{j}{M}}^{s,(i)} \sim Ber\left(\lambda^{s}\left(S_{t+\frac{(j-1)}{M}}^{(i)}, X_{t+\frac{(j-1)}{M}}^{(i)}\right)M^{-1}\right)$$
$$J_{t+\frac{j}{M}}^{f,(i)} \sim Ber\left(\lambda^{f}\left(F_{t+\frac{(j-1)}{M}}^{(i)}\right)M^{-1}\right);$$

   (c) simulate states and prices forward:

$$S_{t+\frac{j}{M}}^{(i)} = S_{t+\frac{j-1}{M}}^{(i)} + \mu^{s}(S_{t+\frac{j-1}{M}}^{(i)}, X_{t+\frac{j-1}{M}}^{(i)})M^{-1} + \sigma^{s}(S_{t+\frac{j-1}{M}}^{(i)}, X_{t+\frac{j-1}{M}}^{(i)})\varepsilon_{t+\frac{j}{M}}^{s,(i)} + Z_{t+\frac{j}{M}}^{s,(i)} J_{t+\frac{j}{M}}^{s,(i)}$$

$$F_{t+\frac{j}{M}}^{(i)} = F_{t+\frac{j-1}{M}}^{(i)} + \mu^{x}\left(F_{t+\frac{j-1}{M}}^{(i)}\right)M^{-1} + \sigma^{x}\left(F_{t+\frac{j-1}{M}}^{(i)}\right)\varepsilon_{t+\frac{j}{M}}^{f,(i)} + Z_{t+\frac{j}{M}}^{f,(i)} J_{t+\frac{j}{M}}^{f,(i)}.$$

2. Collect the new simulated prices and states into

$$\left(X_{t+1}^{M}\right)^{(i)} = \left(\left(F_{t+1}^{M}\right)^{(i)}, \left(S_{t+1}^{M}\right)^{(i)}, \left(Z_{t+1}^{s,M}\right)^{(i)}, \left(Z_{t+1}^{f,M}\right)^{(i)}, \left(J_{t+1}^{s,M}\right)^{(i)}, \left(J_{t+1}^{f,M}\right)^{(i)}\right).$$

3. Evaluate the likelihood function at the new state, $p\left(S_{t+1}|\left(X_{t+1}^{M}\right)^{(i)}\right)$, and set

$$\pi_{t+1}^{(i)}=\frac{p\left(S_{t+1}|\left(X_{t+1}^{M}\right)^{(i)}\right)}{\sum_{i=1}^{N}p\left(S_{t+1}|\left(X_{t+1}^{M}\right)^{(i)}\right)}.$$

4. Finally, resample N particles with replacement from the multinomial distribution of $\left\{\left(X_{t+1}^{M}\right)^{(i)}\right\}_{i=1}^{N}$ with probabilities $\left\{\pi_{t+1}^{(i)}\right\}_{i=1}^{N}$.

Johannes, Polson and Stroud (2003) provide further details into these algorithms. They provide simulation evidence on the efficacy of the algorithm and provide examples using both returns and option price data.

6.2 Practical Filtering

In principle, MCMC could be directly applied to the problem of filtering and sequential parameter estimation. MCMC generates samples from $p\left(\Theta,X^{t}|Y^{t}\right)$. Integrating out the other variables provides estimates of the filtering density, $p\left(X_{t}|Y^{t}\right)$, and the parameters $p\left(\Theta|Y^{t}\right)$. Repeated application of an MCMC algorithm would provide sequential estimates of the parameters and states. Unfortunately, this brute force approach is computationally intractable.

Much like in the case of particle filtering, we can use the structure of the problem to develop an MCMC-based approximation to the sequential estimation problem. Johannes, Polson and Stroud (2002) and Polson, Stroud and Mueller (2002) develop an alternative to particle filtering known as practical filtering. The practical filter relies on the idea of fixed-lag filtering (see, e.g., Andersen and Moore (1978)). The fixed lag filter relies on the observation that, when estimating the current filtering distribution, today's observations provide little information about states in the distant past, beyond the information embedded in past returns. Mathematically, this implies that $p\left(X_{t-k}|Y^{t+1}\right)\approx p\left(X_{t-k}|Y^{t}\right)$, which is likely a reasonable approximation, especially for large k.

