MCMC Maximum Likelihood For Latent State Models

Eric Jacquier, Michael Johannes and Nicholas Polson*

January 13, 2004

Abstract

This paper develops a simulation-based approach for performing maximum likelihood estimation in latent state variable models using Markov Chain Monte Carlo methods (MCMC). The MCMC algorithm simultaneously computes (by numerical integration) and optimizes the marginal likelihood function. The approach relies on data augmentation, combining the insights of simulated annealing and evolutionary or “genetic” algorithms. We prove a limit theorem in the degree of data augmentation and use this to provide standard errors and convergence diagnostics. The estimator inherits all of the usual sampling asymptotic properties of maximum likelihood estimators. We demonstrate the approach on two latent state models central to financial econometrics: a stochastic volatility model and a multivariate jump-diffusion model.

*Jacquier is at HEC Montreal and CIRANO, eric.jacquier@hec.ca. Johannes is at the Graduate School of Business, Columbia University, 3022 Broadway, NY, NY, 10027, mj335@columbia.edu. Polson is at the Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago IL 60637, ngp@gsb.uchicago.edu. We would like thank Ron Gallant for his discussion of the paper, Lars Hansen for his comments, and seminar participants at the 2003 Montreal Financial Econometrics Conference and Columbia University for helpful comments.

MCMC Maximum Likelihood For Latent State Models

JEL Classification: C1, C11, C15, G1

Key words: MCMC, Maximum Likelihood, Optimization, Simulated Annealing, Evolutionary Monte-Carlo, Stochastic volatility, Jumps, Diffusion, Financial Econometrics.

1 Introduction

Computing maximum likelihood estimators (MLE) in latent variable models is a notoriously difficult problem for two reasons. First, the likelihood function is not known in closed form in latent variable models. Computing the likelihood typically requires Monte Carlo methods to draw from the latent state distribution and then approximate the integral that appears in the likelihood. Second, a nonlinear search algorithm optimizes the approximated likelihood over the parameters. In this paper, we provide a Markov Chain Monte Carlo based approach that simultaneously performs the evaluation and likelihood optimization in latent state models.¹ The method provides parameter estimates, standard errors and estimation of latent state variable distribution.

Our approach combines the insights of simulated annealing (see, e.g., Kirkpatrick et al. (1983) and Van Laarhoven and Aarts (1987)), and evolutionary MCMC algorithms (see, e.g., Liu, Liang, and Wong (2000) and Mueller (2000)). Like simulated annealing, our approach shares the goal of simulation-based optimization, but unlike simulated annealing, it does not require that the objective function can be directly evaluated, in our setting the likelihood, $\mathcal{L}(\theta)$. Simulated annealing generates samples from a sequence of densities, $\pi_{J(g)}(\theta) \propto \mathcal{L}(\theta)^{J(g)}$, where g the length of the Markov Chain. The ‘temperature’ schedule, $J(g)$, is chosen such that as g increases, $\pi_{J(g)}(\theta)$ concentrates around its maximum, the MLE. Evolutionary Monte Carlo, on the other hand, generates J copies (or populations) of the parameter θ and generates a Markov chain over these copies. It often has better convergence properties despite the higher dimensionality. Unfortunately, in latent state models, the likelihood function is itself an integral and simulated annealing or evolutionary Monte Carlo methods do not directly apply because the likelihood cannot be directly computed.

Our approach combines the insights from simulated annealing and evolutionary Monte

¹Approaches that have been developed for computing MLE’s include the expectation-maximization algorithm of Dempster, Laird and Rubin (1982), Geyer’s (1991) Monte Carlo maximum likelihood approach, and Besag (1974) and Doucet et al. (2002) for maximum a posteriori estimation.

Carlo. We augment the data by creating J copies of the latent variables (as opposed to J copies of the parameters). This allows the transition kernel of the Markov Chain to depend on other latent state components, the so-called evolutionary property. Specifically, we define a joint distribution $\pi_J(\tilde{X}^J, \theta)$ on the space of the parameters θ and the J copies of the latent state variables, $\tilde{X}^J$. The augmented joint distribution $\pi_J(\tilde{X}^J, \theta)$ has a marginal parameter distribution that is proportional to power transform of the likelihood, that is $\pi_J(\theta) = \int \pi_J(\tilde{X}^J, \theta) d\tilde{X}^J \propto \mathcal{L}(\theta)^J$. As in simulated annealing, as J increases, our algorithm produces a sequence of parameter draws, θ^J , that converge to the maximum of $\mathcal{L}(\theta)$, the MLE. Standard MCMC methods provide samples from $\pi_J(\tilde{X}^J, \theta)$ by, for example, iteratively sampling from the complete conditional distributions, $\pi(\theta|\tilde{X}^J)$ and $\pi(\tilde{X}^J|\theta)$.

Under regularity conditions, we provide an asymptotic in J showing that the marginal distribution (from the higher-dimensional joint) is approximately normal with the asymptotic observed variance-covariance matrix scaled by J . This can be used to calculate standard errors from the observed variance-covariance matrix by appropriately scaling the MCMC draws. This also provides a powerful practical tool to diagnose convergence of the chain and hence the degree of augmentation that is required. Since the researcher controls the degree of augmentation, one increases it until the parameter draws are approximately normal.

Our approach has a number of practical and theoretical advantages. First, unlike other simulated maximum likelihood approaches for state estimation which substitute the final parameter estimates into an approximate filter, our algorithm provides the optimal smoothing distribution of the latent variables. This is especially important in non-linear or non-normal latent variable models, models for which the Kalman filter does not apply. Second, it has all of the advantages of MCMC without any of the, perceived by some, disadvantages. For example, we do not, *per se*, require prior distributions over the parameters, although the conditionals and joint distributions used must be integrable.

Third, like simulated annealing, we provide MLE estimates without resorting to numerical search algorithms, such as inefficient gradient based methods, which often get locked into local maxima. Fourth, the approach handles models with nuisance parameters and latent variables as well as models with constrained parameters or parameters on boundaries. Finally, as an asymptotic in sample size, the estimation procedure inherits all the asymptotic properties of standard MLE methods for these models.

To illustrate our approach, we analyze two important models in financial econometrics. The first is the standard log-stochastic volatility model (SV) of Taylor (1986), analyzed using MCMC first by Jacquier, Polson and Rossi (1994) and others. Here, our approach produces the MLE. The second model is a multivariate version of Merton's (1976) jump-diffusion model. This model is of special interest in asset pricing because it delivers closed form option prices, but is difficult to estimate given the well-known degeneracies of the likelihood, see, for example, Kiefer (1978).

