Exact likelihood inference for C.I.R. stochastic volatility models[1]

Drew D. Creal

Booth School of Business, University of Chicago

May 23, 2012

PRELIMINARY DRAFT: COMMENTS WELCOME

Abstract

This paper introduces new methods for filtering, smoothing, and likelihood evaluation for affine stochastic volatility models. Affine stochastic volatility models are built from the discrete-time autoregressive gamma process, whose transition density over any interval of time is the same as the Cox-Ingersoll-Ross or Feller diffusion. The new methods are based on the insight that it is possible to integrate out the latent variance analytically leaving only a latent discrete variable. The likelihood function of the model can be calculated accurately up to computer tolerance. This allows for the estimation of unknown parameters by maximum likelihood and the filtering and smoothing of the latent variance.

Keywords: stochastic volatility; Cox-Ingersoll-Ross; affine models; autoregressive-gamma process.

JEL classification codes: C32, G32.

[1]I would like to thank Alan Bester, Chris Hansen, Robert Gramacy, Siem Jan Koopman, Neil Shephard, and Jing Cynthia Wu as well seminar participants at the Triangle Econometrics Workshop (UNC Chapel Hill), University of Washington, and Erasmus Universiteit Rotterdam for their helpful comments. Correspondence: Drew D. Creal, University of Chicago Booth School of Business, 5807 South Woodlawn Avenue, Chicago, IL 60637. Email: dcreal@chicagobooth.edu

1 Introduction

Stochastic volatility models are common in finance and increasingly in macroeconomics for modeling the uncertain variation of asset prices and macroeconomic time series. There exists a large literature in econometrics dedicated to estimating both discrete and continuous-time stochastic volatility (SV) models; see, e.g. Shephard (2005). The challenge when estimating SV models stems from the fact that the log-likelihood function of the model is a high-dimensional integral that does not have a closed-form solution. The contribution of this paper is a new method that can calculate the log-likelihood function of the model exactly up to computer tolerance for a class of SV models whose stochastic variance follows a Cox, Ingersoll, and Ross (1985) process. The new algorithms are based on the insight that it is possible to integrate out the latent variance analytically leaving only a latent discrete variable. The resulting model can be approximated accurately by a finite state Markov-switching model, whose likelihood can easily be evaluated. I also illustrate how the ideas developed here can be extended to cover other models whose latent variables have a dependence structure like the Cox-Ingersoll-Ross process, including stochastic intensity, stochastic duration, and stochastic transition models.

Stochastic volatility models belong to the class of nonlinear, non-Gaussian state space models. Two classes of state space models that are routinely used in economics are linear, Gaussian models and Markov-switching models. One reason for their popularity is because the log-likelihood function for these models can be computed in closed-form; see, e.g. Kalman (1960) and Hamilton (1989). The Kalman filter for linear, Gaussian models recursively updates the mean vector and covariance matrix of a normal distribution; see, e.g. Harvey (1989) and Durbin and Koopman (2001). The filtering and smoothing algorithms introduced here do not update a small set of sufficient statistics through time such as the Kalman filter and are therefore not closed-form solutions. Nevertheless, the algorithms can compute the solution up to numerical tolerance on a computer.

The results in this paper follow from the properties of the transition density of the Cox, Ingersoll, and Ross (1985) (CIR) process, which in discrete time is known as the autoregressive gamma (AG) process of Gouriéroux and Jasiak (2006). The transition density of the CIR/AG process is well-known to be a non-central gamma (non-central chi-squared up to scale) distribution, which is a Poisson mixture of gamma random variables. The CIR/AG process actually has two state variables. One is the continuous valued variance h_t which is conditionally a gamma random variable and the other is a discrete mixing variable z_t which is conditionally a Poisson random variable.

I integrate the continuous-valued state variable h_t out of the model analytically. This leaves a state space model whose only state-variable is discrete-valued and defined over the set of non-negative integers. I show that the Markov transition distribution of the discrete variable z_t is a Sichel distribution; e.g. Sichel (1974, 1975). Like the Poisson distribution from which it is built, the Sichel distribution assigns increasingly small probability mass to large integers and the probabilities eventually converge to zero. Although the support of the state variable z_t is technically infinite dimensional, it is finite dimensional for all practical purposes. The state space model for the discrete-variable z_t with the variance h_t marginalized out can be approximated by a finite dimensional Markov-switching model whose log-likelihood can be computed from known recursions; see, e.g. Baum and Petrie (1966), Baum, Petrie, Soules, and Weiss (1970), and Hamilton (1989).[2]

Additional objects of interest are the filtered and smoothed distributions of the latent variance h_t. The filtering distribution characterizes the information about the state variable at time t given information up to and including time t whereas the smoothing distribution uses all the available data to retrospectively estimate past values of the variance. An interesting feature of the CIR/AG process is that once one conditions on the discrete mixing variable z_t and the data y_t, the variance h_t is independent of the variance in all other time periods. This characteristic of the model is critical and ultimately drives the results of this paper. The marginal filtering and smoothing distributions of the variance h_t conditional on z_t and the data y_t are shown to be generalized inverse Gaussian (GIG) distributions whose moments can be computed analytically. Solving the filtering and smoothing problem for the variance h_t reduces to a two step problem. First, the marginal filtering and smoothing distributions for the discrete variable z_t are calculated using the algorithms for Markov-switching. Then, moments and quantiles of the filtering distribution for the continuous-variable h_t can be computed by averaging the moments and quantiles of the conditional GIG distribution by the marginal probabilities of z_t calculated

[2]Technically, the accuracy of the computations are limited to the accuracy of the modified Bessel function of the second kind, which is typically around 10^{-14} or 10^{-15} depending on the point at which it is evaluated. Secondly, I note that all non-linear, non-Gaussian state space models whose state variables are continuous are ultimately approximated by finite-state Markov switching models due to the limitations of a computer.

from the Markov switching algorithm.

A popular method in the literature for approximating the likelihood function and filtering distributions for stochastic volatility models are sequential Monte Carlo methods also known as particle filters. Particle filters were introduced into the economics literature by Kim, Shephard, and Chib (1998). Creal (2012b) provides a recent review of the literature on particle filters. The filtering and smoothing recursions developed in this paper can be related to the particle filter. The algorithms have an interpretation as an optimal particle filter but where there are no Monte Carlo methods involved. Specifically, the algorithms introduced here combine marginalization techniques Chen and Liu (2000) and Klass, de Freitas, and Doucet (2005) with the optimal resampling algorithm of Fearnhead and Clifford (2003). Using this analogy, I demonstrate how an optimal particle filter coincides with the solution introduced here.

This paper continues as follows. In Section 2, I discuss the CIR stochastic volatility model and demonstrate how to reformulate it with the latent variance integrated out. The model reduces to a Markov-switching model, whose transition distribution is characterized. In Section 3, I provide the details of the filtering and smoothing algorithms and a discussion of how to compute them efficiently. In Section 4, I compare standard particle filterings to the new algorithms and illustrate their relative accuracy. The new methods are then applied to several financial time series.

2 C.I.R. stochastic volatility models

I consider the class of CIR-SV models for log-prices p_t given by

$$dp_t = (\mu + \beta h_t)\, dt + \sqrt{h_t} dW_{1t} \tag{1}$$

$$dh_t = \kappa\,(\theta_h - h_t)\, dt + \sigma\sqrt{h_t} dW_{2t}, \tag{2}$$

where W_{1t} and W_{2t} are a pair of independent standard Brownian motions, β is a variance risk premium, and μ is a location parameter. The unconditional mean of the variance is θ_h and κ controls the spead of mean reversion. The process for the variance (2) was considered by Cox, Ingersoll, and Ross (1985) as a model for the short term interest rate and is called the CIR process. It was originally analyzed by Feller (1951) and is also known as a Feller or square-root process. The model is a staple of the literature in continuous-time finance because it is a subset of the class of affine models, which have closed form option pricing formulas; see, e.g. Heston (1993).

In discrete time, the CIR process is known as the autoregressive gamma (AG) process of Gouriéroux and Jasiak (2006), who developed many of its properties. The discrete time AG-SV model for an observable time series y_t is

$$\begin{aligned} y_t &= \mu + \beta h_t + \sqrt{h_t}\varepsilon_t \qquad \varepsilon_t \sim \mathrm{N}(0,1) & (3)\\ h_t &\sim \mathrm{Gamma}\left(\nu + z_t, c\right) & (4)\\ z_t &\sim \mathrm{Poisson}\left(\frac{\phi h_{t-1}}{c}\right), & (5)\end{aligned}$$

where I assume that the initial condition is a draw from the stationary distribution $h_0 \sim \mathrm{Gamma}\left(\nu, \frac{c}{1-\phi}\right)$ with mean $\mathrm{E}\left[h_0\right] = \frac{c\nu}{1-\phi}$. The AG-SV model has two state variables h_t and z_t, where the latter can be regarded as an "auxiliary" variable in the sense that is not of direct interest.

Importantly, the AG(1) and the CIR process have the same transition density. Consider changes in the continuous-time variance process (2) over an interval of time τ. By setting $\phi = \exp\left(-\kappa\tau\right), c = \frac{\sigma^2[1-\exp(-\kappa\tau)]}{2\kappa}$, and $\nu = \frac{2\kappa\theta_h}{\sigma^2}$, the autoregressive gamma process (4) and (5) converges to the CIR process (2) as $\tau \to 0$. Integrating z_t out of (4) and (5), the transition density of h_t is

$$p(h_t|h_{t-1};\theta) = h_t^{\nu-1}\frac{1}{c^{\nu}}\exp\left(-\frac{(h_t+\phi h_{t-1})}{c}\right)\sum_{k=0}^{\infty}\left[\frac{h_t^k}{c^k\Gamma\left(\nu+k\right)}\frac{\left(\phi h_{t-1}/c\right)^k}{k!}\right],$$

which is a non-central gamma (non-central chi-squared up to scale) distribution. The conditional mean and variance of this distribution are

$$\begin{aligned} \mathrm{E}\left[h_t|h_{t-1}\right] &= \nu c + \phi h_{t-1}\\ \mathrm{V}\left[h_t|h_{t-1}\right] &= \nu c^2 + 2c\phi h_{t-1}\end{aligned}$$

The conditional heteroskedasticity of h_t as a function h_{t-1} is a well-known feature of the AG/CIR model that the log-normal SV model does not share; see, e.g. Shephard (2005).

The parameter ϕ controls the autocorrelation of h_t because the conditional mean is linear in h_t. The restriction $\phi < 1$ is required for stationarity. An additional restriction that researchers are likely to impose is the Feller condition $\nu > 1$, which guarantees that the process h_t never reaches zero.

Further details on the limiting behavior as well as other properties of the process can be found in Gouriéroux and Jasiak (2006). For the remainder of the paper, I primarily focus on the discrete time model, whose parameters have domains $0 < \phi < 1, \nu > 1$ and $c > 0$. I let $\theta = (\mu, \beta, \nu, \phi, c)$ denote the vector of unknown parameters of the model.[3] Throughout the paper, I use the notation $x_{t-k:t}$ to denote a sequence of variables $(x_{t-k}, \ldots, x_t)$, which may be either observed or latent.

