State Estimation with Stable Errors and Whisker Particles

J. Huston McCulloch*

March 10, 2021

Keywords: State space estimation, particle filter, stable distributions, infinite variance, whisker particles, paradigm shifts, rank-stratified sampling, stock returns, inflation, Bitcoin returns

JEL Codes: C11 Bayesian Analysis; C15 Statistical Simulation Models; C22 Time Series Models; G12 Asset Pricing

* Adjunct Professor, New York University and Professor Emeritus, Ohio State University. Email mcculloch.2@osu.edu.

The author is indebted to audiences at 25th Annual Conference on Computing in Economics and Finance at Carleton University, the NYU Econometrics Seminar, the NYU Tandon School Financial Engineering Seminar, and the Florida International University Economics Seminar, as well as to Neil Shephard, Michael Pitt, Simon Godsill, and several anonymous NSF referees for helpful comments and suggestions. Special thanks are due to John Nolan of American University for developing his STABLE computer package and for making it available.

The latest version of the paper, together with Matlab programs, will be online via www.asc.ohio-state.edu/mcculloch.2/papers/ParticleFilter/

ABSTRACT

This study estimates the local level model with heavy-tailed stable errors, by means of a particle filter. The infinite-variance stable errors capture both sudden regime shifts and occasional large measurement errors. Stable distributions are chosen because of their role in the Generalized Central Limit Theorem. However, the method is readily adaptable to other heavy-tailed distributions.

The study introduces the use of "Whisker particles" to facilitate the early detection of regime shifts without an excessive number of particles. In order to minimize randomness, the study also employs rank-stratified sampling in the initialization, resampling, and propagation steps. Simulations show that the proposed method dominates the Basic particle filter of Gordon, Salmond and Smith (1993) and Kitagawa (1996), the Adaptive particle filter of Pitt and Shephard (1999), and the Mixture Kalman filter of Chen and Liu (2000) using the skew-stable Gaussian mixture approximation of Lemke, Riabiz and Godsill (2015), in terms of Mean Absolute Error for 10,000 or fewer particles. It is also considerably faster than the Mixture Kalman filter.

The model is fit to stock returns, inflation, and Bitcoin returns, with characteristic exponent α estimates of 1.88, 1.77, and 1.71, respectively. Normality is overwhelmingly rejected in each case. The estimated signal/noise ratio is unexpectedly high in all cases.

"Life is a gamble, at terrible odds;
If it were a bet, you wouldn't take it."
— Tom Stoppard,
Rosenkrantz and Guildenstern are Dead

1. INTRODUCTION AND SUMMARY OF RESULTS

This study estimates a simple state-space model with heavy-tailed stable errors, by means of a particle filter. The infinite-variance stable errors capture both the sudden regime shifts and occasional large measurement errors that are often present in economic and financial data. The study introduces the use of "Whisker particles" in order to facilitate the early detection of regime shifts without an excessive number of particles or over-reaction to observation noise. Stable distributions are chosen because of their role in the Generalized Central Limit Theorem. However, the method is readily adaptable to other heavy-tailed distributions, such as Student's t.

In order to minimize randomness, the study employs rank-stratified sampling in the initialization, resampling, and propagation steps. The only random element is that the propagation step employs random permutations of a non-random stratified sample from the regime-shift distribution.

Simulations show that with 10,000 or fewer particles, the proposed Whisker filter dominates the Basic particle filter of Gordon, Salmond and Smith (1993) and Kitagawa (1996), as well as the Adaptive particle filter of Pitt and Shephard (1999), in terms of Mean Absolute Error. With 100,000 particles, the three methods give similar results.

The Mixture Kalman filter of Chen and Liu (2000), using the skew-stable normal mixture algorithm of Lemke, Riabiz and Godsill (2015), is found to be less accurate than the Whisker filter and to run considerably slower than the Whisker filter with an equal number of components. A modification of the Mixture Kalman filter with unequal probability components governed by a Beta distribution is marginally more accurate, but still is dominated by the Whisker filter.

The model is fit to stock returns, inflation, and Bitcoin returns, with characteristic exponent α estimates of 1.88, 1.77, and 1.71, respectively. Normality ($\alpha = 2$) is overwhelmingly rejected and skewness is pronounced in each case. The estimated signal/noise ratio is unexpectedly high in all cases, which warrants further research.

The simple univariate state variable of the present paper is intended to be a prequel to a more flexible multivariate state variable model. Future research will also implement a two-filter particle smoother.

2. THE LOCAL LEVEL MODEL

The "Local Level Model" (LLM) is the simplest state-space model. In it, an unobserved univariate state variable x_t follows a random walk for $t = 1, \ldots T$, while an observed sequence y_t equals the state variable plus a white-noise measurement error:

$$y_t = x_t + \varepsilon_t \quad (1)$$

$$x_t = x_{t-1} + \eta_t \quad (2)$$

The transitory measurement errors ε_t, sometimes called the "noise," are mutually independent with mean zero, density $\mathrm{f}(\varepsilon)$, and cumulative distribution $\mathrm{F}(\varepsilon)$. The permanent regime shifts η_t, sometimes called the "signal," are independent of one another as well as the measurement errors, with mean zero, density $\mathrm{g}(\eta)$, and cumulative distribution $\mathrm{G}(\eta)$.

The LLM, having a univariate state variable with trivial dynamics, is overly simplistic for most purposes, but serves as a first step to the treatment of more general Time Varying Parameter (TVP) models, e.g. multiple regression with drifting coefficients as in Recursive Least Squares as proposed by Sargent (1999), Evans and Honkapohja (2001), and McCulloch (2005).

The "filter" distribution $\mathrm{p}(x_t|\mathbf{Y}_t)$, where $\mathbf{Y}_t = (y_1, y_2, \cdots y_t)$, gives the posterior distribution of x_t, given the data observed to up to and including time t. If the errors are all Gaussian, the well-known Kalman Filter provides a fast, closed-form solution to this filtering problem. However, if either or both of the errors are non-Gaussian, numerical methods are required.

In the general case of potentially non-Gaussian error distributions, the filtering problem may be solved by the iterative Bayesian procedure due independently to Alspach and Sorenson (1972) and Kitagawa (1987), and presented by Harvey (1990, pp. 162 ff). [1] This involves the following series of Initialization (I), Propagation (P), and Updating (U) steps:

(I) Initialize filter for $t = 1$ using Bayes' Rule with a uniform prior on x_1:

$$\mathrm{p}(x_1|y_1) = \mathrm{f}(y_1 - x_1) \quad (3)$$

[1] The stable signal processing approach of Nolan (2008) provides a rolling-window stable Maximum Likelihood estimate of the signal, as if the signal were constant during the period of the window. However, this approach does not rigorously take into account the time-variation of the signal.

(P) Propagate to obtain the predictive prior for x_{t+1} from the filter for x_t,

$$\mathrm{p}(x_{t+1}|\mathbf{Y}_t) = \int \mathrm{g}(x_{t+1} - x_t)\mathrm{p}(x_t|\mathbf{Y}_t)\mathrm{d}x_t \quad (4)$$

(U) Update to obtain the posterior filter for x_{t+1}, again using Bayes' Rule:

$$p(x_{t+1}|\mathbf{Y}_{t+1}) \propto p(x_{t+1}|\mathbf{Y}_t)\mathrm{f}(y_{t+1} - x_{t+1}) \quad (5)$$

Iteration: Replace t with t+1 and repeat P and U until $t > T$.

3. STABLE DISTRIBUTIONS

Economic shocks are often the cumulative outcome of the decisions of countless individuals over innumerable time sub-increments. It is therefore often appropriate to use a Central Limit Theorem (CLT) to reduce the universe of potential choices for f(ε) and g(η).

According to the Classical (or 19th century) CLT, the sum of a large number of i.i.d. finite variance random variables converges in distribution, after adjusting location and scale, to a standard normal distribution. Unfortunately, measurement errors and regime shifts are commonly too leptokurtic or heavy-tailed to be normal.

However, according to the Generalized (or 20th century) CLT, if the sum of a large number of i.i.d. random variables converges in distribution, after adjusting location and scale, to a limiting distribution, then the limiting distribution must be a member of the *stable* class (Zolotarev 1986, Samoridnitsky and Taqqu 1994). The normal distribution is a member of this class, but it is the thinnest-tailed member, and the only one with a finite variance. This CLT property led Mandelbrot (1963), Fama and Miller (1972), and others to propose the infinite-variance stable distributions as a more realistic alternative to the normal when returns are too leptokurtic for normality. See McCulloch (1996a) for an overview of financial applications of stable distributions and Taleb (2010) for a popularized discussion.

A stable distribution S(α, β, c, μ) with density $s_{\alpha,\beta,c,\mu}(x)$ and cumulative distribution $S_{\alpha,\beta,c,\mu}(x)$ is determined by four parameters: The *characteristic exponent* α lies in the interval (0, 2], and determines the heaviness of the tails. When $\alpha = 2$, the distribution is normal with mean μ and variance $2\sigma^2$. When $\alpha < 2$, the variance of a stable random variable X is infinite and one or both tails have a Pareto-like or "Paretian" power shape:

$$\mathrm{P}(|X| > x) = \mathrm{O}(x^{-\alpha}). \quad (6)$$

The *skewness parameter* $\beta \in [-1, 1]$ determines the relative weight of the two tails:

$$\beta = \lim_{x\to\infty} \frac{1-S(x)-S(-x)}{1-S(x)+S(-x)} . \quad (7)$$

The *scale parameter* $c \in (0, \infty)$ is approximately the semi-interquartile range. The *location parameter* μ is the mean provided $\alpha > 1$ so that the mean exists.

The standard stable density $s_{\alpha,\beta,1,0}(x)$ is determined by its Fourier transform or characteristic function, whose logarithm is:

$$\log \mathrm{cf}_{\alpha,\beta,1,0}(t) = \log \mathrm{E}\, e^{iXt} = \begin{cases} -|t|^{\alpha}\left(1 - i\beta \tan\frac{\pi\alpha}{2}\right), & \alpha \neq 1 \\ -|t|\left(1 + i\beta\frac{2}{\pi}\mathrm{sgn}(t)\log|t|\right), & \alpha = 1 \end{cases}$$

If X is distributed S(α, β, 1, 0), then $Y = cX + \mu$ is distributed S(α, β, c, μ) when c and μ are taken as scale and location parameters.[2] Figure 1 plots the standard symmetric stable densities with $\beta = 0$ for $\alpha = 0.9, 1.0, \ldots 2.0$. Figure 2 plots standard skew-stable densities with $\alpha = 1.5$ for $\beta = -1.0, -0.8, \ldots 1.0$.

[2] This is the Mean-Focus Cartesian parameterization, used by DuMouchel (1975), in which the location parameter μ is the mean for $\alpha > 1$ and what the author has called the *focus of stability* for $\alpha < 1$, and which is a location-scale family for all values of α and β. It corresponds to that of Samorodnitsky and Taqqu (1994) and to the "S1" parameterization of Nolan (2009), except in the *afocal* cases $\alpha = 1$, $\beta \neq 0$, which do not concern us here. See McCulloch (1996b) for details. Other parameterizations include the Continuous Cartesian (Nolan "S0," Matlab default), which is computationally convenient but in which the location parameter has no simple interpretation, and the Polar, which is often mathematically convenient.

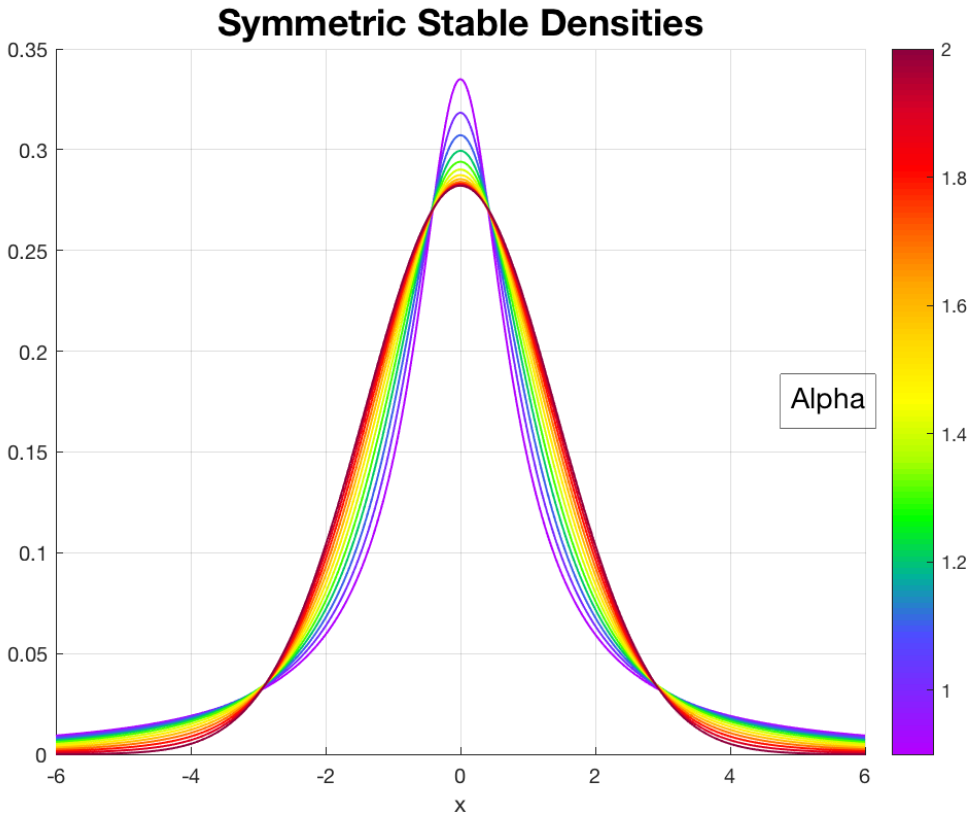

Figure 1

Symmetric Stable densities for $\alpha = 0.9, 1.0, \ldots 2.0$, with $\beta = 0$, $c = 1$, and $\mu = 0$.

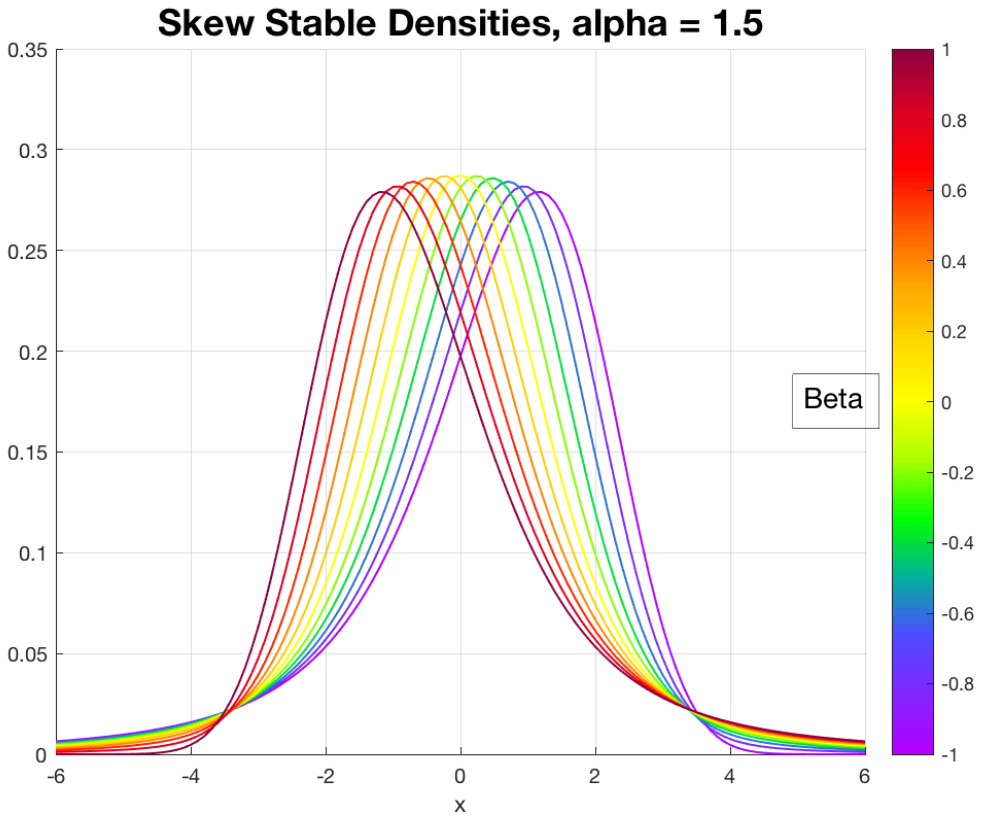

Figure 2

Skew-stable densities, with $\beta = -1.0, -0.8, \ldots 1.0$, $\alpha = 1.5$, $c = 1$, and $\mu = 0$.

If X_1 and X_2 are independent stable random variables distributed S(α, β, c_i, μ_i), i = 1, 2, then their sum $X_3 = X_1 + X_2$ is distributed S(α, β, c_3, μ_3), where

$$\mu_3 = \mu_1 + \mu_2 \quad (\text{unless } \alpha = 1 \text{ and } \beta \neq 0),$$[3]

and scale c_3 determined by the following *scale rule*:

$$c_3^\alpha = c_1^\alpha + c_2^\alpha. \tag{8}$$

If X_1 and X_2 have the same α but different β's, X_3 is distributed S(α, β_3, c_3, μ_3), with c_3 and μ_3 as above, and skewness determined by the following *skewness rule*:

$$\beta_3 = (c_1^\alpha \beta_1 + c_2^\alpha \beta_2)/c_3^\alpha. \tag{9}$$

High-precision numerical approximations to the stable density, distribution, and inverse distribution are now available in John Nolan's STABLE program (2009), as well as in the Matlab 2016+ Statistics Toolbox.[4] The computations in this study for the most part use the Nolan "quick" routines that are very fast and adequately precise. For technical reasons, these programs do not at present compute the stable functions for α within 0.01 (Nolan) or 0.02 (Matlab) of 1.0, unless $\beta = 0$ or $\alpha = 1.0$. However, it is not empirically restrictive in economics and finance to constrain α to be greater than 1.02.

We assume in this study that the measurement error distribution f(ε) and regime shift distribution g(η) are both stable with a common $\alpha \geq 1.02$ and means $\mu = 0$, but with different scales c_ε and c_η. In order to accommodate the skewness that often appears in economic and financial data, we allow the measurement errors to be skewed, with parameter β_ε, but constrain the regime shifts to be symmetric, with $\beta_\eta = 0$.[5]

Stable random variables are easily and precisely generated by means of the algorithm of Chambers, Mallows, and Stuck (1976), which is incorporated into the Nolan and Matlab packages. Figure 3 depicts illustrative stable errors for $T = 100$, $\alpha = 1.7$, $\beta_\varepsilon = 0.3$, $c_\varepsilon = 1.0$, and $c_\eta = 0.25$. The regime shifts η_t are represented by blue diamonds, while the measurement errors ε_t are represented by red stars. This realization happens to have a particularly dramatic (negative) regime shift at $t = 53$, as well as a particularly dramatic (negative) measurement error at $t = 72$.[6] Figure 4 shows the corresponding state variable x_t (blue diamonds), and observations y_t.(red stars).