It is important to recognize that our MCMC approach is distinct from the approach of Chernozhukov and Hong (2003) who propose a quasi-Bayesian MCMC procedure. Their methodology applies to a wide class of models when $\mathcal{L}(\theta)$ could be a likelihood or GMM criterion function. Instead of finding the mode of the target function of interest, e.g. the MLE, they estimate its mean or quantiles. Their estimators have good asymptotic properties as the sample size increases. More precisely, they study the target distribution $\mathcal{L}(\theta)\mu(\theta)/\int_{\Theta}\mathcal{L}(\theta)\mu(d\theta)$ for some measure $\mu(d\theta)$. Of course, this estimation procedure, viewed as a quasi-Bayes posterior inherits the asymptotic properties of Bayes estimators with flat priors: see Dawid (1970), Heyde and Johnstone (1978) or Schervish (1995). Unlike Chernozhukov and Hong (2003), we estimate the optimum of the target function of interest, and our asymptotic is in J and g for a fixed sample size. Chernozhukov and Hong (2003) provide little discussion of the choice of J or the speed of convergence of their Markov Chain as a function of g for a given sample size.

Our approach also applies to many other problems in economics and finance that

require joint integration and optimization. Standard expected utility problems are an excellent example of this, as the agent first integrates the uncertainty to compute expected utility and then maximizes. In Jacquier, Johannes and Polson (2004) we extend the existing approach to maximum expected utility portfolio problems.

The rest of the paper proceeds as follows. Section 2 provides the general methodology together with the convergence proofs and the details of the convergence properties of the algorithm. Section 3 and 4 provides simulation based evidence for two commonly used latent state variable models in econometrics, the log-stochastic volatility model and a multivariate jump-diffusion model. Finally, Section 5 concludes with directions for future research.

2 Simulation-based Likelihood Inference

Models of with latent variables abound throughout finance and economics. In finance, many of the prominent models include latent variables: time-varying equity premium or volatility models, models with jumps, and regime-switching models. In economics, discrete-choice models, censored and truncated regression models, panel data problems with missing data all qualify as missing data problems. When possible, MLE is a preferred estimation methods due to its strong theoretical properties.

Formally, consider a model with observed data $Y = (Y_1, \dots, Y_T)$, latent state variables $X = (X_1, \dots, X_T)$, and parameter vector θ . The marginal likelihood of θ is given by

$$\mathcal{L}_T(\theta) = \int p(Y|X, \theta)p(X|\theta)dX, \quad (1)$$

where we refer to $p(Y|X, \theta)$ as the full-information or augmented likelihood function and $p(X|\theta)$ is the distribution of the latent states, the state-variable evolution.

Directly maximizing $\mathcal{L}_T(\theta)$ is difficult for three reasons. First, since the likelihood in (1) is rarely known in closed form, one must first generate samples from $p(X|\theta)$ and

then approximate the integral with Monte Carlo methods. Unfortunately, it is in fact more complicated because it is rarely possible to directly draw from $p(X|\theta)$ (linear, Gaussian models are an important exception) and thus another layer of approximations is required. To draw from $p(X|\theta)$, it is common to use approximate filters or importance sampling. Second, iterating between approximating and optimizing the likelihood is typically extremely computationally burdensome. Finally, in some latent variable models, the MLE may not exist. For example, in a time-discretization of Merton's (1976) model, the case of a mixture of two normal distributions with Bernoulli mixing probabilities, the likelihood is unbounded for certain parameter values. Our approach deals with all of these issues.

To understand our approach, consider the joint density, $\pi_J(\tilde{X}^J, \theta)$, over the parameter θ and a tensor $\tilde{X}^J = (X^1, \dots, X^J)$, of J -independent copies of the state variables, where $X^j = (X_1^j, \dots, X_T^j)'$. It is important to contrast this with typical Bayesian inference in latent state models which defines a Markov Chain over $\pi(X, \theta)$. Although we drastically increase the dimension of the state space of the chain, the insights of evolutionary MCMC suggest that this has important advantages. The joint distribution over the parameters and augmented state matrix is defined by:

$$\pi_J(\tilde{X}^J, \theta) \propto \prod_{j=1}^J p(Y|X^j, \theta) p(X^j|\theta).$$

In general, the density, $\pi_J(\tilde{X}^J, \theta)$, may not integrate, even for large J , although it could alleviate problems in some cases. In these cases, it is useful to introduce a dominating measure $\mu(\theta)$ and consider the joint distribution $\pi_J^\mu(\tilde{X}^J, \theta)$ defined by

$$\pi_J^\mu(\tilde{X}^J, \theta) \propto \prod_{j=1}^J \pi(X^j|\theta) \mu(\theta),$$

where

$$\pi(X^j|\theta) = p(Y|X^j, \theta)p(X^j|\theta).$$

We later discuss the choice of the dominating measure, typically, it is just Lebesgue measure.

Using MCMC to generate samples from $\pi_J^\mu(\tilde{X}^J, \theta)$ has several advantages. First, it is easy to sample from $\pi_J^\mu(\tilde{X}^J, \theta)$ using MCMC. The Clifford-Hammersley Theorem (see Robert and Casella (1999) or Johannes and Polson (2004)) implies that $\pi(\theta|\tilde{X}^J)$ and $\pi(\tilde{X}^J|\theta)$ are the complete conditionals. An MCMC algorithm consists of iteratively drawing from these two distributions: given the g^{th} draws of the chain, $\tilde{X}^{J,(g)}$ and $\theta^{(g)}$, draw

$$\theta^{(g+1)} \sim \pi(\theta|\tilde{X}^{J,(g)}) \text{ and } \tilde{X}^{J,(g+1)} \sim \pi(\tilde{X}^J|\theta^{(g+1)}).$$

Here $\tilde{X}^{J,(g)}$ denotes the g^{th} draw of the tensor $\tilde{X}^J$, and J is fixed. If these steps can be performed via direct draws, the algorithm is a Gibbs sampler. In other cases, Metropolis-Hastings algorithms can be used to sample from the appropriate distributions.² Drawing the latent states comes from drawing J independent copies of $\pi(\tilde{X}^J|\theta^{(g+1)})$: where

$$\pi(\tilde{X}^J|\theta^{(g+1)}) = \prod_{j=1}^J \pi(X^j|\theta^{(g+1)}).$$

Second, when using Metropolis to sample from $\pi(\tilde{X}^J|\theta)$, the algorithm can have a genetic or evolutionary component. Specifically, when updating X^j , the Metropolis kernel can depend on $(X^1, \dots, X^{j-1}, X^{j+1}, \dots, X^J)$. In difficult problems, this can improve

²An alternative for sampling from $\pi(\theta)^J$ when derivative information is available is the Langevin diffusion which solves

$$d\Theta_t = \frac{J\sigma_J^2}{2} \nabla_{\Theta} \log \pi(\Theta_t) dt + \sigma_J dW_t.$$

The solution to this SDE, Θ_t has a stationary distribution of $\pi(\theta)^J$. Roberts and Rosenthal (1997) show that the optimal scaling parameter is $\sigma_J^2 = O(J^{-1/3})$ and that convergence is in $O(J^{1/3})$, rather than $O(J)$ steps for general Metropolis-Hastings algorithms. Of course, this approach also requires that the target density is known.

convergence. At first this may seem counterintuitive, as the dimension of the state space is higher. However, if the algorithm is genetic, it can be harder for an element of X to get stuck or trapped in a region of the state space “getting stuck” may now require that all J copies get stuck in a certain region of the parameter space. This drastically reduces the probability of having the Markov Chain get trapped.