Econometric analysis of the AG-SV model is challenging because the likelihood of the model $p(y_{1:T}; \theta)$ is in the form of a high-dimensional integral

$$p(y_{1:T}; \theta) \quad = \quad \int_0^\infty \ldots \int_0^\infty \sum_{z_1=0}^{\infty} \ldots \sum_{z_T=0}^{\infty} \prod_{t=1}^{T} p(y_t|h_t; \theta) p(h_t|z_t; \theta) p(z_t|h_{t-1}; \theta) p(h_0; \theta) dh_0, \ldots, dh_T.$$

For linear, Gaussian state space models and Markov-switching models, this integral is solved recursively beginning at the initial iteration using the prediction error decomposition; see, e.g. Durbin and Koopman (2001) and Frühwirth-Schnatter (2006). It is the difficulty of computing this integral in closed-form that has led researchers to search for approximations to the log-likelihood function primarily by Monte Carlo methods. SV models of different types have been estimated by simulated maximum likelihood (SML) by among others Durbin and Koopman (1997), Durham and Gallant (2002),....ADD MORE REFERENCES. These papers use different Monte Carlo methods including importance sampling, particle filters, and Markov chain Monte Carlo (MCMC).

Alternatively, researchers avoid direct approximation of the likelihood function and use either Bayesian methods or (possibly) less-efficient estimators of the model's parameters such

[3] If interest centers on the continuous-time model (1) and (2), it is possible to estimate the parameters $\theta = \left(\mu, \beta, \kappa, \theta_h, \sigma^2\right)$. However, one needs to make an assumption about the integrated variance $\int_t^{t+\tau} h(s)ds$ which does not have a closed-form solution for the CIR model. A standard assumption in the literature is to sub-divide time intervals τ into M equal sub-intervals and approximate this integral by a Riemann sum $\int_t^{t+\tau} h(s)ds \approx \sum_{j=1}^{M} h_{t+j} \frac{\tau}{M}$. The discrete time model here assumes values $M = 1$ and $\tau = 1$ whereas in most applications in continuous-time finance the parameters are reported on an annual basis such that $\tau = \frac{1}{256}$.

as the generalized method of moments, efficient method of moments, or indirect inference; see, Shephard (2005) for a survey. The continuous-time CIR-SV model (possibly with jumps and leverage) has been estimated in applications for the term-structure of interest rates and option pricing. Versions of the continuous-time model appear in Eraker, Johannes, and Polson (2003), and Chib, Pitt, and Shephard (2006),....ADD MORE REFERENCES. Most work in the literature estimates the continuous-time version of the model using an Euler discretization (conditional normal approximation) for the dynamics of the variance instead of the exact non-central gamma distribution.

Apart from calculating the likelihood function, additional objects of interest are the posterior distributions $p(h_t|y_{1:t};\theta)$, $p(h_{t+1}|y_{1:t};\theta)$, and $p(h_t|y_{1:T};\theta)$ which summarize knowledge about the latent state variable h_t based on different information sets. Filtering corresponds to an estimate of the state variable at time t given information up to and including time t whereas smoothed estimates use all the available data.

The key insight for developing recursions that can accurately compute the log-likelihood function for the AG-SV model is to recognize that it is possible to integrate the variance h_t out of the model analytically. The only unobserved state variable remaining is the auxiliary variable z_t, whose Markov transition distribution can be characterized. The following proposition describes the dynamics of the model once the variance h_t has been integrated out.

Proposition 1 *The conditional likelihood $p(y_t|z_t;\theta)$, transition distribution $p(z_t|z_{t-1}, y_{1:t-1};\theta)$, and initial distribution are*

$$\begin{aligned} p(y_t|z_t;\theta) &= \textit{Variance Gamma}\left(\mu, \sqrt{\frac{2}{c}+\beta^2}, \beta, \nu + z_t\right) && (6) \\ p(z_t|z_{t-1}, y_{1:t-1};\theta) &= \textit{Sichel}\left(\nu + z_{t-1} - \frac{1}{2}, \frac{\phi}{c}\left(y_{t-1}-\mu\right)^2, \frac{c}{\phi}\left[\frac{2}{c}+\beta^2\right]\right) && (7) \\ p(z_1;\theta) &= \textit{Negative Binomial}(\nu, \phi) && (8) \end{aligned}$$

Definitions for the distributions are available in the appendix.

Proof: See the Appendix.

Once the variance has been integrated out, the model reduces to a variance gamma (VG) distribution with a Markov-switching state variable z_t. The variance gamma (VG) distribution is also known as a generalized (asymmetric) Laplace distribution; see, e.g. Kotz, Kozubowski, and Podgórski (2001). The symmetric version ($\beta = 0$) of the distribution was introduced into the literature by Madan and Seneta (1990) and subsequently extended to the asymmetric distribution ($\beta \neq 0$) by Madan, Carr, and Chang (1998). The VG model for asset returns (without the Markov-switching variable z_t) is extremely popular among practitioners. Like the CIR-SV model, it has closed-form solutions for option prices. It is often advocated as an alternative to the CIR-SV model. This result illustrates how over discrete time intervals the CIR-SV model generalizes the VG model. Consider the mean and variance of the conditional distribution $p(y_t|z_t;\theta)$ which are

$$
\begin{aligned}
\mathrm{E}\left[y_t|z_t\right] &= \mu + c\beta\left(\nu + z_t\right) \quad \text{(CHECK THIS)} \\
\mathrm{V}\left[y_t|z_t\right] &= c\left(\nu + z_t\right) + c^2\beta^2\left(\nu + z_t\right)
\end{aligned}
$$

A large past value of y_{t-1} or z_{t-1} causes more probability mass to be placed on larger values of z_t in the Sichel distribution (7). This causes the variance of the conditional distribution for y_t to increase.

The state variable z_t has a non-homogenous Markov transition kernel known as a Sichel distribution; see, e.g. Sichel (1974, 1975). A Sichel distribution is a Poisson distribution whose mean is a random draw from a generalized inverse Gaussian (GIG) distribution. The support of a Sichel random variable is therefore over the set of non-negative integers. The state variable z_t technically has a countably infinite dimensional support but like the Poisson distribution the probabilities converge to zero for large integer values. The stationarity of the model ensures that there always exists an integer Z_{θ,y_t} such that the probability $p(z_t = Z_{\theta,y_t}|y_{1:t},\theta)$ is zero up to machine tolerance on a computer. The algorithms developed below will take advantage of this property.

Despite analytically integrating h_t from the model, the filtering distributions for z_t cannot be computed recursively in closed-form using a finite set of sufficient statistics like the Kalman filter. Given that this distribution assigns negligible probability mass to large integers, the distribution of $p(z_t|y_{1:t};\theta)$ can be approximated accurately using a finite dimensional $(Z_{\theta,y_t} + 1) \times 1$ vector of probabilities $\hat{p}(z_t|y_{1:t};\theta)$ whose points of support are the integers from 0 to Z_{θ,y_t}. The Markov transition distribution $p(z_t|z_{t-1}, y_{1:t-1};\theta)$ can also be approximated by a

$(Z_{\theta,y_t} + 1) \times (Z_{\theta,y_t} + 1)$ transition matrix $\hat{p}(z_t|z_{t-1}, y_{1:t-1}; \theta)$. Both approximations will be equivalent to the estimates produced by a closed-form solution as $Z_{\theta,y_t} \to \infty$. Computing the probabilities for extremely large values of z_t is, however, computationally wasteful as these values are numerically zero. The empirical applications in Section 4 illustrate how to accurately compute the log-likelihood function with practical values of Z_{θ,y_t}. To obtain a fixed precision across the parameter space, the integer Z_{θ,y_t} at which the distributions are truncated should be a function of the parameters of the model θ and the data y_t. A discussion on how to determine the value of Z_{θ,y_t} is provided further below. In the following, I eliminate dependence of θ and y_t on Z_{θ,y_t} for notational convenience.

Using the finite representation for the distribution of z_t, the model is now equivalent to a Markov-switching model, which have known filtering and smoothing recursions; see, e.g. Baum and Petrie (1966), Baum, Petrie, Soules, and Weiss (1970), and Hamilton (1989). It follows that the log-likelihood function is available and it is straightforward to estimate the unknown parameters θ by maximum likelihood. The marginal filtering $p(z_t|y_{1:t}; \theta)$, one-step ahead predictive $p(z_{t+1}|y_{1:t}; \theta)$, and smoothing distributions $p(z_t|y_{1:T}; \theta)$ can be also computed from these algorithms. Further computational details for these algorithms are discussed in Section 3.

The Markov switching algorithms provide estimates of the auxiliary variable z_t but in applications interest centers on estimation of the latent variance h_t through filtering and smoothing algorithms similar to the linear, Gaussian state space model; see, e.g. Durbin and Koopman (2001). The marginal filtering and smoothing distributions for h_t are not known analytically but moments and quantiles of the distributions can be accurately computed. Accurate computation of these quantities is an implication of the following result.

Proposition 2 *Conditional on the discrete variable and the data, the predictive* $p(h_{t+1}|z_{1:t}, y_{1:t}; \theta)$*, marginal filtering* $p(h_t|z_{1:t}, y_{1:t}; \theta)$*, and marginal smoothing* $p(h_t|z_{1:T}, y_{1:T}; \theta)$ *distributions as well* *as the distribution of the initial condition are*

$$p(h_{t+1}|z_{1:t+1}, y_{1:t}; \theta) = Gamma\left(\nu - \frac{1}{2} + z_{t+1}, c\right) \tag{9}$$

$$p(h_t|z_{1:t}, y_{1:t}; \theta) = GIG\left(\nu + z_t - \frac{1}{2}, (y_t - \mu)^2, \frac{2}{c} + \beta^2\right) \tag{10}$$

$$p(h_t|z_{1:T}, y_{1:T}; \theta) = GIG\left(\nu + z_t + z_{t+1} - \frac{1}{2}, (y_t - \mu)^2, \frac{2(1+\phi)}{c} + \beta^2\right) \tag{11}$$

$$p(h_0|z_{1:T}, y_{1:T}; \theta) = Gamma\,(\nu + z_1, c) \tag{12}$$

Definitions for the generalized inverse Gaussian (GIG) distribution are available in the appendix.

Proof: See the Appendix.

Conditional on the discrete mixing variable and the data, the marginal filtering and smoothing distributions for the variance h_t are either generalized inverse Gaussian (GIG) or gamma distributions. Useful properties of GIG random variables that are needed for computational purposes are discussed in the appendix. The main point of this proposition is that the distribution of the variance h_t is independent of the variance at other time periods conditional on z_t and y_t. This is due to the fact that h_t depends on h_{t-1} only through z_t. This is critical for the results of this paper. In a general state space model with Markov switching (see e.g. Shephard (1994) and Frühwirth-Schnatter (2006)), the continuous-valued state variable remains dependent on itself after conditioning on the discrete variable. This causes the domain of integration to grow exponentially over time, i.e. a model with Z possible values for z_t would have dimension Z^t at time t. This proposition illustrates that this is not true for models which have a dependence structure like the AG-SV model.