The advantage of this method is that it applies in the presence of unknown parameters and also avoids degeneracies associated with particle filters. The key to the fixed-lag filter is based on the following identity:

$$p\left(\Theta,X_{t+1}|Y^{t+1}\right)=\int p\left(\Theta,X_{t-k+1}^{t+1}|X_{t-k},Y^{t+1}\right)p\left(X_{t-k}|Y^{t+1}\right)dX_{t-k}^{t}$$

where $X_j^k = [X_j, ..., X_k]$. In the filtering recursion, there are samples from $p\left(X_{t-k}|Y^t\right)$ and then, for large enough k, $p\left(X_{t-k}|Y^{t+1}\right) \approx p\left(X_{t-k}|Y^t\right)$ as an additional observation at time $t+1$ has little impact on X_{t-k}. With these samples, we use MCMC methods to generate samples from $p\left(\Theta, X_{t-k+1}^{t+1}|X_{t-k}, Y^{t+1}\right)$ which provides samples from the $p\left(\Theta, X_{t+1}|Y^{t+1}\right)$. If one can efficiently sample from $p\left(\Theta, X_{t-k+1}^{t+1}|X_{t-k}, Y^{t+1}\right)$ the fixed-lag filter provides a computational attractive method for sequential parameter learning and state filtering.

In the filtering recursion, assume that we already have samples $\left\{X_{1:t-k}^{(g)}, \theta^{(g)}\right\}$ from the joint filtering and parameter distribution $p(\Theta, X_{1:t-k}|Y^t)$. Notice that by the Markov property we need only store the samples of $X_{t-k}^{(g)}$ in order to simulate from the next filtering distribution, using MCMC, namely $p(X_{t-k+1}^{t+1}|X_{t-k}, Y^{t+1})$. Moreover, if we assume that the addition of the next data data Y_{t+1} has little influence on the marginal of the lagged-filtering distribution $p(X_{t-k}|Y^{t+1})$ we can use the samples $X_{t-k}^{(g)} \sim p(X_{t-k}|Y^t$ from the previous iteration as approximate draws fro the next step and generate our next set of states $(X_{t-k+1}^{t+1})^{(g)}$ from $p(X_{t-k+1}^{t+1}|X_{t-k}^{(g)}, Y^{t+1})$.

For the sequential parameter updates we notice that many models can exploit a sufficient statistics structure and we only have to keep track of a set of sufficient statistics $t(X_{1:t})$. Hence we generate the next parameter draw $\theta^{(g)}|X_{t-k+1,t+1}, Y^{t+1} \sim p(\Theta|X_{1:t-k}^{(g)}, X_{t-k+1,t+1}, Y^{t+1})$ in an iterative fashion with the new state draw $(X_{t-k+1}^{t+1})^{(g)}$. In the case where there exists a set of sufficient statistics with reduces to a draw of $p(\Theta|t(X_{1:t-k}^{(g)}, X_{t-k+1,t+1}), Y^{t+1})$ and an update of $t(X_{1:t+1})$. Storvik (2002) shows how to use sufficient statistics in particle filtering. Johannes, Polson and Stroud (2003) provide a comparison of the performance of the particle and practical filter in a stochastic volatility model with jumps.

7 Conclusions and Future Directions

This chapter provides an overview of MCMC methods. We discussed the theoretical underpinnings of the algorithms and provided a tutorial on MCMC methods for a number of continuous-time asset pricing models. While MCMC methods have been used for a number of practical problems, we feel there are numerous additional applications in which MCMC methods will be useful. We now briefly outline a number of future directions.

In many problems, economic theory places constraints on parameter values. For example, pricing kernels must be non-negative to exclude arbitrage or equilibrium excess expected returns must be positive. Bayesian and MCMC methods are ideally suited to

handling these difficult problems, which can be intractable using classical methods. For example, the paper by Wang and Zhang (2003) shows how to use MCMC to characterize the Hansen-Jagannathan distance which imposes positivity on the pricing kernel. As the authors show, these constrains can have major implications regarding our ability to discriminate across models.

