MCMC Methods for Financial Econometrics

Michael Johannes and Nicholas Polson*

May 8, 2002

Abstract

This chapter discusses Markov Chain Monte Carlo (MCMC) based methods for estimating continuous-time asset pricing models. We describe the Bayesian approach to empirical asset pricing, the mechanics of MCMC algorithms and the strong theoretical underpinnings of MCMC algorithms. We provide a tutorial on building MCMC algorithms and show how to estimate equity price models with factors such as stochastic expected returns, stochastic volatility and jumps, multi-factor term structure models with stochastic volatility, time-varying central tendancy or jumps and regime switching models.

*We thank the Editors, Yacine Ait-Sahalia and Lars Hansen and Chris Sims for his discussion. We also thank Mark Broadie, Mike Chernov, Anne Gron and Paul Glasserman 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@gsbngp.uchicago.edu.

Contents

1 Introduction

Dynamic asset pricing theory uses equilibrium and arbitrage based arguments to derive asset prices conditional on a state variable model, parameters and market prices of risk. Empirical analysis of dynamic asset pricing models tackles the *inverse problem*: extracting information about the state variables and the parameters from the observed asset prices. The solution to this inverse problem is the distribution of the parameters, Θ, and state variables, X, conditional on observed prices, Y, which we denote by $p(\Theta, X|Y)$. Through the marginal distributions $p(\Theta|Y)$ and $p(X|Y)$ this distribution summarizes parameter estimation and state variable estimation and also provides specification diagnostics.

As an example, consider a model of an equity price, S_t, whose variance, V_t, follows a square-root process

$$\frac{dS_t}{S_t} = (r_t + \eta_v V_t)\,dt + \sqrt{V_t}dW_t^s \quad (1)$$

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

where W_t^s and W_t^v are two scalar Brownian motions with correlation ρ and r_t is the instantaneous spot interest rate. In this model, the goal of empirical asset pricing is to learn about the volatility states, $V = \{V_t\}_{t=1}^T$, the parameters that drive the evolution of the volatility process, κ_v, θ_v, σ_v and ρ, the risk premium, η_v, and assess model specification from observed prices, $Y = \{S_t\}_{t=1}^T$. Parameter inference and volatility estimation are summarized by the distributions $p(\kappa_v, \theta_v, \sigma_v, \rho, \eta_v|Y)$ and $p(V|Y)$, respectively, which are marginals from the joint distribution, $p(\kappa_v, \theta_v, \sigma_v, \rho, \eta_v, V|Y)$.

Characterizing $p(\Theta, X|Y)$ in continuous-time models is difficult for a number of reasons. First, the data are observed discretely while the models specify that asset prices and state variables evolve continuously through time. Second, in most interesting models, there are state variables which are latent from the perspective of the econometrician. Third, in practical problems $p(\Theta, X|Y)$ is typically of very high dimension due to the dimension of the state vector. Fourth, most continuous time models generate transition distributions for prices and state variables that are non-normal and non-standard (e.g., models with stochastic volatility or jumps). 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.

This chapter provides a description of MCMC methods and describes how they overcome these difficulties to generate samples from $p(\Theta, X|Y)$. We provide a general description of the Bayesian approach to estimation, a detailed discussion of the components of MCMC algorithms and their convergence properties and, finally, we show how to estimate a number of popular asset pricing models with MCMC methods.

We consider the following applications: equity price models with time-varying expected returns, stochastic volatility and jumps; dynamic term structure models with time-varying central tendency and stochastic volatility; regime-switching diffusion models and equity option pricing models. In each case, we start with the simplest model to develop intuition on the mechanics of MCMC and then sequentially proceed to more complicated models. For example, with equity price models we start with a geometric Brownian motion model and then consider the addition of Poisson driven jumps, stochastic volatility and stochastic expected returns. Finally, we show how to add option price data.

Our approach begins with an interpretation of continuous-time asset pricing models as state space models (see also Duffie (1996)). Specifically, the observation equation consists of

the conditional distribution of observed asset prices given state variables and parameters and the evolution equation consists of the dynamics of state variables given the parameters. For example, in the stochastic volatility model above, (1) is the observation equation and (2) is the state evolution. All of the asset pricing models we consider take the form of a nonlinear, non-Gaussian state space model.

Inference for parameters and state variables in nonlinear, non-Gaussian state space models is difficult, for the reasons mentioned above. *If* the parameters were known and the state space model were linear and Gaussian, the Kalman filter provides the posterior distribution of the state variables, $p(X|Y)$. Similarly, *if* the state variables were known, parameter estimation, characterizing $p(\Theta|Y)$, is straightforward. MCMC, on the other hand, provides a general methodology that applies in nonlinear and non-Gaussian state models and, unlike classical filtering methods, provides the distribution of both the state variables and the parameters given the data, $p(\Theta, X|Y)$. In fact, the Kalman filter is a MCMC algorithm in the case of a linear and Gaussian state space model with known parameters.

MCMC methods are particularly attractive for practical finance applications for several reasons. First, MCMC is a unified estimation procedure which simultaneously estimates both parameters and state variables. This implies, for example, that it is straightforward to separate out the effects of jumps and stochastic volatility in models of interest rates or equity prices. Instead of resorting to approximate filters or noisy latent variable proxies which are often used in the literature, MCMC computes the distribution of the state variables and parameters given the observed data.

Second, MCMC methods account for estimation and model risk. Estimation risk is the inherent risk present in estimating parameters or state variables. Increasingly in practical problems, estimation risk is serious issue whose impact must be quantified. For example, estimation risk is a very important component of optimal portfolio allocations. MCMC also allows the researcher to quantify model risk, the uncertainty over the choice of model.

Finally, MCMC is solely a conditional simulation methodology, and therefore avoids any maximization and long unconditional state variable simulation. Because of this, MCMC estimation is typically extremely fast in terms of computing time. The practical implication of this is that it allows the researcher to obtain answers in much shorter periods of time than competing methods. Moreover, it allows the researcher to perform Monte Carlo simulations to guarantee that the estimation procedure accurately estimates the objects of interest, a feature

not shared by many other methods.

The chapter is outlined as follows. The next section provides a brief, non-technical overview of MCMC methods highlighting the major components of MCMC algorithms. Section 3 discusses the Bayesian approach to estimation of continuous time asset pricing models. Section 4 discusses a number of prominent models and examples that fit nicely into the state space approach. Section 5 discusses the general mechanics and convergence of MCMC algorithms. Section 6 provides practical recommendations for users of MCMC. Section 7 discusses MCMC inference in models of equity prices and option prices and Section 8 covers a variety of term structure models. Section 9 discusses estimation and model risk and, finally, section 10 concludes.

2 Overview of MCMC

MCMC is a conditional simulation methodology that generates random samples from a given target distribution, in our case $p(\Theta, X|Y)$. The key to MCMC is a remarkable result known as the Hammersley-Clifford theorem. This theorem states that a joint distribution can be characterized by its complete set of conditional distributions. In our setting, the theorem implies that knowledge of $p(X|\Theta, Y)$ 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 characterize $p(\Theta, X|Y)$. Consider the following algorithm: given two initial draws, $X^{(0)}$ and $\Theta^{(0)}$, 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 a *Markov Chain* with attractive limiting properties: the distribution of the chain converges to $p(\Theta, X|Y)$, its equilibrium distribution, under a number of metrics and mild conditions.

The key to the success of MCMC algorithms is that it is easier to characterize two conditional densities, $p(X|\Theta, Y)$ and $p(\Theta|X, Y)$, then one joint density, $p(\Theta, X|Y)$, of higher dimension. In many models, the distribution of state variables conditional on parameters and data, $p(X|\Theta, Y)$, can be computed using standard filtering techniques. For example, in the case of a linear and Gaussian model, the Kalman filter provides $p(X|\Theta, Y)$. Similarly, the distribution of the parameters given observed data and state variables, $p(\Theta|X, Y)$, is typically easy to simulate as the state variables are observed.

If $p(\Theta|X,Y)$ and $p(X|\Theta,Y)$ are known in closed form and can be directly drawn from, the algorithm described above is known as a *Gibbs sampler*. In other cases, it may not be possible to directly sample from one or both of the conditional conditional distributions. In this case, sampling methods known as *Metropolis-Hastings algorithms* apply. These algorithms use a two step procedure. The first step samples a candidate draw from a proposal density which may be chosen to approximate the desired conditional distribution, and, in the second step, accepts or rejects this draw based on a specified acceptance criterion. Together, Gibbs steps and Metropolis-Hastings steps combine to generate what is known as MCMC algorithms.

These random samples can be used for parameter and state variable estimation using the *Monte Carlo* method, hence the name Markov Chain Monte Carlo. For example, a point estimate of the parameter, Θ_i, is typically the marginal posterior mean:

$$E(\Theta_i|Y) = \int p(\Theta_i|Y)\, d\Theta_i \approx \frac{1}{G}\sum_{g=1}^{G} \Theta_i^{(g)}.$$

In many cases, this simple Monte Carlo estimate can be improved using a technique known as Rao-Blackwellization. If there is an analytical form for the conditional density, $p\left(\Theta_i|\Theta_{-i}^{(g)}, X^{(g)}, Y\right)$ where Θ_{-i} refers to all of the elements of Θ except Θ_i, we can take advantage of the conditioning information to estimate the marginal posterior mean as

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

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

Estimation of the marginal posterior distribution for the state variables is similar using $p(X|Y)$, but it is now important to focus on a number of different densities, depending on how the conditioning information is used. To understand the issues, we let Y^T denote the full sample and Y^t denote the sample up to time t. There are three main problems that are associated with state variable estimation:

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

The filtering and forecasting problems are inherently sequential while the smoothing problem is static. With samples from $p(\Theta, X|Y)$ we can only estimate the smoothing distribution. Recent work shows that it is straightforward to obtain the filtering density, $p(X_t|Y^t)$, using MCMC methods (see, Johannes, Polson and Stroud (2001) and Muller, Polson and Stroud (2001)).

Returning to the smoothing problem, notice that the marginal smoothing distribution can be written as

$$p(X_t|Y) = \int p(\Theta, X|Y)\, d\Theta dX_{-t} \approx \frac{1}{G}\sum_{g=1}^{G} p\left(X_t|\Theta_i^{(g)}, X_{-t}^{(g)}, Y\right)$$

and the posterior mean is commonly reported as the estimate of the state variable at time t, $\widehat{X_t} = E[X_t|Y]$. This approach takes into account parameter estimation risk as the posterior integrates out the parameters and should be contrasted with classical methods which estimate state variables conditional on parameters: $\widehat{X_t}\left(\widehat{\Theta}\right) = E\left[X_t|\widehat{\Theta}, Y\right]$. It also easy to estimate the posterior standard deviation which measures estimation risk.

It is also straightforward to build credible sets, an analog to the classic confidence interval. For example, in the case of a parameter, a $(1-\alpha)$ per cent credible set is the subset of the marginal posterior distribution, C, such that $1-\alpha = \int_C p(\Theta_i|Y)\, d\Theta_i$ and can be estimated based on the quantiles of $\left\{\Theta^{(g)}\right\}_{g=1}^{G}$. Since the credible set is not unique, one can report the highest posterior credible set which is the smallest set C of the marginal posterior distribution with $1-\alpha$ per cent credibility.

The limiting properties of MCMC estimators, as $G \to \infty$, are also of interest. There are two types of convergence operating simultaneously. First, there is convergence in distribution of the Markov Chain $\left\{X^{(g)}, \Theta^{(g)}\right\}_{g=1}^{G}$ 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 $E[f(\Theta, X)|Y]$. It turns out that convergence of both of these is guaranteed by the Ergodic Theorem for Markov Chains, 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. In practice, geometric or even polynomial convergence rates are common.

Finally, it is also straightforward to the use the output of MCMC algorithms to provide model specification diagnostics. 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. These

provide finite sample results for model comparison. Estimates of state variables from MCMC algorithms also provide informal, but powerful diagnostics, as the properties of the estimated path of the state variables can be contrasted the model implied properties.

We now discuss the general estimation problem in greater detail.

3 Bayesian Inference and Asset Pricing Models

The Bayesian approach to inference in asset pricing models requires a number of modeling components: the likelihood, the state variable dynamics and parameter distributions. Formally, empirical asset pricing takes a given set of discretely observed asset prices, $Y = \{Y_t\}_{t=1}^T$ and seeks to learn about the parameters and the state variables, that is, to characterize $p(\Theta, X|Y)$. It is useful to decompose $p(\Theta, X|Y)$ via Bayes rule into

$$p(\Theta, X|Y) \propto p(Y|X, \Theta)\, p(X|\Theta)\, p(\Theta). \tag{3}$$

Here $p(Y|X, \Theta)$ is the likelihood function, $p(X|\Theta)$ is the distribution of the state variables arising from the parametric model specification and $p(\Theta)$ is the prior distribution of the parameters. The distribution $p(X|\Theta)$ is given by a parametric model specification and $p(\Theta)$ summarizing any non-sample information about the parameters.

The likelihood function and the state variables can be decomposed, assuming a Markov evolution, into its components:

$$p(Y|X, \Theta) = p(Y_1, ..., Y_T|X, \Theta) = \prod_{t=1}^{T} p(Y_t|Y_{t-1}, X_{t-1}, \Theta)$$

and

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

where, for simplicity, we normalize the time span between observations to 1.

Since Y are observed prices, the likelihood function is just the distribution of the increments in the asset price conditional on the parameters and state variables. At this stage, it is important to recognize that we utilize the full-information or data augmented likelihood function, $p(Y|X, \Theta)$, which conditions on the state variables in addition to the parameters.

This differs from the marginal or classical likelihood function, $p(Y|\Theta)$, which integrates or marginalizes out 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 of interest in finance, there are latent variables (at least from the perspective of the econometrician) and analytical computation of the likelihood function is generally intractable, making maximum likelihood unattractive. MCMC methods, in contrast, directly deal with the latent variables by using the augmented likelihood and provide a likelihood based approach when direct maximum likelihood is infeasible.

In order to construct an MCMC algorithm, we need to be able to evaluate the conditional distributions of interest in the likelihood and state dynamics, $p(Y_{t+1}|Y_t, X_t, \Theta)$ and $p(X_{t+1}|X_t, \Theta)$. In the setting of continuous-time models, the conditional distributions of the asset prices and state variables arise as solutions to a parameterized stochastic differential equations. We now discuss the connection between the likelihood, state variables dynamics, asset pricing models and stochastic differential equations. We also discuss the role of the parameter distribution.

3.1 Prices and the Likelihood Function

There are two types of likelihoods that arise in practice. In the first case, which is common in models of equity prices or exchange rates, the dynamics of the prices are directly modeled as the solution to a stochastic differential equation. In the second case, which is common in the case of option pricing and term structure modeling, there is a deterministic function between prices and state variables and parameters which requires special attention.

In the first case, asset prices solve the parameterized stochastic integral equation

$$Y_{t+1} = Y_t + \int_t^{t+1} \mu_y(Y_s, X_s, \Theta)\, ds + \int_t^{t+1} \sigma_y(Y_s, X_s, \Theta)\, dW_s + \sum_{j=N_t}^{N_{t+1}} \xi_j,$$

where the dynamics are driven by the state variables, a vector of Brownian motions $\{W_t\}_{t\geq 0}$, and vector point process $\{N_t\}_{t\geq 0}$ with stochastic intensity λ_t, and ξ_j is a jump with $\mathcal{F}_{\tau_j -}$ distribution $\Pi_{\tau_j -}$. All of these random variables are defined on a filtered probability space

$\left(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P\right)$ and we assume that characteristics have sufficient regularity for a well-defined solution to exist. The distribution implied by the solution of the stochastic differential equation, $p\left(Y_{t+1}|Y_t, X_t, \Theta\right)$, generates the likelihood function.

In the second case, at least one of the asset prices is a known function of the state variables and parameters, $Y_t = f\left(X_t, \Theta\right)$. The econometrician observes neither the parameters nor the state variables. Classic examples are multifactor term structure and option pricing models. In multi-factor term structure models, the short rate process, r_s, is often assumed to be a function of a set of state variables, $r_s = r\left(X_s\right)$, and bond prices are given by

$$f\left(X_t, \Theta\right) = E^Q\left[e^{-\int_t^T r(X_s)ds}|X_t\right]$$

where Q is an equivalent martingale measure on the original probability space and the function f can be computed either analytically or as the solution to ordinary or partial differential equation. In models of option prices, $f\left(X_t, \Theta\right)$, is given as

$$f\left(X_t, \Theta\right) = E^Q\left[e^{-\int_t^T r(X_s)ds}\left(X_T - K\right)_+|X_t\right]$$

in the case of a call option where $(\cdot)_+$ denotes the positive component. More generally, the Fundamental Theorem of Asset pricing (see, e.g, Duffie (2001)), asserts the existence of a probability measure Q, equivalent to P, such that prices are discounted expected values of payoffs under Q.

In term structure and option pricing applications, the observation equation is technically a degenerate distribution as the prices are known, with probability one conditional on state variables and parameters. In this case, if the parameters were actually known, the state variables can often be inverted from observed prices, $X_t = f^{-1}\left(Y_t, \Theta\right)$. An example of this is implying volatility from observed option prices. However, in practice, the parameters are not known and researchers commonly assume there exists a pricing error, ε_t. In this case of an additive pricing error,

$$Y_t = f\left(X_t, \Theta\right) + \varepsilon_t$$

where $\varepsilon_t \sim N\left(0, \Sigma_\varepsilon\right)$.

There are two justifications for using a pricing error. First, there is often a genuine concern with noisy price observations generated by the bid-ask spread. An excellent example of this is an at-the-money option price on an equity index like the S&P 100 or 500 which typically

has a bid-ask spread equal to at least 5-10% of the value of the option. In fixed income, zero coupon or par bond yields are often extrapolated which also generates measurement error.

Second, the pricing error can be useful as a tool to break stochastic singularities between the prices and the state variables which occurs when there are more observed asset prices than state variables. This is often used in option pricing and fixed income settings, see, for example, Chen and Scott (1993), Bates (1996, 2000) or Pan (2001). 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.

3.2 State Variable Dynamics

Explicit in the asset prices are a set of state variables, X_t, that are also modeled as solutions to stochastic differential equations (SDEs). We now provide a discussion of the three types of state variable specifications most often used in applications: first, diffusion models; second jump-diffusion models and third, Markov switching diffusions.

