Quantile Filtering and Learning

Michael Johannes, Nicholas Polson and Seung M. Yae*

October 5, 2009

Abstract

Quantile and least-absolute deviations (LAD) methods are popular robust statistical methods but have not generally been applied to state filtering and sequential parameter learning. This paper introduces robust state space models whose error structure coincides with quantile estimation criterion, with LAD a special case. We develop an efficient particle based method for sequential state and parameter inference. Existing approaches focus solely on the problem of state filtering, conditional on parameter values. Our approach allows for sequential hypothesis testing and model monitoring by computing marginal likelihoods and Bayes factors sequentially through time. We illustrate our approach with a number of applications with real and simulated data. In all cases we compare our results with existing algorithms where possible and document the efficiency of our methodology.

1 Introduction

This paper provides two contributions to the literature on sequential inference in state space models. First, we introduce a class of robust state space models with error structures that

*Johannes is at the Graduate School of Business, Columbia University, 3022 Broadway, NY, NY, 10027, mj335@columbia.edu. Polson and Yae are at the Booth School of Business, University of Chicago, 5807 S. Woodlawn, Chicago IL 60637, ngp@chicagobooth.edu, yae@chicagobooth.edu.

generate common robust estimation criterion such as quantile, least absolute deviations (LAD). Second, we develop efficient particle filtering algorithms for sequential learning of both parameters and state variables, modifying the exact or perfect particle sampling approach of Johannes and Polson (2007). The particle based algorithm provides samples from the posterior distribution of the parameters and states, conditional on the observed data, and is fully recursive, a key to its computational attractiveness.

Robust procedures such as LAD or quantile estimation (Huber, 1981, and Koenker and Hallock, 2001) have been successfully applied in a wide range of settings. One of the main reasons for the popularity of these methods are their parsimony: they generally do not introduce any new, difficult-to-estimate, parameters. This can be contrasted with approaches that specify flexible error distributions such as t-distributed errors or discrete mixtures of normal distributions.

Surprisingly, these robust statistical methods have not been widely applied in state space models. Existing robust filtering procedures with either non-Gaussian state or observation noise use theoretical results in Masreliez (1975), see for example, Masreliez and Martin (1977), West (1981), Martin and Raftery (1987), Meinhold and Singpurwalla (1987), Gordon and Smith (1993), Le, Martin, and Raftery, (1996). These approaches typically assume the parameters are known, which is not the case in most practical applications.

In this paper, we introduce state space models with observation and state errors whose error distribution generates objective functions that coincide with quantile-norm estimation with LAD a special case. This allows robustness in the observation errors, the state evolution errors, or both. We prove a key result that quantile errors can be expressed as a scale mixture of normal distributions, and the mixing variable is used as an auxiliary variable in filtering and learning.

For inference, we develop an efficient particle filtering approach for sequential inference on both parameters and states. Previous robust procedures, cited above, use approximate

or linear methods and focus exclusive on estimating latent states, abstracting from the difficult problem of incorporating parameter learning. To do this, we combine scale mixture error distributions, particle approximations, and exact sampling methods. First, scale mixture error representations, combined with data augmentation, generate a sufficient statistic structure for parameter posteriors and aid in state propagation. This step has been widely used in the Bayesian literature using MCMC methods (see, for example, Carlin and Polson, 1991, Carlin, Polson and Stoffer, 1992). Second, particle approximations discretize the conditional distribution of the parameters and latent states given observed data, providing a computationally feasible approach for sequential inference. Third, we adapt the approach developed in Johannes and Polson (2007) for sequential inference to these robust settings. The approach requires a fixed-dimensional sufficient statistic for the parameters, which are generated conditional on the states, data, and auxiliary scale mixture variables. A number of comparisons, detailed in the following paragraphs, show that our particle filtering algorithm is more efficient than existing analytical approximations or standard particle filtering algorithms such as Gordon et al. (1993) or Stovik (2002).

We consider a number of empirical applications. First, in the context of LAD estimation we compare our optimal filters with the approximation nonlinear filters of Le, Martin, and Raftery (1996) and Gordon and Smith (1993) using simulated data. This allows us to compare the particle filtering approach with analytical approximations. The approximate nonlinear filters require known parameters, whereas our approach allows for sequential learning of both parameters and states. As a handicap, we assume the parameters are known when using the Le, Martin, and Raftery (1996) for state filtering, but assume both the parameters and states are unknown when using our approach. In simulated data, our approach outperforms these nonlinear approximations by a wide margin.

Next, we consider an example of an autoregressive model without latent states and an application to U.S. short term interest rates that was considered by Koenker and Xiao

(2006) in the context of quantile estimation. In this model, we document that the posterior distribution of the static model parameters vary substantially over time. In order to capture that variable, we extend the model to consider quantile state filtering and parameter learning using short term US interest rates. In the context of this model, in addition to sequentially learning and parameters and state variable, we also compare our algorithm to Storvik's (2002) algorithm, an alternative algorithm for sequentially learning parameters and states. We show that our algorithm is more efficient, as it generates a greater effective sample size.

Another advantage of our particle filtering approach is that it provides Monte Carlo estimates of marginal likelihoods. Unlike MCMC which computes a single marginal likelihood for the entire dataset, these marginal likelihoods can be computed sequentially as new data arrives. This provides a mechanism to perform sequential "model monitoring," as discussed, for example, by West (1984). As an example of this, we compare L^p-norm estimation to LAD and normal errors. As in the previous cases and using simulated data, our algorithm efficiently learns the parameters in the L^p-norm or bridge estimator setting. In terms of sequential model choice, we simulate data assuming LAD errors, and the sequential likelihood ratios can successfully discriminate the LAD case from the case of L^p and normal errors.

The rest of the paper is outlined as follows. Section 2 develops the models and algorithms for robust state and parameter learning. Section 3 provides simulated and real data examples. Finally, Section 4 concludes.

2 Robust Particle Filtering and Sequential Learning

2.1 Robust state space models

Consider the following class of state space models:

$$y_{t+1} = Fx_{t+1} + \sigma\eta_{t+1} \tag{1}$$

$$x_{t+1} = Gx_t + \sigma_x\eta^x_{t+1}, \tag{2}$$

where y_{t+1} is observed, x_{t+1} is a persistent latent state variable, η_{t+1} and η_{t+1} are errors, and the parameters are *not* assumed to be known. Thus, we are interested in learning the state variables and the parameters sequentially through time as new data arrives.

We assume that either or both of the errors have a check exponential or asymmetric Laplacean distribution, which is indexed by τ and defined by the density function

$$\mathcal{CE} : p(x) = \mu_\tau \exp(-2\rho_\tau(x))$$

where $\rho_\tau(x) = \frac{1}{2}|x| + \left(\tau - \frac{1}{2}\right)x$ and $\mu_\tau = \tau(1-\tau)$. LAD or Double-exponential ($\mathcal{DE}$) errors corresponds to the case of $\tau = 1/2$. We will show below in Theorem 1 that check-exponential random variables can be expressed as a scale mixture of normals distributions.

Different assumptions about the error distributions generate different robust estimation criterion. For example, it is well known that there is mapping between LAD estimation and likelihood estimation assuming $\mathcal{DE}$ errors, and similar mapping between quantile estimation and the likelihood estimation assuming asymmetric Laplacean errors (Koenker and Portnoy, 1999). A novel feature of our framework is that the choice of filtering robustness to observation and state errors can be made separately. For example, $\eta_t \sim \mathcal{DE}$ and $\eta^x_t \sim \mathcal{CE}$ corresponds to an LAD estimation criterion for the observation equation and quantile estimation criterion for the state equation. This flexibility is new, as it provides researchers with methods to learn about different aspects of the distributions of unknown states and parameters.

To see the connections more clearly, we can express the conditional densities generated by the observation and state equations as

$$p\left(y_{t+1}|\theta, x_{t+1}\right) = \sigma^{-1}\mu_\tau \exp\left(-\frac{2}{\sigma}\rho_\tau\left(y_{t+1} - Fx_{t+1}\right)\right)$$
$$p\left(x_{t+1}|\theta, x_t\right) = \sigma_x^{-1}\mu_\tau \exp\left(-\frac{2}{\sigma}\rho_\tau\left(x_{t+1} - Gx_t\right)\right).$$

For example, $\tau = 1/2$ generates the familiar LAD criterion

$$p\left(y_{t+1}|\theta, x_{t+1}\right) = \sigma^{-1}\mu_\tau \exp\left(-\frac{1}{\sigma}\left|y_{t+1} - Fx_{t+1}\right|\right).$$

For all of these models, the sequential inference problem is solved by $p\left(x_t, \theta|y^t\right)$, which is defined recursively via updating and prediction. Given an initial distribution of the parameters and states, $p\left(x_0, \theta\right)$, the relationship between $p(x_t, \theta|y^t)$ and $p(x_{t+1}, \theta|y^{t+1})$ are given by

$$\begin{aligned} p\left(x_{t+1}, \theta|y^{t+1}\right) &= \frac{p\left(y_{t+1}|x_{t+1}, \theta\right) p\left(x_{t+1}, \theta|y^t\right)}{p(y_{t+1}|y^t)} \\ &= \int p(y_{t+1}|x_t, \theta) p(x_{t+1}|x_t, \theta, y_{t+1}) p\left(x_t, \theta|y^t\right) dx_t d\theta, \end{aligned}$$

which uses the the fact that $p\left(x_{t+1}, \theta|y^t\right) = \int p\left(x_{t+1}, \theta|x_t, y^t\right) p\left(x_t|y^t\right) dx_t$ and by Bayes rule

$$p\left(y_{t+1}|x_{t+1}, \theta\right) p\left(x_{t+1}|x_t, \theta\right) = p\left(y_{t+1}|x_t, \theta\right) p\left(x_{t+1}|x_t, \theta, y_{t+1}\right)$$

In the models under consideration here, many of the distributions in the above relations are not known analytically. For example, neither $p(y_{t+1}|x_t, \theta)$ nor $p(x_{t+1}|x_t, \theta, y_{t+1})$ are known analytically, and the joint distribution, $p\left(x_{t+1}, \theta|y^{t+1}\right)$, is certainly not known. To generate approximate samples from these distributions, we use auxiliary variables and particle filtering methods. Auxiliary variables provide a mechanism to break an intractable distribution using a greater number of tractable distributions, and is widely used in

2.2 Scale mixture representation

The key to Bayesian inference in heavy-tailed error models is a representation of errors as a scale mixture of normals for the error distributions of η_t and η_t^x and the use of data augmentation. The simplest of the representations is for double exponential errors, as first given in Andrews and Mallows (1974). In this case, the representation implies that if $\lambda_t \sim \mathcal{E}(2)$, where $\mathcal{E}(\mu)$ denotes an exponential distribution with parameter μ, then $\eta_t = \sqrt{\lambda_t}\varepsilon_t$ has a double exponential distribution. With LAD errors in both the observation and state equation, this implies that the state space model is given by

$$y_{t+1} = Fx_{t+1} + \sigma\sqrt{\lambda_{t+1}}\varepsilon_{t+1} \tag{3}$$

$$x_{t+1} = Gx_t + \sigma_x\sqrt{\omega_{t+1}}\varepsilon^x_{t+1}, \tag{4}$$

where λ_{t+1} and ω_{t+1} are independent $\mathcal{E}(2)$ random variables and ε_{t+1} and ε^x_{t+1} are independent standard normal random variables. Given the scale mixture representation, the model becomes a conditionally Gaussian state space model. Carlin, Polson and Stoffer (1992) used this representation and MCMC simulation for smoothing problems.

