

Quasi Monte Carlo Kalman filter for Nonlinear and Non-Gaussian State Space Models*

Guang Zhang[†]

Department of Economics, Boston University, Boston, MA, 02215

September 10, 2020

Abstract

In this study, we present a new filtering approach for nonlinear and non-Gaussian state space models. This approach builds on the well-established Kalman filter, combined with approximations of least square linearization for nonlinear function and Gaussian mixture for non-Gaussian noises, and applies the quasi Monte Carlo method for numerical integration during computation. We compare our approach with other existing methods, such as Particle filter, using simulated data, and we show the proposed approach can outperform these methods in terms of speed and accuracy. This study also provides analysis on the stability of this new filtering approach. In addition, we propose two methods to estimate the unknown parameters in the model, and show the consistency of the proposed quasi-maximum likelihood estimator under general conditions. To illustrate the proposed approach, we discuss several numerical examples. We also introduce two applications of our approach. The first one is a popular stochastic volatility model, and we apply it to foreign exchange data between Sterling and Dollar. In the second application, we discuss a jump model, and show the jump size has a Gaussian mixture representation. We use the 3-month T-bill data to estimate the jump probability and investigate the jump sources based on macroeconomic events.

Keywords: Quasi Monte Carlo Kalman filter, nonlinear and non-Gaussian state space models, stability

JEL codes: C11, C32, C41

*I am deeply indebted to my main adviser Zhongjun Qu for his guidance and constant encouragement throughout this project. I am grateful to Iván Fernández-Val, Jean-Jacques Forneron, Hiroaki Kaido, Pierre Perron for their invaluable suggestions. I thank Undral Byambadalai, Shuowen Chen, Taosong Deng, Anlong Qin, and the seminal participants at the Boston University Economics Department, and BU-BC Econometrics Workshop. All errors are my own and comments are welcome.

[†]Email: gzhang46@bu.edu.

1 Introduction

In 1960, R. E. Kalman published his seminal paper on the linear filtering problem. Since then, due to its simplicity, optimality, tractability and robustness, Kalman filter (KF) has been applied extensively in a wide range of economics, especially in macroeconomics to estimate the dynamic structural economic models (Watson (1980), Harvey (1994), and Hamilton (1994)). Kalman filter, conceptualized as two distinct steps: “predict step” and “update step”, is an efficient recursive filter for a linear dynamic system perturbed by Gaussian noises. It is well-known that Kalman filter is the optimal linear filter in the sense of minimizing the mean square error of the estimation of unobservable state variable. However, the basic Kalman filter is limited to the linear and Gaussian assumptions on the model, and the application of the Kalman filter to the nonlinear and non-Gaussian dynamic systems is difficult.

For nonlinear systems, the most common approach is the Extended Kalman filter (EKF) developed by NASA Ames. By Taylor expansion of the nonlinear systems around the current estimation of the state variables, Extended Kalman filter is obtained based on a linear approximation of the original nonlinear system. Higher order Extended Kalman filter can be found in Jazwinski (1970) and Maybeck (1982). Extended Kalman filter has been proven to be a useful method for filtering of nonlinear systems. However, there is no guarantee of the optimality in the mean square error sense for the EKF. Also, the filter would be unstable if the assumptions of local linearity is violated.

A recent development on nonlinear filtering problem is the Unscented Kalman filter (UKF) by Julier and Uhlmann (1997). UKF is based on the unscented transformation that uses a set of appropriately chosen weighted points (called ”sigma points”) to parameterize the means and covariances of probability distributions. The details of UKF could be found in Julier and Uhlmann (1997), Wan and van der Merwe (2000) and van der Merwe and Wan (2001). UKF is based on a small set of trial points, thus its is not a truly global approximation of the nonlinear system. In addition, the application of UKF is limited to models driven by Gaussian noises.

Particle filtering (Gordon et al. (1993), Del Moral (1996, 1998)) is a simulation based technique for filtering problem and has gained in popularity in the last 30 years. Particle filtering represents the posteriori distribution of the state variable based on a set of particles. The state space model could be nonlinear and the distribution of the noises can take any form required. Variants of Particle filter include the bootstrap filter (BF), proposed by Gordon et al. (1993), the auxiliary particle filter (APF) proposed by Pitt and Shephard (1999), the Unscented Particle filter (UPF) proposed by van der Merwe et al. (2000) and the Rao-Blackwellised particle filter proposed by Doucet et al. (2000). The mathematical foundations are established in Del Moral (1996, 1998), Del Moral and Guionnet (1999ab, 2001). The application of Particle filter in economics and econometrics can be found in Kim, Shephard and Chib (1998), Flury and Shephard (2011), Creal et al. (2013),

Fernández-Villaverde and Rubio-Ramírez (2007), Fernández-Villaverde et al. (2011), and others. Lopes and Tsay (2011) provides a survey of application of Particle filter in financial econometrics. Particle filter can be applied to much wider range of dynamic systems than Kalman filter and its variants. However, Particle filter suffers from two problems: sampling degeneracy and impoverishment. In the particle propagation process, the weight will concentrate on a few particles after only several iterations and the weight of the most particles will be close to zero. This is the so-called sampling degeneracy. To address this problem, resampling procedure is suggested by Kitagawa (1996) among others. In resampling, a particle with a larger weight is more likely to be drawn and particles with very small weights are likely to be abandoned. This introduces the problem of sample impoverishment, which could negatively impact the quality of Particle filter as we are trying to represent the posteriori distribution with very few number of distinct particles.

In this paper, we propose a new approach, Quasi Monte Carlo Kalman filter (QMCKF), to nonlinear and non-Gaussian models. We divide this problem into two cases: Case 1 where the model is nonlinear and the noises are Gaussian; Case 2 where the noises to the model are non Gaussian. As we will show later with more details, our approach to Case 2 will be based on our approach to Case 1.

In Case 1, when the model is nonlinear and Gaussian, we apply the least square approach to approximate the nonlinear function in the model with a linear one. We show that the predicting distribution and filtering distribution can be approximated by a Gaussian distribution, and the mean and variance of this distribution can be derived in a closed-form way, with a simple recursive two-step format, like the Kalman filter. Essentially, this approximation can be considered as a Gram-Charlie series of the true filtering distribution with the true first two moments of the state variable. The benefits of this least square-based approximation are not only in the computation of the predicting and filtering distributions. We show that the nonlinear system in the problem can be transformed into a linear one, thus, we can apply the well-established theories on the stability of the Kalman filter to discuss the stability of our proposed nonlinear filter.

To effectively compute the true moments of state variable in the predicting and filtering distributions, we propose to use the quasi Monte Carlo method (Niederreiter (1978, 1992), Morokoff and Catflisch (1995)). This is an efficient method for numerical integration using low-discrepancy sequences, such as the Halton sequence and the Sobol' sequence.

In Case 2, when the dynamic system in the model is perturbed by non-Gaussian noises, we propose and apply the Gaussian mixture model to approximate the distribution of non-Gaussian noises. For either a linear model or a nonlinear model, we show that based on this approximation, the predicting and filtering distributions can be approximated by a mixture of Gaussian distributions. The means and variances of the Gaussian distributions are derived in a recursive two-step

way as those in Case 1, and can be computed in a closed-form way for linear model, or based on quasi Monte Carlo simulation for nonlinear models. The merit of our approach for case 2 is that we can transform the non-Gaussian model into several Gaussian models, and solve every Gaussian model simultaneously. We show this can heavily simplify the computation of the predicting and filtering distributions and achieve a faster rate of convergence.

We also discuss the stability of our proposed filtering approach for both Gaussian models and non-Gaussian models. More specifically, we show that under some mild and regular assumptions, our filter for Gaussian models is exponentially stable, and our filter for non-Gaussian models is asymptotically stable.

In the basic filtering problem of a state space model, we assume the parameters in the model are known, however, in some cases, we need to estimate the unknown parameters in the model based on the observations. To solve this problem, we propose two methods to estimate the parameters. The first approach is to construct the likelihood function based on our approximation of the filtering distribution, and then derive the maximum likelihood estimator of the unknown parameters. We show that this estimator is consistent. The second approach is to take the unknown parameters as new unobservable state variables, and augment the original model to include the parameters into a new model. Then, we apply the filtering to the new state variables and consider the mean of the state variable under filtering distribution as the estimation of the parameters.

To test the performance of our new filtering approach, we apply it to several state space models with simulated data. We compare the results of our proposed approach to the popular existing filtering such as Kalman filter and Particle filter, and we show our filtering is competitive in both accuracy and speed.

Also, we introduce two applications of our proposed filtering approach in this study. In the first application, we discuss the filtering and estimation of a stochastic volatility model based on Kim, Shephard and Chib (1998). We show that this model can be transformed to a non-Gaussian model. We use the Sterling/Dollar exchange rate data, and apply our approach to estimate the unobservable log volatility and unknown parameters in the model. In the second application, we discuss a jump model, and show that the jump size can be represented by a finite Gaussian mixture model. We further use the 3-month Treasury bill data to estimate its jump probability and jump size, and connect the jumps with major macroeconomic events.

We begin in Section 2 with a description of our approach for nonlinear and Gaussian state space models. In Section 3, we describe our filtering for nonlinear and non-Gaussian models. Section 4 discusses the stability of our filter for both Gaussian and non-Gaussian models. Section 5 provides two approaches for parameter estimation of the state space models. In Section 6, we provide various simulation examples of the state space models and examine the performance. In Section 7,

we present two applications of our proposed approach. And Section 7 concludes.

2 Nonlinear and Gaussian state space models

2.1 Examples

We first provide two examples for nonlinear and Gaussian state space models. The first example is the Quadratic Term Structure Model (QTSM) by Monfort, Renne and Rousset (2015). And the second example is the Heston's model by Heston (1993).