First, since the seminal work of Merton (1971) and Black and Scholes (1973), researchers commonly specify state variables as diffusions. Here, the state variables dynamics are generated by

$$X_{t+1} = X_t + \int_t^{t+1} \mu\left(X_s, \Theta\right) ds + \int_t^{t+1} \sigma\left(X_s, \Theta\right) dW_s$$

where again W_t is a vector of Brownian motions under the P-measure and we assume sufficient regularity on μ, σ and X_0 for a well-behaved solution to exist.

The defining characteristics of a diffusion are the continuous sample path and its Markov structure. A common way to relax the continuity assumption is to introduce a marked point process in addition to the diffusion component. Specifically, a point process (N_t) counts the number of jump times $\{\tau_j\}_{j=1}^{\infty}$ prior to time t. At each time τ_j a mark or jump, ξ_j arrives and induces a discontinuity in the state variables, $X_{\tau_j} - X_{\tau_j -} = \xi_j$. Between jumps, the state variables diffuse which implies that X_t solves

$$X_{t+1} = X_t + \int_t^{t+1} \mu\left(X_s, \Theta\right) ds + \int_t^{t+\Delta} \sigma\left(X_s, \Theta\right) dW_s + \sum_{j=N_t}^{N_{t+1}} \xi_j. \tag{4}$$

Now, in addition to μ and σ, the dynamics of the state variables are characterized by the arrival intensity of point process, $\lambda_t = \lambda\left(X_t\right)$, and the $F_{\tau_j -}$ conditional distribution of the

jump sizes, $\Pi\left(X_{\tau_j-}, \Theta\right)$.

Third, the drift and diffusion coefficients could randomly jump over time which leads to a regime switching model. This is in sharp contrast to the previous model where the levels of the state variables jump. We assume that the drift and diffusion coefficients are functions of a continuous-time, discrete state Markov chain, Z_t, which is characterized by a set of states, $Z = \{Z_1, ..., Z_k\}$, and a transition rate function $\lambda(i, j)$ which is interpreted as the conditional probability of the chain moving from state i to state j in an instant:

$$P\left[Z_{t+dt} = j | Z_t = i\right] = \lambda(i, j)\, dt.$$

In this case, X_t solves

$$X_{t+\Delta} = X_t + \int_t^{t+\Delta} \mu\left(X_s, Z_s, \Theta\right) ds + \int_t^{t+\Delta} \sigma\left(X_s, Z_s, \Theta\right) dW_s$$

where again, we assume sufficient regularity for well-behaved solutions to exist.

3.3 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 we always choose a parameterized distribution. This implies that the researcher must choose both a distribution for the prior and the parameters that index the distribution. Through both the choice of distribution and prior parameters, the researcher can impose non-sample information or, alternatively, choose to impose little information. In the latter case, an "uninformative" or diffuse prior is one that provide little or no information regarding the location of the parameters.

When possible we use standard conjugate prior distributions (see, for example Raiffa and Schaeffer (1961) or DeGroot (1970)) which 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, consider a geometric Brownian motion model for returns which implies that continuously compounded returns, Y_t, are normally distributed, $Y_t \sim N(\mu, \sigma^2)$. Assuming a normal prior on μ, $\mu \sim N(a, A)$, the conditionl posterior distribution $p(\mu|\sigma^2, Y)$ is also normally distributed, $N(a^*, A^*)$, where the starred parameters depend on the data, sample

size and on 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 captures generates an uninformative prior. For the variance parameter, the inverted gamma distribution is also conjugate (see Appendix A). 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 is not proper, that is, it 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 $\int_\Theta p(\Theta|Y)\, d\Theta = 1$. In many cases, flat priors can lead to improper posterior. This is especially true in state space models where the the marginal likelihood, $p(Y|\Theta)$, is unavailable in closed form and it is impossible to check the propriety of the posterior. Additionally, as we discuss later, joint posterior propriety is a necessary condition for MCMC algorithms to converge. This implies that one motivation for using diffuse proper priors is a computational tool for implementation via MCMC. For a recent review of noninformative priors, see Kass and Wasserman (1996).

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

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

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.

Another use of informative priors is using the prior 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, κ_v is often very small which 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 i and j. 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 application to impose economically motivated priors. For example, Pastor and Stambaugh (2000) use the prior to represent an investor's degree of belief over a multifactor 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. The extreme of this is when the parameters are not identified by the likelihood and the posterior is equal to the prior.

3.4 Time-Discretization

Before applying MCMC methods in specific cases, there is an important issue which must be discussed. The state variable dynamics and the likelihood are both abstractly given as conditional distribution arising from the solution of stochastic differential equations, $p(X_{t+1}|X_t, \Theta)$ and $p(Y_{t+1}|Y_t, X_t, \Theta)$. Only in a few simple cases, e.g., a square root process, Gaussian process or geometric Brownian motion, is the transition density of the prices or the state variables known in closed form.

To see the problem, consider a diffusive specification for the state variables. We know explicitly recognize the length of time between observations and denote it by Δ. Previously, we specified generically that $\Delta = 1$. The conditional distribution of $X_{t+\Delta}$ given X_t is generated

by

$$X_{t+\Delta} = X_t + \int_t^{t+\Delta} \mu\left(X_s, \Theta\right) ds + \int_t^{t+\Delta} \sigma\left(X_s, \Theta\right) dW_s.$$

The induced distribution of the state increments, $X_{t+\Delta} - X_t$, is difficult to characterize because the distributions of $\int_t^{t+\Delta} \mu\left(X_s, \Theta\right) ds$ and $\int_t^{t+\Delta} \sigma\left(X_s, \Theta\right) dW_s$ are not generally known. However, if the drift and diffusion functions are continuous functions of the state, it is not unreasonable to assume for short time increments that

$$\begin{aligned} \int_t^{t+\Delta} \mu\left(X_s, \Theta\right) ds &\approx \mu\left(X_t, \Theta\right) \Delta \text{ and} \\ \int_t^{t+\Delta} \sigma\left(X_s, \Theta\right) dW_s &\approx \sigma\left(X_t, \Theta\right) \left(W_{t+\Delta} - W_t\right). \end{aligned}$$

These approximations lead to the following "Euler" approximation for the state variables:

$$X_{t+\Delta} = X_t + \mu\left(X_t, \Theta\right) \Delta + \sigma\left(X_t, \Theta\right) \left\{W_{t+\Delta} - W_t\right\}.$$

This discretization implies that the distribution of the increments is conditionally normal,

$$p\left(X_{t+\Delta} - X_t | X_t, \Theta\right) \sim N\left(\mu\left(X_t, \Theta\right) \Delta, \Sigma\left(X_t, \Theta\right) \Delta\right),$$

where $\Sigma = \sigma\sigma'$ and the state dynamics, $p\left(X|\Theta\right)$, are given by the product of normal distributions.

Poisson driven jumps or discrete-state Markov can similarly be dealt with. Consider, for example, the case of jump-diffusion state variables. For example, in the case of a jump-diffusion, (4). The only difference in this case is that we time-discretize the point process that generates jump times. The point process, N_t, has the property that

$$Prob(N_{t+\Delta} - N_t = 1) \approx \lambda_t \Delta.$$

To discretization of the point process, we define an indicator variable $J_{t+\Delta}$ and assume that $J_{t+\Delta} = 1$ (with probability $\lambda_t \Delta$). Similarly, the jump size distribution is approximated by $\xi_{t+\Delta} \sim \Pi\left(X_t, \Theta\right)$. This implies that a time-discretization of the jump-diffusion model is given by:

$$X_{t+\Delta} = X_t + \mu\left(X_t, \Theta\right) \Delta + \sigma\left(X_t, \Theta\right) \left\{W_{t+\Delta} - W_t\right\} + J_{t+\Delta} \xi_{t+\Delta}.$$

Platen and Rebolledo (1985) or Liu and Li (2000) provide formal justifications for these approximations.

Given the discretization, it is common in models with jumps to expand the state space to include the jump times and the jump sizes. The jump-augmented state vector now consists of $[X_t, J_t, \xi_t]$. However, and this is the key to the MCMC approach, the distribution of the increment is normally distributed, conditional on current state and the jump times and sizes:

$$X_{t+\Delta}|X_t, J^x_{t+\Delta}, \xi^x_{t+\Delta} \sim N\left(X_t + \mu_t\Delta + J^x_{t+\Delta}\xi^x_{t+\Delta}, \sigma_t\sigma_t'\Delta\right)$$

where we have suppressed the dependence of the drift and diffusion on the parameters and state variables. Time discretization of a Markov chain proceeds similarly.

The time discretization generates a much simpler conditional distribution structure and allows the use of standard MCMC techniques. As an approximation, it is important to check that this does not introduce any systematic biases. One method of checking the accuracy of this Euler approximation is to perform a simulation study to check that the discretization is not introducing any substantive biases. For example, in an equity price model with stochastic volatility, jumps in returns and jump in volatility, Eraker, Johannes and Polson (2001) document that there is not any systematic biases in parameter estimates, while Johannes, Kumar and Polson (1998) show that in Merton's jump-diffusion model, the discretization bias is negligible. As noted by Pritsker (1997, 1998) and Johannes (2001), 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 Euler approximation may not provide an accurate approximation to the true dynamics. MCMC solves this problem by using a novel method developed by Eraker (1997), Elerian Shephard and Chib (2000) and Jones (1999) to "fill in" asset price or state variable values at times in between observation dates. To see how this works, assume that data is observed at a frequency of Δ. The augmentation introduces an additional "missing" observation $X_{t+\frac{\Delta}{2}}$ of the state variable at time $t + \Delta/2$ thus discretizing at a finer partition.

Now, given an Euler approximation for the dynamics over the intervals $\left[t, t+\frac{\Delta}{2}\right]$ and $\left[t+\frac{\Delta}{2}, t+\Delta\right]$, the dynamics of the state variables between the observation dates is

$$\begin{aligned} X_{t+\Delta} &= X_t + \left[\mu\left(X_t, \Theta\right) + \mu\left(X_{t+\frac{\Delta}{2}}, \Theta\right)\right]\frac{\Delta}{2} \\ &\quad + \sigma\left(X_t, \Theta\right)\left\{W_{t+\frac{\Delta}{2}} - W_t\right\} + \sigma\left(X_{t+\frac{\Delta}{2}}, \Theta\right)\left\{W_{t+\Delta} - W_{t+\frac{\Delta}{2}}\right\}. \end{aligned}$$

If we treat $X_{t+\frac{\Delta}{2}}$as an additional state variable that we simulate along with the other state variables and the parameters then MCMC algorithms are straightforward to apply. A number

of multi-factor interest rate models have been applied by Eraker (1997) and Jones (2001) and in the scalar diffusion case by Elerian Shephard and Chib (2000).

4 Asset Pricing Models

This section describes models that conveniently fit into framework developed above and are especially well-suited for MCMC estimation.

4.1 Continuous-time Equity Price Models

Following Merton (1969, 1971) and Black and Scholes (1973), equity prices are typically modeled in continuous-time as this specification often leads to analytically tractable solutions in portfolio allocation and option pricing applications. The first generation of models specified that equity prices were geometric Brownian motions

$$dS_t = \mu S_t dt + \sigma S_t dW_t^s \tag{5}$$

where, for simplicity, W_t^S is a scalar Brownian motion.

Empirical tests typically reject the geometric Brownian motion model and lead researchers to consider models with jumps, stochastic expected returns and volatility:

$$dS_t = \mu_t S_{t-} dt + S_{t-}\sqrt{V_{t-}} dW_t^s + d\left(\sum_{j=1}^{N_t^s} S_{\tau_j -}\left(e^{\xi_j^s} - 1\right)\right) \tag{6}$$

where the expected returns are typically assumed to follow a Gaussian diffusion process and the volatility is a jump-diffusion:

$$\begin{aligned} d\mu_t &= \kappa_\mu\left(\theta_\mu - \mu_t\right) dt + \sigma_\mu dW_t^\mu \\ dV_t &= \kappa_v\left(\theta_v - V_{t-}\right) dt + \sigma\sqrt{V_{t-}} dW_t^v + d\left(\sum_{j=1}^{N_t^v} \xi_j^v\right) \end{aligned}$$

In this model, the observed data is typically the log-returns, $Y_t = \log\left(S_t/S_{t-1}\right)$, and the state variables are the time-varying mean and volatility, $X_t = [\mu_t, V_t]$. An alternative, the

log-volatility model, $d\log(V_t) = \kappa_v(\theta_v - \log(V_t))dt + \sigma_v dW_t^v$ is also popular for empirical applications.

The mean-reverting specification for expected returns is popular in the portfolio choice literature and was introduced by Merton (1971) (see also Kim and Omberg (1996), Liu (1999), Wachter (2000)). 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 (2002) estimate stochastic volatility models with jumps in returns and volatility using MCMC methods. Eraker (2002) extends Eraker, Johannes and Polson (2002) to incorporate option prices. Liu, Longstaff and Pan (2001) analyze the portfolio implications of models with jumps in stock prices and in volatility. Duffie and Pan (2001) consider multivariate version of Merton's jump-diffusion model for equity prices evaluating market and credit risk and Glasserman and Kou (2000) use it as a model of forward rates.

4.2 Affine Diffusion Term structure models

Affine term structure models specify that the instantaneous spot rate, r_t, and the drift and diffusion coefficients of the state variables are affine functions of the states:

$$\begin{aligned} r(X_t) &= \gamma_0 + \gamma_1' X_t \\ dX_t &= (a + bX_t)\,dt + \sigma(X_t)\,dW_t \end{aligned}$$

where

$$\sigma(X_t)\,\sigma(X_t)' = \Sigma(X_t) = \Sigma^0 + \sum_{k=1}^{N} \Sigma^k X_t^k$$

and where X_t and a are $K \times 1$ vectors and Σ^0, Σ^k and b are $K \times K$ matrices.

The advantage of these models are that the continuously compounded yields on a zero coupon bond maturing at time τ_1 is linear in the states:

$$Y_{t,\tau_1} = \alpha(\tau_1, \Theta) + \alpha^x(\tau_1, \Theta)' X_t$$

where, as shown by Cox, Ingersoll and Ross (1981), Brown and Schaefer (1994) or Duffie and Kan (1994, 1996), α and α^x solve ordinary differential equations. In addition to continuously-compounded zero-coupon bond yields, coupon bond prices, par bond yields, discretely com-

pounded interest rates such as Libor and futures prices of discretely compounded interest rates are also exponential affine.

The observed data are typically a vector of continuously compounded yields with maturities $\tau = [\tau_1, ..., \tau_n]'$ which we denote by $Y_t = [Y_{t,\tau_1}, ..., Y_{t,\tau_n}]$ where $Y_{t,\tau_i} = \tau_i^{-1} \log P(t, \tau, X_t, \Theta)$. The term structure state space to be estimated is given by

$$\begin{aligned} Y_{t,\tau} &= \alpha(\tau, \Theta) + \alpha^x(\tau, \Theta)' X_t \\ dX_t &= (a + bX_t)\, dt + \sigma(X_t)\, dW_t. \end{aligned}$$

Notice that the state and observation equation are linear in the state variables. Although the parameters are linear in the evolution equation, the parameters enter the loading functions, $\alpha(\tau, \Theta)$ and $\alpha^x(\tau, \Theta)$, which are generally highly nonlinear and/or non-analytic. We show later how the Metropolis algorithm handles this problem.

4.3 Continuous-time Markov Switching Models

An alternative to the continuous-state space in the previous section is to assume that the coefficients are driven by a continuous-time, discrete state Markov Chain. For example, consider the following extension of the Vasicek short rate model:[1]

$$dr_t = \kappa_r(\mu(Z_t) - r_t)\, dt + \sigma_r(Z_t)\, dW_t^r$$

where $Z_t \in [Z_1, ..., Z_K]$ is a continuous-time Markov chain. Assuming the transition probabilities are state invariant, then the state space is given by

$$\begin{aligned} Y_t &= \alpha(\tau, Z_t, \Theta) + \alpha^r(\tau, Z_t, \Theta)\, r_t \\ dr_t &= \kappa_r(\mu(Z_t) - r_t)\, dt + \sigma_r(Z_t)\, dW_t^r \end{aligned}$$

where the functions α and α^r conditional on the given Markov state, solve standard Riccatti ordinary differential equations; $Y_t = [Y_{t,\tau_1}, ..., Y_{t,\tau_n}]$, is a vector of observed continuously

[1] Lee and Naik (1994), Landen (2000) and Dai and Singleton (2002) provide theoretical results on bond pricing when interest rates are subject to regime switches. Landon (2000) also considers the incomplete information case where the Markov state is not observed by the agents pricing bonds and must be filtered from observed prices. Ang and Bekaert (2000), Gray (1996) and Lee and Naik (1994) provide empirical analyses of regime switching interest rate models.

compounded yields. The state vector consists of the current Markov state, Z_t, and the instantaneous spot rate: $X_t = [r_t, Z_t]$.

Regime switches also provide an alternative to diffusive stochastic volatility and expected returns, $d\log(S_t) = \mu(Z_t)\,dt + \sigma(Z_t)\,dW_t$. For example, Naik (1993) derives the prices of options when volatility switches between high and low states and Kim, Mark and Lam (1998) estimate a regime-switching model of stochastic volatility.

4.4 Equity index option pricing models

In previous sections, we discussed equity models where the only observed data were the continuously compounded equity returns. Option prices sharpens inference by providing information about the market prices of volatility and jump risks that are embedded only in derivative prices.

Adding an option adds one level to the state space model:

$$\begin{aligned}
C_t &= E^Q\left[e^{-r(T-t)}\left(S_T - K\right)_+ | V_t, S_t\right] = f\left(S_t, V_t, K, T-t, \Theta\right) \\
dS_t &= \mu_t S_{t-}dt + S_{t-}\sqrt{V_{t-}}dW_t^s + d\left(\sum_{j=1}^{N_t^s} S_{\tau_j -}\left(e^{\xi_j^s} - 1\right)\right) \\
dV_t &= \kappa_v\left(\theta_v - V_{t-}\right)dt + \sigma_v\sqrt{V_{t-}}dW_t^v + d\left(\sum_{j=1}^{N_t^v}\xi_j^v\right)
\end{aligned}$$

where C_t is the price of a call option struck at K maturing at time T. In this case, the observed data is $Y_t = [C_t, \log(S_t/S_{t-1})]$. The fact that the option price is only known up to a numerical integration poses no problems for an MCMC based estimation approach as shown by Eraker (2001).

4.5 Structural Models of Default

A related problem is pricing corporate debt in a structural model of default. As an example of this, consider the structural default model of Merton (1974). In this case, the firm has assets with a market value of V_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

$(V_T - B)_+$. In this case, standard arguments imply that the value of equity, S_t, is given by $S_t = E_t^Q \left[e^{-r(T-t)} (V_T - B)_+ | V_t \right]$.