A further implication of this result is that moments and quantiles of the marginal distribution $p(h_t|y_{1:t}; \theta)$ can be computed by weighting the expectations of the conditional distribution $p(h_t|z_t, y_t; \theta)$ by the discrete probabilities $p(z_t|y_{1:t}; \theta)$ calculated from the Markov switching algorithm. At each iteration, the joint distribution can be decomposed as

$$p(h_t, z_t|y_{1:t}; \theta) = p(h_t|y_{1:t}, z_t; \theta)p(z_t|y_{1:t}; \theta) \approx p(h_t|y_t, z_t; \theta)\hat{p}(z_t|y_{1:t}; \theta)$$

Estimates of the moments of the marginal filtering and predictive distributions for the variance are

$$\mathrm{E}\left[h_t^\alpha|y_{1:t};\theta\right] = \sum_{z_t=0}^{\infty}\mathrm{E}\left[h_t^\alpha|y_t,z_t;\theta\right]p(z_t|y_{1:t};\theta) \approx \sum_{z_t=0}^{Z}\mathrm{E}\left[h_t^\alpha|y_t,z_t;\theta\right]\hat{p}(z_t|y_{1:t};\theta) \tag{13}$$

$$\mathrm{E}\left[h_{t+1}^\alpha|y_{1:t};\theta\right] = \sum_{z_t=0}^{\infty}\mathrm{E}\left[h_{t+1}^\alpha|z_{t+1};\theta\right]p(z_{t+1}|y_{1:t};\theta) \approx \sum_{z_t=0}^{Z}\mathrm{E}\left[h_{t+1}^\alpha|z_{t+1};\theta\right]\hat{p}(z_{t+1}|y_{1:t};\theta) \tag{14}$$

where the orders of integration for z_t and h_t have been exchanged. In these expresssions, the conditional expectations $\mathrm{E}\left[h_t^\alpha|y_t,z_t;\theta\right]$ and $\mathrm{E}\left[h_{t+1}^\alpha|z_{t+1};\theta\right]$ are with respect to the GIG and gamma distributions from Proposition 2. These can be calculated analytically for any value of z_t and are then weighted by their marginal probabilities $\hat{p}(z_t|y_{1:t};\theta)$ produced from the Markov-switching algorithm; see Section 3.1 below. As $Z\to\infty$, the solution converges to the truth whereas numerically accurate estimates can be computed for finite values of Z due to the fact that the probabilities $p(z_t|y_{1:t};\theta)$ converge to zero for large integer values.

If interest centers on the quantiles of the marginal filtering or predictive distributions of the variance, these can be estimated analagously as

$$\begin{aligned} P\left(h_t<x|y_{1:t};\theta\right) &= \sum_{z_t=0}^{\infty}\left[\int_0^x p\left(h_t|y_t,z_t;\theta\right)dh_t\right]p(z_t|y_{1:t};\theta) \\ &\approx \sum_{z_t=0}^{Z}\left[\int_0^x p\left(h_t|y_t,z_t;\theta\right)dh_t\right]\hat{p}(z_t|y_{1:t};\theta) \qquad (15) \\ P\left(h_{t+1}<x|y_{1:t};\theta\right) &= \sum_{z_t=0}^{\infty}\left[\int_0^x p\left(h_{t+1}|z_{t+1};\theta\right)dh_t\right]p(z_{t+1}|y_{1:t};\theta) \\ &\approx \sum_{z_t=0}^{Z}\left[\int_0^x p\left(h_{t+1}|z_{t+1};\theta\right)dh_t\right]\hat{p}(z_{t+1}|y_{1:t};\theta) \qquad (16) \end{aligned}$$

where the terms in brackets are the cumulative distribution functions of the GIG and gamma distributions from Proposition 2. For a given quantile and with a fixed value of $\hat{p}(z_t|y_{1:t};\theta)$ or $\hat{p}(z_{t+1}|y_{1:t};\theta)$, the value of x in (15) and (16) can be determined with a simple zero-finding algorithm. Similar ideas can be applied to estimate moments and quantiles of the marginal smoothing distributions, which are also GIG and gamma distributions as shown in Proposition 2. Further details are discussed in Section 3.2.

Another contribution of this paper is to recognize the importance of the conditional dependence structure of the CIR/AG dynamics. The ideas developed in this section can be used to analyze other models whose dependence structure is like the CIR/AG dynamics as long as it is possible to analytically calculate the distributions $p(y_t|z_t;\theta)$ and $p(z_t|z_{t-1}, y_{1:t-1};\theta)$. Important models for which this is possible are when the distribution of $p(y_t|h_t;\theta)$ is Poisson with time-varying mean h_t, exponential with time-varying intensity h_t, and gamma with time-varying mean h_t. The exponential model with time-varying intensity is an important model in the literature on credit risk; see, e.g. Duffie and Singleton (2003) and Duffie, Eckner, Horel, and Saita (2009). These results are developed further in Creal (2012c) who builds methods for Cox processes with CIR intensities, marked Poisson processes with CIR intensities, and duration models built from the AG process.

3 Filtering and smoothing recursions

In this section, I describe the filtering and smoothing algorithms in further detail including a discussion of how to implement them efficiently. I also discuss the relationship of the filtering algorithm with the particle filter.

3.1 Marginal prediction and filtering recursion

The recursions for the marginal filtering and one-step ahead predictive distributions proceed forward in time from $t = 1, \dots, T$ as follows. Assuming that the initial condition is a draw from the stationary distribution $h_0 \sim \text{Gamma}\left(\nu, \frac{c}{1-\phi}\right)$, the distributions for the initial iteration are computed analytically in the following proposition.

Proposition 3 *The initial contribution to the likelihood $p(y_1;\theta)$, the marginal filtering distri-*

butions $p(z_1|y_1;\theta)$, $p(h_1|y_1;\theta)$, and the predictive distribution $p(z_2|y_1;\theta)$ are

$$\begin{aligned} p(y_1;\theta) &= \textit{Variance Gamma}\left(\mu, \sqrt{\frac{2(1-\phi)}{c}+\beta^2}, \beta, \nu\right) \\ p(h_1|y_1;\theta) &= \textit{GIG}\left(\nu-\frac{1}{2}, (y_1-\mu)^2, \frac{2(1-\phi)}{c}+\beta^2\right) \\ p(z_1|y_1;\theta) &= \textit{Sichel}\left(\nu-\frac{1}{2}, \frac{\phi}{c}(y_1-\mu)^2, \frac{c}{\phi}\left[\frac{2(1-\phi)}{c}+\beta^2\right]\right) \\ p(z_2|y_1;\theta) &= \textit{Sichel}\left(\nu-\frac{1}{2}, \frac{\phi}{c}(y_1-\mu)^2, \frac{c}{\phi}\left[\frac{2(1-\phi)}{c}+\beta^2\right]\right) \end{aligned} \tag{17}$$

Proof: See Appendix.

From this point forward, the distributions cannot be solved analytically because the integrals have an unknown solution. I proceed by approximating them by a finite-state Markov switching model. Proposition 3 provides a way to initial the algorithm. I proceed by computing the $(Z+1)\times 1$ vector of predictive probabilities $\hat{p}(z_2|y_1;\theta)$ using the Sichel distribution (17). Computing the filtering probabilities $\hat{p}(z_t|y_{1:t};\theta)$ and the predictive probabilities $\hat{p}(z_{t+1}|y_{1:t};\theta)$ for $t=2,\ldots,T$ follows by recursively applying Bayes' rule. The filtering and prediction steps are

$$\hat{p}(z_t=i|y_{1:t};\theta) = \frac{p(y_t|z_t=i;\theta)\hat{p}(z_t=i|z_{t-1},y_{t-1};\theta)}{\sum_{j=0}^{Z} p(y_t|z_t=j;\theta)\hat{p}(z_t=j|y_{1:t-1};\theta)} \tag{18}$$

$$\hat{p}(z_{t+1}=i|y_{1:t};\theta) = \sum_{j=0}^{Z}\hat{p}(z_{t+1}=i|z_t=j,y_t;\theta)\hat{p}(z_t=j|y_{1:t};\theta) \tag{19}$$

At each iteration of the algorithm, the distributions $\hat{p}(z_t|y_{1:t};\theta)$ and $\hat{p}(z_{t+1}|y_{1:t};\theta)$ are self-normalized to ensure that they are well-defined probability distributions summing to one. The term $\hat{p}(y_t|y_{1:t-1};\theta)=\sum_{j=0}^{Z}p(y_t|z_t=j)\hat{p}(z_t=j|y_{1:t-1};\theta)$ is the contribution to the likelihood function for date t. The log-likelihood function is given by

$$\log\hat{p}\left(y_{1:T};\theta\right) = \sum_{t=2}^{T}\log\hat{p}(y_t|y_{1:t-1};\theta)+\log p(y_1;\theta)$$

During the algorithm, expectations of the marginal predictive and filtering distributions for the variance can be computed using (13) and (14). Further discussion on Markov-switching algorithms can be found in Hamilton (1994), Cappé, Moulines, and Rydén (2005), and Frühwirth-Schnatter (2006).

At this point, it is important to note that the prediction step in (19) is a vector-matrix multiplication which is an $O(Z^2)$ operation. Fortunately, the Markov transition distribution $p(z_t|z_{t-1}, y_{1:t-1};\theta)$ is a sparse, non-symmetric band matrix. In addition, many of the entries in the vector of filtered probabilities $p(z_t|y_{1:t};\theta)$ are close to zero. This feature of the model means that the $O(Z^2)$ computation in the prediction step of the algorithm can be avoided. It is only necessary to compute those elements in the vector-matrix operation that when multiplied together are greater than machine zero on a computer. In other words, if for any combination of states i and j the product $p(z_{t+1} = i|z_t = j, y_{1:t-1};\theta)p(z_t = j|y_{1:t};\theta)$ is numerically zero, then it is unnecessary to compute the probability for other transitions whose probabilities are known to be smaller. In addition, each row times column operation in the vector-matrix multiplication of the prediction step can be computed in parallel on multi-core processors. The sparse nature of the transition matrix and the ability to compute the operations in parallel make computing the likelihood easily feasible for macroeconomic and financial data sets using daily, monthly, or quarterly data.

An important practical issue is the choice of the threshold Z that determines where the infinite sums are truncated. First, it is possible to determine the values of Z_{θ,y_t} endogenously at each time period within the algorithm such that $\hat{p}(z_t = Z_{\theta,y_t}|y_{1:t};\theta)$ is numerically equal to zero or some other prespecified tolerance level. In this way, the values of Z_{θ,y_t} can be automatically determined. The size of the vectors $\hat{p}(z_t|y_{1:t};\theta)$ and $\hat{p}(z_{t+1}|y_{1:t};\theta)$ as well as the Markov transition matrix are then stochastic. Stochastic truncation of the infinite sums will ensure a uniform approximation across the parameter space. Alternatively, a researcher can determine a fixed value of Z a priori by carefully considering the properties of the model. First, one can find initial values for the parameters (ϕ, c, ν) by comparing moments of the stationary distribution of volatility and its persistence to their sample counterparts based on squared returns. Next, note that the stationary distribution of z_t is a negative binomial(ν, ϕ) distribution whose mean and variance are $\mathrm{E}[z] = \frac{\nu\phi}{1-\phi}$ and $\mathrm{V}[z] = \frac{\nu\phi}{(1-\phi)^2}$. Plugging the values of (ϕ, ν) into these formulas and noting that as the mean of the negative binomial distribution gets larger it becomes approximately normally distributed, we can see that almost all of the probability mass for z_t will be within 3-4 standard deviations of the mean.