Example 1 (Quadratic Term Structure Models)

As an extension to the traditional affine term structure models (Duffie and Kan (1996), Duffie and Singleton (1997), and Dai and Singleton (2000, 2002)), Monfort, Renne and Rousset (2015) consider the following state space representation of the quadratic term structure model. Suppose X_t is an unobservable state variable of size n , and Z_t is an observable variable of size m . Z_t can be considered as, for example, the nominal instantaneous interest rate. The quadratic term structure model is defined as:

$$X_t = \phi + \varphi X_{t-1} + \varepsilon_t$$

$$Z_t = A + BX_t + \sum_{k=1}^m e_k X'_t C^{(k)} X_t + \eta_t$$

where ε_t and η_t are independent Gaussian white noise with unit variance matrices. e_k is the column selection vector of size m whose components are 0 except the k -th one, which is equal to 1. ϕ , φ , A , and B are parameters in the model.

A quadratic term structure framework with a component-by-component version of the measurement equation is also explored by Ahn, Dittmar and Gallant (2002). The framework designates the yield on a bond as a quadratic function of underlying state variables. They demonstrate in their paper that the quadratic term structure models can overcome limitations inherent in affine term structure models. They also test the empirical performance of the above model in determining bond prices and show that the quadratic term structure models outperform the affine term structure models in explaining historical bond price behavior in the United States. The same framework is also applied in Kim and Singleton (2012) and Doshi et al. (2013) among others.

Example 2 (Discrete-time Heston model)

Heston (1993) propose a stochastic volatility model based on the Black-Scholes (1973) model for the dynamic of stock price and derive a closed-form solution for the price of a European call option on an asset with stochastic volatility. The following is a discrete-time version of Heston's model. We denote S as the stock price and s as the return with respect to this stock. Also, we define Y as the

log volatility of stock price. The discrete-time Heston's model is defined as:

$$s_t = \ln \frac{S_t}{S_{t-1}} = \left(r - \frac{\exp(2 * Y_t)}{2} \right) + \exp(Y_t) \left[\sqrt{1 - \rho^2} W_{1,t} + \rho W_{2,t} \right]$$

$$Y_{t+1} = \beta^* + e^{-\alpha} (Y_t - \beta^*) + \gamma \sqrt{\left(\frac{1 - e^{-2\alpha}}{2\alpha} \right)} W_{2,t}$$

where W_1 and W_2 are two Gaussian noises and correlated with correlation coefficient ρ . r , β^* , γ , and α are parameters.

The above model and its continuous-time version allow arbitrary correlation between volatility and spot asset returns and have been applied in a number of literature in finance. For example, Ang et al. (2006), Bakshi, Cao, and Chen (1997), and Bates (1996) among others. The above model can be interpreted as a nonlinear and Gaussian state space model by considering s as the observable variable or measurement and Y as the unobservable state variable, and Y entering into the measurement equation through a nonlinear exponential transformation.

2.2 The Model

In this section, we consider the following state space model for $(\mathbf{X}_t, \mathbf{Z}_t)$ with respect to natural filtration $\mathcal{F}_t \equiv \sigma((\mathbf{X}_s, \mathbf{Z}_s) : s \leq t)$:

$$\mathbf{X}_{t+1} = \mathbf{F}(\mathbf{X}_t; \theta) + \boldsymbol{\varepsilon}_t \quad (1)$$

$$\mathbf{Z}_t = \mathbf{H}_1(\mathbf{X}_t; \theta) + \mathbf{H}_2(\mathbf{X}_t; \theta) \boldsymbol{\eta}_t \quad (2)$$

We call \mathbf{X} the unobservable state variable and \mathbf{Z} the observation or measurement. \mathbf{F} , \mathbf{H}_1 and \mathbf{H}_2 are nonlinear functions of the state variable with parameter $\theta \in \Theta$. We assume $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ are mutually independent random variables such that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_{\varepsilon})$ and $\boldsymbol{\eta} \sim \mathcal{N}(0, \Sigma_{\eta})$, and $\boldsymbol{\eta}_s$ and $\boldsymbol{\eta}_t$, $s \neq t$, are serially independent. In addition, we assume the initial state \mathbf{X}_0 follows a probability measure μ with finite mean μ_0 and variance Σ_0 .

Due to separability between state variable and noise and Gaussianity of noise in (1), it's obvious that the space of \mathbf{X} , which we denote as \mathcal{X} , is non-compact. The non-compactness of \mathcal{X} makes the prove of stability of nonlinear filtering complicated. When \mathcal{X} is compact, such as in regime switching model (Hamilton, 1989), the stability of nonlinear filtering is well established in Douc et al. (2004). We leave the discussion of the stability of the nonlinear filtering when \mathcal{X} is non-compact in Section 4. Note that the space of observation, \mathcal{Z} , would also be non-compact and we assume \mathbf{H}_2 to be invertible.

For nonlinear and Gaussian state space model (1)-(2), our interests are twofold. First, we want to obtain the filtering distribution of the state variable, that is $\mathbf{P}(\mathbf{X}_t | \mathcal{F}_t^Z)$, where $\mathcal{F}_t^Z \equiv \sigma(\mathbf{Z}_s : s \leq t)$, given the initial probability measure on \mathbf{X}_0 . The filtering distribution can be applied

to construct the likelihood of the observation. Second, we want to estimate $\mathbf{E}(f(\mathbf{X}_t) | \mathcal{F}_t^Z)$ given some function f on \mathbf{X} . For example, we are interested in the estimation of the first moment and second moment of the state variable based on the filtering distribution.

2.3 Filtering approach

2.3.1 Linear approximation

We denote $\mathfrak{Z}_t \equiv \{\mathbf{Z}_t, \mathbf{Z}_{t-1}, \dots, \mathbf{Z}_0\}$. Based on the model (1)-(2), the filtering distribution $p(\mathbf{X}_t | \mathfrak{Z}_t)$ and the one-step prediction distribution $p(\mathbf{X}_t | \mathfrak{Z}_{t-1})$ can be computed as

$$p(\mathbf{X}_t | \mathfrak{Z}_t) = c_t p(\mathbf{X}_t | \mathfrak{Z}_{t-1}) p(\mathbf{Z}_t | \mathbf{X}_t) \quad (3)$$

$$= c_t p(\mathbf{X}_t | \mathfrak{Z}_{t-1}) \phi\left(\mathbf{Z}_t - \mathbf{H}_1(\mathbf{X}_t), \mathbf{H}_2(\mathbf{X}_t) \boldsymbol{\Sigma}_{\eta} \mathbf{H}_2(\mathbf{X}_t)^T\right) \quad (4)$$

$$1/c_t = \int p(\mathbf{X}_t | \mathfrak{Z}_{t-1}) \phi\left(\mathbf{Z}_t - \mathbf{H}_1(\mathbf{X}_t), \mathbf{H}_2(\mathbf{X}_t) \boldsymbol{\Sigma}_{\eta} \mathbf{H}_2(\mathbf{X}_t)^T\right) d\mathbf{X}_t \quad (5)$$

and

$$\begin{aligned} p(\mathbf{X}_t | \mathfrak{Z}_{t-1}) &= \int p(\mathbf{X}_{t-1} | \mathfrak{Z}_{t-1}) p(\mathbf{X}_t | \mathbf{X}_{t-1}) d\mathbf{X}_{t-1} \\ &= \int p(\mathbf{X}_{t-1} | \mathfrak{Z}_{t-1}) \phi(\mathbf{X}_t - \mathbf{F}(\mathbf{X}_{t-1}), \boldsymbol{\Sigma}_{\varepsilon}) d\mathbf{X}_{t-1} \end{aligned} \quad (6)$$

where ϕ is the pdf of normal distribution, i.e.,

$$\phi(\nu, \Sigma) = \frac{1}{(2\pi)^{\dim(\nu)/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} v^T \Sigma^{-1} v\right).$$

If \mathbf{F} (and \mathbf{H}_1) is a linear function, then according to the theory on compound distribution, $p(\mathbf{X}_t | \mathfrak{Z}_{t-1})$ is Gaussian, thus $p(\mathbf{X}_t | \mathfrak{Z}_t)$ is also Gaussian, and then a recursive algorithm based on above two equations can be used to derive the filtering distribution and predication distribution. This is the well-known Kalman filter (Kalman, 1960; Kalman and Bucy, 1961). However, when \mathbf{F} (or \mathbf{H}_1) is nonlinear, $p(\mathbf{X}_t | \mathfrak{Z}_{t-1})$ and $p(\mathbf{X}_t | \mathfrak{Z}_t)$ are non-Gaussian even if the noises are Gaussian, and simple recursive algorithm can not be used to get the filtering and predication distribution.

Our approach is based on the idea of Halbert White (1980). In that paper, White demonstrated that a nonlinear function can be approximated by applying a least square, and the properties of this approximation were discussed. More specifically, we approximate \mathbf{F} and \mathbf{H}_1 through least square by assuming:

$$\mathbf{F}(\mathbf{X}_{t-1}) \approx \mathbf{B}_{t-1} \mathbf{X}_{t-1} + \mathbf{b}_{t-1} + \mathbf{e}_{t-1} \quad (7)$$

$$\mathbf{H}_1(\mathbf{X}_t) \approx \mathbf{A}_{1t} \mathbf{X}_t + \mathbf{a}_t + \mathbf{c}_t \quad (8)$$

where \mathbf{A}_1 , \mathbf{B} , \mathbf{a} , and \mathbf{b} are deterministic time-varying coefficients. \mathbf{c} and \mathbf{e} are random residuals and are assumed to be Gaussian with mean being 0 and variance being $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}_e$ respectively.