Quantile errors are more challenging than LAD errors due to the kink in the likelihood. To develop the particle filtering algorithms, we need the scale mixture representation in the following theorem provides the requisite scale mixture representation

Theorem : *The check exponential errors are a mixture of normals where*

$$\eta_{t+1} = (1-2\tau)\lambda_{t+1} + \sqrt{\lambda_{t+1}}\varepsilon_{t+1},$$

and if $\lambda_{t+1} \sim \mathcal{E}\left(\mu_\tau^{-1}\right)$, *then* $\eta_{t+1} \sim \mathcal{CE}$.

Proof: From the representation of the double exponential distribution as a scale mixture of normals (see Andrews and Mallows, 1974),

$$\int_0^\infty \frac{\exp\left(-\frac{y_{t+1}^2}{2\sigma^2\lambda_{t+1}} - \frac{\lambda_{t+1}}{2}\right)}{\sigma\sqrt{2\pi\lambda_{t+1}}} d\lambda_{t+1} = \frac{1}{\sigma}\exp\left(-\left|\frac{y_{t+1}}{\sigma}\right|\right).$$

For check-exponential $\mathcal{CE}$ errors, first multiply through by $\mu_\tau e^{-(2\tau-1)\frac{y}{\sigma}}$ with $\mu_\tau = 2\tau(1-\tau)$ giving

$$\int_0^\infty \frac{\mu_\tau \exp\left(-\frac{y_{t+1}^2}{2\sigma^2\lambda_{t+1}} - (2\tau-1)\frac{y_{t+1}}{\sigma} - \frac{\lambda_{t+1}}{2}\right)}{\sigma\sqrt{2\pi\lambda_{t+1}}} d\lambda_{t+1} = \frac{\mu_\tau}{\sigma}\exp\left(-\left|\frac{y_{t+1}}{\sigma}\right| - (2\tau-1)\frac{y_{t+1}}{\sigma}\right).$$

Completing the square,

$$\int_0^\infty \frac{\mu_\tau \exp\left(-\frac{(y_{t+1}+(2\tau-1)\lambda_{t+1}\sigma)^2}{2\lambda_{t+1}\sigma^2} - \mu_\tau\lambda_{t+1}\right)}{\sigma\sqrt{2\pi\lambda_{t+1}}} d\lambda_{t+1} = \frac{\mu_\tau}{\sigma}\exp\left(-\left|\frac{y_{t+1}}{\sigma}\right| - (2\tau-1)\frac{y_{t+1}}{\sigma}\right)$$

Therefore, the asymmetric Laplace distribution or check-exponential distribution $\mathcal{CE}(0,\sigma)$ with density $\sigma^{-1}\mu_\tau e^{-\frac{2}{\sigma}\rho_\tau(y)}$ is a scale mixture of normals with mixing measure $p(\lambda_{t+1}) \sim \mathcal{E}\left(\mu_\tau^{-1}\right)$ as required. ■

This result, which is a modification of the results for double-exponential distributions, implies that the model can be equivalently written in state space form as

$$y_{t+1}^* = y_{t+1} + \sigma(2\tau-1)\lambda_{t+1} = Fx_{t+1} + \sigma\sqrt{\lambda_{t+1}}\varepsilon_{t+1} \tag{5}$$

$$x_{t+1}^* = x_{t+1} + \sigma_x(2\tau_x-1)\omega_{t+1} = Gx_t + \sigma_x\sqrt{\omega_{t+1}}\varepsilon_{t+1}^x, \tag{6}$$

where $\lambda_{t+1} \sim \mathcal{E}\left(\mu_\tau^{-1}\right)$ and $\omega_{t+1} \sim \mathcal{E}\left(\mu_{\tau_x}^{-1}\right)$. This allows for different quantiles for the observation and state equations. Parameter updating in this setting is slightly more difficult due to the fact that the mixing variable appears both in the conditional mean and the conditional variance.

Scale mixture results have been widely used in statistics, although not in either for quantile estimation or for sequential parameter and state inference in state space models. Carlin and Polson (1991) and Philips (2002) use scale mixtures and MCMC methods and the EM algorithm, respectively, for parameter estimation in models with LAD errors. Carlin, Polson, and Stoffer (1992), Buckle (1995), and Godsill and Kuruoglu (2002) have used these mixture results to develop MCMC algorithms in a non-sequential setting. None of

these papers incorporated unobserved state variables or provide sequential inference for parameters and latent state variables.

2.3 Sequential model choice

Our particle filtering algorithm provides a means to perform sequential Bayesian model choice. West (1984) is an early reference on the sequential nature of the model choice problem. Beginning with Jeffreys (1961), Bayesian model choice relies on comparing the posterior probabilities of two models. The posterior odds of model $\mathcal{M}_i$ to $\mathcal{M}_j$ are defined as

$$odds\left(\mathcal{M}_i \text{ vs. } \mathcal{M}_j|y^t\right) = odds_t^{i,j} = \frac{p\left(\mathcal{M}_i|y^t\right)}{p\left(\mathcal{M}_j|y^t\right)} = \frac{p\left(y^t|\mathcal{M}_i\right)}{p\left(y^t|\mathcal{M}_j\right)}\frac{p\left(\mathcal{M}_i\right)}{p\left(\mathcal{M}_j\right)},$$

where the priors odds ratio is $p\left(\mathcal{M}_i\right)/p\left(\mathcal{M}_j\right)$ and the Bayes factor is the likelihood ratio

$$\mathcal{BF}_{i,j}^t = \mathcal{LR}_{i,j}^t = \frac{p(y^t|\mathcal{M}_i)}{p(y^t|\mathcal{M}_j)}.$$

All of these quantities can be defined and analyzed sequentially. The Bayes factor can be recursively defined as

$$\mathcal{BF}_{i,j}^{t+1} = \frac{p(y_{t+1}|y^t,\mathcal{M}_i)}{p(y_{t+1}|y^t,\mathcal{M}_j)}\mathcal{BF}_{i,j}^t,$$

where $p(y_{t+1}|y^t,\mathcal{M}_i)$ is the predictive likelihood under model i, which is given by

$$p(y_{t+1}|y^t,\mathcal{M}_i) = \int p\left(y_{t+1}|x_{t+1},\theta,\mathcal{M}_i\right)p\left(x_{t+1},\theta|y^t,\mathcal{M}_i\right)d\left(x_{t+1},\theta\right). \tag{7}$$

The relative predictive likelihoods determine any changes to the Bayes factor. The central challenge in Bayesian model choice is one of computation: how to compute the marginal predictive likelihoods, integrating out all of the parameter and state variable uncertainty. Computing these quantities is a particular challenge in state space models, because of the time series dependence. There is a large literature developing MCMC methods for model choice.

Our particle filtering algorithms provide approximate samples from $p\left(x_{t+1}, \theta | y^t, \mathcal{M}_i\right)$, which makes it straightforward to estimate the marginal likelihoods using Monte Carlo. In fact, the marginal likelihoods are a natural by-product of the particle filtering algorithm. The models that we consider in this paper essentially differ in terms of their error terms, and the marginal likelihoods provide a mechanism for comparing the different specifications over time. We provide an application below comparing LAD, L^p-norm, and normal errors.

3 Sequential Learning Algorithms

The posterior distribution, $p\left(\theta, x_t | y^t\right)$, computed across time as new data arrives, solves the sequential learning problem. Notice that we focus on $p\left(\theta, x_t | y^t\right)$ instead of $p\left(\theta, x^t | y^t\right)$. The former distribution has a fixed dimension, whereas the dimension of the latter distribution increases over time, leading to the curse of dimensionality. This, not surprisingly, leads to more efficient particle filtering approximations.

To generate approximate samples from $p\left(\theta, x_t | y^t\right)$, we use data augmentation and additionally track the scale variables, λ_t. For notational simplicity, we denote all of the additional scale variables as λ_t, and we do not explicitly reference ω_t. These scale variables are important because they also induce a conditional sufficient statistic structure for the parameter posteriors. Sufficient statistics are defined as

$$p(\theta | x^t, \lambda^t, \omega^t, y^t) = p(\theta | s_t),$$

where superscripts denote histories up to time t, for example, $x^t = (x_1, ..., x_t)$. The key to the use of sufficient statistics is the fact that they can be recursively defined,

$$s_{t+1} = \mathcal{S}\left(s_t, x_{t+1}, \lambda_{t+1}, y_{t+1}\right),$$

conditional on the latent states and auxiliary variables. The mapping $\mathcal{S}$ is analytical and the sufficient statistic structure leads to draws from standard distributions.