TO BE COMPLETED. IT MAY NOT BE POSSIBLE BECAUSE THE DISTRIBUTION MAY NOT EVENTUALLY LEAD TO A PROPER DENSITY. To this point, I have assumed that the initial condition h_0 is a draw from the stationary distribution $h_0 \sim \text{Gamma}\left(\nu, \frac{c}{1-\phi}\right)$. For linear, Gaussian state space models, some components of the state vector may be non-stationary and the Kalman filter can be initialized with a non-informative or diffuse initial condition; see, e.g. de Jong (1991) and Koopman (1997). Even when a model is stationary, it is often sensible to use a diffuse initialization to minimize the impact of the initial condition on the likelihood function as well as estimates of the state variables. It is possible to initialize the filtering algorithm for AG-SV models using a diffuse initialization. The gamma distribution for the initial condition becomes diffuse as both $\nu \to 0$ and $\frac{c}{1-\phi} \to 0$. In this case, the density is improper and is proportional to the function h_0^{-1}. With a diffuse initialization, the initial contribution to the log-likelihood, the filtering distributions, and prediction distribution in Proposition 3 are in the same family of distributions but with different parameters. Importantly, the one-step ahead predictive probabilities $\hat{p}(z_2|y_1;\theta)$ need to be initialized from a different Sichel distribution; see the appendix for details.

3.1.1 Relationship to the particle filter

The filtering recursions developed in this paper can be compared to several techniques developed in the literature on particle filters; see, e.g. Creal (2012b) for a survey. Particle filters are Monte Carlo algorithms that approximate distributions whose support is infinite by a finite set of points and probability masses. The algorithm from the previous section can be interpreted as an optimal particle filter in the following way. First, the continuous-valued variance h_t is analytically integrated out of the model through an application of Rao-Blackwellisation, e.g. Chen and Liu (2000). Next, the filtering distribution $p(z_{t-1}|y_{1:t-1};\theta)$ is represented by a collection of "particles" given by $\left\{w_{t-1}^{(i)}, z_{t-1}^{(i)}\right\}_{i=1}^{Z}$. The values $z_{t-1}^{(i)}$ are locations on the support of the distribution and the $w_{t-1}^{(i)}$ are weights representing the probability mass at that point. In the current model, the locations on the support of the distribution are the integers $z_{t-1}^{(i)} = i$ for $i = 0\ldots, Z$ and the weights are the probabilities $\hat{p}(z_{t-1}|y_{1:t-1};\theta)$.

A particle filtering algorithm updates the particles' locations and weights from one time period $\left\{w_{t-1}^{(i)}, z_{t-1}^{(i)}\right\}_{i=1}^{Z}$ to the next time period $\left\{w_{t}^{(i)}, z_{t}^{(i)}\right\}_{i=1}^{Z}$ using Monte Carlo methods. In a typical particle filter where the support of z_t is continuous, the transitions for each particle from

$z_{t-1}^{(i)}$ to $z_t^{(i)}$ are determined randomly and the weights are updated by importance sampling. The new set of particles $\left\{w_t^{(i)}, z_t^{(i)}\right\}_{i=1}^{Z}$ form a discrete distribution that approximates the next filtering distribution $p(z_t|y_{1:t};\theta)$. When the support of the state variable is discrete, the transitions do not need to be random but can be deterministic in order not to allow particles to duplicate transitions. Duplicating transitions is redundant and computationally wasteful. Suppose for simplicity that the sparse matrix operation discussed in the previous section is ignored when computing the prediction step. Then the filtering algorithm of Section 3.1 forces each particle to transition deterministically through all of the future Z states creating a total of Z^2 offspring. This particle filtering algorithm is an example of a marginalized particle filter; see, e.g. Klass, de Freitas, and Doucet (2005). In a particle filter, particles are resampled when their weights become imbalanced. In this setting, the Z^2 particles are resampled into Z unique particles with unequal weights $\left\{w_t^{(i)}, z_t^{(i)}\right\}_{i=1}^{Z}$ using an optimal resampling algorithm, e.g. Fearnhead and Clifford (2003). In the algorithm outlined in Section 3.1, the resampling step is not explicitly carried out. It is implicitly part of the vector-matrix multiplication in the prediction step when the row vector of filtering probabilities are multiplied by each column of the Markov transition matrix and summed to compute the predictive probabilities.

By interpreting the algorithm as a particle filter, it is easy to see that the only way to (recursively) compute the solution to the filtering problem more accurately is to increase the number of "particles" Z, i.e. by truncating the infinite sums at larger integers. In addition, it is also easy to see that the algorithm will be consistent in the sense that as $Z \to \infty$ the solution will converge to the true value.

3.2 Marginal smoothing recursion

An object of independent interest are smoothed or two-sided estimators of the latent variance. It is also possible to recursively compute the marginal smoothing distributions for $p(z_t|y_{1:T};\theta)$ and moments and quantiles of the marginal distribution $p(h_t|y_{1:T};\theta)$. The solution for smoothing is analagous to the solution for forward filtering. First, the joint distribution can be decomposed as $p(h_{0:T}, z_{1:T}|y_{1:T};\theta) = p(h_{0:T}|z_{1:T}, y_{1:T};\theta)p(z_{1:T}|y_{1:T};\theta)$ and the smoothing algorithms for discrete-state Markov switching models can be applied to calculate the marginal probabilities $p(z_t|y_{1:T};\theta)$. Moments and quantiles of the marginal smoothing distribution $p(h_t|y_{1:T};\theta)$

can be computed using the results from Proposition 2.

First, the filtering algorithm (18) and (19) is run forward in time and the one-step ahead prediction probabilities and the conditional likelihoods $p(y_t|z_t;\theta)$ are stored for $t = 1,\dots,T$. The backwards smoothing algorithm is initialized using the final iteration's filtering probabilities $\hat{p}(z_T|y_{1:T})$. For $t = T-1,\dots,1$, the smoothing distributions for z_t are computed as

$$\hat{p}\left(z_t = i|z_{t+1} = j, y_{1:T};\theta\right) = \frac{p\left(y_t|z_t = i;\theta\right)\hat{p}(z_t = i|z_{t+1} = j, y_{1:t+1};\theta)}{\sum_{k=0}^{Z} p(y_t|z_t = k;\theta)\hat{p}(z_t = k|z_{t+1} = j, y_{1:t+1};\theta)} \tag{20}$$

$$\hat{p}\left(z_t = i, z_{t+1} = j|y_{1:T};\theta\right) = \hat{p}\left(z_{t+1} = j|y_{1:T};\theta\right)\hat{p}\left(z_t = i|z_{t+1} = j, y_{1:T};\theta\right) \tag{21}$$

$$\hat{p}\left(z_t = i|y_{1:T};\theta\right) = \sum_{j=0}^{Z}\hat{p}\left(z_t = j|y_{1:T};\theta\right)\hat{p}\left(z_t = i|z_{t+1} = j, y_{1:T};\theta\right) \tag{22}$$

At each iteration, the distributions are self-normalized to ensure that they are well-defined probability distributions. Computation of (20) also requires an $O(Z^2)$ operation on the backwards pass. This can be reduced substantially by taking into account the sparse matrix structure of the transition probabilities.

This algorithm also computes the bivariate smoothing probabilities $\hat{p}\left(z_t = i, z_{t+1} = j|y_{1:T};\theta\right)$ which are used to calculate the moments and quantiles of the marginal smoothing distribution for the variance

$$\mathrm{E}\left[h_t^{\alpha}|y_{1:T}\right] \approx \sum_{z_t=0}^{Z}\sum_{z_{t+1}=0}^{Z}\mathrm{E}\left[h_t^{\alpha}|y_t, z_t, z_{t+1}\right]\hat{p}(z_t, z_{t+1}|y_{1:T};\theta)$$

$$P\left(h_t < x|y_{1:T};\theta\right) \approx \sum_{z_t=0}^{Z}\sum_{z_{t+1}=0}^{Z}\left[\int_0^x p\left(h_t|y_t, z_t, z_{t+1};\theta\right)dh_t\right]\hat{p}(z_t, z_{t+1}|y_{1:T};\theta)$$

Conditional expectations and quantiles of the marginal smoothing distribution $p(h_t|y_t, z_t, z_{t+1};\theta)$ are calculated from the properties of the GIG and gamma distributions of Proposition 2. Smoothed estimates of the variance are of independent interest but the algorithm can also be used as part of the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977) as an alternative means of computing the maximum likelihood estimator.

Finally, we note that the unique conditional independence structure highlighted in Proposition 2 makes it easy to draw samples from the joint distribution $p(h_{0:T}, z_{1:T}|y_{1:T},\theta)$ and thus the marginal distribution of the variance. This follows by decomposing the joint distribution into a conditional and a marginal as $p(h_{0:T}, z_{1:T}|y_{1:T};\theta) = p(h_{0:T}|y_{1:T}, z_{1:T};\theta)p(z_{1:T}|y_{1:T};\theta)$. A draw is taken from the marginal distribution $z_{1:T} \sim \hat{p}(z_{1:T}|y_{1:T};\theta)$ using standard results on Markov-switching algorithms. Then, conditional on the draw of $z_{1:T}$, a draw is taken from $h_{0:T} \sim p(h_{0:T}|y_{1:T}, z_{1:T};\theta)$ which are the distributions from Proposition 2. Algorithms that draw samples from the joint distribution of the latent variables given the data are known as simulation smoothers or alternatively as forward-filtering backward sampling (FFBS) algorithms; see, e.g. Carter and Kohn (1994), Frühwirth-Schnatter (1994), de Jong and Shephard (1995) and Durbin and Koopman (2002) for linear, Gaussian models and Chib (1996) for Markov-switching models. Simulations smoothers are useful for frequentist and Bayesian analysis of more complex models. In addition, they are useful for calculating moments and quantiles of nonlinear functions of the state variable such as the volatility $\sqrt{h_t}$. Creal (2012a) shows how these results can be used for Bayesian inference of the AG-SV model and its extensions.

4 Application of the new methods

4.1 Comparisons with the particle filter

In this section, I compare the log-likelihood functions calculated from the new algorithm with a standard particle filtering algorithm to demonstrate their relative accuracy. The algorithms are compared by running them on the S&P500 series discussed further below in Section 4.2. All filtering algorithms were run with the parameter values fixed at the ML estimates from the next section. I run the filtering algorithm of Section 3.1 with the truncation parameter set at $Z = 3500$. For comparison purposes, I implement a particle filter with the transition density as a proposal where the particles are resampled at random times according to the effective sample size (ESS), see Creal (2012b). I use the residual resampling algorithm of Liu and Chen (1998), which is known to be unbiased. This is a simple extension of the original particle filter of Gordon, Salmond, and Smith (1993).[4]

To compare the methods, I compute slices of the log-likelihood function for the parameters ϕ and ν over the regions $[0.97, 0.999]$ and $[1.0, 2.2]$, respectively. Cuts of the log-likelihood function for the remaining parameters (μ, β, c) are similar and are available in an online appendix. Each

[4] I implemented additional particle filtering algorithms including the auxiliary particle filter of Pitt and Shephard (1999). The results were similar and are available in an online appendix.