This joint distribution has a very special property: the marginal distribution of θ has the same form of the objective function as the objective used in simulated annealing. For notational simplicity, assume that $\mu(\theta)$ is a Lebesgue measure. Now, the marginal distribution is given by

$$\pi_J(\theta) = \int \pi_J(\theta, \tilde{X}^J) d\tilde{X}^J = \prod_{j=1}^J \int p(Y|X^j, \theta) p(X^j|\theta) dX^j$$

where we recall that $\mathcal{L}_T(\theta) = \int p(Y|X^j, \theta) p(X^j|\theta) dX^j$. Assuming that $\int \mathcal{L}_T(\theta)^J d\theta < \infty$, then we can write the marginal distribution as

$$\pi_J(\theta) = \frac{\mathcal{L}_T(\theta)^J}{\int \mathcal{L}_T(\theta)^J d\theta}.$$

If we re-write the likelihood as $\pi_J(\theta) \propto \exp(J \log \mathcal{L}_T(\theta))$, the main insight of simulated annealing implies that as we increase J , $\pi_J(\theta)$ collapses onto the maximum of $\log \mathcal{L}_T(\theta)$. Hence, by a careful choice of the degree of augmentation J we will be able to recover the maximum likelihood estimator and its asymptotic standard errors as required.

In summary the approach provides the following. First, the parameter draws $\theta^{(g)}$ converge to the finite-sample MLE estimate $\hat{\theta}$. Given that the approach deals with a fixed sample, the approach inherits all of the classical asymptotic properties (in the sample size) of the MLE. In contrast, Chernozhokov and Hong (2003) propose an estimation procedure different from MLE for which they give asymptotic properties. Second, we show below that, by appropriately scaling the parameter draws and looking at $\psi^{(g)} =$

$\sqrt{J}(\theta^{(g)} - \hat{\theta})$, one obtains an estimate of the observed Fisher's information matrix. Again, the observed Fisher's information matrix converges asymptotically, as is known, to the true information. We also use the simulated distribution of $\psi^{(g)}$ to provide a useful diagnostic of how to choose J . In many cases, due to the data augmentation, our approach will result in a fast mixing chain and a low value of J will be sufficient. Quantile plots will be used to assess the convergence to normality of $\psi^{(g)}$.

2.1 The Choice of J and $\mu(\theta)$

We now discuss the choice of $\mu(d\theta)$ and the properties of $\pi_J^\mu(\theta)$. Recall that we choose the measure $\mu(\theta)$ so that π_J^μ is integrable. In many cases, we can assume that μ is Lebesgue measure and $\mu(\theta) \equiv 1$. However, as mentioned above, $\mu(\theta)$ can be used to avoid non-integrability that may arise in some state models, e.g., the jump model. Three special cases are worth discussing.

- $J = 1$ and $\mu(\theta) = p(\theta)$ is a subjective prior distribution: Then, π is the posterior distribution of the states and parameters given the data. The approach collapses to MCMC Bayes estimation.
- $J = 1$ and $\mu(\theta) = 1$: In this case, there is a danger of non-integrability of the objective function, as in the case of some Bayesian problems with flat priors.
- $J > 1$: As J increases, the effect of $\mu(\theta)$ disappears on the range of values where it assigns positive mass.

To understand the role of the dominating measure, we now provide two illustrative examples documenting how the dominating measure, $\mu(\theta)$, and raising the likelihood to the power J can be used to overcome deficiencies in the likelihood. Both examples are highly stylized. Since the marginal likelihood is rarely available in closed form in latent variable models, it is difficult to find examples among commonly analyzed models.

Consider first the simplest random volatility model: $y_t = \sqrt{V_t}\varepsilon_t$, $\varepsilon_t \sim N(0, 1)$, and $V_t \sim \mathcal{IG}(\alpha, \beta)$, where α is known and $\mathcal{IG}$ denotes the inverse Gamma distribution. The joint distribution of volatilities and parameters, $\pi(V, \beta)$ is

$$\pi(V, \beta) \propto \prod_{t=1}^T \frac{1}{V_t} e^{-\frac{y_t^2}{2V_t}} \beta^\alpha V_t^{\alpha+1} e^{-\frac{\beta}{2V_t}}$$

which implies that the marginal likelihood for β is

$$\pi(\beta) \propto \prod_{t=1}^T \left(\frac{\beta}{y_t^2 + \beta} \right)^\alpha.$$

$\pi(\beta)$ does not integrate in the right tail for any α . In this case, a dominating measure that downweights the right tail is required to generate a well-defined likelihood. A similar degeneracy occurs a time-discretization of Merton's (1976) jump-diffusion model. In this model, when one of the volatility parameters is driven to zero, the likelihood function increase without bound, and thus the likelihood has no maximum. In this case, $\mu(\theta)$ will bound this parameter away from the origin.

The second case is a two-factor volatility model, where $y_t = v_t + \sigma\varepsilon_t$, $v_t \sim N(0, \tau_t^2)$. The joint distribution is

$$\pi(\tau, \sigma) \propto \left[\prod_{t=1}^T \frac{1}{\sqrt{\tau_t^2 + \sigma^2}} \right] e^{-\sum_{t=1}^T \frac{y_t^2}{(\tau_t^2 + \sigma^2)}}.$$

In the right tail of the distribution (for a fixed σ), $\pi(\tau, \sigma)$ behaves like τ_t^{-1} , which is not integrable. On the other hand, $\pi^J(\tau, \sigma)$ behaves like τ_t^{-J} in the tail and is integrable. These examples provide an example how the dominating measure and raising the likelihood to a power can overcome integrability problems. It is difficult to make general statements regarding these issues as integrability is model dependent and, in models with latent variables, one is rarely able to integrate the likelihood analytically (the Gamma

stochastic volatility model above is one example).

2.2 Convergence Properties of the Algorithm

This section describes the convergence properties of the Markov chain as a function of g and the augmentation parameter J . First, for a fixed J , we have the standard MCMC convergence properties; $\{\tilde{X}^{J,(g)}, \theta^{(g)}\}_{g=1}^G \rightarrow \pi_J(\tilde{X}, \theta)$ as $G \rightarrow \infty$, see Casella and Robert (2002). Then as J increases, we can argue heuristically as follows.