We will use \mathbf{A}_1 , \mathbf{B} , \mathbf{a} , \mathbf{b} , Σ_c , Σ_e , and \mathbf{A}_2 to match the true mean and variance of \mathbf{F} and \mathbf{H}_1 , and the second moment of \mathbf{H}_2 with those of the approximation. That is, we match the right hand side with left hand side for the following equations (9)-(13) by choosing values for \mathbf{A}_1 , \mathbf{A}_2 , \mathbf{B} , \mathbf{a} , \mathbf{b} , Σ_c and Σ_e .

$$E(\mathbf{H}_1(\mathbf{X}_t)|\mathfrak{Z}_{t-1}) = \mathbf{A}_{1t}\mathbf{X}_t + \mathbf{a}_t \quad (9)$$

$$\text{Var}(\mathbf{H}_1(\mathbf{X}_t)|\mathfrak{Z}_{t-1}) = \mathbf{A}_{1t} \text{Var}(\mathbf{X}_t) \mathbf{A}_{1t}^T + \Sigma_{c_t} \quad (10)$$

$$E(\mathbf{F}(\mathbf{X}_{t-1})|\mathfrak{Z}_{t-1}) = \mathbf{B}_{t-1}\mathbf{X}_{t-1} + \mathbf{b}_{t-1} \quad (11)$$

$$\text{Var}(\mathbf{F}(\mathbf{X}_{t-1})|\mathfrak{Z}_{t-1}) = \mathbf{B}_{t-1} \text{Var}(\mathbf{X}_{t-1}) \mathbf{B}_{t-1}^T + \Sigma_{e_{t-1}} \quad (12)$$

$$\left(E(\mathbf{H}_2(\mathbf{X}_t)\mathbf{H}_2(\mathbf{X}_t)^T)|\mathfrak{Z}_{t-1} \right)^{\frac{1}{2}} = \mathbf{A}_{2t} \quad (13)$$

The benefit of linearization (7)-(8) is threefold. First, by this linearization, we can transform a nonlinear state-space model to a linear one, which the well-established Kalman filter can be applied to. We denote $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_{t-1})$ and $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_t)$ as the approximation of the prediction distribution and filtering distribution when we plug the linearization (7)-(8) into (3)-(6), and we present the property of $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_{t-1})$ and $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_t)$ in Proposition 1.

Proposition 1

For model (1)-(2), the approximated prediction distribution $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_{t-1})$ and approximated filtering distribution $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_t)$ are Gaussian if the probability measure on initial state \mathbf{X}_0 (i.e., μ) is Gaussian.

By matching the true mean and variance of nonlinear function in (9)-(13), we can match the true mean and variance of filtered and predicted estimates of the state variable. We show this by calculating, for example, the approximation of $E(\mathbf{X}_t|\mathfrak{Z}_{t-1})$ and comparing it with the true as following:

$$\begin{aligned} E(\widetilde{\mathbf{X}_t|\mathfrak{Z}_{t-1}}) &= \iint X_t \tilde{p}(\mathbf{X}_{t-1}|\mathfrak{Z}_{t-1}) \tilde{p}(\mathbf{X}_t|\mathbf{X}_{t-1}) d\mathbf{X}_{t-1} d\mathbf{X}_t \\ &= \iint X_t \tilde{p}(\mathbf{X}_t|\mathbf{X}_{t-1}) d\mathbf{X}_t \tilde{p}(\mathbf{X}_{t-1}|\mathfrak{Z}_{t-1}) d\mathbf{X}_{t-1} \\ &= \int \hat{\mathbf{X}}_{t|t-1} \tilde{p}(\mathbf{X}_{t-1}|\mathfrak{Z}_{t-1}) d\mathbf{X}_{t-1} \\ &= E(\mathbf{X}_t|\mathfrak{Z}_{t-1}) \end{aligned}$$

The detail of computing $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_{t-1})$ and $\tilde{p}(\mathbf{X}_t|\mathfrak{Z}_t)$ will be introduced later in this section and the algorithm for the implementation will be discussed in Section 2.4. In the remainder of this paper, we drop the tilde for \tilde{p} , and denote p as \tilde{p} .

Most importantly, through the linearization, the stability of a nonlinear filtering for (1)-(2) can be established through the analysis of the stability of linear filtering based on (7)-(8). We leave the discussion of stability in Section 4.

We are now in a position to discuss how to solve for the values of the parameters in (9)-(13). With the help of OLS, these values can be simply computed, and we summarize the results in the following proposition.

Proposition 2

For a nonlinear and Gaussian state space model (1)-(2), we apply (7)-(13) to approximate the nonlinear function \mathbf{F} , \mathbf{H}_1 and \mathbf{H}_2 by matching the approximated mean and variance of \mathbf{F} and \mathbf{H}_1 , and variance of \mathbf{H}_2 with the true ones. The time-varying parameters \mathbf{A}_1 , \mathbf{A}_2 , \mathbf{B} , \mathbf{a} , \mathbf{b} , Σ_c and Σ_e can be computed as in (14)-(20)

$$\hat{\mathbf{A}}_{1t} = \psi_{\mathbf{H}_{1t}} (\mathbf{P}_t^-)^{-1} \quad (14)$$

$$\hat{\mathbf{a}}_t = \bar{\mathbf{Z}}_t - \hat{\mathbf{A}}_{1t} \bar{\mathbf{X}}_t^- \quad (15)$$

$$\hat{\Sigma}_{c_t} = \Phi_{\mathbf{H}_{1t}} - \hat{\mathbf{A}}_{1t} \mathbf{P}_t^- \hat{\mathbf{A}}_{1t}^T \quad (16)$$

$$\hat{\mathbf{B}}_t = \psi_{\mathbf{F}_t} (\mathbf{P}_t)^{-1} \quad (17)$$

$$\hat{\mathbf{b}}_t = \bar{\mathbf{X}}_{t+1}^- - \hat{\mathbf{B}}_t \bar{\mathbf{X}}_t^- \quad (18)$$

$$\hat{\Sigma}_{e_t} = \Phi_{\mathbf{F}_t} - \hat{\mathbf{B}}_t \mathbf{P}_t \hat{\mathbf{B}}_t^T \quad (19)$$

$$\hat{\mathbf{A}}_{2t} = \Phi_{\mathbf{H}_{2t}} \quad (20)$$

where

$$\psi_{\mathbf{H}_{1t}} = E_{p_t^-} \left[\left(\mathbf{X}_t - \bar{\mathbf{X}}_t^- \right) (H_1(\mathbf{X}_t) - \bar{\mathbf{Z}}_t)^T \right] \quad (21)$$

$$\bar{\mathbf{Z}}_t = E_{p_t^-} [H_1(\mathbf{X}_t)] \quad (22)$$

$$\mathbf{P}_t^- = E_{p_{t-1}} \left[\left(\mathbf{F}(\mathbf{X}_{t-1}) - \bar{\mathbf{X}}_t^- \right) \left(\mathbf{F}(\mathbf{X}_{t-1}) - \bar{\mathbf{X}}_t^- \right)^T \right] + \Sigma_\epsilon \quad (23)$$

$$\Phi_{\mathbf{H}_{1t}} = E_{p_t^-} \left[(H_1(\mathbf{X}_t) - \bar{\mathbf{Z}}_t) (H_1(\mathbf{X}_t) - \bar{\mathbf{Z}}_t)^T \right] \quad (24)$$

$$\psi_{\mathbf{F}_t} = E_{p_t} \left[\left(\mathbf{X}_t - \bar{\mathbf{X}}_t^- \right) \left(\mathbf{F}(\mathbf{X}_t) - \bar{\mathbf{X}}_{t+1}^- \right)^T \right] \quad (25)$$

$$\mathbf{P}_t = E_{p_t^-} \left[(\mathbf{X}_t - \bar{\mathbf{X}}_t) (\mathbf{X}_t - \bar{\mathbf{X}}_t)^T \right] \quad (26)$$

$$\bar{\mathbf{X}}_{t+1}^- = E_{p_t} [\mathbf{F}(\mathbf{X}_t)] \quad (27)$$

$$\bar{\mathbf{X}}_t = E_{p_t^-} [\mathbf{X}_t] \quad (28)$$

$$\Phi_{\mathbf{F}_t} = E_{p_t} \left[\left(\mathbf{F}(\mathbf{X}_t) - \bar{\mathbf{X}}_{t+1}^- \right) \left(\mathbf{F}(\mathbf{X}_t) - \bar{\mathbf{X}}_{t+1}^- \right)^T \right] \quad (29)$$

$$\Phi_{\mathbf{H}_{2t}} = \left(E_{p_t} \left[\mathbf{H}_2(\mathbf{X}_t) \mathbf{H}_2(\mathbf{X}_t)^T \right] \right)^{\frac{1}{2}} \quad (30)$$

In the above proposition, we use $\hat{\cdot}$ to denote the estimate of the parameter in (9)-(13), and $\bar{\cdot}$ in the superscript to denote the predicted estimate of the variable (filtered estimate if no superscript $\bar{\cdot}$). Also, $E_{p_t^-}$ (or E_{p_t}) are denoted as the expectation based on the prediction distribution (or filtering distribution), i.e., we use p_t^- (or p_t) to represent $p(\mathbf{X}_t | \mathcal{Z}_{t-1})$ (or $p(\mathbf{X}_t | \mathcal{Z}_t)$) in the Proposition 2

2.3.2 Filter

Based on the linearization (7)-(8), we can rewrite the model (1)-(2) as:

$$\mathbf{X}_{t+1} = \hat{\mathbf{B}}_t \mathbf{X}_t + \hat{\mathbf{b}}_t + \mathbf{e}_t + \boldsymbol{\varepsilon}_t \quad (31)$$

$$\mathbf{Z}_t = \hat{\mathbf{A}}_{1t} \mathbf{X}_t + \hat{\mathbf{a}}_t + \mathbf{c}_t + \hat{\mathbf{A}}_{2t} \boldsymbol{\eta}_t \quad (32)$$

where $\mathbf{e}_t \sim N(0, \hat{\Sigma}_{\mathbf{e}_t})$ and $\mathbf{c}_t \sim N(0, \hat{\Sigma}_{\mathbf{c}_t})$. The time-varying parameters $\hat{\mathbf{A}}_1$, $\hat{\mathbf{A}}_2$, $\hat{\mathbf{B}}$, $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}$, $\hat{\Sigma}_{\mathbf{e}}$ and $\hat{\Sigma}_{\mathbf{c}}$ in (31)-(32) are based on (14)-(20) in Proposition 2. It can be verified that $\hat{\mathbf{B}}_t$, $\hat{\mathbf{b}}_t$, and $\hat{\Sigma}_{\mathbf{e}_t}$ in (31) are calculated based on the approximated filtering distribution $\tilde{p}(\mathbf{X}_t | \mathcal{Z}_t)$, and $\hat{\mathbf{A}}_{1t}$, $\hat{\mathbf{A}}_{2t}$, $\hat{\mathbf{a}}_t$, and $\hat{\Sigma}_{\mathbf{c}_t}$ in (32) are calculated based on the approximated prediction distribution $\tilde{p}(\mathbf{X}_t | \mathcal{Z}_{t-1})$.