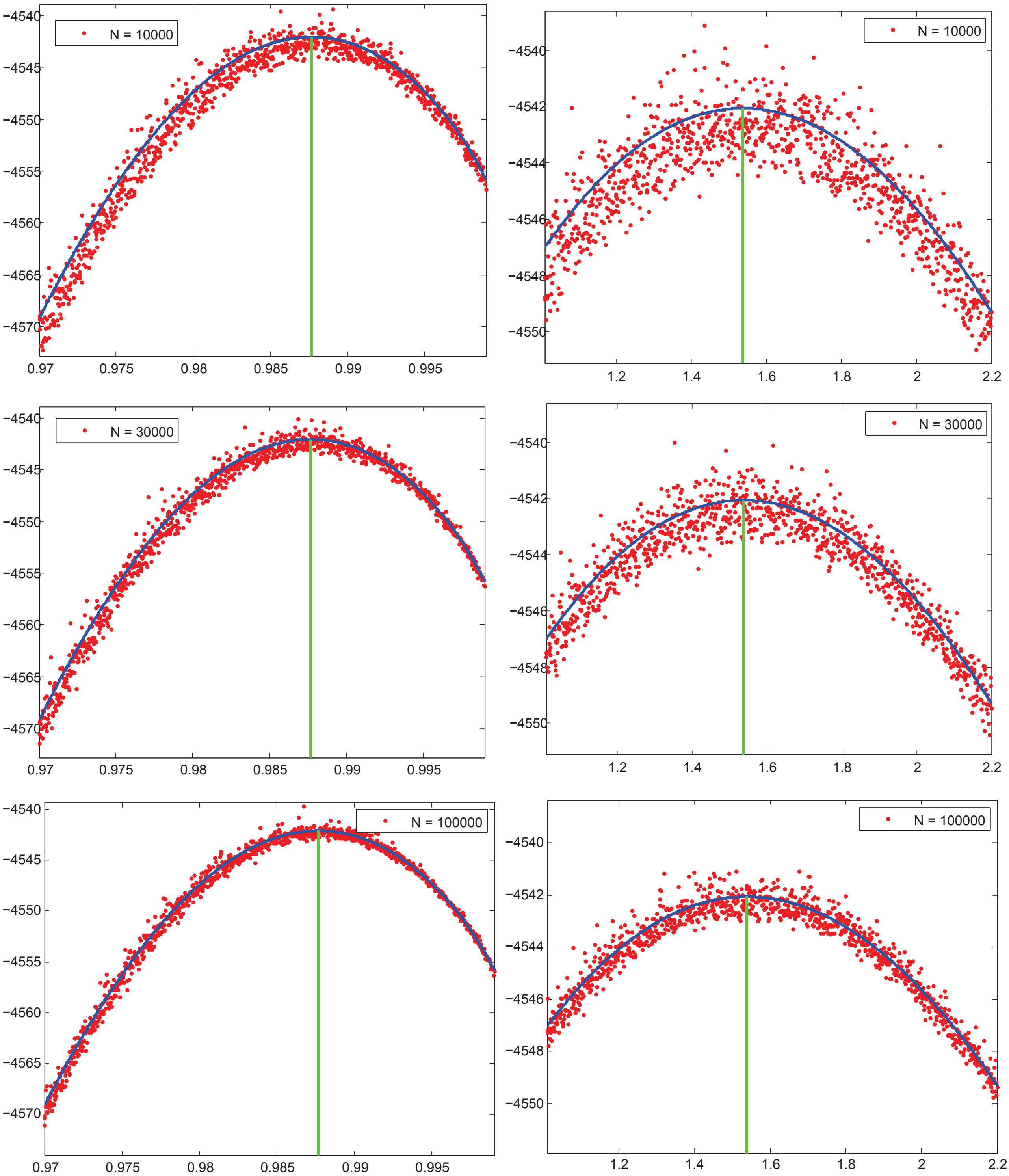

Figure 1: *Slices of the log-likelihood function for* ϕ *(left) and* μ *(right) for the new algorithm (blue line) versus the particle filter (red dots). The particle sizes are* $N = 10000$ *(top row),* $N = 30000$ *(middle row), and* $N = 100000$ *(bottom row). The vertical (green) lines are the ML estimates of the model reported in Table 1.*

of these regions are divided into 1000 equally spaced points, where the log-likelihood function is evaluated. While the values of ϕ and ν change one at a time, the remaining parameters are held fixed at their ML estimates. The log-likelihood functions from the particle filter are shown for different numbers of particles $N = 10000$, $N = 30000$ and $N = 100000$. These values were selected because they are representative of values of N used throughout the literature. One limitation of the particle filter that was raised early in that literature can be seen from Figure 1. The particle filter does not produce an estimate of the log-likelihood function that is smooth in the parameter space. Consequently, standard derivative-based optimization routines may have trouble converging to the maximum. The smoothness of the log-likelihood function produced by the new algorithm is an attractive feature. It makes calculation of the ML estimates easily feasible as shown in the next section.

Secondly, in this application, there is a sizeable downward bias in the particle filter's estimates of the log-likelihood function as well as a reasonably large amount of variability in the estimates. It is known that the particle filter's estimator of the likelihood function is unbiased, as long as the resampling algorithm used within a particle filter is unbiased. The particle filter's estimate of the likelihood is also consistent and asymptotically normal; see, e.g. Chapter 11 of Del Moral (2004). However, taking logarithms causes the particle filter's estimate to be biased downward for finite values of the particles due to Jensen's inequality.

4.2 Applications to financial data

In this section, I apply the new methods to several daily financial time series. These include the S&P 500 index, the MSCI-Emerging Markets Asia index, and the Euro-to-U.S. dollar exchange rate. The first two series were downloaded from Bloomberg and the latter series was taken from the Board of Governers of the Federal Reserve. All series are from January 3rd, 2000 to December 16th, 2011 making for 3009, 3118, and 3008 observations for each series, respectively. A value of $Z = 3500$ was used for all series. Starting values for the parameters of the variance (ϕ, ν, c) were obtained by matching the unconditional mean, variance, and persistence of average squared returns to the unconditional distribution. The values of (μ, β) were initialized at zero. The Feller condition $\nu > 1$ as well as the constraints $0 < \phi < 1$ and $c > 0$ were imposed throughout the estimation.

Table 1: Maximum likelihood estimates for the AG-SV model.

	μ	β	ϕ	c	ν	θ_h	κ	σ^2	log-like
S&P500	0.102	-0.061	0.988	0.015	1.539	1.815	3.168	7.470	-4542.1
	(0.020)	(0.018)	(0.004)	(0.003)	(0.193)	——	——	——	
MSCI-EM-ASIA	0.250	-0.115	0.978	0.024	1.963	2.115	5.736	12.364	-5213.2
	(0.031)	(0.020)	(0.006)	(0.005)	(0.229)	——	——	——	
Euro/$	0.073	-0.148	0.994	0.0007	3.740	0.442	1.489	0.359	-2862.2
	(0.022)	(0.058)	(0.004)	(0.0002)	(1.263)	——	——	——	

Maximum likelihood estimates of the parameters $\theta = (\mu, \beta, \phi, c, \nu)$ of the discrete-time AG-SV model on three data sets of daily returns. The series are the S&P500, the MSCI emerging markets Asia index, and the EURO-$ exchange rate. The data cover January 3, 2000 to December 16, 2011. Asymptotic standard errors are reported in parenthesis. The continuous-time parameters are reported for intervals $\tau = \frac{1}{256}$.

Estimates of the parameters of the discrete-time model as well as standard errors are reported in Table 1. The standard errors are calculated by numerically inverting the Hessian at the ML estimates. The implied parameters of the continuous-time model are also reported in Table 1 assuming a discretization step size of $\tau = \frac{1}{256}$. For all series, the risk premium parameters β are estimated to be negative and significant implying that the distribution of returns are negatively skewed. Estimates of the autocorrelation parameter ϕ for the S&P500 and MSCI series are slightly smaller than is often reported in the literature for log-normal SV models. The dynamics of volatility for the Euro/$ series is substantially different than the other series. The mean θ_h of volatility is much lower and the volatility is more persistent.

Figure 2 contains output from the estimated model for the S&P500 series. The top left panel is a plot of the filtered and smoothed estimates of the volatility $\sqrt{h_t}$ over the sample period. The estimates are consistent with what one would expect from looking at the raw return series. There are large increases in volatility during the U.S. financial crisis of 2008 followed by another recent spike in volatility during the European debt crisis.

To provide some information on how the truncation parameter Z impacts the estimates, the top right panel of Figure 2 is a plot of the marginal filtering distributions for the discrete mixing variable $p(z_t|y_{1:t};\theta)$ on three different days. The three distributions that are pictured are representative of the distribution $p(z_t|y_{1:t};\theta)$ during low (6/27/2007), medium (10/2/2008), and high volatility (12/1/2008) periods. The distribution of z_t for the final date (12/1/2008) was