While a number of authors have analyzed term structure models with Gaussian or square-root factors using MCMC, there are a number of other areas that need to be analyzed. There is very little work of jump-diffusion term structure models, and MCMC methods are ideally suited to answering a number of interesting questions. Do multiple factors jump, or is it only the short rate? Does the market price diffusive and jump risks differently in the term structure? How do predictable jumps affect the term structure?

On a mechanical level, there are a number of issues that have not been resolved. First, in many stochastic volatility models (e.g., square-root models), MCMC algorithms update volatility in a single-state manner. While accurate, it would be preferable to have algorithms to update the volatilities in blocks. If efficient blocking routines were developed, MCMC algorithms would be less computationally intensive and allow a far wider range of models to be analyzed sequentially. Second, in term structure models, the observable yields are often take the form of par rates. These models are nonlinear in the states, but it should be possible to tailor MCMC algorithms to handle this specific form of nonlinearity. Third, there is little work on sequential inference. The filtering distribution of parameters and states is far more relevant than the smoothing distribution for financial applications, it is important to develop and test sequential algorithms.

8 References

Aït-Sahalia, Yacine, 2003, Disentangling Jumps from Volatility, forthcoming, *Journal of Financial Economics*.

Aldous, David, 1987, On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing, *Probability in Engineering Information Systems* 1, 33-46.

Aldous, David, 1989, *Probability Approximations via the Poisson Clumping Heuristic*, New York, Springer-Verlag.

Andersen, Torben., Bollerslev, Tim and Diebold, Frank, 2002, Parametric and Nonparametric Volatility Measurement, in Lars Peter Hansen and Yacine Ait-Sahalia (editors), *Handbook of Financial Econometrics*, Amsterdam: North-Holland, forthcoming.

Andersen, Torben, Hyung-Jin Chung, and Bent Sorensen, 1999, Efficient method of moments estimation of a stochastic volatility model, *Journal of Econometrics* 91, 61-87.

Anderson, Theodore, 1984, *An Introduction to Multivariate Statistical Analysis*, 2nd Edition, John Wiley & Sons.

Barndorff-Nielsen, Ole Barndorff-Nielson and Neil Shephard, 2003, Impact of jumps on returns and realised variances: econometric analysis of time-deformed Lévy processes, working paper, Oxford University.

Bates, David, 1996, Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options, *Review of Financial Studies* 9, 69-107.

Bates, David, 2000, Post-'87 Crash Fears in S&P 500 Futures Options, *Journal of Econometrics* 94, 181-238.

Bernardo, Jose and Adrian Smith, 1995, *Bayesian Theory*, Wiley, New York.

Besag, Julian, 1974, Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Association Series B* 36, 192-236.

Besag, Julian and Green, Peter, 1993, Spatial Statistics and Bayesian Computation (with discussion). *Journal of the Royal Statistical Association Series B* 55, 25-37.

Bester, Alan, 2003, Random Fields and Affine Models of Interest Rates, working paper, Duke University.

Black, Fischer and Myron S. Scholes, 1973, The pricing of options and corporate liabilities, *Journal of Political Economy* 81, 637-654.

Carlin, Bradley, and Sidhartha Chib, 1995, Bayesian Model Choice through Markov Chain Monte Carlo, *Journal of the Royal Statistical Association Series B, 57,* 473–484.

Carlin, Bradley, and Nicholas Polson, 1991). Inference for Nonconjugate Bayesian Models using the Gibbs sampler. *Canadian Journal of Statistics*, 19, 399-405.

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

Carlin, Bradley, and Nicholas Polson, and David Stoffer, 1992, A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling, *Journal of the American Statistical Association*, 87, 493-500.

Carpenter, James, Peter Clifford, and Paul Fearnhead, 1999, An Improved Particle Filter for Nonlinear Problems. *IEE Proceedings – Radar, Sonar and Navigation*, 1999, 146, 2–7.

Carter, C.K., and Robert Kohn, 1994, On Gibbs Sampling for State Space Models, *Biometrika*, 81, 541-553.