The models are closed via the prior distribution, $p(\theta, x_0)$. We specify the priors in a hierarchical form via $p(x_0|\theta)\, p(\theta)$. To do this, we follow the standard procedures used in the filtering literature and initialize the state distribution using the stationary distribution of the state variables. To sample from this distribution, we note that $p(x_0|\theta) = \int p(x_0|\theta, \lambda_0)\, p(\lambda_0)\, d\lambda_0$, where $p(x_0|\theta, \lambda_0)$ is a Gaussian distribution with the unconditional mean and variance. Alternatively, the state variables could be initialized using any other distribution that can be easily sampled. For the parameters, we use standard prior distributions. Since all of the models are linear and, conditional on the auxiliary variables, Gaussian, the priors are normal-inverse Gamma:

$$p\left(F, \sigma^2\right) \sim \mathcal{N}(a_0, A_0)\, \mathcal{IG}(b_0, B_0)$$
$$p\left(G, \sigma_x^2\right) \sim \mathcal{N}(c_0, C_0)\, \mathcal{IG}(d_0, D_0).$$

Given the auxiliary scale variables, sufficient statistics, and priors, we use particle methods to generate approximate samples from $p\left(\theta, s_t, x_t, \lambda_t|y^t\right)$. At this point, notice that the dimensionality of this distribution is fixed and does not grow as more data arrives. As noted earlier, this is key for developing efficient particle filtering algorithms. The target distribution is

$$\begin{aligned} p\left(\theta, s_{t+1}, x_{t+1}, \lambda_{t+1}|y^{t+1}\right) &= p(\theta|s_{t+1})\, p\left(s_{t+1}, x_{t+1}, \lambda_{t+1}|y^{t+1}\right) \\ &= p(\theta|s_{t+1})\, p\left(s_{t+1}, x_{t+1}|y^{t+1}\right) p\left(\lambda_{t+1}|s_{t+1}, x_{t+1}, y^{t+1}\right), \end{aligned}$$

decomposing the joint learning problem into a filtering problem, learning s_{t+1}, x_{t+1} and λ_{t+1}, and a standard sampling problem, drawing from $p(\theta|s_{t+1})$. Let $H_t = (\theta, s_t, x_t)$. The key our approach is to represent $p\left(s_{t+1}, x_{t+1}|y^{t+1}\right)$ as

$$p\left(s_{t+1}, x_{t+1}|y^{t+1}\right) = \int p(y_{t+1}|\theta, x_t, \lambda_{t+1})\, p(s_{t+1}, x_{t+1}|H_t, \lambda_{t+1}, y_{t+1})\, dp\left(H_t, \lambda_{t+1}|y^t\right) \quad (8)$$

and $dp\left(\lambda, H_t|y^t\right) = dp\left(\lambda_{t+1}\right) dp\left(H_t|y^t\right)$ since λ_{t+1} is independent of the past. The distribution $p\left(y_{t+1}|\theta, x_t, \lambda_{t+1}\right)$ is known analytically and $p\left(\lambda_{t+1}\right)$ can be directly sampled. The final piece can be expressed as

$$\begin{aligned} p\left(s_{t+1}, x_{t+1}|H_t, \lambda_{t+1}, y_{t+1}\right) &= p\left(s_{t+1}|x_{t+1}, H_t, \lambda_{t+1}, y_{t+1}\right) p\left(x_{t+1}|H_t, \lambda_{t+1}, y_{t+1}\right) \\ &= p\left(s_{t+1}|x_{t+1}, s_t, \lambda_{t+1}, y_{t+1}\right) p\left(x_{t+1}|\theta, x_t, \lambda_{t+1}, y_{t+1}\right), \end{aligned}$$

where we abuse notation a bit since $p\left(s_{t+1}|x_{t+1}, s_t, \lambda_{t+1}, y_{t+1}\right)$ is a degenerate distribution. For all of the models under consideration, $p\left(x_{t+1}|\theta, x_t, \lambda_{t+1}, y_{t+1}\right)$ is normally distributed,

$$p\left(x_{t+1}|\theta, x_t, \lambda_{t+1}, y_{t+1}\right) \sim \mathcal{N}\left(\mu_{x,t+1}, \sigma^2_{x,t+1}\right),$$

where the hyperparameters are updated as

$$\frac{\mu_{x,t+1}}{\sigma^2_{x,t+1}} = \frac{Fy_{t+1}}{\sigma^2\lambda_{t+1}} + \frac{Gx_t}{\sigma^2_x\omega_{t+1}} \text{ and } \frac{1}{\sigma^2_{x,t+1}} = \frac{F^2}{\sigma^2\lambda_{t+1}} + \frac{1}{\sigma^2_x\omega_{t+1}}.$$

This representation of $p\left(s_{t+1}, x_{t+1}, \lambda_{t+1}|y^{t+1}\right)$ has the advantage that it takes the form of a standard continuous-mixture distribution, where $p\left(y_{t+1}|\theta, x_t, \lambda_{t+1}\right)$ are the weights or mixture indicators.

We approximate the integral in equation 8 using particle filtering methods, essentially utilizing Monte Carlo approximations to the integrals. We sample from a particle approach to $p\left(\theta, s_{t+1}, x_{t+1}, \lambda_{t+1}|y^{t+1}\right)$ consisting of weights, which are always equal to 1/N for our approach, and particles $\left\{(\theta, s_t, x_t, \lambda_t)^{(i)}\right\}_{i=1}^N$, denoted by $p^N\left(\theta, s_t, x_t, \lambda_t|y^t\right)$. The following algorithm provides the required samples:

Algorithm: Robust filtering and learning

Step 1: Draw $\lambda_{t+1}^{(i)} \sim p\left(\lambda_{t+1}\right)$ and $\omega_{t+1}^{(i)} \sim p\left(\omega_{t+1}\right)$

Step 2: (Resampling) Choose mixture indices via

$$z_{t+1}^{(i)} \sim Multi_N\left(\left\{w\left(\theta, x_t, \lambda_{t+1}, \omega_{t+1}\right)^{(i)}\right\}_i\right)$$

where

$$w\left(\theta, x_t, \lambda_{t+1}\right)^{(i)} = \frac{p\left(y_{t+1}|\left(\lambda_{t+1}, x_t, \theta\right)^{(i)}\right)}{\sum_{j=1}^{N} p\left(y_{t+1}|\left(\lambda_{t+1}, x_t, \theta\right)^{(j)}\right)}$$

Step 3: Propagate state variables

$$x_{t+1}^{(i)} \sim \mathcal{N}\left(\mu_{x,t+1}^{(i)}, \left(\sigma_{x,t+1}^2\right)^{(i)}\right)$$

Step 4: Propagate parameter sufficient statistics:

$$s_{t+1} = \mathcal{S}\left(\left(x_{t+1}, \omega_{t+1}, \lambda_{t+1}, s_t\right)^{(i)}, y_{t+1}\right).$$

Step 5: Update parameters: Draw $\theta^{(i)} \sim p\left(\theta|s_{t+1}^{(i)}\right)$.

The final step of parameter updating requires some discussing, as there are minor differences required for the different model specifications. For models with qunatile errors, the sufficient statistics and parameter posteriors are easy to compute. Conditional on λ_{t+1}, x_{t+1}, the models in equations 3 and 4 are linear and Gaussian, which leads to standard conjugate posterior distributions:

$$p\left(F, \sigma^2|s_{t+1}\right) \sim \mathcal{N}\left(a_{t+1}, A_{t+1}\sigma^2\right) \mathcal{IG}\left(b_{t+1}, B_{t+1}\right)$$
$$p\left(G, \sigma_x^2|s_{t+1}\right) \sim \mathcal{N}\left(c_{t+1}, C_{t+1}\sigma_x^2\right) \mathcal{IG}\left(d_{t+1}, D_{t+1}\right).$$

The hyperparameters are the sufficient statistics and are easy to compute using standard Bayesian updating in dynamic linear models and are given in Harrison and West (1981) or Johannes and Polson (2007).

For the quantile case, parameter updating is slightly more difficult. Conditional on x_{t+1} and λ_{t+1}, the models are given by

$$y_{t+1} = Fx_{t+1} + \sigma\left(1-2\tau\right)\lambda_{t+1} + \sigma\sqrt{\lambda_{t+1}}\varepsilon_{t+1} \tag{9}$$

$$x_{t+1} = Gx_t + \sigma_x\left(1-2\tau_x\right)\omega_{t+1} + \sigma_x\sqrt{\omega_{t+1}}\varepsilon^x_{t+1}, \tag{10}$$

which are somewhat non-standard, since the volatility parameters appear multiplied by the scale variable as well as the Gaussian shock. Focussing on the observation equation, we note that if $p\left(F|s_{t+1},\sigma^2\right) \sim \mathcal{N}\left(a_{t+1}, A_{t+1}\sigma^2\right)$, then F can be marginalized out of the observation equation to generate

$$y_{t+1} = \left(a_{t+1} + \sigma\sqrt{A_{t+1}}\right)x_{t+1} + \sigma\left(1-2\tau\right)\lambda_{t+1} + \sigma\sqrt{\lambda_{t+1}}\varepsilon_{t+1} \tag{11}$$

$$= a_{t+1}x_{t+1} + \sigma\left(1-2\tau\right)\lambda_{t+1} + \sigma\sqrt{\lambda^2_{t+1} + A_{t+1}}\widetilde{\varepsilon}_{t+1}, \tag{12}$$

where $\widetilde{\varepsilon}_{t+1}$ is standard normal. Re-writing this, we have that

$$\tilde{y}_{t+1} \equiv y_{t+1} - a_{t+1}x_{t+1} = \sigma\left(1-2\tau\right) + \sigma\lambda_{t+1}\sqrt{\lambda^2_{t+1} + A_{t+1}}\widetilde{\varepsilon}_{t+1}.$$

For the scale parameters, we use a size-biased normal distribution, which generalizes the inverse gamma and has density given by

$$p(x) \sim \mathcal{SN}\left(a,b,B\right) = C\left(a,b,B\right)x^a e^{-\frac{1}{2B}(x-b)^2},$$

where C is a normalizing constant. We use this distribution as a prior for the parameter σ^{-1}. Given this, the family admits conditional sufficient statistics with a recursive updating structure as follows: the next likelihood is given by

$$\sigma^{-1}\exp\left(-\frac{\tilde{y}^2_{t+1}}{2\sigma^2 C_{t+1}}\right)$$

where $C_{t+1} = \lambda_{t+1}^2 \left(\lambda_{t+1}^2 + A_{t+1}\right)$.

