Econometric Reviews, 0(0):1–26, 2016

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2015.1092809

Taylor & Francis
Taylor & Francis Group

Particle Learning for Fat-Tailed Distributions

Hedibert F. Lopes[1] and Nicholas G. Polson[2]
[1]*INSPER Institute of Education and Research, São Paulo, Brazil*
[2]*University of Chicago Booth School of Business, Chicago, Illinois, USA*

It is well known that parameter estimates and forecasts are sensitive to assumptions about the tail behavior of the error distribution. In this article, we develop an approach to sequential inference that also simultaneously estimates the tail of the accompanying error distribution. Our simulation-based approach models errors with a t_ν-distribution and, as new data arrives, we sequentially compute the marginal posterior distribution of the tail thickness. Our method naturally incorporates fat-tailed error distributions and can be extended to other data features such as stochastic volatility. We show that the sequential Bayes factor provides an optimal test of fat-tails versus normality. We provide an empirical and theoretical analysis of the rate of learning of tail thickness under a default Jeffreys prior. We illustrate our sequential methodology on the British pound/U.S. dollar daily exchange rate data and on data from the 2008–2009 credit crisis using daily S&P500 returns. Our method naturally extends to multivariate and dynamic panel data.

Keywords Bayesian inference; Credit crisis; Dynamic panel data; Kullback–Leibler, MCMC.

JEL Classification C01; C11; C15; C16; C22; C58.

1. INTRODUCTION

Fat-tails are an important statistical property of time series prevalent in many fields, particularly economics and finance. Fat-tailed error distributions were initially introduced by Edgeworth (1888) and explored further by Jeffreys (1961) who once remarked that "... *all data are* t_4." They can be incorporated into dynamic models as latent variable scale mixtures of normals (Carlin et al., 1992). In this article, we develop a simulation-based sequential inference procedure for estimating the tail behavior of a time series using the t_ν-distribution. This family is attractive for this purpose due to its flexibility with normality ($\nu = \infty$) and Cauchy ($\nu = 1$) errors as special cases. Our method complements the existing literature by estimating the set of sequential posterior distributions $p(\nu|y^t)$ Q1

Address correspondence to Hedibert F. Lopes, INSPER Institute of Education and Research, Rua Quatá 300, Vila Olímpia, São Paulo, SP, 04546-042, Brazil; E-mail: hedibertFL@insper.edu.br
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lecr.

for data $y^t = (y_1, \ldots, y_t)$ and $t = 1, \ldots, T$, as opposed to Markov chain Monte Carlo (MCMC) which estimates v given the full data history $p(v|y^T)$ (see Geweke, 1993; Eraker, Jacquier, and Polson (JPR), 1998; Jacquier et al., 2004; Fonseca et al., 2008). In other words, our methodology allows the researcher to estimate and update the tail-thickness of the error distribution as new data arrives. Q2

The novel feature of our approach are the on-line estimates of the tail thickness of the error distribution using the marginal posterior distribution of the degrees of freedom parameter v. Being able to sequentially assess the degree of tail-heaviness is particularly important for dynamic portfolio and risk management strategies. For instance, $p(v|y^{t_0})$ and $p(\mu|y^{t_1})$ might resemble standard normal and t_4 distributions, respectively, for say t_0 much smaller than t_1, which in turn would potentially affect decision making at both time points.

Our method is based on particle learning (PL, see Carvalho et al., 2010, and Lopes et al., 2010). We analyze two cases in detail: In the first observations y_t follow the independent and identically distributed (iid) standard t_v-distribution, i.e., $y_t \sim t_v(0, 1)$ (iid-t case), and in the second observations follow a non-identically distributed stochastic volatility model with fat-tails (SV-t case), i.e., $y_t|h_t \sim t_v(0, \exp\{h_t\})$ are conditionally independent given the T-dimensional latent vector of log-volatilities $h^T = (h_1, \ldots, h_T)$, see JPR (2004) and Chib et al. (2002).

Our posterior distribution $p(v|y^t)$ on the tail thickness is sensitive to the choice of prior distribution, $p(v)$. We model the prior on the degrees of freedom v using a default Jeffreys prior (Fonseca et al., 2008) . In this setting, we show that the Jeffreys prior has desirable properties. Primarily, it reduces bias for estimating the tail thickness in small sized data sets. Moreover, it is well known that more data helps to discriminate similar error distributions. Hence *a priori* we know that we will need a larger dataset to discriminate a t_{20}-distribution from a normal distribution than a t_4-distribution from a normal. We develop a metric based on the asymptotic Kullback–Liebler rate of learning of tail thickness that can guide the amount of data required to discriminate two error distributions. Given the observed data, we then develop an empirical and theoretical analysis of the sequential Bayes factor which provides the optimal test of normality versus fat-tails in our sequential context.

Recent estimation approaches for fat-tails use approximate latent Gaussian models (McCausland, 2012). We use the traditional data augmentation with a vector of latent scale variables λ_t to avoid evaluating the likelihood (a T-dimensional integral). We develop a particle learning algorithm for sampling from the sequential set of joint posterior distributions $p(\lambda_t, v|y^t)$, for the iid-t case, and from $p(\lambda_t, h_t, v|y^t)$, for the SV-t case, for $t = 1, \ldots, T$. The marginal posterior distribution $p(v|y^t)$ provides estimates of the tail-thickness of the error distribution. The purpose for developing new estimation methods is apparent from a remark of Smith (2000) who warns that the likelihood for non-Gaussian models can have several local maxima, be very skewed, or have modes on the boundary of the parameter space, making estimating tail behavior a complex statistical problem.

The rest of the article is outlined as follows. Section 2 describes how to sequentially learn the tail of the t_v-distribution under iid-t and SV-t models. Section 3 discusses our particle learning implementation. We focus on using a default Jeffreys prior, showing that this has a number of desirable properties when learning the fat-tailed error distribution with finite samples. Section 4 provides an analysis of the sequential Bayes factor for testing normality versus fat-tails. Section 5 provides our empirical analysis and comparisons including an analysis of the British pound and U.S. dollar daily exchange rate and daily S&P500 returns from the credit crisis. Jacquier et al. (2004) apply MCMC methods to the SV-t model to daily exchange rate on the British pound versus the U.S. dollar, and we provide a sequential analysis for comparative purposes. Finally, Section 6 concludes.

2. T_v-DISTRIBUTED ERRORS

Consider data $y^t = (y_1, \ldots, y_t)$ arising from a fat-tailed t_v-distribution. The data are observed on-line, and we wish our estimation procedure to take this into account. Given a prior distribution $p(v)$, the aim is to compute a set of sequential marginal posterior distributions $p(v|y^t)$ which are given by Bayes rule

$$p(v|y^t) = \frac{p(y^t|v)p(v)}{\int p(y^t|v)p(v)dv}.$$

The marginal likelihood is given by $p(y^t|v)$. In an iid setting, this likelihood is simply $p(y^t|v) = \prod_{i=1}^t p(y_i|v)$, a product of marginals. In the SV-t setting, it is more complicated and requires integrating out the unobserved t-dimensional vector of log-volatilities $h^t = (h_1, \ldots, h_t)$, namely

$$p(y^t|v) = \int \prod_{i=1}^{t} p(y_i|h_i, v)p(h^t)dh^t,$$

where $p(y_i|h_i, v) \sim t_v(0, \exp\{h_i\})$. One advantage of particle methods is that this computation will naturally occur within the procedure. Our task is to provide sequential inference for the degrees of freedom or tail thickness parameter, v, via the set of marginal posterior distributions $p(v|y^t)$, for $t = 1, \ldots, T$. To do this, we will first use a standard data augmentation and then provide a sequential Monte Carlo algorithm to sample from $p(\lambda_t, v|y^t)$ which we now describe for the iid-t and SV-t models.

2.1. The iid-t Model

Consider iid observations y_t, for $t = 1, \ldots, T$, from a fat-tailed location-scale model

$$y_t = \mu + \sigma\eta_t \quad \text{where} \quad \eta_t \overset{\text{iid}}{\sim} t_v(0, 1).$$

Data augmentation uses a scale mixture of normals representation by writing η_t in the following two steps: i) $\eta_t = \sqrt{\lambda_t}\epsilon_t$ and ii) $\lambda_t \overset{\text{iid}}{\sim} IG(v/2, v/2)$, where IG denotes the inverse gamma distribution. The marginal data distribution, integrating out λ_t, is then the fat-tailed t_v-distribution $p(y_t|v, \mu, \sigma^2) \sim t_v(\mu, \sigma^2)$, where σ^2 can be interpreted as a scale parameter. This leads to a hierarchical specification of the model

$$y_t = \mu + \sigma\sqrt{\lambda_t}\epsilon_t \qquad \text{where } (\lambda_t|v) \overset{\text{iid}}{\sim} IG(v/2, v/2) \text{ and } \epsilon_t \overset{\text{iid}}{\sim} N(0, 1).$$

These specifications lead to a likelihood function $p(y|\mu, \sigma^2, v)$ of the form

$$p(y|\mu, \sigma^2, v) = \prod_{t=1}^{T} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v}\Gamma\left(\frac{v}{2}\right)} \left[1 + \frac{1}{v}\left(\frac{y_t - \mu}{\sigma}\right)^2\right]^{-\frac{v+1}{2}}$$

with marginal distribution $p(y_t|v) = \int p(y_t|v, \mu, \sigma^2)p(\mu, \sigma^2)d\mu d\sigma^2$. Fonseca et al. (2008) make the important observation that the marginal likelihood for v becomes unbounded as $v \to \infty$ and the maximum likelihood estimator is not well defined. This leads us to further develop an approach based on prior regularization, namely that the degree of freedom parameter v is random with a prior distribution $p(v)$ which we further discuss in Section 2.3.

Inference on the parameters (μ, σ^2) is not the focus of our study, and for simplicity we assume that either they are known quantities or taken from a standard diffuse prior, $p(\mu) \propto 1$, and inverse-gamma prior $\sigma^2 \sim IG(n_0/2, n_0\sigma_0^2/2)$ given hyper-parameters n_0 and σ_0^2. These parameters control, respectively, the shape and the location of the distribution.

2.2. The SV-t Model

A common model of time-varying volatility is the stochastic volatility model with fat-tails (SV-t) for returns and volatility (see Lopes and Polson, 2010a, for a recent review). The basic SV model is specified by evolution dynamic

$$\begin{aligned} y_t &= \exp\{h_t/2\}\epsilon_t \qquad \epsilon_t \overset{\text{iid}}{\sim} N(0, 1), \\ h_t &= \alpha + \beta h_{t-1} + \tau u_t \qquad u_t \overset{\text{iid}}{\sim} N(0, 1). \end{aligned}$$

The fat-tailed SV-t is obtained by adding an extra random scale parameter λ_t and, as described in the conditionally iid setting, is equivalent to assuming that $\epsilon_t \sim t_v(0, 1)$ (see, for example, JPR, 2004). The model can then be expressed as

$$\begin{aligned} y_t &= \exp\{h_t/2\}\sqrt{\lambda_t}\epsilon_t \qquad \epsilon_t \overset{\text{iid}}{\sim} N(0, 1) \\ h_t &= \alpha + \beta h_{t-1} + \tau u_t \qquad u_t \overset{\text{iid}}{\sim} N(0, 1) \\ \lambda_t &\overset{\text{iid}}{\sim} IG(v/2, v/2). \end{aligned}$$

The parameter β is the persistence of the volatility process and τ^2 the volatility of the log-volatility. Estimation of these parameters will be greatly affected by the fat-tail error assumptions which in turn will affect predicting price and volatility (see, for example, Jacquier and Polson, 2000).