Now, since the model is linear and Gaussian, the Kalman filter can be applied to the linearized system and we have the following priori predication and posteriori estimation of the mean and variance of the state variable through “Prediction step” and “Updating step”:

- Prediction step:

$$\bar{\mathbf{X}}_t^- = \hat{\mathbf{B}}_{t-1} \bar{\mathbf{X}}_{t-1} + \hat{\mathbf{b}}_{t-1} = E_{p_{t-1}}(\mathbf{F}(\mathbf{X}_{t-1})) \quad (33)$$

$$\mathbf{P}_t^- = \hat{\mathbf{B}}_{t-1} \mathbf{P}_{t-1} \hat{\mathbf{B}}_{t-1}^T + \hat{\Sigma}_{\mathbf{e}_{t-1}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \quad (34)$$

$$= E_{p_{t-1}} \left[(\mathbf{F}(\mathbf{X}_{t-1}) - \bar{\mathbf{X}}_t^-) (\mathbf{F}(\mathbf{X}_{t-1}) - \bar{\mathbf{X}}_t^-)^T \right] + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \quad (35)$$

- Updating step:

$$\bar{\mathbf{Z}}_t = \hat{\mathbf{A}}_t \bar{\mathbf{X}}_t^- + \hat{\mathbf{a}}_t = E_{p_t^-}[\mathbf{H}_1(\mathbf{X}_t)] \quad (36)$$

$$\mathbf{S}_t = \hat{\mathbf{A}}_{1t} \mathbf{P}_t^- \hat{\mathbf{A}}_{1t}^T + \hat{\Sigma}_{\mathbf{c}_t} + \hat{\mathbf{A}}_{2t} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \hat{\mathbf{A}}_{2t}^T \quad (37)$$

$$= E_{p_t^-} \left[(\mathbf{H}_1(\mathbf{X}_t) - \bar{\mathbf{Z}}_t) (\mathbf{H}_1(\mathbf{X}_t) - \bar{\mathbf{Z}}_t)^T \right] + \hat{\mathbf{A}}_{2t} \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \hat{\mathbf{A}}_{2t}^T \quad (38)$$

$$\mathbf{C}_t = \hat{\mathbf{A}}_{1t} \mathbf{P}_t^- = E_{p_t^-} \left[(\mathbf{X}_t - \bar{\mathbf{X}}_t^-) (\mathbf{H}_1(\mathbf{X}_t) - \bar{\mathbf{Z}}_t)^T \right] \quad (39)$$

$$\mathbf{K}_t = \mathbf{C}_t \mathbf{S}_t^{-1} \quad (40)$$

$$\bar{\mathbf{X}}_t = \bar{\mathbf{X}}_t^- + \mathbf{K}_t (\mathbf{Z}_t - \bar{\mathbf{Z}}_t) \quad (41)$$

$$\mathbf{P}_t = \mathbf{P}_t^- - \mathbf{K}_t \mathbf{S}_t \mathbf{K}_t^T \quad (42)$$

We denote the above filtering (33)-(42) as \mathfrak{F}_μ^g , where μ is the initial measure on \mathbf{X}_0 and g indicates this filtering is for Gaussian model. In (33)-(42), \mathbf{S} is the variance of \mathbf{Z} , and \mathbf{C} is the covariance between \mathbf{Z} and \mathbf{X} . Following the literature on Kalman filter, we denote \mathbf{K} as the Kalman gain matrix. Unlike a true linear state space model where all of the coefficients are known, the coefficients in (31)-(32) are time-varying and dependent on the state variable, and their values need to be determined within the filtering process. We provide another expression for each estimation in (33)-(39) after the second equation mark, and these expressions are more convenient for the evaluation of interests. Note that (33)-(39) contain expectation of nonlinear function under Gaussian distribution, which generally does not admit closed solution, and in this paper, we will use Quasi Monte Carlo approach to evaluate the expectation in (33)-(39). The detail of Quasi Monte Carlo will be discussed later.

Based on our Gaussian assumption on e and c , the predication distribution and filtering distribution would also be Gaussian with mean and variance determined in (33)-(35) and (41)-(42). That is,

$$\begin{aligned} p(\mathbf{X}_t | \mathcal{Z}_{t-1}) &= \phi\left(\mathbf{X}_t - \bar{\mathbf{X}}_t^-, \mathbf{P}_t^-\right), \\ p(\mathbf{X}_t | \mathcal{Z}_t) &= \phi\left(\mathbf{X}_t - \bar{\mathbf{X}}_t, \mathbf{P}_t\right). \end{aligned}$$

2.4 Quasi Monte Carlo

We briefly introduce the idea of Quasi Monte Carlo, and the details about this approach can be found in Niederreiter (1978, 1992), Morokoff and Caflisch (1994) and Caflisch (1998).

Quasi Monte Carlo approach is a numerical approach to evaluate integral based on deterministic quasi-random sequences. In contrast to the pseudo-random sequences used in Monte Carlo method, quasi-random sequences are designed to provide better uniformity than a random sequence. In the literature of quasi-random sequences, uniformity of a sequence is measured in terms of its discrepancy defined as below:

Definition 1 (Discrepancy)

The discrepancy of a set $x = \{x_1, \dots, x_G\}$ is defined, using the notation of Niederreiter (1978), as

$$D_G(x) = \sup_{J \in \mathcal{J}} \left| \frac{\chi(J; x)}{G} - \lambda_s(J) \right| \quad (43)$$

where λ_s is the s -dimensional Lebesgue measure, $\chi(J; x)$ is the number of points in x that fall into J , and \mathcal{J} is the set of s -dimensional boxes of the form

$$\prod_{i=1}^s [\varsigma_i, \kappa_i) = \{\omega \in \mathbb{R}^s : \varsigma_i \leq \omega_i \leq \kappa_i\}$$

where $0 \leq \varsigma_i < \kappa_i \leq 1$.

Definition 2 (low-discrepancy sequence (quasi-random sequence))

Sequence x is said to be low-discrepancy or quasi-random if

$$D_G(x) \leq c_{\dim(x)} (\log G)^{\dim(x)} G^{-1}$$

Popular examples of quasi-random sequence are van der Corput sequence¹, Halton sequence². Quasi-random sequences are efficient for integration because they lead to smaller error than standard Monte Carlo method. This is summarized in the well-known Koksma-Hlawka inequality.

Theorem 1 (Koksma-Hlawka inequality, Koksma (1942/43), Hlawka (1964,1971))

Let \bar{I}^s be the s -dimensional unit cube, $\bar{I}^s = [0, 1] \times \dots \times [0, 1]$. Let f have bounded variation $W(f)$ on \bar{I}^s in the sense of Hardy and Krause (K. Basu and A. Owen, 2015). Then for any $x = \{x_1, \dots, x_G\}$ such that x_i is in $\bar{I}^s = [0, 1] \times \dots \times [0, 1]$,

$$\left| \frac{1}{G} \sum_{i=1}^G f(x_i) - \int_{\bar{I}^s} f(u) du \right| \leq W(f) D_G(x) \quad (44)$$

where

$$W(f) = \int_0^1 \left| \frac{df}{dt} \right| dt. \quad (45)$$

2.5 Algorithm

We finish this section by proposing a recursive algorithm for the implementation of (33)-(42). We use Quasi Monte Carlo approach to evaluate the expectations in (33)-(42), and call this filter as

¹Suppose the b -ary representation of a positive integer n is

$$n = \sum_{l=0}^L d_l(n) b^l,$$

where b is the base and $d_l(n)$ is the l -th digit in the expansion of n , then the van der Corput sequence is defined as $x = (\rho_b(n))_n$ such that:

$$\rho_b(n) = \sum_{l=0}^{L-1} d_l(n) b^{-l-1}.$$

²Based on the van der Corput sequence, an s -dimensional Halton sequence $x = (x_n)_n$ is defined as

$$x_n = (\rho_{b_1}(n), \dots, \rho_{b_s}(n)), \quad n = 0, 1, \dots,$$

where b_1, \dots, b_s are the first s prime numbers. For example, the following is a 1-dimensional Halton sequence:

$$1/2, 1/4, 3/4, 1/8, 5/8, 3/8, 7/8, 1/16, 9/16, \dots$$

where

$$\begin{aligned} 6 &= 1 * 2^2 + 1 * 2^1 + 0 * 2^0 \\ 0 * 2^{-1} + 1 * 2^{-2} + 1 * 2^{-3} &= 3/8 \end{aligned}$$