If we assume a size-biased normal distribution where $p(\sigma^{-1}|s_t) \sim \mathcal{SN}(a_t, b_t, B_t)$ then we have a conjugate posterior of the form

$$p(\sigma^{-1}|s_{t+1}) \propto \sigma^{-1} \exp\left(-\frac{\sigma^{-2} y_{t+1}^2}{2C_{t+1}}\right) \sigma^{-a_t} e\left(-\frac{1}{2B_t}(\sigma^{-1} - b_t)^2\right) \sim \mathcal{SN}(a_{t+1}, b_{t+1}, B_{t+1})$$

where we have the recurrence relation

$$\begin{aligned} a_{t+1} &= a_t + 1 \\ b_{t+1} &= B_{t+1} B_t^{-1} b_t \\ B_{t+1}^{-1} &= B_t^{-1} + \frac{y_{t+1}}{C_{t+1}}^2 \end{aligned}$$

can be used to update the scale σ. As a prior, we use the traditional inverse gamma density ($b = 0$). A number of comments are warranted.

- The algorithm provides an exact or i.i.d. sample from p^N. This can be contrasted with the typical approach using importance sampling to generate an *approximate* sample from p^N, as in Storvik (2002) and Fearnhead (2002). This is important because importance sampling has inherent degeneracies, as reviewed in Li, Bengtsson, and Bickel (2006). In this regard, the algorithm can be viewed as an extension of Pitt and Shephard's (1999) "fully-adapted" auxiliary particle filtering algorithm to the case of parameter uncertainty. In the case of known parameters, Pitt and Shephard (1999) show that in many models, it is possible to generate particle filtering algorithms that provide i.i.d. samples from p^N, without using importance sampling, resulting in a substantial improvement of the efficiency of the algorithms. In our numerical example below, we show that the algorithm given above outperforms Storvik's (2002) algorithm, the only competing algorithm for models of this form that has been developed in the literature.

- A key to the algorithm is the initial re-sampling using the predictive distribution, $p\left(y_{t+1}|\lambda_{t+1},\omega_{t+1},x_t,\theta\right)$. This implies that only high-likelihood states, parameters, and sufficient statistics are propagated forward. This is in the opposite order of traditional particle filtering algorithms (following Gordon, Salmond, and Smith, 1993) that first propagate and then resample using importance sampling. Storvik (2002) is an example of the traditional approach applied to parameter learning. These traditional algorithms can perform poorly because they do not use the information in the new observation, y_{t+1}, when propagating.

- The algorithm is related to the resample-move algorithms. They developed a particle filtering algorithm to sample from $p\left(x^t|y^t\right)$, in the case with no additional auxiliary variables, and then used an MCMC step to generate samples from $p\left(\theta|x^t,y^t\right)$, to generate approximate samples from $p\left(\theta,x^t|y^t\right)$. Our approach focuses on the filtering distribution, $p\left(\theta,x_t|y^t\right)$, and utilizes sufficient statistics. Fearnhead (2002) and Storvik (2002) provide formal algorithms incorporating sufficient statistics.

- After Step 5, it is possible to include an extra 'MCMC step' to replenish the auxiliary state variables. This requires the distributions $p\left(\lambda_{t+1}|\theta,x_{t+1},y_{t+1}\right)$ and $p\left(\omega_{t+1}|\theta,x_{t+1},y_{t+1}\right)$, which are both known. For the case of LAD, the first is given by

$$\begin{aligned} p\left(\lambda_{t+1}|\theta,x_{t+1},y_{t+1}\right) &\propto p\left(y_{t+1}|x_{t+1},\lambda_{t+1}\right)p\left(\lambda_{t+1}\right) \\ &\propto \lambda_{t+1}^{-\frac{1}{2}}\exp\left(-\frac{1}{2\lambda_{t+1}}\left(\frac{y_{t+1}-Fx_{t+1}}{\sigma}\right)^2-\frac{\lambda_{t+1}}{2}\right) \\ &\sim \mathcal{IN}\left(\sigma\left|y_{t+1}-Fx_{t+1}\right|^{-1},1\right), \end{aligned}$$

  $\mathcal{IN}$ is the inverse Gaussian distribution. The case of updating ω_{t+1} is similar.

- Finally, the efficiency of the algorithm will depend on the effective sample size (ESS) which in turn is related to the variance of the re-sampling weights. In practice, the

efficiency of our approach depends on the variance of the weights. In our algorithm, the weights are proportional to $p(y_{t+1}|x_t, s_t, \theta)$. In Storvik's algorithm, the weights are proportional to $p(y_{t+1}|x_{t+1}, s_{t+1}, \theta)$ and since

$$p(y_{t+1}|x_t, s_t, \theta) = E\left[p(y_{t+1}|x_{t+1}, s_{t+1}, \theta)\,|s_t, x_t, \theta\right],$$

our weights have lower variance using the Rao-Blackwellization result. Since the effective sample size, ESS, is defined as N divided by the variance of the weights, our algorithm has a higher effective sample size. We quantify the differences in examples below.

3.1 Alternative Algorithms

Our setting is novel because the specification allows robust errors in the observation equation, the state equation, or both. Thus, we cover the additive outliers (fat-tailed observation noise) and innovations outliers (fat-tailed state noise) cases discussed by Martin and Raftery (1987). Martin and Raftery argue that the case with outliers in both equations is central in many applied fields, and is important for both estimation and forecasting. Most of the previous filtering approaches in the literature, which rely on analytical approximations, follow Masreliez (1975) and allow for robust errors in one equation, but not both. Kitagawa (1987) provides an alternative approximate filter using local linear approximations to filtering densities in robust settings. Meinhold and Singpurwalla (1987) consider a robust state filtering setting with t-distributed errors.

Our modeling framework departs from the existing literature in two important ways. First, we introduce the concept of quantile filtering in state space models, and allow for more flexible error distributions. Second, we assume the parameter vector, $\theta = (F, G, \sigma, \sigma_x)$, is unknown. All of the above mentioned robust filtering papers do not consider the problem of sequential parameter learning. Most of these filters require parameters to be known to derive

analytical approximations, which implies that our setting requires a new methodological approach. Koenker and Hallock (2001) and Yu, Lu, and Stander (2003) provide recent reviews of quantile parameter estimation.

We compare our approach to existing filters. The first is the approximate nonlinear filter of Le, Martin and Raftery (LMR, 1996), which is based on Masreliez (1975) and Masreliez and Martin (1977). They assume that the state noise is Gaussian and observation noise is non-Gaussian. The filter approximates the score function of the non-Gaussian observation noise via Huber's monotonic and Hempel's re-descending function in a manner similar to Martin (1979). If we assume the density of x_t given y^{t-1} is multivariate normal with mean vector $\hat{x}_{t-1}$ and covariance matrix M_t, then the approximate filter's estimate of x_t is given by

$$\hat{x}_t = F\hat{x}_{t-1} + M_t G^T \Psi_t(y_t),$$

where $M_{t+1} = FP_tF^T + Q_t$, $P_t = M_t - M_t G^T \Psi'(y_t) G M_t$ and

$$\Psi_t(y_t) = -\left(\frac{\partial}{\partial y_t}\right) \log f_Y(y_t|y^{t-1}).$$

Here Q_t is the known covariance of ε_t^x. The key property is the approximation of $\Psi_t(y_t)$ via Huber's monotonic and Hempel's re-descending score function.

In contrast, our approach allows for non-normal errors in both the observation and the state equation. For comparison purposes, we consider Gaussian state errors and LAD observation errors along the dimension of the mean-squared error of the state estimates. For the LMR filter these are defined as the minimum values obtained using the optimal choice of a, α and c, the parameters indexing the score functions in their specification.

Another approximate nonlinear filter based on Masreliez (1975) is developed in West (1981). Here the posterior distribution is approximated via a Taylor series expansion of log-likelihood, under the assumption that the observation density error is symmetric and unimodal at zero and twice-piecewise differentiable. This algorithm requires the Hessian

of the log-likelihood function $\Psi'(\cdot)$ to be always non-negative and so we truncate $\Psi'(\cdot)$ at zero. When $F = G = 1$, the filter is

$$\hat{x}_t = \hat{x}_{t-1} + C_t \Psi_t (y_t - \hat{x}_{t-1})$$
$$C_t^{-1} = (C_{t-1} + W_t)^{-1} + \Psi_t'(y_t - \hat{x}_{t-1})$$

This can be used in the case of t-distributed or Cauchy errors. Gordon and Smith (1993) point out that this filter is often unstable in many applications. To resolve this problem, they introduce the notion of exceedance number and curve. To ensure modal consistency, we keep the exceedance number below one and the maximum exceedance curve of zero.

We also compare our algorithm to Storvik's (2002) algorithm. Storvik implements an algorithm that (a) blindly propagates states; (b) resamples the states and old parameters and then (c) updates parameters. For comparison purposes, we consider an extension of Storvik's algorithm that incorporates a look-ahead auxiliary-particle filtering step for state propagation. For state filtering, Kuruoglu et al. (1998) and Kuruoglu (2002) discuss state filtering for L^p-norm errors, but do not discuss sequential parameter learning.

4 Empirical Results

This section provides our empirical results using simulated and real data. We consider a simulation study to compare our algorithm to the existing algorithms discussed in the previous section in two robust settings: LAD and Cauchy errors. We also incorporate sequential parameter learning. We consider two real data examples of quantile state filtering and sequential parameter estimation to show the flexibility of our approach. We consider sequential parameter estimation of the QAR(1) model of Koenker and Xiao (2004, 2006) and a state space extension with latent variables, analyzed using US interest rate data. Finally, we analyze sequential learning in the case of L^p-norm errors and sequential model choice comparing L^p-norm, LAD, and Gaussian error distributions.