To complete the model specification, we need a prior distribution for the parameters (α, β, τ^2) given v. For simplicity, we take a conditionally conjugate normal-inverse-gamma-type prior. Specifically, $(\alpha, \beta)|\tau^2 \sim N(b_0, \tau^2 B_0)$ and $\tau^2 \sim IG(c_0, d_0)$, for known hyper-parameters b_0, B_0, c_0, and d_0. Lack of prior information is achieved when $B_0^{-1} \approx 0$ and $c_0 \approx 0$. The marginal prior distribution for (α, β) is, therefore, a Student's t distribution. This conditionally conjugate structure will aid in the development of our particle learning algorithm as it leads to conditional sufficient statistics. Nonconjugate prior specifications can also be handled in our framework, see Lopes et al. (2010) for further discussion.

2.3. Priors on v

In the models considered so far, an important modeling assumption is the regularization penalty $p(v)$ on the tail thickness. A default Jeffreys-style prior was developed by Fonseca et al. (2008) and, we will see, with a number of desirable properties—particularly when learning a fat-tail (e.g., a t_4-distribution) from a finite dataset. The default Jeffreys prior for v takes the form

$$p(v) = \frac{1}{\sigma}\left(\frac{v}{v+3}\right)^{1/2}\left\{\psi'\left(\frac{v}{2}\right) - \psi'\left(\frac{v+1}{2}\right) - \frac{2(v+3)}{v(v+1)^2}\right\}^{1/2}, \tag{1}$$

where $\psi'(a) = d\{\psi(a)\}/da$ and $\psi(a) = d\{\log\Gamma(a)\}/da$ are the trigamma and digamma functions, respectively. The interesting feature of this prior is its behavior as v goes to infinity and it has polynomial tails of the form $p(v) \sim v^{-4}$. This is in contrast to commonly used priors such as Fernandez and Steel (1999) and Geweke (1993) who essentially specify priors with exponential tails of the form $v\exp\{-\lambda v\}$, for a subjectively chosen hyper-parameter, λ. In this case, the tail of the prior decays rather fast for large values of v and assessing the degree of tail thickness can require prohibitively large samples.

Table 1 compares Fonseca's robust prior to several exponential priors, including the exponential prior with mean 20 (rate $\lambda = 0.05$), which was advocated, for instance, by Geweke (1993). As it can be seen, despite its higher mass for heavy tailed distributions (small values of v), Fonseca's prior also places higher mass for normality (large values of v), when compared to the exponential with mean 20. The exponential priors, with high mean, essentially place zero mass on normality, whereas Fonseca's prior places approximately 0.01 probability on normality, which although being small can still be overwhelmed by an informative likelihood.

In our empirical analysis, we will show how this prior reduces bias in the posterior mean $E(v|y^t)$ and also how it helps discriminate a fat-tailed t_4-distribution from

TABLE 1
Fonseca's Prior and Geweke's Prior

v_0	10	20	50	150	190
$P_{\mathcal{E}}(v < v_0 \mid \lambda = 0.01)$	0.01	0.20	0.45	0.8959	0.98180
$P_{\mathcal{E}}(v < v_0 \mid \lambda = 0.05)$	0.36	0.61	0.91	0.9995	0.99997
$P_{\mathcal{E}}(v < v_0 \mid \lambda = 0.20)$	0.85	0.98	1.00	1.0000	1.00000
$P_J(v < v_0)$	0.85	0.93	0.98	0.9972	0.99952

normality. On the other hand, the flat uniform prior suffers from placing too much mass on high values of v—which are close to normality—making the inference problem harder for finite samples.

3. PARTICLE LEARNING FOR FAT-TAILS

We now provide a discussion of particle learning with particular reference to estimating fat-tails. Sequential Bayesian computation requires calculation of a set of posterior distributions $p(v|y^t)$, for $t = 1, \ldots, T$, where $y^t = (y_1, \ldots, y_t)$.

Loosely speaking, particle learning is a sequential Monte Carlo scheme that sequentially learns a low dimensional vector of essential states, usually comprising a combination of a few latent states of the state-space model along with conditional sufficient statistics for fixed, time-invariant parameters. Section 3.1 provides a thorough explanation and implementation of particle learning for the iid-t case. See Carvalho et al. (2010), Lopes et al. (2010), Lopes and Tsay (2011), and Lopes and Carvalho (2013) for extended discussion and several examples of PL in action.

Central to PL is the creation of a *essential state vector* Z_t to be tracked sequentially. We assume that this vector is conditionally sufficient for the parameter of interest; so that $p(v|Z_t)$ is either available in closed-form or can easily be sampled from. More precisely, given samples $\{Z_t^{(i)}\}_{i=1}^N \sim p(Z_t|y^t)$ and a Rao–Blackwellized identity, then a simple mixture approximation to the set of posteriors is given by

$$p^N(v|y^t) = \frac{1}{N}\sum_{i=1}^{N} p(v|Z_t^{(i)}).$$

Here the conditional posterior $p(v|Z_t^{(i)})$ will include the dependence on σ^2 for the iid-t case and (α, β, τ^2) and the latent volatilities $h^t = (h_1, \ldots, h_t)$ for the SV-t case through the essential state vector.

The task of sequential Bayesian computation is then equivalent to a filtering problem for the essential state vector, drawing $\{Z_t^{(i)}\}_{i=1}^N \sim p(Z_t|y^t)$ sequentially from the set of posteriors. To this end, PL exploits the following sequential decomposition of Bayes' rule

$$p(Z_{t+1}|y^{t+1}) = \int p(Z_{t+1}|Z_t, y_{t+1})\, d\mathbb{P}(Z_t|y^{t+1})$$
$$\propto \int \underbrace{p(Z_{t+1}|Z_t, y_{t+1})}_{\text{propagate}}\, \underbrace{p(y_{t+1}|Z_t)}_{\text{resample}}\, d\mathbb{P}(Z_t|y^t).$$

The distribution $d\mathbb{P}(Z_t|y^{t+1}) \propto p(y_{t+1}|Z_t)d\mathbb{P}(Z_t|y^t)$ is a 1-step smoothing distribution. Here $\mathbb{P}(Z_t|y^t)$ denotes the current distribution of the current state vector and in particle form corresponds to $\frac{1}{N}\sum_{i=1}^{N}\delta_{Z_t^{(i)}}$, with δ a Dirac measure.

Bayes rule above then gives us a prescription for constructing a sequential simulation-based algorithm: given $\mathbb{P}(Z_t|y^t)$, find the smoothed distribution $\mathbb{P}(Z_t|y^{t+1})$ via resampling and then propagate forward using $p(Z_{t+1}|Z_t, y_{t+1})$. This simply finds draws from the next filtering distribution $\mathbb{P}(Z_{t+1}|y^{t+1})$. Parameter inference is then achieved offline using $p(\theta|Z_{t+1})$.

From a sampling perspective, this leads to a very simple algorithm for updating particles $\{Z_t\}_{i=1}^N$ to $\{Z_{t+1}\}_{i=1}^N$ in the following three steps:

1. *Resample:* with replacement from a multinomial with weights proportional to the predictive distribution $p(y_{t+1}|Z_t^{(i)})$ to obtain $\{Z_t^{\zeta(i)}\}_{i=1}^N$;
2. *Propagate:* with $Z_{t+1}^{(i)} \sim p(Z_{t+1}|Z_t^{\zeta(i)}, y_{t+1})$ to obtain $\{Z_{t+1}^{(i)}\}_{i=1}^N$;
3. *Learning:* ν from $p(\nu|Z_{t+1})$.

The ingredients of particle learning are the essential state vector Z_t, a predictive probability rule $p(y_{t+1}|Z_t^{(i)})$ for resampling $\zeta(i)$, and a propagation rule to update particles: $Z_t^{\zeta(i)} \rightarrow Z_{t+1}^{(i)}$. The essential state vector will include the necessary conditional sufficient statistics for parameter learning given a model specification.

3.1. PL for the iid-t Case

First, we consider the normal location-scale model of Section 2.1 with $\mu = 0$ for simplicity. The model corresponds to a data augmentation scheme $(y_t|\sigma^2, \lambda_t) \sim N(0, \sigma^2\lambda_t)$ with $(\lambda_t|\nu) \sim IG(\nu/2, \nu/2)$. To complete the model, we assume priors of the form $\sigma^2 \sim IG(n_0/2, n_0\sigma_0^2/2)$ and Jeffreys prior $p(\nu)$ for ν (Eq. 1).

Now, the key to our approach is the use of an essential state vector Z_t. The algorithm requires the following distributions: $p(y_{t+1}|Z_t)$, $p(\nu, \sigma^2|Z_t)$, and $p(\lambda_t|\sigma^2, \nu, y_t)$. Bayes rule yields

$$p\left(\nu|\lambda^t\right) \equiv p(\nu|Z_{t1}, Z_{t2}) \propto p(\nu)\left(\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\right)^t Z_{t1}^{-(\nu/2+1)}\exp\{-\nu Z_{t2}/2\} \tag{2}$$

and

$$p(\sigma^2|y^t,\lambda^t) \equiv p(\sigma^2|Z_{t3},Z_{t4}) \sim IG(Z_{t3}/2, Z_{t4}/2) \tag{3}$$

with recursive updates for the parameter sufficient statistics

$$Z_{t1} = Z_{t-1,1}\lambda_t \quad \text{and} \quad Z_{t2} = Z_{t-1,2} + 1/\lambda_t,$$
$$Z_{t3} = Z_{t-1,3} + 1 \quad \text{and} \quad Z_{t4} = Z_{t-1,4} + y_t^2/\lambda_t,$$

with initial values $Z_{01} = 1$, $Z_{02} = 0$, $Z_{03} = n_0$, and $Z_{04} = n_0\sigma_0^2$.

Additionally, the predictive distribution for resampling and the latent state conditional posterior for propagation are directly available as

$$p(y_{t+1}|\lambda_{t+1}, Z_t) \sim t_{Z_{t3}+2}\left(0, \frac{Z_{t4}}{Z_{t3}+2}\lambda_{t+1}\right), \tag{4}$$

$$p(\lambda_t|\sigma^2, v, y_t) \sim IG\left(\frac{v+1}{2}, \frac{v + y_t^2/\sigma^2}{2}\right). \tag{5}$$

Therefore, we use an essential state vector given by $Z_t = (\lambda_{t+1}, Z_{t1}, Z_{t2}, Z_{t3}, Z_{t4})$. We are now ready to outline the steps of the PL scheme (see Panel A).