---

[3] See McCulloch (1996b) concerning the "afocal" cases with $\alpha = 1$ and $\beta \neq 0$.

[4] Matlab uses the Continuous Cartesian (Nolan S0) parameterization, in which the location parameter differs from μ by a location shift equal to $\beta c \tan(\pi\alpha/2)$ when $\beta \neq 0$ and $\alpha \neq 1$.

[5] In the multivariate sequel to this study discussed in section 15, it will be expedient to assume that the regime shifts have an elliptical joint stable distribution, and therefore will in any event be symmetrical.

[6] Although this illustration was randomly generated, it was selected from several such realizations for its particularly dramatic outliers. Although $\beta_\varepsilon = 0.3$, the largest ε_t happens to be negative. In a large sample, the probability that the largest draw and β will have opposite signs is $(1-\beta)/2 = 0.35$ in this case, so that this is not a rare occurrence.

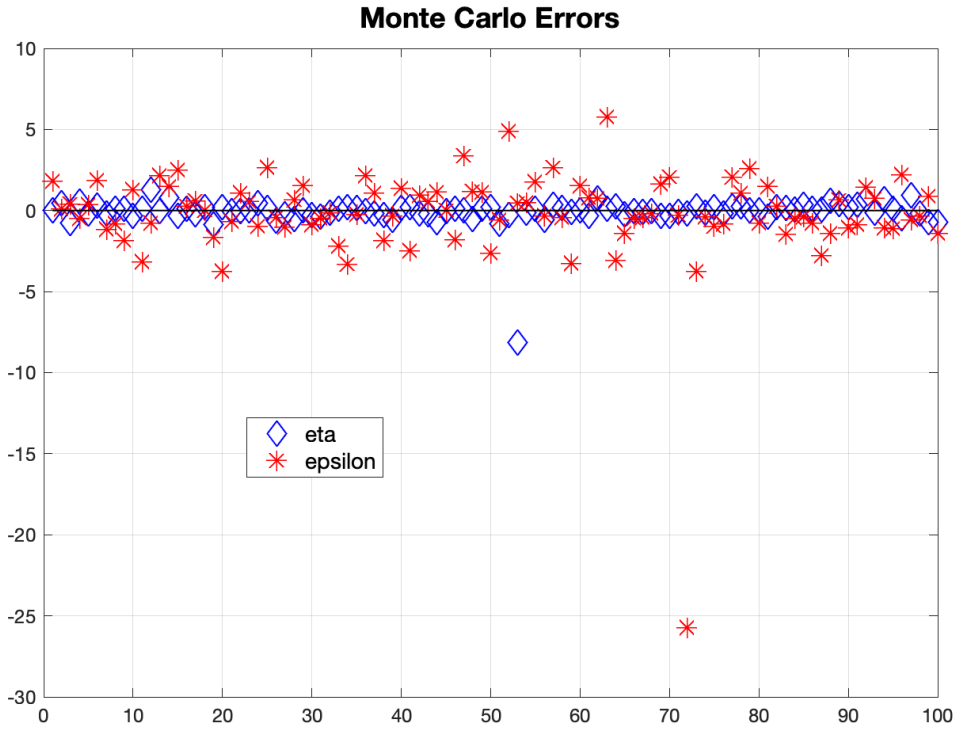

Figure 3

Illustrative stable errors with $T = 100$, $\alpha = 1.7$, $\beta_\varepsilon = 0.3$, $c_\varepsilon = 1.0$, and $c_\eta = 0.25$. Regime shifts η_t are represented by blue diamonds, while measurement errors ε_t are represented by red stars.

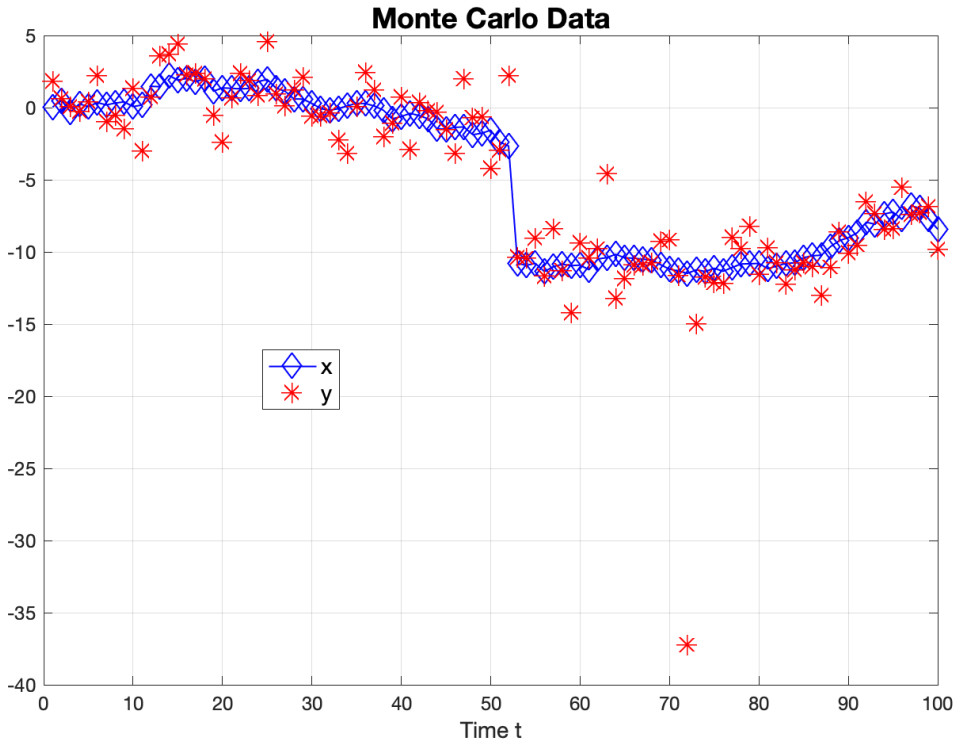

Figure 4

Illustrative state variable x_t (blue diamonds), and observations y_t.(red stars) corresponding to the errors in Figure 3.

4. PARTICLE FILTERING

When the state variable is one-dimensional, evaluating the propagation integral (4) with N-point numerical integration at each of N x-values and each of T points in time requires $O(TN^2)$ operations. This is tedious but feasible (e.g. Oh 1994, Bidarkota and McCulloch 1998). However, in the more interesting cases when the state variable is k-dimensional, this integral would require N^k operations at each of N^k points, for a total of $O(TN^{2k})$ operations, which quickly becomes intractable.

Particle Filtering instead approximates the requisite distributions with N mass points, and replaces the propagation integral with a simulation requiring only Nk operations, or $Nk \log(N)$ after the sorting that is required by the stratified resampling employed in the present paper, for a total of only $O(TNk \log(N))$ operations. The present

paper therefore implements particle filtering with a univariate state variable, as a first step toward an ultimately multivariate model.

In Particle Filtering, introduced independently by Gordon, Salmond and Smith (1993) and Kitagawa (1996), each of the required distributions is approximated by a step function determined by a set of N ordinates x_i and corresponding probabilities p_i summing to 1. We will denote such a particle representation by $\langle x_i, p_i \rangle = \{(x_i, p_i), i = 1, \ldots N\}$.

Such a set of particles may be said to represent a continuous distribution $\mathrm{H}(x)$ if, after sorting the particles in increasing order by x and defining $P_i = \sum_{j=1}^{i} p_j$, the step (S) function defined by

$$\widehat{\mathrm{H}}_{\mathrm{S}}(x) = \begin{cases} 0, x \in (-\infty, x_1] \\ P_i, x \in (x_i, x_{i+1}] \\ 1, \ x \in (x_N, \infty) \end{cases} \qquad (10)$$

approximates $\mathrm{H}(x)$. It is common in the filtering literature to select the particles randomly, as follows:

$$x_i = \mathrm{H}^{-1}(u_i), u_i \sim \mathrm{U}(0,1); \quad p_i = 1/N.$$

This random sampling implies

$$\text{s. d. } \widehat{\mathrm{H}}_{\mathrm{S}}(x) = \big(\mathrm{H}(x)\big(1 - \mathrm{H}(x)\big)/N\big)^{1/2} = \mathrm{O}\big(N^{-1/2}\big)$$

This approximation error can be greatly reduced by instead employing *rank-stratified sampling,* as used already by Malik and Pitt (2011), as follows:

$$x_i = \mathrm{H}^{-1}\left(\tfrac{i-0.5}{N}\right); \quad p_i = 1/N$$

This rank-stratified sampling implies

$$\sup_{x \in \mathbb{R}} \left|\widehat{\mathrm{H}}_{\mathrm{S}}(x) - \mathrm{H}(x)\right| \leq 1/N$$

Clearly rank-stratified sampling is more accurate than random sampling. Note that in rank-stratified sampling, each ordinate x_i represents the median of its respective probability interval $[P_{i-1}, P_i]$.

The *Effective Sample Size* (ESS) of a particle approximation is defined by

$$\mathrm{ESS} = 1/\sum_{i=1}^{N} p_i^2 \ .$$

Clearly ESS = N if all particles have equal probability $1/N$. If the particles have been drawn randomly and independently, ESS is inversely proportional to the variance of the particle estimate of the mean of a finite-variance distribution.[7] However, if the particles are not independent, as is often the case in particle filtering, equal weights are not

[7] Although stable shocks have infinite variance when $\alpha < 2$, the filter distribution has finite variance after $t = 1$ so long as $\alpha > 1$, because its tail densities are then the product of the $\mathrm{O}(x^{-(\alpha+1)})$ power tails of the predictive density and those of the likelihood of the newest observation.

necessarily optimal. A low ESS is clearly undesirable, but a high ESS may be deceptive if the particles are not independent.

The Basic particle filter (with rank-stratification) proceeds as follows:

I (Initialization): The particle filter is initialized for $t = 1$ by a stratified draw $< x^{\mathrm{F}}_{1,i}, p^{\mathrm{F}}_{1,i} >$ from the initial filter density $\mathrm{f}(y_1 - x_1)$:

$$x^{\mathrm{F}}_{1,i} = y_1 - \mathrm{F}^{-1}\left(1 - \tfrac{i-0.5}{N}\right);\ \ p^{\mathrm{F}}_{1,i} = 1/N,\ \ i = 1, \ldots N$$

R (Resampling): Given the time t filter $< x^{\mathrm{F}}_{t,i}, p^{\mathrm{F}}_{t,i} >$, the resampled particles $< x^{\mathrm{R}}_{t,i}, p^{\mathrm{R}}_{t,i} >$ are an equal-weighted rank-stratified draw from the interpolated step-function distribution defined by the sorted filter particles. Many prior studies instead inefficiently draw a random sample with replacement from the multinomial distribution defined by the filter particles.[8] (This resampling step is redundant when $t = 1$, since the particles already have equal weights.)

P (Propagation): The time t predictive (P) or prior distribution is simulated with particles $< x^{\mathrm{P}}_{t,i}, p^{\mathrm{P}}_{t,i} >$, where

$$x^{\mathrm{P}}_{t,i} = x^{F}_{t,i} + \hat{\eta}_{\mathrm{J}_t(i)};\qquad p^{\mathrm{P}}_{t,i} = p^{\mathrm{F}}_{t,i},$$

the $\hat{\eta}_i, i = 1, \ldots N$ are a perfectly representative stratified sample of size N from the signal distribution $\mathrm{G}(\eta)$, and $\mathrm{J}_t(i)$ is a random permutation of the first N integers. Most, if not all, previous studies instead employ an imperfectly representative random sample from $\mathrm{G}(\eta)$.

U (Updating): The new time $t + 1$ posterior filter distribution is simulated with particles $< x^{\mathrm{F}}_{t+1,i}, p^{\mathrm{F}}_{t+1,i} >$, where

$$p^{\mathrm{F}}_{t+1,i} \propto p^{\mathrm{P}}_{t,i}\ \mathrm{f}(y_{t+1} - x^{\mathrm{P}}_{t,i}\ );\quad x^{\mathrm{F}}_{t+1,i} = x^{\mathrm{P}}_{t,i};$$

and the probabilities are normalized to sum to 1.

Iteration: Replace t with $t + 1$, and repeat steps R, P, and U until $t > T$.

Bayesian credible intervals (CIs) for the state variable x_t may be derived by inverting the interpolated cumulative distribution defined by the filter particles. Many

[8] Significant exceptions are Kitagawa (1996), Carpenter et al. (1999), and Malik and Pitt (2011). However, Kitagawa and Carpenter et al. take the particles in arbitrary order, and hence do not consolidate adjacent weak particles. Although their sampling is stratified, it is not rank-stratified. The same is also true of the "stratified," "systematic" and "residual" methods described by Hol et al. (2006). Malik and Pitt (2011) do use rank-stratified resampling, with interpolation, as in the present paper.

authors instead use the resampled filter particles, but this is a mistake: Although the resampled filter has a "perfect" ESS of *N*, it obviously contains no more information than the raw filter. Its high ESS is deceptive, because its particles are even less independent than those of the raw filter. Even if the resampling is rank-stratified, the resampled filter is a corrupt, "lossy" version of the raw filter, that effectively rounds large filter probabilities to an integer multiple of 1/*N*, and that erases much of the detailed information contained in the weak particles. Going forward, it is essential to resample in order to maximize the independence of the propagation step and to eliminate weak particles, but the raw filter is the definitive simulation of the filter distribution.

For the purpose of resampling and computing credible intervals, the "lumpiness" of the step-function approximation to the filter distribution may be mitigated inside the interval $[x_1^{\mathrm{F}}, x_N^{\mathrm{F}}]$ by considering each particle to represent the median of its probability interval, and then linearly interpolating between the midpoints $Q_{t,i}^{\mathrm{F}}$ of the risers in the step function approximation, as follows:

$$\widehat{\mathrm{H}}_t^{\mathrm{F}}(x) = Q_{t,i-1}^{\mathrm{F}} + \left(Q_{t,i}^{\mathrm{F}} - Q_{t,i-1}^{\mathrm{F}}\right) \frac{x - x_{t,i-1}^{\mathrm{F}}}{x_{t,i}^{\mathrm{F}} - x_{t,i-1}^{\mathrm{F}}}, \quad x \in [x_{t,i-1}^{\mathrm{F}}, x_{t,i}^{\mathrm{F}}], \tag{11}$$

where

$$Q_{t,i}^{\mathrm{F}} = \left(P_{t,i-1}^{\mathrm{F}} + P_{t,i}^{\mathrm{F}}\right)/2; \quad P_{t,i}^{\mathrm{F}} = \sum_{j=1}^{i} p_{t,j}^{\mathrm{F}}; \quad P_{t,0}^{\mathrm{F}} \equiv 0.$$

Malik and Pitt (2011) employ such an interpolation in their resampling step and note the additional advantage that it makes the estimated likelihood a continuous function of the underlying hyperparameters.

However, if we interpret $x_{t,1}^{\mathrm{F}}$ as telling us the median of the probability interval $(0, p_{t,1}^{\mathrm{F}})$, probability $p_{t,1}^{\mathrm{F}}/2$ could lie anywhere in $(-\infty, x_{t,1}^{\mathrm{F}})$, and similarly for probability $p_{t,N}^{\mathrm{F}}/2$ above $x_{t,N}^{\mathrm{F}}$. It would clearly be overly confident to place all this probability at $x_{t,1}^{\mathrm{F}}$ or $x_{t,N}^{\mathrm{F}}$ as per the step-function approximation, and ordinarily even to extrapolate linearly below $\left[x_{t,1}^{\mathrm{F}},\ x_{t,2}^{\mathrm{F}}\right]$ or above $\left[x_{t,N-1}^{\mathrm{F}},\ x_{t,N}^{\mathrm{F}}\right]$. In the interest of erring on the side of caution, we therefore consider a credible interval boundary to be undefined (i.e. $\pm\infty$) if its probability is less than $p_{t,1}^{\mathrm{F}}/2$ or greater than $1 - p_{t,N}^{\mathrm{F}}/2$. It may therefore be informative to have some small probability particles at each end of the filter distribution, rather than striving for equal probabilities throughout.

Figure 5 shows the Bayesian 50% and 95% credible intervals that result when the Basic filter is applied to the data in Figure 4, with $N = 100$ and rank-stratification. The bold line is the posterior median. It may be seen that the Basic filter is quite slow to detect the pronounced regime shift that occurs at $t = 53$. On close inspection, it may also be seen that for several periods after this shift, the lower bound of the 95% CI is

undefined, indicating that the particles are so uneven in probability that $p_{t,1}^{\mathrm{F}}/2$ is greater than 0.025, despite having used 100 particles.

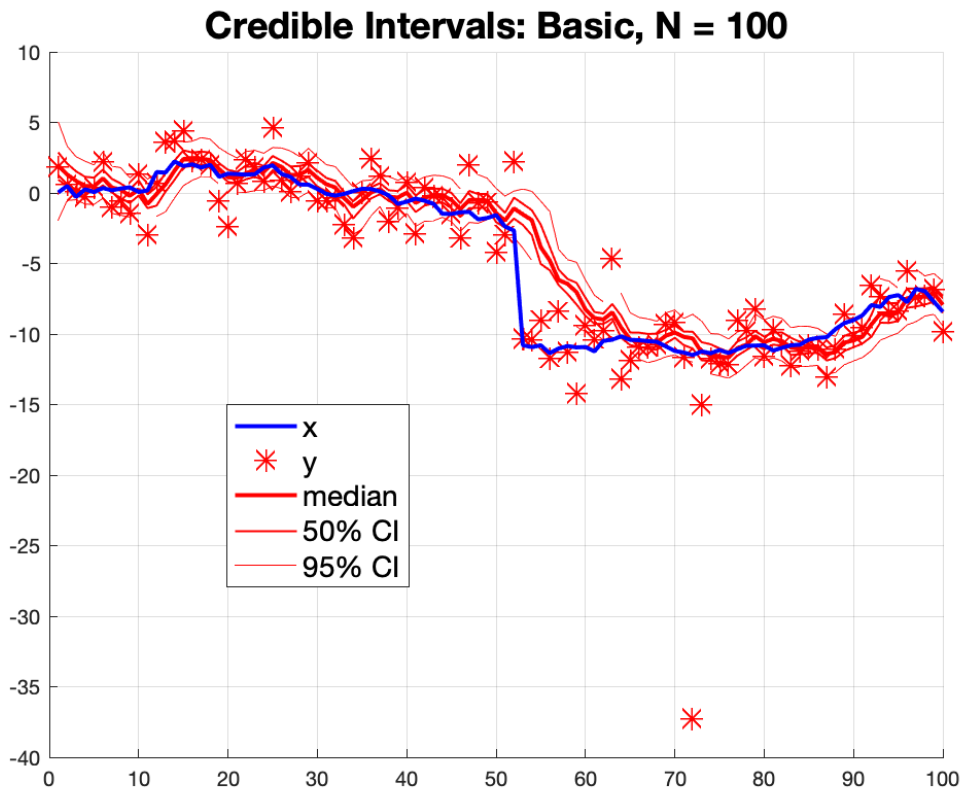

Figure 5

Bayesian 50% and 95% credible intervals for the data of Figure 4, using the Basic filter, with $N = 100$. The heavy red line is the posterior filter median. The blue line represents the unobserved state variable x_t. The red stars are the observations y_t.