Using insights from simulated annealing, for example, Pincus (1968) and Robert and Casella (1999), for sufficiently smooth densities, we also know that

$$\lim_{J \rightarrow \infty} \frac{\int \theta \mathcal{L}_T(\theta)^J \mu(d\theta)}{\int \mathcal{L}_T(\theta)^J \mu(d\theta)} = \hat{\theta}$$

where $\hat{\theta}$ is the maximizer of $\mathcal{L}_T(\theta)^J$, the MLE. Hence, following Van Laarhoven and Aarts (1987), for a suitable chosen sequence $J^{(g)}$, we have that $\lim_{J,g \rightarrow \infty} \theta^{J^{(g)}} = \hat{\theta}$, as g increases.

Consider now the convergence of the distribution of the latent state. Since each vector X^j is conditionally independent, we can first fix j and consider what the convergence of the marginal distribution of X_t^j . As $g \rightarrow \infty$, we have that

$$p(X_t^j) = E_{\theta^{(g)}}[p(X_t^j | \theta^{(g)})] \rightarrow p(X_t^j | \hat{\theta}),$$

which implies the algorithm recovers the exact smoothing distribution of the state variables. The argument underlying this is as follows. First, by the Ergodic Theorem, we know that for any function with a finite mean, we have

$$\frac{1}{G} \sum_{g=1}^G f(X^{J^{(g)}}, \theta^{(g)}) \rightarrow E[f(X^J, \theta)].$$

Applying this for a fixed time t , we have that

$$\frac{1}{G} \sum_{g=1}^G p \left(X_t^{j,(g)} | \theta^{(g)} \right) \rightarrow E_{\theta^{(g)}} [p(X_t^j, \theta)].$$

Since $\theta^{(g)} \rightarrow \hat{\theta}$, we also have that $p(X_t^j) = \lim_{J,g \rightarrow \infty} p(X_t^j | \theta^{(g)}) = p(X_t^j | \hat{\theta})$. Hence, each of the latent variable draws comes from the smoothing distribution of X_t^j conditional on $\hat{\theta}$.

We now consider the asymptotic in J more formally. We proceed as follows. Let $\sigma_T = V(\hat{\theta})^{-\frac{1}{2}}$ where $V(\hat{\theta}) = \mathcal{L}_T''(\hat{\theta})$ be the observed sample standard deviation. Formally, we show the following result.

Theorem: *Suppose that the following regularity conditions hold:*

(A1) *The density $\mu(\theta)$ is continuous and positive at $\hat{\theta}_T$;*

(A2) *$\mathcal{L}_T(\theta)$ is almost surely twice differentiable in some neighborhood of $\hat{\theta}_T$;*

(A3) *Define the neighborhood $N_{\hat{\theta}_T}^{(a,b)}(J) = \left(\hat{\theta}_T + \frac{a\sigma_T}{\sqrt{J}}, \hat{\theta}_T + \frac{b\sigma_T}{\sqrt{J}} \right)$. For any (a,b) there exists a J such that for $\theta \in N_{\hat{\theta}_T}^{(a,b)}(J)$ there exists a ϵ_J where $\epsilon_J \rightarrow 0$ as $J \rightarrow \infty$ such that $\sup_{N_{\hat{\theta}_T}^{(a,b)}(J)} |R_T(\theta)| < \epsilon_J < 1$, where $R_T(\hat{\theta}_T) = - \left(\mathcal{L}_T''(\hat{\theta}) \right)^{-1} \left\{ \mathcal{L}_T''(\theta^{(g)}) - \mathcal{L}_T''(\hat{\theta}_T) \right\}$.*

Then,

$$\psi^{(g)} = \sqrt{J^{(g)}} \left(\theta^{(g)} - \hat{\theta}_T \right) \Rightarrow N \left(0, V(\hat{\theta}_T) \right).$$

Hence,

$$\text{Var}(\psi^{(g)}) \rightarrow V(\hat{\theta})$$

Proof: See the Appendix.

2.3 Details of the MCMC algorithm

To simulate the joint distribution, (X^J, θ) , standard MCMC techniques can be used, see, e.g. Johannes and Polson (2004) for a survey. MCMC algorithms typically use the Gibbs sampler or the Metropolis algorithm. In the case of the Gibbs sampler, at step

$g+1$, we generate independent draws of each copy $j = 1, \dots, J$ of the state variable vector:

$$X^{j,(g+1)} \sim p(X^j | \theta^{(g)}) \propto p(Y | \theta^{(g)}, X^j) p(X^j | \theta^{(g)}),$$

and a single draw of the parameter given the J copies.

$$\theta^{(g+1)} | \tilde{X}^{J,(g)}, Y \sim \prod_{j=1}^J p(Y | \theta^{(g)}, X^{j,(g+1)}) p(X^{j,(g+1)} | \theta^{(g)}).$$

The key step is verifying that $p(\theta | X, Y)$ is integrable and that the Hammersley-Clifford theorem applies.

A useful alternative to the Gibbs sampler approach outlined above is a Metropolis approach. The advantage is that Metropolis allows the use of information across the J samples in updating the remaining states. Instead of drawing each of the X^j 's independently, we first could propose from a transition kernel

$$Q(X^{(g+1)}, X^{(g)}) = \prod_{j=1}^J Q(X^{j,(g+1)}, X^{(g)})$$

and accepting with the probability

$$\alpha(X^{(g)}, X^{(g+1)}) = \min \left[1, \frac{p(X^{(g+1)} | \theta, Y) Q(X^{(g)}, X^{(g+1)})}{p(X^{(g)} | \theta, Y) Q(X^{(g+1)}, X^{(g)})} \right].$$

The key here is that the Metropolis kernel can depend on the entire history of the X 's. Unlike typical Metropolis algorithms, this generates a time-inhomogeneous Markov Chain, but standard results in the convergence of time inhomogeneous chains still apply, again, see Van Laarhoven and Aarts (1987). The intuition for this result is as follows. Consider the typical random walk Metropolis proposal, $X^{j,(g+1)} = X^{j,(g)} + \tau \varepsilon$. A well-known problem with this algorithm is that the random walk step can wander too far and the choice of τ is problematic. Drawing instead, using the information in the other J

samples:

$$X^{j,(g+1)} = \frac{1}{J} \sum_{i=1}^J X^{i,(g)} + \tau \varepsilon$$

and similarly, allows us to adjust the variance of the random walk error. Johannes and Polson (2004) provide a further discussion of Metropolis algorithms in general, and more specifically, scaling random-walk algorithms.

We now determine the analytical properties of $\pi_J(\tilde{X}^J, \theta)$ and the MCMC algorithm for two important latent state models in financial econometrics; a stochastic volatility and a multivariate jump-diffusion models.

3 Application to the Stochastic Volatility Model

We first consider the benchmark log-stochastic volatility model. Here, returns r_t follow a latent state model of the form:

$$\begin{aligned} y_t &= \log(S_t/S_{t-1}) = \sqrt{V_t} \varepsilon_t \\ \log(V_t) &= \alpha + \delta \log(V_{t-1}) + \sigma u_t, \end{aligned}$$

where V_t is an unobserved volatility, S_t is the continuously-compounded return, the shocks are uncorrelated, and $\theta = (\alpha, \delta, \sigma)$ is the parameter governing the evolution of volatility. This model has been analyzed with a number of different econometric techniques: MCMC (Jacquier, Polson and Rossi (1994)), GMM (Taylor (1986), Andersen, Sorensen and Chung (1998)), simulated maximum likelihood (Danielson (1995)) and simulated method of moments (Gallant, Hsieh, and Tauchen (1997)).