Chamberlain, Gary, 2001, Econometrics and decision theory, *Journal of Econometrics* 95, 255-283

Chib, Sidhartha, 1995, Marginal Likelihood From the Gibbs Output, *Journal of the American Statistical Association*, 90, 1313-1321.

Chib, Sidhartha, 1996, Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models, *Journal of Econometrics*, 75, 79-97.

Chib, Sidhartha, 1998, Estimation and Comparison of Multiple Change Point Models, *Journal of Econometrics* 86, 221-241.

Clifford, Peter, 1993, Discussion on the Meeting on the Gibbs Sampler and other Markov Chain Monte Carlo methods, *Journal of the Royal Statistical Society Series B, 55, 53-54.*

Collin-Dufresne, Pierre, Robert Goldstein and Chris Jones, 2003, Identification and Estimation of 'Maximal' Affine Term Structure Models: An Application to Stochastic Volatility, working paper, USC.

Cox, John, Jonathan Ingersoll and Stephen Ross, A Theory of the Term Structure of Interest Rates, *Econometrica* 53, 385-407.

Dai, Qiang and Kenneth Singleton, 2000, Specification Analysis of Affine Term Structure Models, *Journal of Finance* 55 1943-1978.

Dai, Qiang and Kenneth Singleton, 2003, Term Structure Modeling in Theory and Reality, *Review of Financial Studies* 16, 631-678.

DeGroot, Morris, 1970, *Optimal statistical decisions*, McGraw-Hill, New York.

Devroye, Luc, 1986, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

Doucet, Arnaud., Nando de Freitas, and Neil Gordon, 2001, Sequential Monte Carlo methods in practice. Springer, New York.

Duarte, Jefferson, 2003, Evaluating An Alternative Risk Preference in Affine Term Structure Models, forthcoming, *Review of Financial Studies.*

de Finetti, Bruno, 1931, Sul Significato Soggettivo della Probabilità, Fundamenta Mathematicae 17, 298-329. Translated into English, On the subjective meaning of probability, in Paola Monari and Daniela Cocchi (Editors), *Probabilitá e Induzione*, 1993, Clueb, Bologna, 291-321.

Diaconis, Persi and Daniel Stroock, 1991, Geometric bounds for eigenvalues of Markov chains, *Annals of Applied Probability* 1, 36-61.

Duffie, Darrell, 1996, State-Space Models of the Term Structure of Interest Rates, in H.Körezlioglu, B. Øksendal, and A. Üstünel, editors, *Stochastic Analysis and Related Topics V*: The Silivri Workshop, Boston: Birkhauser.

Duffie, Darrell, Damir Filipovic, and Walter Schachermayer, 2003, Affine Processes and Applications in Finance, *Annals of Applied Probability* 13, 984-1053.

Duffie, Darrell, Kenneth Singleton and Jun Pan, 2000, Transform Analysis and Asset Pricing for Affine Jump−Diffusions, *Econometrica* 68, 1343–1376.

Duffie, Darrell and Jun Pan, 1997, An Overview of Value at Risk, *Journal of Derivatives* 4, 7-49.

Elekes, György, 1986, A geometric inequality and the complexity of computing volume, *Discrete Computing in Geometry* 1, 289–292.

Eraker, Bjørn, 2001, MCMC Analysis of Diffusion Models with Applications to Finance, *Journal of Business and Economic Statistics* 19-2, 177-191.

Elerian, Ola, Sidhartha Chib and Neil Shephard, 2001, Likelihood Inference for Discretely Observed Nonlinear Diffusions, *Econometrica* 69, 959-994.

Eraker, Bjørn, 2003, Do Equity Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices, forthcoming, *Journal of Finance*.

Eraker, Bjorn, Michael Johannes and Nicholas Polson, 2003, The Impact of Jumps in Equity Index Volatility and Returns, *Journal of Finance* 58, 1269-1300.

Frieze, Alan, Ravi Kannan, and Nichoals Polson, 1994, Sampling from log-concave distributions. *Annals of Applied Probability* 4, 812-834

Gelfand, Alan, Susan Hills, Amy Racine-Poon, and Adrian Smith, 1990, Illustration of Bayesian inference in normal data models using Gibbs Sampling, *Journal of the American Statistical Association* 85, 972-982.