Quasi Monte Carlo Kalman filter (QMCKF) for nonlinear and Gaussian state space model (1)-(2). More specifically, QMCKF, denoted as $\hat{\mathfrak{F}}_\mu^g$, is defined as the followings with $\mathbf{X}_0 \sim \mu$ and simulation size G :

$$\mathbf{P}_t^{G-} = \hat{\Phi}_{\mathbf{F}_{t-1}}^G + \boldsymbol{\Sigma}_\epsilon \quad (46)$$

$$\mathbf{S}_t^G = \hat{\Phi}_{\mathbf{H}_{1t}}^G + \hat{\Phi}_{\mathbf{H}_{2t}}^G \boldsymbol{\Sigma}_\eta \left(\hat{\Phi}_{\mathbf{H}_{2t}}^G \right)^T \quad (47)$$

$$\mathbf{C}_t^G = \hat{\psi}_{\mathbf{H}_{1t}}^G \quad (48)$$

$$\mathbf{K}_t^G = \mathbf{C}_t^G \mathbf{S}_t^{G-1} \quad (49)$$

$$\bar{\mathbf{X}}_t^G = \bar{\mathbf{X}}_{t-1}^{G-} + \mathbf{K}_t^G \left(\mathbf{Z}_t - \bar{\mathbf{Z}}_t^G \right) \quad (50)$$

$$\mathbf{P}_t^G = \mathbf{P}_t^{G-} - \mathbf{K}_t^G \mathbf{S}_t^G \left(\mathbf{K}_t^G \right)^T \quad (51)$$

And the algorithm for (46)-(51) is summarized in Algorithm 1.

Algorithm 1: QMCKF for nonlinear and Gaussian case

- Step 1: Generate a low-discrepancy sequence (e.g., a Halton sequence) $\left\{ \mathbf{X}_{t-1}^{(g)} \right\}_{g=1}^G$ from $\mathcal{U}(\mathbf{0}, \mathbf{1})$ and use Box-Muller transform to make the sequence be generated from $\mathcal{N}(\bar{\mathbf{X}}_{t-1}^G, \mathbf{P}_{t-1}^G)$.
- Step 2: Compute and store $\bar{\mathbf{X}}_t^{G-}$, $\hat{\Phi}_{\mathbf{F}_{t-1}}^G$ and \mathbf{P}_t^{G-} according to $\bar{\mathbf{X}}_t^{G-} = \frac{1}{G} \sum_{i=1}^G \mathbf{F} \left(\mathbf{X}_{t-1}^{(g)} \right)$, $\hat{\Phi}_{\mathbf{F}_{t-1}}^G = \frac{1}{G} \sum_{i=1}^G \left(\mathbf{F} \left(\mathbf{X}_{t-1}^{(g)} \right) - \bar{\mathbf{X}}_t^{G-} \right) \left(\mathbf{F} \left(\mathbf{X}_{t-1}^{(g)} \right) - \bar{\mathbf{X}}_t^{G-} \right)^T$ and $\mathbf{P}_t^{G-} = \hat{\Phi}_{\mathbf{F}_{t-1}}^G + \boldsymbol{\Sigma}_\epsilon$.
- Step 3: Generate a Halton sequence $\left\{ \mathbf{X}_t^{(g)-} \right\}_{g=1}^G$ from $\mathcal{N}(\bar{\mathbf{X}}_t^{G-}, \mathbf{P}_t^{G-})$ as Step 1.
- Step 4: Compute and store $\bar{\mathbf{Z}}_t^G$, $\hat{\psi}_{\mathbf{H}_t}^G$, $\hat{\Phi}_{\mathbf{H}_{1t}}^G$ and $\hat{\Phi}_{\mathbf{H}_{2t}}^G$ according to $\bar{\mathbf{Z}}_t^G = \frac{1}{G} \sum_{i=1}^G \mathbf{H}_1 \left(\mathbf{X}_t^{(g)-} \right)$, $\hat{\psi}_{\mathbf{H}_{1t}}^G = \frac{1}{G} \sum_{i=1}^G \left(\mathbf{X}_t^{(g)-} - \bar{\mathbf{X}}_t^{G-} \right) \left(\mathbf{H}_1 \left(\mathbf{X}_t^{(g)-} \right) - \bar{\mathbf{Z}}_t^G \right)^T$, $\hat{\Phi}_{\mathbf{H}_{1t}}^G = \frac{1}{G} \sum_{i=1}^G \left(\mathbf{H}_1 \left(\mathbf{X}_t^{(g)-} \right) - \bar{\mathbf{Z}}_t^G \right) \left(\mathbf{H}_1 \left(\mathbf{X}_t^{(g)-} \right) - \bar{\mathbf{Z}}_t^G \right)^T$ and $\hat{\Phi}_{\mathbf{H}_{2t}}^G = \left(\frac{1}{G} \sum_{i=1}^G \left(\mathbf{H}_2 \left(\mathbf{X}_t^{(g)-} \right) \right) \left(\mathbf{H}_2 \left(\mathbf{X}_t^{(g)-} \right) \right)^T \right)^{\frac{1}{2}}$
- Step 5: Based on Step 4, compute and store \mathbf{S}_t^G , \mathbf{C}_t^G , \mathbf{K}_t^G , $\bar{\mathbf{X}}_t^G$ and \mathbf{P}_t^G .
- Step 6: Repeat Step 1-5 for $t = 2, \dots, T$.

Note that When $t = 1$, $\left\{ \mathbf{X}_0^{(g)} \right\}_{g=1}^G$ in Step 1 will be sampled from μ .

3 Nonlinear and Non-Gaussian state space models

3.1 Examples

Before we present the nonlinear and non-Gaussian state space models in a general form, we first provide two examples of such models. The first example is the regime switching model by Hamilton (1994). We show that we can transform the regime switching model into a linear and non-Gaussian state space model. The second example is the fat-tailed stochastic volatility model by Jacquier, Polson and Rossi (2004).

Example 3 (Regime switching model)

This example is from Hamilton (1994). We consider an unobservable state variable s_t that can take finite number of values $1, 2, \dots, S$ corresponding to S different possible regimes. Also we think of the following regression model:

$$y_t = \mathbf{x}'_t \boldsymbol{\beta}_{s_t} + w_t \quad (52)$$

where \mathbf{x}_t is a $(k \times 1)$ vector of exogenous variables and $w_t \sim i.i.d.\mathcal{N}(0, \sigma^2)$. The value of the coefficient $\boldsymbol{\beta}_{s_t}$ is dependent on the regime s_t . Denote $\zeta_t \equiv (y_t, y_{t-1}, \dots, y_1, \mathbf{x}'_t, \mathbf{x}'_{t-1}, \dots, \mathbf{x}'_1)'$, and assume

$$\text{Prob}(s_{t+1} = j | s_t = i, s_{t-1} = i_1, s_{t-2} = i_2, \dots, \zeta_t) = \text{Prob}(s_{t+1} = j | s_t = i).$$

Let $\boldsymbol{\Pi}$ be an $(S \times S)$ matrix such that

$$\boldsymbol{\Pi} = (\text{Prob}(s_{t+1} = j | s_t = i))_{ij}$$

Let $\boldsymbol{\iota}_i$ denote the i -th column of the I_S and construct $\boldsymbol{\xi}_t$ equal to $\boldsymbol{\iota}_i$ when $s_t = i$. Then the Markov chain of state implies

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\Pi} \boldsymbol{\xi}_t + \boldsymbol{\nu}_{t+1} \quad (53)$$

and

$$E(\boldsymbol{\xi}_{t+j} | \boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \dots, \boldsymbol{\xi}_1, \zeta_t) = \boldsymbol{\Pi}^j \boldsymbol{\xi}_t$$

Then the regression model could be written as

$$y_t = \mathbf{x}'_t \mathcal{B} \boldsymbol{\xi}_t + w_t \quad (54)$$

where \mathcal{B} is a $(k \times S)$ matrix whose i -th column is given by $\boldsymbol{\beta}_i$.

(53)-(54) is an example of linear and non-Gaussian state space model. The implication of this example is that a regime switching model can be represented as a non-Gaussian state space model. However, due to the non-Gaussianity of noise, Kalman filter cannot be applied to this model to estimate the filtering distribution of the state variable. A popular tool for this regime switching model

is the Hamilton filter developed by Hamilton (1989). More detail on the theories and applications of regime switching model can be found in Engel and Hamilton (1990), Lam (1990), Goodwin (1993), Engel (1994), Filardo (1994), Ghysels (1994), Kim and Yoo (1995), Garcia and Perron (1996), Kim and Nelson (1998), Bai and Perron (1998), Carrasco, Hu, and Ploberger (2014), and Qu and Zhuo (2020).

Example 4 (Fat-tailed stochastic volatility model)

As an extension to the basic log-normal auto regressive stochastic volatility model (Taylor (1986), and Kim, Shephard and Chib (1998)), Jacquier, Polson and Rossi (2004) consider the following fat-tailed stochastic volatility model:

$$y_t = \sqrt{h_t} \bar{\varepsilon}_t = \sqrt{h_t} \sqrt{\lambda_t} \varepsilon_t, \quad (55)$$

$$\log h_t = \alpha + \beta \log h_{t-1} + \sigma_\eta \eta_t \quad t \geq 1, \quad (56)$$

$$\lambda_t \sim p(\lambda_t | \nu). \quad (57)$$

where y is the return from an asset and h is its volatility. α , β , σ_η , and ν are parameters in the model. Jacquier, Polson and Rossi (2004) assume ε_t and η_t are dependent random variables with $\varepsilon_t \sim \mathcal{N}(0, 1)$ and $\eta_t \sim \mathcal{N}(0, 1)$. λ_t is assumed to follow i.i.d inverse gamma distribution, or $\nu/\lambda_t \sim \chi_\nu^2$. This implies the marginal distribution of $\bar{\varepsilon}_t \equiv \sqrt{\lambda_t} \varepsilon_t$ is Student's t_ν . They argue that the fat-tailed stochastic volatility models can be viewed as more outlier resistant than the basic stochastic volatility models. A likelihood-based inference method based on MCMC for this model is discussed in Jacquier, Polson and Rossi (2004), which is based on Jacquier, Polson and Rossi (1994).