Mathematically, the filter distribution has infinite support, so that when an outlier like y_{53} or y_{72} appears, a weak, albeit perhaps microscopic, mode will immediately appear in the filter in its vicinity. If subsequent observations reinforce this outlier, this second mode will become stronger, and eventually there will be a "paradigm shift" in the sense of Kuhn (1962), in which the second mode abruptly and nonlinearly becomes stronger than the first and eventually takes over.[9] However, the Basic resampling step treats the

[9] If the errors are all Gaussian as assumed by the familiar Kalman filter, the posterior filter is also Gaussian and therefore unimodal, the filter mean reacts linearly to the new observation, and such nonlinear Kuhnian paradigm shifts do not occur.

support of the filter distribution as $[x^{\mathrm{F}}_{t,1}, x^{\mathrm{F}}_{t,N}]$ without extrapolation, or as a small extension of this interval if linear extrapolation is employed, so that it is literally impossible that a regime shift outside this range could ever have occurred. Zero times anything is still zero, so the second mode never develops, no matter how many times it is reinforced with observations. Eventually, the propagation step will generate particles in the vicinity of the data, but it may take an inordinately large number of particles to capture the regime shift in a timely manner through this diffusion alone.

5. ADAPTIVE RESAMPLING

In order to increase the number of predictive particles in the vicinity of a new observation, and at the same time to make the t+1 filter probabilities as nearly equal as possible, Pitt and Shephard (1999) propose "adaptively" resampling $x^{\mathrm{F}}_{t,i}$ with probability proportional to $p^{\mathrm{F}}_{t,i}\mathrm{a}_t(x^{\mathrm{F}}_{t,i})$, where $\mathrm{a}_t(x)$ is an "auxiliary distribution" centered on y_{t+1}, rather than with probability $p^{\mathrm{F}}_{t,i}$ itself. In order to preserve the filter probabilities, they then assign each draw $x^{\mathrm{R}}_{t,j}$ obtained from $x^{\mathrm{F}}_{t,i}$ a resampled probability $p^{\mathrm{R}}_{t,j}$ proportional to $1/\mathrm{a}_t(x^{\mathrm{F}}_{t,i})$.

The time t+1 updated filter weights will be equal in expectation (given x_t and y_{t+1} but not η_{t+1} or x_{t+1}), if

$$\begin{aligned}\mathrm{a_t}(x_t) &= \mathrm{E}(\mathrm{f}(y_{t+1} - x_{t+1})|x_t, y_{t+1}) \\ &= \int \mathrm{f}(y_{t+1} - x_t - \eta_{t+1})\mathrm{g}(\eta_{t+1})d\eta_{t+1} \\ &= \varphi(y_{t+1} - x_t)\ ,\end{aligned}$$

where the "bridge" distribution $\varphi(\bullet)$ is the convolution of $\mathrm{f}(\varepsilon)$ and $\mathrm{g}(\eta)$. For most choices of $\mathrm{f}(\varepsilon)$ and $\mathrm{g}(\eta)$, this convolution is intractable and an approximation must be substituted.[10] However, if $\mathrm{f}(\varepsilon)$ and $\mathrm{g}(\eta)$ are stable with our assumptions, the above scale and skewness rules (8) and (9) imply

$$\varphi(\bullet) \sim \mathrm{S}(\alpha, \beta_\varphi, c_\varphi, 0),$$

where

$$\begin{aligned}c^\alpha_\varphi &= c^\alpha_\varepsilon + c^\alpha_\eta, \\ \beta_\varphi &= \beta_\varepsilon c^\alpha_\varepsilon / c^\alpha_\varphi,\end{aligned}$$

so that

$$\mathrm{a}_t(x) \sim \mathrm{S}(\alpha, -\beta_\varphi, c_\varphi, y_{t+1})\ .$$

[10] For example, if $\mathrm{f}(\varepsilon)$ and $\mathrm{g}(\eta)$ both have finite variance, it would ordinarily be adequate to scale $\mathrm{f}(\varepsilon)$ up in proportion to the standard deviation of $\varepsilon + \eta$.

Note that although this Adaptive resampling increases the ESS of the time t+1 filter, it at the same time reduces the independence of the resampled particles, as well as the ESS of the time t predictive density. It therefore provides at best a mixed benefit.

Figure 6 shows 50% and 95% filter credible intervals for the state variable, for the data of Figure 4 and $N = 100$, using Adaptive resampling with rank-stratification from the implied step function. Again, the filter is slow to detect the obvious regime shift at $t = 53$, although it is not quite as slow as the Basic filter. Although this Adaptive filter does generate more particles near to y_{53}, it is still constrained to draw resampled points entirely from the interval $[x^{\mathrm{F}}_{t,1}, x^{\mathrm{F}}_{t,N}]$, under the usual interpretation that it is impossible that x_t could ever be outside this interval. Despite the effort to make the filter particles as equal in weight as possible, the lower 95% CI boundary is again undefined as the filter catches up with the data after $t = 53$, and again at $t = 72$.

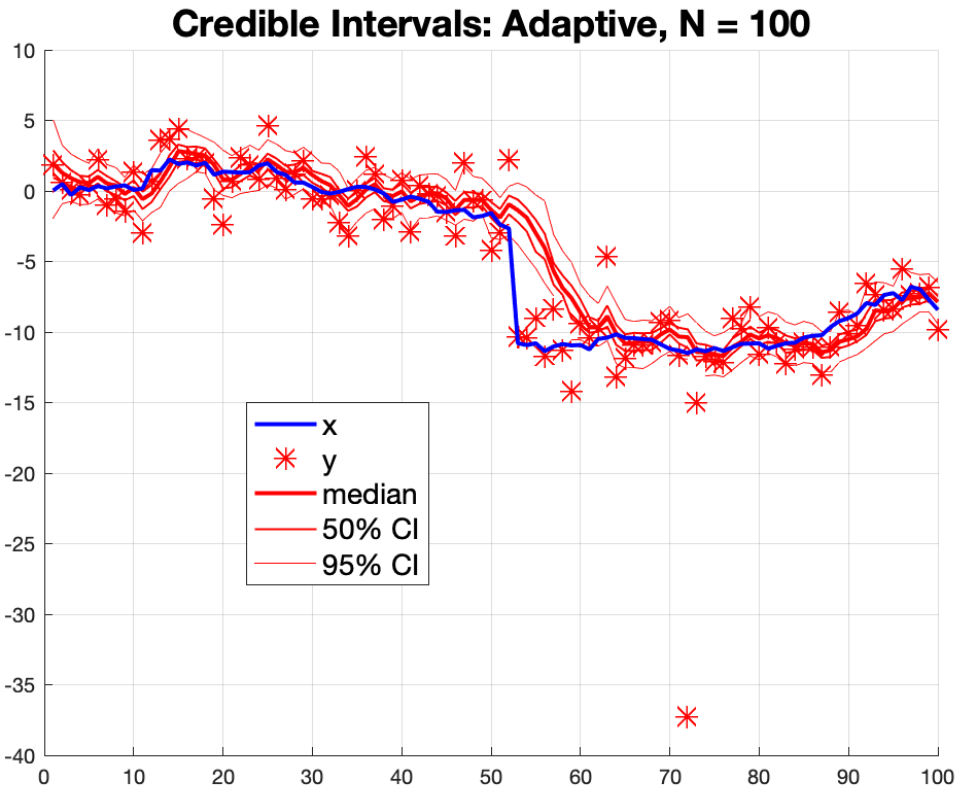

Figure 6

Bayesian 50% and 95% credible intervals for the data of Figure 4, using the Adaptive filter with $N = 100$. The heavy red line is the posterior filter median. The blue line represents the unobserved state variable x_t. The red stars are the observations y_t.

6. WHISKER PARTICLES

In nature, many species have evolved negligible-mass extensions of their bodies, called "whiskers," which enable them to detect objects in their vicinity, even in complete darkness, before they actually bump into them. Taking a cue from nature, the present study therefore adds minute "Whisker particles" outside the range of the filter particles to the particle set when resampling, in order to provide early detection of possible regime shifts.

In order to concentrate some precision in the neighborhood of y_{t+1} in the spirit of Adaptive resampling, while at the same time preserving much of the independence in the time t filter, we will resample the time t filter distribution by drawing $N_A = \theta_A N$ points from $A_t(x)$, the cumulative of $a_t(x)$, and $N_B = N - N_A$ points from the time t filter itself, for some $\theta_A \in [0, 1]$. It is easy enough to draw N_A stratified points from $A_t(x)$, but the ones we are primarily interested in may lie outside the interval $[x^F_{t,1}, x^F_{t,N}]$, and even outside its extension by linear extrapolation. In order to calibrate the probability of these points to the filter distribution, we observe that if we interpret $x^F_{t,1}$ as representing the median of the probability interval $(0, p^F_{t,1})$, probability $p^F_{t,1}/2$ could lie anywhere below $x^F_{t,1}$, and similarly for probability $p^F_{t,N}/2$ above $x^F_{t,N}$. It would clearly be overly informative to place all this probability at $x^F_{t,1}$ or $x^F_{t,N}$ as is implicit in the Basic and Adaptive filters. At the other extreme, if we were to a spread it uniformly (with zero density) over the interval $(-\infty, x^F_{t,1})$ or $(x^F_{t,1}, \infty)$, as implicit in the conservative rule for CIs set out above, the Whisker particles in this range would all have 0 probability. The true time t filter density as given by (5) is proportional to the product of the time t–1 predictive density, which ordinarily decreases slowly outside $\left[x^P_{t-1,1}, x^P_{t-1,N}\right] = \left[x^F_{t,1}, x^F_{t,N}\right]$, times the time t likelihood $f(y_t - x)$. A reasonable (albeit still somewhat conservative) way to allocate the filter density in these tail ranges is therefore in proportion to the likelihood, so that it cumulates as follows:

$$\widehat{H}^F_t(x) = \begin{cases} \left(1 - F(y_t - x)\right) \frac{p^F_{t,1}/2}{1-F(y_t - x^F_{t,1})}, & x < x^F_{t,1} \\ 1 - F(y_t - x) \frac{p^F_{t,N}/2}{F(y_t - x^F_{t,N})}, & x > x^F_{t,N}. \end{cases}$$

Inside $[x^F_{t,1}, x^F_{t,N}]$, we may linearly interpolate $\widehat{H}^F_t(x)$ between the midpoints of the risers in the step function approximation, as in equation (11) above.

In order to calibrate the N_A particles representing $A_t(x)$ to this extended filter distribution $\widehat{H}^F_t(x)$, we introduce a cumulative form of importance sampling, rather than

using the customary density form as outlined by Geweke (1989):[11] First, we observe that in general we want each particle $x_{t,i}$ to represent the median of its respective probability interval $[P_{t,i\text{-}1}, P_{t,i}]$, in terms of the calibrating filter distribution. We therefore use the auxiliary distribution to partition the real line into N_A intervals $[z_{t,i\text{-}1}, z_{t,i}]$ with equal probability under the auxiliary distribution, so that

$$z_{t,i} = \text{A}_t^{-1}(i/N_\text{A}), \quad i = 0, \ldots N_\text{A}; \quad z_{t,0} \equiv -\infty, \quad z_{t,N_\text{A}} \equiv +\infty,$$

and then use the interpolated and extended filter distribution to compute calibrated cumulative probabilities for these boundaries:

$$P_{t,i}^\text{A} = \widehat{\text{H}}_t^\text{F}\big(z_{t,i}\big); \quad P_{t,0}^\text{A} \equiv 0, \quad P_{t,N_\text{A}}^\text{A} \equiv 1 \, .$$

At this point, we could use these cumulative probabilities to compute particle probabilities, with ordinates assigned at the midpoints of these probability ranges per the calibrating filter distribution, and then merge these particles with $N_\text{B} = N - N_\text{A}$ basic particles drawn with equal probabilities from the interpolated and extended filter distribution in the usual manner. However, merging the auxiliary and basic particles directly would result in auxiliary particles that had very unequal probabilities in the resampled filter and therefore in the subsequent predictive distribution, along with basic particles that had very unequal expected filter probabilities in the next updating step. We can improve the ESS of both the predictive and expected future filter particles by instead also characterizing the basic particles in terms of their probability boundaries

$$P_i^\text{B} = i/(N_\text{B} + 1), \quad i = 1, \ldots N_\text{B},$$

and then interleaving the two sets of probability boundaries by merging and sorting them, to obtain N+1 Resampled probability boundaries $P_{t,i}^\text{R}$, $i = 0, \ldots N$. We then define

$$p_{t,i}^\text{R} = P_{t,i}^\text{R} - P_{t,i-1}^\text{R}, \; i = 1, \ldots N,$$

$$x_{t,i}^\text{R} = \widehat{\text{H}}_t^{\text{F}\,-1}(P_{t,i}^\text{R} - p_{t,i}^\text{R}/2).$$

(Because the auxiliary probability boundaries already contain the default boundaries 0 and 1, it would be redundant to include these in the basic boundaries. If $N_\text{A} = 0$, we would instead set $P_i^\text{R} = i/N$, $i = 0, \ldots N$.)

Because of the basic boundaries, no interleaved resampled particle will have probability greater than $1/N_\text{B}$, while because of the auxiliary boundaries, no filter particle in the next update step will have expected probability greater than $1/N_\text{A}$. However, it may happen that some of the probability boundaries as computed above will be very close to one of their neighbors, and hence will produce particles with near-zero weight that

[11] Although a particle representation can consistently estimate the cumulative distribution, it does not consistently estimate the density function, since even with interpolation, the arc derivatives between particles tend either to zero or infinity as the number of particles increases. The customary density-based importance sampling therefore will not work if the particles are actually drawn, either randomly or with stratification, from $a_t(x)$. In the Adaptive filter of Section 5, the particles are in fact being drawn from a discrete distribution governed by $a_t(x)$, and not from $A_t(x)$ itself.

essentially go to waste. In order to prevent this and thereby modestly increase the resampled ESS, we "equilibrate" the boundaries, by replacing half of the merged boundaries with the average of their two neighbors before computing the resampled particle probabilities and ordinates. This equilibration step gives at best only a slight improvement in filter performance, but is virtually costless, and unexpectedly turns out to make the computed likelihood a substantially smoother function of the hyperparameters.

Figure 7 shows credible intervals using the Whisker particle filter with $N = 100$, $\theta_A = 0.25$ (as recommended in section 9 below), and using the data of Figure 4. Note that the filter quickly identifies the regime shift at $t = 53$ despite the small number of particles, while recognizing $t = 72$ as a measurement error. We would ordinarily want to use far more than 100 particles in order to reduce granularity and to obtain precise estimates of the filter probabilities, but it is clearly beneficial to start with a method that gives robust results with only a modest value of N.

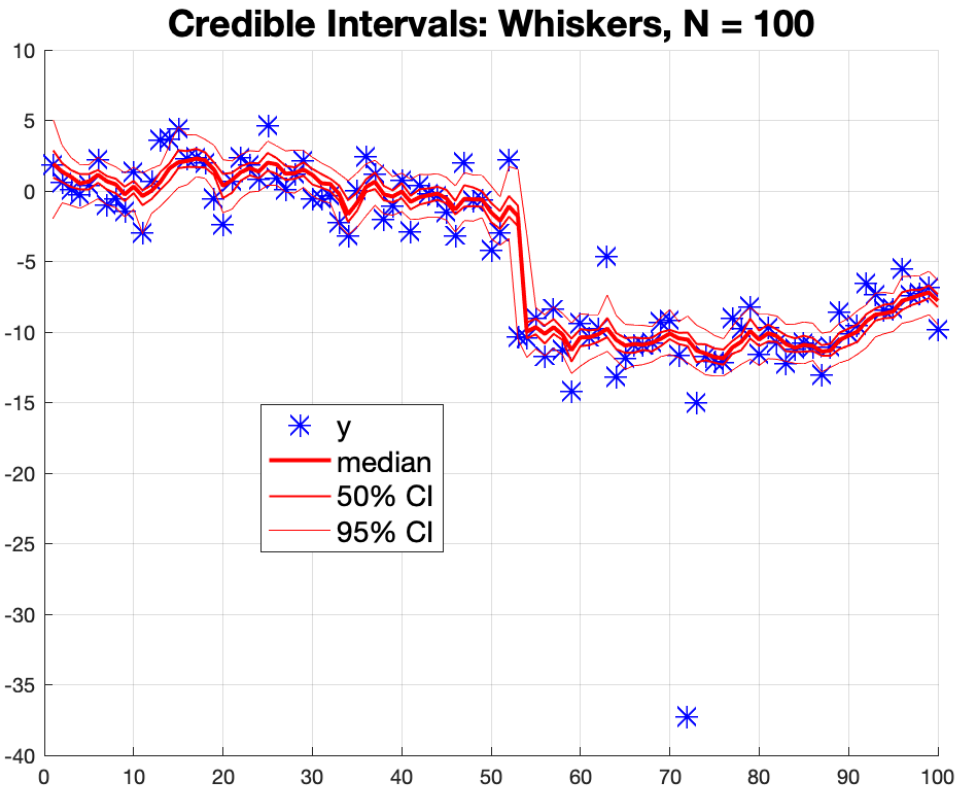

Figure 7

Bayesian 50% and 95% credible intervals for the data of Figure 4, using the Whisker filter, with $N = 100$ and $\theta_A = 0.25$. The heavy red line is the posterior filter median. The blue stars are the observations y_t.

When particle filtering, it is very useful to monitor ESS, since it provides an upper bound on the effective independent sample size and therefore the precision of results. Figure 8 plots the Effective Sample Size (ESS) corresponding to the Whisker filter estimates in Figure 7. At $t = 1$, this necessarily equals the actual number of particles, $N = 100$, since the filter is initialized with equal weights. Afterwards, it never rises above 85, and has an average value of 68.2, indicated by the horizontal red line. It often falls below 50, and the extraordinary regime shift at $t = 53$ causes it to fall to 23.1 immediately afterwards at $t = 54$. However, despite the occasionally low value of ESS, the 95% CI boundaries in Figure 11 are defined for all t. Table 6 in section 9 below shows that occasionally the ESS can fall precariously low, even with $N = 100{,}000$.