When $\mu \neq 0$ and a conditionally conjugate prior for location μ is used, say $N(\mu_0, \sigma^2 C_0)$, it follows that Eq. (3) is replaced by $p(\sigma^2|y^t,\lambda^t)p(\mu|\sigma^2,y^t,\lambda^t)$, while the vector Z_t is expanded accordingly. If instead the prior for μ is $N(\mu_0, C_0)$, independent of σ^2, then the essential vector Z_t would include one of the two parameters, most likely μ since in practice location parameters are easier to update.

3.2. PL for the SV-t Case

Particle learning for the SV-t model is similar to the iid-t model despite being somewhat more elaborated with the latent state now being the scale mixture λ_t as well as the log-volatilities h_t. In addition, there are three parameters (α, β, τ^2) driving the log-volatility dynamic behavior, as opposed to σ^2 in the iid-t model.

Static Parameters. Let us first deal with $\theta = (\alpha, \beta, \tau^2)$ the vector of fixed parameters driving the log-volatility equation (see Section 2.2). Conditional on the latent volatilities $h^t = (h_1, \ldots, h_t)$, sampling θ is rather straightforward since it is based on the conjugate Bayesian analysis of the normal linear regression with $x_t' = (1, h_{t-1})$ (Gamerman and Lopes, 2006, Chapter 2), i.e., $(\alpha, \beta|\tau^2) \sim N(b_t, \tau^2 B_t)$ and $\tau^2 \sim IG(c_t, d_t)$. The parameter sufficient statistics are $Z_t^\theta = (b_t, B_t, c_t, d_t)$, and they can determined recursively as

$$\begin{aligned}
B_t^{-1}b_t &= B_{t-1}^{-1}b_{t-1} + h_t x_t,\\
B_t^{-1} &= B_{t-1}^{-1} + x_t x_t',\\
c_t &= c_{t-1} + 1/2,\\
d_t &= d_{t-1} + (h_t - b_t' x_t)h_t/2 + (b_{t-1} - b_t)' B_{t-1}^{-1} b_{t-1}/2.
\end{aligned} \tag{6}$$

Start at time $t = 0$ with particle set $\{(\nu, \sigma^2, Z_{01}, Z_{02}, Z_{03}, Z_{04})^{(i)}\}_{i=1}^N$.

Step 1. For $i = 1, \ldots, N$,

- Sample $\lambda_{t+1}^{(i)} \sim IG(\nu^{(i)}/2, \nu^{(i)}/2)$,
- Set $Z_t^{(i)} = (\lambda_{t+1}, Z_{t1}, Z_{t2}, Z_{t3}, Z_{t4})^{(i)}$.

Step 2. Resample particles $\{(\tilde{\nu}, \tilde{\sigma}^2, \tilde{Z}_{t1}, \tilde{Z}_{t2}, \tilde{Z}_{t3}, \tilde{Z}_{t4})^{(i)}\}_{i=1}^N$ with weights proportional to $p(y_{t+1}|Z_t^{(i)})$ (equation 4),

Step 3. For $i = 1, \ldots, N$,

- Sample $\lambda_{t+1}^{(i)} \sim p(\lambda_{t+1}|\tilde{\sigma}^{2(i)}, \tilde{\nu}^{(i)}, y_{t+1})$ (equation 5),
- Update the essential state vector:

$$\begin{aligned} Z_{t+1,1} &= \tilde{Z}_{t1}^{(i)}\lambda_{t+1}^{(i)} \quad \text{and} \quad Z_{t+1,2} = \tilde{Z}_{t2}^{(i)} + 1/\lambda_{t+1}^{(i)} \\ Z_{t+1,3} &= \tilde{Z}_{t3}^{(i)} + 1 \quad \text{and} \quad Z_{t+1,4} = \tilde{Z}_{t4}^{(i)} + y_t^2/\lambda_{t+1}^{(i)} \end{aligned}$$

- Sample $\nu^{(i)} \sim p(\nu|Z_{t+1}^{(i)})$ (equation 2),
- Sample $\sigma^{2(i)} \sim p(\sigma^2|Z_{t+1}^{(i)})$ (equation 3).

Set $t = t + 1$ and return to step 1.

Q13

PANEL A Particle learning for the iid-t model.

Resampling Step. To sequentially resample the log-volatility h_t and propagate a new volatility state h_{t+1}, we use the Kim, Shephard, and Chib (1998) strategy of approximating the distribution of $\log \tilde{y}_t^2$, where $\tilde{y}_t^2 = y_t^2/\lambda_t$, by a carefully tuned seven-component mixture of normals[1]. Then, a standard data augmentation argument allows the mixture of normals to be conditionally transformed in individual normals, i.e., $(\varepsilon_t|k_t) \sim N(\mu_{k_t}, v_{k_t}^2)$, such that $k_t \sim \text{Mult}(\pi)$. Conditionally on k^t, the SV-t model for $z_{k_t} = \log y_t^2 - \log \lambda_t - \mu_{k_t}$ can be rewritten as a standard first order dynamic linear model, i.e., Q3

$$(z_{k_t}|h_t, \lambda_t, k_t) \sim N(h_t, v_{k_t}^2),$$

$$(h_t|h_{t-1}, \theta) \sim N(\alpha + \beta h_{t-1}, \tau^2),$$

with conditional state sufficient statistics $Z_t^h = (m_t, C_t)$ given by the standard Kalman recursions (West and Harrison, 1997). More explicitly, the conditional posterior Q4

[1]More precisely, $\log \tilde{y}_t^2 = h_t + \varepsilon_t$, where $\varepsilon_t = \log \epsilon_t^2$ follows a $\log \chi_1^2$ distribution, a parameter-free left skewed distribution with mean -1.27 and variance 4.94. They show that the $\log \chi_1^2$ can be well approximated by $\sum_{j=1}^7 \pi_j N(\mu_j, v_j^2)$, where $\pi = (0.0073, 0.1056, 0.00002, 0.044, 0.34, 0.2457, \ 0.2575)$, $\mu = (-11.4, -5.24, -9.84, 1.51, -0.65, 0.53, -2.36)$, and $v^2 = (5.8, 2.61, 5.18, \ 0.17, 0.64, \ 0.34, 1.26)$.

$(h_t|Z_t^h, \theta) \sim N(m_t, C_t)$ with moments given by

$$m_t = (1 - A_t)a_t + A_t z_{k_t} \quad \text{and} \quad C_t = (1 - A_t)R_t, \tag{7}$$

where $a_t = (\alpha + \beta m_{t-1})$, $A_t = R_t/Q_t$, $R_t = \beta^2 C_{t-1} + \tau^2$ and $Q_t = R_t + v_{k_t}^2$.

Essential State Vector. We will take advantage of the above Kalman recursions in the resampling step. We use an essential state vector of the form

$$Z_t = (\lambda_{t+1}, Z_t^\theta, Z_t^v, Z_t^h),$$

where the subset (Z_t^θ, Z_t^v) of Z_t is essentially the set $(Z_{t1}, \ldots, Z_{t4})$ derived from the iid-t model.

There are many efficiencies to be gained with this approach over traditional SMC approaches. For example, we only need to sample h_{t-1} and h_t (Step 2) in order to propagate Z_t^θ and sample θ (Step 4). In other words, PL does not necessarily need to keep track of the log-volatilities. For instance, point-wise evaluations of $p(h_t|y^t)$ can be approximated by the Monte Carlo average of the Kalman filter densities, i.e., $p^N(h_t|y^t) = \frac{1}{N}\sum_{i=1}^N p(h_t; m_t^{(i)}, C_t^{(i)})$.

For estimation of the fat-tails, we can use a Rao–Blackwellized density estimate. For example in the SV-t case, in order to reduce Monte Carlo error, we use an estimate of the form

$$p(v|y^t) = \mathbb{E}\{p(v|\lambda_t, h_t, y^t)\} \approx \frac{1}{N}\sum_{i=1}^N p(v|(\lambda^t, h^t)^{(i)}, y^t),$$

where $\{(\lambda^t, h^t)^{(i)}\}_{i=1}^N$ are draws from $p(\lambda^t, h^t|y^t)$. This leads to efficiency gains as the conditional $p(v|\lambda^t, h^t, y^t)$ and conditional mean $\mathbb{E}(v|\lambda^t, h^t, y^t)$ are known in closed form. We are now ready to outline the steps of the PL scheme for the SV-t model (see Panel B).

PL and MCMC. Although direct comparison with MCMC (Verdinelli and Wasserman, 1991) is not the focus of this article, we observe that MCMC is inherently a nonsequential procedure. MCMC provides the full joint distribution $p(h^T, \theta, v|y^T)$ including smoothing of the initial volatility states particle learning only computes $p(h_T, \theta|y^T)$—the distribution of the final state h_T and parameters θ. Another difference is in the assessment of MC errors. MCMC generates a dependent sequence of draws, PL has standard $\sqrt{N}$ MC bounds, but can suffer from accumulation of MC error for larger T. MCMC for learning fat-tails v can exhibit low conductance (Eraker et al., 1998), having difficulty escaping lower values of v in the chain, and can lead to poor convergence. Computationally speaking, the cost of performing MCMC sequentially is prohibitive high when compared to PL. Carvalho et al. (2010, Example 7) compares MCMC to PL in the simple first order normal dynamic linear model and show that, for $T = 1{,}000$ time periods and $N = 500$ particles, PL is roughly one order of magnitude faster than MCMC.

Step 0. Sample $\lambda_t^{(i)} \sim IG(\nu^{(i)}/2, \nu^{(i)}/2)$,

Step 1. Resample particles $\{(\tilde{Z}_{t-1}^{\theta}, \tilde{Z}_{t-1}^{h}, \tilde{\lambda}_t, \tilde{\theta})\}_{i=1}^N$ with weights

$$w_t^{(i)} \propto \sum_{k_t=1}^{7} \pi_i p_N(z_{k_t}^{(i)}; a_t^{(i)}, Q_t^{(i)}),$$

Step 2. Sample (h_{t-1}, h_t) from $p(h_{t-1}, h_t | Z_{t-1}^h, \lambda_t, \theta, y^t)$:

Step 2.1. Sample h_{t-1} from $\sum_{j=1}^7 \pi_j f_N(h_{t-1}; \hat{h}_{t-1,j}, V_{t-1,j})$, where

$$\begin{aligned} \hat{h}_{t-1,j} &= V_{t-1,i}(m_{t-1}/C_{t-1} + z_{ti}\beta/(v_i^2 + \tau^2)) \\ V_{t-1,j} &= 1/(1/C_{t-1} + \beta^2/(v_j^2 + \tau^2)) \end{aligned}$$

for $z_{ti} = \log y_t^2 - \log \lambda_t - \mu_i - \alpha$,

Step 2.2. Sample h_t from $= \sum_{j=1}^7 \pi_j f_N(h_t; \tilde{h}_{tj}, W_{tj})$, where

$$\begin{aligned} \tilde{h}_{ti} &= W_{ti}(\tilde{z}_{ti}/v_i^2 + (\alpha + \beta h_{t-1})/\tau^2) \\ W_{ti} &= 1/(1/v_i^2 + 1/\tau^2) \end{aligned}$$

for $\tilde{z}_{ti} = \log y_t^2 - \log \lambda_t - \mu_i$,

Step 3. Update $Z_{t+1}^{\nu(i)}$ (equation 4); sample $\nu^{(i)} \sim p(\nu | Z_{t+1}^{\nu(i)})$ (equation 2),

Step 4. Update $Z_t^{\theta(i)}$ (equation 6); sample $\theta \sim p(\theta | Z_t^{\theta(i)})$,

Step 5. Propagate $Z_t^{h(i)}$ (equation 7).