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

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

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, V_t, is an unobserved state variable and estimating it is one of the primary objectives.

5 MCMC methods: Theory

MCMC methods have a number of important theoretical underpinnings. The Hammersley-Clifford Theorem provides a characterization of $p(\Theta, X|Y)$ into its complete conditional distributions. An MCMC algorithm iteratively samples from these conditional distributions. When these complete conditionals can be directly sampled, the MCMC algorithm is known as a *Gibbs sampler*. When some or all cannot be directly sampled, we discuss a number of different approaches based on the widely applicable *Metropolis-Hastings* algorithm. MCMC algorithms, in general, possess attractive limiting properties. The general theory of Markov chains can be used to prove convergence and find convergence rates. For example, MCMC algorithms deliver an ergodic and central limit theorem. These can be used to judge convergence and provide Monte Carlo standard errors. We now discuss these issues in turn.

5.1 Hammersley-Clifford Theorem

Since $p(\Theta, X|Y)$ is rarely known in closed form, we resort to numerical methods to obtain samples from $p(\Theta, X|Y)$. The key to generating samples from $p(\Theta, X|Y)$ is to break $p(\Theta, X|Y)$ into a number of lower dimensional distributions which are easier to characterize.

The theoretical guide for simplifying the problem into a number of problems of lower dimension is a remarkable theorem by Hammersley and Clifford.[2] The general version of the

[2]Somewhat suprising, Clifford and Hammersley never published there results as they could not relax the

Hammersley-Clifford theorem (Besag, 1974) provides conditions for when a set of conditional distributions characterizes a unique joint distribution. In our setting, as mentioned earlier, its first application states that knowledge of $p(\Theta|X,Y)$ and $p(X|\Theta,Y)$ uniquely determines the joint distribution $p(\Theta,X|Y)$. Moreover, each of these distributions can be further simplified. For example, if we partition $\Theta=(\Theta_1,\Theta_2)$, knowledge of $p(\Theta_1|X,\Theta_2,Y)$, $p(\Theta_2|X,\Theta_1,Y)$ and $p(X|\Theta,Y)$ is equivalent to knowledge of $p(\Theta,X|Y)$. Repeated application of the Hammersley-Clifford theorem implies that, for example, drawing a $T+K$ state and parameter vector can be simplified by iteratively drawing from the $T+K$ one dimensional distributions.

This result is based on the following identity for $p(\Theta,X|Y)$. In this case, the Hammersley-Clifford theorem is based on the Besag formula (Besag (1974)) and states that for any pair (Θ^0,X^0) of points, the joint density $p(\Theta,X|Y)$ is determined as

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

as long as a *positivity* condition is satisfied. Thus knowledge of $p(\Theta|X,Y)$ and $p(X|\Theta,Y)$ is, 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.

5.2 Gibbs Sampling

We now have all of the necessary tools to build MCMC algorithms. Given the decomposition of the posterior into the complete set of conditionals, we need only sample from these distributions. When each of these distributions can be directly sampled using standard methods (see Devroye (1986) or Ripley (1992)), the algorithm is known as the Gibbs sampler. Often the posterior conditionals require draws from standard distributions such as the normal, beta, gamma or binomial.

positivity condition. For a discussion of the circumstances surrounding this, see the interesting discussion by Clifford (1974) and Hammersley (1974) after the paper by Besag (1974).

The Gibbs sampler is defined by the following algorithm:

1. Given a set of initial states $\left(\Theta^{(0)}, X^{(0)}\right)$
2. Draw $\Theta^{(1)} \sim p\left(\Theta|X^{(0)}, Y\right)$
3. Draw $X^{(1)} \sim p\left(X|\Theta^{(0)}, Y\right)$.

This algorithm generates a sequence of random variables, $\left\{\Theta^{(g)}, X^{(g)}\right\}_{g=1}^{G}$, that has $p\left(\Theta, X|Y\right)$ as its equilibrium distribution. If it is not possible to conveniently draw from these conditional distributions, another application of Hammersley-Clifford can be used to further simplify the algorithm.

For example, consider a blocking or partition of $\Theta \in R^K$ in $r \leq K$ components $\Theta = (\Theta_1, ..., \Theta_r)$ where each component Θ_j could be multidimensional. Given a partition, the Hammersley-Clifford theorem implies that the following set of conditional distributions

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

uniquely determines $p\left(\Theta|X, Y\right)$. The Gibbs sampler is now

1. Given a set of initial states $\left(\Theta^{(0)}, X^{(0)}\right)$
2. Draw $\Theta_i^{(1)} \sim p\left(\Theta_i|\Theta_1^{(1)}, ..., \Theta_{i-1}^{(1)}, \Theta_{i+1}^{(0)}, \ldots, \Theta_r^{(0)}, X^{(0)}, Y\right)$ for $i = 1, ..., r$
3. Draw $X^{(1)} \sim p\left(X|\Theta^{(0)}, Y\right)$.

If $p\left(X|\Theta, Y\right)$ cannot be conveniently sampled from, one can again decompose this distribution down into another block of conditionals. The standard way to do this is sample one state variable at a time.

Notice how the Gibbs sampler takes advantage of conditional probabilistic structure of the problem. The key is that the conditional distribution of a given state variable or parameter given all other variables can be directly sampled from. In many cases, the Gibbs sampler will end up drawing from standard distributions such normal, gamma or beta. This implies that complete characterization of the posterior distribution involves only repeatedly sampling from distributions that are easy to sample from.

5.3 Metropolis-Hastings

The Gibbs sampler requires that each of the conditional distributions can be easily sampled from. What happens in the case where one of the conditional distributions, generically, $\pi(\Theta_i^{(g+1)}) \triangleq p(\Theta_i|\Theta_1^{(g+1)}, \ldots, \Theta_{i-1}^{(g+1)}, \Theta_{i+1}^{(g)}, \ldots, \Theta_r^{(g)} X, Y)$, for example, is not easy to sample from? In this case, we use what is known as the Metropolis or Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm requires the researcher to specify a recognizable proposal density $q\left(\Theta^{(g+1)}|\Theta^{(g)}\right)$ where for simplicity we suppress the dependence of the proposal on the other parameters and the state variables. We also require that we can compute the posterior density ratio $\pi(\Theta^{(g+1)})/\pi(\Theta^{(g)})$ which is nearly always available in models of interest.

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:

$$\text{Step 1} \quad : \quad \text{Draw } \Theta_i^{(g+1)} \text{ from the proposal density } q(\Theta_i^{(g+1)}|\Theta_i^{(g)}) \tag{7}$$

$$\text{Step 2} \quad : \quad \text{Accept } \Theta_i^{(g+1)} \text{ with probability } \alpha\left(\Theta_i^{(g+1)}, \Theta_i^{(g)}\right) \tag{8}$$

where

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

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 will again generate samples $\left\{\Theta^{(g)}, X^{(g)}\right\}_{g=1}^{G}$.

The Metropolis-Hastings algorithm significantly extends the number of applications that can be analyzed as the conditional density need not be known in closed form. A number of points immediately emerge: (1) 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 or partial or ordinary differential equations. One only has to evaluate the true density at two given points; (2) Gibbs sampling is a special case of Metropolis-Hastings, where $q(\Theta_i^{(g+1)}|\Theta_i^{(g)}) \propto \pi(\Theta_i^{(g+1)})$ and from (7) this implies that the acceptance probability is always one and the algorithm always moves; (3) Because 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; (4) there is an added advantage when there are constraints in the parameter space — one can just reject these draws. Alternatively, one can sample conditional on that region (see Gelfand, Smith and Lee, 1993). This provides a convenient approach for analyzing nonlinear parameter restrictions imposed by economic models.

Although the theory places no restrictions on the proposal density q, it is important to note that the choice of proposal density will generally 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 separate attention.

5.3.1 Independence Metropolis-Hastings

In the general Metropolis-Hastings algorithm above, the candidate draw, $\Theta^{(g+1)}$ was drawn from proposal density, $q(\Theta^{(g+1)}|\Theta^{(g)})$, which depended on the previous Markov state $\Theta^{(g)}$. 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:

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

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

where

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

Even though the candidate draws, $\Theta_i^{(g+1)}$, 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.

5.3.2 Random-Walk Metropolis

Random walk Metropolis, the original algorithm considered by Metropolis, et al (1953), is the mirror image of the independence Metropolis-Hastings algorithms. 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 the t-distribution. Due to the symmetry in the proposal density, the algorithm simplifies to

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

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

where

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

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

5.4 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.

5.4.1 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.

5.4.2 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

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

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 $\pi(y) = \int P(x,y)\,\pi(dx)$.

Checking time reversibility for the Metropolis-Hastings algorithm, is straightforward. The transition function (or conditional probability of moving from x to y in the Metropolis-Hastings algorithm is

$$P(x,y) = \alpha(x,y)\,Q(x,y) + (1 - r(x))\,\delta_x(y) \tag{13}$$

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 (13) is similar. Thus Metropolis-Hastings algorithms generate Markov Chains that are time-reversible and have the target distribution as an invariant distribution.

Verifying $\pi-$irreducibility is always straightforward (see Roberts and Polson (1994) for the Gibbs samplers and Roberts and Smith (1994) and Robert and Casella (1999) for Metropolis-Hastings). One sufficient condition (see Mengerson and Tweedie (1996)) is that $\pi(y) > 0$ implies that $Q(x,y) > 0$. 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 on can eventually reach state y (see Roberts and Polson (1996)). To verify aperiodicity,

we 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 the Metropolis-Hastings algorithm converges, thus the chain is ergodic.

Having discusses this result, it is important to note that we are rarely solely interested in convergence of the Markov chain. In practice (see the first section in this section) 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 $\frac{1}{G}\sum_{g=1}^{G}\theta^{(g)}$. Note that there are in fact two subtle forms of convergence in the sample average: first the convergence of the chain, and second the convergence of the sample average. Fortunately, the following result guarantees convergence:

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 $\Theta^{(g)}$*

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

almost surely.

We can in fact go a bit further with an ergodic CLT:

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 $\sigma\left(f\right)$ such that*

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

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

In spite of these limiting results, we are typically interested in the speed of convergence of the chain. 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 (1996) 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. Mengerson 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 (see also Roberts and Tweedie (1992) and Tierney (1994)). As shown by Polson (1996), careful use of data augmentation can also aid in improving convergence. For example, the data augmentation in Swendsen and Wang (1987) is a useful alternative to direct Metropolis. Mengerson and Tweedie (1996) show that *random walk* algorithms converge at a geometric rate if the target density has geometric tails. Jarner and Roberts (2001) discuss Metropolis convergence rates in the case the target has polynomial tails.

There are a number of caveats in order. First, these results are limiting theorems and therefore do not guarantee provable convergence. Provable convergence, on the other, guarantees a pre-specified accuracy with arbitrarily high probability. For further discussion of this line of research, see Diaconis and Stroock (1991), Frieze, Kannan and Polson (1994), Polson (1996) and Rosenthal (1995, 1996). For example, Frieze, Kannan and Polson (1994) show that MCMC algorithms for log-concave densities are converge in polynomial time.

Second, in addition to the formal convergence theory, there is a large literature that studies the information content of sequence $\left\{\Theta^{(g)}\right\}_{g=1}^{G}$. While theory is clear that the chains converge, it is impossible to formally diagnose convergence from the realized output of the chain. 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. Popular observed-chain based diagnostics include calculating parameter trace plots (plots of $\Theta_i^{(g)}$ as a function of g), analyzing the correlation structure of draws and Monte Carlo estimates for the standard errors of $\frac{1}{G}\sum_{g=1}^{G} f(\Theta^{(g)})$ (see Mengerson, et al (1996) and Robert and Casella (1999)). 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 (1991), 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. Also notice that Gelfand and Smith (1990) to also estimate conditional

and marginal posterior distributions using Rao-Blackwellized averages across the chain.

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 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.

6 MCMC Methods: Practical Recommendations

While the theory behind MCMC algorithms is very clear, there are number of issues that arise in practice that must be addressed. In this section we provide a list of practical recommendations for researchers in order to avoid common errors.

First, one must be careful when using non-informative priors. Without care, conditional or joint posteriors can be improper violating the Hammersley-Clifford 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, we recommend, as discussed earlier, careful simulation studies. We recommend that proper priors, typically diffuse, always be used unless there is a very strong justification for doing otherwise.

Second, provable convergence rates are always important. This implies that algorithms that are provably geometric convergent are preferred to those that are not. This implies that one should be careful in using normal proposal densities when the target has known fat tails. If possible, using independence-Metropolis, one should find a proposal which bounds the tails of the target density. Similarly, one should be careful using random-walk Metropolis with normal proposals and should, alternatively, use fat tailed distributions such as a t-distribution (see,

e.g., Tierney (1994) and Mengerson and Tweedie (1996)).

Third, due to modular nature of MCMC algorithms, we recommend building the algorithms bottom-up. 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.

Fourth, careful blocking of parameters and use of latent variables can improve the convergence properties. As shown by Kong, Liu and Wong (1993), drawing correlated parameters in blocks can improve the speed of convergence. Also, as shown by Polson (1996), the introduction of additional latent state variables (data augmentation) can also 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.

Fifth, 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 two 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. For example, see Johannes, Kumar and Polson (1998) and Eraker, Johannes and Polson (2000) who perform simulation studies on models with jumps and stochastic volatility.

Finally, careful examination of parameter and state variable trace plots (as a function of G) provides the information content of the Markov Chain and, specifically, Monte Carlo standard errors. Whenever the convergence of the MCMC algorithm is in question, careful simulation studies can provide reassurance that the MCMC algorithm is providing reliable inference.

7 MCMC Inference in Equity Price Models

This section explicitly derives MCMC algorithms for a number of different equity price models that are often used in applied financial modeling. Each model contains a novel factor, and, for simplicity, we focus solely on that factor. For example, we analyze jumps, stochastic volatility

and time-varying expected returns in the absence of the other factors. This allows us to isolate the impact of each of the factors.

The advantage of MCMC is that one can piece algorithms together: to estimate a model with jumps and stochastic volatility, the researcher needs to just cut and paste different portions of the algorithms together, typically with minimal adjustments required.

7.1 Geometric Brownian Motion

To understand MCMC methods, we first review the simplest possible case, estimation of a geometric Brownian motion model of an asset price, S_t. The price solves the familiar SDE

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

which has a closed form solution for log-returns ($Y_t = \log(S_t/S_{t-1})$):

$$Y_t = \mu + \sigma(W_t - W_{t-1})$$

where, for notational ease, we have redefined the drift to avoid explicitly accounting for the variance correction. This model can be easily estimated by maximum likelihood, but we consider MCMC estimation to develop the intuition of MCMC algorithms.

MCMC estimation provides samples from the distribution of the parameters given the observed return data

$$p(\Theta|Y) = p\left(\mu, \sigma^2|Y\right)$$

where $Y = [Y_1, ..., Y_T]$ is a vector containing the time series of continuously compounded returns. The Hammersley-Clifford theorem implies that knowledge of $p(\mu|\sigma^2, Y)$ and $p(\sigma^2|\mu, Y)$ fully characterizes the joint distribution $p(\mu, \sigma^2|Y)$. The MCMC algorithm therefore just iteratively draws from these distributions. Given $\mu^{(g)}$ and $(\sigma^2)^{(g)}$, the algorithm updates sequentially by drawing

$$\mu^{(g+1)} \sim p\left(\mu|\left(\sigma^2\right)^{(g)}, Y\right) \text{ and } \left(\sigma^2\right)^{(g+1)} \sim p\left(\sigma^2|\mu^{(g+1)}, Y\right).$$

What are the densities $p(\mu|\sigma^2, Y)$ and $p(\sigma^2|\mu, Y)$? Assuming the priors on the parameters

are independent,[3] Bayes rule implies that

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

where the likelihood function is given by

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

and $p\left(\mu\right)$ and $p\left(\sigma^2\right)$ are prior distributions. If prior on μ is normal and the prior on the σ^2 is inverse Gamma, the posteriors are conjugate, that is $p\left(\mu|\sigma^2, Y\right)$ is normal and $p\left(\sigma^2|\mu, Y\right)$ is inverse Gamma.

This previews the general approach to MCMC estimation: (1) write out the state space model and the posterior; (2) using the Hammersley-Clifford theorem, characterize a joint distribution by its complete set of conditionals; (3) using Bayes rule, write out the conditional distributions of the parameters and state variables and (4) using standard methods for drawing random variables, sequentially update the parameters and the state variables.

7.2 Option Pricing

Consider an extension of the previous section to include equity option prices. This is a special case of Eraker (2002). The price of a call option struck at K is given by the familiar Black-Scholes formula

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

where

$$d_1 = \frac{\log\left(S_t/K\right) + \left(r + 0.5\sigma^2\right)\left(T - t\right)}{\sigma\sqrt{T-t}}.$$

For notational simplicity, we have suppressed the dependency of K, r and $T-t$ in $BS\left(\sigma, S_t\right)$.

[3] Alternatively, one could use dependent conditional conjugate priors such as $p\left(\mu, \sigma^2\right) = p\left(\mu|\sigma^2\right) p\left(\sigma^2\right)$ where $p\left(\mu|\sigma^2\right) \sim N\left(a, \sigma^2 A\right)$ and $p\left(\sigma^2\right) \sim IG\left(b, B\right)$. This leads to closed form conditional posteriors for $p\left(\sigma^2|Y\right)$ and $p\left(\mu|\sigma^2, Y\right)$ and also allows for a direct block update from $p\left(\mu, \sigma^2|Y\right)$.

Assuming the option prices are observed with a normally distributed error, the state space is

$$\begin{aligned} Y_t &= \mu + \sigma\left(W_t - W_{t-1}\right) \\ C_t &= BS\left(\sigma, S_t\right) + \varepsilon_t^c. \end{aligned}$$

As mentioned earlier, there is a strong justification for using a pricing error in the option equation as at-the-money index equity options have a bid-ask spread of around 10% of the contract value. The econometrician observes the time series of equity prices, $S = S_1, ..., S_T$, option prices, $C = [C_1, .., C_T]$, and continuously compounded returns, Y, as defined earlier. The joint posterior is $p\left(\mu, \sigma^2 | S, C, Y\right)$ and Hammersley-Clifford implies that $p\left(\mu|\sigma^2, S, C, Y\right)$ and $p\left(\sigma^2|\mu, S, C, Y\right)$ fully characterize the joint posterior. Since the option price does not depend on μ, $p\left(\mu|\sigma^2, S, C, Y\right) = p\left(\mu|\sigma^2, Y\right)$ and is normal provided the prior distribution of μ is normal as in the previous section.