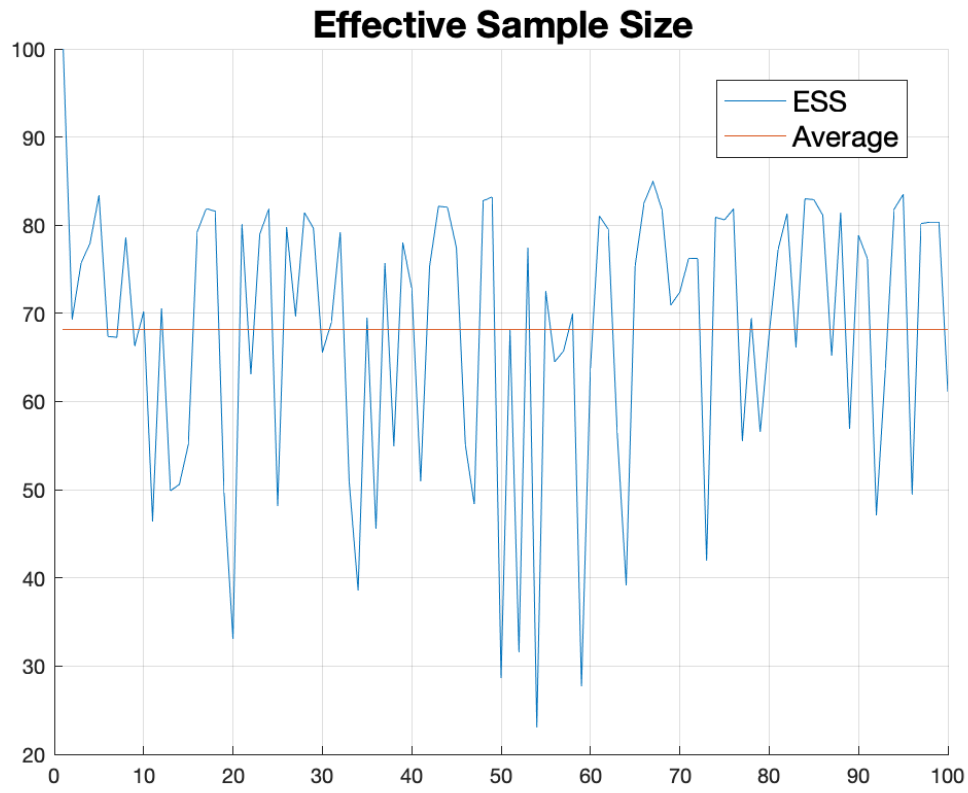

Figure 8

Effective Sample Size (ESS) for the Whisker filter estimates of Figure 7. Red horizontal line indicates average value.

Figure 9 plots credible interval half-ranges (i.e. measured from the posterior median, both above and below) for the whisker filter estimates in Figure 7. The thin red and blue lines show the upper and lower half-ranges, respectively, for the 95% credible interval, while the bold red and blue lines show the upper and lower half-ranges for the 50% credible interval. As in the Gaussian Kalman filter case, the credible interval ranges start off high and then fall as more data is observed. However, unlike the Gaussian case governed by the Kalman filter, the scales do not settle down to a constant and symmetrical value, but rather respond asymmetrically to outliers in both the signal and noise errors. The 95% CIs are particularly wide immediately after the $t = 53$ regime shift, reflecting temporary uncertainty as to which type of shock has just occurred.

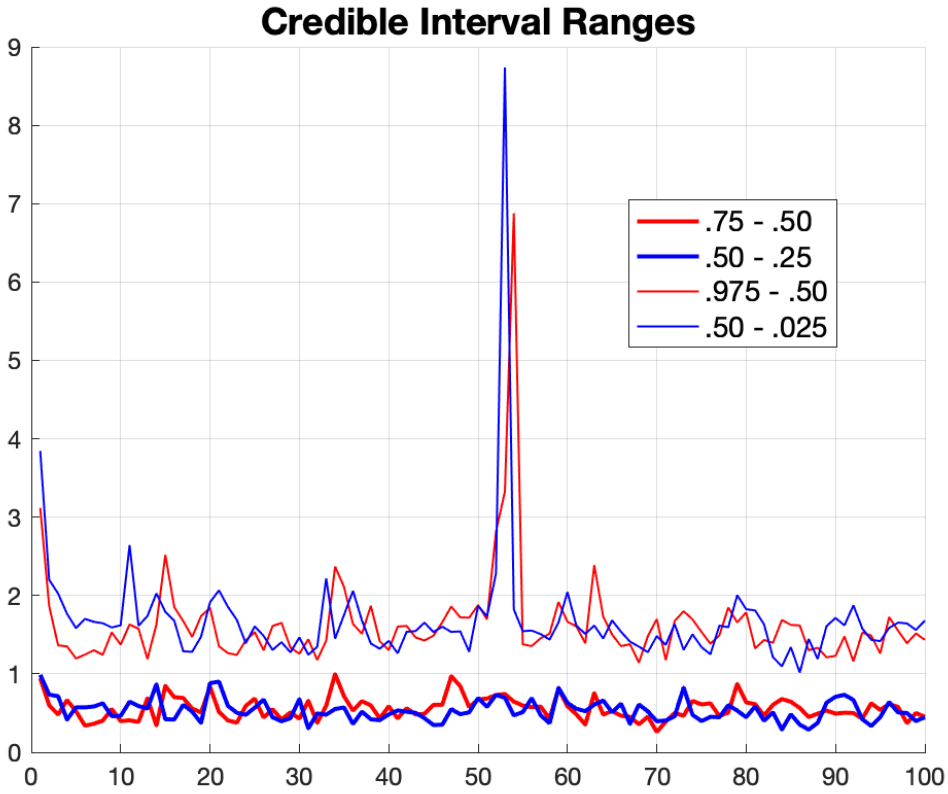

Figure 9

Credible interval half-ranges for the Whisker filter estimates of Figure 7.

Table 1 gives the computation time for one pass through the Whisker Filter algorithm using the data of Figure 4 for $T = 100$, and the same data extended to $T = 1000$, with $N = 100$, 1000, 10,000, 100,000, and 1,000,000. In each case, the algorithm is given

the true value of the parameters, and times do not include any plots or computation of data or credible intervals. Calculations were performed in 64-bit Matlab R2017a on a MacBook Pro with a 2.5GHz Quad-Core Intel Core i7 processor and 16 GB of memory, using the "quick" routines s*tableqkpdf*, *stableqkcd*f, and *stableqkinv* from Nolan's (2009) STABLE 5.1 for Matlab package.

Although the time is theoretically O($TN \log N$), it may be seen that it barely increases by a factor of 2 in passing from $N = 100$ to $N = 1000$, for either value of T, consistent with fixed startup costs. It does increase by almost the predicted factor of 13.3 between $N = 1000$ and 10,000 for either value of T, and 12.5 between $N = 10{,}000$ and 100,000 for $T = 1000$. However, in passing from $N = 100{,}000$ to 1,000,000, the time increases by a factor of 33 for $T = 100$ and 83 for $T = 1000$, far in excess of the predicted factor of 12.0. This unexpected slowing is probably due to the computer running out of high-speed memory, requiring it to page intermediate results out into flash storage, since no attempt was made to conserve on memory use. This issue may also account for the unexpected slowing between $N = 10{,}000$ and 100,000 with $T = 100$, though if anything memory should be less of an issue with $T = 100$ than 1000. In passing from $T = 100$ to $T = 1000$, the time does increase roughly by the expected factor of 10, at least for $N \leq 10{,}000$.

Table 1

Whisker Filter Computation Times

(Seconds)

N	$T = 100$	$T = 1000$
100	0.0327	0.263
1000	0.0545	0.565
10,000	0.584	4.55
100,000	14.1	49.7
1,000,000	1177.	1652.

7. RPU VS. PRU

The traditional sequence of particle filtering steps as described in section 4 is Resample – Propagate – Update, or what might be called RPU. It has recently been proposed by Carvalho et al. (2010) and Singpurwalla et al. (2017) to switch the R and P

steps when using Adaptive resampling as follows: Propagate – Resample – Update, or what might be called PRU. This allows Resampling to be "completely adapted," by setting

$$a_t(x) = \mathrm{f}(y_{t+1} - x),$$

so that the time t+1 filter weights are all 1/N and the Update step becomes trivial.

However, even though PRU makes the filter particle weights all equal, it also makes them even less independent than they are with conventional RPU Adaptive resampling, since many time t+1 filter particles now contain the same time t propagation errors in addition to the same time t filter values. Equal weights do give ESS = N, but this is only optimal to the extent that the particles are independent. This apparent advantage of PRU is therefore illusory. It is more important that Adaptive resampling be modified to generate Whisker particles potentially outside the range of the filter particles as in Section 6.

8. MIXTURE KALMAN FILTERING

Chen and Liu (2000) observe that if f(ε) and g(η) can both be represented as a mixture of normal distributions, with random variances and/or means, the filter distribution can also be approximated by a finite mixture of normal distributions in which the particle mass points are replaced by Gaussian densities, and a large part of the work of filtering can be efficiently performed by the Kalman filter. Chen and Liu call this the "Mixture Kalman Filter" (MKF).

MKF has the theoretical advantage over particle filtering that the implied filter distribution will be smooth and will have unbounded support without the *ad hoc* extensions that we used to calibrate any whisker particles that lie outside the range of the raw filter particles.

In MKF, there is little, if any, advantage to rank-stratifying the resampling over the density means before Resampling, since the density means are not unambiguous indicators of their locations. Resampling should still be stratified, however, but taking the components in their received order with no ranking or shuffling. As long as descendants of a common ancestor are left together sequentially, not shuffling will tend to make the resampling of a common ancestor's descendants more representative. If the resampling is not rank-stratified, there is no point to using interpolation.

Perhaps the biggest disadvantage of MKF is that its Update step relies on a random pairing of prediction and observation densities, whereas any of the particle filter supdates each particle precisely using the observation density with no additional random

element in the Update step. This random pairing greatly increases the random element of the MKF estimates. Likewise, although in theory it would be advantageous use adaptive resampling of the time *t* filter densities, resampling each in proportion to its expected contribution to the time $t+1$ likelihood, this would not be helpful in practice, since the required auxiliary proposal distribution could only be simulated with a great deal of noise or additional computation.

Symmetric stable random variables with exponent α can be precisely represented as scale mixtures of normals, using positive stable random variables with exponent $\alpha/2$ and skewness $\beta = 1$ as the subordinating variances, per a 1955 theorem of Salomon Bochner, given in Proposition 1.3.1 of Samorodnitsky and Taqqu (2004), with their $\alpha' = 2$. If both the observation and transition errors are symmetric stable, it therefore is possible to directly employ the MKF of Chen and Liu, as has already been implemented by Lombardi and Godsill (2006).[12] The Bochner approach has the advantage that it can generate a single, perfectly representative, rank-stratified sample that can be reused with random permutations for each *t*.

However, economic data often are significantly skewed. McCulloch and Percy (2013), e.g., find, using an Extended Neyman Smooth goodness-of-fit test, that four competing symmetric heavy-tailed distributions, including the symmetric stable, can all be rejected as a model for US stock returns, primarily because of the presence of skewness. It follows that at least the observation errors should not be constrained to be symmetric. In the empirical results below, we find β_ε estimates of −0.62, +0.57, and +1.00 for stock returns, inflation, and Bitcoin returns, respectively. Unfortunately, there is no direct extension of the Bochner stable subordination theorem to the skew-stable case at the present time.

Nevertheless, Lemke, Riabiz and Godsill (LRG, 2015) have exploited a representation of skew-stable distributions as an infinite sum of normal random variables, with random variances and means governed by two simple transforms of a realization of an infinite Poisson series, as given by Proposition 1.4.2 of Samorodnitsky and Taqqu (2004), to develop an approximation based on a finite truncation of the Poisson series, with an adjustment that matches the first two moments of the truncated terms. Although the resulting mixture of normals based on the finite truncation is not precisely skew-stable, they report that it can yield a very good approximation.

---

[12] In the model of Lombardi and Godsill (2006), the state variable is a musical signal that they assume to be Gaussian, but its innovations could as easily have been non-Gaussian symmetric stable instead. Their equation (2) erroneously states, in their notation, $X_1X_2 \sim \mathrm{S}(\alpha, 0)$. This should be $X_1\sqrt{X_2} \sim \mathrm{S}(\alpha, 0)$. However, the correct formula is applied in their equation (8) and presumably in their calculations.

One disadvantage of the LRG approach in the context of filtering is that rank-stratification of the observation components in the Update step is no longer feasible. This means that the simulated observation components must be laboriously computed from scratch for each t, and their magnitudes will be erratically random rather than perfectly representative.

Details of our implementation of the MKF using the LRG simulation of skew-stable observation errors are given in the Appendix.

Figure 10 shows credible intervals for the Mixture Kalman Filter estimates of the state variable, using the data of Figure 4 and $N = 100$ components. The MKF quickly detects the regime shift at t = 53, without being unduly distracted by the big observation error at t = 72. For this example, at least, the MKF behaves very similarly to the Whisker filter in Figure 7.

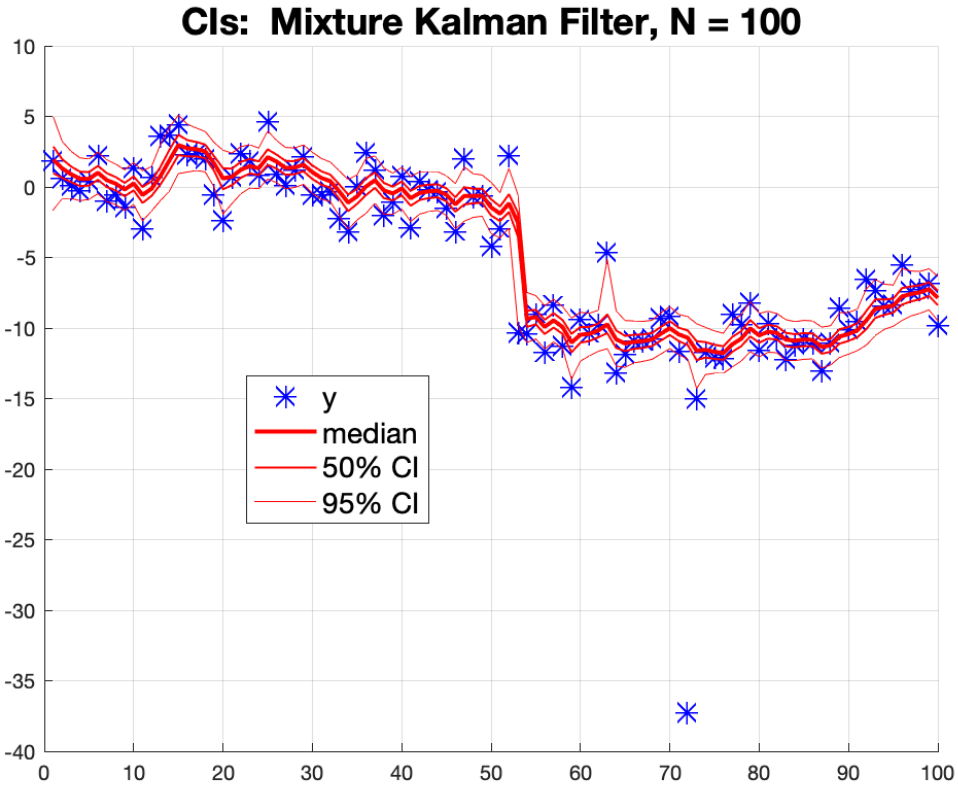

Figure 10

Bayesian 50% and 95% credible intervals for the data of Figure 4, using the Mixture Kalman filter, with $N = 100$. The heavy red line is the posterior filter median. The blue stars are the observations y_t.

Since the MKF does not employ a sorting step when Resampling, its theoretical computation time is O(TN), rather than O(TN log$N$) as for the Whisker Filter. Table 2 below shows the time, in seconds, for one pass through the filter as in Table 1. In most cases, the elapsed time is roughly proportional to both T and N up to N = 100,000. As is the case with the Whisker filter, there is some unpredicted slowing down at N = 1,000,000 though it is relatively minor.

Table 2

Mixture Kalman Filter ComputationTimes
(Seconds)

N	T = 100	T = 1000
100	0.306	2.02
1000	2.08	17.1
10,000	20.1	171.
100,000	218.	1856.
1,000,000	2968.	39,875.

Comparing Tables 1 and 2, it may be seen that the MKF is always slower than the Whisker Filter for equal values of N and T. In the vicinity of N = 10,000, it is about 35 times slower, despite our attempt to vectorize the calculation as much as possible. This relative slowness is most likely due to the calculation-intensive Poisson series method of simulating skew-stable distributions proposed by LRG. Unless the MKF is substantially more accurate than the Whisker Filter for equal values of N, it is clearly dominated by it.

9. SIMULATIONS

In order to compare the relative accuracies of the filtering methods described, we construct Monte Carlo simulations of the LLM and then, for each simulation, compute the Mean Absolute Error (MAE) over all but the first 10 observations, computed as follows:

$$\text{MAE} = \frac{1}{T-10}\sum_{t=11}^{T}\left|\sum_{i=1}^{N} x_{t,i}^{\text{F}} p_{t,i}^{\text{F}} - x_t\right| .$$

The first 10 observations are excluded as being unrepresentative, since any filter necessarily takes some time to "learn" a representative amount about the state variable.

Table 3 tabulates the average MAE over 1000 Monte Carlo simulations, each of length $T = 100$, with $\alpha = 1.1$, 1.3, 1.5, 1.7, and 1.9, and with $N = 100$, 1000, 10,000, and 100,000 particles. In each simulation, $\beta_\varepsilon = 0.3$, $\beta_\eta = 0.0$, and $c_\eta / c_\varepsilon = 0.25$, as in the above illustration. The first two columns show results for the conventional Basic and Adaptive filters as described above, using rank-stratified rather than the customary random sampling and resampling. The next two columns use the proposed Whisker filter, first with 50% adaptive particles, and then with 25% adaptive particles. The same seed was used to generate the data for each of the simulations, so that the different methods use exactly the same simulated series for each α and N. Since the Chambers et al. (1976) stable random number generator is a continuous function of α and β, this also makes the simulations for different values of α as similar as possible. In every case, the filter is given the true values of the stable parameters, so that no hyperparameter estimation error is involved.