Direct MLE is impossible because the marginal likelihood, $\mathcal{L}_T(\theta) = \int p(Y|V, \theta) p(V|\theta) dV$, where V is the time series of latent volatilities, is not known in closed form as V is a high-dimensional vector and its distribution is non-standard. Maximizing an approximate likelihood is also complicated as $p(V|\theta)$ is a T -dimensional distribution and the

evolution for V_t is non-normal and non-Gaussian. Moreover, MLE does not provide a method to estimate the latent states. In the next subsection, we describe our MCMC maximum likelihood approach.

3.1 Algorithm

In this subsection, we derive the conditional distributions required for the algorithm. In this setting, we need to be able to draw from $p(\theta|\tilde{V}^J, Y)$ and $p(V^j|\theta, Y)$ for $j = 1, \dots, J$. To derive these, let $V^j = [V_1^j, \dots, V_T^j]$ be a $1 \times T$ vector of volatilities, $V_{t,J} = [V_t^1, \dots, V_t^J]'$ is a vector of J copies of V_t and $\tilde{V}^J = [V^1, \dots, V^J]'$ is the $J \times T$ matrix of stacked volatilities.

$$\log(V_{t,J}) = \alpha + \delta \log(V_{t-1,J}) + \sigma_v u_{t,J} = [\mathbf{1}_J, \log(V_{t-1,J})] \begin{pmatrix} \alpha \\ \delta \end{pmatrix} + \sigma_v u_{t,J},$$

where $\mathbf{1}_J$ is a $J \times 1$ vector of 1's. If we denote $X_{t-1} = [\mathbf{1}_J, \log(V_{t-1,J})]$. Stacking, we have a regression of the form, $\log V = X\beta + \sigma_v u$, where

$$\begin{pmatrix} \log(V_{1,J}) \\ \vdots \\ \log(V_{T,J}) \end{pmatrix} = \begin{pmatrix} \mathbf{1}_J & \log(V_{0,J}) \\ \vdots & \vdots \\ \mathbf{1}_J & \log(V_{T-1,J}) \end{pmatrix} \begin{pmatrix} \alpha \\ \delta \end{pmatrix} + \sigma_v \begin{pmatrix} u_{1,J} \\ \vdots \\ u_{T,J} \end{pmatrix}.$$

Then we can write the joint density $\pi_J(\tilde{V}^J, \theta)$ of the stacked system:

$$p(Y|\alpha, \delta, \sigma_v, \tilde{V}^J) \propto (\sigma^2)^{-\frac{JT}{2}} \exp\left(\frac{1}{2\sigma_v^2} \left[(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) + S\right]\right)$$

where $S = (\log V - X\hat{\beta})' (\log V - X\hat{\beta})$, and $\hat{\beta} = (X'X)^{-1} \log V$.

The algorithm for the evaluation of the maximum likelihood by MCMC is then:

1. Given initial parameter values, $\theta^{(0)} = (\alpha^{(0)}, \delta^{(0)}, \sigma_v^{(0)})$. Then for $g = 1, \dots, G$
2. $p(V^j | \theta, Y)$: For $j = 1, \dots, J$, draw J copies of the volatilities:

$$(V^j)^{(g+1)} \sim p(V^j | \alpha^{(g)}, \delta^{(g)}, \sigma_v^{(g)}, Y),$$

independently for each j , using for example, a Metropolis Hastings algorithm as in Jacquier, Polson and Rossi (1994).

3. $p(\theta | \tilde{V}^J, Y)$: This is a simple regression. One can update the parameters in two steps. First update $(\alpha, \delta | \sigma_v, \tilde{V}^J)$, then draw from $p(\sigma_v | \tilde{V}^J)$. Namely, we draw $\alpha^{(g+1)}, \delta^{(g+1)}$ from

$$p(\alpha, \delta | \sigma_v, \tilde{V}^J) \sim N(\hat{\beta}^{(g+1)}, (\sigma_v^2)^{(g)} [X^{(g+1)'} X^{(g+1)}]^{-1})$$

where $\hat{\beta}^{(g)}$ is the OLS estimate $(X^{(g+1)'} X^{(g+1)})^{-1} X^{(g+1)'} \log V^{(g+1)}$. Similarly, we draw $\sigma^{(g+1)}$ from

$$p(\sigma | \alpha^{(g+1)}, \delta^{(g+1)}, \tilde{V}^{J, (g+1)}) \propto IG(J, S^{(g+1)})$$

One noticeable feature of the algorithm is that there is no need for priors for $J \geq 2$. This occurs because the posterior for σ_v , $p(\sigma_v | \alpha, \delta, \tilde{V}^J)$, is proper for $J \geq 2$. This contrast this with standard MCMC where we require a proper prior for σ_v .

3.2 Performance

We demonstrate the behavior of the algorithm for the basic stochastic volatility model with parameters $\alpha = -0.363$, $\delta = 0.95$, $\sigma_v = 0.26$. These parameters are consistent with empirical estimates for financial equity return series and often used in simulation studies. We simulate one series of $T = 1000$ observations. We then run the MCMC algorithm for

$N = 25000$ draws, for four different values of $J = 1, 2, 10, 20$. For $J = 1$, the algorithm is essentially identical to that in JPR (1994). That is, it converges to draws of the posterior distribution of the parameters and volatilities. In the other cases, as J becomes large, the algorithm produces a sequence of draws of the parameters which average converges to the ML point estimate. As J increases, the variance of the sequence of draws decreases.

The convergence results in the previous sections indicate that, as J increases, the draws will converge to the MLE. The theory does not however indicate at what rate this occurs. In practice, these results would not be very useful if an inordinately high J was required for the algorithm to approach the MLE. We now empirically show that this is not the case, at least with the SV and jump-diffusions models. Namely, the algorithm is quite effective for even moderate values J . To study this we look at the distribution of the scaled draws of $\psi^{(g)}$. Recall that looking at whether $\theta^{(g)}$ is close or far from the true θ is not useful in itself since $\theta^{(g)}$ for large g is only an estimate of $\hat{\theta}$ which we know converges to the true parameter only with the sample size.