Updating volatility is slightly more difficult as both the option price and the equity returns contain information about this parameter. More specifically, we have that

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

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

$$p\left(C|\sigma^2, S\right) \propto \prod_{t-1}^{T} p\left(C_t|\sigma^2, S_t\right) \propto \prod_{t-1}^{T} \exp\left(-\frac{1}{2}\left(\frac{C_t - BS\left(\sigma, S_t\right)}{\sigma^c}\right)^2\right)$$

We consider a independence Metropolis algorithm to update this parameter. The algorithm proposes using data from the returns and then accepts/rejects based on the information contained in the option prices. Specifically, proposing from:

$$q\left(\sigma^2\right) = p\left(\sigma^2|\mu, Y\right) \propto p\left(Y|\mu, \sigma^2\right) p\left(\sigma^2\right).$$

If we assume the prior on σ^2 is inverted gamma, the proposal is inverted Gamma. This implies the following Metropolis step:

$$\text{Step 1} \quad : \quad \text{Draw } \left(\sigma^2\right)^{(g+1)} \text{ from } q\left(\sigma^2\right) \sim IG \tag{14}$$

$$\text{Step 2} \quad : \quad \text{Accept } \left(\sigma^2\right)^{(g+1)} \text{ with probability } \alpha\left(\left(\sigma^2\right)^{(g+1)}, \left(\sigma^2\right)^{(g)}\right) \tag{15}$$

where

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

Proceeding, this algorithm will provide samples from the joint posterior. This example clearly shows the power of the Metropolis algorithm: the conditional density of $p\left(\sigma^2|\mu, Y, S, C\right)$ can be evaluated but cannot be directly sampled from due to the nonlinear nature in which σ^2 enters the Black-Scholes option price. Moreover, since the Black-Scholes price is always bounded by the underlying price, $BS\left(\sigma, S_t\right) \leq S_t$, the tail behavior of $\pi\left(\sigma^2\right)$ is determined by the likelihood piece and the algorithm is geometrically convergent.

7.3 Multivariate Jump-Diffusion Models

A multivariate jump-diffusion model provides an excellent example of data augmentation and the Gibbs sampler. Consider a multivariate version of Merton's jump diffusion

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

where $\sigma\sigma' = \Sigma \in R^K \times R^K$ is the diffusion matrix, N_t is a Poisson process with constant intensity λ and the jump sizes, $\xi_j \in R^K$ are multivariate normal with mean μ_J and variance-covariance matrix Σ_J. Solving this stochastic differential equation, continuously compounded equity returns (Y_t) over a daily interval $(\Delta = 1)$ are

$$Y_{t+1} = \mu + \sigma\left(W_{t+1} - W_t\right) + \sum_{j=N_t}^{N_{t+1}} \xi_j$$

where, again, we have redefined the drift vector to account for the variance correction.

We consider a time-discretization of this model which implies that at most a single jump can occur over each time interval:

$$Y_{t+1} = \mu + \sigma\left(W_{t+1} - W_t\right) + J_{t+1}\xi_{t+1}$$

where $P[J_t = 1] = \lambda \in (0,1)$ and the jumps 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. The parameters and state variable vectors are given by

$$\begin{aligned} \Theta &= \{\mu, \Sigma, \lambda, \mu_J, \Sigma_J\} \\ X &= \{J_t, \xi_t\}_{t=1}^T . \end{aligned}$$

Our MCMC algorithm samples from $p(\Theta, X|Y) = p(\Theta, J, \xi|Y)$ where J and ξ are vectors containing the time series of jump times and sizes.

Our MCMC algorithm draws Θ, ξ and J sequentially. Each of posterior conditionals are standard distributions that can easily be sampled from, and thus the algorithm is a Gibbs sampler. This occurs because the augmented likelihood function

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

where

$$p(Y_t|\Theta, J_t, \xi_t) = N(\mu + \xi_t J_t, \Sigma)$$

which is conditionally Gaussian. On the other hand, the observed likelihood, $p(Y_t|\Theta)$, is difficult to deal with because it is a mixture of multivariate normal distributions. In the univariate case, the observed likelihood has degeneracies (for certain parameter values, the likelihood is infinite). There are also well-known multimodalities. Multivariate mixtures are even more complicated and direct maximum likelihood is rarely attempted.

Assuming standard conjugate prior distributions for the parameters,

$$\begin{aligned} \mu &\sim \mathcal{N}(a, A), \ \Sigma \sim \mathcal{W}^{-1}(b, B) \\ \mu_J &\sim \mathcal{N}(c, C), \Sigma_J \sim \mathcal{W}^{-1}(d, D) \\ \lambda &\sim \mathcal{B}(e, E) . \end{aligned}$$

where $\mathcal{W}^{-1}$ is an inverted Wishart (multivariate inverted gamma) and $\mathcal{B}$ is the beta distribu-

tion, our MCMC algorithm iteratively draws the parameters and the state variables:

$$\begin{aligned}
\text{Diffusive Parameters} &: p(\mu|\Sigma, J, \xi, Y) \propto N(a^*, A^*) \\
&: p(\Sigma|\mu, J, \xi, Y) \propto \mathcal{W}^{-1}(b^*, B^*) \\
\text{Jump Size Parameters} &: p(\mu_J|\Sigma_J, J, \xi) \propto N(c^*, C^*) \\
&: p\left(\Sigma_J|\mu_\xi, J, \xi\right) \propto \mathcal{W}^{-1}(d^*, D^*) \\
\text{Jump Time Parameters} &: p(\lambda|J) \propto \mathcal{B}(e^*, E^*) \\
\text{Jump Sizes} &: p(\xi_t|\Theta, J_t, Y_t) \propto N(m_t^*, V_t^*) \\
\text{Jump Times} &: p(J_t|\Theta, \xi_t, Y_t) \propto Binomial(\lambda_t^*)
\end{aligned}$$

All of these conditional posteriors are easy to derive. In this model the augmented likelihood function is

$$p(Y|J, \xi, \Theta) \propto |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}(y_t - \mu - \xi_t J_t)' \Sigma^{-1}(y_t - \mu - \xi_t J_t)\right\}.$$

By Bayes rule, the posterior for the μ is given by

$$p(\mu|\Sigma, J, \xi, Y) \propto p(\mu|\Sigma, J, \xi, Y) \propto p(Y|\mu, \Sigma, J, \xi)\, p(\mu).$$

Since the likelihood, $p(Y|\mu, \Sigma, J, \xi)$, and the prior, $p(\mu)$, are both normal, a standard calculation (completing the square) implies that $p(\mu|\Sigma, J, \xi, Y) \propto N(a^*, A^*)$, where a^* and A^* are defined in Appendix A. Similarly, the conditional posterior for the variance is given by:

$$p(\Sigma|\mu, J, \xi, Y) \propto p(Y|\mu, \Sigma, J, \xi)\, p(\Sigma)$$

which standard calculations indicate is given by $p(\Sigma|\mu, J, \xi, Y) \propto W^{-1}(b^*, B^*)$ where again b^* and B^* are defined Appendix A. Similar arguments show that the conditional distribution of the jump mean, μ_ξ, and the jump variance, Σ_J, are also normal and inverse Wishart.

Since the jump intensity λ is between zero and one, we use a beta prior, $p(\lambda) = \mathcal{B}(e, E) \propto \lambda^{e-1}(1-\lambda)^{E-1}$. Conditional on the jump times, the posterior for the arrival intensity is independent of all other parameters and jump sizes which implies that the conditional posterior is

$$\begin{aligned}
p(\lambda|J) &\propto p(J|\lambda)\, p(\lambda) \propto \left[(\lambda)^{\sum_{t=1}^{T} J_t}(1-\lambda)^{T-\sum_{t=1}^{T} J_t}\right]\lambda^{e-1}(1-\lambda)^{E-1} \\
&\propto \mathcal{B}(e^*, E^*)
\end{aligned}$$

where $e^* = \sum_{t=1}^{T} J_t + e$ and $E^* = T - \sum_{t=1}^{T} J_t + E$.

The last step in the Gibbs sampler is to update the latent variables: $p(\xi|\Theta, J, Y)$ and $p(J|\Theta, \xi, Y)$. The conditional posterior for ξ_t is given by:

$$\begin{aligned} p(\xi_t|r_t, J_t, \Theta) &\propto \exp\left(-\frac{1}{2}\left[(y_t - \mu - \xi_t J_t)' \Sigma^{-1} (y_t - \mu - \xi_t J_t) + (\xi_t - \mu_\xi)' \Sigma_J^{-1} (\xi_t - \mu_\xi)\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[(\xi_t - m_t^*)' V_t^{-1} (\xi_t - m_t^*)\right]\right) \end{aligned}$$

where

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

The jump size update is simply a normal draw. For the jump times, since J_t can only take two values, the posterior is Bernoulli. The conditional posterior probability that $J_t = 1$ is

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

Computing $p(J_t = 0|\Theta, \xi_t, Y_t)$ then provides the Bernoulli probability. This completes the specification of our MCMC algorithm.

Iteratively drawing from these distributions generates a the following Markov Chain

$$\left\{\mu^{(g)}, \mu_J^{(g)}, \Sigma^{(g)}, \Sigma_J^{(g)}, \lambda^{(g)}, J^{(g)}, \xi^{(g)}\right\}_{g=1}^{G}$$

The arguments in Rosenthal (1995*a*,*b*) show that the algorithm is in fact polynomial time convergent, and thus, converges quickly.

To illustrate our methodology, we consider a simple 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

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

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

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

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. In choosing the prior parameters, we were informative only for the jump intensity by specifying that jumps are rare. 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 time-independent. Figure 1 provides parameter trace plots and shows how, after burn-in, Gibbs sampler moves around the posterior distribution.

Figure 2 provides Monte Carlo estimates of the jump sizes in returns ($\xi_t J_t$). 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 in the next section by introducing time-varying and stochastic volatility.

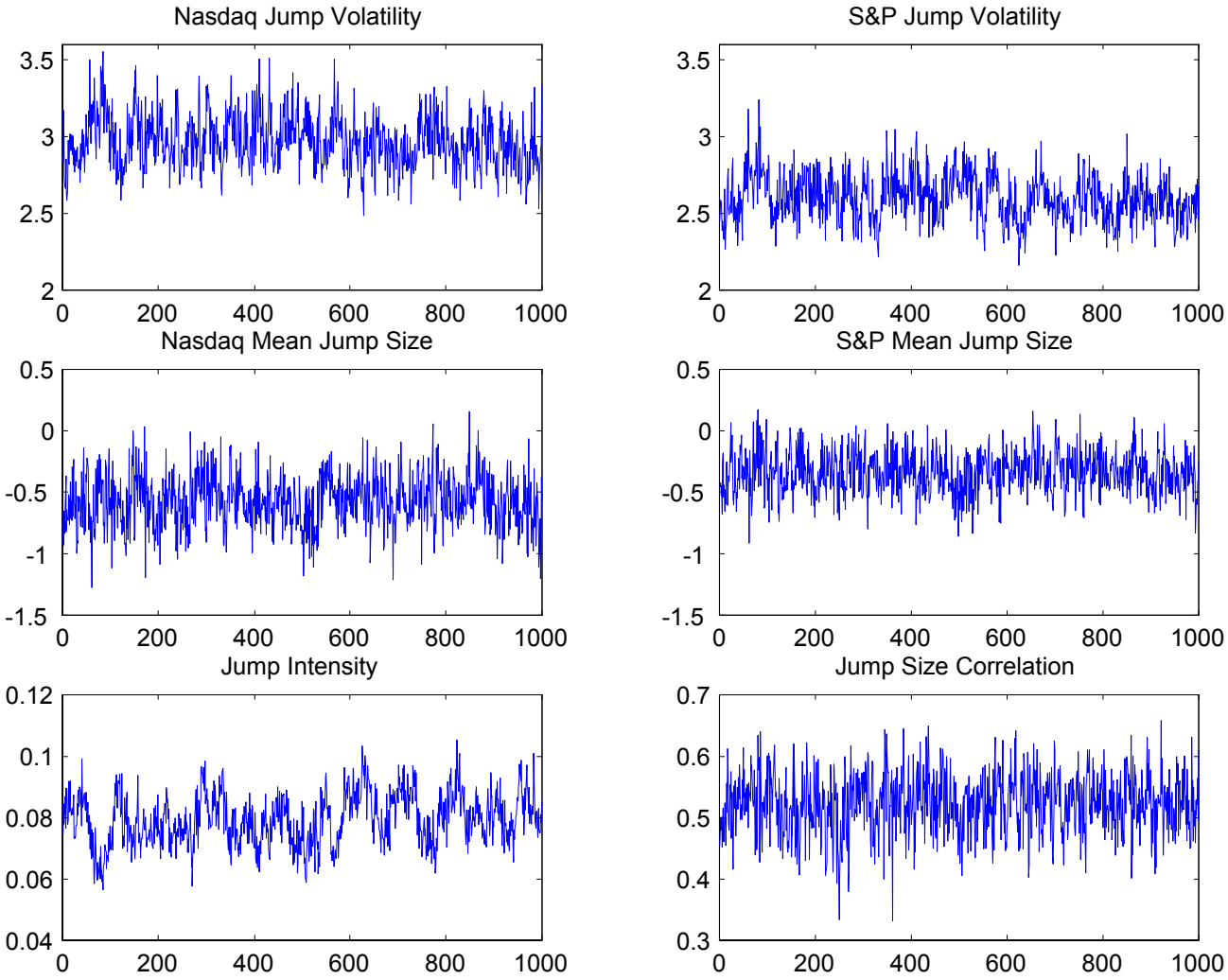

Figure 1: Parameter trace plots for the jump parameters. Each panel shows $\left\{\Theta^{(g)}\right\}_{g=1}^{G}$ for the individual parameters.

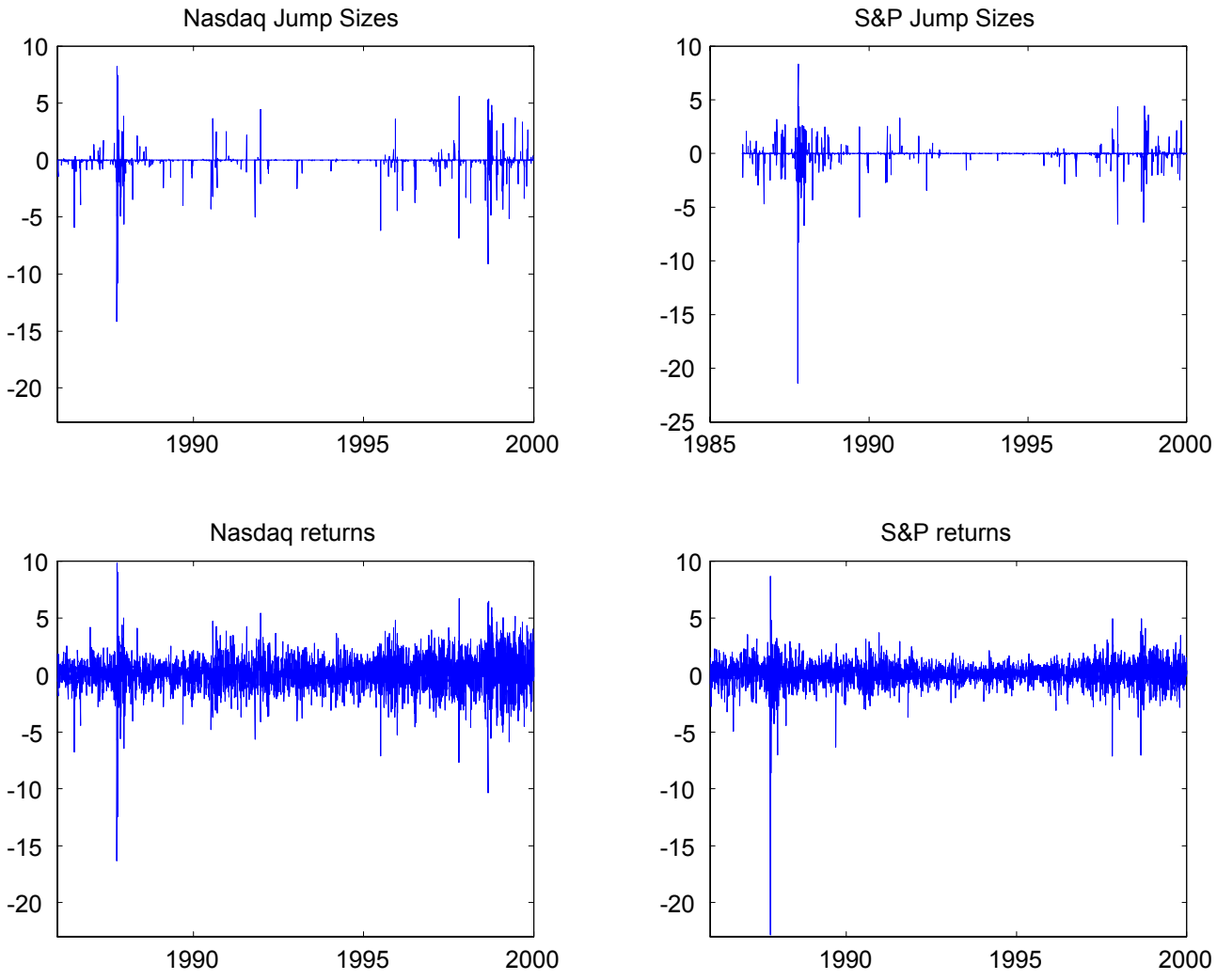

Figure 2: Estimated jump sizes in returns for the Nasdaq and S&P 500 and actual returns over the same period.

7.4 Stochastic Volatility Models

Consider a continuous-time log-stochastic volatility model for equity returns:

$$\begin{aligned} Y_{t+1} &= \int_t^{t+1} \mu_s ds + \int_t^{t+1} \sqrt{V_s} dW_s^s \\ \log(V_{t+1}) &= \int_t^{t+1} \kappa_v (\theta_v - \log(V_s)) ds + \int_t^{t+1} \sigma_v dW_s^v. \end{aligned}$$

where, for simplicity, we assume that the Brownian motions are independent, although this assumption is easy to relax (Jacquier, Polson and Rossi (2000)). For simplicity, we assume that the drift of the returns equation is zero. The distribution of the continuously-compounded returns is not known in closed form, due to the randomization of the Brownian increment by the volatility process, $\int_t^{t+1} \sqrt{V_s} dW_s^s$.