All the methods considered are consistent, so that as the number of particles N becomes large, the particle or Mixture Kalman filter approximation converges in distribution to the mathematical filter distribution determined by Equations (3), (4) and (5) above. In all cases, this limit appears to have been well approached with 100,000 particles. However, for $\alpha = 1.1$, 1.3 or 1.5 and $N = 100$, 1000 or 10,000, the Whisker filter is clearly more accurate than either the Basic or Adaptive filters. The same is true for $\alpha = 1.7$ with $N = 100$ or 1000. With $\alpha = 1.9$, the advantage of the Whisker filter is greatly reduced, but still is present. It is interesting that the Adaptive filter has only a small advantage over the Basic filter, and then only for the smaller values of α.

Comparison of columns 3 and 4 of Table 3 shows that while there is not much difference between $\theta_A = 0.50$ and $\theta_A = 0.25$, the latter is slightly better more often than not, and therefore is to be preferred. It should be noted that $\theta_A = 0.0$ is not much different from the Basic filter, so that there is probably not much to be gained from further reduction in θ_A below 0.25. We also attempted 100% Adaptive Whisker particles (θ_A = 1.0, not tabulated), but this occasionally gave a disastrous fall in ESS to within rounding error of 1.0000, and hence is not recommended.

Table 4 tabulates the worst-case maximum value of MAE over the 1000 replications of Table 3. For $N = 100$, 1000, and 10,000, the Whisker filter performs dramatically better than the Basic and Adaptive filters, confirming the results in Table 3. The Basic and Adaptive filters admittedly do happen to outperform the Whisker filter in

this extreme scenario when $N = 100{,}000$. Nevertheless, the average MAE results in Table 4 show that there is essentially no difference across methods in these cases when all 1000 replications are averaged together.

Table 5 gives the average, over the 1000 replications of Tables 3 and 4, of ESS_{min}, the minimum of ESS over the $T = 100$ observations of each simulation. It may be seen that average ESS_{min} is only a small fraction of the actual number of particles N, and generally is smaller, the smaller is α. Even $N = 1000$ often gives precariously small values of ESS_{min}, particularly if α is less than 1.7. Average ESS_{min} does increase with N, but not quite in proportion to it. It is clear that one should use at least 1000, and preferably even 10,000 particles for reliable credible intervals with $T = 100$. As T increases, the odds of encountering difficult observations increase, so that N should prudently be a large multiple of T.

Table 6 gives the minimum over the 1000 replications of ESS_{min}. It is apparent that in these 1000 replications, there is at least one scenario which even the Whisker filter has great trouble analyzing for all t. Although the Whisker filter is generally more robust than the Basic or Adaptive filters, it is clearly not entirely foolproof. These worst-case scenarios deserve further exploration that goes beyond the scope of the present paper.

Table 3

Average Mean Absolute Error
1000 Replications
$T = 100$, $\beta_\varepsilon = 0.3$, $\beta_\eta = 0.0$, $c_\eta / c_\varepsilon = 0.25$

	N	Basic	Adaptive	Whiskers $\theta_A = 0.50$	Whiskers $\theta_A = 0.25$	Mixture Kalman	Beta MKF
$\alpha = 1.1$							
	100	17.80	17.53	2.15	1.90	6.46	2.97
	1000	11.33	11.25	2.37	2.18	4.71	2.40
	10,000	6.53	6.15	2.24	2.03	1.89	2.12
	100,000	1.93	1.87	2.02	2.11	NA	NA
$\alpha = 1.3$							
	100	4.31	4.19	1.14	1.08	2.36	1.42
	1000	2.78	2.76	1.20	1.14	1.91	1.20
	10,000	1.81	1.69	1.14	1.11	1.05	1.11
	100,000	1.06	1.04	1.10	1.10	NA	NA
$\alpha = 1.5$							
	100	1.64	1.60	0.82	0.79	1.21	0.92
	1000	1.22	1.21	0.83	0.81	0.91	0.83
	10,000	0.95	0.91	0.80	0.80	0.78	0.79
	100,000	0.78	0.78	0.79	0.79	NA	NA
$\alpha = 1.7$							
	100	0.89	0.87	0.67	0.66	0.75	0.71
	1000	0.78	0.78	0.67	0.66	0.70	0.67
	10,000	0.69	0.68	0.66	0.66	0.65	0.65
	100,000	0.65	0.65	0.65	0.65	NA	NA
$\alpha = 1.9$							
	100	0.61	0.61	0.58	0.57	0.58	0.59
	1000	0.60	0.60	0.57	0.57	0.58	0.58
	10,000	0.58	0.58	0.57	0.57	0.57	0.57
	100,000	0.57	0.57	0.57	0.57	NA	NA

Note: Average of simulated Mean Absolute Error over 1000 replications with sample size T = 100 each, for $\alpha = 1.1$, 1.3, 1.5, 1.7, and 1.9, using $N = 100$, 1000, 10,000, and 100,000 particles. In all cases, $\beta_\varepsilon = 0.3$, $\beta_\eta = 0.0$, and $c_\eta / c_\varepsilon = 0.25$. In each replication, MAE is computed for t = 11, ... 100. NA indicates Not Attempted.

Table 4

Maximum Mean Absolute Error
1000 Replications
$T = 100$, $\beta_\varepsilon = 0.3$, $\beta_\eta = 0.0$, $c_\eta / c_\varepsilon = 0.25$

	N	Basic	Adaptive	Whiskers $\theta_A = 0.50$	Whiskers $\theta_A = 0.25$	Mixture Kalman	Beta MKF
$\alpha = 1.1$							
	100	2333.7	2333.9	106.2	123.9	946.4	179.8
	1000	2328.9	2329.1	141.4	156.4	1307.8	179.8
	10,000	1720.8	1672.8	157.0	142.4	65.1	179.5
	100,000	91.1	71.2	145.8	148.0	NA	NA
$\alpha = 1.3$							
	100	501.7	501.9	20.4	23.9	288.6	38.4
	1000	495.5	496.1	38.6	27.9	584.4	38.4
	10,000	289.5	220.1	38.5	38.6	14.0	35.7
	100,000	17.9	14.4	27.6	30.5	NA	NA
$\alpha = 1.5$							
	100	147.8	148.1	7.8	6.7	79.2	11.7
	1000	141.0	139.4	12.2	12.1	54.8	11.7
	10,000	63.3	39.1	12.2	12.2	4.5	7.6
	100,000	4.6	4.7	8.3	9.2	NA	NA
$\alpha = 1.7$							
	100	48.9	49.4	3.3	2.7	22.5	4.3
	1000	43.6	43.6	3.5	3.3	19.7	4.4
	10,000	15.0	13.9	4.7	4.6	2.0	2.0
	100,000	2.0	2.0	3.3	3.4	NA	NA
$\alpha = 1.9$							
	100	11.3	11.8	1.5	1.3	3.8	1.8
	1000	10.7	9.7	1.9	1.9	6.2	1.8
	10,000	4.8	4.7	1.5	1.9	1.1	1.1
	100,000	1.1	1.1	1.5	1.5	NA	NA

Note: Maximum of Mean Absolute Error over the 1000 simulations of Table 3. In each replication, MAE is computed for t = 11, ... 100. NA indicates Not Attempted.

Table 5

Average ESS_{min}
1000 Replications
$T = 100$, $\beta_\varepsilon = 0.3$, $\beta_\eta = 0.0$, $c_\eta / c_\varepsilon = 0.25$

	N	Basic	Adaptive	Whiskers $\theta_A = .50$	Whiskers $\theta_A = .25$	Mixture Kalman	Beta MKF
$\alpha = 1.1$							
	100	5.4	7.4	5.4	5.9	3.3	3.3
	1000	23.5	29.0	21.5	26.4	32.7	23.2
	10,000	130.8	144.1	96.0	119.7	326.0	221.5
	100,000	842.3	906.1	543.8	718.7	NA	NA
$\alpha = 1.3$							
	100	10.8	13.2	8.6	10.9	4.1	3.7
	1000	44.7	58.4	36.8	50.1	40.5	27.7
	10,000	260.4	304.3	194.1	242.1	402.9	365.9
	100,000	2017.0	2229.0	1253.0	1734.9	NA	NA
$\alpha = 1.5$							
	100	20.2	33.1	16.1	21.0	5.2	4.4
	1000	75.7	101.1	65.2	80.8	51.5	34.2
	10,000	514.1	622.6	356.2	491.8	513.5	332.6
	100,000	4144.4	4847.1	2642.1	3627.8	NA	NA
$\alpha = 1.7$							
	100	28.9	63.1	23.3	29.4	8.1	6.3
	1000	116.3	168.3	101.7	120.1	78.4	51.2
	10,000	951.2	1283.3	676.5	910.1	780.4	501.3
	100,000	8139.	10,624.	5326.	7379.	NA	NA
$\alpha = 1.9$							
	100	30.9	75.1	23.3	30.1	23.5	15.9
	1000	249.1	496.7	184.4	241.4	220.1	140.2
	10,000	1795.4	3287.6	1419.5	1779.0	2198.0	1383.8
	100,000	16,537.	28,116.	11,830.	15,709.	NA	NA

Note: Average, over the 1000 simulations of Table 3, of ESS_{min}, the minimum of Effective Sample Size over $t = 1, \ldots 100$. NA indicates Not Attempted.

Table 6

Minimum ESS_{min}
1000 Replications
$T = 100$, $\beta_\varepsilon = 0.3$, $\beta_\eta = 0.0$, $c_\eta / c_\varepsilon = 0.25$

	N	Basic	Adaptive	Whiskers $\theta_A = .50$	Whiskers $\theta_A = .25$	Mixture Kalman	Beta MKF
$\alpha = 1.1$							
	100	1.28	1.74	1.34	1.64	1	1
	1000	1.13	1.16	1.06	1.15	1	1.11
	10,000	1.04	1.07	1.08	1.03	1	1.78
	100,000	1.04	2.30	1.37	1.47	NA	NA
$\alpha = 1.3$							
	100	1.55	2.21	1.39	1.84	1	1
	1000	1.56	1.64	1.30	1.41	1	1.00
	10,000	1.11	1.14	1.07	1.08	1.00	1.94
	100,000	1.22	1.31	1.10	1.10	NA	NA
$\alpha = 1.5$							
	100	1.29	6.56	2.00	3.27	1	1
	1000	1.68	4.23	1.90	2.26	1	1.00
	10,000	1.33	1.53	1.56	1.61	1.00	1.59
	100,000	2.41	1.85	1.84	4.11	NA	NA
$\alpha = 1.7$							
	100	1.65	20.21	4.35	5.66	1	1
	1000	2.33	20.36	5.06	5.48	1	1.02
	10,000	1.23	3.62	2.31	2.67	1.00	1.41
	100,000	5.52	15.96	5.28	11.80	NA	NA
$\alpha = 1.9$							
	100	3.36	60.19	5.54	6.71	1	1
	1000	8.24	267.42	18.22	33.19	1	1.03
	10,000	1.59	21.81	6.78	9.73	1.00	1.30
	100,000	3.98	4.64	5.70	3.58	NA	NA

Note: Minimum, over the 1000 simulations of Table 3, of ESS_{min}, the minimum of Effective Sample Size over $t = 1, \ldots 100$. A value of "1" indicates computed value is precisely unity to within machine precision. NA indicates Not Attempted.

Column 5 of Tables 3 through 6 give results for the Mixture Kalman Filter, using the LRG skew-stable normal mixture simulation method as described in the Appendix. Since the MKF is considerably slower than the particle filters, it was not attempted with $N = 100{,}000$ (indicated NA).

Table 3 indicates that the MKF does always perform better than either the Basic or Adaptive filter in terms of Average MAE for equal values of N. However, although the MKF does do as well or slightly better than the Whisker filter with $N = 10{,}000$, the Whisker filter always performs better with $N = 100$ or 1000. The MKF does close in on the common value for $N = 100{,}000$ by $N = 10{,}000$.

The maximum MAE values for the MKF in Table 4 show that the MKF usually does better than the Basic and Adaptive filters by this metric as well, but is outperformed by the Whisker filter again for $N = 100$ and 1000. By $N = 10{,}000$, the MKF approaches the very large N limit more closely than the Whisker filter. With $\alpha \geq 1.7$ and $N = 10{,}000$, the MKF does outperform the Whisker filter by this metric.

The average and minimum ESS_{min} values for the MKF in Tables 5 and 6 are not directly comparable to those for the particle filters, since with the MKF an ESS of 1 is not a degenerate outcome as it is with the particle filters. When $\alpha = 2$, all components of the MKF are identical, so that their probabilities and therefore the ESS are immaterial. However, it is still true that with $\alpha < 2$ a small ESS_{min} is less desireable, and particularly so for smaller values of α. Nevertheless, the minimum ESS_{min} values in Table 6 are all either "1" (indicating an answer within machine precision of 1) or "1.00" (indicating an answer different than 1 but within rounding error of 1.00), indicating that there is at least one of the 1000 scenarios that the MKF has great difficulty analyzing.

The last column in Tables 3 through 6 give results for a modification of the MKF in which the variances in the Propagation and Resampling steps are drawn with unequal probabilities governed by a Beta(1, 0.5) distribution, rather than equal probabilities as governed by a U(0, 1) distribution. This ensures that the MKF has some very high variance (but low probability) components. Details are given in the Appendix.

It may be seen that this "Beta-MKF" filter greatly outperforms the Basic and Adaptive Filters in terms of both average and maximum MAE. However, although it is competitive with the Whisker Filter in terms of MAE, the Whisker Filter almost always outperforms it in terms of average MAE for equal values of N. Since the Beta-MKF is no faster than the standard MKF, the Whisker Filter is still more cost-effective in terms of computation time.

In summary, the proposed Whisker filter clearly outperforms the conventional Basic and Adaptive filters. The Mixed Kalman filter generally does do better than the Basic and Adaptive filters with equal values of *N*, but worse than the Whisker filter except with a very large value of *N*. Given that it is about 35 times slower than the Whisker filter for the same *N*, its use is not justified, let alone mandated, despite its theoretical appeal. The proposed Beta modification of the MKF is marginally dominated by the Whisker filter for equal values of *N*, and is just as slow as the standard MKF, and therefore is not cost-effective.

10. PARAMETER ESTIMATION

With empirical data, the stable parameters are ordinarily unknown and must be estimated from the data itself. A straightforward approach is to numerically maximize the log likelihood

$$\mathcal{L}\left(\alpha, \beta_\varepsilon, c_\varepsilon, c_\eta | \mathbf{Y}_t\right) = \sum_{t=2}^{T} \log\left(l_t\right) \tag{12}$$

where the likelihood contributions

$$l_t = \sum_{i=1}^{N} p_{t-1,i}^{\mathrm{P}} \mathrm{f}(y_t - x_{t-1,i}^{\mathrm{P}}), \quad t = 2, \ldots T$$

are the normalization factors required in the update step. Because of the uniform prior on x_1, the first observation makes no contribution to the likelihood.

However, even though the exact mathematical likelihood is continuous in the hyperparameters α, β_ε, c_ε, and c_η and ordinarily has a unique maximum,[13] the numerical likelihood simulated by a particle filter is full of small discontinuities if the filter distribution is not interpolated when resampling, and may be serrated even if the filter distribution is interpolated. In either case, the numerical likelihood can have numerous local maxima that may be nowhere near the global maximum. As the number of particles increases, the discontinuities or serrations become smaller, but they also become more numerous. The numerical global maximum consistently estimates the exact global maximum, but care must be taken that a numerical maximization routine does not settle on a greatly inferior local maximum. Furthermore, since the derivatives of the simulated likelihood function do not consistently estimate the derivatives of the mathematical likelihood function, Newton-Raphson methods may well founder.

An alternative and much easier likelihood-based approach that largely avoids this problem is what we call the "short-cut method": Under the LLM with our stable

[13] See DuMouchel (1971). If α and/or T are very small, it is possible for the mathematical likelihood to be multimodal with stable distributions. However, so long as $\alpha > 1$ and $T > 100$, this should not be an issue.

assumptions, the first differences of the observed series y_t have a symmetric stable distribution with mean 0 and scale c_1:

$$\Delta y_t = y_t - y_{t-1} = \varepsilon_t - \varepsilon_{t-1} + \eta_t \sim S(\alpha, 0, c_1, 0),$$

where, by the scale rule (8),

$$c_1^\alpha = 2\, c_\varepsilon^\alpha + c_\eta^\alpha. \tag{13}$$

Furthermore, the first differences of y_t at lag 2 will also have a symmetric stable distribution with mean 0 and somewhat larger scale c_2:

$$\Delta_2 y_t = y_t - y_{t-2} = \varepsilon_t - \varepsilon_{t-2} + \eta_t + \eta_{t-1} \sim S(\alpha, 0, c_2, 0),$$

where, by the scale rule,

$$c_2^\alpha = 2\, c_\varepsilon^\alpha + 2\, c_\eta^\alpha. \tag{14}$$

Equations (13) and (14) imply

$$c_\varepsilon^\alpha = c_1^\alpha - c_2^\alpha/2, \tag{15}$$

$$c_\eta^\alpha = c_2^\alpha - c_1^\alpha, \tag{16}$$

so that if we can estimate α, c_1, and c_2, (15) and (16) will give us estimates of c_ε and c_η.

Nolan's package STABLE includes a routine *stablefitmle* that estimates the four stable parameters of an i.i.d. stable sample by Maximum Likelihood, along with their standard errors, as computed from the information matrix.[14] A companion routine *stablefitmlerestricted* allows specified parameters to be fixed, e.g. $\beta = 0$ and $\mu = 0$ in the present application. Adjacent first differences are not independent, but they are independent modulo 2. That is, the first differences are independent for t = 2, 4, ... T (assuming T is even), as well as for t = 3, 5, ... $T-1$. This means that standard ML estimates with asymptotically valid standard errors may be obtained from either the even- or odd-numbered first differences. The two subsets will give somewhat different estimates of α and c_1, but essentially the same standard errors, since both samples are approximately $T/2$ in size. A valid consensus of the two estimates, with all the desirable ML properties of either, may be obtained by adding the two likelihoods together and maximizing their sum. Conveniently, this will give exactly the same point estimate as simply submitting all $T-1$ first differences to the program as if they were independent. The program will return standard errors based on the false assumption that all $T-1$ observations are independent, which will be smaller than those obtained using either half of the data by a factor of approximately $1/\sqrt{2}$, but these may be conservatively corrected (if it is desired to use them), simply by multiplying the reported values by $\sqrt{2}$. We will call these consensus estimates $\hat{\alpha}$ and $\hat{c}_1$.

Similarly, the lag 2 first differences are independent modulo 3, so that an ML estimate of c_2 is valid using every third value, starting with $t = 3$, 4, or 5. A consensus

[14] See DuMouchel (1975).

estimate $\hat{c}_2$ may be obtained as above simply by submitting all $T-2$ lag 2 first differences as if they were independent, constraining $\alpha = \hat{\alpha}$ as well as $\beta = 0$ and $\mu = 0$, and then adjusting the standard error up by a factor of $\sqrt{3}$.[15]

Having thus estimated α, c_ε and c_η, β_ε may be estimated by a univariate search over the log likelihood (12), and then examining the profile to ensure that the discontinuities are not an issue. Inference on β_ε (conditional on the estimated values of the other three parameters) may then be performed using the likelihood ratio test in the usual manner, provided the null hypothesis is not on the boundary of the parameter space $[-1, 1]$. (See Moran 1971, DuMouchel 1973.)