Gelfand, Alan. and Adrian Smith, 1990, Sampling Based approaches to calculating Marginal densities. *Journal of the American Statistical Association* 85, 398-409.

Gelfand, Alan, Adrian Smith, and T.M. Lee, 1992, Bayesian Analysis of constrained parameters and truncated data problems using Gibbs Sampling, *Journal of the American Statistical Association* 87, 523-532.

Geman, Stuart and Don Geman, 1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 6, 721-741.

Geyer, Charles, 1993, Practical Markov chain Monte Carlo. *Statistical Science* 7, 473-511

Glasserman, Paul and Nicholas Merener, 2003, Numerical Solution of Jump-Diffusion LIBOR Market Models, *Finance and Stochastics* 7, 1-27.

Gordon, N., Salmond, D. and Smith, Adrian, 1993, Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings*, F-140, 107–113.

Gray, Stephen, 1996, Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process, *Journal of Financial Economics* 42, 27-62.

Hammersley, John and Peter Clifford, 1970, Markov fields on finite graphs and lattices, Unpublished Manuscipt.

Hammersley, John, 1974, Discussion of Besag's paper, *Journal of the Royal Statistical Society. Series B, 36, 230-231.*

Han, Cong and Bradley Carlin, 2000, MCMC Methods for Computing Bayes Factors: A Comparative Review, working paper, University of Minnesota.

Hastings, W. Keith, 1970, Monte Carlo sampling Methods using Markov Chains and their Applications. *Biometrika* 57, 97-109.

Heston, Steven, 1993, A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options, *Review of Financial Studies* 6, 327-343.

Hobert, J.P. and George. Casella, 1996, The effect of improper priors on Gibbs sampling in hierarchical linear models. *Journal of the American Statistical Association* **91**, 1461-1473.

Honore, Peter, 1998, Pitfalls in estimating jump-diffusion models, working paper, University of Aarhus.

Jacquier, Eric and Robert Jarrow, 2000, Bayesian Analysis of Contingent Claim Model Error, *Journal of Econometrics* 94, 145-180. 2000.

Jacquier, Eric, Nicholas Polson, and Peter Rossi, 1994, Bayesian analysis of Stochastic Volatility Models, (with discussion). *Journal of Business and Economic Statistics* 12, 4.

Jacquier, Eric, Nicholas Polson, and Peter Rossi, 1995, Models and Priors for Multivariate Stochastic Volatility Models, working paper, University of Chicago.

Jacquier, Eric, Nicholas Polson, and Peter Rossi, 2004, Bayesian Inference for SV models with Correlated Errors, forthcoming, *Journal of Econometrics*.

Johannes, Michael, 2004, The Statistical and Economic Role of Jumps in Interest Rates, forthcoming *Journal of Finance*.

Johannes, Michael, Nicholas Polson and Jonathan Stroud, 2002, Volatility Timing and Portfolio Returns, working paper, Columbia University.

Johannes, Michael, Nicholas Polson and Jonathan Stroud, 2003, Optimal filtering of jump-diffusions: extracting jumps and volatility from prices, working paper, Columbia University.

Johannes, Michael, Nicholas Polson and Jonathan Stroud, 2003, Sequential parameter estimation in stochastic volatility jump-diffusion models, working paper, Columbia University.

Jones, Christopher, 2003, The Dynamics of Stochastic Volatility: Evidence from Underlying and Options Markets, forthcoming, *Journal of Econometrics*.

Kass, Robert and Adrian Raftery, 1995, Bayes Factors," *Journal of the American Statistical Association* 90, 773-795.

Kiefer, Nicholas, 1978, Discrete parameter variation: efficient estimation of a switching regression mode, *Econometrica*, 46, 427–434.

Kim, Sangyoon, Neil Shephard and Siddhartha Chib, 1998, Stochastic volatility: likelihood inference and comparison with ARCH models, *Review of Economic Studies* 65, 361-93.

Kim, Tong and Edward Omberg, 1996, Dynamic Non-Myopic Portfolio Behavior, *Review of Financial Studies*, 9, 141-161.