4.1 LAD Filtering and Parameter Learning

We simulate data from a model with LAD observation noise and Gaussian state noise:

$$y_t = Fx_t + \sigma\sqrt{\lambda_t}\varepsilon_t$$

$$x_t = Gx_{t-1} + \sigma_x\varepsilon_t^x,$$

where $\varepsilon_t^y, \varepsilon_t^x$ are i.i.d. standard normal and $\lambda_t \sim \mathcal{E}(2)$. We simulate 100 datasets with $T = 500$ assuming $F = 1, G = 0.95, \sigma = \sigma_x = 1$. The priors are

$$F|\sigma^2 \sim \mathcal{N}(1, 0.05\sigma^2),\ \sigma^2 \sim \mathcal{IG}(5, 4),$$

$$G|\sigma_x^2 \sim \mathcal{N}(0.95, 0.025\sigma_x^2),\ \sigma_x^2 \sim \mathcal{IG}(5, 4).$$

We compute two versions of our robust particle filter, one assuming all parameters are unknown, and another assuming only the scale parameters are known. This provides a particularly stringent comparison, as our algorithm must estimate parameters in addition to the states. We use a particle size of $N = 30,000$.

Table 1 documents that our LAD particle filter outperforms the alternative algorithms, even if we assume that some of the parameters are unknown. For example, the MSE of our filter with all parameters unknown is 0.9437, compared to 0.9763 for the LMR (1996) approximate filter, which assumes the parameters are known. Also note the ranking of the approximate algorithms as LMR (1996) outperforms Gordon and Smith (1993), which in turn outperforms West (1981). This occurs despite the fact that the competitor algorithms know the true parameters and is remarkable. Our particle filtering algorithm more efficiently adapts to new data based on the exact structure of the model, instead of an approximate structure, provide large efficiency gains.

Our robust filter with parameter learning generates parameter estimates, summarized in Table 2 and Figures 1. Table 2 indicates that we accurately estimate all of the parameters, despite the extreme non-normality generated by the LAD errors. Figure 1 provides a

		LMR		West	GS
	Particle Filter	Huber′s monotonic	Hempel's re-descending		
F, G, σ_x, σ unknown	0.9437	—	—	—	—
F,G known σ_x,σ unknown	0.9295	—	—	—	—
All parameters known		0.9763	1.1086	1.4314	1.1607
		(0.0096)	(0.0125)	(0.0162)	(0.0130)

Table 1: Mean squared error of filtered state variables in a model with LAD noise. To optimize the result of alternative models, we set $a = 1.5$ for Huber's and $\alpha c = 2$, $c = 6$ for Hempel's. Numbers in parenthesis are standard error for mean of MSEs from 100 time-series of length 500.

Parameter	True value	Posterior median	2.5%	97.5%
F	1	1.0165	0.8653	1.1227
G	0.95	0.9490	0.9164	0.9773
σ^2	1	0.9868	0.7607	1.2296
σ_x^2	1	1.0133	0.6957	1.4511

Table 2: Parameter posterior summaries in a model with LAD noise. The parameter posteriors correspond to the full sample posteriors computed at the last data points.

graphical summary of the posterior distribution and sequential parameter estimates for one representative sample. The right-hand panels show that the parameter posteriors at the end of the sample are extremely non-normal, even multi-modal. For example, the posterior $p(F|y^T)$ in the upper right panel shows clear multi-modality. This contrasts strongly with models with normally distributed errors, which generate posteriors that are approximately normal for these sample sizes. The other parameters also have non-normal posteriors.

We also provide the sequential parameter learning plots for $t = 1, \ldots, T$ in the left-hand panels of Figure 1. Despite the fat-tailed errors in the observation equation, sequential parameter learning proceeds quite smoothly, as these errors are properly accounted for. Note, there are a few minor spikes; for example, just after time period 250 in $p\left(F|y^t\right)$. This is due to a large observation and not Monte Carlo error. We conclude: our approach outperforms existing filters even if we assume parameters are unknown and they condition on the true parameters and accurately estimates parameter sequentially.

Table 3 provides a summary of the effective sample size for our algorithm, compared to the original Storvik algorithm and an extension discussed earlier using the auxiliary particle filter for state updating. The effective sample size for our algorithm is more than twice the ESS of Stovik's original algorithm and almost twice that of the extension of Stovik's

algorithm using the APF. This shows that our algorithm more effectively generates samples from the particle approximation.

4.2 Sequential Quantile Filtering and Learning

We now consider a real data example using two different models both motivated by Koenker and Xiao (2004). The first considers the realistic setting of sequentially learning about static parameters in the setting using a model of short term interest rates with a potential unit root. Short term rates are highly persistent, but also have highly non-normal innovations motivating a robust estimation setting. Due to this, many continuous-time models of short term interest rates include randomly arriving jumps (Johannes, 2004) to capture the large movements and to generate the non-normality.

The quantile-autoregression QAR(1) model of Koenker and Xiao (2004) is

$$r_t = \alpha + \beta r_{t-1} + \sigma \eta_t$$

where $\eta_t \sim \mathcal{CE}$. We used the following priors: $p\left(\alpha|\sigma^2\right) \sim \mathcal{N}(0.096, \sigma^2)$ and $p\left(\beta|\sigma^2\right) \sim \mathcal{N}(0.98, \sigma^2)$. For this initial experiment, we fix $\sigma = 0.5$, which is motivated by the fact that the marginal posterior distribution on the parameters is the quantile objective function, when we constrain $\sigma = 0.5$. This matches exactly the common quantile objective function (see, e.g., Chernozhukov and Hong, 2003). We have also learned this parameter and the results are qualitatively unchanged.

In our setting, inference is summarized by the posterior distribution which is identical regardless of whether there is a unit root or not. This can be contrasted with classical inference for unit roots based on asymptotics. Sims (1988) discusses the differences between Bayesian and classical approaches to unit root inference. Another advantage of our approach is that our particle samples provide a direct estimate of parameter standard errors via the posterior standard deviation.

	Storvik	Storvik+APF	New PF
Mean of effective	374	439	861
particle size	(10)	(11)	(11)

Table 3: Effective particle size of pure filtering with fixed parameters. Numbers in parenthesis are standard errors for the mean of effective particle size from 100 time-series of length 200 with LAD errors. Physical particle size is 1000

Sequential learning in models with potential unit roots is difficult as each additional observation provides little new information, given the high level of persistence. Here, the experiment is to track the views of an investor who is computing the posterior distribution of the parameters at each data point, $p\left(\alpha, \beta | r^t\right)$, and who monitor how they learn over time. In particular, we are interested in how their perceptions change in the volatile period of the late 1970s and early 1980s, and also how their views vary across error specification (quantile vs. normal). The data we use is the same as in Koenker and Xiao (2006) and is described in detail there.

We consider two cases. In one case, we assume that the errors are normal, and in the other we assume the errors are given by the quantile specification for various values of τ. Figure 2 summarizes the results. In the top panel, we display the posterior parameter distribution at the end of the sample, $p\left(\alpha | r^T\right)$ and $p\left(\beta | r^T\right)$. These figures closely match those in Koenker and Xiao (2004), but will not be identical due to the fact that we assume all of the parameters are unknown and $p\left(\alpha | r^T\right)$ and $p\left(\beta | r^T\right)$ integrate out the uncertainty in the other parameter.

Consider the sequential learning plots given in the bottom panel of Figure 2. The solid lines summarize $p\left(\alpha, \beta | r^t\right)$ in the quantile setting (for simplicity we consider the LAD case, assuming $\tau = 0.5$) and the shaded lines summarize $p\left(\alpha, \beta | r^t\right)$ for the case of normal errors

for each time point t. There are two main notables. First, for both specifications, there is substantial learning over time. This dispels the argument that outlying observations are responsible for the substantial variation in parameter estimate, as our $\tau = 0.5$ specification will heavily down-weight these data points. Thus, even properly accounting for large observations, there is substantial sequential variation in parameter estimates.

Second, the parameter estimates differ, both during the sample and at the end. For example, a comparison of the bottom panels in Figure 4 shows that in the mid-1970s, the models, not surprisingly, behaved differently when very large interest rate changes arrived. When large shocks arrived, the $\tau = 0.5$ quantile specification assumes many of the large shocks are due to the errors, while the Gaussian model has trouble fitting these observations. The net result is that the Gaussian specification imparts more mean-reversion (β is lower) than in the quantile specification.

At the end of the sample cumulative effect of is quite severe: while the Gaussian model is statistically different from a unit root the least absolute deviation coefficient is not. The shaded $(5\%, 95\%)$ band for normal errors does not contain $\beta = 1$, while the bands for the quantile case with $\tau = 1/2$ contain $\beta = 1$. The posterior means for β are just slightly less than 1 for $\tau = 1/2$, but are about 0.98 for normal errors. Thus, inference changes on the unit root behavior. This shows how sensitive unit root conclusions are to the assumption of normal errors. The top right panel shows the posterior means for β for various quantiles. The quantile approach sheds light on the complicated nature of unit roots in interest rates: for $\tau = 0.25$, the estimate of β is about 0.96, but for $\tau = 0.75$ the estimate is over unity.

Finally, to illustrate quantile filtering of state variables, we consider the same interest rate data but allow for an extension of the QAR(1) model of Koenker and Xiao (2006) to

	Blind	Storvik+APF	J-P PF
Mean of effective	$3,638$	$4,751$	$6,941$
particle size	(158)	(176)	(139)

Table 4: Effective particle size of the filtering with sequential parameter learning. Numbers in parenthesis are standard errors for the mean of effective particle size from 100 time-series of length 200 with $\mathcal{DE}$ errors. Physical particle size is $10,000$

incorporate a time-varying latent mean:

$$r_{t+1} - r_t = x_t + \sigma\eta_{t+1}$$
$$x_{t+1} = \alpha + \beta x_t + \sigma_x \eta^x_{t+1}.$$

Such time-varying means are often used in the term structure literature in finance to capture slow-moving trends. A related model is that of De Rossi and Harvey (2006), who assume that the quantiles themselves vary over time via in an autoregressive manner.