Consider an Euler-time discretization of the model:

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

where we have reparameterized the model by setting $\alpha_v = \kappa_v \theta_v$ and $\beta_v = 1 - \kappa_v$. This allows us to use standard conjugate updating theory for the parameters. Given the reparameterization, the parameters and state variables are

$$\begin{aligned} \Theta &= \left\{ \alpha_v, \beta_v, \sigma_v^2 \right\} \\ X &= V = \{V_t\}_{t=1}^T \end{aligned}$$

and the posterior distribution is given by $p(\Theta, X|Y) = p\left(\Theta, \{V_t\}_{t=1}^T | Y\right)$.

The Hammersley-Clifford theorem implies that $p\left(\Theta, \{V_t\}_{t=1}^T | Y\right)$ is completely characterized by the following distributions

$$\begin{aligned} \text{Regression Parameters} &: p(\alpha_v, \beta_v | \sigma_v, V, Y) \\ \text{Volatility of volatility} &: p\left(\sigma_v^2 | \alpha_v, \beta_v, V, Y\right) \\ \text{Volatility States} &: p(V | \Theta, Y). \end{aligned}$$

Assuming standard conjugate priors for the regression parameters, $p(\alpha_v, \beta_v) \sim N(a, A)$, and the volatility of volatility parameter, $p(\sigma_v^2) \sim IG(b, B)$, the parameter draws are all conjugate and thus are Gibbs steps in the algorithm.

To see this, note that, conditional on volatility, the volatility parameters are just regression parameters

$$\log(V_{t+1}) = \alpha_v + \beta_v \log(V_t) + \sigma_v \varepsilon^v_{t+1}.$$

Specifically, the conditional posterior for the volatility regression parameters is

$$\begin{aligned} p(\alpha_v, \beta_v | \sigma_v, V, Y) &\propto \prod_{t=1}^{T} p(V_t | V_{t-1}, \Theta)\, p(\alpha_v, \beta_v) \\ &\propto N(a^*, A^*) \end{aligned}$$

which follows from standard regression theory where

$$\begin{aligned} (A^*)^{-1} &= (A)^{-1} + \sigma_v^{-2} (W'W)^{-1}, \\ a^* &= \sigma_v^{-2} W' V^*, \\ W &= [1, \log(V_1), ..., \log(V_{T-1})] \end{aligned}$$

and V^* is the vector $[\log(V_2), ..., \log(V_T)]$. For σ_v, a straightforward calculation shows that

$$\begin{aligned} p\left(\sigma_v^2 | \alpha_v, \beta_v, V, Y\right) &\propto \prod_{t=1}^{T} p(V_t | V_{t-1}, \alpha_v, \beta_v, \sigma_v)\, p\left(\sigma_v^2\right) \\ &\sim IG\left(T + b, B + \Sigma_{t=1}^{T} e_t(V_t)^2\right). \end{aligned}$$

where $e_t(V_t) = \log(V_t) - \alpha_v - \beta_v \log(V_{t-1})$.

The only difficult step arises in updating the volatility states. We first consider "single state" updating as the joint volatility posterior, $p(V|\Theta, Y)$, cannot directly drawn from without approximations. Jacquier, Polson and Rossi (JPR) (1994) first considered this model and there are now numerous other ways of updating volatility, some of which we will describe below. The full joint posterior for volatility is

$$\begin{aligned} p(V|\Theta, Y) &\propto p(Y|\Theta, V)\, p(V|\Theta) \propto \prod_{t=1}^{T} p(V_t | V_{t-1}, V_{t+1}, \Theta, Y) \\ &\propto \prod_{t=1}^{T} p(y_t | V_t, \Theta)\, p(V_{t+1} | V_t, \Theta)\, p(V_t | V_{t-1}, \Theta). \end{aligned}$$

For a single state update, we use the fact that

$$\begin{aligned} & p\left(y_t|V_t,\Theta\right)p\left(V_t|V_{t-1},\Theta\right)p\left(V_{t+1}|V_t,\Theta\right) \\ \propto \;\; & V_t^{-\frac{1}{2}}\exp\left(-\frac{Y_t^2}{2V_t}\right)\exp\left(-\frac{e_t\left(V_t\right)^2}{2\sigma_v^2}\right)V_t^{-1}\exp\left(-\frac{e_{t+1}\left(V_{t+1}\right)^2}{2\sigma_v^2}\right) \end{aligned}$$

As this distribution is not recognizable, we use a Metropolis-Hastings algorithm to sample from it.

Following JPR (1994), we use an independence Metropolis-Hastings algorithm with an inverse 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 $q\left(V_t\right)$ and the true conditional density as $\pi(V_t) \triangleq p\left(V_t|V_{t-1},V_{t+1},\Theta,Y\right)$, this implies the Metropolis-Hastings step is given by:

1. Draw $V_t^{(g+1)}$ from $q\left(V_t\right)$
2. Accept $V_t^{(g+1)}$ with probability $\alpha\left(V_t^{(g+1)},V_t^{(g)}\right)$

where

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

Iterating between the parameter and volatility updates, the algorithm generates

$$\left\{\Theta^{(g)},V^{(g)}\right\}_{g=1}^{G}=\left\{\alpha_v^{(g)},\beta_v^{(g)},\left(\sigma_v^2\right)^{(g)},V^{(g)}\right\}_{g=1}^{G}.$$

Given that the gamma distribution bounds the tails of the true conditional density, the algorithm is geometrically convergent.

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

Since the volatility states are correlated, one would ideally like to update them in a block. Unfortunately, direct block updating is 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. The approximation uses the fact that the distribution of $log\left(\left(\varepsilon_{t+1}^s\right)^2\right)$ can be approximated

by a mixture of normals (see Carter and Kohn (1994) and Kim and Shephard (1995)). The advantage of this is that volatility can be drawn in a block and the disadvantage is that it approximates the true model and drastically increases the state space by indicator variables (see Rosenthal (1995) for a discussion of convergence problems with discrete state spaces).

Other volatility specifications are easy to deal with. Eraker, Johannes and Polson (2000) consider a square-root stochastic volatility model and also add Poisson driven, exponentially distributed jumps in volatility

$$dV_t = (\alpha + \beta V_t)\, dt + \sigma_v \sqrt{V_t} dW_t^v + d\left(\sum_{j=0}^{N_t^v} \xi_j^v\right)$$

The only difficulty in estimating this model, like the log-volatility case, is updating volatility. Eraker, et al (2000) use a random walk Metropolis-Hastings algorithm and provide a simulation study to evaluate the ability of the MCMC algorithm to estimate the parameters of the underlying continuous-time process. By simulating off of the continuous-time model and estimating with discretely sampled data, they find that the MCMC algorithm provides reliable inference on the volatility parameters, even in the presence of jumps in returns or jumps in volatility. Incorporating jumps is also straightforward by applying the results of the previous section. Jones (2000) uses MCMC methods to estimate a call of volatility specifications with constant-elasticity:

$$dV_t = (\alpha + \beta V_t)\, dt + \sigma_v V_t^{\gamma} dW_t^v$$

and finds evidence that $\gamma > 1/2$.

7.5 Time-Varying Expected Returns

For portfolio applications, a number of authors, beginning with Merton (1971) consider a continuous-time model of time-varying expected returns:

$$\begin{pmatrix} d\log(S_t) \\ d\mu_t \end{pmatrix} = \begin{pmatrix} \mu_t \\ \kappa_\mu(\theta_\mu - \mu_t) \end{pmatrix} dt + \begin{pmatrix} \sigma dW_t^s \\ \sigma_\mu dW_t^\mu \end{pmatrix}.$$

where, again, for simplicity we assume the Brownian motions are uncorrelated and we abstract from stochastic volatility. Time-discretized versions of this model (with various extensions)

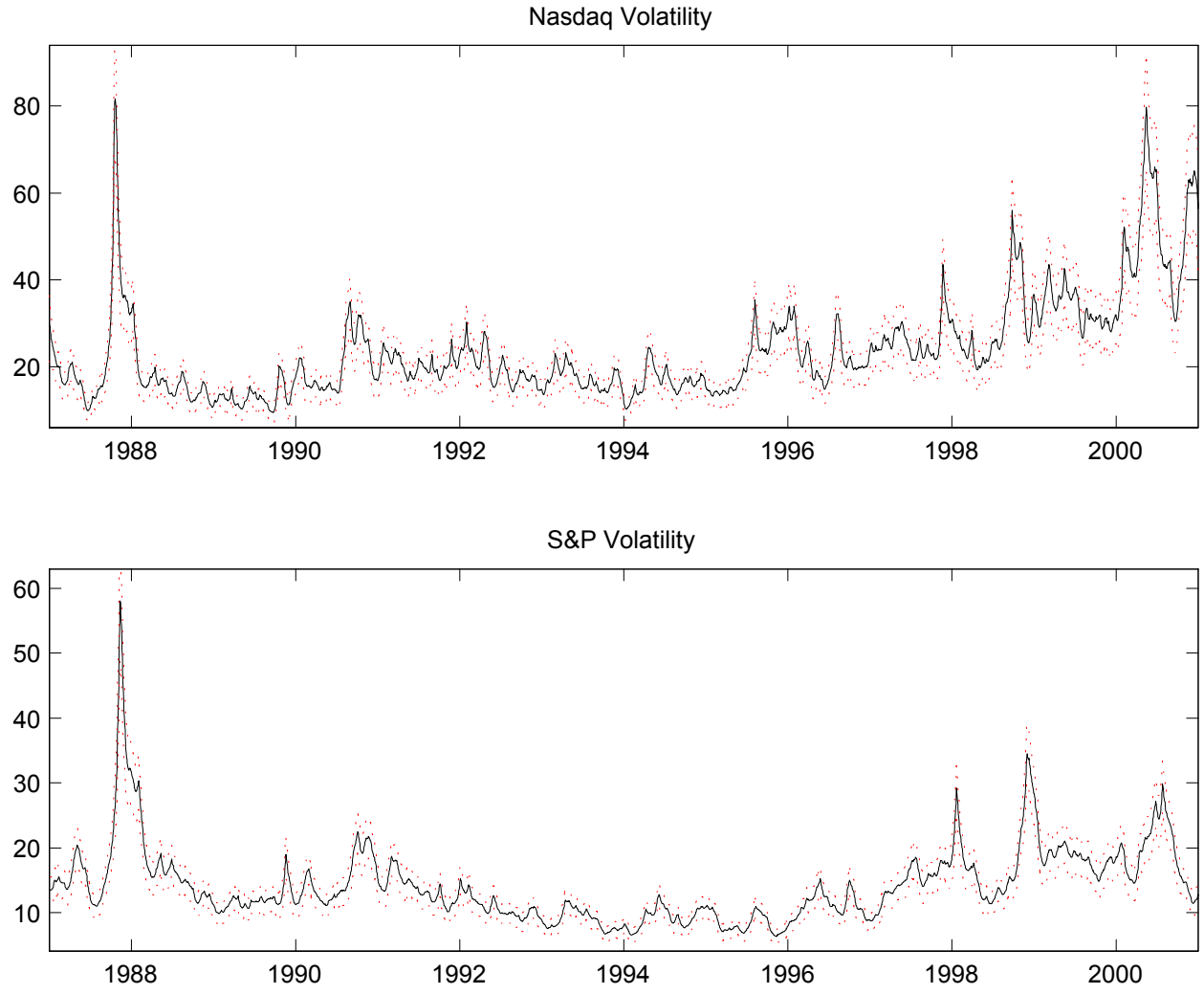

Figure 3: Smoothed volatility paths (with confidence bands) for the S&P 500 and Nasdaq 100 from 1987-2001.

have recently been examined by Brandt and Kang (2000) and Johannes, Polson and Stroud (2001) and are given by:

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

where we again have redefined the parameters. The general form of the model is therefore

$$\begin{aligned} \Theta &= \left\{\alpha_\mu, \beta_\mu, \sigma_\mu^2, \sigma^2\right\} \\ X &= \mu = \{\mu_t\}_{t=1}^T \end{aligned}$$

and the posterior distribution is $p(\Theta, \mu|Y)$ where $\mu = \{\mu_t\}_{t=1}^T$.

Our MCMC algorithm will alternate by updating the parameters and the states:

$$\begin{aligned} \text{Regression Parameters} &: \alpha_\mu, \beta_\mu|\mu, \sigma_\mu^2, \sigma^2, Y \\ \text{Variance parameters} &: \sigma_\mu^2|\mu, \alpha_\mu, \beta_\mu, \sigma^2, Y \\ &: \sigma^2|\mu, \alpha_\mu, \beta_\mu, \sigma_\mu^2, Y \\ \text{State Variables} &: \mu|\alpha_\mu, \beta_\mu, \sigma_\mu^2, \sigma^2, Y. \end{aligned}$$

The parameter posteriors follow immediately from normal regression theory and are thus omitted. The regression parameters $\left[\alpha_\mu, \beta_\mu\right]$ have a normal conditional posterior and the volatility parameters, σ_μ and σ, both have inverted Gamma distributions.

Drawing from $\mu|Y, \alpha, \beta, \sigma_v^2$ might at first glance seem to be difficult because it is a high-dimensional problem, however, we can use the Kalman filter to obtain this density via the forward-filtering backward sampling (FFBS) algorithm described in Carter and Kohn (1993). The mechanics of this are quite simple.

Consider the following decomposition of the joint expected returns posterior:

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

where $Y^t = [Y_1, ..., Y_t]$. To simulate from this, consider the following procedure:

1. Run the Kalman filter to get the moments of $p(\mu_t|Y, \Theta)$
2. Sample the last state from $\widehat{\mu}_T \sim p\left(\mu_T|Y^t, \Theta\right)$
3. Sample backward through time: $\widehat{\mu}_t \sim p\left(\mu_t|\widehat{\mu}_{t+1}, Y^t, \Theta\right)$

Then the samples $(\widehat{\mu}_1, ..., \widehat{\mu}_T)$ are direct block draw from $p(\mu|Y, \Theta)$. It is important to recognize that the Kalman filter is just one step in the algorithm, and the other steps (parameter updating) indicates that we are taking into account parameter uncertainty.

We now provide some empirical results. Using S&P 500 returns and Nasdaq 100 returns from 1973-2000 and 1987-2000, we report estimates of the time series of the latent expected returns over the period from 1987-2000. The results take into account parameter uncertainty, although we do not report parameter estimates. Due to identification concerns, we for simplicity set $\beta_\mu = 0.98$. Figure 3 provides posterior estimates of μ_t, $E[\mu_t|Y]$ and we also include a symmetric one standard deviation confidence interval.

It is also possible to perform filtering, that is, compute $p(\Theta, \mu_t|Y^t)$ for $t = 1, ..., T$ (see Johannes, Polson and Stroud (2001) for further details that are beyond this paper.

8 MCMC Inference in Term Structure Models

Term structure models pose a number of difficult problems for estimation. The first problem is that the parameters enter the state space in a highly nonlinear fashion, often non-analytically. Second, there are often issues of stochastic singularities as researchers typically use low-dimensional factor models to model the yield curve and there are a large number of observable interest rates. To address these issues, we begin with the simplest model, Vasicek's (1977) model, and then proceed to analyze the square-root model of CIR (1985) and then multi-factor models.

8.1 Vasicek's Model

Vasicek's (1997) model assumes that the instantaneous spot rate is a Gaussian diffusion:

$$\begin{aligned} dr_t &= \kappa_r(\theta_r - r_t)\,dt + \sigma_r dW_t^r \\ dr_t &= (a_r - b_r r_t)\,dt + \sigma_r dW_t^r \end{aligned}$$

where we will work with the (a_r, b_r) parameterization of the drift. If we assume the risk premium is constant, the evolution of the spot rate under Q is

$$dr_t = (a_r + \lambda_r\sigma_r - b_r r_t)\,dt + \sigma_r dW_t^r\,(Q)$$

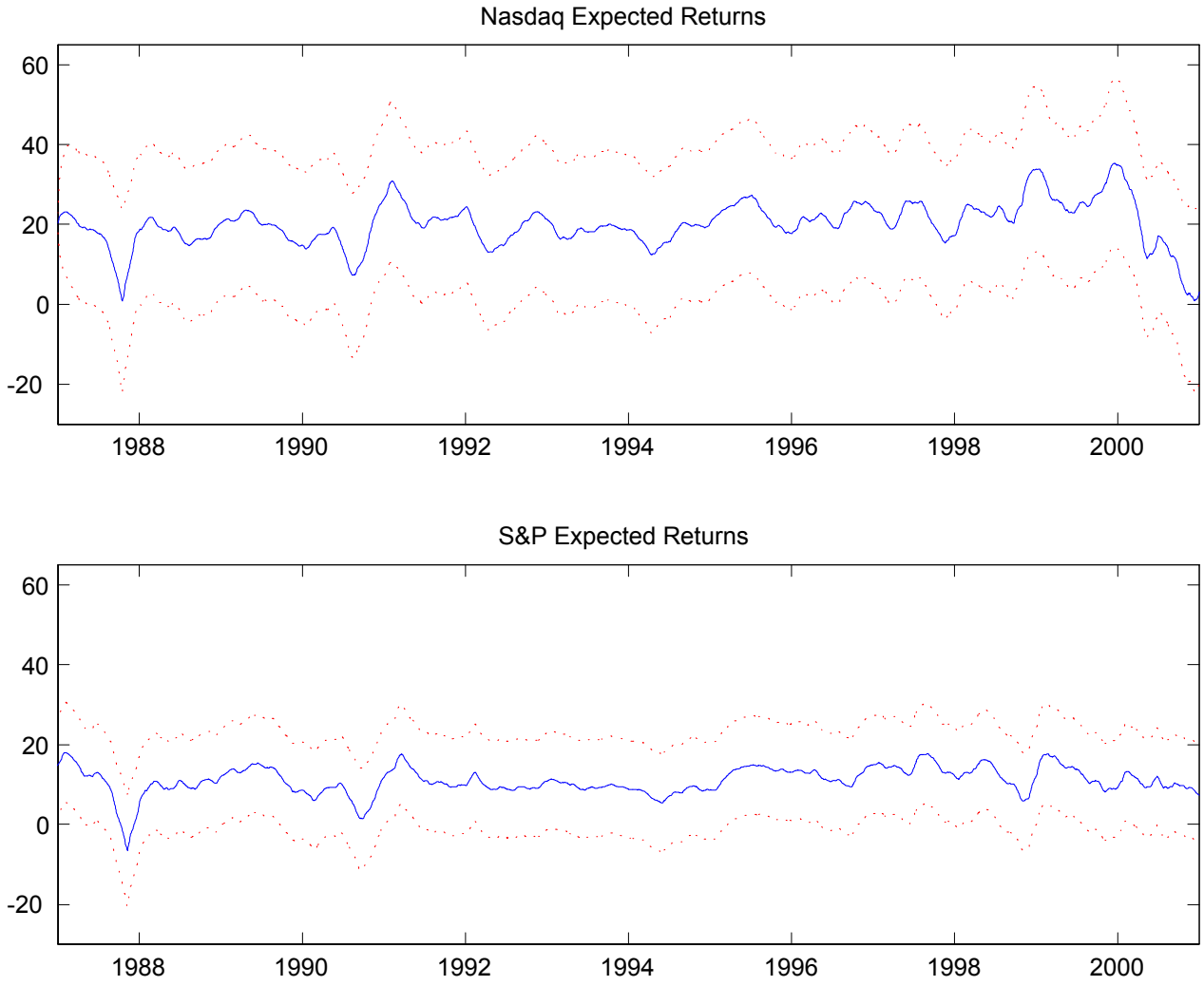

Figure 4: Smoothed expected return paths (with confidence bands) for the S&P 500 and Nasdaq 100 from 1987-2001.

where $W_t^r(Q)$ is a standard Brownian motion under the equivalent martingale measure Q. In this model, zero coupon bond prices have a closed form solution:

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

where the loading functions are given by:

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

The observed data is a panel of yields at maturities $\tau_1,...,\tau_n$, $Y_t = [Y_{t,\tau_1},...,Y_{t,\tau_k}]$, where $y_{t,\tau} = -\log\left(P(\Theta,r_t,\mu_t,\tau)\right)/\tau$.