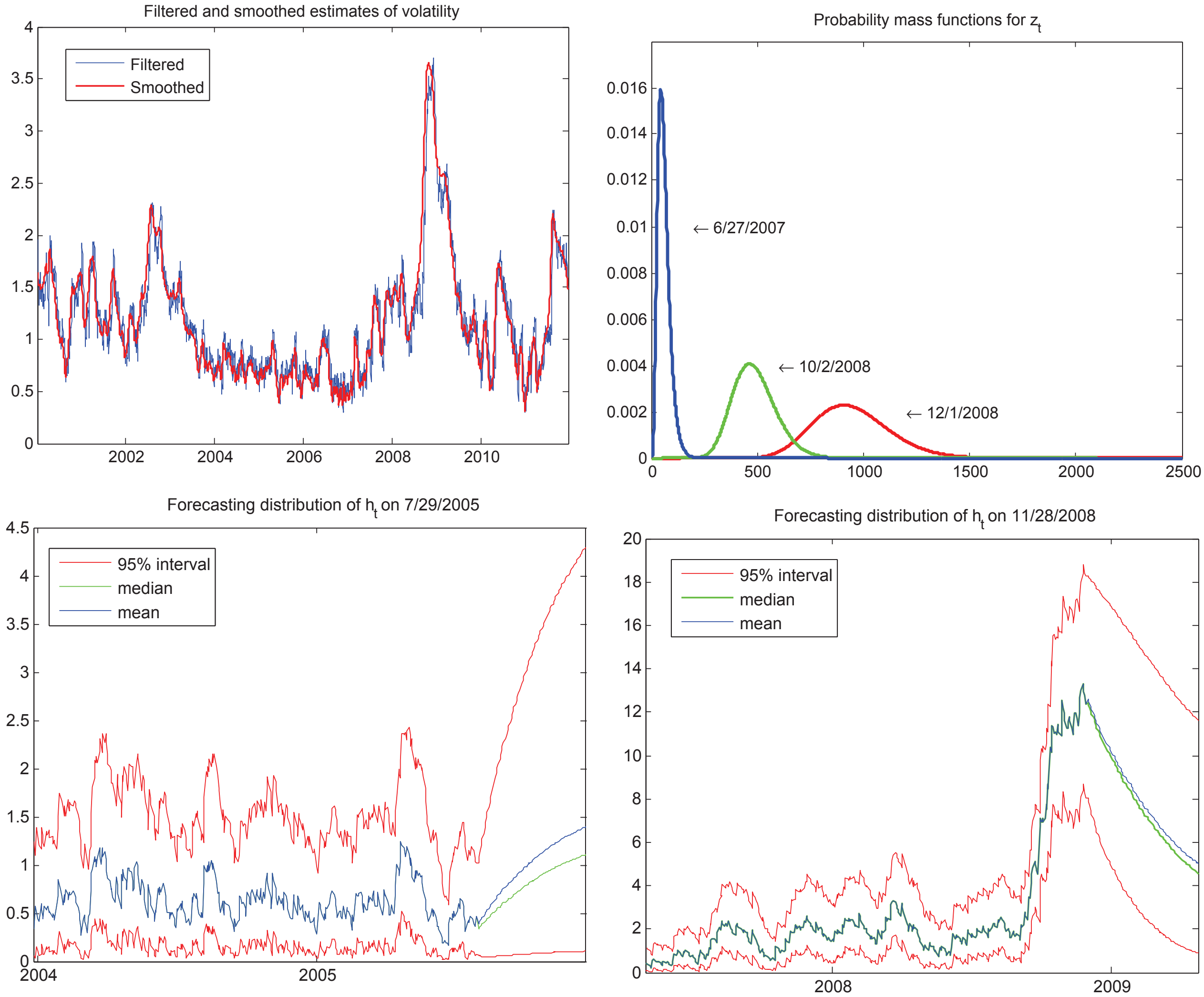

Figure 2: *Estimation results for the AG-SV model on the S&P500 index from 1/3/2000 to 12/16/2011. Top left: filtered and smoothed estimates of the volatility* $\sqrt{h_t}$*. Top right: marginal probability mass functions* $p(z_t|y_{1:t};\theta)$ *on three different dates. Bottom left: forecasting distribution of the variance* h_t *for 100 days beginning on 7/29/2005. Bottom right: forecasting distribution of the variance* h_t *for 100 days beginning on 11/28/2008.*

chosen because it is the day at which the mean of z_t is estimated to be the largest throughout the sample. Consequently, it is the distribution where the truncation will have the largest impact. This time period also corresponds to the largest estimated value of the variance h_t. The graph illustrates visually that the impact of truncating the distribution appears to be negligible as the truncation point is far into the tails of the distribution.

Table 2: Cumulative distribution functions $P(z_t|y_{1:t};\theta)$ for different values of Z.

	$P(z_t \leq 1500|y_{1:t};\theta)$	$P(z_t \leq 2000|y_{1:t};\theta)$	$P(z_t \leq 2500|y_{1:t};\theta)$	log-like
$Z = 2500$	0.996057165621493	0.999995671659054	1.000000000000000	-4542.062558886049
$Z = 3000$	0.996057162486533	0.999995668594590	0.999999998627221	-4542.062558891409
$Z = 3500$	0.996057162486131	0.999995668594186	0.999999998626824	-4542.062558891406
$Z = 5000$	0.996057162486131	0.999995668594186	0.999999998626824	-4542.062558891406

This table contains the cumulative distribution functions of the filtering distributions $P(z_t|y_{1:t};\theta)$ and the log-likelihood as a function of the truncation parameter for $Z = 2500, 3000, 3500, 5000$ on 12/1/2008. The table also reports the log-likelihood function for the entire dataset.

In Section 3.1.1, I argued that the only way to compute the log-likelihood function more accurately was to increase the value of Z. To quantify this further, the algorithms were run again and the log-likelihood functions as well as the c.d.f.s for z_t were calculated for a series of different truncation values. The cumulative distribution functions for the filtering distribution $P(z_t|y_{1:t};\theta)$ on 12/1/2008 and the overall log-likelihoods are reported in Table 2 for values of $Z = 2500, 3000, 3500, 5000$. The first row of the table contains the cumulative probabilities for values of $z_t = 1500, 2000, 2500$ when the algorithm was run with $Z = 2500$. When the value of Z is set at 2500, there is 0.996057165621493 cumulative probability to the left of $z_t = 1500$. When $z_t = 2500$ and $Z = 2500$, the c.d.f. reaches a value of one due to self-normalizing the probabilities. The purpose of the table is to illustrate that the value of the log-likelihood function converges (numerically) between the value of $Z = 3000$ and $Z = 3500$. Therefore, it is possible to calculate the log-likelihood function of the model accurately for plausible values of Z.

An important feature of the AG-SV model that is useful in practice is that forecasts of the variance can be computed accurately without simulation. These H-step ahead forecasts of the variance are produced by iterating on the vector of probabilities $\hat{p}(z_{t+1}|y_{1:t};\theta)$ from the last iteration of the Markov-switching algorithm using the transition distribution of the latent variable z_t. When there are no observations available out of sample, the transition distribution of z_t was shown by Gouriéroux and Jasiak (2006) to be a negative binomial distribution $p(z_t|z_{t-1};\theta) = \text{Neg. Bin.}\left(\nu + z_{t-1}, \frac{\phi}{1+\phi}\right)$. Iterating for H periods on this distribution produces the distribution $\hat{p}(z_{t+H}|y_{1:t};\theta)$. This distribution can be combined with the conditional distribution of the variance at any horizon $p(h_{t+H}|z_{t+H};\theta) = \text{Gamma}\left(\nu + z_{t+H}, c\right)$ to construct forecasting intervals. For example, to calculate quantiles of the H-step ahead forecasting distribution, one can solve for x given a fixed value of Z and the ML estimator $\hat{\theta}$ in the equation

$$\hat{P}\left(h_{t+H} < x|y_{1:t};\theta\right) = \sum_{z_t=0}^{Z}\left[\int_0^x p\left(h_{t+H}|z_{t+1};\hat{\theta}\right)dh_t\right]\hat{p}(z_{t+H}|y_{1:t};\hat{\theta}) \quad (23)$$

using a simple zero-finding algorithm.

The bottom two plots in Figure 2 are forecasts of the future variance h_t beginning on two different dates 7/29/2005 and 11/28/2008. Forecasts for the mean, median, and 95% intervals for h_t are produced for H =100 days. These dates were selected to illustrate the substantial difference in both asymmetry and uncertainty in the forecasts during low and high periods of volatility. When volatility is low (7/29/2005), the forecasting distribution of h_t is highly asymmetric and the mean and median of the distribution differ in economically important ways. Conversely, when volatility is high (11/28/2008), the distribution of the variance is roughly normally distributed. The difference in width of the 95% error bands between the two dates also illustrates how much more uncertainty exists in the financial markets during a crisis.

5 Conclusion

In this paper, I developed algorithms that can compute the log-likelihood function of affine stochastic volatility models exactly up to computer tolerance. The new algorithms are based on the insight that it is possible to integrate out the latent variance analytically leaving only a discrete mixture variable. The discrete variable is defined over the set of non-negative integers but for practical purposes it is possible to approximate the distributions involved by a finite-dimensional Markov switching model. Consequently, the log-likelihood function can be computed using standard algorithms for Markov-switching models. Furthermore, filtered and smoothed estimates of the latent variance can easily be computed as a by-product.

References

Abramowitz, M. and I. A. Stegun (1964). *Handbook of Mathematical Functions.* New York, NY: Dover Publications Inc.

Barndorff-Nielsen, O. E. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics 5*(3), 151–157.

Baum, L. E. and T. P. Petrie (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics 37*, 1554–1563.

Baum, L. E., T. P. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics 41*, 164–171.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in Hidden Markov Models*. New York: Springer Press.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*(3), 541–553.

Chen, R. and J. S. Liu (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society, Series B 62*(3), 493–508.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics 75*(1), 79–97.

Chib, S., M. K. Pitt, and N. Shephard (2006). Likelihood inference for diffusion driven state space models. Working Paper. Department of Economics, Nuffield College, University of Oxford.

Cox, J. C., J. E. Ingersoll, and S. A. Ross (1985). A theory of the term structure of interest rates. *Econometrica 53*(2), 385–407.

Creal, D. D. (2012a). A simulation smoother for CIR stochastic volatility models. Working paper, University of Chicago, Booth School of Business.

Creal, D. D. (2012b). A survey of sequential Monte Carlo methods for economics and finance. *Econometric Reviews 31*(3), 245–296.

Creal, D. D. (2012c). Maximum likelihood estimation of affine stochastic intensity models. Working paper, University of Chicago, Booth School of Business.

de Jong, P. (1991). The diffuse Kalman filter. *The Annals of Statistics 19*(2), 1073–83.

de Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*(2), 339–350.

Del Moral, P. (2004). *Feyman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* New York, NY: Springer.

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 39*(1), 1–38.

Duffie, D., A. Eckner, G. Horel, and L. Saita (2009). Frailty correlated default. *The Journal of Finance 64*(5), 2089–2123.

Duffie, D. and K. J. Singleton (2003). *Credit Risk.* Princeton, NJ: Princeton University Press.

Dugué, D. (1941). Sur un nouveau type de courbe de fréquence. *C.R. Academie Science Paris 213*, 634–635.

Durbin, J. and S. J. Koopman (1997). Monte Carlo maximum likelihood estimation of non-Gaussian state space models. *Biometrika 84*(3), 669–84.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods.* Oxford, UK: Oxford University Press.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*(3), 603–616.

Durham, G. and A. R. Gallant (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes (with discussion). *Journal of Business and Economic Statistics 20*(3), 297–338.

Eraker, B., M. Johannes, and N. G. Polson (2003). The impact of jumps in returns and volatility. *The Journal of Finance 53*(3), 1269–1300.

Fearnhead, P. and P. Clifford (2003). On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society, Series B 65*(4), 887–899.

Feller, W. S. (1951). Two singular diffusion problems. *Annals of Mathematics 54*(1), 173–182.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis 15*(2), 183–202.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York, NY: Springer Press.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). A novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE Proceedings. Part F: Radar and Sonar Navigation 140*(2), 107–113.

Gouriéroux, C. and J. Jasiak (2006). Autoregressive gamma processes. *Journal of Forecasting 25*(2), 129–152.

Hamilton, J. D. (1989). A new approach to the econometric analysis of nonstationary time series and the business cycle. *Econometrica 57*(2), 357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge, UK: Cambridge University Press.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies 6*(2), 327–343.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions of the ASME 82*(1), 35–46.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies 65*(3), 361–393.

Klass, M., N. de Freitas, and A. Doucet (2005). Toward practical N^2 Monte Carlo: the marginal particle filter. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*.

Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association 92*(440), 1630–1638.

Kotz, S., T. J. Kozubowski, and K. Podgórski (2001). *The Laplace Distribution and Generalizations*. Boston, MA: Birkhauser.