We apply the algorithm outlined in Section 3. with $N = 30,000$. Figure 3 shows the changes $r_{t+1} - r_t$ together with state filtered estimates for the following quantile values: $\tau = 0.2, 0.5, 0.8$. Given the volatility of the state it is difficult to see the differences in the state estimates, although they are clearly different. Figure 4 provides an easier way to see the differences. Here, we track the sequential parameter posteriors for $\tau = (0.2, 0.8)$, and additionally learn the parameter σ. The results are generally similar for for σ_x and β, but dramatically different for α. This occurs because different quantile criterion differentially downweighting large negative and large positive innovations. This has a dramatic impact on the long-run mean variable, α, as large movements are captured by the errors and are down-weighted for parameter inference differently for different quantiles.

Table 4 provides a comparison of effective sample sizes for the SIR algorithm, Storvik with adaption and our method for a total of $10,000$ particles. Our approach has the highest

efficient ratio of 69%, generating an effective sample size more than twice the ESS generated by Storvik's algorithm.

4.3 Robust L^p-norm learning

We also consider L^p estimation, although this is not the main focus of this paper. Robust L^p-norm criterion are based on minimizing a criterion function of the form $|\cdot|^p$. In our setting, we can incorporate an L^p-norm criterion by assuming that the error distribution is an exponential power family, whose density is given by

$$\mathcal{EP} : p\left(x\right) = \frac{\exp\left\{-\left|\frac{x}{\sigma}\right|^p\right\}}{2\sigma\Gamma(1+p^{-1})},$$

where Γ is the Gamma function. Applied to the state space model, the distribution induced by the observation equation is given by

$$p(y_t|x_t,\sigma,p) = \frac{\Gamma\left(1+p^{-1}\right)^{-1}}{2\sigma}\exp\left\{-\left|\frac{y_t - Fx_t}{\sigma}\right|^p\right\}. \tag{13}$$

This parameterization has the advantage that special cases are given by the double exponential ($p = 1$) and normal distribution ($p = 2$).

L^p norm criterion, we can adapt a scale mixture result from West (1987). The only modification from the previous results is that the distribution of the mixing variable is more complicated,

$$p(\lambda_{t+1}) \propto \lambda_{t+1}^{-3/2}\mathrm{St}^+_{\frac{p}{2}}\left(\lambda_{t+1}^{-1}\right),$$

where St^+_a it is the positive stable distribution with index a. Although this distribution is analytically intractable, it is easy to generate random samples from this distribution using the algorithm in Chambers et al. (1976).

Consider an autoregressive model with L^p-norm errors in the observation equation and

Gaussian errors in the state equation. The evolution is given by

$$y_{t+1} = x_{t+1} + \sigma\eta_{t+1}$$
$$x_{t+1} = \alpha + \beta x_t + \sigma_x \eta^x_{t+1}.$$

For an example of this model, we simulate $T = 300$ data points assuming $\alpha = 0$, $\beta = 0.9$, $\sigma_x = 0.2$, $\sigma = 0.3$, and $p = 1.2$. This generates observation errors that are between LAD and Gaussian errors. The particle algorithm is run with $N = 30{,}000$ particles.

Results of the sequential parameter estimates are in Figure 6, and they are qualitatively similar to those reported for the other specifications. Although the mixing variable has a more complicated distribution in the L^p-norm case, the algorithm performs well since it generates an exact sample from the particle filtering distribution. It is also possible to learn the parameter p, assuming a discrete support. In this case, it is possible to compute marginal likelihoods for various values of p, as discussed in the next section, which provides the posterior distribution of the discrete values of p.

4.4 Sequential model choice

As a final example, we consider sequential model choice between three competing specifications: in each case, the specification is a linear state model with Gaussian state errors, with either LAD, L^p-norm, or Gaussian observations errors. We simulate a time series of 300 data points using the structural parameters in the previous section, assuming the true model has LAD errors. For the L^p-norm errors, we chose a value of p that is close to the LAD case, $p = 1.2$.

Figure 6 provides a time series of the Bayes factors for a pointwise comparison of the models. The solid line is $p\left(L^p|y^t\right)/p\left(LAD|y^t\right)$ and the shaded line is $p\left(\text{Gaussian}|y^t\right)/p\left(LAD|y^t\right)$. Focussing on the shaded line, the Bayes factor quickly identifies the preferred LAD specification, after only a few data points. Since the true model is LAD, the first large outlying

observation generates a very small predictive likelihood, given the exponential tails of the Gaussian distribution. Future observations just confirm this, as the Bayes factor quickly converges to zero. For the case of L^p-norm errors with $p = 1.5$, the Bayes factor takes longer to distinguish between the two model specifications, but does so by about data point 50. Noticeably, between data points 200 and about 240, there is a rise in the Bayes factor, which occurs because there no large observations occur and the L^p-norm errors, for this series of observations, does a marginally better job of fitting these observations. Two large outliers arrive after this pushing the Bayes factor down to effectively zero.

In the context of these state space models, this previous example is just one of many potential hypothesis testing and model choice problems than can be performed using the marginal likelihoods. Three examples that arise in applications are testing for unit roots in the state variable, comparing the relative merits of different quantile criterion (e.g., $\tau = 1/2$ vs. $\tau = 1/3$), and learning the parameter p in the case of L^p-norm errors.

5 Conclusion

This paper provides methods for robust sequential parameter learning and state filtering. We introduce robust state space models, based on quantile and L^p-norm estimation criterion and develop algorithms for sequential inference using particle filters. This fills an important gap in the literature as robust methods (especially quantile methods) have not been widely applied in state space models.

We provide a number of comparisons to existing approximate analytical filters and particle filtering algorithms using simulated data and real data examples. We find that our algorithms outperform existing approximate filters, with the additional advantage that our approach sequentially estimates unknown parameters. Our algorithm also generates larger effective sample sizes than Storvik's (2002) algorithm. In a real data example us-

ing short term interest rates, we document the time-variation in posterior estimates in a quantile-autoregression, and examine an extension that allows for a time-varying mean. In a simulated example, we show that the algorithm efficiently learns the parameters in an L^p- norm example, and using sequential Bayes factors, the algorithm correctly identifies the true model in comparisons between LAD, L^p- norm, and Gaussian errors.

References

Andrews, D. F. and Mallows, C. L. (1974) Scale mixtures of normal distributions. *J. R. Statist. Soc. B* 36, 99-102.

Buckle, D. (1995). Bayesian inference for stable distributions, *Journal of American Statistical Association* 90, 605-613.

Carlin, B. P. and Polson, N. G. (1991). Monte Carlo Bayesian methods for Discrete Regression Models and Categorical Time Series. In *Bayesian Statistics 4* (eds Bernardo et al). Oxford University Press, 577-586.

Carlin, B.P., Polson, N.G. and 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,

Chambers, J.M., C.L. Mallows and B.W. Stuck. A Method for Simulating Stable Random Variables. *Journal of the American Statistical Association*, 71, 340-344.

Chen, R, and J. Liu (2000). Mixture Kalman Filters. *Journal of Royal Statistical Society B*, 62, 493-508.

Chernozhukov, V. and H. Hong. (2003). An MCMC Approach to Classical Estimation. *Journal of Econometrics* 115, 293-346

De Rossi, G. and A. Harvey (2006). Time-varying Quantiles. *Working Paper.*

Fearnhead, P. (2002). MCMC, Sufficient statistics and Particle Filter. *Journal of Computuational and Graphical Statistics*, 11, 848-862.

Godsill, S. J. and E. E. Kuruoglu. (2002). Bayesian inference for time series with heavy-tailed symmetric alpha -stable noise processes. CUED Tech rep INFENG.

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

Gordon, N. and A.F.M. Smith (1993). Approximate Non-Gaussian Bayesian Estimation and modal consistency. *Journal of Royal Statistical Society B,* 55, 913-918.

Huber, P.J. (1981). *Robust Statistics*. Wiley: New York.

Johannes, M. (2004). The economic and statistical role of jumps to interest rates. *Journal of Finance*, 59, 227-260.

Johannes, M. and N.G. Polson (2007). Exact Particle Filtering and Learning. *Working paper*, University of Chicago.

Kitagawa G. (1987) Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association* 94, 82: 1032–1063.

Koenker, R. and Macado, J.A.F. (1999). Goodness of fit amd related inference processes for quantile regression.*Journal of the American Statistical Association*, 94, 1296-1310.

Koenker, R. and K. Hallock. (2001) Quantile Regression: An Introduction, *Journal of Economic Perspectives* 15, 143-156.

Koenker, R. and S. Portnoy. (1997). The Gaussian Hare and the Laplacean Tortoise: Computability of Squared-error vs Absolute Error Estimators. *Statistical Science,* 12, 279-300.

Koenker, R. and Xiao, Z. (2004). Unit root quantile regression inference. *Journal of the American Statistical Association*, 99, 775-787.

Koenker, R. and Xiao, Z. (2006). Quantile Autoregression. *Journal of the American Statistical Association*, 96, 980-1006.

Kuruoglu, E.E. (2002). Nonlinear least L^p-norm filters for nonlinear autoregressive processes. *Digital Signal Processing*, 12(1), 119-142.

Kuruoglu, E.E., P. Rayner and W. J. Fitzgerald (1998). Least L^p-norm implusive noise cancellation using polynomial filters. *Signal Processing*, 69(1), 1-14.

Le, N.D., R.D. Martin and A.E. Raftery (1996). Robust Order Selection in Autoregressive Models using Robust Bayes factors. *Journal of the American Statistical Association*, 91, 123-131.

Li, B., T. Bengtsson and P. Bickel, (2006), Curse of Dimensionality Revisited: the Collapse of Importance Sampling in Very Large Scale Systems. *Working Paper*.

Martin, R.D. (1979). Robust estimation for Time series autoregressions. In: *Robustness in Statistics* (eds R.L. Launer and G.N. Wilkinson), 147-148.

Martin, R.D. and A.E. Raftery (1987) Robustness, Computation and non-Euclidean Models (comment). *Journal of American Statistical Association*, 82, 1044-50.

Masreliez, C.J. (1975) Approximate Non-Gaussian Filtering with linear state and observation relations. *IEEE Transactions on Automatic Control,* 20, 107-110.