PANEL B Particle learning for the SV-t model

4. MODEL ASSESSMENT WITH A SEQUENTIAL BAYES FACTOR

Sequential model determination is performed using a Bayes factor $\mathcal{B}_T$ (Jeffreys, 1961; West, 1984). This naturally extends to a sequential version for an infinite sequence of (dependent) data we will still identify the "true" model. A probabilistic approach for determining how quickly you can learn the tail of the error distribution is to use the recursion Q6

$$\mathcal{B}_{T+1} = \frac{p(y_{T+1}|y_1, \ldots, y_T)}{q(y_{T+1}|y_1, \ldots, y_T)} \mathcal{B}_T.$$

Blackwell and Dubins (1962) provide a general discussion of the merging of opinions under Bayesian learning. They show that for any two models $p(y_1, \ldots, y_T)$ and $q(y_1, \ldots, y_T)$ that are absolutely continuous with respect to each other, opinions that merge in the following sense. First, $\mathscr{B}_T$ is a martingale, $\mathscr{F}_T$-measurable and under the true model Q,

$$\mathbb{E}_Q\left(\frac{p(y_{T+1}|y_1,\ldots,y_T)}{q(y_{T+1}|y_1,\ldots,y_T)}\middle|\mathscr{F}_T\right) = 1 \text{ so that } \mathbb{E}\left(\mathscr{B}_{T+1}|\mathscr{F}_T\right) = \mathscr{B}_T,$$

where $\mathscr{F}_T$ represents all the information up to time t.

By the martingale convergence theorem, $\mathscr{B}_\infty = \lim_{T\to\infty}\mathscr{B}_T$ exists almost surely under Q and in fact $\mathscr{B}_\infty = 0$ a.s. Q. Put simply, the sequential Bayes factor will correctly identify the "true" model Q under quite general data sequences include the SV-t model we consider here in detail. Furthermore, by the Shannon–McMillan–Breiman theorem (see, for example, Cover and Thomas, 2006), we can analyze the rate of learning via the quantity

$$\lim_{T\to\infty}\frac{1}{T}\ln q(y_1,\ldots,y_T) \to H \qquad \text{a.s. } Q,$$

where H is the entropy rate defined by $H = \lim_{T\to\infty}\mathbb{E}_Q\left(-\ln p(y_{T+1}|y_1,\ldots,y_T)\right) < 0$. Hence as $H \in [-\infty, 0)$, we have that $\mathscr{B}_\infty = 0$. A similar result for the marginal likelihood ratio

$$\lim_{T\to\infty}\frac{1}{T}\ln\frac{p(y_1,\ldots,y_T)}{q(y_1,\ldots,y_T)} \to \lim_{k\to\infty}\mathbb{E}_Q\left(\ln\frac{p(y_{k+1}|y_k,\ldots,y_1)}{q(y_{k+1}|y_k,\ldots,y_1)}\right) < 0 \qquad \text{a.s. } Q.$$

We will use this in the next subsection.

Bayes factors have a number of attractive features as they can be converted into posterior model probabilities when the model set is exhaustive. Lopes and Tobias (2011) provide a recent survey including computational strategies based on the Savage–Dickey density ratio. These results are only asymptotic, and with a finite amount of data, it helps to analyze the rate of learning using a Kullback–Leibler metric.

4.1. Discriminating a t_4 from a Gaussian

We can use these theoretical insights (see also Edwards et al., 1963; Lindley, 1956) to address the question *a priori* of *"how long a time series one would have to observe* Q7 *to have strong evidence of a t_4 versus a Gaussian*?" Jeffreys observed that one needs data sequences of length $T = 500$ to be able to discriminate the tails of an underlying probability distribution. We now formalize this argument using our sequential Bayes factor. One is motivated to define *a priori* the "expected" log-Bayes factor for a given data length, $\overline{\mathscr{BF}}_T$, under the Gaussian model

$$\frac{1}{T}\ln\overline{\mathscr{B}}_T = \mathbb{E}_{t_\infty}\ln\frac{t_v}{t_\infty} = KL(t_v, t_\infty)$$

under the Gaussian t_∞-model where KL denotes Kullback–Leibler divergence. Then, *a priori*, if we are given a level of Bayes factor discrimination $\overline{\mathcal{BF}}_T$, we then have to observe on average $T^\star$ observation to be able to discriminate the two models where

$$T^\star = \frac{1}{KL(t_\nu, t_\infty)} \ln \overline{\mathcal{BF}}_T.$$

This measure is asymmetric, as if the data is generated by a t_ν distribution, the constant changes to $KL(t_\infty, t_\nu)$. Q8

To illustrate the magnitudes of these effects, if we take $\nu = 3$ and $\mathcal{B} = 10$ (strong evidence), for example, this argument would suggest that on average $T = 150$ observations from a standard normal are needed to strongly reject the t_3 model, and on average $T = 20$ observations from the t_3 to strongly reject the standard normal distribution. This is borne out in our empirical study. Figure 1 plots the first factor in the above expression, namely, the Kullback–Leibler divergence between the t_ν-family and the Gaussian.

This also confirms the analysis in Gramacy and Pantaleo (2010). In a multivariate regression setting, they perform a Monte Carlo experiment where T and ν varies with $T \in \{30, 75, 100, 200, 500, 1000\}$ and $\nu \in \{3, 5, 7, 10, \infty\}$. They observed the frequency of time the $\mathcal{B}$ indicated *strong* preference ($\mathcal{B} > 10$) for a model. Under normal errors, $\nu = 3$ could be determined with high accuracy for $T \leq 200$, $\nu = 5$ took $T \leq 1{,}000$, and for $10 \leq \nu < \infty$

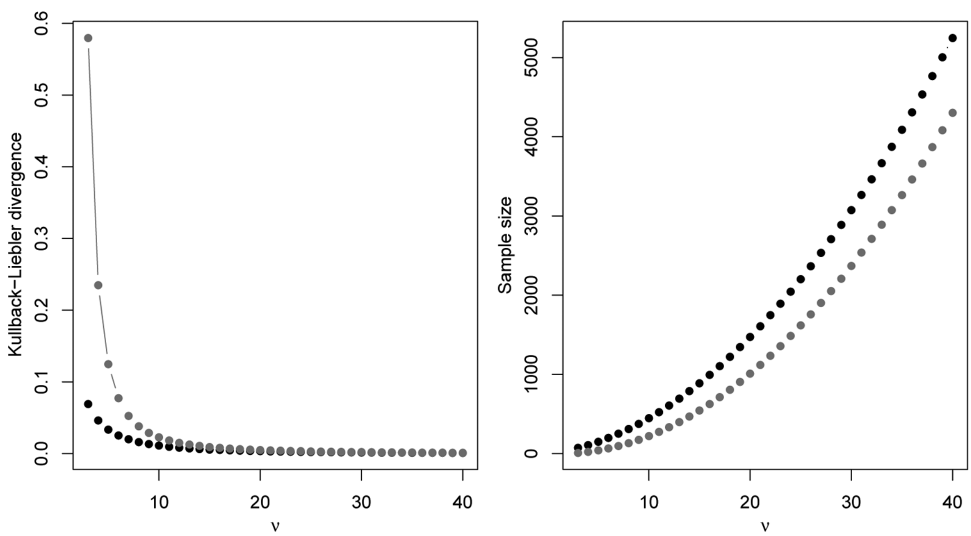

FIGURE 1 *i.i.d. model.* Discriminating a t_ν from a Gaussian. $KL(t_\nu, t_\infty)$ (black) and $KL(t_\infty, t_\nu)$ (grey). For $\nu = 4, 10, 20$, theoretical sample sizes are $T^* = 108, 446, 1{,}473$ for strong evidence against normality and $T^* = 22, 220, 1{,}009$ for strong evidence against t_ν.

very large samples would be required to discriminate the tails with any degree of posterior accuracy. Of course, for a given dataset, the Bayes factor might provide strong evidence even for small samples. The Jeffreys prior then has the nice property (by definition of the inverse of the Fisher information matrix) of down-weighting these regions of the parameter space where it is hard to learn the parameters.

It is also interesting to address the asymptotic behavior of the fat-tailed posterior distribution when the true model is not in the set of models under consideration. Berk (1966, 1970) assumes that the data generating process comes from $y_t \sim q(y)$— a model outside our current consideration. Given our fat-tailed model $p(y|\theta, \nu)$, Berk shows that under mild regularity conditions the posterior distribution $p(\theta, \nu|y)$ will asymptotically concentrate with probability one on the subset of parameter values where the Kullback–Leibler divergence between $p(y|\theta, \nu)$ and $q(y)$ is minimized or equivalently $\int \log p(y|\theta, \nu)q(y)dy$ is maximized.

5. EMPIRICAL RESULTS

We now illustrate our methodology for iid SV-Student's t error distributions (see Sections 2.1 and 2.2 for the specifications). The iid-t model illustration will serve the additional and important purpose of showing that the uniform prior is not necessarily always a harmless prior. The SV-t model will be estimated sequentially on the British pound/U.S. dollar daily exchange rate series and daily returns on the S&P500 from a period in 2007–2010 that includes the credit crisis. Resulting inferences will be compared with MCMC at the end of the sample.

5.1. The iid-t Model

To illustrate the efficiency of our approach, we simulate a sample of size $T = 200$ from a Student's t_4 distribution, centered at zero and unit scale, i.e., $\sigma^2 = 1$. Figures 2 and 3 show the joint posterior distributions of $p(\sigma^2, \nu|y^t)$ for $t = 50, 100, 150$, and 200 under, respectively, the uniform prior and the Jeffreys prior of Fonseca et al. (2008). As the model implies that $Var(y_t) = \sigma^2\nu/(\nu - 2)$, one should not be too surprised that there is a posterior correlation between σ^2 and ν for small values of ν.