Figure 1 shows these sequences of draws of δ for the four runs of the algorithm with $J = 1, 2, 10, 20$. Each draw of delta is conditional on a vector of volatilities of length TJ . The plots confirm that an increasing J quickly reduces the variance of such draws, with a dramatic effect on the resulting sequence. Figure 2 shows a similar result for σ_v . Note that even for J as large as 20, the algorithm dissipates initial conditions very quickly. Although the variance of the draws vanishes as J increases, there is no fear that the algorithm will fail to move promptly to the MLE. The horizontal lines show the true parameter value as well as the average of the draw sequence for the last 24000 observations. While the estimate of the MLE for σ_v , $J = 20$, appears close to the true value of 0.26, that for δ is quite different from 0.95. These results, obtained for this one sample do not constitute a sampling experiment. One expects the MLE itself to converge to the true value only as the sample size T goes to infinity.

These first two figures indicate that even a moderate increase in J , from 1 to 10, rad-

ically changes the nature of the algorithm. We complete this diagnostic by documenting the rate at which the draws converge to normality in distribution. The left and right plots in Figure 3 show the normal probability plots for δ and σ_v , and $J = 1, 2$, and 10. Recall that for $J = 1$, the algorithm produces the Bayesian estimator. Panels (a) and (d), $J = 1$, exhibit very strong non-normality and skewness, reflecting the well known non-normality of the posterior distribution. The convergence to normality is all the more remarkable as J increases to 2, panels (b) and (e), and then 10, panels (c) and (f). With as few as J states, the algorithm produces samples of draws very close to normality, which is consistent with a very rapid convergence to the MLE as J increases.

While Figures 1 to 3 confirm that the algorithm effectively estimates the MLE for even small values of J , we now turn to the effect of J on the volatility smoother itself. Again, for $J = 1$, the algorithm produces the Bayes estimator of the volatility. For example, averaging the draws yields a Monte Carlo estimate of the posterior mean. Consider a specific draw, for example the last one $G = 25000$. For this draw, the algorithm makes J draws of V_t . Compute the average of these J draws. Figure 4 plots the time series of these draws of the V_t 's for the four values of J . Panel (a) - the Bayesian case of $J = 1$, shows the large noise in one draw. This noise obscures time series pattern in the true V_t 's. As the J -averaging is effected for an increasing J , the noise decreases and a time series pattern of the volatilities emerges. Figure 4 shows this happen immediately for relatively small values of J . Indeed, panels (c) and (d) - $J = 10$ and 20, present dramatically similar time series even though they plot two unrelated draws from two independent runs. This tremendous increase in precision is at the core of the commensurate increase in precision in the draws of the parameters α, δ, σ_v . Recall that the draw of the parameters effectively uses the information in JT volatilities.

The algorithm also produces smoothed estimates of the volatilities V_t : simply average all the draws of V_t over both J , as in Figure 4, and N . Figure 5 shows that the resulting smoothed estimates are nearly identical for all values of J . Panel (a) follows from an

averaging over $G = 25000$ draws while panel (c) is over $GJ = 250000$ draws. Their smoothed estimates are identical because the precision in the averaging in the Bayesian case is high enough to make any further increase in precision - via a larger J , insignificant. Effectively, the small changes in the parameter estimates make even smaller changes in the volatility estimates. This is confirmed in Figure 6 which shows the plot of the smoothed volatilities versus the true volatilities. Panel (a) and (b) represent $J = 1$ and 2. They are identical, with a cross-correlation of 0.74. So, our algorithm preserves the efficient smoother originally produced by the Bayesian MCMC algorithm. This is in sharp contrast with an approximate smoother which would for example substitute the MLE of the parameters into a Kalman filter.

4 Application to Merton's Jump-Diffusion Model

A multi-variate version of Merton's jump-diffusion model specifies that a vector of asset prices, S_t , solve the stochastic differential equation:

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

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, $Z_j \in R^K$ $Z_j \sim N(\mu_z, \Sigma_z)$. Solving this stochastic differential equation, continuously compounded equity returns (Y_t) over a daily interval are

$$Y_{t+1} = \mu + \sigma (W_{t+1} - W_t) + \sum_{i=N_t}^{N_{t+1}} Z_j$$

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

In the univariate case, this model is commonly used for option pricing, following Merton (1976). The model is maybe more useful in the multivariate case. For a vector

of risky assets, the multivariate jump model generates fat tails and allows for differential correlation structure between ‘normal’ movements (σ) and large market movements (Σ_z). For example, this allows for large returns to be more highly correlated than small returns. Duffie and Pan (2001) provide closed form approximations for the value-at-risk (VAR) of a portfolio of the underlying or of a portfolio of options. They do not provide an empirical analysis of the model.

Likelihood based estimation of the model is extremely difficult for two reasons. First, random mixtures of normal have well-known degeneracies in 1-dimension, and the degeneracies are likely much worse in higher dimensions. Second, even for moderate K , there are a large number of parameters and gradient based optimization of a complicated likelihood surface is rarely attempted. Our approach, relies on the insights of SA, which is often a preferable method to optimize complicated functions. The next section describes the algorithm and provides simulation based evidence on the algorithm’s performance.

4.1 Algorithm

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

$$Y_{t+1} = \mu + \sigma (W_{t+1} - W_t) + I_{t+1}Z_{t+1}$$

where $P[I_t = 1] = \lambda \in (0, 1)$ 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

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

Our MCMC algorithm samples from $p(\theta, X|Y) = p(\theta, I, \xi|Y)$ where I 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

$$p(Y|\theta, I, \xi) = \prod_{t=1}^T p(Y_t|\theta, I_t, \xi_t)$$

where $p(Y_t|\theta, I_t, \xi_t) = N(\mu + \xi_t I_t, \Sigma)$ 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 multi-modalities. Multivariate mixtures are even more complicated and direct maximum likelihood is rarely attempted.

Assuming standard conjugate prior distributions for the parameters,

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

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

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

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

MCMC algorithm: sample from $p(\theta, \tilde{I}, \tilde{Z}|Y)$ by the iterative Gibbs sampler

$$\begin{aligned}
\theta^{(g+1)} & \sim p(\theta|\tilde{I}^{(g)}, \tilde{Z}^{(g)}, Y) \\
\tilde{I}^{(g+1)} & \sim p(\tilde{I}|\theta^{(g+1)}, \tilde{Z}^{(g)}, Y) \\
\tilde{Z}^{(g+1)} & \sim p(\tilde{Z}|\theta^{(g+1)}, \tilde{I}^{(g+1)}, Y)
\end{aligned}$$

where we note that the last two draws are just J draws from the same distribution.

4.2 Performance

We next consider a three-dimensional version of Merton's model introduced earlier, we simulate a vector of 1000 returns using the following parameter values, scaled to daily units: $\mu = (0.2, 0.15, 0.1)$, $\mu_z = (-3, -3.5, -4)$, $\lambda = 0.10$, $\sigma_z^{11} = 3$, $\sigma_z^{22} = 4$, $\sigma_z^{33} = 5$, $\sigma^{11} = 1.5$, $\sigma^{22} = 1$, $\sigma^{33} = 0.5$, the off diagonal elements of the diffusive and jump-covariance matrix are such that the diffusive or jump covariance between two assets is fifty percent. These parameters are typically to those that would be found of an analysis of large, volatile equity indices in the United States.