Masreliez, C.J. and Martin, R.D. (1977) Robust Bayesian Estimation for the linear model and Robustifying the Kalman Filter. *IEEE Transactions on Automatic Control*, 22, 361-371.

Meinhold, R.J. and Singpurwalla, N.D. (1987). Robustification of Kalman Filter Models. *Journal of American Statistical Association*, 84, 479-486.

Philips, R. (2002). Least absolute deviations estimation via the EM algorithm. *Statistics and Computing*, 12(3). 281-285.

Pitt, M. and N. Shephard (1999). Filtering via simulation: auxiliary particle filter, *Journal of the American Statistical Association*, 590-599.

Sims, Christopher. (1988) Bayesian Skepticism on Unit Root Econometrics, *Journal of Economic Dynamics and Control* 12, p. 463-474.

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters, *IEEE Trans. on Signal Processing,* 50, 281-289.

West, M. (1981) Robust Sequential Approximate Bayesian Estimation. *Journal of Royal Statistical Society B*, 43(2), 157-166.

West, M. (1986) Bayesian model monitoring, *Journal of Royal Statistical Society B*, 48, 70-78.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika 75*, 646-648.

West, M. and J. Harrison. (1987). *Bayesian Forecasting and Dynamic Models*, Springer.

Yu, K., Z. Lu, and J. Stander. (2003) Quantile regression: applications and current research areas, *The Statistician 52, 331-250.*

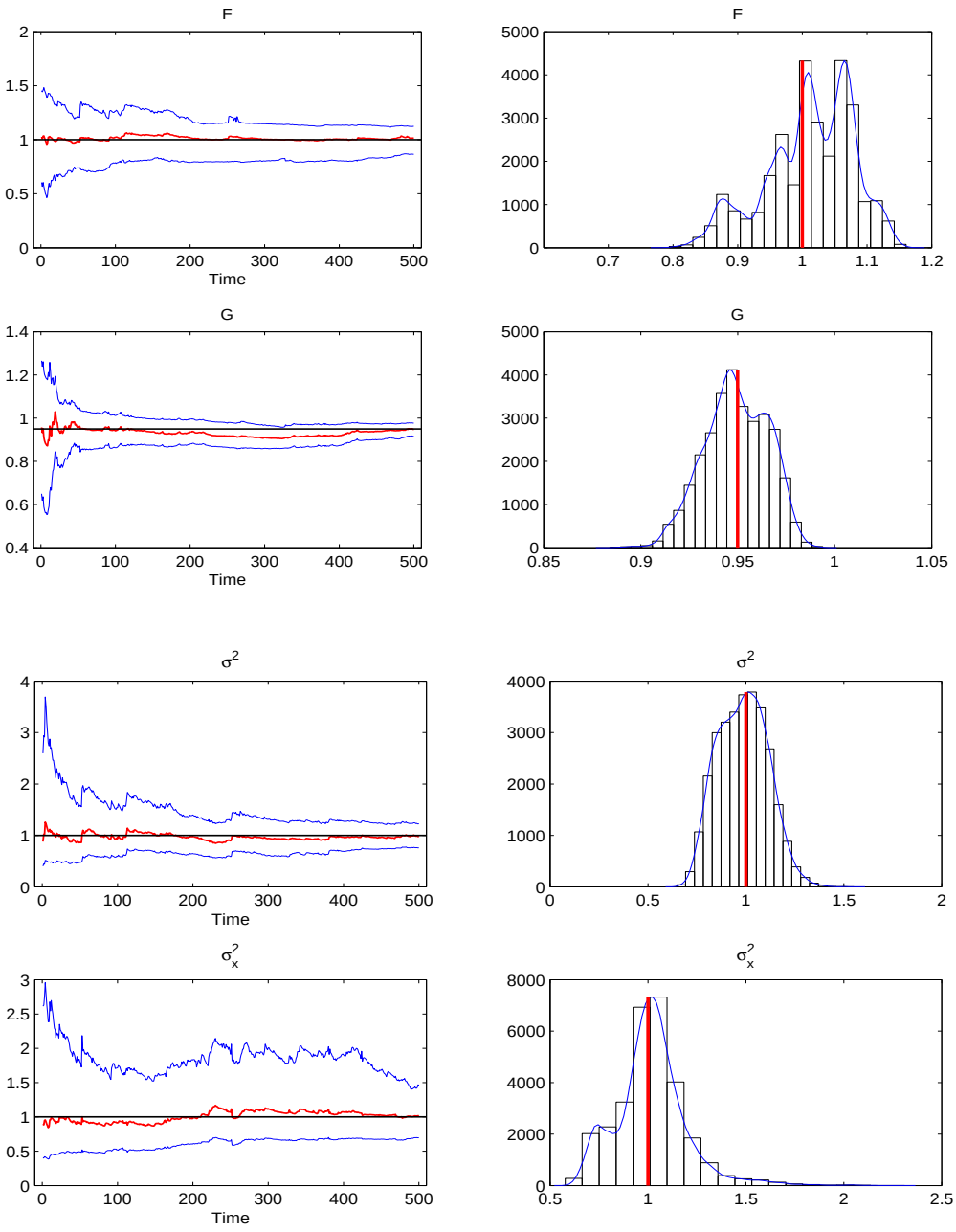

Figure 1: Left column : Parameter posterior distributions for learning and filtering with LAD errors. For each parameter, we provide a histogram (raw and smoothed) of the posteriors, as well as the true values which are given by the solid line. Right column : Sequential parameter learning. Posterior mean and 95% Bayesian confidence interval are provided sequentially.

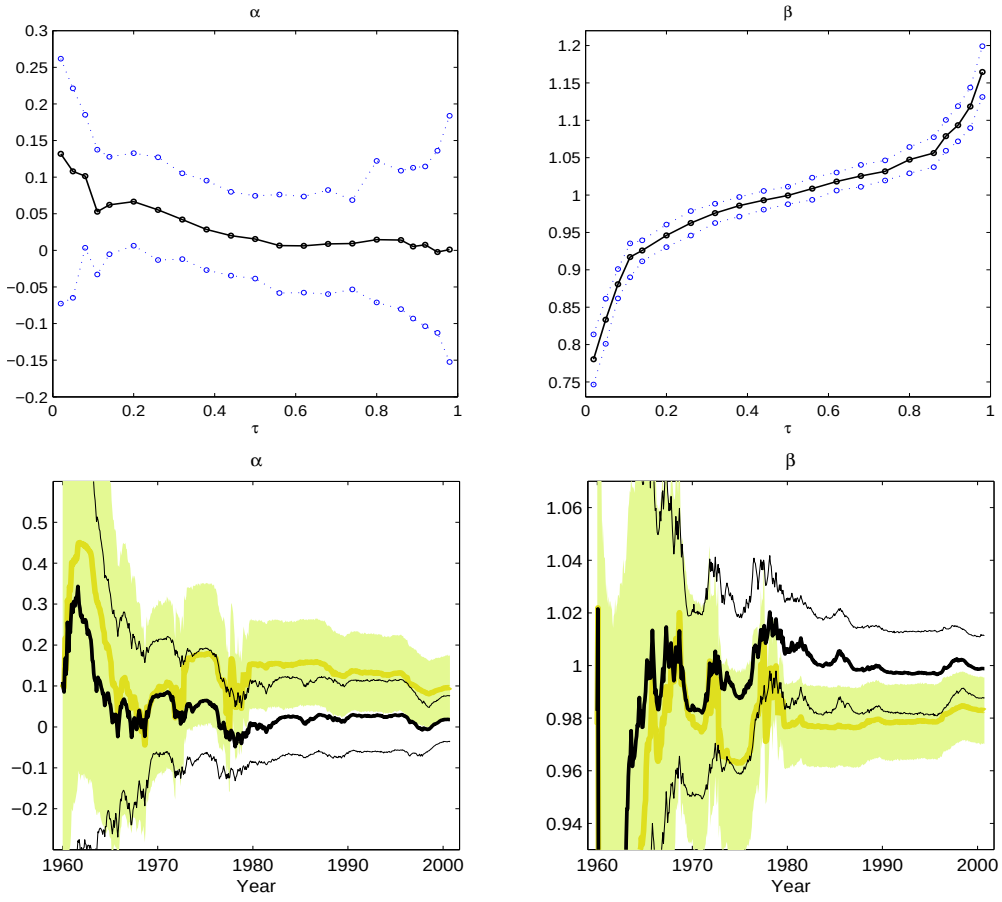

Figure 2: Top panel : Posterior parameter estimates in QAR(1) by quantile filtering for short term interest rate data. Dotted lines denote 90% BCI. Bottom panel : Comparison of sequential parameter learning between quantile filtering (solid line) and Gaussian filtering (shaded region)

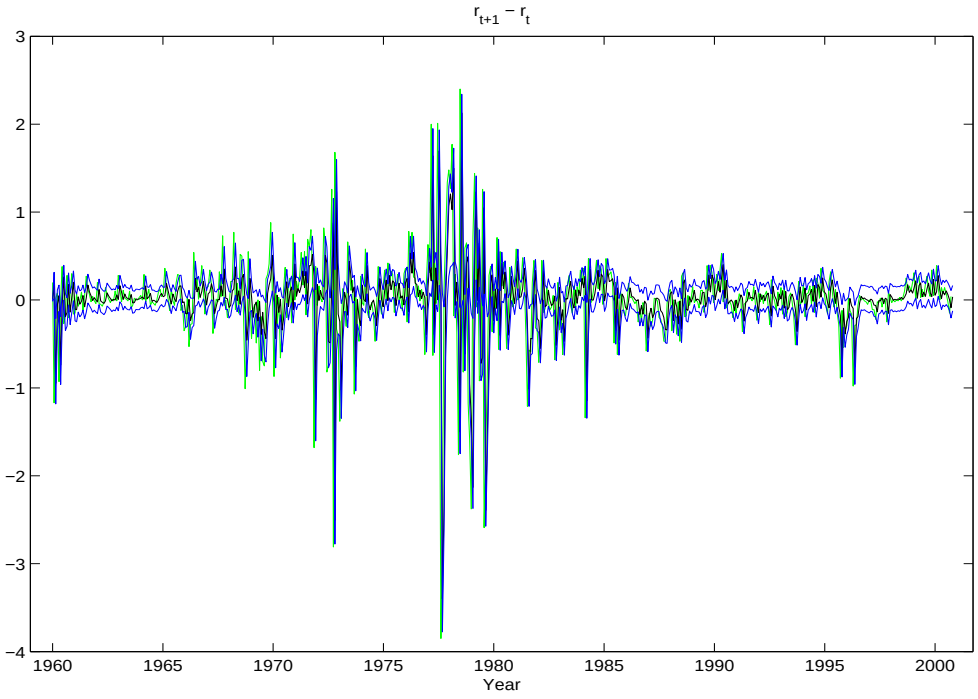

Figure 3: Filtered quantiles of state variable x_t with $\tau = 0.2, 0.5, 0.8$ in an extension of QAR model. Green line is the first difference in the observed short term interest rate.