It is clear that the posterior provides fairly accurate sequential estimates for the joint as well as the marginal distributions (the exact posterior probabilities are computed on a fine bivariate grid). On the one hand, the Jeffreys prior, as anticipated, penalizes larger values of ν with the penalization slightly decreasing as the sample size increases. On the other hand, the uniform prior is impartial with respect to the number of degrees of freedom, so any information regarding ν comes exclusively from the likelihood which, in turn, is fairly uninformative about ν for $t = 50, 100$, and 150. Even when $t = 200$, there is still no

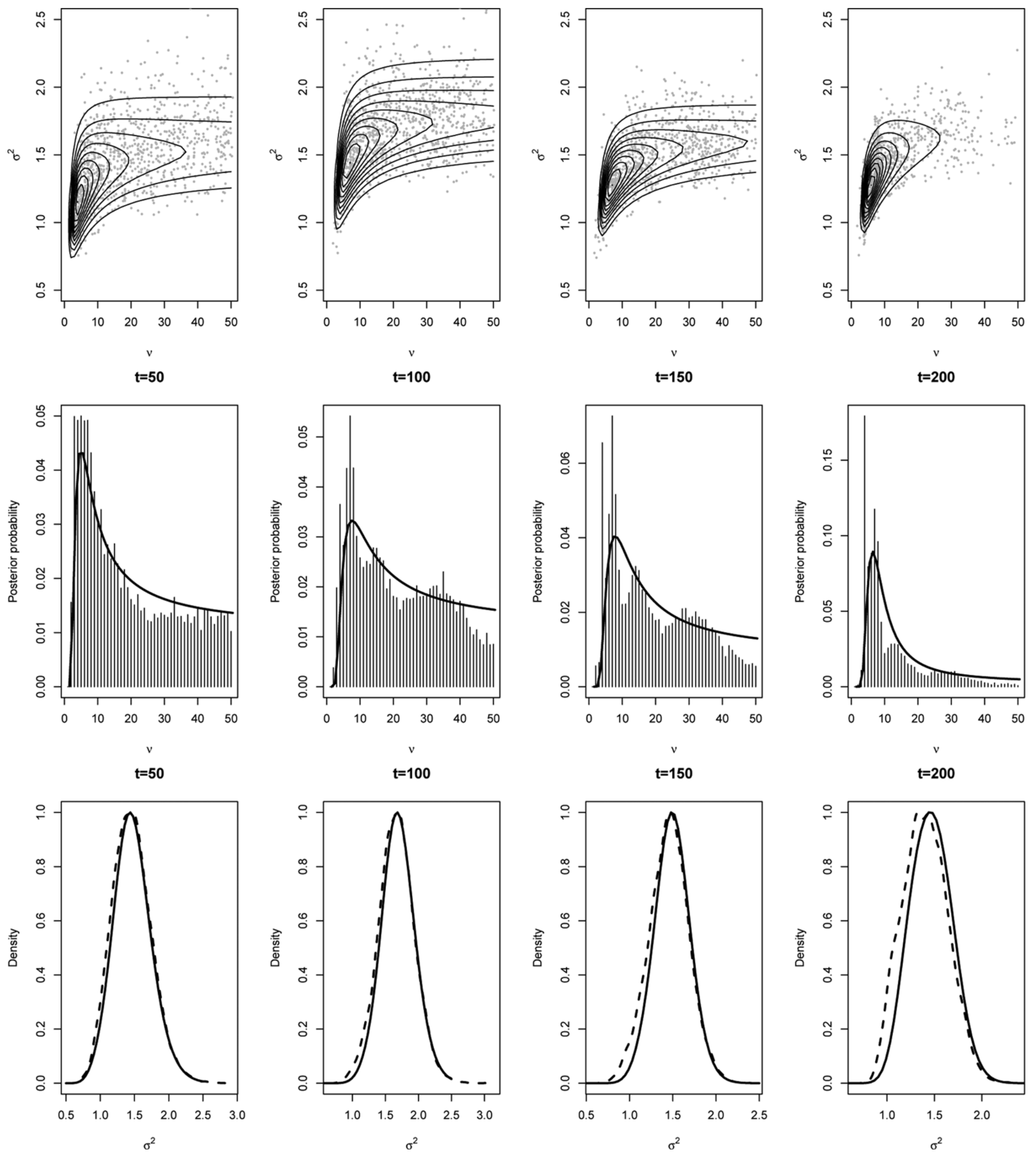

FIGURE 2 *i.i.d. model.* Sequential posterior inference for (σ^2, ν) based on PL for $T = 200$ iid observations drawn from t_4 with uniform prior for ν. PL is based on $N = 10{,}000$ particles.

negligible mass for values $\nu > 10$. Figure 4 shows that PL estimates are still accurate when $n = 1{,}000$. It also shows that the marginal posterior of ν is highly concentrated around the true value for $t > 500$, as theoretically predictive in Section 4.1 and Fig. 1.

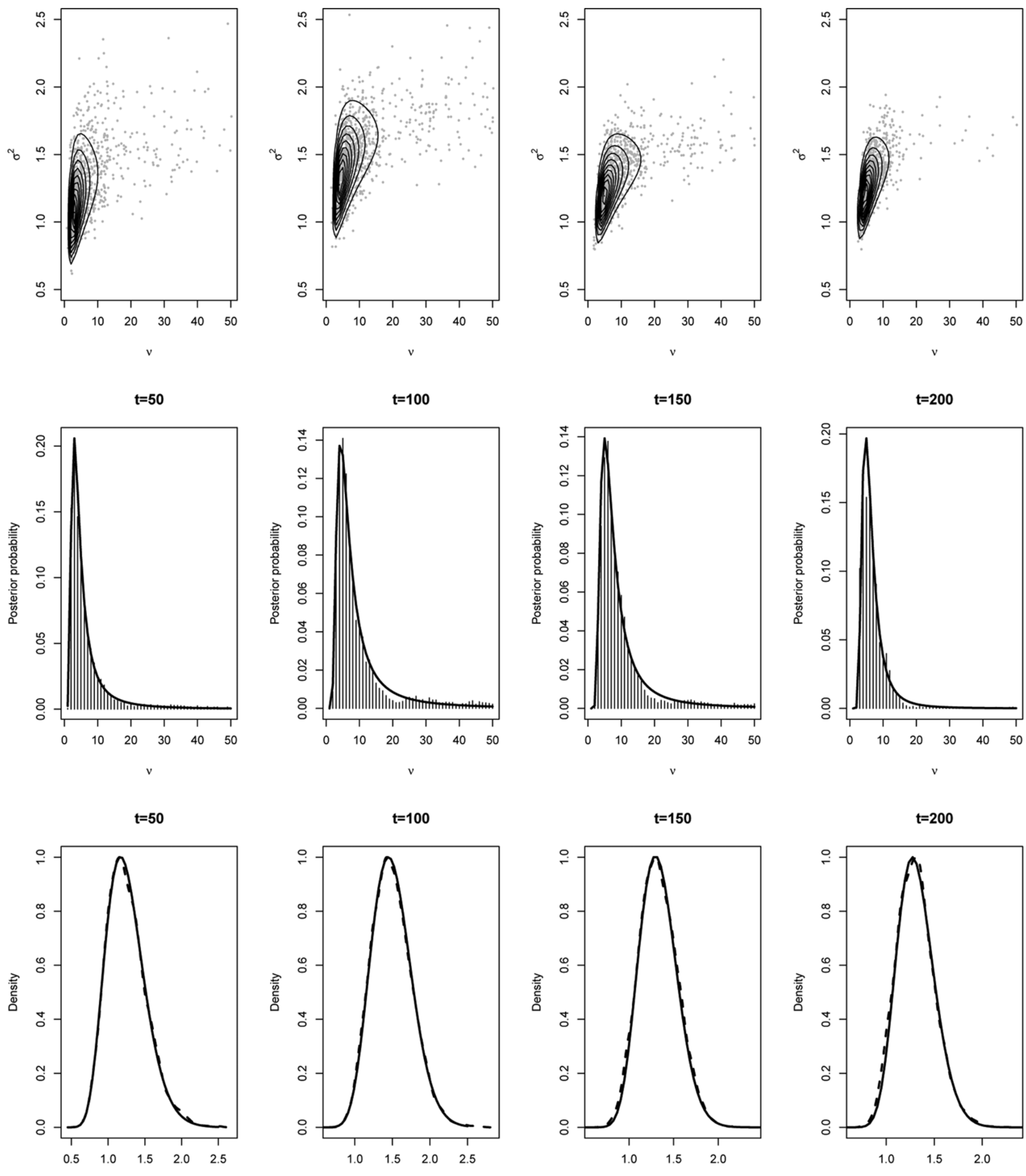

FIGURE 3 *i.i.d. model.* Sequential posterior inference for (σ^2, ν) based on PL for $T = 200$ iid observations drawn from t_4 with Jeffreys prior for ν. PL is based on $N = 10{,}000$ particles.

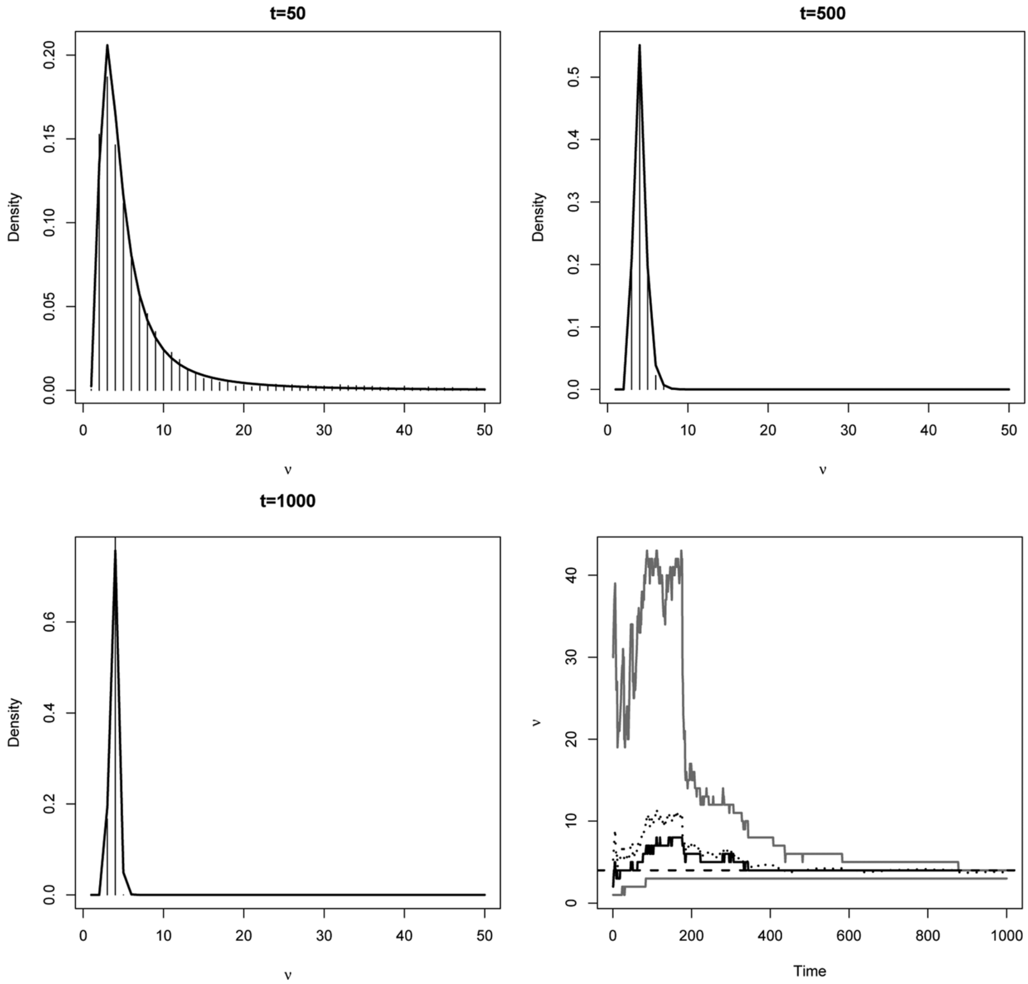

FIGURE 4 *i.i.d. model.* Sequential posterior inference for ν based on PL for $T = 1{,}000$ iid observations drawn from t_4 with Jeffreys prior for ν. PL is based on $N = 10{,}000$ particles.