3.2 The Model

We consider the following nonlinear and non-Gaussian state-space model:

$$\mathbf{X}_{k+1} = \mathbf{F}(\mathbf{X}_k; \theta) + \mathbf{u}_k \quad (58)$$

$$\mathbf{Z}_k = \mathbf{H}(\mathbf{X}_k; \theta) + \mathbf{v}_k \quad (59)$$

We assume the initial state \mathbf{X}_0 follows the probability measure μ which is Gaussian with mean μ_0 and variance Σ_0 . For the sake of simplicity, we assume $\mathbf{u}_k \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{u}, \boldsymbol{\Sigma}_\mathbf{u})$ and allow for non-Gaussianity of \mathbf{v}_k , that is $\mathbf{v}_k \sim p_\mathbf{v}(\cdot)$ where $p_\mathbf{v}(\cdot)$ is a known parametric pdf with finite first two moments. As in Section 2, we assume \mathbf{u} and \mathbf{v} are independent and \mathbf{v}_s and \mathbf{v}_t , $s \neq t$, are serially independent. For simplicity, in the non-Gaussian model, we assume the volatility of the observation in the measurement equation to be independent of the state variable. That is, we assume the \mathbf{H}_2 function in the Gaussian model (i.e., equation (2)) degenerates to be $\mathbf{1}$ in non-Gaussian model,

and we denote \mathbf{H}_1 function as \mathbf{H} . As we will discuss with more details, our approach to non-Gaussian model will treat the Gaussian model (1)-(2) as a special case, and it is easy to see that our approach can be simply extended to include more general case that \mathbf{H}_2 is not $\mathbf{1}$. \mathcal{X} , the space of state variable, is non-compact because \mathbf{u} is Gaussian, however, for that stability of nonlinear filtering, we limit \mathcal{Z} , the space of observation, to be a locally compact subspace of $\mathbb{R}^{\dim(\mathbf{Z})}$. The Gaussian assumption on μ and \mathbf{u} is not necessary and would be relaxed later in this section. When \mathbf{u} is non-Gaussian, we also need to limit \mathcal{X} to be a locally compact subspace of $\mathbb{R}^{\dim(\mathbf{X})}$.

Examples of locally compact space would be the set of real numbers (e.g., the Gaussian state space model in Section 2), a set of finite number of points (e.g., the regime switching models in Example 3) or a compact set (e.g., the stochastic volatility models where state variable is non-negative in Example 2 and 4)

3.3 Filtering Approach

We denote $\mathfrak{Z}_k \equiv \{\mathbf{Z}_k, \mathbf{Z}_{k-1}, \dots, \mathbf{Z}_0\}$. To derive the prediction distribution $p(\mathbf{X}_{k+1}|\mathfrak{Z}_k)$ and the filtering distribution $p(\mathbf{X}_{k+1}|\mathfrak{Z}_{k+1})$ in (58)-(59), we first approximate $p_{\mathbf{v}}$ using a series of Gaussian distributions:

$$p_{\mathbf{v}}(\mathbf{v}) \approx \tilde{p}_{\mathbf{v}}(\mathbf{v}) = \sum_{i=1}^m \alpha_i \phi(\mathbf{v} - \mathbf{a}_i, \mathbf{P}_i), \quad \sum_{i=1}^m \alpha_i = 1 \quad (60)$$

where

$$\phi(v, \Sigma) = \frac{1}{(2\pi)^{\dim(v)/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} v^T \Sigma^{-1} v\right).$$

This Gaussian mixture approximation of non-Gaussian density was also applied in Kim, Shaphad and Chib (1998) to improve the performance of estimating a stochastic volatility model. In their approach, one of the m mixtures, $(\mathbf{a}_i, \mathbf{P}_i)$, would first be drawn with the relative frequency of α_i , and then the non-Gaussian state space model would degenerate to a Gaussian model and the Kalman filter can be applied accordingly. However, their approach is limited to univariate case. Also, when the model is nonlinear, simple Kalman filter can not be directly applied for filtering. In our approach, instead of sampling $(\mathbf{a}_i, \mathbf{P}_i)$, we can derive the predicting distribution and filtering distribution in a tractable way. Our approach can be applied to nonlinear or multivariate state-space models.

To get the filtering distribution, we can plug this approximation (60) into (3):

$$\begin{aligned} p(\mathbf{X}_k|\mathfrak{Z}_k) &= c_k p(\mathbf{X}_k|\mathfrak{Z}_{k-1}) p_{\mathbf{v}}(\mathbf{Z}_k - \mathbf{H}(\mathbf{X}_k) | \mathbf{X}_k) \\ &\approx c_k p(\mathbf{X}_k|\mathfrak{Z}_{k-1}) \sum_{i=1}^m \alpha_i \phi(\mathbf{Z}_k - \mathbf{H}(\mathbf{X}_k) - \mathbf{a}_i, \mathbf{P}_i) \\ &= \sum_{i=1}^m c_{ki} p(\mathbf{X}_k|\mathfrak{Z}_{k-1}) \phi(\mathbf{Z}_k - \mathbf{H}(\mathbf{X}_k) - \mathbf{a}_i, \mathbf{P}_i), \end{aligned} \quad (61)$$

where

$$1/c_k = \int p(\mathbf{X}_k | \mathfrak{Z}_{k-1}) p_{\mathbf{v}}(\mathbf{Z}_k - \mathbf{H}(\mathbf{X}_k) | \mathbf{X}_k) d\mathbf{X}_t \quad (62)$$

and

$$c_{ki} = c_k * \alpha_i.$$

Notice that (61) is similar to (3), and the approach in Section 2 can also be applied for linearization of nonlinear function \mathbf{H} and \mathbf{F} . In Proposition 2-4, we show that the prediction distribution and filtering distribution can be written as a mixture of Gaussian distributions if we start from Gaussian $\mu = p(\mathbf{X}_0)$.

Proposition 3

For a nonlinear and non-Gaussian state space model given in (58)-(59), the filtering distribution when time is 0, i.e., $p(\mathbf{X}_0 | \mathbf{Z}_0)$ is given in the following:

$$p(\mathbf{X}_0 | \mathbf{Z}_0) = \sum_{i=1}^m \beta_{0i} \phi(\mathbf{X}_0 - \mathbf{b}_{0i}, \mathbf{P}_{0i}) \quad (63)$$

where

$$\mathbf{b}_{0i} = \mu_0 + \mathbf{K}_{0i} (\mathbf{Z}_0 - E_{\mu}[\mathbf{H}(\mathbf{X}_0)] - \mathbf{a}_i) \quad (64)$$

$$\mathbf{S}_{0i} = E_{\mu} [(\mathbf{H}(\mathbf{X}_0) - \mathbf{H}(\mu_0)) (\mathbf{H}(\mathbf{X}_0) - \mathbf{H}(\mu_0))^T] + \mathbf{P}_i \quad (65)$$

$$\mathbf{C}_{0i} = E_{\mu} [(\mathbf{X}_0 - \mu_0) (\mathbf{H}(\mathbf{X}_0) - \mathbf{H}(\mu_0))^T] \quad (66)$$

$$\mathbf{K}_{0i} = \mathbf{C}_{0i} \mathbf{S}_{0i}^{-1} \quad (67)$$

$$\mathbf{P}_{0i} = \Sigma_0 - \mathbf{K}_{0i} \mathbf{S}_{0i} \mathbf{K}_{0i}^T \quad (68)$$

$$\beta_{0i} = \frac{\alpha_i \phi(\mathbf{Z}_0 - E_{\mu}[\mathbf{H}(\mathbf{X}_0)] - \mathbf{a}_i, \mathbf{S}_{0i})}{\sum_{j=1}^I \alpha_j \phi(\mathbf{Z}_0 - E_{\mu}[\mathbf{H}(\mathbf{X}_0)] - \mathbf{a}_j, \mathbf{S}_{0j})}. \quad (69)$$

In Proposition 3, we denote E_{μ} as the expectation under the distribution μ on \mathbf{X}_0 with mean μ_0 and variance Σ_0 . \mathbf{a}_i in (64) and (69), \mathbf{P}_i in (65), and α_i in (69) for $i = 1, \dots, m$ are from the approximation (60). Also, ϕ represents the Gaussian pdf. The next two propositions will provide the prediction distribution and filtering distribution at arbitrary time k .