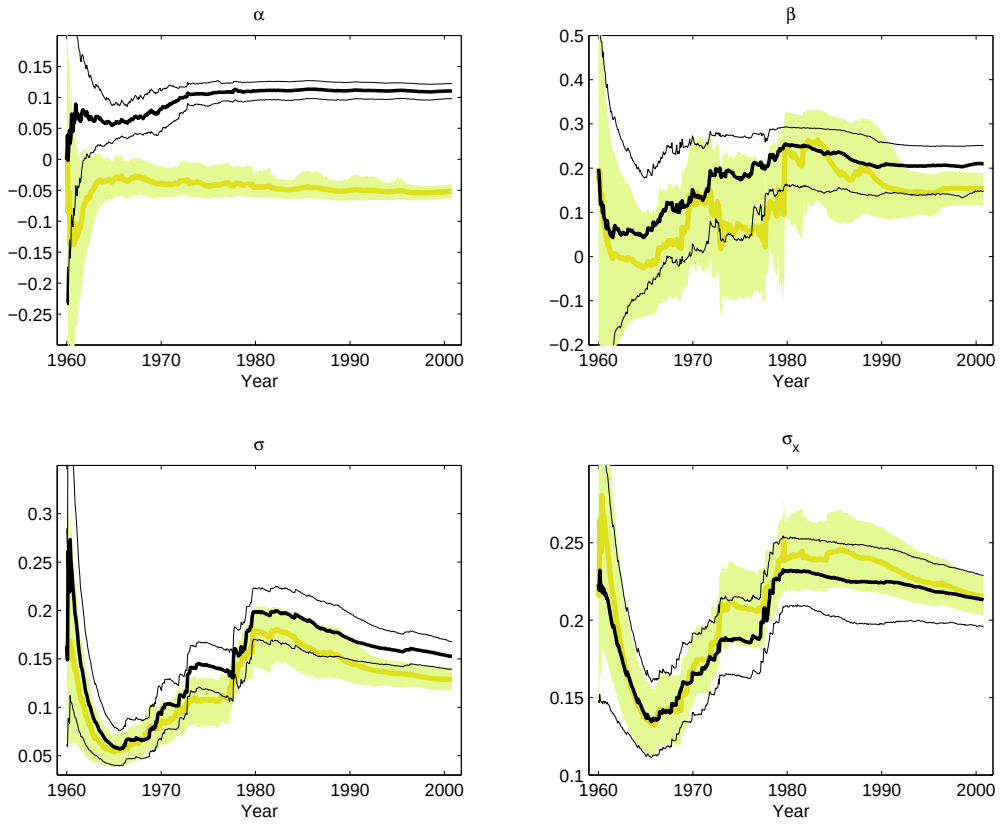

Figure 4: Comparison of sequential parameter learning between quantile filtering with $\tau = 0.8$ (solid line) and with $\tau = 0.2$ (shaded region) in an extension of QAR model. The center line denotes posterior mean and the band 90% BPR. Particle size is 20,000.

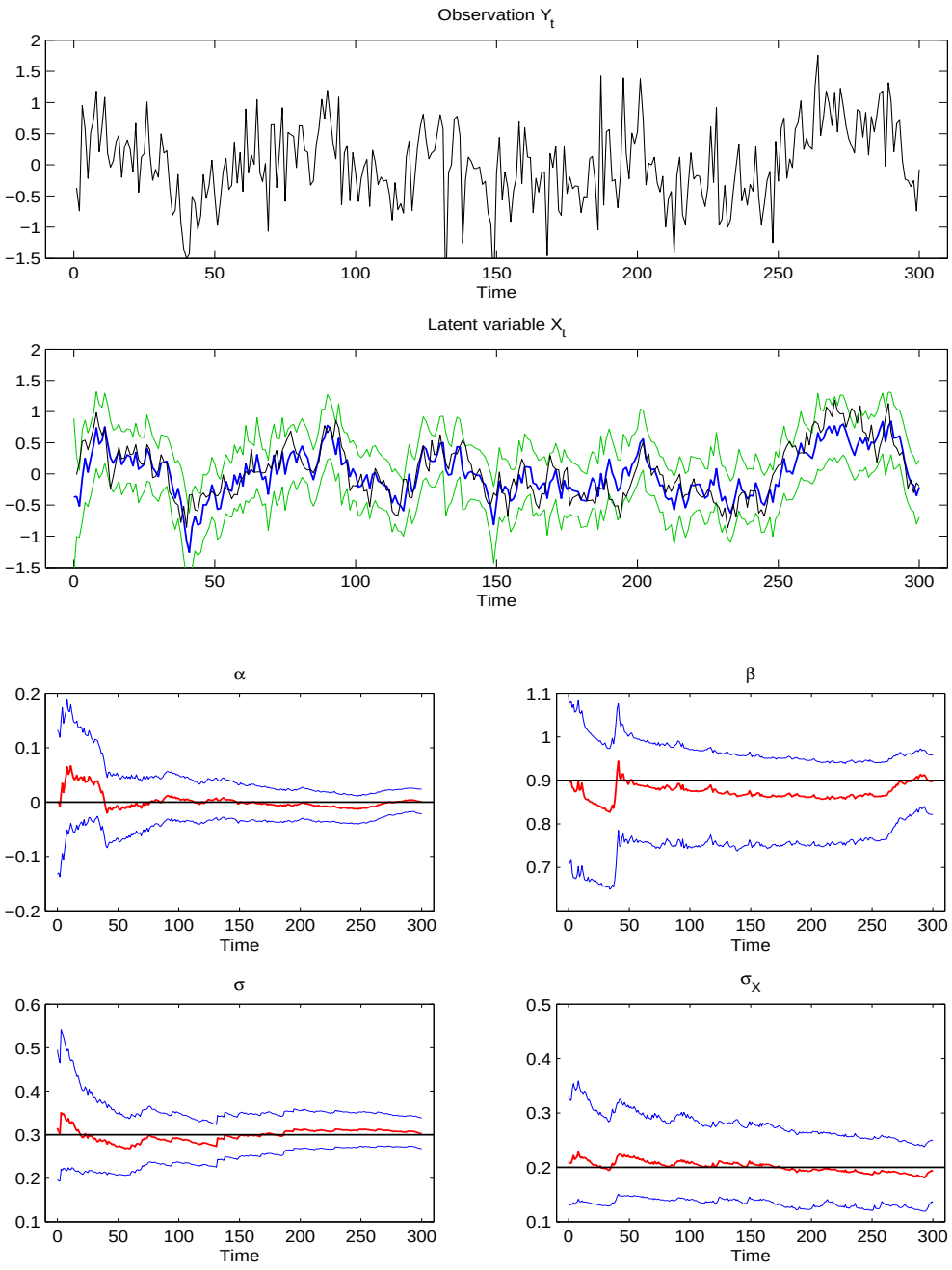

Figure 5: AR(1) with L^p norm errors in the observations. Particle size is $30k$. $p = 1.2$

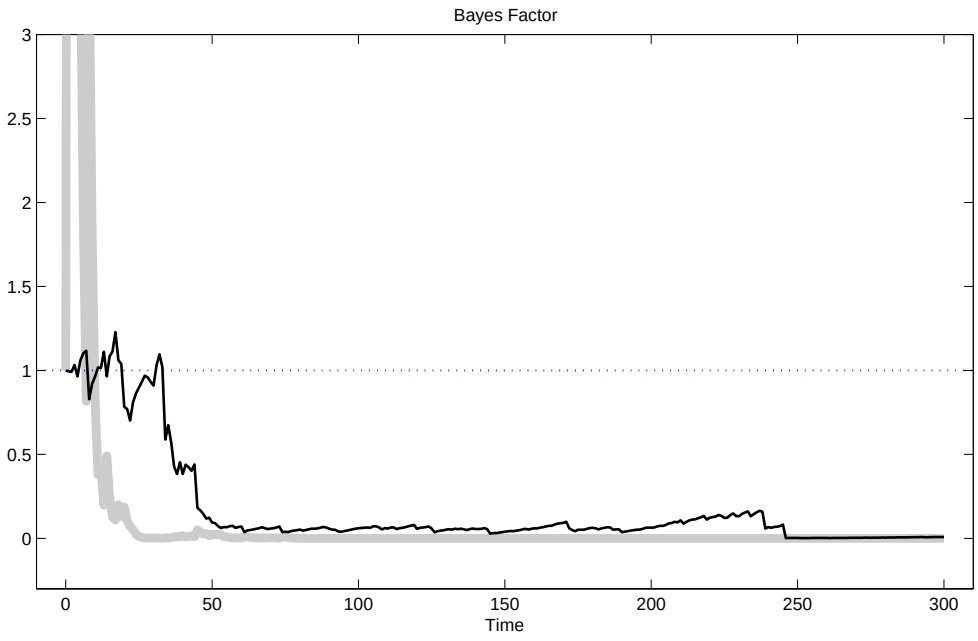

Figure 6: Sequential Bayes Factor comparison. The solid and shaded lines denote p(L_p|y)/p(LAD|y) and p(Normal|y)/p(LAD|y) respectively. Particle size is $100k$. $p = 1.5$