The state space in this case is given by:

$$\begin{aligned}Y_{t,\tau} &= -\frac{1}{\tau}\beta(\Theta,\tau) - \frac{1}{\tau}\beta^r(\Theta,\tau) r_t + \varepsilon_t\\ r_{t+1} &= a_r + (1-b_r) r_t + \sigma_r \varepsilon_{t+1}^r\end{aligned}$$

where $\varepsilon_{t+1}^r$is a mean zero normal random variable with variance matrix Σ_ε. Since the spot rate evolution is Gaussian, an alternative approach would use the exact transitions for the spot rate: for any Δ

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

This is approach is recommended when Δ is large. For parameters typically estimated from data and for common $\Delta's$ (daily or weekly), the discretization bias is negligible and thus we proceed using a time-discretization.

In this model, the posterior distribution is given by $p(a_r,b_r,\sigma_r,\lambda_r,r|Y)$ where $r = \{r_t\}_{t=1}^T$ and Y is a matrix of observed yields. The main difficulty in sampling the posterior is the nonlinearity of the parameters in the observation equation as the loading functions $\beta(\Theta,\tau)$ and $\beta^r(\Theta,\tau)$ are highly nonlinear functions of the parameters. To see how to handle this

problem, consider the following MCMC algorithm:

$$\begin{aligned}
\text{Drift parameters} &: a_r, b_r|\sigma_r, \lambda_r, \Sigma_\varepsilon, r, Y \\
\text{Volatility parameters} &: \sigma_r^2|a_r, b_r, \lambda_r, \Sigma_\varepsilon, r, Y \\
\text{Pricing error variance} &: \Sigma_\varepsilon|a_r, b_r, \lambda_r, \sigma_r, r, Y \\
\text{Market Price of Risk parameters} &: \lambda_r|a_r, b_r, \sigma_r, \Sigma_\varepsilon, r, Y \\
\text{State variables} &: r|a_r, b_r, \sigma_r, \lambda_r, \Sigma_\varepsilon, Y.
\end{aligned}$$

With the exception of Σ_ε, updating the parameters requires Metropolis algorithms as the conditional posteriors are not recognizable distributions. First, consider the parameters that determine the interest rate drift under the P-measure. The posterior is given by

$$\begin{aligned}
p\left(a_r, b_r|\sigma_r, \lambda_r, \Sigma_\varepsilon, r, Y\right) &\propto p\left(r, Y|a_r, b_r, \sigma_r, \lambda_r, \Sigma_\varepsilon\right) p\left(a_r, b_r\right) \\
&\propto p\left(r|a_r, b_r, \sigma_r\right) p\left(Y|a_r, b_r, \sigma_r, \lambda_r, \Sigma_\varepsilon, r\right) p\left(a_r, b_r\right).
\end{aligned}$$

This suggests the following independence Metropolis algorithm:

1. Draw $\left(a_r^{(g+1)}, b_r^{(g+1)}\right)$ from $\pi\left(a_r, b_r\right) = p\left(r|a_r, b_r, \sigma_r\right) p\left(a_r, b_r\right)$
2. Accept $\left(a_r^{(g+1)}, b_r^{(g+1)}\right)$ with probability $\alpha\left(\left(a_r^{g+1}, b_r^{g+1}\right), \left(a_r^{(g)}, b_r^{(g)}\right)\right)$

where

$$\alpha\left(\left(a_r^{g+1}, b_r^{g+1}\right), \left(a_r^{(g)}, b_r^{(g)}\right)\right) = \min\left(\frac{p\left(Y|\left(a_r^{(g)}, b_r^{(g)}\right), \sigma_r^{(g)}, \lambda_r^{(g)}, \Sigma_\varepsilon^{(g)}, r\right)}{p\left(Y|\left(a_r^{(g+1)}, b_r^{(g+1)}\right), \sigma_r^{(g)}, \lambda_r^{(g)}, \Sigma_\varepsilon^{(g)}, r\right)}, 1\right).$$

This algorithm, intuitively, consists of a proposal that uses the information contained in the state evolution and then accepting/rejecting based on the information in the yields, or full-information likelihood. The parameter σ_r can be handled in a similar manner. Updating Σ_ε is straightforward as it is a conjugate draw.

The market price of risk parameter is, in general, difficult to estimate.[4] As the state space formulation shows, it only enters the state space through $\beta\left(\Theta, \tau\right)$ as it does not affect the evolution of the spot rates under the P-measure. Conditional on the other parameters, λ_r

[4]See, for example, the discussion of identification in the appendix of Dai and Singleton (2001).

enters linearly in $\beta(\Theta, \tau)$. This is not the general case, but it allows for standard conjugate updating.[5]

The last step is drawing the spot rates. Since the Vasicek model is a linear, Gaussian state space model, the spot rates can be updated using the forward-filtering backward sampling (FFBS) algorithm as described in the previous section. Again, this provides a sample that is a direct block draw from $p(r|Y, \Theta)$.

8.2 Cox, Ingersoll and Ross's (1985) square root model

The Vasicek model assumes that the volatility of interest rate increments is constant, which implies that interest rate increments are normally distribution and that 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's (1985) assume the spot rate follows a Feller (1951) square root process

$$\begin{aligned} dr_t &= \kappa_r(\mu_r - r_t)\,dt + \sigma_r\sqrt{r_t}dW_t^r \\ dr_t &= (a_r + b_r r_t)\,dt + \sigma_r\sqrt{r_t}dW_t^r \end{aligned}$$

where W_t^r is a Brownian motion under the objective measure, P. Note that as interest rates fall to zero, the Brownian increment becomes smaller which allows the mean-reverting drift increase spot rates. Under regularity conditions on the parameters and the initial condition, the CIR models insures interest rates are positive.

To value bonds, we assume the market price of interest rate risk is $\lambda_t = \lambda_r\sqrt{r_t}$ which implies that the dynamics of the spot rate under Q-measure are

$$dr_t = (a_r + (b_r + \lambda_r)\,r_t)\,dt + \sigma_r\sqrt{r_t}dW_t^r(Q).$$

Again, the price of a zero coupon, default-free bond maturing at time $t+\tau$ is given by

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

[5]For example, if we assume a market price of risk was linear in the spot rate, $\lambda_t = \lambda_0 + \lambda_1 r_t$, then the proportional term, λ_1, enters nonlinearly in the loading functions. In this case, as discussed below, standard updating is not possible and a Metropolis step is required.

where β and β^r are

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

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

Given an observed panel of yields at maturities $\tau_1,...,\tau_n$, $Y_t=[Y_{t,\tau_1},...,Y_{t,\tau_k}]$ and assuming a time-discretization[6] of the interest rate increments, the state space is given by:

$$\begin{aligned}Y_{t,\tau} &= -\frac{1}{\tau}\beta(\Theta,\tau)-\frac{1}{\tau}\beta^r(\Theta,\tau)r_t+\varepsilon_t \\ r_{t+1} &= a_r+(1-b_r)r_t+\sigma_r\sqrt{r_t}\varepsilon_{t+1}^r.\end{aligned}$$

The state space is still linear and Gaussian in the states, but there is now conditional heteroskedasticity in the evolution equation for the spot rates.

The posterior distribution is given by $p(a_r,b_r,\sigma_r,\lambda_r,r|Y)$ where r and Y are the spot rates and observed yields. The MCMC algorithm we consider is similar to the one in the previous section

$$\begin{aligned}\text{Structural Parameters} &: a_r,b_r,\sigma_r^2|\lambda_r,\Sigma_\varepsilon,r,Y \\ \text{Market Prices of Risk} &: \lambda_r|a_r,b_r,\sigma_r,\Sigma_\varepsilon,r,Y \\ \text{Pricing error variance} &: \Sigma_\varepsilon|a_r,b_r,\lambda_r,\sigma_r,r,Y \\ \text{State variables} &: r|a_r,b_r,\sigma_r,\lambda_r,\Sigma_\varepsilon,Y.\end{aligned}$$

The only change from the Vasicek model is that the functional forms of the conditional posteriors will be different due to the different functional forms in the market price of risk and the evolution equation of the spot rate.

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

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

where $u=cr_te^{-b_r}$, $v=cr_{t+1}$ and $c=\frac{2b_r}{\sigma_r^2(1-e^{-b_r})}$. 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.

Updating the structural parameters of the spot rate evolution, a_r, b_r and σ_r^2, and the pricing error variance, Σ_ε, proceed in the same manner as in the Vasicek model. For the structural parameters, an independence Metropolis algorithm can be used where one proposes from the structural evolution and accepts/rejects based on the yields. Drawing the market price of risk parameters from $p\left(\lambda_r | a_r, b_r, \sigma_r, \Sigma_\varepsilon, r, Y\right)$ is more difficult because the spot rate evolutions do not provide any information about this parameter. To update λ_r, we use a random walk Metropolis algorithm.

$$\text{Step 1} \quad : \quad \text{Draw } \lambda_r^{(g+1)} \text{ from the proposal density } q(\lambda_r^{(g+1)} | \lambda_r^{(g)}) \tag{16}$$

$$\text{Step 2} \quad : \quad \text{Accept } \lambda_r^{(g+1)} \text{ with probability } \alpha\left(\lambda_r^{(g+1)}, \lambda_r^{(g)}\right) \tag{17}$$

where

$$\alpha\left(\Theta_i^{(g+1)}, \Theta_i^{(g)}\right) = \min\left(\frac{p\left(\lambda_r^{(g+1)} | a_r, b_r, \sigma_r, \Sigma_\varepsilon, r, Y\right)}{p\left(\lambda_r^{(g)} | a_r, b_r, \sigma_r, \Sigma_\varepsilon, r, Y\right)}, 1\right).$$

We we recommend the proposal density to fat-tailed, such as a $t-$distribution.

The last step is updating the spot rates, $\{r_t\}_{t=1}^T$. To do this, note that the observation equation is linear in the state variables, but that the state evolution has conditional heteroskedasticity. This poses no problem, as the forward-filtering, backward sampling algorithm can be modified in a straightforward manner to use a heteroskedastic version of the Kalman filter (see Kalman (1960)). This provides a block update for the short rates and completes the MCMC algorithm.

8.3 Vasicek with Jumps

Baz and Das (1996) consider an extension of Vasicek's (1977) model to incorporate jumps in the short rate:

$$dr_t = (a_r - b_r r_{t-})\, dt + \sigma_r dW_t^r + d\left(\sum_{j=1}^{N_t} \xi_{\tau_j}\right)$$

where we assume that N_t is a Poisson process with constant intensity h and the jumps sizes are i.i.d. normal, $\xi_{\tau_j} \sim N\left(\mu_J, \sigma_J^2\right)$. If we assume the diffusive risk premium is constant, that N_t^Q is a Poisson process under Q with constant intensity h^Q and that the jump sizes are

normally distributed under Q, $\xi_{\tau_j}^Q \sim N\left(\mu_J^Q, (\sigma_J^2)^Q\right)$, the evolution of the spot rate under Q is

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

where $W_t^r(Q)$ is a standard Brownian motion under the equivalent martingale measure Q.[7] In this model, zero coupon bond prices have a closed form solution:

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

where the loading functions now solve the system of ordinary differential equations:

$$\begin{aligned} \frac{d\beta^r(\Theta,\tau)}{d\tau} &= \beta^r[a + \lambda_r b_r] + \frac{1}{2}(\sigma_r \beta^r)^2 + h^Q\left[e^{\beta^r \mu_J^Q + \frac{1}{2}\left(\beta^r \sigma_r^Q\right)^2} - 1\right] \\ \frac{d\beta(\Theta,\tau)}{d\tau} &= 1 + \beta^r b \end{aligned}$$

subject to the terminal condition $\beta^r(\Theta, 0) = \beta(\Theta, 0) = 0$.

The observed data is a panel of continuously compounded yields Y_t and the state space in this case is given by:

$$\begin{aligned} Y_{t,\tau} &= -\frac{1}{\tau}\beta(\Theta,\tau) + -\frac{1}{\tau}\beta^r(\Theta,\tau) r_t + \varepsilon_t \\ r_{t+1} &= a_r + (1 - b_r) r_t + \sigma_r \varepsilon_{t+1}^r + \xi_t J_t \end{aligned}$$

where $\varepsilon_{t+1}^r$is a mean zero normal random variable with variance matrix Σ_ε, $\xi_t \sim N(\mu_J, \sigma_J^2)$, and $J_t = 1$ with probability h. Our MCMC algorithm consists of drawing from the following

[7] These are, of course, very restrictive assumptions on the market prices of risk, especially for the jump components.

conditional distributions:

$$\begin{aligned}
\text{Structural Parameters} &: a_r, b_r, \sigma_r^2 | \lambda_r, h, \mu_J, \sigma_J, h^Q, \mu_J^Q, \sigma_J^Q, \Sigma_\varepsilon, r, \xi, J, Y \\
&: \mu_J | a_r, b_r, \sigma_r^2, h, \sigma_J, \lambda_r, h^Q, \mu_J^Q, \sigma_J^Q, \Sigma_\varepsilon, r, \xi, J, Y \\
&: \sigma_J^2 | a_r, b_r, \sigma_r^2, h, \mu_J, \lambda_r, h^Q, \mu_J^Q, \sigma_J^Q, \Sigma_\varepsilon, r, \xi, J, Y \\
&: h | a_r, b_r, \sigma_r^2, \mu_J, \sigma_J, \lambda_r, h^Q, \mu_J^Q, \sigma_J^Q, \Sigma_\varepsilon, r, \xi, J, Y \\
\text{Market Prices of Risk} &: \lambda_r, h^Q, \mu_J^Q, \sigma_J^Q | a_r, b_r, \sigma_r, \Sigma_\varepsilon, r, \xi, J, Y \\
\text{Pricing error variance} &: \Sigma_\varepsilon | a_r, b_r, \lambda_r, \sigma_r, r, \xi, J, Y \\
\text{State variables} &: r | a_r, b_r, \sigma_r, \lambda_r, h, \mu_J, \sigma_J, h^Q, \mu_J^Q, \sigma_J^Q, \Sigma_\varepsilon, \xi, J, Y \\
&: \xi | a_r, b_r, \sigma_r, h, \mu_J, \sigma_J, J, Y \\
&: J | a_r, b_r, \sigma_r, h, \mu_J, \sigma_J, \xi, Y
\end{aligned}$$

Sampling from these distributions is straightforward using the results from the previous sections. The structural parameters $(a_r, b_r, \sigma_r^2, h, \mu_J, \sigma_J)$ can be sampled using the independence Metropolis strategy given in the previous section: proposing from the structural evolution and then accepting/rejecting based on the information in the yields. The market price of risk parameters also require a Metropolis step, typically a random-walk algorithm.

Updating the state variables is also straightforward. Conditional on the jump times and the jumps sizes, we can rewrite the interest rate increments as

$$r_{t+1} = \xi_t J_t + a_r + (1 - b_r) r_t + \sigma_r \varepsilon_{t+1}^r$$

and view the jump times and sizes as regressors. Since they are known, we return to a linear, Gaussian state space model and we can use the forward-filtering backward sampling algorithm of Carter and Kohn (1993) to update the spot rates. The conditional posteriors for the jump times and sizes are straightforward to derive and sample from given the results in Section 7.3.

8.4 Time-Varying Central Tendency

The time-varying central tendency model specifies that the level to which the short rate mean reverts is random:

$$\begin{pmatrix} dr_t \\ d\mu_t \end{pmatrix} = \begin{pmatrix} \kappa_r (\mu_t - r_t) \\ \kappa_\mu (\theta_\mu - \mu_t) \end{pmatrix} dt + \begin{pmatrix} \sigma_r \sqrt{r_t} dW_t^r \\ \sigma_\mu dW_t^\mu \end{pmatrix}$$

where the Brownian motions are uncorrelated. In this case, the affine structure implies that continuously compounded yields are given by:

$$Y_{t,\tau} = \alpha(\tau, \Theta) + \beta^r(\tau, \Theta) r_t + \beta^\mu(\tau, \Theta) \mu_t.$$

To implement the model, we consider adding a normally distributed pricing error (ε_t) on the bond yields and a time-discretization for the state variable evolution. In this case, again, the observed data is typically a vector of continuously-compounded yields,$Y_{t,\tau}$, the state space is

$$\begin{aligned} Y_{t,\tau} &= \alpha(\tau, \Theta) + (\beta^r(\tau, \Theta), \beta^\mu(\tau, \Theta)) \begin{pmatrix} r_t \\ \mu_t \end{pmatrix} + \varepsilon_t \\ \begin{pmatrix} r_{t+1} \\ \mu_{t+1} \end{pmatrix} &= \begin{pmatrix} \kappa_r(\mu_t - r_t) \\ \alpha_\mu + \beta_\mu \mu_t \end{pmatrix} + \begin{pmatrix} \sigma_r \sqrt{r_t} \varepsilon^r_{t+1} \\ \sigma_\mu \varepsilon^\mu_{t+1} \end{pmatrix}. \end{aligned}$$

where $\beta^j(\tau, \Theta) = [\beta^r(\tau_1, \Theta), ..., \beta^r(\tau_n, \Theta)]'$ for $j = r, \mu$.