Proposition 4

For a nonlinear and non-Gaussian state space model given in (58)-(59), the prediction distribution when time is k , i.e., $p(\mathbf{X}_{k+1} | \mathfrak{Z}_k)$ is given in the following:

$$p(\mathbf{X}_{k+1} | \mathfrak{Z}_k) = \sum_{i=1}^{J_{k+1}} \gamma_{k+1i} \phi(\mathbf{X}_{k+1} - \mathbf{c}_{k+1i}, \mathbf{Q}_{k+1i}) \quad (70)$$

where

$$\mathbf{c}_{k+1i} = E_{\phi_i} [\mathbf{F}(\mathbf{X}_k)] + \boldsymbol{\mu}_{\mathbf{u}} \quad (71)$$

$$\mathbf{Q}_{k+1i} = E_{\phi_i} \left[(\mathbf{F}(\mathbf{X}_k) - \mathbf{F}(\mathbf{b}_{ki})) (\mathbf{F}(\mathbf{X}_k) - \mathbf{F}(\mathbf{b}_{ki}))^T \right] + \boldsymbol{\Sigma}_{\mathbf{u}} \quad (72)$$

$$\gamma_{k+1i} = \beta_{ki} \quad (73)$$

$$J_{k+1} = I_k. \quad (74)$$

In Proposition 4, we denote E_{ϕ_i} as the expectation under the i -th mixture of the filtering distribution $p(\mathbf{X}_k|\mathcal{Z}_k)$ computed according to Proposition 5 where we have I_k mixtures in total. Similarly, β_{ki} in (73) is the i -th weight with respect to the i -th mixture of $p(\mathbf{X}_k|\mathcal{Z}_k)$. Also, \mathbf{b}_{ki} in (72) is the filtered mean of the state variable under the i -th mixture of $p(\mathbf{X}_k|\mathcal{Z}_k)$. $\boldsymbol{\mu}_{\mathbf{u}}$ in (71), and $\boldsymbol{\Sigma}_{\mathbf{u}}$ in (72) are the mean and variance of the noise \mathbf{u} in (58). The result of Proposition 4, $p(\mathbf{X}_{k+1}|\mathcal{Z}_k)$, will have J_{k+1} mixtures in total, and its i -th mixture, $i = 1, \dots, J_{k+1}$, will be a Gaussian pdf with the mean \mathbf{c}_{k+1i} and the variance \mathbf{Q}_{k+1i} , associated with the weight γ_{k+1i} .

Proposition 5

For a nonlinear and non-Gaussian state space model given in (58)-(59), the filtering distribution when time is $k+1$, i.e., $p(\mathbf{X}_{k+1}|\mathcal{Z}_{k+1})$ is given in the following:

$$p(\mathbf{X}_{k+1}|\mathcal{Z}_{k+1}) = \sum_{i=1}^{I_{k+1}} \beta_{k+1i} \phi(\mathbf{X}_{k+1} - \mathbf{b}_{k+1i}, \mathbf{P}_{k+1i}) \quad (75)$$

where

$$\mathbf{b}_{k+1i} = \mathbf{c}_{k+1s} + \mathbf{K}_{k+1i} \left(Z_{k+1} - E_{\phi_s^-} [\mathbf{H}(\mathbf{X}_{k+1})] - \mathbf{a}_t \right) \quad (76)$$

$$\mathbf{S}_{k+1i} = E_{\phi_s^-} \left[(\mathbf{H}(\mathbf{X}_{k+1}) - \mathbf{H}(\mathbf{c}_{k+1s})) (\mathbf{H}(\mathbf{X}_{k+1}) - \mathbf{H}(\mathbf{c}_{k+1s}))^T \right] + \mathbf{P}_t \quad (77)$$

$$\mathbf{C}_{k+1i} = E_{\phi_s^-} \left[(\mathbf{X}_{k+1} - \mathbf{c}_{k+1s}) (\mathbf{H}(\mathbf{X}_{k+1}) - \mathbf{H}(\mathbf{c}_{k+1s}))^T \right] \quad (78)$$

$$\mathbf{K}_{k+1i} = \mathbf{C}_{k+1i} \mathbf{S}_{k+1i}^{-1} \quad (79)$$

$$\mathbf{P}_{k+1i} = \mathbf{Q}_{k+1s} - \mathbf{K}_{k+1i} \mathbf{S}_{k+1i} \mathbf{K}_{k+1i}^T \quad (80)$$

$$\beta_{k+1i} = \frac{\alpha_t \phi \left(Z_{k+1} - E_{\phi_s^-} [\mathbf{H}(\mathbf{X}_{k+1})] - \mathbf{a}_t, \mathbf{S}_{k+1i} \right)}{\sum_{s=1}^{J_{k+1}} \sum_{t=1}^m \alpha_t \phi \left(Z_{k+1} - E_{\phi_s^-} [\mathbf{H}(\mathbf{X}_{k+1})] - \mathbf{a}_t, \mathbf{S}_{k+1i} \right)} \quad (81)$$

$$I_{k+1} = J_{k+1} * m, \quad (82)$$

$$I_0 = m, \quad (83)$$

for $s = 1, \dots, J_{k+1}$ and $t = 1, \dots, m$.

In Proposition 5, we denote $E_{\phi_s^-}$ as the expectation under the s -th mixture of the prediction distribution $p(\mathbf{X}_{k+1}|\mathcal{Z}_k)$ computed in Proposition 4 where we have J_{k+1} mixtures, and \mathbf{c}_{k+1s} and \mathbf{Q}_{k+1s} are its mean and variance. \mathbf{a}_t in (76) and (81), \mathbf{P}_t in (77), and α_t in (81) for $t = 1, \dots, m$ are the mean, variance and weight associated with the t -th mixture from the approximation (60). The result of Proposition 5, $p(\mathbf{X}_{k+1}|\mathcal{Z}_{k+1})$, will have I_{k+1} mixtures in total, and its i -th mixture, $i = 1, \dots, I_{k+1}$, will be a Gaussian pdf with the mean \mathbf{b}_{k+1i} and the variance \mathbf{P}_{k+1i} , associated with the weight β_{k+1i} .

Through the Proposition 2 to 4, we keep the notation the same as that in linear model (33)-(42): we denote \mathbf{S} as the variance of the state variable, \mathbf{C} as the covariance between the state variable and observation and \mathbf{K} as the Kalman gain matrix. The idea of above propositions is if we use a Gaussian mixture model to approximate the non-Gaussian distribution of the noise, then the prediction distribution and filtering distribution can also be approximated by mixture of Gaussian distributions with parameters updated recursively based on a series of Kalman filters (or QMCKF if the model is nonlinear) working simultaneously.

To start the procedure of Proposition 2-4, we need to first implement the approximation $p_{\mathbf{v}}(\mathbf{v}) \approx \tilde{p}_{\mathbf{v}}(\mathbf{v}; \boldsymbol{\vartheta}) = \sum_{i=1}^m \alpha_i \phi(\mathbf{v} - \mathbf{a}_i, \mathbf{P}_i)$ with $\boldsymbol{\vartheta} = \{\alpha_i, \mathbf{a}_i, \mathbf{P}_i\}_i^m$. Kim, Shephard and Chib (1998), following Titterington, Smith, and Makov (1985), matched the first four moments of the Gaussian mixture approximation to those of the true density. This approach is limited in accuracy, speed and applicability. As a more general approach, we first define the relative entropy \mathcal{H} of $p_{\mathbf{v}}$ with respect to $\tilde{p}_{\mathbf{v}}$ as:

$$\mathcal{H}(p_{\mathbf{v}}|\tilde{p}_{\mathbf{v}}) = \int \left(\log \frac{p_{\mathbf{v}}(\mathbf{v})}{\tilde{p}_{\mathbf{v}}(\mathbf{v})} \right) \times p_{\mathbf{v}}(\mathbf{v}) d\mathbf{v}. \quad (84)$$

Then $\boldsymbol{\vartheta}$ is going to be determined to minimize $\mathcal{H}(p_{\mathbf{v}}|\tilde{p}_{\mathbf{v}})$.

In practice, $\mathcal{H}(p_{\mathbf{v}}|\tilde{p}_{\mathbf{v}})$ usually doesn't admit closed form, so we use Monte Carlo approach to evaluate to integration:

$$\mathcal{H}(p_{\mathbf{v}}|\tilde{p}_{\mathbf{v}}) = E \left[\left(\log \frac{p_{\mathbf{v}}(\mathbf{v})}{\tilde{p}_{\mathbf{v}}(\mathbf{v})} \right) \times p_{\mathbf{v}}(\mathbf{v}) \right] \quad (85)$$

$$\approx \frac{1}{G} \sum_{g=1}^G \left(\log \frac{p_{\mathbf{v}}(\mathbf{v}^{(g)})}{\tilde{p}_{\mathbf{v}}(\mathbf{v}^{(g)})} \right) \times p_{\mathbf{v}}(\mathbf{v}^{(g)}) \quad \{\mathbf{v}^{(g)}\} \sim \mathcal{U} \quad (86)$$

If $p_{\mathbf{v}}$ is easy to sample, $\mathcal{H}(p_{\mathbf{v}}|\tilde{p}_{\mathbf{v}})$ can also be approximated by

$$\mathcal{H}(p_{\mathbf{v}}|\tilde{p}_{\mathbf{v}}) = E_{p_{\mathbf{v}}} \left[\left(\log \frac{p_{\mathbf{v}}(\mathbf{v})}{\tilde{p}_{\mathbf{v}}(\mathbf{v})} \right) \right] \quad (87)$$

$$\approx \frac{1}{G} \sum_{g=1}^G \left(\log \frac{p_{\mathbf{v}}(\mathbf{v}^{(g)})}{\tilde{p}_{\mathbf{v}}(\mathbf{v}^{(g)})} \right), \quad \{\mathbf{v}^{(g)}\} \sim p_{\mathbf{v}}(\cdot) \quad (88)$$

If not, we can apply importance sampling to reduce the approximation error, i.e., for some Gaussian

$\psi_{\mathbf{v}}$:

$$\mathcal{H}(p_{\mathbf{v}}|\tilde{p}_{\mathbf{v}}) = E_{\psi_{\mathbf{v}}} \left[\left(\log \frac{p_{\mathbf{v}}(\mathbf{v})}{\tilde{p}_{\mathbf{v}}(\mathbf{v})} \right) \times \frac{p_{\mathbf{v}}(\mathbf{v})}{\psi_{\mathbf{v}}(\mathbf{v})} \right] \quad (89)$$

$$\approx \frac{1}{G} \sum_{g=1}^G \left(\log \frac{p_{\mathbf{v}}(\mathbf{v}^{(g)})}{\tilde{p}_{\mathbf{v}}(\mathbf{v}^{(g)})} \right) \times \frac{p_{\mathbf{v}}(\mathbf{v}^{(g)})}{\psi_{\mathbf{v}}(\mathbf{v}^{(g)})}, \quad \{\mathbf{v}^{(g)}\} \sim \psi_{\mathbf{v}}(\cdot) \quad (90)$$

Other distance criterion can also be applied, such as

- Hellinger distance: $d_H^2(p_{\mathbf{v}}, \tilde{p}_{\mathbf{v}}) = \int \left(\sqrt{p_{\mathbf{v}}(\mathbf{v})} - \sqrt{\tilde{p}_{\mathbf{v}}(\mathbf{v})} \right)^2 d\mathbf{v}$
- Total variation distance: $d_{TV}(p_{\mathbf{v}}, \tilde{p}_{\mathbf{v}}) = \frac{1}{2} \int |p_{\mathbf{v}}(\mathbf{v}) - \tilde{p}_{\mathbf{v}}(\mathbf{v})| d\mathbf{v}$

The property of \mathcal{H} , d_H^2 and d_{TV} for density approximation of Gaussian mixture models have been discussed in Li and Barron (1999), Genovese and Wasserman (2000), Ghosal (2001), Ghosal and van der Vaart (2001, 2007) and we will discuss the detail in Section 4. Numerical approach for the implementation based on importance sampling can be found in Joseph (2012) and Rue, Martino and Chopin (2009).