The undesirable bias of the not-so-harmless uniform prior is highlighted in the Monte Carlo exercise summarized by Figs. 5 and 6. The posterior means, medians, and modes of ν based on $p(\nu|y^t)$, $t = 30, 50, 100, 300, 400$, and 500 are compared across $R = 50$ samples. As it can be seen, the bias of the uniform prior is striking for samples of size up to $T = 100$, when compared to those of the Jeffreys prior. For samples of size $T = 400$ and $T = 500$, the bias is much smaller, but a closer look reveals its presence. For example, the 25th percentiles of the mean, median, and mode box-plots when $T = 500$ are all above the true value $\nu = 4$ for the uniform prior.

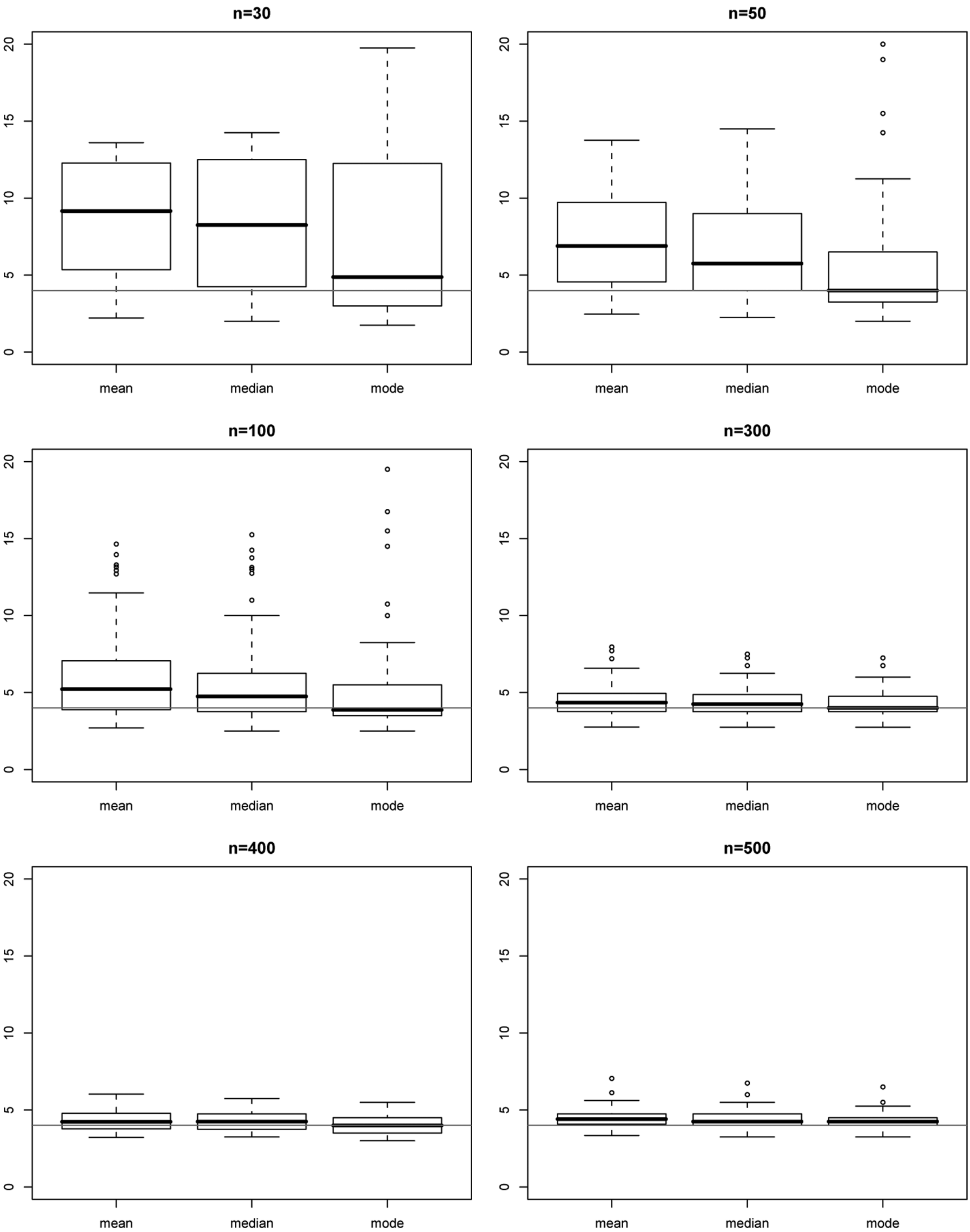

FIGURE 5 *i.i.d. model.* Posterior mean, median, and mode for the number of degrees of freedom v under the uniform prior, for different sample sizes and based on a Gibbs sampler of length $M = 1{,}000$ after a burn-in period of M_0 draws. Boxplots are based on $R = 50$ datasets.

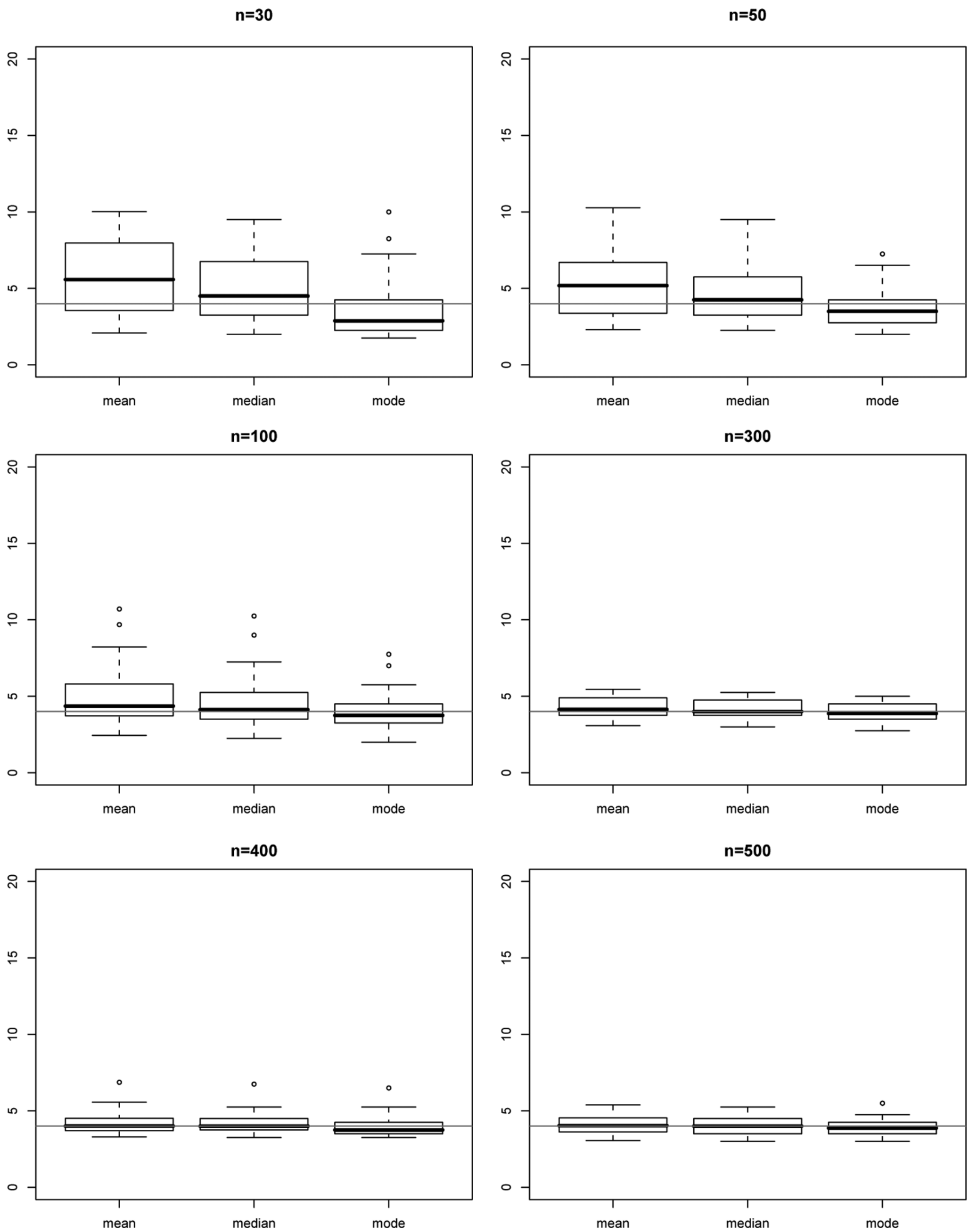

FIGURE 6 *i.i.d. model.* Posterior mean, median, and mode for the number of degrees of freedom ν under the Jeffreys prior, for different sample sizes and based on a Gibbs sampler of length $M = 1{,}000$ after a burn-in period of M_0 draws. Boxplots are based on $R = 50$ datasets.

5.2. The SV-t Model

We now revisit the well-known British pound versus U.S. dollar exchange rate data of Jacquier et al. (2004). The data consists of $T = 937$ daily rates form October 1st, 1981 to June, 28th 1985. For illustration purposes, we simulated data with exactly the same length from a SV-t_4 model with parameters $(v, \alpha, \beta, \tau^2) = (4, -0.202, 0.980, 0.018)$ and initial value $h_0 = -8.053$. Both simulated and real data sets are presented in Fig. 7.