This model has a linear state evolution and one of the state variables, r_t, is heteroskedastic. A further complication comes from the fact that although the parameters are linear in the state evolution, they appear nonlinearly in the observation equation. In fact, they are not even analytical as the loading functions α, β^r and β^μ solve ordinary differential equations. If the observations are discretely-compounded rates such as Libor or Eurodollar futures, a simple log transformation results in a linear-in-the-state variables model.

We specify conjugate normal priors for κ_r and α_μ, β_μ and inverse Gamma priors for the variance parameters, σ_r^2, σ_μ^2 and σ_ε^2 and consider the following updating scheme:

$$\begin{aligned} \text{Regression parameters} &: \alpha_\mu, \beta_\mu, \kappa_r | r, \mu, \Theta_-, Y \\ \text{Volatility parameters} &: \sigma_\mu^2, \sigma_r^2, \Sigma_\varepsilon | r, \mu, \Theta_-, Y \\ \text{State Variables} &: r, \mu | , \Theta, Y \end{aligned}$$

where $Y = (Y_1, Y_2, ..., Y_T)'$ is the $T \times k$ matrix of observations.

Updating the parameters in this case is not as easy as in the previous cases because the parameters are not only in the state evolution equation, but are also in the loading functions, α, β^r and β^μ. Consider for example, the parameters α_μ, β_μ. The conditional posterior is given

by:

$$\begin{aligned} p\left(\alpha_\mu, \beta_\mu | r, \mu, \Theta_-, Y\right) &\propto p\left(Y, r | \mu, \Theta\right) p\left(\alpha_\mu, \beta_\mu | \mu\right) \\ &\propto p\left(Y | r, \mu, \Theta\right) p\left(r | \mu, \Theta\right) p\left(\alpha_\mu, \beta_\mu | \mu\right) \\ p\left(\alpha_\mu, \beta_\mu | r, \mu, \Theta_-, Y\right) &\propto p\left(Y | r, \mu, \Theta\right) p\left(\alpha_\mu, \beta_\mu | \mu\right) \end{aligned}$$

since as a function of α_μ, β_μ, $p\left(r | \mu, \Theta\right)$ is constant. The augmented likelihood function is given by

$$p\left(Y | r, \mu, \Theta\right) \propto \prod_{t=1}^{T} p\left(y_t | r_t, \mu_t, \Theta\right)$$

where, for $\tau = (\tau_1, \tau_2, ..., \tau_n)$

$$p\left(y_t | r_t, \mu_t, \Theta\right) = N\left(\alpha\left(\tau, \Theta\right) + \beta^r\left(\tau, \Theta\right) r_t + \beta^\mu\left(\tau, \Theta\right) \mu_t, \Sigma_\varepsilon\right).$$

Since the loading functions are not analytical, direct updating from the posterior is not feasible. Instead, we consider a Metropolis-Hastings algorithm where we propose from $p\left(\alpha_\mu, \beta_\mu | \mu\right)$ which is normal. Note that the target

$$\begin{aligned} \pi\left(\alpha_\mu, \beta_\mu | r, \mu, \Theta_-, Y\right) &\propto p\left(\alpha_\mu, \beta_\mu | r, \mu, \Theta_-, Y\right) \\ &\propto p\left(Y | r, \mu, \Theta\right) p\left(\alpha_\mu, \beta_\mu | \mu\right) \end{aligned}$$

and we let the conditional likelihood, as a function of α_μ, β_μ be denoted as $g\left(\alpha_\mu, \beta_\mu\right)$. This leads to the following Metropolis-Hastings algorithm:

1. Draw $\alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}$ from $p\left(\alpha_\mu, \beta_\mu | \mu^{(g+1)}\right)$
2. Accept $\alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}$with probability $\alpha\left(\alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}, \alpha_\mu^{(g)}, \beta_\mu^{(g)}\right)$

where

$$\alpha\left(\alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}, \alpha_\mu^{(g)}, \beta_\mu^{(g)}\right) = \min\left(\frac{g\left(\alpha_\mu^{(g+1)}, \beta_\mu^{(g+1)}\right)}{g\left(\alpha_\mu^{(g)}, \beta_\mu^{(g)}\right)}, 1\right)$$

In order to implement this requires an evaluation of the likelihood at the new parameters which requires us to solve the ODE's given the new parameter draws. Updating the other parameter proceeds in the same way.

We update the latent states together in one block. To see how, we write

$$p\left(r,\mu|\Theta,Y\right)\propto p\left(r_T,\mu_T|\Theta,Y\right)\prod_{t=1}^{T}p\left(r_t,\mu_t|r_{t+1},\mu_{t+1},\Theta,Y^t\right).$$

Since the joint distribution, $p\left(r_t,\mu_t|r_{t+1},\mu_{t+1},\Theta,Y^t\right)$, is Gaussian, we can apply the FFBS algorithm which relies, in this case, on a bivariate Kalman filter. In the case of the $r_t\prime s$ we rely on the heteroskedastic Kalman filter (as the variance of the evolution of the $r_t's$ depends on r_t). The algorithm is:

1. Run the Kalman filter to get the moments of $p\left(r_t,\mu_t|Y,\Theta\right)$
2. Sample the last state from $\widehat{r}_T,\widehat{\mu}_T\sim p\left(r_T,\mu_T|y^T,\Theta\right)$
3. Sample backward through time: $\widehat{r}_t,\widehat{\mu}_t\sim p\left(r_t,\mu_t|\widehat{r}_{t+1},\widehat{\mu}_{t+1},Y^t,\Theta\right)$.

This provides a direct block update from $p\left(r,\mu|\Theta,Y\right)$.

8.5 Regime Switching

Consider the following regime switching model term structure model:

$$dr_t=\kappa\left(\mu\left(Z_t\right)-r_t\right)dt+\sigma\left(Z_t\right)dW_t$$

where there is a continuous-time discrete state Markov Chain Z_t taking values in the $Z=\{1,...,K\}$ with a state independent $P=\{P_{ij}\}$. This model incorporates a time-varying central tendency taking values $\mu_1,...,\mu_K$ and a stochastic volatility factor taking values $\sigma_1,...,\sigma_K$.

Assuming the Markov chain is observed by the agents in the economy, Lee and Naik (1994) and Landen (2000) show that zero coupon bond prices in this model are given by

$$Y_{t,\tau}=\alpha\left(\tau,Z_t,\Theta\right)+\beta\left(\tau,Z_t,\Theta\right)r_t$$

where to coefficient functions α and β solve Riccatti ordinary differential equations conditional on the Markov state. That is, for each state $Z_t=j$, one needs to solve an ordinary differential equation to get $\alpha\left(\tau,\Theta,j\right)$ and $\beta\left(\tau,\Theta,j\right)$.

Again, consider a discretized version of the continuous-time model, where we abuse notation by redefining the parameters in the drift where necessary:

$$\begin{aligned}Y_{t,\tau}&=\alpha\left(\tau,\Theta,Z_t\right)+\beta\left(\tau,\Theta,Z_t\right)r_t+\varepsilon_t\\r_{t+1}&=\mu\left(Z_t\right)+\kappa r_t+\sigma\left(Z_t\right)\varepsilon_t^r.\end{aligned}$$

where it is convenient to think of the model conditional on the Markov state: conditional on state j,

$$\begin{aligned} y_t &= \alpha(\tau, \Theta, j) + \beta(\tau, \Theta, j) r_t + \varepsilon_t \\ r_{t+1} &= \mu_j + \kappa r_t + \sigma_j \varepsilon_t^r. \end{aligned}$$

Finally, we need to specify prior distributions. As mentioned earlier, informative priors are required to avoid degeneracies (as in all regime switching models, there is a labeling problem: there is no unique way to identify the states). One approach to overcome this problem is to order the means or variances. A common noninformative reference prior that orders the means is

$$p(\mu, \sigma) \propto 1_{[\mu_1 < \ldots < \mu_k]} \prod_{i=1}^{K} \frac{1}{\sigma_i}.$$

However, this price leads to a conditional distribution which violates Hammersley-Clifford since $p\left(\mu_i, \sigma_i^2 | Z_t \neq i \ \forall i, \mu_{-i}, \sigma_{-i}\right) = p\left(\mu_i, \sigma_i^2 | \mu_{-i}\right)$, which is not a proper density. To remedy the problem of improper posterior conditionals, we propose the following prior:

$$\begin{aligned} p\left(\mu_i | \mu_{-i}\right) &= N(a_i, A_i) 1_{[\mu_{i-1} < \mu_i < \mu_{i+1}]} \text{ and} \\ p\left(\sigma_i^2\right) &= IG(b_i, B_i). \end{aligned}$$

We also place a normal prior on κ.

The analysis of regime switching models with MCMC methods have received an extraordinary amount of attention in the statistics and econometrics literature, in part due to the need for a formal avenue for imposing prior information due to degeneracies in the likelihood function. For additional references and discussions, see ....

Our MCMC algorithm is be defined over $[r, Z, P, \Theta]$ where $\Theta = (\mu, \sigma, \kappa)$. We consider the following updating scheme:

$$\begin{aligned} \text{Parameters} &: p(\mu, \sigma | \kappa, r, P, Z, Y) \\ &: p(P | \mu, \sigma, \kappa, r, Z, Y) \\ &: p(\kappa | \mu, \sigma, P, r, Z, Y) \\ \text{State Variables} &: p(Z | \kappa, \sigma, \mu, \kappa, P, r, Y) \\ &: p(r | \kappa, \sigma, \mu, \kappa, P, Z, Y) \end{aligned}$$

As in other term structure models, updating the parameters is difficult to the non-analytical functional form of the loading functions, $\alpha(\tau, \Theta, j)$ and $\beta(\tau, \Theta, j)$. Because of this, we will use a Metropolis-Hastings algorithm.

To update the parameters, $\mu, \underline{\sigma}$ and κ, we proceed as follows. For updating, μ, we update μ_i separately. To do this, we first find all times such that $Z_t = i$ as the conditional posterior for μ_i will only depend on these observations. Given this, we have that

$$\pi(\mu_i) \triangleq p\left(\mu_i|\kappa, \mu_{-i}, \sigma, Z, \kappa, r, Y\right) \propto \prod_{t:Z_t=i} p(y_t|r_t, \mu_i, \sigma_i)\, p\left(\mu_i|\mu_{-i}, \sigma_i, \kappa, \underline{r}\right).$$

where

$$p\left(\mu_i|\mu_{-i}, \sigma_i, \kappa, r\right) \propto \prod_{t:Z_t=i} p(r_t|r_{t-1}, \mu_i, \kappa)\, p\left(\mu_i|\mu_{-i}\right)$$

If we let $p\left(\mu_i|\mu_{-i}\right)$ be a truncated normal (to impose the ordering) then $p\left(\mu_i|\mu_{-i}, \sigma_i, \kappa, r\right)$ is also a truncated normal (Geweke (1994)). Defining

$$g(\mu_i) = \prod_{t:Z_t=i} p(y_t|r_t, Z_t = i, \mu_i, \sigma_i)$$

then an independence Metropolis-Hastings algorithm (equation 4) is given by:

1. draw $\mu_i^{(g+1)} \sim q(\mu_i) \propto p\left(\mu_i|\mu_{-i}, \sigma_i, \kappa, r\right)$
2. Accept with probability $\alpha\left(\mu_i^{(g+1)}, \mu_i^{(g)}\right)$

where

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

For the $\sigma_i's$, updating proceeds with a similar Metropolis-Hastings where we specify standard inverse gamma distributions for each σ_i. Updating κ follows in a similar fashion.

Updating the transition probabilities, $p(Z|\kappa, \sigma, \mu, \kappa, P, r, Y)$, also requires a Metropolis-Hastings step. We use a standard Dirichlet prior on the elements of P. That is, $\{p_{ij}\}_{i,j=1}^{K} \propto \mathcal{D}\left(\gamma_{ij}\right)$. The conditional posterior for the transition probabilities is given by:

$$p(Z|\Theta, P, r, Y) \propto p(Y|\sigma, \mu, \kappa, P, r, Z)\, p(P|Z)$$

The distribution $p(P|Z)$ is $\mathcal{D}\left(\gamma_{ij}+n_{ij}\right)$ where n_{ij} is the number of transitions from i to j, that is $n_{ij}=\sum_{t=1}^{T}1_{[Z_{t-1}=i,Z_t=j]}$. Proceeding as above, if we set $g(P)=p(Y|\Theta,P,r,Z)$, we use a independence Metropolis-Hastings step with proposal density $q(P)\sim p(P|Z)$.

Now we turn to the states. First, updating r is straightforward using the FFBS algorithm (heteroskedastic Kalman filter) as r_t enters linearly in the observation equation. Therefore we can directly draw from $p(r|\kappa,\sigma,\mu,\kappa,Z,Y)$.

To draw the Markov states, we consider both a single-state update and a block-update, both Gibbs steps. The single state update draws from

$$\begin{aligned}\pi(\mu_i) &\triangleq p(Z_t=i|Z_{t+1}=j,Z_{t-1}=k,\Theta,P,r,Y)\\ &\propto p(Y_t|Z_t=i,\Theta,r_t,P)\,p(Z_{t+1}=j|Z_t=i)\,p(Z_t=i|Z_{t-1}=k)\\ &\propto p(Y_t|Z_t,\Theta,r_t,P)\,p_{ki}p_{ij}\end{aligned}$$

The block update is computationally more expensive and is based on a recursive formula for the conditional state transitions, $p(Z_t|Z_{t-1},\Theta,r,Y)$ which based on a formula of see Lindgren (1978). Since

$$p(Z_t=j|Z_{t-1}=i,\Theta,r,Y)\propto c_j(t)\,p(Y_t|Z_t=j,r_t,\Theta)\,p_{ij}$$

where

$$\begin{aligned}c_j(t) &= p(Y_{t+1},..,Y_T|r_{t+1},...,r_T,Z_t=j,\Theta)\\ &= \sum_{i=1}^{K}p(Y_{t+1}|r_{t+1},Z_{t+1}=i,\Theta)\,c_i(t+1)\,p_{ij}.\end{aligned}$$

This can be recursively found by starting at $c_j(T)=prob(Y_T|r_T,Z_T=j,\Theta)$. To draw the vector of Markov states in one block, sample forward using the conditional distribution

$$p(Z_t=j|Z_{t-1}=i,\Theta,\underline{r},Y).$$

9 Estimation and Model Risk

Estimation corresponds to the uncertainty present regarding the parameter or state variables values and model corresponds to the uncertainty over which model is the most accurate

description of the data. We now discuss each of these issues and how the output of MCMC algorithms can be used to quantify the issues.

As mentioned, inference on the parameters and the state variables is summarized by $p(\Theta|Y)$ and $p(X|Y)$. There are two ways in which these distributions quantify and account for estimation risk. First, $p(\Theta|Y)$ and $p(X|Y)$ are distributions and, in contrast to point estimates of parameters or state variables, provide all of the sample based information regarding their values. Thus it is easy to quantify the uncertainty in estimating Θ and X. Second, these distributions that into account the uncertainty present in estimating the other quantity. We can represent, for example, the marginal parameter posterior as

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

and it is clear that the parameter estimates take into account the fact that the state variables are also imperfectly estimated.

Model risk relates to the problem of understanding how inference and pricing change under different model specifications. It provides a framework for performing sensitivity analysis and providing specification diagnostics for a given model. For example, EJP(2001) consider the effect of adding jumps to returns and volatility on inference and option pricing. To fully account for model risk, Bayes rule provides the natural solution of the following form. For the purpose of illustration, suppose that there are two models, or hypotheses, under consideration denoted by M_0 and $\mathcal{M}_1$, respectively.

First, the conditional posterior probability distributions for the parameters Θ under the two model specifications are given by $p(\Theta|X,Y,\mathcal{M}_0)$ and $p(\Theta|X,Y,\mathcal{M}_1)$. There is an equivalent set of conditionals for the state vector X. If the researcher wishes to fully account for model risk or uncertainty they must calculate the posterior probabilities of each model or hypothesis being true given the data, namely $p(\mathcal{M}_i|Y)$. By Bayes rule

$$p(\mathcal{M}_i|Y) = \frac{p(Y|\mathcal{M}_i)p(\mathcal{M}_i)}{p(Y)}$$

where

$$p(Y|\mathcal{M}_i) = \int_{X,\Theta} p(Y|X,\Theta,\mathcal{M}_i)p(X,\Theta|\mathcal{M}_i)dXd\Theta$$

is the marginal likelihood of the observed data.

The marginal posterior $p(\Theta|Y)$ is then given by the weighted average of the conditional posteriors

$$p(\Theta|Y) = p(\Theta|Y, \mathcal{M}_0)p(\mathcal{M}_0|Y) + p(\Theta|X, Y, \mathcal{M}_1)p(\mathcal{M}_1|Y)$$

where

$$p\left(\Theta|Y, \mathcal{M}_i\right) = \int p\left(\Theta|X, Y, \mathcal{M}_i\right) p\left(X|\mathcal{M}_i\right) dX$$

It is more common to compute the ratio of posterior model probabilities via an Odds ratio identity. For model comparison, notice that we do not have to assume models fully exhaustive. For parameter inference via the marginal posterior, $p(\Theta|Y)$, Bayes rule requires that all candidate models to be accounted for.

Our approach to model comparison uses an Odds ratio identity to quantify the probability that one model generated the data relative to another model. Formally, the posterior odds of $\mathcal{M}_0$ being correct versus $\mathcal{M}_1$ is given by

$$\frac{p\left(\mathcal{M}_0|Y\right)}{p\left(\mathcal{M}_1|Y\right)} = \frac{p\left(Y|\mathcal{M}_0\right)}{p\left(Y|\mathcal{M}_1\right)} \frac{p\left(\mathcal{M}_0\right)}{p\left(\mathcal{M}_1\right)}$$

where the equality follows from Bayes rule and $p\left(\mathcal{M}_0\right)$ is the prior probability that $\mathcal{M}_0$ is correct.