Liu, J. S. and R. Chen (1998). Sequential Monte Carlo computation for dynamic systems. *Journal of the American Statistical Association 93*(443), 1032–1044.

Madan, D. B., P. P. Carr, and E. C. Chang (1998). The variance gamma process and option pricing. *Review of Finance 2*(1), 79–105.

Madan, D. B. and E. Seneta (1990). The variance gamma (V.G.) model for share market returns. *Journal of Business 63*(4), 511–524.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: auxiliary particle filters. *Journal of the American Statistical Association 94*(446), 590–599.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*(1), 115–131.

Shephard, N. (2005). *Stochastic Volatility: Selected Readings.* Oxford, UK: Oxford University Press.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society, Series A 135*(1), 25–34.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association 70*(350), 542–547.

Slevinsky, R. M. and H. Safouhi (2010). A recursive algorithm for the G transformation and accurate computation of incomplete Bessel functions. *Applied Numerical Mathematics 60*, 1411–1417.

Appendix:

Notation for the VG, GIG, and Sichel distributions

In this appendix, I define notation for the VG, GIG, and Sichel distributions that are used in the paper. The notation for the parameters of the distributions is local to this section of the appendix. A variance gamma random variable $Y \sim \text{V. G.}(\mu, \alpha, \beta, \nu)$ is obtained by taking a normal distribution $Y \sim \text{N}(\mu + \beta\sigma^2, \sigma^2)$ and allowing its variance to have a gamma distribution $\sigma^2 \sim \text{Gamma}\left(\nu, \frac{2}{\alpha^2-\beta^2}\right)$. The p.d.f. of a V.G. random variable is

$$p(y|\mu,\alpha,\beta,\nu) = \left(\frac{\alpha^2-\beta^2}{2}\right)^{\nu} \frac{\sqrt{2}\,|y_t-\mu|^{(\nu-\frac{1}{2})}\,K_{\nu-\frac{1}{2}}\left(\alpha|y_t-\mu|\right)}{\sqrt{\pi}\Gamma(\nu)\,\alpha^{(\nu-\frac{1}{2})}} \exp\left(\beta|y_t-\mu|\right)$$

where $K_\lambda(x)$ is the modified Bessel function of the second kind; see, e.g. Abramowitz and Stegun (1964). The mean and variance of the distribution are

$$\mathrm{E}[y] \;=\; \mathrm{E}_{\sigma^2}\left[\mathrm{E}\left(y|\sigma^2\right)\right] \;=\; \mathrm{E}_{\sigma^2}\left[\mu+\beta\sigma^2\right] \;=\; \mu+\frac{2\beta\nu}{\alpha^2-\beta^2}$$

$$\mathrm{V}[y] \;=\; \mathrm{E}\left[\mathrm{V}\left(y|\sigma^2\right)\right]+\mathrm{V}\left[\mathrm{E}\left(y|\sigma^2\right)\right] \;=\; \mathrm{E}\left[\sigma^2\right]+\mathrm{V}\left[\mu+\beta\sigma^2\right] \;=\; \frac{2\nu}{\alpha^2-\beta^2}+\frac{4\nu\beta^2}{\left(\alpha^2-\beta^2\right)^2}$$

see, e.g. Kotz, Kozubowski, and Podgórski (2001). The parameter β controls the symmetry of the distribution such that $\beta < 0$ is left-skewed, $\beta > 0$ is right skewed, and a value of $\beta = 0$ is symmetric.

A generalized inverse Gaussian random variable $X \sim \mathrm{GIG}(\lambda, \chi, \psi)$ has p.d.f.

$$p(x|\lambda,\chi,\psi) \;=\; \frac{\chi^{-\lambda}\left(\sqrt{\chi\psi}\right)^{\lambda}}{2K_\lambda\left(\sqrt{\chi\psi}\right)}x^{\lambda-1}\exp\left(-\frac{1}{2}\left[\chi x^{-1}+\psi x\right]\right)$$

The GIG distribution was first proposed by E. Halphen under the pseudonym Dugué (1941) and it was popularized in statistics by Barndorff-Nielsen (1978). Moments of the GIG distribution are $\mathrm{E}\left[X^k\right]=\left(\frac{\chi}{\psi}\right)^{\frac{k}{2}}\frac{K_{\lambda+k}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}$, which are needed throughout the paper for computations of the filtering and smoothing distributions for the latent variance. The c.d.f. of a GIG random variable needed for the calculation of the quantiles can be expressed in terms of the incomplete Bessel function. Slevinsky and Safouhi (2010) provide a routine for its calculation that does not use numerical integration.

A random variable $Z \sim \mathrm{Sichel}\left(\lambda,\chi,\psi\right)$ is obtained by taking a Poisson random variable $Z \sim \mathrm{Poisson}\left(X\right)$ and allowing the mean X to be a random draw from a GIG distribution $X \sim \mathrm{GIG}(\lambda,\chi,\psi)$. The mass function for a Sichel random variable is

$$p(z|\lambda,\chi,\psi) \;=\; \left(\sqrt{\frac{\psi}{\psi+2}}\right)^{\lambda}\left(\sqrt{\frac{\chi}{\psi+2}}\right)^{z}\frac{1}{z!}\frac{K_{\lambda+z}\left(\sqrt{\chi\left(\psi+2\right)}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}$$

The first two moments of a Sichel distribution are often reported incorrectly in the literature. They follow from the law of iterated expectations as

$$\mathrm{E}[z] \;=\; \mathrm{E}_x\left[\mathrm{E}\left(z|x\right)\right] \;=\; \mathrm{E}_x[x] \;=\; \left(\frac{\chi}{\psi}\right)^{\frac{1}{2}}\frac{K_{\lambda+1}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}$$

$$\mathrm{E}\left[z^2\right] \;=\; \mathrm{E}_x\left[\mathrm{E}\left(z^2|x\right)\right] \;=\; \mathrm{E}_x\left[x+x^2\right] \;=\; \left(\frac{\chi}{\psi}\right)^{\frac{1}{2}}\frac{K_{\lambda+1}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}+\left(\frac{\chi}{\psi}\right)\frac{K_{\lambda+2}\left(\sqrt{\chi\psi}\right)}{K_\lambda\left(\sqrt{\chi\psi}\right)}$$

The probabilities of the Sichel distribution and the density of the V.G. distribution can be computed in a computationally stable manner using a recursive relationship for the ratio of modified Bessel functions.

$$\frac{K_{\lambda+1}(x)}{K_\lambda(x)} \;=\; \frac{K_{\lambda-1}(x)}{K_\lambda(x)}+\frac{2\lambda}{x}$$

Proof of Proposition 1:

The conditional likelihood obtained by integrating the variance h_t out of the original measurement equation is

$$\begin{aligned}
p(y_t|z_t;\theta) &= \int_0^\infty p(y_t|h_t;\theta)p(h_t|z_t;\theta)dh_t \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} h_t^{-1/2} \exp\left(-\frac{1}{2}\left[y_t-\mu-\beta h_t\right]^2 h_t^{-1}\right) \frac{1}{\Gamma(\nu+z_t)}\left(\frac{1}{c}\right)^{(\nu+z_t)} h_t^{(\nu+z_t)-1}\exp\left(-\frac{h_t}{c}\right)dh_t \\
&= \left(\frac{1}{c}\right)^{\nu+z_t} \frac{\sqrt{2}\left|y_t-\mu\right|^{\nu+z_t-\frac{1}{2}} K_{\nu+z_t-\frac{1}{2}}\left(\sqrt{\frac{2}{c}+\beta^2}|y_t-\mu|\right)}{\sqrt{\pi}\Gamma(\nu+z_t)\left(\sqrt{\frac{2}{c}+\beta^2}\right)^{\nu+z_t-\frac{1}{2}}} \exp\left(|y_t-\mu|\beta\right) \\
&= \text{V.G.}\left(\mu, \sqrt{\frac{2}{c}+\beta^2}, \beta, \nu+z_t\right)
\end{aligned}$$

see the appendix above for the definition of the density.

By Proposition 2 below, we know that $p(h_{t-1}|z_{t-1}, y_{1:t-1};\theta) = \text{GIG}\left(\nu+z_t-\frac{1}{2}, (y_t-\mu)^2, \frac{2}{c}+\beta^2\right)$. The transition distribution of the discrete mixing variable conditional on observing $y_{1:t-1}$ is given by

$$\begin{aligned}
p(z_t|z_{t-1}, y_{1:t-1};\theta) &= \int_0^\infty p(z_t|h_{t-1};\theta)p(h_{t-1}|z_{t-1}, y_{1:t-1};\theta)dh_{t-1} \\
&= \int_0^\infty \left(\frac{\phi h_{t-1}}{c}\right)^{z_t}\frac{1}{z_t!}\exp\left(-\frac{\phi h_{t-1}}{c}\right) \frac{\left[(y_{t-1}-\mu)^2\right]^{-(\nu+z_{t-1}-\frac{1}{2})}\left(\sqrt{(y_{t-1}-\mu)^2\left(\frac{2}{c}+\beta^2\right)}\right)^{\nu+z_{t-1}-\frac{1}{2}}}{2K_{\nu+z_{t-1}-\frac{1}{2}}\left(\sqrt{(y_{t-1}-\mu)^2\left(\frac{2}{c}+\beta^2\right)}\right)} \\
&\quad h_{t-1}^{\nu-\frac{1}{2}+z_{t-1}-1}\exp\left(-\frac{1}{2}\left[(y_{t-1}-\mu)^2 h_{t-1}^{-1}+\left(\frac{2}{c}+\beta^2\right)h_{t-1}\right]\right)dh_{t-1} \\
&= \left(\frac{\frac{2}{c}+\beta^2}{\frac{2(1+\phi)}{c}+\beta^2}\right)^{(\nu+z_{t-1}-\frac{1}{2})/2}\left(\frac{\phi}{c}\right)^{z_t}\left(\frac{|y_{t-1}-\mu|}{\sqrt{\frac{2(1+\phi)}{c}+\beta^2}}\right)^{z_t} \\
&\quad \frac{1}{z_t!}\frac{K_{\nu+z_{t-1}+z_t-\frac{1}{2}}\left(\sqrt{(y_{t-1}-\mu)^2\left(\frac{2(1+\phi)}{c}+\beta^2\right)}\right)}{K_{\nu+z_{t-1}-\frac{1}{2}}\left(\sqrt{(y_{t-1}-\mu)^2\left(\frac{2}{c}+\beta^2\right)}\right)} \\
&= \text{Sichel}\left(\nu+z_{t-1}-\frac{1}{2}, \frac{\phi}{c}(y_{t-1}-\mu)^2, \frac{c}{\phi}\left[\frac{2}{c}+\beta^2\right]\right)
\end{aligned}$$

The Markov transition distribution of z_t is non-homogenous as it depends on the most recent observation y_{t-1}.

The initial distribution of z_1 was given by Gouriéroux and Jasiak (2006) but we provide it for completeness.