We report results of sampling experiments similar to those in the previous subsection. Figures 7 to 10 display a summary of the MCMC output for $G = 5000$, and $J = 1, 2, 10$ and 20 . The results are largely consistent with those seen for the SV model. For example, consider Figure 7, which contains trace plots for the parameters for λ . As J increases, the variability of the draws reduces drastically, collapsing on the true value of 0.10 . As a comparison, the volatility of the draws for λ for $J = 1$ is 0.00705 and 0.00160 , a reduction of approximately 4.41 . This is right in line with the implications of our central limit theorem as $\sqrt{20} = 4.47$. Figure 8 provides QQ or normality plots the draws for $J = 1, 2, 10$ and 20 . As J increases, the draws also converge to their limiting standard normal distribution.

Figure 9 and 10 provide trace plots for two other parameters of interest, μ_z^1 and α_z^1 . In both of these figures, the plots collapse as predicted, albeit with slight biases. For example, for μ_z^1 the average of the draws is roughly -2.7 , lower than the true value of -3 . Similarly, for σ_z^1 , the mean of the draws for $J = 20$ is 3.15 , which is slightly above the true value of 3.0 . In both cases, the estimates appear to a slightly biased. However, there is no reason to believe that either Bayes estimators (in the case $J = 1$) or maximum likelihood estimators (as J increases) maximum likelihood are unbiased in finite samples. Both estimators are asymptotically unbiased, and the divergence in our sampling experiments of the estimates from their true value is merely an implication of this finite sample bias.

5 Conclusion

In this paper, we develop MCMC algorithms for computing finite sample MLE and standard errors in latent state models. The likelihood function requires marginalization of the state variables and then optimization over the parameters. We design MCMC algorithms that simultaneously perform the integration and optimization.

Our approach makes use of data augmentation and evolutionary MCMC. We use MCMC methods to simulate from a high-dimensional joint distribution $\pi^\mu(\theta, \tilde{X})$ that

arises for the parameters and J copies of the latent state variables. We also discuss issues of how to avoid singularities in the marginal likelihood by a suitable choice of the dominating measure for this joint density π_J^μ . We estimate the multivariate jump-diffusion model that illustrates this issue. We also estimate a stochastic volatility model. While our asymptotics is in J , our implementation provides convincing evidence that convergence occurs quickly for low values of J .

There are at least three directions in which the current work can be extended. First, it would be interesting to extend our work to the case where the MLE is on the boundary of the parameter space using similar arguments as in Erkanli et al. (1994). Second, while we focussed on latent state variable models in financial econometrics, there are numerous models throughout economics with latent state variables. Often, researchers do not even consider likelihood based estimation schemes, turning instead to various simulated method of moments estimators. Our approach provides an attractive alternative to method-of-moments based estimators.

Third, as mentioned in the introduction, in addition to econometrics, there are many problems that require joint integration and optimization. For example, solving for optimal portfolio rules in myopic or dynamic settings with parameter uncertainty require integrating the parameter uncertainty out of the utility. Jacquier, Johannes, and Polson (2004) uses the algorithms presented here to construct maximum expected utility portfolios, for example, optimal portfolios under stochastic volatility.

References

- Besag, J., 1974, Spatial Interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society, Series B* 36, 192-326.
- Robert, C. and G. Casella, 1999, *Monte Carlo Statistical Methods*, Springer-Verlag, New York.
- Chernozhukov, V. and H. Hong, 2003, Likelihood inference for some nonregular econometric models, *Journal of Econometrics* 115, 293-346.
- Dawid, A. P., 1970. On the limiting normality of posterior distributions. *Proceedings of the Cambridge Philosophical Society* 67, 625-633.
- Dempster, A.P., N.M. Laird and D. B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B*, 339, 1-38.
- Doucet, A., Godsill, S and C. Robert, 2002, Marginal maximum posterior estimation using Markov Chain Monte Carlo, *Statistics and Computing* 12, 77-84.
- Erkanli A., 1994, Laplace Approximations for Posterior Expectations When the Mode Occurs on the Boundary of the Parameter Space, *Journal of the American Statistical Association* 89, 250-258.
- Gallant, A. Ronald, David A. Hsieh and George Tauchen, 1997, Estimation of Stochastic Volatility Models with Diagnostics, *Journal of Econometrics*, 81(1), 159-192.
- Geyer, C.J., 1991, Markov Chain Monte Carlo Maximum Likelihood. In: Keramidas E.M.(Ed.), *Computing Science and Statistics: Proceedings of the 23rd symposium on the Interface*, Interface Foundation, 156-163
- Heyde, C., and I.M. Johnstone, 1979, On asymptotic posterior normality for stochastic processes, *Journal of the Royal Statistical Society Series B*, 41, 184-189.
- Jacquier, E., Johannes, M., and N.G. Polson, 2003, MCMC Methods for Expected Utility Problems, working paper.
- Jacquier, E., Polson, N.G., and P. Rossi, 1994, Bayesian Analysis of Stochastic Volatility Models (with discussion). *Journal of Business and Economic Statistics* 12, 4, 371-417.

- Jacquier, E., Polson, N.G., and P. Rossi, 2003, Bayesian Analysis of Stochastic Volatility Models with Fat Tails and Leverage Effects, forthcoming *Journal of Econometrics*.
- Johannes, M and N.G. Polson, 2004, MCMC methods for Financial Econometrics. To appear in *Handbook of Financial Econometrics*, Y. Ait-Sahalia and L. Hansen eds.
- Johannes, M and N.G. Polson, 2004, Discussion of Pastorello S., Patilea V., and E. Renault, "Iterative and Recursive Estimation in Structural Non-Adaptive Models," *Journal of Business and Economic Statistics* 21, No 4, 449-509.
- Kiefer, N.M., 1978, Discrete parameter variation: efficient estimation of a switching regression mode, *Econometrica*, 46:427-434, 1978.
- Kirkpatrick, S., 1984, Optimization by simulated annealing: quantitative studies, *Journal of Statistical Physics* 34, 975-986.
- Kirkpatrick, S., C.D. Gelatt and M.P. Vecchi, 1983, Optimization by simulated annealing, *Science* 220, 671-680.
- Liang, F., and W.H. Wong, 2001, Real Parameters Evolutionary Monte Carlo with Applications to Bayesian Mixture Models, *Journal of the American Statistical Association* 96, 653-666.
- Lindley, D., 1976, A class of utility functions, *Annals of Statistics* 4, 1-10.
- Liu, J., Liang F., and W.H. Wong, 2001, A theory for Dynamic Weighting in Monte Carlo Computation, *Journal of the American Statistical Association* 96, No.454, 561-573.
- Mueller, P., 2000, Simulation based optimal design, *Bayesian Statistics 6*, Bernardo, et al. eds., (Oxford).
- Pastorello S., Patilea V., and E. Renault, 2003, Iterative and Recursive Estimation in Structural Non-Adaptive Models, *Journal of Business and Economic Statistics* 21, No 4, 449-509.
- Pincus, M., 1968, A Closed Form Solution of Certain Programming Problems, *Operation Research* 18, 1225-1228.
- Polson, N., 1996, Convergence of MCMC algorithms, *Bayesian Statistics 5*, Bernardo, et

al. eds., Oxford, 297-321.