Normality may be tested using the kurtosis statistic K. If an i.i.d. sample of size n has a normal distribution with unknown mean and variance to be estimated, the normalized K-statistic

$$Z_K = (K - 3)/\sqrt{24/n}$$

is known to be asymptotically distributed N(0, 1). As noted above, even and odd numbered values of Δy_t are serially independent, so that either set of first differences could be used to compute K and Z_K, with $n = T/2$ or $T/2-1$. A consensus of the two K and Z_K statistics may be found simply by computing K from the full set of first differences, but then computing Z_K using $T/2$ (which is asymptotically equivalent to $T/2-1$) as a conservative estimate of the effective independent sample size in place of n.

11. STOCK RETURNS

Figure 11 show monthly logarithmic percent returns in excess of the risk-free 1-month Treasury bill rate on the Center for Research on Security Prices (CRSP) Value-Weighted stock price index, with dividends, for Jan. 1950 to Dec. 2016.[16] While there were extreme returns in both directions, October 1987 and October 2008 had particularly large negative returns.

---

[15] The characteristic exponent α could be estimated from the lag 2 differences along with c_2. However, the lag 2 first differences have an effective independent sample size of approximately $T/3$, whereas the first differences themselves have an effective independent sample size of approximately $T/2$, and therefore provide a superior estimate. It is therefore preferable to first estimate α from the first differences along with c_1, and then to estimate c_2 with α constrained to this $\hat{\alpha}$.

[16] The CRSP monthly index represents the last trading day of each month, so that no spurious serial correlation of returns is introduced, as would be the case if it were the monthly average of daily values.

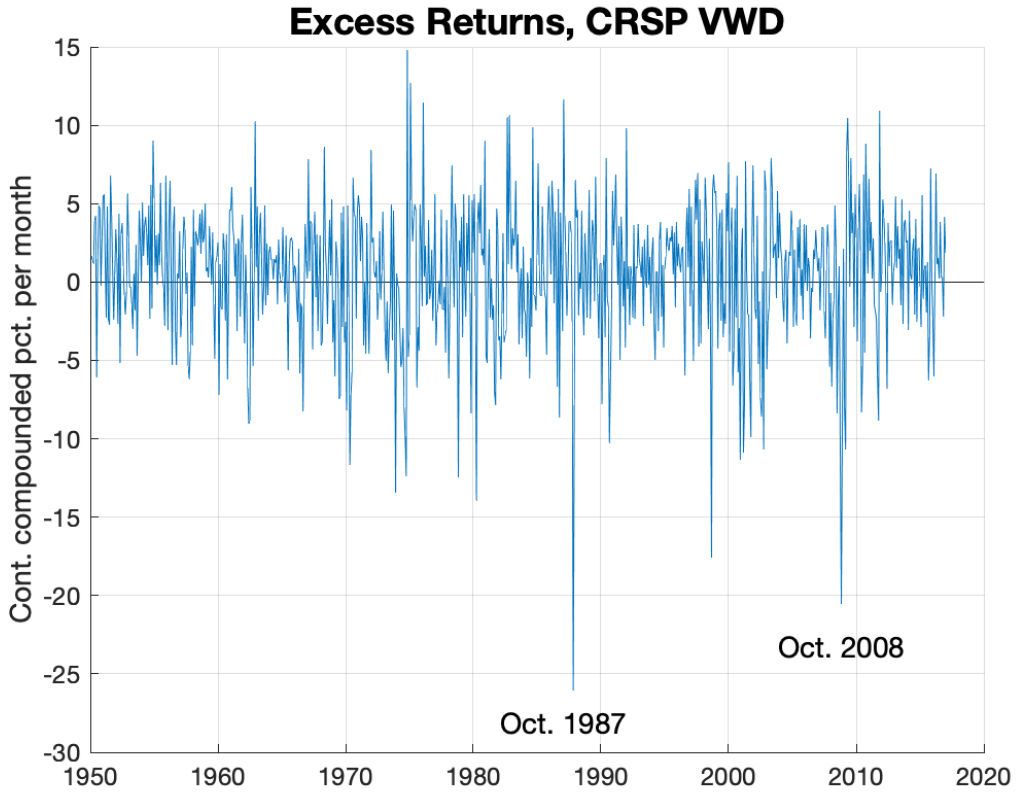

Figure 11

Monthly logarithmic percent returns on the CRSP Value-Weighted stock price index, with dividends, in excess of the 1-month Treasury bill return, for Jan. 1950 to Dec. 2016.

The benchmark martingale difference model of stock returns assumes that the equity premium in excess stock returns is a constant. However, it is conceivable that the observed equity premium in fact varies from decade to decade due to demographic changes, changes in investor attitudes toward risk, and/or changes in the tax code. The LLM should be capable of picking up such changes.[17]

The first data column of Table 7 shows the kurtosis test for normality of stock returns, along with "short-cut" estimates of α, c_ε, and c_η, and a grid search estimate of β_ε

[17] Most financial returns in fact exhibit volatility clustering and therefore are not independent. While it is feasible to model returns as conditionally stable with GARCH-like time-varying scale (see McCulloch 1985, Oh 1994, McCulloch and Bidarkota 1998), the present paper assumes for simplicity that errors are homoskedastic and serially independent.

using N = 10,000 particles. I.i.d. normality is overwhelmingly rejected, with a 2-tailed p-value of 1.16 x 10^{-8}. The estimated stable characteristic exponent α is 1.88. The estimated signal/noise scale ratio c_η / c_ε is 0.302, which is surprisingly high and indicates a much stronger departure from pure martingale-difference behavior than was anticipated. This finding warrants further investigation that goes beyond the scope of the present paper.

Table 7

Normality tests and parameter estimates
for stock returns, CPI inflation, and Bitcoin returns.

	Stocks	Inflation	Bitcoin
T	804	709	101
K	4.36	4.99	6.88
Z_K	5.59	7.66	5.63
p_K	1.16 x 10^{-8}	1.90 x 10^{-14}	1.88 x 10^{-8}
α	1.88	1.77	1.71
c_ε	2.55	0.737	13.0
c_η	0.77	0.749	9.7
c_η / c_ε	0.302	1.02	0.75
β_ε	−0.62	+0.57	+1.00
ESS_{min}	16.0	39.7	325.5

Note: T is the number of monthly observations. Kurtosis K is computed from first differences per text. Standardized kurtosis Z_K is computed with $n = T/2$ per text. Probability value p_K is 2-tailed test for normality. Parameters α, c_ε, and c_η are estimated by "shortcut method" per text. Parameter β_ε is estimated by grid search of Whisker particle filter likelihood with N = 10,000 particles and step size of 0.01. ESS_{min} is the minimum Effective Sample Size for the whisker particle filter over t = 1, ... T, at the estimated value of β_ε.

Figure 12 plots the log likelihood for the stock returns as a function of β_ε, as computed using the whisker particle filter with N = 10,000 particles, in steps of .01 from −1 to +1, and conditional on the "short-cut" estimates of α, c_ε, and c_η. The short vertical green line indicates the likelihood-maximizing value of –0.62. The horizontal red line is 1.92 = 3.84/2 below the likelihood maximum, so that β_ε values whose log likelihood lie below this line may be rejected at the .05 confidence level by the χ^2 Likelihood Ratio test

with one degree of freedom. It may be seen that almost all positive values may be rejected, but that symmetry is just on the borderline of acceptability. Almost all negative values are acceptable. [18] On close examination, the blue line is "jittery" because of serrations in the simulated log likelihood induced by the particle filter, yet smooth enough to ascertain a global maximum near –0.62.

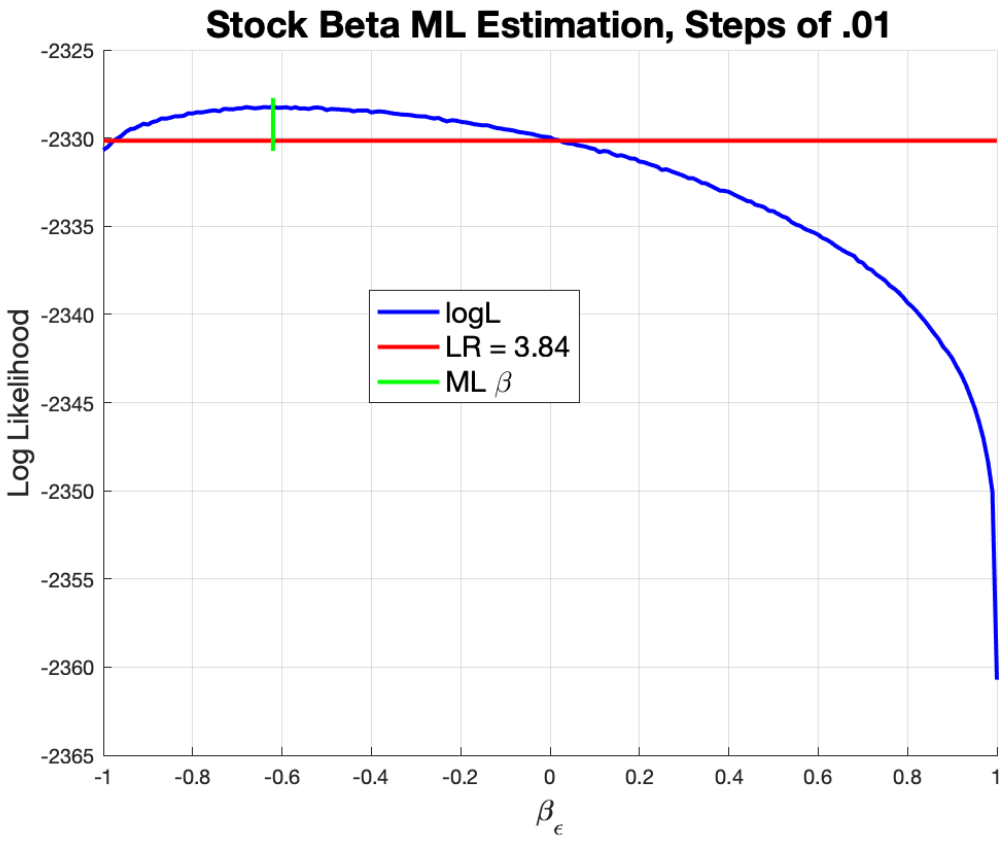

Figure 12

Maximum likelihood search for stock noise skewness parameter β_ε. Blue line is log likelihood, conditional on "short-cut" estimates of other stock stable parameters. Vertical green line indicates likelihood-maximizing value of −0.62. Horizontal red line is 1.92 below maximized log likelihood, so that LR statistic is 3.84.

[18] Moran (1971) notes that the LR test does not have its customary asymptotic normality when the null hypothesis is at the boundary of the parameter space, as is the case for $\beta_0 = \pm 1$, so that caution must be exercised in making claims about these cases. DuMouchel (1973) observes that when the null is $\beta_0 = \pm 1$ or $\alpha_0 = 2$, MLE is in fact super-efficient, in that it converges on the true value at a rate faster than $n^{-1/2}$. McCulloch (1997) finds that the LR statistic is informative for the boundary null $\alpha_0 = 2$, but that the customary chi-square critical values are too conservative.

Figure 13 plots the median, 50% and 95% credible intervals for the expected logarithmic stock excess return, as simulated with the whisker filter with $N = 10{,}000$ particles, $\theta_A = 0.25$, and equilibration. In order to concentrate on the financial crisis of 2008, only Feb. 2004 (month 650) through June 2012 (month 750) are plotted. Months 650 through 700 are typical of the portion of the series that is not plotted – the filter median is more often positive than not, but zero ordinarily lies within the 95% credible interval. This failure to find significantly positive expected log returns is apparently due to the surprisingly high estimated signal/noise ratio, which gives the filter a very short memory. This feature of the data deserves further study. During and shortly after the October 2008 crash (in month 706), the expected log excess return was actually significantly negative, again perhaps due to the short memory of the process.

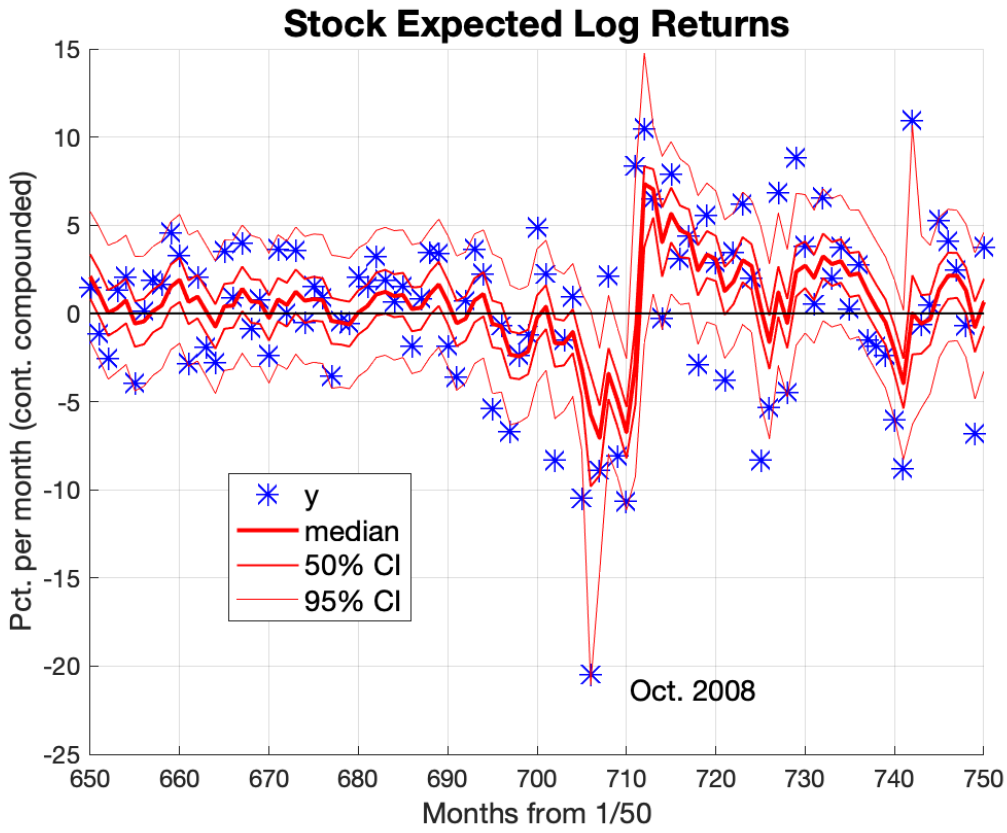

Figure 13

Whisker filter distribution for expected log excess stock returns for Feb. 2004 (month 650) through June 2012 (month 750). Blue stars are observed excess logarithmic returns in percent per month. Heavy red line is posterior median. Thinner lines are 50% and 95% credible intervals.

The October 1987 crash (not plotted) was even larger than the October 2008 crash, and did create a similarly strong spike in the lower 95% CI bound. However, because it was not reinforced with adjacent negative returns, it only barely pulled the 50% CI into the negative range, and hence, unlike the 2008 crash, did not result in a significantly negative expected log excess return at the 95% credible level.

Even though ESS_{min} was only 16.0 despite using $N = 10{,}000$ particles, the extreme particles always had probability less than 0.05, so that the 95% CI bounds were never undefined. This was possible because the addition of "whisker" particles ensures that there will be relatively weak particles at each end of the distribution.

12. INFLATION

In the case of inflation, the Local Level Model corresponds to the Adaptive Expectations (AE) hypothesis of Cagan (1956), as established by Muth (1960). It is now generally recognized that AE is overly simplistic as a model of inflationary expectations: Inflation has richer dynamics than implied by the LLM, and perhaps other observable variables such as the money supply, unemployment, and/or the Fed's interest rate targets are empirically significant marginal predictors of inflation. It is anticipated that a future multivariate extension of the present study will enable the incorporation such variables. Nevertheless, the LLM remains a useful univariate first approximation to the inflation process.

Annualized percent inflation was computed from the logarithm of the seasonally adjusted, chain-type Personal Consumption Expenditures (PCE) price index from Feb. 1959 to Feb. 2018 (709 monthly observations).[19] The kurtosis test for normality reported in the second data column of Table 7 rejects the null of i.i.d. normality with a 2-sided *p*-value of 1.90×10^{-14}. The "short-cut" estimate of α is 1.77. The estimated signal/noise scale ratio of 1.02 is surprisingly high, as it again implies a very short memory to the process. The estimated noise skewness parameter β_ε is +0.57, indicating that unusually high inflation (relative to the expectation) is more likely than comparably low inflation, even though the inflation rates are computed logarithmically.

Figure 14 plots the median and 50% and 95% credible intervals for the whisker filter, using $N = 10{,}000$ particles, along with the observed annualized monthly PCE

[19] The seasonal adjustments in the official data are themselves a set of 12 constrained time-varying unobserved state variables that ideally should be estimated simultaneously with the seasonally adjusted inflation trend x_t as in McCulloch (2005). For the sake of simplicity, we use the official seasonal adjustments, even though these are computed after the fact using subsequent data.

inflation (blue stars). In order to focus in on the financial crisis, the figure is restricted to Jan. 2004 (month 540) through May 2012 (month 640). November 2008 (month 598) stands out as an extremely negative observation.