Kitagawa, Genshiro, 1994, The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother, *Annals of the Institute of Statistical Mathematics* 46, 605-623.

Kitagawa, Genshiro, 1996, Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, Journal of Computational Graphics and Statististics 5, 1–25.

Kloeden, Peter and Eckhard Platen, 1995, *Numerical Solution of Stochastic Differential Equations*, Springer Verlag.

Lamoureux, Chris and Doug Witte, 2002, Empirical Analysis of the Yield Curve: The Information in the Data Viewed through the Window of Cox, Ingersoll, and Ross, *Journal of Finance* 57, 1479–1520.

Landen, Camilla, 2000, Bond pricing in a hidden Markov model of the short rate, *Finance and Stochastics* 4, 371-389.

Lindgren, Georg, 1978, Markov regime models for mixed distributions and switching regressions, *Scandanavian Journal of Statistics* 5, 81-91.

Lindley, Dennis, 1971, *Bayesian Statistics: A Review*, SIAM, Philadelphia.

Liu, Jun, 1999, Portfolio Selection in Stochastic Environments, Working paper, UCLA.

Liu, Jun, Francis Longstaff and Jun Pan, 2003, Dynamic Asset Allocation with Event Risk, *Journal of Finance 58,* 231–259.

Liu, Jun, Wing Wong and Agustine Kong , 1994, Covariance Structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes, *Journal of the Royal Statistical Association Series B*, 57, 157-169.

Mengersen, Kerrie and Robert, Christian, 1998, MCMC Convergence Diagnostics: A Review (with discussion). In *Bayesian Statistics 6*, Jose Bernardo et al (Eds.), Oxford University Press, Oxford, 399-432.

Mengersen, Kerrie and Richard Tweedie, 1996, Rates of convergence of the Hastings and Metropolis algorithms, *Annals of Statistics* 24, 101–121.

Merton, Robert, 1976, Option pricing when the underlying stock returns are discontinuous, *Journal of Financial Economics 3,* 125-144.

Merton, Robert, 1980, Estimating the expected return on the market, *Journal of Financial Economics* 8, 323-363.

Metropolis, Nicholas, Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, Edward, 1953, Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, **21**, 1087-1091.

Meyn, Sean and Richard Tweedie, 1995, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York.

Mikulevicius, R., Platen, Eckhart, 1988, Time Discrete Taylor Approximations for Ito Processes with Jump Component, *Math. Nachr.*138, 93 - 104.

Naik, Vasant and Moon Hoe Lee, 1997, Yield Curve Dynamics with Discrete Shifts in Economic Regimes: Theory and Estimation, Working paper, University of British Columbia.

Nummelin, E., 1984, *General irreducible Markov chains and non-negative operators*, Cambridge University Press.

Pan, Jun, 2002, The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study, *Journal of Financial Economics* 63, 3–50.

Pastor, Lubos, and Robert Stambaugh, 2000, Comparing asset pricing models: An investment perspective , with Robert F. Stambaugh, *Journal of Financial Economics 56,* 335–381.

Piazzesi, Monika, 2004, Bond yields and the Federal Reserve, forthcoming *Journal of Political Economy.*

Pitt, Michael, and Neil Shephard, 1999, Filtering via simulation: auxiliary particle fillters, *Journal of the American Statistical Association* 94, 590-599.

Platen, Echard and Ricardo Rebolledo, 1985, Weak Convergence of Semimartingales and Discretization Methods, *Stochastic Processes and Their Application* 20, 41 - 58.

Poincare, Henri, 1901, *Science and Hypothesis*, New York, Dover.

Polson, Nicholas, 1992, Comment on Practical Markov chain Monte Carlo by Charles Geyer, *Statistical Science* 7, 490-491.

Polson, Nicholas, 1996, Convergence of Markov Chain Monte Carlo Algorithms (with discussion). In *Bayesian Statistics 5*, J.M. Bernardo et al (Eds.). Oxford University Press, Oxford, 297-323.

Polson, Nicholas and Jonathan Stroud, 2003, Bayesian Inference for Derivative Prices, In *Bayesian Statistics 7* (Bernardo et al., eds.), Oxford University Press, 641-650.