Recall that h_0 is a draw from the stationary distribution $\text{Ga}\left(\nu, \frac{c}{1-\phi}\right)$.

$$
\begin{aligned}
p(z_1;\theta) &= \int_0^\infty p(z_1|h_0;\theta)p(h_0;\theta)dh_0 \\
&= \int_0^\infty \frac{1}{z_1!}\left(\frac{\phi h_0}{c}\right)^{z_1} \exp\left(-\frac{\phi h_0}{c}\right) \frac{h_0^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{h_0(1-\phi)}{c}\right)\left(\frac{c}{1-\phi}\right)^{-\nu} dh_0 \\
&= \frac{\Gamma(\nu+z_1)}{\Gamma(z_1+1)\Gamma(\nu)}(1-\phi)^\nu \phi^{z_1} \\
&= \text{Negative Binomial}(\nu,\phi)
\end{aligned}
$$

where $E[z_1] = \frac{\nu\phi}{1-\phi}$. This completes the proof. ∎

Proof of Proposition 2:

The one-step ahead predictive distribution $p(h_{t+1}|z_{1:t}, y_{1:t}\theta) = \text{Gamma}(\nu + z_t, c)$ follows immediately from the definition of the model. The marginal filtering distribution of the variance is

$$
\begin{aligned}
p(h_t|z_{1:t}, y_{1:t};\theta) &\propto p(y_t|h_t;\theta)p(h_t|z_t;\theta) \\
&\propto h_t^{-1/2} \exp\left(-\frac{1}{2}\left[y_t - \mu - \beta h_t\right]^2 h_t^{-1}\right) \frac{1}{\Gamma(\nu+z_t)} c^{-(\nu+z_t)} h_t^{(\nu+z_t)-1} \exp\left(-\frac{h_t}{c}\right) \\
&\propto h_t^{\nu-1/2+z_t-1} \exp\left(-\frac{1}{2}\left[(y_t-\mu)^2 h_t^{-1} + \left(\frac{2}{c}+\beta^2\right)h_t\right]\right) \\
&= \text{GIG}\left(\nu + z_t - \frac{1}{2}, (y_t-\mu)^2, \frac{2}{c}+\beta^2\right)
\end{aligned}
$$

The conditional distribution only depends on z_t and y_t. The final distribution at time $t = T$ is just a special case of the former result

$$
p(h_T|z_{1:T}, y_{1:T};\theta) = \text{GIG}\left(\nu + z_T - \frac{1}{2}, (y_T-\mu)^2, \frac{2}{c}+\beta^2\right).
$$

The marginal smoothing distribution for the variance is

$$
\begin{aligned}
p(h_t|z_{1:T}, y_{1:T};\theta) &\propto p(y_t|h_t;\theta)p(h_t|z_t;\theta)p(z_{t+1}|h_t) \\
&\propto h_t^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[y_t - \mu - \beta h_t\right]^2 h_t^{-1}\right) \frac{1}{\Gamma(\nu+z_t)} c^{-(\nu+z_t)} h_t^{(\nu+z_t)-1} \exp\left(-\frac{h_t}{c}\right) \\
&\quad \left(\frac{\phi h_t}{c}\right)^{z_{t+1}} \frac{1}{z_{t+1}!} \exp\left(-\frac{\phi h_t}{c}\right) \\
&\propto h_t^{(\nu+z_t+z_{t+1}-\frac{1}{2})-1} \exp\left(-\frac{1}{2}\left[(y_t-\mu)^2 h_t^{-1} + \left(\frac{2(1+\phi)}{c}+\beta^2\right)h_t\right]\right) \\
&= \text{GIG}\left(\nu + z_t + z_{t+1} - \frac{1}{2}, (y_t-\mu)^2, \frac{2(1+\phi)}{c}+\beta^2\right)
\end{aligned}
$$

The marginal smoothing distribution of h_t at intermediate dates only depends on z_t, z_{t+1} and y_t. The distribution for the initial condition is independent of the data conditional on z_1. I find

$$\begin{aligned} p(h_0|z_{1:T}, y_{1:T}\theta) &\propto p(z_1|h_0;\theta)p(h_0;\theta) \\ &\propto \frac{1}{z_1!}\left(\frac{\phi h_0}{c}\right)^{z_1} \exp\left(-\frac{\phi h_0}{c}\right)\frac{h_0^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{h_0(1-\phi)}{c}\right)\left(\frac{c}{1-\phi}\right)^{-\nu} \\ &\propto h_0^{\nu+z_1-1} \exp\left(-\frac{h_0}{c}\right) \\ &= \text{Gamma}\left(\nu + z_1, c\right) \end{aligned}$$

This completes the proof. ∎

Proof of Proposition 3:

Recall that the stationary distribution of the variance is a gamma distribution $h_0 \sim \text{Ga}\left(\nu, \frac{c}{1-\phi}\right)$. One iteration of the Markov transition kernel leaves the distribution invariant such that $h_1 \sim \text{Ga}\left(\nu, \frac{c}{1-\phi}\right)$. The first contribution to the log-likelihood is

$$\begin{aligned} p(y_1;\theta) &= \int_0^\infty p(y_1|h_1;\theta)p(h_1;\theta)dh_1 \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} h_1^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(y_1-\mu-\beta h_1\right)^2 h_1^{-1}\right) \exp\left(-\frac{h_1(1-\phi)}{c}\right)\left(\frac{c}{1-\phi}\right)^{-\nu} \frac{1}{\Gamma(\nu)} h_1^{\nu-1} dh_1 \\ &= \left(\frac{1-\phi}{c}\right)^{\nu} \frac{1}{\Gamma(\nu)}\frac{1}{\sqrt{2\pi}} \exp\left(\beta|y_1-\mu|\right) \int_0^\infty h_1^{\nu-\frac{1}{2}-1} \exp\left(-\frac{1}{2}\left[\left(y_1-\mu\right)^2 h_1^{-1} + \left[\frac{2(1-\phi)}{c}+\beta^2\right] h_1\right]\right) dh_1 \\ &= \left(\frac{1-\phi}{c}\right)^{\nu} \frac{\sqrt{2}\left|y_1-\mu\right|^{\nu-\frac{1}{2}} K_{\nu-\frac{1}{2}}\left(\left(\sqrt{\frac{2(1-\phi)}{c}+\beta^2}\right)\left|y_1-\mu\right|\right)}{\sqrt{\pi}\Gamma(\nu)\left(\sqrt{\frac{2(1-\phi)}{c}+\beta^2}\right)^{\nu-\frac{1}{2}}} \exp\left(\beta\left|y_1-\mu\right|\right) \\ &= \text{V.G.}\left(\mu, \sqrt{\frac{2(1-\phi)}{c}+\beta^2}, \beta, \nu\right) \end{aligned}$$

The filtering distribution $p(h_1|y_1;\theta)$ immediately follows from Bayes' rule

$$\begin{aligned} p(h_1|y_1;\theta) &\propto p(y_1|h_1;\theta)p(h_1;\theta) \\ &\propto h_1^{-1/2} \exp\left(-\frac{1}{2}\left[\left(y_1-\mu\right)^2 h_1^{-1} + \beta^2 h_1\right]\right) \exp\left(-\frac{h_1(1-\phi)}{c}\right) h_1^{\nu-1} \\ &\propto h_1^{\nu-\frac{1}{2}-1} \exp\left(-\frac{1}{2}\left[\left(y_1-\mu\right)^2 h_1^{-1} + \left(\frac{2(1-\phi)}{c}+\beta^2\right) h_1\right]\right) \\ &= \text{GIG}\left(\nu-\frac{1}{2}, \left(y_1-\mu\right)^2, \frac{2(1-\phi)}{c}+\beta^2\right) \end{aligned}$$

The conditional filtering distribution of z_1 is

$$\begin{aligned}
p(z_1|y_1;\theta) &\propto p(y_1|z_1;\theta)p(z_1;\theta) \\
&\propto \frac{1}{\Gamma(\nu+z_1)}\left(\frac{\phi}{c}\right)^{z_1}\frac{K_{\nu+z_1-\frac{1}{2}}\left(\sqrt{\frac{2}{c}+\beta^2}|y_t-\mu|\right)}{\left[(y_1-\mu)^2\right]^{-z_1}\left(\sqrt{\frac{2}{c}+\beta^2}|y_1-\mu|\right)^{z_1}}\frac{1}{z_1!}\Gamma(\nu+z_1) \\
&\propto \left(\frac{\phi}{c}\frac{|y_1-\mu|}{\sqrt{\frac{2}{c}+\beta^2}}\right)^{z_1}\frac{1}{z_1!}K_{\nu+z_1-\frac{1}{2}}\left(\sqrt{\frac{2}{c}+\beta^2}|y_t-\mu|\right) \\
&= \text{Sichel}\left(\nu-\frac{1}{2},\frac{\phi}{c}(y_1-\mu)^2,\frac{c}{\phi}\left[\frac{2(1-\phi)}{c}+\beta^2\right]\right)
\end{aligned}$$

A Sichel distribution is typically derived as a Poisson random variable with a GIG mixing distribution. Here, it is the posterior distribution of a variance gamma likelihood combined with a negative binomial prior. The one-step ahead predictive distribution of z_2 is

$$\begin{aligned}
p(z_2|y_1;\theta) &= \int_0^\infty p(z_2|h_1;\theta)p(h_1|y_1;\theta)dh_1 \\
&= \int_0^\infty \frac{1}{z_2!}\left(\frac{\phi h_1}{c}\right)^{z_2}\exp\left(-\frac{\phi h_1}{c}\right) \\
&\quad \frac{\left[(y_1-\mu)^2\right]^{-(\nu-\frac{1}{2})}\left(|y_1-\mu|\sqrt{\frac{2(1-\phi)}{c}+\beta^2}\right)^{\nu-\frac{1}{2}}}{2K_{\nu-\frac{1}{2}}\left(|y_1-\mu|\sqrt{\frac{2(1-\phi)}{c}+\beta^2}\right)}h_1^{(\nu-\frac{1}{2})-1} \\
&\quad \exp\left(-\frac{1}{2}\left[(y_1-\mu)^2h_1^{-1}+\left(\frac{2(1-\phi)}{c}+\beta^2\right)h_1\right]\right)dh_1 \\
&= \left(\frac{\frac{2(1-\phi)}{c}+\beta^2}{\frac{2}{c}+\beta^2}\right)^{(\nu-\frac{1}{2})/2}\left(\frac{\phi}{c}\right)^{z_2}\left(\frac{\sqrt{(y_1-\mu)^2}}{\sqrt{\frac{2}{c}+\beta^2}}\right)^{z_2}\frac{1}{z_2!}\frac{K_{\nu+z_2-\frac{1}{2}}\left(|y_1-\mu|\sqrt{\frac{2}{c}+\beta^2}\right)}{K_{\nu-\frac{1}{2}}\left(|y_1-\mu|\sqrt{\frac{2(1-\phi)}{c}+\beta^2}\right)} \\
&= \text{Sichel}\left(\nu-\frac{1}{2},\frac{\phi}{c}(y_1-\mu)^2,\frac{c}{\phi}\left[\frac{2(1-\phi)}{c}+\beta^2\right]\right)
\end{aligned}$$

This is the usual construction of a Sichel distribution as a GIG mixture of Poisson random variables. This completes the proof. ∎