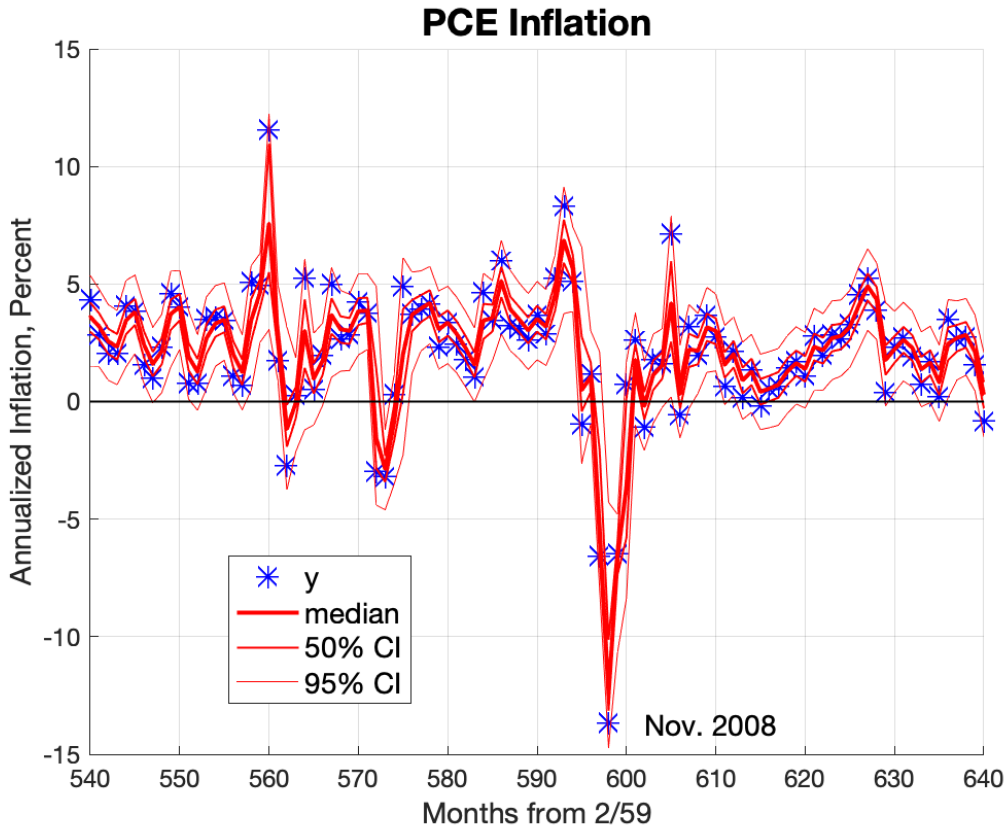

Figure 14

Whisker filter estimates of expected annualized percent PCE inflation using parameter estimates of Table 7 with $N = 10{,}000$ particles. Detail shown is for Jan. 2004 (month 540) through May 2012 (month 640). Red lines indicate the simulated filter median and 50% and 95% credible intervals. Blue stars are observed annualized monthly inflation.

Despite the short memory implied by the high estimated signal/noise ratio, expected inflation is ordinarily significantly positive up until August of 2008. After the extremely negative observation of November 2008, expected inflation is significantly less than −4% per annum, and remains there through December 2008. Although in my view the extreme reaction of the Fed at the time was unwarranted, one can empathize that a sense of urgency might have prevailed at the time.

A multivariate model of inflation that is richer than the LLM, e.g. in which x_t depends on lagged inflation with time-varying parameters as in McCulloch (2005), and perhaps also is predicted by other variables, might tell a different story during this important episode. Such a model must await a multivariate extension of the present paper.

13. BITCOIN RETURNS

Percent logarithmic returns were computed from month-end Bitcoin prices obtained from coindesk.com, for December 2010 through May 2019, yielding 101 returns. The prices used were weekly averages for the week beginning with the last Monday of each month. These weekly averages largely iron out transitory fluctuations that may be an artifact of any thinness of the market. However, since they are separated by more than one week, no spurious serial correlation in the monthly returns is induced by the averaging.[20]

Parameters for the returns are tabulated in the last column of Table 7. I.i.d. normality may be rejected with a 2-tailed p-value of 1.88×10^{-8}. The short-cut estimate of α is 1.71, the lowest of the three series. The signal/noise scale ratio is 0.75, which again is unexpectedly high, implying a very short memory. The observation error skewness estimate is +1.00.

Figure 15 plots the Bitcoin returns (blue stars), along with the Whisker filter median and 50% and 95% credible intervals, using $N = 10{,}000$ particles. Although the median expected log return is ordinarily positive, it is significantly positive only briefly, in early 2013 (near month 27) and late 2017 (near month 83), and in fact is briefly significantly negative in late 2011 (near month 9).[21]

Takeaway: Despite its extraordinarily high average performance, Bitcoin is very risky, with an only questionably positive expected log return.

---

[20] The data were first collected from coindesk.com on 2/27/18, and updated 2/6/19 and 5/27/19. The updates revealed that some of the earlier data had been slightly revised, but this effect was negligible. The data begins with Dec. 2010 because prior to that date the price was under $1.00, and hence returns would have been very sensitive to rounding error. Furthermore, the earlier market was likely relatively thin.

[21] Figure 15 suggests that there is pronounced volatility clustering in additional to conditional leptokurtosis, that could be modeled wither with a GARCH-type model or an additional stochastic volatility state variable. The present paper makes no attempt to model this additional complication.

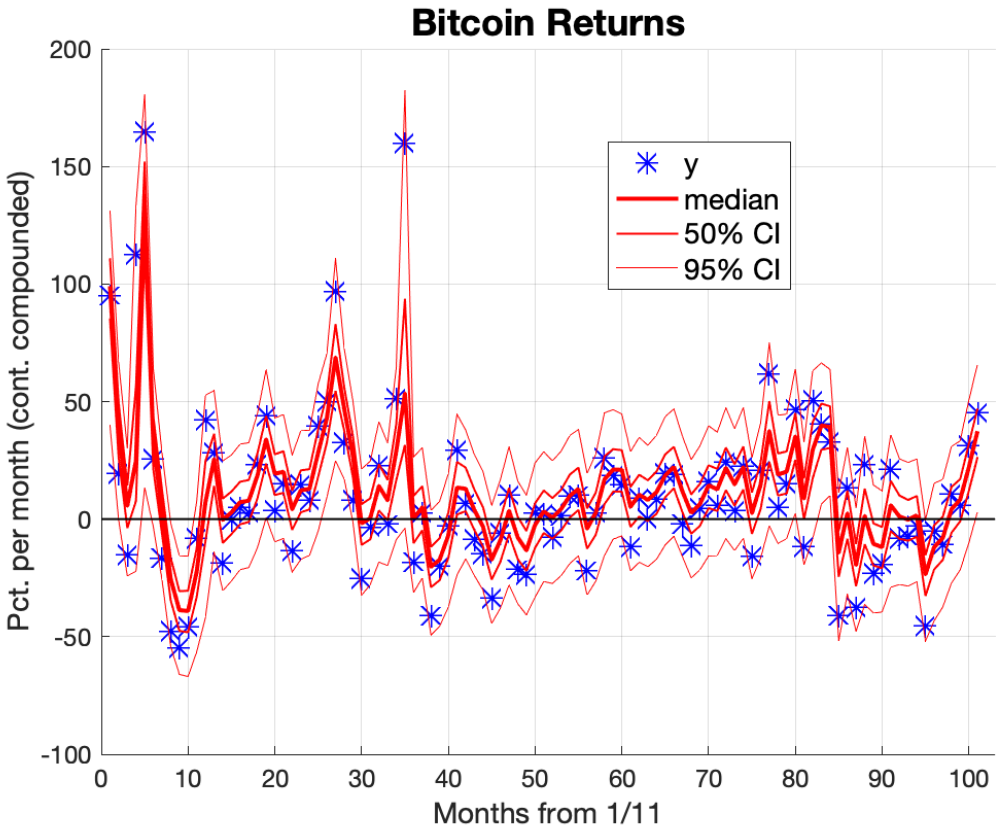

Figure 15

Monthly percent logarithmic Bitcoin returns (blue stars), together with Whisker filter median and 50% and 95% credible intervals.

14. PARTICLE SMOOTHING

The Whisker particle filter, which simulates the backward-looking "filter" density $p(x_t|\mathbf{Y}_t)$, may easily be extended to simulate the backward- and forward-looking "smoother" density $p(x_t|\mathbf{Y}_T)$, where $\mathbf{Y}_T = (y_1, y_2, \ldots y_T)$, by means of the two-filter smoother introduced by Kitagawa (1987) and reported by Harvey (1990, 162 ff.). Setting $\mathbf{Y}_{t:T} = (y_t, y_{t+1}, \ldots y_T)$, $p(x_t \mid \mathbf{Y}_{t:T})$ is the backward filter, obtained by running the filter backward from time T to time t, and $p(x_t \mid \mathbf{Y}_{t+1:T})$ is the backward predictive density for x_t given y_{t+1} through y_T, as computed in the course of the backward filter. The smoother may then be computed and simulated in either of two ways.

The first, or "S_1," method computes the smoother at time t from the time t forward filter and time t+1 backward filter using

$$\begin{aligned} \mathrm{p}(x_t|\mathbf{Y}_T) &\propto p(x_t|\mathbf{Y}_t)p(x_t|\mathbf{Y}_{t+1:T}) \\ &= p(x_t|\mathbf{Y}_t)\int \mathrm{g}(x_{t+1}-x_t)p(x_{t+1}|\mathbf{Y}_{t+1:T})dx_{t+1}. \end{aligned} \tag{17}$$

This could be simulated, albeit with $O(N^2)$ operations for each time t, with

$$\begin{aligned} p_{t,i}^{\mathrm{S}_1} &\propto p_{t,i}^{\mathrm{F}} \sum_{j=1}^{N} \mathrm{g}\left(x_{t+1:T,j}^{\mathrm{R}} - x_{t,i}^{\mathrm{F}}\right) p_{t+1:T,j}^{\mathrm{R}} , \\ x_{t,i}^{\mathrm{S}1} &= x_{t,i}^{\mathrm{F}} , \end{aligned} \tag{18}$$

where $p_{t+1:T,j}^{\mathrm{R}}$ etc. pertains to the resampled reverse filter for x_{t+1} based on y_{t+1} through y_T. However, in the spirit of particle filtering, the expense of this convolution can be reduced to $O(N)$ by replacing the sum with one randomly selected element, as follows:

$$p_{t,i}^{\mathrm{S}_1} \propto p_{t,i}^{\mathrm{F}} \mathrm{g}\left(x_{t,i}^{\mathrm{F}} - x_{t+1:T,j(i)}^{\mathrm{R}}\right) p_{t+1:T,j(i)}^{\mathrm{R}} ,$$

where $j(i)$ is a random permutation of the first N integers. It is preferable to use the resampled backward filter particles rather than the raw backward filter particles for this purpose, as these will place more appropriate weights on the transition densities. Also, sampling without replacement by means of the permutation $j(i)$ rather than with replacement ensures that each resampled backward filter particle will be used exactly once for some i.[22]

The second, or "S_2," method instead computes the smoother at time t from the time t backward and time $t-1$ forward filter:

$$\begin{aligned} \mathrm{p}(x_t|\mathbf{Y}_T) &\propto p(x_t|\mathbf{Y}_{t:T})p(x_t|\mathbf{Y}_{t-1}) \\ &= p(x_t|\mathbf{Y}_{t:T})\int \mathrm{g}(x_t - x_{t-1})p(x_{t-1}|\mathbf{Y}_{t-1})dx_{t-1}, \end{aligned} \tag{19}$$

and simulates this with

$$\begin{aligned} p_{t,i}^{\mathrm{S}_2} &\propto p_{t:T,i}^{\mathrm{F}} \mathrm{g}\left(x_{t:T,i}^{\mathrm{F}} - x_{t-1,J(i)}^{\mathrm{R}}\right) p_{t-1,J(i)}^{\mathrm{R}} , \\ x_{t,i}^{\mathrm{S}_2} &= x_{t:T,i}^{\mathrm{F}} , \end{aligned}$$

where $J(i)$ is a different random permutation of the first N integers.

Although (17) and (19) are mathematically equivalent, the S_1 and S_2 smoothers differ because they use different sets of ordinates which may result in different Effective Sample Sizes, $\mathrm{ESS}_t^{\mathrm{S}_1}$ and $\mathrm{ESS}_t^{\mathrm{S}_2}$. If η_{t+1} is unusually large, the forward filter and hence S_1 will ordinarily be the more accurate estimate of x_t. and will have the higher ESS. Or, if η_t is unusually large, the reverse filter and hence S_2 will be better and have the higher ESS.

---

[22] See Fearnhead et al. (2010). Kantas et al. (2015) note that J. Olsson and J Westerborn have proposed instead drawing K elements from the sum in (18), where $1 < K < N$ for some K that does not increase with N. However, given that the filter has already proxied each propagation convolution with a single, randomly chosen element, it is not clear that this would be worth the expense. If K times more computation is warranted, N should simply be replaced with NK.

The detailed information in the better estimate may be automatically exploited simply by merging the two sets of particles, weighted by their respective ESS:

$$< x_{t,i}^{S}, p_{t,i}^{S} > = < x_{t,i}^{S_1}, p_{t,i}^{S_1} \frac{\mathrm{ESS}_t^{S_1}}{\mathrm{ESS}_t^{S_1}+\mathrm{ESS}_t^{S_2}} > \cup < x_{t,i}^{S_2}, p_{t,i}^{S_2} \frac{\mathrm{ESS}_t^{S_2}}{\mathrm{ESS}_t^{S_1}+\mathrm{ESS}_t^{S_2}} > .$$

This gives the smoother $2N$ particles rather than N, but this is not a great burden, since most of the calculations, including any hyperparameter estimation, have already been done with N particles. In any event, the smoother is often of primary importance and hence may merit the additional detail.

The particle smoother proposed here has not yet been implemented, and is left for future research.

15. TIME-VARYING PARAMETER (TVP) SEQUEL

The Local Level Model can be generalized to a more flexible Time Varying Parameter (TVP) model in which an observed variable y_t obeys a linear regression, each of whose regression coefficients follows a random walk. When the errors are Gaussian and the covariance matrix of the regime shifts is constrained in a certain way, the very useful Recursive Least Squares model results, as proposed by Ljung (1992), Sargent (1999), Evans and Honkapohja (2001), and McCulloch (2005).

One purpose of this paper has been to be a prequel to a subsequent study that will develop such a TVP model with stable observation errors and regime shifts, as follows:

$$y_t = \sum_{j=1}^{k} b_{t,j} x_{t,j} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{S}(\alpha, \beta_\varepsilon, c_\varepsilon, 0),$$

$$\mathbf{b}_t = \mathbf{b}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \text{MV Elliptical S}(\alpha, \mathbf{C}_{\boldsymbol{\eta},t}, \mathbf{0}),$$

where the $x_{t,j}$ are exogenous observed regressors, $\mathbf{b}_t = (b_{t,1}, \ldots b_{t,k})'$ is a vector containing the regression coefficients at time t, and $\boldsymbol{\eta}_t = (\eta_{t,1}, \ldots \eta_{t,k})'$ is a vector containing the shocks to each regression coefficients.

Multivariate stable distributions can be very complicated, but assuming that the contours of the joint distribution are elliptical greatly simplifies the possibilities. In this case, the scales and covariations are determined by a $k \times k$ *coscale* matrix $\mathbf{C}$ that determines the shape of the ellipses much like the covariance matrix in the Gaussian special case.[23] In the spirit of Recursive Least Squares, it will be assumed that $\mathbf{C}_{\boldsymbol{\eta},t}$ is proportional to the uncertainty in the estimate of $\mathbf{b}_{t-1}$, as measured by the covariance of the time $t-1$ particle filter. Elliptical stable distributions are necessarily symmetrical, but this does not prevent the observation errors from being skewed.

[23] See Nolan (2006).

In his book *Fractals*, Benoît Mandelbrot (1977: 131-42) refers to such a multivariate random walk with elliptical stable disturbances as a "Levy Flight." A simulation of a bivariate Levy Flight with $\mathbf{C} = \mathbf{I}$, $\alpha = 1.7$, and $T = 1000$ is depicted in Figure 16. Mandelbrot erroneously states that such a process would have step lengths that are stably distributed. In fact, it must be the marginals in each direction that are stably distributed in order for it to be a self-similar "fractal," in his terminology.

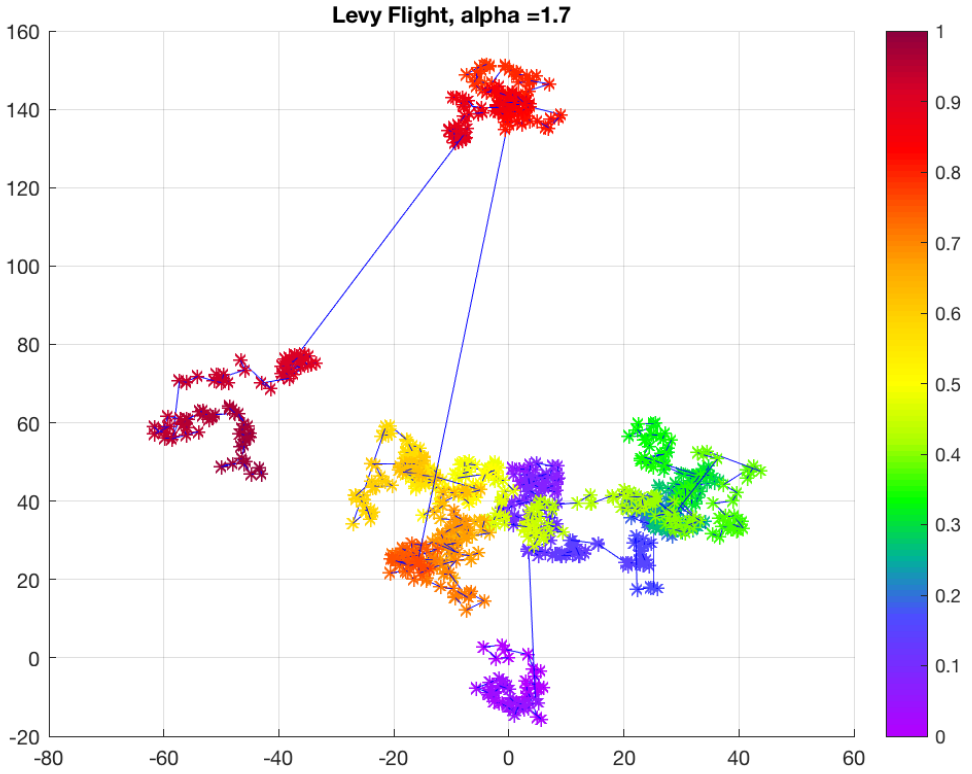

Figure 16

Isotropic Levy Flight of length 1000 with $\mathbf{C} = \mathbf{I}$ and $\alpha = 1.7$. The color axis indicates time, scaled to [0, 1].

16. CONCLUSION

A particle filter is used to estimate the unobserved state variable in the Local Level Model with infinite variance stable disturbances. The proposed Whisker Filter method, which adds minimal-weight particles outside the range of the received particle filter when resampling, robustly detects large regime shifts without being unduly distracted by extreme observation errors or requiring an excessive number of particles. While all methods considered consistently estimate the filter distribution, the Whisker Filter gives the best results in terms of Mean Absolute Error with a moderate number of particles. The proposed method is suitable for use with other heavy-tailed distributions, such as Student's t.

The paper employs rank-stratified sampling in the Initialization, Resampling and Propagation steps in order to make the method as deterministic as possible. Although the Propagation step requires some randomness, this is minimized by using a random permutation of a stratified sample from the signal distribution. A Mixture Kalman filter as proposed by Chen and Liu (2000), using the skew-stable Poisson series simulation of Lemke, Riabiz and Godsill (2015), was implemented, but was much slower, often failed, and had no clear advantage over the Whisker Filter in terms of MAE, even when modified by including more high-variance components as governed by a Beta distribution.