Polson, Nicholas, Jonathan Stroud and Peter Muller, 2002, Nonlinear State-Space Models with State-Dependent Variances, *Journal of the American Statistical Association* 98, 377-386.

Polson, Nicholas, Jonathan Stroud and Peter Muller, 2002, Affine state dependent variance models, working paper, University of Chicago.

Polson, Nicholas, Jonathan Stroud and Peter Muller, 2003, Practical Filtering for Stochastic Volatility Models, working paper, University of Pennsylvania.

Pritsker, Matt, 1998, Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models, *Review of Financial Studies*, 449-87.

Raiffa, Howard and Robert Schlaifer, 1961, *Applied Statistical Decision Theory*, Harvard University., Boston, MA.

Ramsey, Frank P. 1931, Truth and Probability, In The Foundations of Mathematics and other Logical Essays, 156-198, Routledge and Kegan Paul, London.

Ripley, Brian, 1992, *Stochastic Simulation*, Wiley, New York.

Ritter, Charles and Martin Tanner, 1991, Facilitating the Gibbs sampler: the Gibbs stopper and the Griddy-Gibbs sampler, *ournal of the American Statistical Association* 87, 861-868.

Robert, Christian and George Casella, 1999, Monte Carlo Statistical Methods, New York, Springer.

Roberts, Gareth and Nicholas Polson, 1994, On the Geometric Convergence of the Gibbs sampler. *Journal of the Royal Statistical Association, Series* B, 377-384.

Roberts, Gareth and Jeffrey Rosenthal, 1998, Markov chain Monte Carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics*, 26, 4-31.

Roberts, Gareth and Jeffrey Rosenthal, 2001, Optimal scaling for various Metropolis-Hastings algorithms, *Statistical Science* 16:351-367.

Roberts, Gareth and Jeffrey Rosenthal, Optimal scaling of discrete approximations to Langevin diffusions. (JRSSB 60:255-268, 1998.

Roberts, Gareth and Adrian Smith, 1994, Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms, *Stochastic Processes and Their Application,* 49, 207–216.

Roberts, Gareth and Richard Tweedie, 1996, Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms, *Biometrika*, 83, 95–110.

Roberts, Gareth and Richard Tweedie, 1999, Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and Their Application,* 80 211–229.

Rosenthal, Jeffrey, 1995a, Rates of convergence for Gibbs sampling for variance component models. *Annals of Statistics*, 23, 740–761.

Rosenthal, Jeffrey, 1995b, Minorization Conditions and Convergence Rates for MCMC, *Journal of the American Statistical Association* 90, 558-566.

Runggaldier, Wolfgang, 2003, Jump Diffusion Models, In *Handbook of Heavy Tailed Distributions in Finance* (S.T. Rachev, ed.), Handbooks in Finance, Book 1 Elesevier/North-Holland, 169-209

Savage, Leonard, 1964, *The Foundations of Statistics,* John Wiley, New York.

Scott, Steven, 2002, Bayesian Methods for Hidden Markov Models, *Journal of the American Statistical Association* 97, 337-351.

Shimony, Abner, 1955, Coherence and the axioms of confirmation, *The Journal of Symbolic Logic* 20, 1-28.

Smith, Adrian and Roberts, Gareth, 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.

Stanton, Richard, 1997, A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk, *Journal of Finance* 52, 1973-2002.

Storvik, Geir, 2002, Particle filters in state space models with the presence of unknown static parameters, *IEEE Transactions of Signal Processing* 50, 281–289.

Stroud, Jonathan, Peter Müller, Peter and Nicholas Polson, 2001, Nonlinear State-Space Models with State-Dependent Variance Functions, forthcoming, *Journal of the American Statistical Association.*

Tanner, Martin and Wing Wong, 1987, The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.

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

Tweedie, Richard and K.L. Mengersen, 1996, Rates of convergence in the Hastings-Metropolis algorithm, *Annals of Statistics,* 24 101-121.

Vasicek, Oldrich, 1977, An equilibrium characterization of the term structure, *Journal of Financial Economics* 5, 177–188.