Schervish M.J., 1995, *Theory of Statistics*, Springer Verlag, New York.

Taylor, S., 1986, *Modeling Financial Time Series*, Wiley, New York.

Van Laarhoven, P.J., and E.H.L Aarts, 1987, *Simulated Annealing: Theory and Applications*, CWI Tract 51, Reidel, Amsterdam.

Appendix: Convergence in J

First, define $\ell_T(\theta) \equiv \log \mathcal{L}_T(\theta)$, and write the target marginal density

$$\pi_J(\theta) = \frac{\exp^{J\ell(\theta)} \mu(\theta^{(g)})}{m_J}$$

where $m_J = \int \exp^{J\ell_T(\theta)} \mu(\theta^{(g)}) d\theta$. Now write

$$\pi_J(\theta) = \pi_J(\hat{\theta}) e^{J(\ell(\theta) - \ell(\hat{\theta}))}.$$

By Taylor's theorem we can write

$$\ell(\theta) = \ell(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})\ell''(\theta^*)$$

where $\theta^* = \theta + \gamma(\hat{\theta} - \theta)$ where $0 < \gamma < 1$.

Let $R_n(\theta) = \sigma_T^2(\ell''(\theta^*) - \ell''(\hat{\theta}))$. Then we can write

$$\pi_J(\theta) = \pi_J(\hat{\theta}) e^{-\frac{J}{2\sigma_T^2}(\theta - \hat{\theta})^2(1 - R_n(\theta))}$$

By definition $\psi_J = \sqrt{J}\sigma_T^{-1}(\theta - \hat{\theta})$ and we need to show that for any $a < b$ we have that

$$\lim_{J \rightarrow \infty} P(a < \psi_J < b) = \Phi(b) - \Phi(a)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Now

$$P(a < \psi_J < b) = P\left(\hat{\theta} + \frac{a\sigma_T}{\sqrt{J}} < \theta < \hat{\theta} + \frac{b\sigma_T}{\sqrt{J}}\right)$$

Let

$$= \int_{N_{\hat{\theta}}^{(a,b)}(J)} \frac{e^{J\ell(\theta)}}{m_J} \mu(\theta) d\theta$$

Now for any $\varepsilon > 0$, by continuity of $\mu(\theta)$ (assumption (A1)) at $\hat{\theta}$ we can find J so that

$$(1 - \varepsilon) < \inf_{N_{\hat{\theta}}^{(a,b)}(J)} \frac{\mu(\theta)}{\mu(\hat{\theta})} < \sup_{N_{\hat{\theta}}^{(a,b)}(J)} \frac{\mu(\theta)}{\mu(\hat{\theta})} < 1 + \varepsilon$$

Hence we need only consider

$$I = \int_{N_{\hat{\theta}}^{(a,b)}(J)} \frac{e^{J\ell(\theta)}}{m_J} d\theta = \frac{\pi_J(\hat{\theta})}{m_J} \int_{N_{\hat{\theta}}^{(a,b)}(J)} e^{-\frac{J}{2\sigma_n^2}(\theta - \hat{\theta})^2(1 - R_n(\theta))} d\theta$$

By assumption (A3) we can also find J such that there exists an $0 < \varepsilon_J < 1$ where $\varepsilon_J \rightarrow 0$ as $J \rightarrow \infty$ and

$$\sup_{N_{\hat{\theta}}^{(a,b)}(J)} |R_T(\theta)| < \varepsilon_J < 1$$

Therefore

$$\int_{N_{\hat{\theta}}^{(a,b)}(J)} e^{-\frac{J}{2\sigma_T^2}(\theta - \hat{\theta})^2(1 - \varepsilon_J)} d\theta < I < \int_{N_{\hat{\theta}}^{(a,b)}(J)} e^{-\frac{J}{2\sigma_T^2}(\theta - \hat{\theta})^2(1 + \varepsilon_J)} d\theta$$

now as $N_{\hat{\theta}}^{(a,b)}(J) = \left(\hat{\theta} - \frac{a\sigma_T}{\sqrt{J}}, \hat{\theta} + \frac{a\sigma_T}{\sqrt{J}}\right)$ we have

$$\int_{N_{\hat{\theta}}^{(a,b)}(J)} e^{-\frac{J}{2\sigma_T^2}(\theta - \hat{\theta})^2(1 + \varepsilon_J)} d\theta = \sqrt{2\pi}\sigma_T(1 + \varepsilon_J)^{-\frac{1}{2}} \left(\Phi(\sqrt{J}\sigma_T^{-1} \left(\frac{b\sigma_T}{\sqrt{J}}\right)) - \Phi(\sqrt{J}\sigma_T^{-1} \left(\frac{a\sigma_T}{\sqrt{J}}\right)) \right)$$

Taking the limit as $J \rightarrow \infty$ and noting that $\varepsilon_J \rightarrow 0$ we have that

$$P(a < \psi_J < b) \rightarrow \Phi(b) - \Phi(a)$$

as required.

Figure 1: Draws of δ for the SV model, $\delta = 0.95$, $T = 1000$, $G = 25000$

Figure 2: Draws of σ for the SV model, $\sigma = 0.26, T = 1000, G = 25000$

Figure 3: Normality of draws of δ and σ , SV model, $T = 1000, G = 25000$

Figure 4: Last draw ($G = 25000$) of $\sqrt{v_t}$, SV model, $\delta = 0.95$, $T = 1000$

Figure 5: Smoothed estimates of $\sqrt{v_t}$, SV model, $\delta = 0.95$, $T = 1000$, $G = 25000$

Figure 6: Smoothed estimates versus true $\sqrt{v_t}$, SV model $\delta = 0.95, T = 1000, G = 25000$

Figure 7: MCMC draws for the jump intensity, λ , from the three-dimensional Merton's model for $G = 5000$, and the cases $J = 1, 2, 10$, and 20 . The true $\lambda = -0.10$.

Figure 8: QQ or normality plots for the MCMC draws for the jump intensity, λ , from the three-dimensional Merton's model for $G = 5000$, and the cases $J = 1, 2, 10$, and 20 .

Figure 9: MCMC draws for the jump intensity, μ_z^1 , from the three-dimensional Merton's model for $G = 5000$, and the cases $J = 1, 2, 10$, and 20 . The true value is -3 .

Figure 10: MCMC draws for the jump intensity, σ_z^1 , from the three-dimensional Merton's model for $G = 5000$, and the cases $J = 1, 2, 10$, and 20 . The true value is 3.