Excess stock returns, inflation, and Bitcoin returns overwhelmingly reject i.i.d. normality (stable $\alpha = 2.00$), and yield stable α estimates of 1.88, 1.77, and 1.71, respectively. The estimated signal/noise ratios were unexpectedly high in all three cases.

It is anticipated that the methods developed in the present paper will be adapted in future research to smoothing, as well as to richer, multivariate state-space models.

Appendix: The Mixture Kalman Filter

A Gaussian mixture representation of a distribution H(x), h(x) is a set of N triples $<\mu_i, v_i, p_i> = \{(\mu_i, v_i, p_i), i = 1, \ldots N\}$, each giving the mean, variance, and probability weight of a Gaussian distribution, such that

$$\mathrm{H}(x) \approx \sum_{i=1}^{N} p_i \mathrm{N}(x;\ \mu_i, v_i),$$
$$\mathrm{h}(x) \approx \sum_{i=1}^{N} p_i \mathrm{n}(x;\ \mu_i, v_i),$$

where N(x; μ, v) and n(x; μ, v) are the normal CDF and PDF with mean μ and variance v.

According to a 1955 theorem of S. Bochner (see Samorodnitsky and Taqqu 1994, theorem 1.3.1 with their $\alpha' = 2$), if

$$A \sim \mathrm{S}\left(\frac{\alpha}{2}, 1, \cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}, 0\right),$$

and, independently,

$$W \sim \mathrm{N}(0, 1),$$

then

$$X = (2A)^{1/2} W \sim \mathrm{S}(\alpha, 0, 1, 0).$$

It follows that a stratified N-component Gaussian mixture representation of the symmetric stable signal distribution $\mathrm{g}(\eta) = \mathrm{s}_{\alpha,0,c_\eta,0}(\eta)$ may be obtained by setting

$$\mu_i^\eta = 0,$$
$$v_i^\eta = 2\, c_\eta^2\, \mathrm{S}^{-1}\left(\frac{i-0.5}{N}; \frac{\alpha}{2}, 1, \cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}, 0\right),$$
$$p_i^\eta = 1/N.$$

Unfortunately, the Bochner theorem can only be used to generate *symmetric* stable random variables as mixtures of normals. However, following LRG (2015) with some slight rearrangement, an approximate N-component Gaussian mixture representation of $\mathrm{f}(\varepsilon) = \mathrm{s}_{\alpha,\beta_\varepsilon,c_\varepsilon,0}(\varepsilon)$ may be obtained, for $\alpha > 1$, by setting

$$\mu_i^\varepsilon = c_\varepsilon\, \mu_W \left(m_i + R_{i,1}\right),$$
$$v_i^\varepsilon = c_\varepsilon^2\, \sigma_{\mathrm{W}}^2 \left(s_i + R_{i,2}\right),$$
$$p_i^\varepsilon = 1/N,$$

where

$$m_i = \sum_{j=1}^{M_i} \Gamma_{i,j}^{-1/\alpha},$$
$$s_i = \sum_{j=1}^{M_i} \Gamma_{i,j}^{-2/\alpha},$$

$\Gamma_{i,j}$, $j = 1, \ldots M_i$ is a truncated Poisson series (the same for s_i and m_i, but drawn independently for each i), M_i is the largest j for which $\Gamma_{i,j}$ is less than some threshold g,

$$\mathbf{R}_i = \begin{pmatrix} R_{i,1} \\ R_{i,2} \end{pmatrix} \sim N(\boldsymbol{\mu}_\mathbf{R}, \boldsymbol{\Sigma}_\mathbf{R}) \text{ (independently of } \Gamma_{i,j}),$$

$$\boldsymbol{\mu}_\mathbf{R} = \begin{pmatrix} \frac{\alpha}{1-\alpha} g^{\frac{\alpha-1}{\alpha}} \\ \frac{\alpha}{2-\alpha} g^{\frac{\alpha-2}{\alpha}} \end{pmatrix},$$

$$\mathbf{\Sigma_R} = \begin{pmatrix} \frac{\alpha}{2-\alpha} g^{\frac{\alpha-2}{\alpha}} & \frac{\alpha}{3-\alpha} g^{\frac{\alpha-3}{\alpha}} \\ \frac{\alpha}{3-\alpha} g^{\frac{\alpha-3}{\alpha}} & \frac{\alpha}{4-\alpha} g^{\frac{\alpha-4}{\alpha}} \end{pmatrix},$$

and for $|\beta_\varepsilon| < 1$,

$$\sigma_W = \left(\frac{C_\alpha}{\mathrm{E}|W^*|^\alpha}\right)^{1/\alpha},$$
$$\mu_W = \sigma_W \mu^*,$$

where, $\Gamma(x)$ here being the gamma function,

$$C_\alpha = \frac{1-\alpha}{\Gamma(2-\alpha)\cos\left(\frac{\pi\alpha}{2}\right)},$$

and μ^* is implicitly determined from β_ε by numerically inverting

$$\beta_\varepsilon = \frac{\mathrm{E}\left(|W^*|^\alpha \mathrm{sign}(W^*)\right)}{\mathrm{E}(|W^*|^\alpha)},$$

where $W^* \sim \mathrm{N}(\mu^*, 1)$. We found that the above LRG approximation with $N = 1000$ and their suggested value of $g = 100$ gives a good visual match to $\mathrm{S}(\alpha, \beta_\varepsilon, c_\varepsilon, 0)$, with assorted parameter values. (Our g is LRG's c, but we reserve c for the stable scale parameter.)

For $|\beta_\varepsilon| = 1$, we instead must use the limiting values

$$\mu_W = \mathrm{sign}(\beta_\epsilon)\, {C_\alpha}^{1/\alpha},$$
$$\sigma_W = 0,$$

which effectively turn the Gaussian mixture representation for ε into an unstratified particle representation.

Our application of the Mixture Kalman Filter then proceeds as follows:

Preliminary calculations: Compute $\mu_W, \sigma_W,$ and v_i^η as above.

Initialization (I): Draw $\mu_{1,i}^\varepsilon$ and $v_{1,i}^\varepsilon$ by the LRG method as above and, using a uniform prior for x_1, set $t = 1$ and the initial filter (F) components to

$$\mu_{1,i}^{\mathrm{F}} = y_1 - \mu_{1,i}^\varepsilon,$$
$$v_{1,i}^{\mathrm{F}} = v_{1,i}^\varepsilon,$$
$$p_{1,i}^{\mathrm{F}} = 1/N.$$

The minus sign on $\mu_{1,i}^\varepsilon$ is required because the skewness of the likelihood $\mathrm{f}(x_t - y_t)$ as a function of x_t is opposite that of the measurement error density $\mathrm{f}(y_t - x_t)$ as a function of y_t.

Resampling (R): Draw an equal-weighted stratified sample $\mu_{t,i}^{\mathrm{R}}, v_{t,i}^{\mathrm{R}}$ from the time t filter components, without sorting or interpolating, and set

$$p_{t,i}^{\mathrm{R}} = 1/N.$$

Propagation (P): Draw a random permutation $\mathrm{J}_t(i)$ of the first N integers and simulate the predictive (P) density for y_{t+1} conditional on observations through y_t with components

$$\mu_{t,i}^{\mathrm{P}} = \mu_{t,i}^{\mathrm{R}},$$

$$v_{t,i}^{\mathrm{P}} = v_{t,i}^{\mathrm{R}} + v_{\mathrm{J}_t(i)}^{\eta},$$
$$p_{t,i}^{\mathrm{P}} = 1/N.$$

Updating (U): Draw a new set of observation error components $\mu_{t,i}^{\varepsilon}$ and $v_{t,i}^{\varepsilon}$ as above with $p_{t,i}^{\varepsilon} = 1/N$ and, for $|\beta_\varepsilon| \neq 1$, perform a Kalman filter update pairing each predictive density with its randomly determined observation density by setting

$$v_{t+1,i}^{\mathrm{F}} = 1/\left(1/v_{t,i}^{\mathrm{P}} + 1/v_{t,i}^{\varepsilon}\right),$$
$$\mu_{t+1,i}^{\mathrm{F}} = v_{t+1,i}^{\mathrm{F}}\left(\mu_{t,i}^{\mathrm{P}}/v_{t,i}^{\mathrm{P}} + \left(y_{t+1} - \mu_{t,i}^{\varepsilon}\right)/v_{t,i}^{\varepsilon}\right),$$
$$p_{t+1,i}^{\mathrm{F}} \propto p_{t,i}^{\mathrm{P}}\, p_{t,i}^{\varepsilon}\, \mathrm{n}\left(y_{t+1} - \mu_{t,i}^{\varepsilon};\, \mu_{t,i}^{\mathrm{P}},\, v_{t,i}^{\mathrm{P}} + v_{t,i}^{\varepsilon}\right),$$

and normalizing the weights to sum to unity. (The two probability weights in the last expression are computationally redundant, but are included here for the sake of generality.)

Iteration: Replace t by t+1 and repeat steps P and U until $t > T$.

In the limits $|\beta_\varepsilon| = 1$, the Update step becomes

$$\mu_{t+1,i}^{\mathrm{F}} = y_{t+1} - \mu_{t,i}^{\varepsilon},$$
$$v_{t+1,i}^{\mathrm{F}} = 0,$$
$$p_{t+1,i}^{\mathrm{F}} \propto p_{t,i}^{\mathrm{P}}\, p_{t,i}^{\varepsilon}\, \mathrm{n}\left(y_{t+1} - \mu_{t,i}^{\varepsilon};\, \mu_{t,i}^{\mathrm{P}},\, v_{t,i}^{\mathrm{P}}\right).$$

This effectively turns the filter distribution into a step function, as if we had used a particle filter. However, the outcome will be noisier than any of the particle filters considered because the likelihood of each predictive particle is being simulated randomly rather than computed exactly.

Our simulations show that the above standard Mixture Kalman Filter occasionally fails when the simulated data contains an extreme outlier that happens to be far outside the range of the variances in the simulated predictive and observation distributions. I am grateful to Simon Godsill for noting that these computational failures can be easily avoided by first computing the logarithm of $p_{t+1,i}^{\mathrm{F}}$ and then subtracting out the maximum of these logarithms before exponentiating and normalizing.

In order to avoid these perhaps ill-conditioned cases, we also modified the standard MKF to artificially add high-variance but low-probability signal components governed by a Beta(a, b) distribution with a = 1 and b = 0.5, to obtain the "Beta Mixture Kalman Filter" in the tables, as follows. Define

$$P_i^{\mathrm{Beta}} = \mathrm{B}_{1,0.5}^{-1}\left(\frac{i}{N}\right), \qquad i = 0, \ldots N,$$
$$p_i^{\mathrm{Beta}} = P_i^{\mathrm{Beta}} - P_{i-1}^{\mathrm{Beta}}, \qquad i = 1, \ldots N,$$
$$Q_i^{\mathrm{Beta}} = P_i^{\mathrm{Beta}} - \frac{p_i^{\mathrm{Beta}}}{2}, \qquad i = 1, \ldots N,$$

where $\mathrm{B}_{a,b}(x)$ is the CDF of the Beta(a, b) distribution. For the Propagation step, set

$$v_i^{\eta} = 2\, c_\eta^2\, \mathrm{S}^{-1}\left(Q_i^{\mathrm{Beta}}; \frac{\alpha}{2}, 1, \cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}, 0\right),$$
$$p_i^{\eta} = p_i^{\mathrm{Beta}}.$$

In the Resampling step, first sort the components of the time *t* Filter in increasing order by variance. Then draw rank-stratified $v_{t,i}^{\mathrm{R}}$ from the step function defined by the sorted filter variances at probabilities Q_i^{Beta}, without interpolating or extrapolating. Set the $\mu_{t,i}^{\mathrm{R}}$ equal to the corresponding filter means, and set

$$p_i^{\mathrm{R}} = p_i^{\mathrm{Beta}}.$$

Ideally, we would like to use a similar procedure to generate high-variance, low-probability observation error components in the Update step, but this is not possible with the LRG method. The Beta parameter choice $a = 1$ is natural, since we only want to dilate the distribution of variances at the upper end. The choice $b = 0.5$ is arbitrary, but robustly eliminates the underflow cases and substantially improves the MAE with only a small loss of ESS.

REFERENCES

Alspach, D.L., and H.W. Sorenson. (1972). Nonlinear Bayesian Estimation using Gaussian sum approximations. *IEEE Transactions on Automatic Control*, **AC-17**, 439–448.

Bidarkota, P.V., and J.H. McCulloch. (1998). "Optimal Univariate Inflation Forecasting with Symmetric Stable Shocks," *Journal of Applied Econometrics* **13**: 65–70.

Cagan, Phillip (1956). The monetary dynamics of hyperinflation. In Milton Friedman, ed., *Studies in the Quantity Theory of Money*. University of Chicago Press.

Carpenter, J., P. Clifford, and P. Fearnhead. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings – Radar and Sonar Navigation*. **146**: 2-7.

Carvalho, C.M., M.S. Johannes, H.F. Lopes, and N.G. Polson. (2010). Particle learning and smoothing. *Statistical Science* **25**: 88–106.

Chambers, J.M., C.L. Mallows, and B.W. Stuck (1976). A method for simulating stable random variables. *J. American Statistical Assn.* **71**: 340–4.

Chen, Rong, and Jun S. Liu (2000). Mixture Kalman filters. *J. R. Statist. Soc. B* **62** *Part 3*: 493-508.

DuMouchel, William H. (1973). On the asymptotic normality of the maximum-likelihood estimate when sampling from a stable distribution. *Annals of Statistics* **1**: 948–57.

DuMouchel, William H. (1975). Stable distributions in statistical inference: 2. Information from stably distributed samples. *J. American Statistical Assn.* **11**: 1019-31.

Evans, George W., and Seppo Honkapohja (2001). *Learning and Expectations in Macroeconomics*. Princeton University Press.

Fearnhead, P., D. Wyncoll, and J. Tawn (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika* **97**: 447–464.

Fama, Eugene F., and Merton H. Miller (1972). *The Theory of Finance*. Dryden Press.

Geweke, John (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**: 1317-39.

Gordon, N.J., D.J. Salmond, and A.F.M. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F* **140**: 107–13.

Harvey, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Hol, Jeroen D., Thomas B. Schön, and Fredrik Gustafsson (2006). On resampling algorithms for particle filters. *IEEE Nonlinear Statistical Signal Processing*. DOI: 10.1109/NSSPW.2006.4378824.

Kantas, Nikolas, Arnaud Doucet, Sumeetpal S. Singh, Jan Maciejowski and Nicolas Chopin (2015). On particle methods for parameter estimation in state-space models. *Statistical Science* **30** (no. 3): 328-351.

Kitagawa, Genshiro. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion*). J. American Statistical Assn*. **82**: 1032–1063.

Kitagawa, Genshiro. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models*. J. Computational and Graphical Statistics* 5: 1–25.

Kuhn, Thomas S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lemke, Tatjana, Marina Riabiz, and Simon J. Godsill (2015). Fully Bayesian inference for α-stable distributions using a Poisson series representation. *Digital Signal Processing* **47** (Oct. 22): 96-115.

Ljung, Lennart (1992). Applications to adaptive algorithms, in L. Ljung, G. Pflug, and H. Walk, eds., *Stochastic Approximations and Optimization of Random Systems*, Birkhäuser, pp. 95-113.

Lombardi, Marco J., and Simon J. Godsill (2006). On-line Bayesian estimation of signals in symmetric α-stable noise. *IEEE Transactions on Signal Processing* **54** (no. 2, Feb.): 775-779.

Malik, Sheheryar, and Michael K. Pitt (2011). Particle filters for continuous likelihood evaluation and maximization. *Journal of Econometrics* **165**: 190-209.

Mandelbrot, Benoît (1963). The variation of certain speculative prices. *J. Political Economy* **71**:421–40.

Mandelbrot, Benoît (1977). *Fractals: Form, Chance, and Dimension*. S.F. Freeman and Co.

McCulloch, J. Huston (1985). Interest-risk sensitive deposit insurance premia: Stable ACH estimates. *Journal of Banking and Finance* **9**: 137–156.

McCulloch, J. Huston (1996a). Financial applications of stable distributions. *Handbook of Statistics*, **14**, ch. 13.

McCulloch, J. Huston (1996b). On the parametrization of the afocal stable distributions. *Bulletin of the London Mathematical Society* **28**: 651-655.

McCulloch, J. Huston (1997). Measuring tail thickness to estimate the stable index α: A critique. *Journal of Business and Economic Statistics* **15**: 74-81.

McCulloch, J. Huston (2005). The Kalman foundations of adaptive least squares, with application to U.S. inflation. http://www.econ.ohio-state.edu/jhm/papers/KalmanAL.pdf.

McCulloch, J. Huston, and E. Richard Percy (2013). Extended Neyman goodness-of-fit tests, applied to competing heavy-tailed distributions. *Journal of Econometrics* **172** (2): 275-82.

Moran, P.A.P. (1971). The uniform consistency of maximum-likelihood estimators. *Proceedings of the Cambridge Philosophical Society* **17**: 435-439.

Muth, John F. (1960). Optimal properties of exponentially weighted forecasts. *J. American Statistical Assn.* **55** (290): 299-306.

Nolan, John P. (2006). Multivariate elliptically contoured stable distributions: theory and estimation. < http://academic2.american.edu/~jpnolan/stable/EllipticalStable.pdf>

Nolan, John P. (2008). Advances in nonlinear signal processing for heavy tailed noise. *Proceedings of the International Workshop in Applied Probability 2008*. <http://fs2.american.edu/jpnolan/www/stable/NolanIWAP2008.pdf>

Nolan, John P. (2009). STABLE 5.1 for Matlab. Washington, D.C.: RobustAnalysisInc. < http://www.robustanalysis.com/>

Oh, Chang Seok (Francis) (1994). Estimation of time varying term premia for U.S. interest rates using a state space-GARCH model. Ohio State University Economics Department Ph.D. dissertation.

Pitt, Michael K., and Neil Shephard (1999). Filtering via simulation: Auxiliary particle filters. *J. American Statistical Assn.* 94: 590–99.

Samorodnitsky, Gennady, and Murad S. Taqqu (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. New York: Chapman & Hall.

Sargent, Thomas J. (1999). *The Conquest of American Inflation*. Princeton University Press.

Singpurwalla, Nozer D., Nicholas G. Polson, and Refik Soyer (2017). From least squares to signal processing and particle filtering. *Technometrica*. DOI: 10.1080/00401706.2017.1341341.

Taleb, Naseem N. (2010). *The Black Swan: The Impact of the Highly Improbable*, Second edition. Random House.

Zolotarev, Vladimir M. (1986*). One-Dimensional Stable Laws*. Providence, R.I.: American Mathematical Society. Russian original 1983.