The prior distribution of v is given by the discretized version of Fonseca et al.'s (2008) Jeffreys prior, similar to the approach taken in Section 5.1 (see Eq. 1). The vector log-volatility parameters (α, β, τ^2) are independent, *a priori*, of v and its prior distribution is given by $(\alpha, \beta)|\tau^2 \sim N(b_0, \tau^2 B_0)$ and $\tau^2 \sim IG(\eta_0/2, \eta_0\tau_0^2/2)$, while the posterior for the log-volatility at time $t = 0$ is given by $h_0 \sim N(m_0, C_0)$. The hyper-parameters are set at

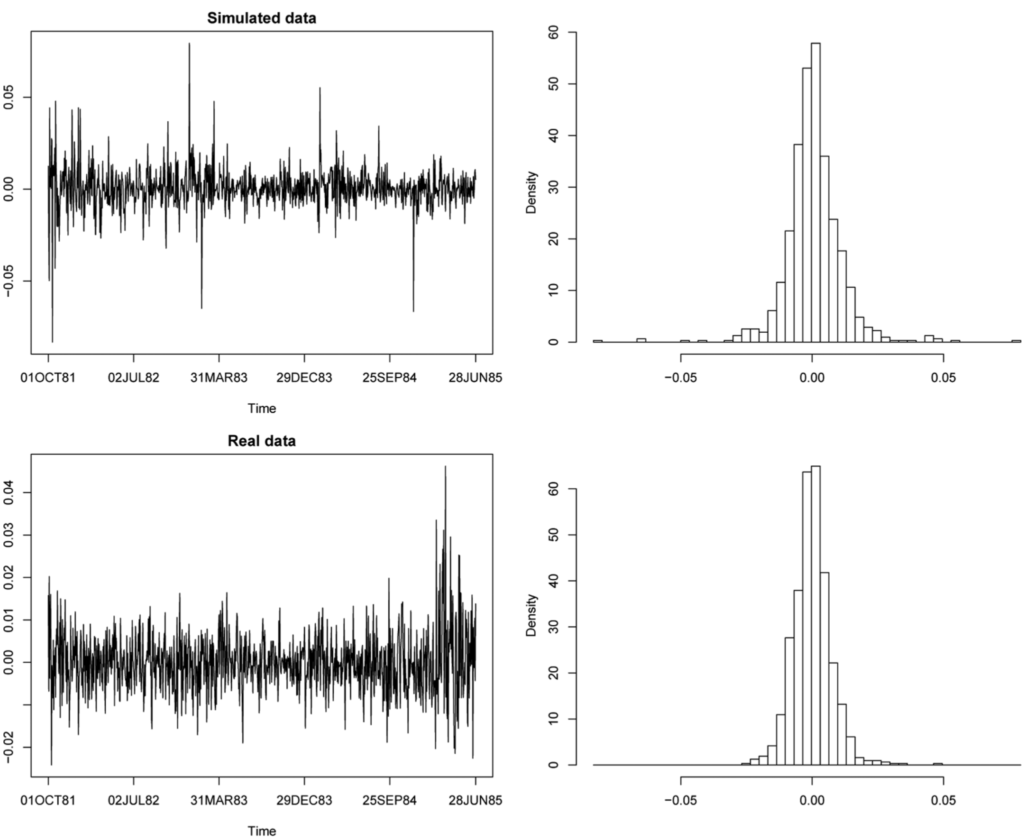

FIGURE 7 *SV-t model.* The top row corresponds to simulated data ($T = 937$) from the SV-t_v model with parameters $v = 4$, $\alpha = -0.202$, $\beta = 0.980$, $\tau^2 = 0.018$, and $x_0 = -8.053$. The bottom row corresponds to JPR's (1994) British pound vs. U.S. dollar exchange ($T = 937$) daily rates from go from October 1, 1981 to June 28, 1985.

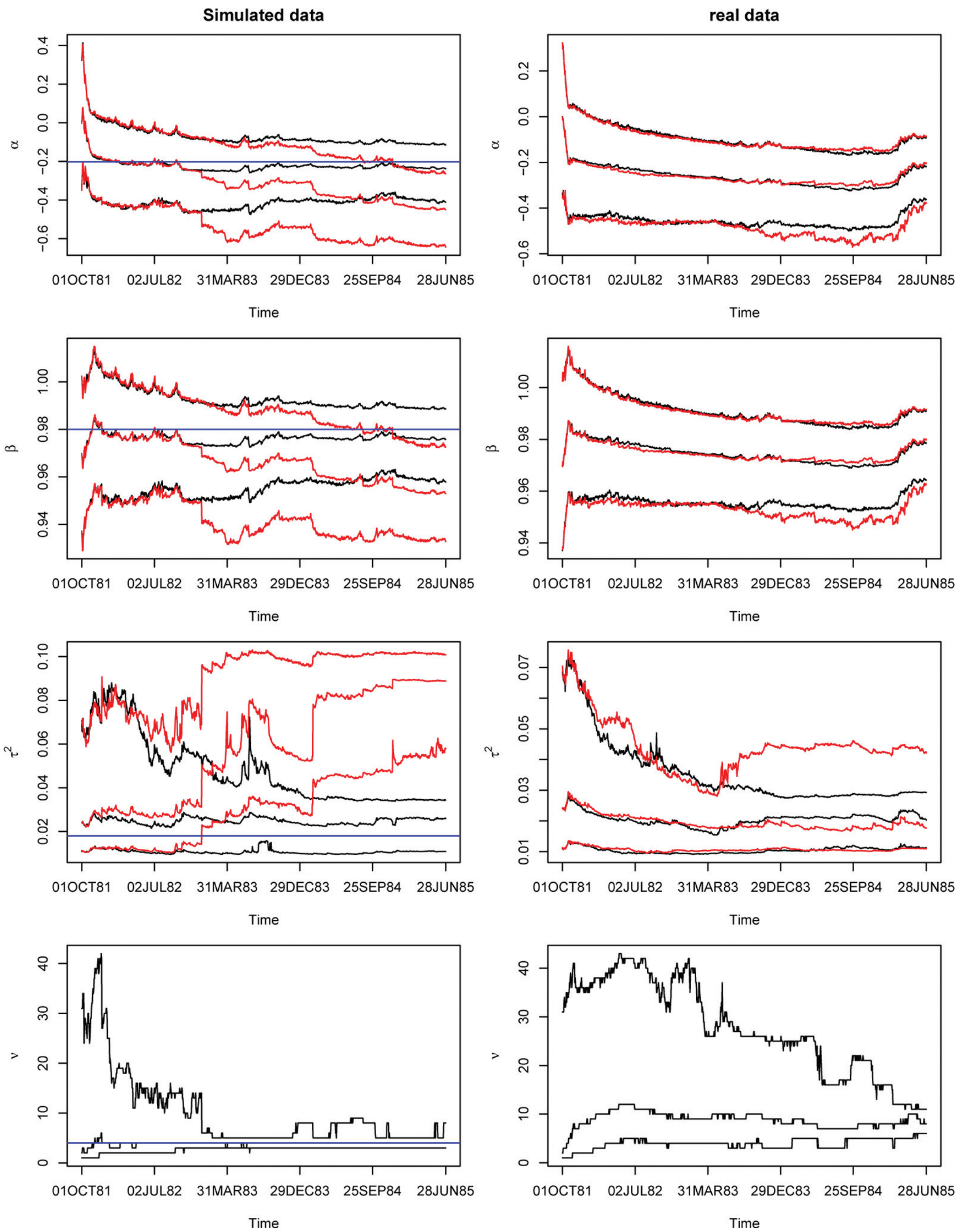

FIGURE 8 *SV-t model.* (2.5, 50, 97.5)th percentiles of the sequential marginal posterior distributions of α, β, τ^2, and ν for the normal (red lines) and Student's t (black lines) models.

the values $m_0 = \log y_1^2$, $C_0 = 1.0$, $b_0 = (-0.002, 0.97)$, $B_0 = \text{diag}(1.0, 0.01)$, $c_0 = 5.0$, and $d_0 = 0.1125$.

Posterior inference is based on PL with $N = 10,000$ particles. Figures 8 presents 2.5th, 50th and 97.5th percentiles of the sequential marginal distributions of α, β, τ^2, and v for both simulated and real data sets. For the simulated data, the posterior distribution of v concentrates around the true value $v = 4$ after about 350 observations. For the real data, v is highly concentrated with around ten degrees of freedom at the end of the sample; however, the right tail of the distribution, i.e., large degrees of freedom, is fairly long for most of the sample. Another interesting fact is that both normal and Student's t model learn about α and β in a similar manner, while the same cannot be said for the volatility of the log-volatility parameter, τ^2. This is perhaps not surprising as the normal model overestimates the volatility of log-volatility to accommodate the fact that daily rates violate the plain normality assumption. The same behavior is present in our simulated data exercise. In fact, the posterior distribution for the log-volatilities, $p(h_t|y^t)$, for the simulated data based on the normal model has larger uncertainty than for the t_v model (figure not shown here). Finally, at the end of the sample we can calculate the marginal posterior on the tail-thickness $p(v|y^T)$, our sequential particle approach agrees with the MCMC analysis of Jacquier et al. (2004). This suggests that the MC accumulation error inherent in our particle algorithm is small for these types of data length and models.

5.2.1. *S&P500: Credit Crisis 2008–2009*

To study the effect of the credit crisis on stock returns, we revisit daily S&P500 returns previously studied, amongst many others, by Abanto-Valle et al. (2010) and Lopes and Polson (2010b). The former article estimates SV models with errors in the class of symmetric scale mixtures of normal distributions and also base their illustration on the S&P500 index from January 1999 to September 2008, therefore missing most of the credit crunch crisis and its aftermath. We concentrate our analysis on the period starting on January 3, 2007 and ending on October 14, 2010 ($T = 954$ observations). We sequentially fit the normal model to this data set as well as the t_v model for $v \in \{5, 10, 50\}$. Figure 9 summarizes our findings. The three Student's t models have higher predictive power than the normal model when measured in terms of log-Bayes factors. This distinction is particularly strong when comparing the t_5 (or t_{10}) model with the normal model. Interestingly, the t_5 model becomes gradually closer to the normal model from July 2008 to July 2010, when again it distances itself from normality.