The Odd's ratio consists of two components, the likelihood ratio, $p\left(Y|\mathcal{M}_0\right)/p\left(Y|\mathcal{M}_1\right)$, which is known as the Bayes Factor and the prior odds $p\left(\mathcal{M}_0\right)/p\left(\mathcal{M}_1\right) = 1$. The marginal likelihood ratio is given by

$$p\left(Y|\mathcal{M}_i\right) = \int p\left(Y|\Theta, \mathcal{M}_i\right) p\left(\Theta|\mathcal{M}_i\right) d\Theta.$$

Note that this is an averaged and not a maximized likelihood. This implies that Bayes Factor's are an automatic "Occam's Razor" in that they account for the complication of the model by integrating out parameter uncertainty.

As in all likelihood based approaches, model comparison is typically straightforward when models are nested. For example, suppose that $\mathcal{M}_0$ corresponds to the case where $\Theta = \Theta_0$ in model $\mathcal{M}_1$. The Bayes Factors then takes a particularly simple form known as the Savage density ratio:

$$\frac{p\left(Y|\mathcal{M}_0\right)}{p\left(Y|\mathcal{M}_1\right)} = \frac{p\left(\Theta = \Theta_0|Y, \mathcal{M}_1\right)}{p\left(\Theta_0|\mathcal{M}_1\right)}.$$

That is the ratio of the posterior ordinate, $p\left(\Theta = \Theta_0 | Y, \mathcal{M}_1\right)$, to the prior ordinate, $p\left(\Theta_0 | \mathcal{M}_1\right)$ both calculated under the model $\mathcal{M}_1$. A simple MCMC estimator of this quantity

$$\frac{p\left(Y|\mathcal{M}_0\right)}{p\left(Y|\mathcal{M}_1\right)} \approx \frac{1}{G}\sum_{g=1}^{G} \frac{p\left(\Theta = \Theta_0 | X^{(g)}, Y, \mathcal{M}_1\right)}{p\left(\Theta_0|\mathcal{M}_1\right)}.$$

We now turn to commonly used asset pricing models.

10 Conclusions

MCMC methods provide an attractive simulation-based approach for inference in empirical asset pricing models. They provide a solution to the inverse problem of drawing inference about state variables and parameter values given observed price data. They provide a natural framework for assessing issues such as estimation and model risk in financial decision making.

This chapter provides a tutorial on building MCMC algorithms for applications in continuous-time finance. In the case of equity prices, we show how to use MCMC to estimate models with jumps, stochastic volatility and time-varying expected returns. These algorithms can be extended in a straightforward manner to include additional observations such as option prices. We show how to build MCMC algorithms to estimate multi-factor term structure models with time-varying central tendancies, stochastic volatility or jumps.

These algorithms have a number of attractive convergence and estimation properties that they inherit from the general theory of Markov chain convergence. For example, they are typically at least geometrically convergent and have an ergodic averaging and central limit theorem available in wide generality.

References

[1] Anderson, T. (1984). An Introduction to Multivariate Statistical Analysis. 2nd Edition. John Wiley & Sons.

[2] Besag, J. (1974), Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society. Series B, 36, 192-236.*

[3] Besag, J. and Green, P. J. (1993). Spatial Statistics and Bayesian Computation (with discussion). *J. R. Statist. Soc.*, B, **55**, 25-37.

[4] Carlin, B.P. and Polson, N.G. (1991). Inference for Nonconjugate Bayesian Models using the Gibbs sampler. *Cand. J. Statistics*, **19**, 399-405.

[5] Carlin, B. and N.G. 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.

[6] Carlin, B.P., and Polson, N.G. and Stoffer, D.S (1992), "A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling," J. Amer. Stat. Ass., 87, 493-500.

[7] Carter, C.K., and Kohn, R. (1994), "On Gibbs Sampling for State Space Models" *Biometrika*, 61.

[8] Carter, C.K., and Kohn, R. (1994), "Markov chain Monte Carlo in conditionally Gaussian state space models" *Biometrika*, 83, 589-601

[9] Chib, S. (1996), "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models." Journal of Econometrics, 75, 79-97.

[10] Chib, S. (1995), "Marginal Likelihood From the Gibbs Output" Journal of the American Statistical Association, 90, 1313-1321.

[11] Chib, S.(1998), "Estimation and Comparison of Multiple Change Point Models" Journal of Econometrics, (1998), 86, 221-241.

[12] Diaconis, P. and D. Stroock, (1991), "Geometric bounds for eigenvalues of Markov chains," Ann. Appl. Prob., vol. 1, 36-61 (1991)

[13] Duffie, D., (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, 1994, Boston: Birkhauser, 1996,

[14] Edwards, W., H. Lindman and L. J. Savage (1963), Bayesian Statistical Inference for Psychological Research. *Psychological Research*, 70, 193.

[15] Elerian, O., S.Chib and N.Shephard "Likelihood Inference for Discretely Observed Non-linear Diffusions" *Econometrica*, (2001), 69, 959-994.

[16] Frieze, A., Kannan, R. and Polson, N.G. (1994). Sampling from log-concave distributions. *Annals of Applied Probability.*

[17] Gelfand, A.E., Hills, S., Racine-Poon, A and Smith, A.F.M., (1990), "Illustration of Bayesian inference in normal data models using Gibbs Sampling," J. Amer. Statist. Associ., 85, 972-982.

[18] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling Based approaches to calculating Marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398-409.

[19] Gelfand, A.E., Smith, A.F.M. and Lee, T.M., "Bayesian Analysis of constrained parameters and truncated data problems using Gibbs Sampling," *J. Amer. Statist. Assoc.*, **87**, 523-532.

[20] Gelman, A. and Rubin, D. (1992). Inference from Iterative simulation using Multiple sequences. *Statistical Science*, **7**, 457-473.

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

[22] Geyer, C.J. (1992). Practical Markov chain Monte Carlo. *Stat. Sci.*, **7**, 473-511

[23] Hammersly, J.M. and Clifford, M.S., (1970), "Markov fields on finite graphs and lattices," Unpublished.

[24] Hammersley, J.M. (1974), "Discussion of Besag's paper," *Journal of the Royal Statistical Society. Series B, 36, 230-231*..

[25] Hastings, W.K. (1970), Monte Carlo sampling Methods using Markov Chains and their Applications. *Biometrika*, 57, 97-109.

[26] Hobert, J.P. and G. Casella (1996) The effect of improper priors on Gibbs sampling in hierarchical linear models. *J. Amer. Statist. Assoc.*, **91**, 1461-1473.

[27] Jacquier, E., Polson, N. and Rossi, P. (1994), "Bayesian analysis of Stochastic Volatility Models", (with discussion). *J. Business and Economic Statistics.*, 12, 4.

[28] Jacquier, E., Polson, N. and Rossi, P. (1994), Models and Priors for Multivariate Stochastic Volatility. Working Paper, U. of Chicago.

[29] Johannes, M. (2001). "Models of Eurodollar Futures: Forward-Futures Spreads and Extracting Zeroes." Working paper, Columbia University.

[30] Johannes, M., N. Polson and J. Stroud. (2001), "Volatility Timing and Portfolio Returns." Working paper, Columbia University.

[31] Johannes, M., N. Polson and J. Stroud. (2001), "The TED Spread." Work in progress.

[32] Jones, C. (2001). "The Dynamics of Stochastic Volatility: Evidence from Underlying and Options Markets" working paper, Rochester University.

[33] Kass, R.E. and Raftery, A.E., (1995), "Bayes Factors," *J. Amer. Statist. Assoc.*, **90**, 773-795.

[34] Liu, J.S., W.H. Wong and Kong, A., (1994) "Covariance Structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes," *J. R. Statist. Soc.*, B, 57, 157-169.

[35] Mengersen, K. and Robert, C. (1998), MCMC Convergence Diagnostics: A Review (with discussion). In *Bayesian Statistics 6*, J.M. Bernardo et al (Eds.). *Oxford University Press*, Oxford, 399-432.

[36] Mengersen, K. L.; Tweedie, R. L. Rates of convergence of the Hastings and Metropolis algorithms. Ann. Statist. 24 (1996), no. 1, 101–121.

[37] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *J. Chemical Physics*, **21**, 1087-1091.

[38] Meyn, Sean P.; Tweedie, R. L. Computable bounds for geometric convergence rates of Markov chains. Ann. Appl. Probab. 4 (1994), no. 4, 981–1011.

[39] Polson, N.G. (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.

[40] Ripley, B. (1976) *Stochastic Simulation*

[41] Roberts, G.O. and Polson, N.G. (1994). On the Geometric Convergence of the Gibbs sampler. *J. R. Statist. Soc.*, B, 377-384.

[42] Roberts, Gareth O. and Rosenthal, Jeffrey S. Geometric ergodicity and hybrid Markov chains. Electron. Comm. Probab. 2 (1997), no. 2, 13–25

[43] Roberts, G.O. and Rosenthal, J.S. (1998). Markov chain Monte Carlo: Some practical implications of theoretical results. Can. J. Statis., 26, 4-31.

[44] Roberts, G. O.; Smith, A. F. M. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic Process. Appl. 49 (1994), no. 2, 207–216.

[45] Roberts, G. O.; Tweedie, R. L. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Biometrika 83 (1996), no. 1, 95–110.

[46] Roberts, G. O. and Tweedie, R. L. Bounds on regeneration times and convergence rates for Markov chains. Stochastic Process. Appl. 80 (1999), no. 2, 211–229.

[47] Rosenthal, Jeffrey S. Rates of convergence for Gibbs sampling for variance component models. Ann. Statist. 23 (1995a), no. 3, 740–761.

[48] Rosenthal, J.S. (1995b). Minorization Conditions and Convergence Rates for MCMC, Journal of the American Statistical Association, vol. 90, 558-566.

[49] Smith, A.F.M. and Roberts, G.O. (1993). Bayesian Computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc.*, B, **55**, 3-23.

[50] Stroud, J.R., Müller, Peter and Polson, N.G. (2001) Nonlinear State-Space Models with State-Dependent Variance Functions *Working Paper*.

[51] Tanner, M.A., and Wong, W.H. (1987). The Calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, **82**, 528-550.

[52] Tierney,L. (1994), Markov Chains for exploring Posterior Distributions (with discussion). *Ann. Statist.*, 22, 1701-1786.

A Regression Analysis

In this appendix, we review standard conjugate analysis for univariate, multivariate and vector autoregression. The conditional posteriors we derive here can be found in any standard text on Bayesian methods, see, e.g., O'Hagan (1997) or Bernardo and Smith (1996).

First consider a univariate regression model for observations, Y, and regressor matrix X:

$$Y = X\beta + \varepsilon$$

where Y is a $T \times 1$ vector, β is a $K \times 1$ vector and X is a $T \times K$ matrix of observations and we assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Assuming standard conjugate independent priors for the regression parameters and the variance

$$\begin{aligned} p(\beta) &= N(a, A) \\ p(\sigma^2) &= IG(b, B). \end{aligned}$$

The inverse Gamma distribution is often used as a prior specification for univariate variance parameters. It has support over the positive real line and if $\sigma^2 \sim \mathcal{IG}(b, B)$, its density (with respect to Lesbesgue measure on [0,∞], is given by:

$$\mathcal{IG}(b, B) \sim \frac{\beta^b e^{-B/x}}{\Gamma(b) x^{b+1}}.$$

The likelihood function is given by:

$$\begin{aligned} p(Y|X, \beta, \sigma^2) &\propto (\sigma^2)^{-\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right) \\ &= (\sigma^2)^{-\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\left(\beta - \widehat{\beta}\right) X'X \left(\beta - \widehat{\beta}\right) + S\right]\right) \end{aligned}$$

where $\widehat{\beta} = (X'X)^{-1} XY$ and $S = \left(Y - X\widehat{\beta}\right)' \left(Y - X\widehat{\beta}\right)$. The conditional posteriors are standard from a number of textbooks on Bayesian methods (see. e.g., O'Hagan (1997)). The conditional posterior for the regression coefficients is given by:

$$\begin{aligned} p(\beta|X, Y, \sigma^2) &\propto p(Y|X, \beta, \sigma^2) p(\beta) \\ &\propto N(a^*, A^*) \end{aligned}$$

where

$$\begin{aligned} a^* &= \left(A^{-1} + \frac{1}{2}X'X\right)^{-1}\left(A^{-1}a + \sigma^{-2}X'Y\right) \\ A^* &= \left(A^{-1} + \sigma^{-2}X'X\right)^{-1} \end{aligned}$$

and the conditional posterior for the variance is similarly

$$\begin{aligned} p\left(\sigma^2|X,Y,\beta\right) &\propto p\left(Y|X,\beta,\sigma^2\right)p\left(\sigma^2\right) \\ &\propto IG\left(b^*,B^*\right) \end{aligned}$$

where

$$\begin{aligned} b^* &= T + b \\ B^* &= \left(Y - X\beta\right)'\left(Y - X\beta\right) + B \end{aligned}$$

Second, consider a multivariate normal model for a vector Y_t :

$$Y_t = N\left(\mu, \Sigma\right),$$

where Y_t and μ are a $k \times 1$ vectors and Σ is a $k \times k$ symmetric covariance matrix. We assume the following prior distributions for the parameters:

$$\begin{aligned} \Sigma &\sim \mathcal{W}^{-1}\left(b, B; k\right) \\ \mu &\sim N\left(a, A\right) \end{aligned}$$

The likelihood function is given by:

$$\begin{aligned} p\left(Y|\mu,\Sigma\right) &= \prod_{t=1}^{T} p\left(Y_t|\mu,\Sigma\right) \\ &\propto |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}\left(Y_t - \mu\right)'\Sigma^{-1}\left(Y_t - \mu\right)\right\} \end{aligned}$$

To simplify this, note that

$$\sum_{t=1}^{T}\left(Y_t - \mu\right)'\Sigma^{-1}\left(Y_t - \mu\right) = T(\mu - \overline{Y})'\Sigma^{-1}(\mu - \overline{Y}) + Ttr\left(\Sigma^{-1}S\right)$$

where $\overline{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$ and $S = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \overline{Y})(Y_t - \overline{Y})'$ is the sample covariance matrix. In order to model variance-covariance matrices, we use a generalization of the inverse gamma known as an inverse Wishart distribution. An $n \times n$ matrix Σ has an inverse Wishart distribution with parameter b and matrix parameter B, its density is given by:

$$\mathcal{W}^{-1}(b, B) \propto |B|^{\frac{(b-n-1)}{2}} |\Sigma|^{-\frac{b}{2}} \exp\left(-\frac{1}{2} tr\left(\Sigma^{-1} B\right)\right).$$

This implies that the likelihood function is given by:

$$p(Y|\xi, J, \Theta) = |\Sigma|^{\frac{T}{2}} \exp\left\{-\frac{1}{2} T(\mu - \overline{Y})'\Sigma^{-1}(\mu - \overline{Y}) - \frac{1}{2} T * tr\left(\Sigma^{-1} S\right)\right\}$$

The conditional posterior for the means is given by:

$$\begin{aligned} p(\mu|Y, \Sigma) &\propto p(Y|\xi, J, \Theta)\, p(\mu) \\ &\propto \exp\left\{-\frac{1}{2} T(\mu - \overline{Y})'\Sigma^{-1}(\mu - \overline{Y}) - \frac{1}{2}(\mu - a)'A^{-1}(\mu - a)\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\mu - a^*)'(A^*)^{-1}(\mu - a^*)\right\} \end{aligned}$$

which is $N(a^*, A^*)$ where

$$\begin{aligned} a^* &= A^*\left(A^{-1}a + T\Sigma^{-1}\overline{Y}\right) \\ A^* &= \left(A^{-1} + T\Sigma^{-1}\right)^{-1}. \end{aligned}$$

The condtional posteriors for the variance is similarly straightforward:

$$\begin{aligned} p(\Sigma|Y, \mu) &\propto p(Y|\mu, \Sigma)\, p(\Sigma) \\ &\propto |\Sigma|^{-\frac{1}{2}(b+T)} \exp\left\{-\frac{1}{2} tr\left(\Sigma^{-1} B\right) - \frac{1}{2} T(\mu - \overline{Y})'\Sigma^{-1}(\mu - \overline{Y}) - \frac{1}{2} T * tr\left(\Sigma^{-1} S\right)\right\} \\ &\propto |\Sigma|^{-\frac{1}{2}(b+T)} \exp\left\{-\frac{1}{2}\left[tr\left(\Sigma^{-1}\left[B + T * S\right]\right) + T(\mu - \overline{Y})'\Sigma^{-1}(\mu - \overline{Y})\right]\right\} \end{aligned}$$

since $T(\mu - \overline{Y})'\Sigma^{-1}(\mu - \overline{Y}) = tr\left(\Sigma^{-1} T(\mu - \overline{Y})(\mu - \overline{Y})'\right)$, we have that

$$p(\Sigma|Y, \mu) \propto |\Sigma|^{-\frac{1}{2}(b+T)} \exp\left\{-\frac{1}{2} tr\left\{\left(\Sigma^{-1}\left[B + T * S + T(\mu - \overline{Y})(\mu - \overline{Y})'\right]\right)\right\}\right\}$$

which implies that

$$\begin{aligned} p\left(\Sigma|Y,\mu\right) &\sim \mathcal{W}^{-1}\left(b^{*},B^{*}\right) \\ b^{*} &= b+T \\ B^{*} &= \left[B+TS+T(\mu-\overline{Y})(\mu-\overline{Y})'\right] \end{aligned}$$

To draw from an inverse Wishart, $\Sigma \sim \mathcal{W}^{-1}(b,B)$, the first step is to form the Cholesky factorization of B. That is, find C such $C\prime C = B$. Now draw a matrix R of i.i.d. $N\left(0,1\right)$ random variables, then let $V = RC$ which has independent rows $v_i \sim N\left(0,B\right)$. Then

$$\sum_{i=1}^{b} v_i v_i' \sim \mathcal{W}^{-1}(b,B).$$

Finally consider a vector autoregression where Y_t is a $n \times 1$ vector of dependent variables and the regressor variable

$$\begin{aligned} Y_t &= X_t\beta+\varepsilon_t \\ &= \left[1,Y_{t-1}\right]'\beta+\varepsilon_t \end{aligned}$$

which implies that $Y_t|Y_{t-1} \sim N\left(\left[1,Y_{t-1}\right]'\beta,\Sigma\right)$. With a noninformative prior on β and Σ,

$$\beta|\Sigma,Y \sim N\left(\widehat{\beta},\Sigma^{-1}\otimes\left(X'X\right)^{-1}\right)$$

and

$$\Sigma|Y \sim \mathcal{W}^{-1}(T-n-1,S^{-1})$$

where S is the sample covariance matrix.