Before the onset of the credit crisis in July 2008, the model with the largest Bayes factor (relative to a normal), and hence the largest posterior model probability (under a uniform prior on v) is the t_5-distribution. This is not surprising as the previous time period consisted of little stochastic volatility and the occasional outlying return—which is nicely accomodated by a t_5 error distribution, in the spirit of Jeffreys initial observation about "real" data. The interesing aspects of Bayesian learning occur in the period of the crisis

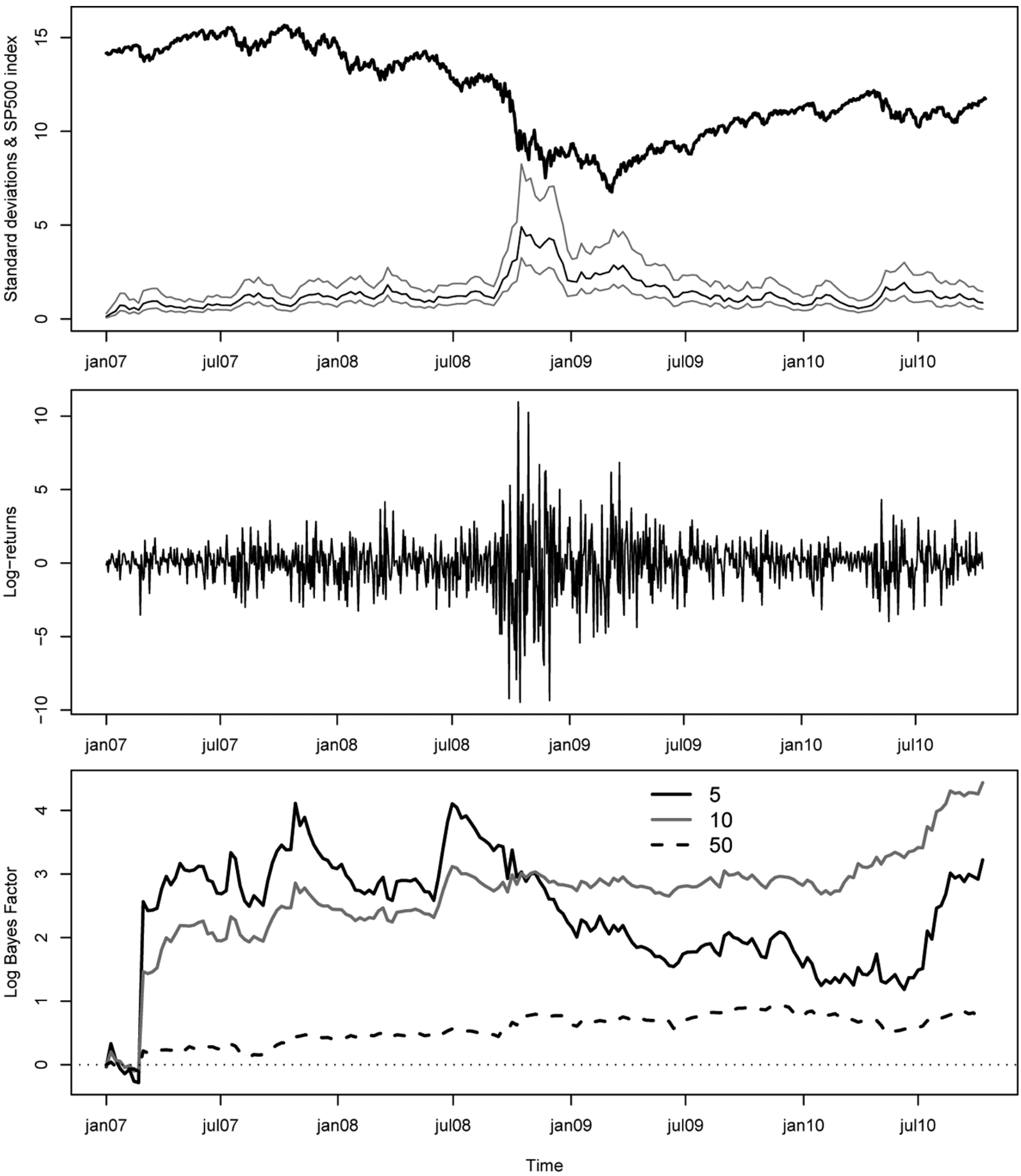

FIGURE 9 *SV-t model for S&P500 returns.* Top frame: S&P500 daily closing price (divided by 100: solid thick line) along with PL approximations to the $(2.5, 50, 97.5)$th percentiles of the posterior distributions of the time-varying standard deviations $p(\exp\{x_t/2\}|y^t)$, for $t = 1, \ldots, T$, under the SV-t_{10} model. Middle frame: Log returns. Bottom frame: Logarithm of the Bayes factors of t_ν against normality for $\nu \in \{5, 10, 50\}$.

from July 2008 to March 2009. One immediately sees a dramatic increase in the stochastic volatility component of the model and the clustering of a high period of volatility. In and of itself, this is sufficient to "explain" the extreme moves in the market. Correspondingly, in terms of online estimation of the fat-tails, the Bayes factor quickly moves to favor the model with light tails, here the t_{10}-distribution. Finally, as the crisis subsides, the volatility mean reverts and the returns again look like they exhibit some outlying behavior (relative to the level of volatility) and the sequential Bayes again starts to move to favor the fatter-tailed t_5-distribution.

6. DISCUSSION

Estimating tail-thickness of the error distribution of an economic or financial time series is an important problem as estimates and forecasts are very sensitive to the tail behavior. Moreover, we would like an on-line estimation methodology that can adaptively learn the tail-thickness and provide parameter estimates that update as new data arrives. We model the error distribution as a t_ν-distribution where $\nu \sim p(\nu)$, and we adopt a default Jeffreys prior on the tail-thickness parameter ν. We show that this has a number of desirable properties when performing inference with a finite amount of data. We use the sequential Bayes factor to provide an on-line test of normality versus fat-tails, and we derive its optimality properties asymptotically and in finite sample using a Kullback–Leibler metric. We illustrate these effects in the credit crisis of 2008–2009 with daily S&P500 stock return data. Our analysis shows how quickly an agent can dynamically learn the tail of the error distribution whilst still accounting for parameter uncertainty and time-varying stochastic volatility. Figures 2–4 and 8 all show that estimating ν is in fact rather difficult. Figure 8, in particular, shows that when the data is not normal it takes several time periods for the parameter ν be stably estimated.

Whilst MCMC is computationally slow for solving the online problem, it does also provide the full smoothing distribution at the end of the sampler. This would require $O(N^2)$ particles in our approach (see Carvalho et al., 2010, for further discussion), and therefore, if smoothed states are required, we recommend filtering forward with particles and smoothing with MCMC. Other estimation methods such as nested Laplace approximation (Smith, 2000) seem unable to identify the true error structure due to the multimodalities present in the posterior and particle methods provide a natural alternative. Clearly, there are a number of extensions of our approach, for example, to multivariate and dynamic panel data.

REFERENCES

Abanto-Valle, C. A., Bandyopadhyay, D., Lachos, V. H., Enriquez, I. (2010). Robust Bayesian analysis of heavy-tailed stochastic volatility models using scale mixtures of normal distributions. *Computational Statistics and Data Analysis* 54:2883–2898.

Andrews, D. F., Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of Royal Statistical Society, Series B* 36:99–102. Q9

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* 37:51–58.

Berk, R. H. (1970). Consistency a posteriori. *Annals of Mathematical Statistics* 41:894–906.

Blackwell, D., Dubins, L. (1962). Merging of opinions with increasing information. *Annals of Mathematical Statistics* 33:882–886.

Carlin, B. P., Polson, N. G., Stoffer, D. S. (1992). A Monte Carlo approach to nonlinear and non-normal state space models. *Journal of the American Statistical Association* 87:493–500.

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

Chib, S., Nardari, F., Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics* 108:281–316.

Cover, T. M., Thomas, J. A. (2006). *Elements of Information Theory*. 2nd ed. New York: Wiley.

Edgeworth, F. Y. (1888). On a new method of reducing observations relating to several quantities. *Philosophical Magazine* 25:184–191.

Edwards, W., Lindman, H., Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* 70:193–242.

Eraker, B., Jacquier, E., Polson, N. G. (1998). The pitfalls of MCMC algorithms. Technical Report, The University of Chicago Booth School of Business.

Fernandez, C., Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 93:359–371.

Fonseca, T., Ferreira, M. A. R., Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika* 95:325–333.

Gamerman, D., Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Baton Rouge: Chapman & Hall/CRC.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear linear model. *Journal of Applied Econometrics* 8:19–40.

Gordon, N., Salmond, D., Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings* F-140:107–113. Q10

Gordon, N., Smith, A. F. M. (1993). Approximate non-Gaussian Bayesian estimation and modal consistency. *Journal of Royal Statistical Society, Series B* 55:913–918. Q10

Gramacy, R., Pantaleo, E. (2010). Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. *Bayesian Analysis* 5:237–262.

Jacquier, E., Polson, N. G. (2000). Discussion of "Time series analysis of non-Gaussian observations". *Journal of Royal Statistical Society, B* 62:44–45.

Jacquier E., Polson, N. G., Rossi, P. E. (2004). Bayesian analysis of stochastic volatility with fat tails and leverage effect. *Journal of Econometrics* 122:185–212.

Jeffreys, H. (1961). *Theory of Probability*. New York: Oxford University Press.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics* 27:986–1005.

Lopes, H. F., Carvalho, C. M. (2013). Online Bayesian learning in dynamic models: An illustrative introduction to particle methods. In: West, M., Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. *Bayesian Theory and Applications*. Clarendon: Oxford University Press, pp. 203–228.

Lopes, H. F., Carvalho, C. M., Johannes, M. S., Polson, N. G. (2010). Particle learning for sequential Bayesian computation (with discussion). In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M., eds. *Bayesian Statistics*, Vol. 9. Oxford: Oxford University Press. To appear. Q11

Lopes, H. F., Polson, N. G. (2010a). Bayesian inference for stochastic volatility modeling. In: Böcker, K., ed. *Re-Thinking Risk Measurement, Management and Reporting Measurement Uncertainty, Bayesian Analysis and Expert Elicitation*. Riskbooks, pp. 515–551.

Lopes H. F., Polson, N. G. (2010b). Extracting SP500 and NASDAQ volatility: The credit crisis of 2007-2008. In: O'Hagan, A., West, M. eds. *Handbook of Applied Bayesian Analysis*. Oxford: Oxford University Press, pp. 319–342.

Lopes, H. F., Tobias, J. (2011). Confronting prior convictions: On issues of prior and likelihood sensitivity in Bayesian analysis. *Annual Review of Economics* 3:107–131.

Lopes, H. F., Tsay, R. (2011). Particle filters and Bayesian inference in financial econometrics. *Journal of Forecasting* 30:168–209.

McCausland, W. (2012). The HESSIAN method: Highly efficient simulation smoothing, in a nutshell. *Journal of Econometrics* 168:189–206.

Smith, A. F. M. (1983). Bayesian approaches to outliers and robustness. In: Florens, J. P., Mouchart, M., Raoult, J. P., Simar, L., Smith, A. F. M. eds. *Specifying Statistical Models: From Parametric to Nonparametric, Using Bayesian or Non-Bayesian Approaches*. New York: Springer-Verlag, pp. 13–35. Q12

Smith, J. Q. (2000). In discussion of "Time series analysis of non-Gaussian observations". *Journal of Royal Statistical Society, B* 62:29–20.

Verdinelli, I., Wasserman, L. (1995). Computing Bayes factors by using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association* 90:614–618.

West, M. (1981). Robust sequential approximate Bayesian estimation. *Journal of Royal Statistical Society, Series B* 43:157–166. Q12

West, M. (1984). Bayesian model monitoring. *Journal of Royal Statistical Society, Series B* 48:70–78.