Remark 1

Note that $(\mathbf{b}_{k+1i}, \mathbf{P}_{k+1i})$ and $(\mathbf{c}_{k+1i}, \mathbf{Q}_{k+1i})$ in Proposition 3 and 4 are the posterior and priori estimates about the mean and variance of the state variable under a single mixture. Instead, the mean and error covariance of the state variable in priori and posteriori estimation, in univariate case, can be computed as following:

$$\hat{\mathbf{X}}_{k+1}^- = \sum_{i=1}^{J_{k+1}} \gamma_{k+1i} \mathbf{c}_{k+1i}, \quad (91)$$

$$\hat{\mathbf{X}}_{k+1} = \sum_{i=1}^{I_{k+1}} \beta_{k+1i} \mathbf{b}_{k+1i}, \quad (92)$$

$$\hat{\mathbf{P}}_{k+1}^- = \sum_{i=1}^{J_{k+1}} \gamma_{k+1i} (\mathbf{Q}_{k+1i} + \mathbf{c}_{k+1i}^2) - \left(\sum_{i=1}^{J_{k+1}} \gamma_{k+1i} \mathbf{c}_{k+1i} \right)^2, \quad (93)$$

$$\hat{\mathbf{P}}_{k+1} = \sum_{i=1}^{I_{k+1}} \beta_{k+1i} (\mathbf{P}_{k+1i} + \mathbf{b}_{k+1i}^2) - \left(\sum_{i=1}^{I_{k+1}} \beta_{k+1i} \mathbf{b}_{k+1i} \right)^2. \quad (94)$$

Also, note that since the predicting distribution and filtering distribution are non-Gaussian, we do not have $p(\mathbf{X}_{k+1}|\mathfrak{Z}_k) = \phi(\mathbf{X}_{k+1} - \hat{\mathbf{X}}_{k+1}^-, \hat{\mathbf{P}}_{k+1}^-)$ or $p(\mathbf{X}_{k+1}|\mathfrak{Z}_{k+1}) = \phi(\mathbf{X}_{k+1} - \hat{\mathbf{X}}_{k+1}, \hat{\mathbf{P}}_{k+1})$.

Remark 2

The number of mixtures for the prediction distribution (in Proposition 4) is J_{k+1} and that for the

filtering distribution (in Proposition 5) is $I_{k+1} = J_{k+1} * m = m^{k+2}$. The exponential increase of the number of mixtures hinders the application of this approach. However, most of the weights (γ and β in Proposition 4 and 5) will become a very small positive number which is close to 0, thus dropping some of the mixtures by setting the corresponding weights to be 0 can significantly reduce the number of mixtures and not hurt the accuracy of the approximation. In practice, we can order the weights and keep only the biggest m^M weights (and mixtures) for a small natural number M , and then we normalize the weights and carry the procedure to the next period. A more robust approach would be first to standardize the Gaussian density of every mixture and then order the standardized weights to keep the biggest m^M mixtures, instead of ordering the raw weights³.

If \mathbf{u}_k is non-Gaussian, more specifically, if $\mathbf{u}_k \sim p_{\mathbf{u}}(\cdot)$, then we can approximate $p_{\mathbf{u}}(\cdot)$ using weighted Gaussians:

$$p_{\mathbf{u}}(\mathbf{u}) \approx \tilde{p}_{\mathbf{u}}(\mathbf{u}) \sum_{i=1}^{\check{m}} \check{\alpha}_i \phi\left(\mathbf{u} - \check{\mathbf{a}}_i, \check{\mathbf{P}}_i\right), \quad \sum_{i=1}^{\check{m}} \check{\alpha}_i = 1 \quad (95)$$

Then the predicting distribution in Proposition 4 should be modified, and we present the result in the following proposition.

Proposition 6

For a nonlinear and non-Gaussian state space model given in (58)-(59), if the distribution of the noise \mathbf{u} is non-Gaussian, and approximated by (95), then, the prediction distribution when time is k , i.e., $p(\mathbf{X}_{k+1}|\mathcal{Z}_k)$, is given by:

$$p(\mathbf{X}_{k+1}|\mathcal{Z}_k) = \sum_{i=1}^{J_{k+1}} \gamma_{k+1i} \phi(\mathbf{X}_{k+1} - \mathbf{c}_{k+1i}, \mathbf{Q}_{k+1i}) \quad (96)$$

where

$$\mathbf{c}_{k+1i} = E_{\phi_s}[\mathbf{F}(\mathbf{X}_k)] + \check{\mathbf{a}}_t \quad (97)$$

$$\mathbf{Q}_{k+1i} = E_{\phi_s}\left[(\mathbf{F}(\mathbf{X}_k) - \mathbf{F}(\mathbf{b}_{ks}))(\mathbf{F}(\mathbf{X}_k) - \mathbf{F}(\mathbf{b}_{ks}))^T\right] + \check{\mathbf{P}}_t \quad (98)$$

$$\gamma_{k+1i} = \beta_{ks} * \check{\alpha}_t \quad (99)$$

$$J_{k+1} = I_k * \check{m}. \quad (100)$$

for $s = 1, \dots, I_k$ and $t = 1, \dots, \check{m}$.

In Proposition 6, we denote E_{ϕ_s} as the expectation under s -th mixture of the filtering distribution $p(\mathbf{X}_k|\mathcal{Z}_k)$. And, β_{ks} in (99) is the s -th weight with respect to the s -th mixture of $p(\mathbf{X}_k|\mathcal{Z}_k)$.

³Thanks to Jean-Jacques Forneron for pointing this out.

$\check{\mathbf{a}}_t$ in (97), $\check{\mathbf{P}}_t$ in (98), and $\check{\alpha}_t$ in (99) for $t = 1, \dots, \check{m}$ are the mean, variance and weight associated with the t -th mixture from the approximation (95).

For both the Gaussian model in Section 2 and the non-Gaussian model in Section 3, we assume we start the filtering with Gaussian μ on \mathbf{X}_0 . However, this assumption can be relaxed. If μ is non-Gaussian, then we can approximate it by the following finite Gaussian mixture model:

$$\mu \approx \tilde{\mu} = \sum_{i=1}^{\check{m}} \check{\alpha}_i \phi \left(\mathbf{X}_0 - \check{\mathbf{a}}_i, \check{\mathbf{P}}_i \right). \quad (101)$$

And the filtering distribution in Proposition 3 can be further generalized accordingly. The next proposition describes this result.

Proposition 7

For a nonlinear and non-Gaussian state space model given in (58)-(59), if the probability measure on the initial state, μ is non-Gaussian, and approximated by (101), then the filtering distribution when time is 0, i.e., $p(\mathbf{X}_0 | \mathbf{Z}_0)$ is given by:

$$p(\mathbf{X}_0 | \mathbf{Z}_0) = \sum_{i=1}^{I_0} \beta_{0i} \phi(\mathbf{X}_0 - \mathbf{b}_{0i}, \mathbf{P}_{0i}) \quad (102)$$

where

$$\mathbf{b}_{0i} = \check{\mathbf{a}}_s + \mathbf{K}_{0i} \left(\mathbf{Z}_0 - E_{\phi_s^-} [\mathbf{H}(\mathbf{X}_0)] - \mathbf{a}_t \right) \quad (103)$$

$$\mathbf{S}_{0i} = E_{\phi_s^-} \left[(\mathbf{H}(\mathbf{X}_0) - \mathbf{H}(\check{\mathbf{a}}_s)) (\mathbf{H}(\mathbf{X}_0) - \mathbf{H}(\check{\mathbf{a}}_s))^T \right] + \mathbf{P}_t \quad (104)$$

$$\mathbf{C}_{0i} = E_{\phi_s^-} \left[(\mathbf{X}_0 - \check{\mathbf{a}}_s) (\mathbf{H}(\mathbf{X}_0) - \mathbf{H}(\check{\mathbf{a}}_s))^T \right] \quad (105)$$

$$\mathbf{K}_{0i} = \mathbf{C}_{0i} \mathbf{S}_{0i}^{-1} \quad (106)$$

$$\mathbf{P}_{0i} = \check{\mathbf{P}}_s - \mathbf{K}_{0i} \mathbf{S}_{0i} \mathbf{K}_{0i}^T \quad (107)$$

$$\beta_{0i} = \frac{\alpha_t \phi \left(\mathbf{Z}_0 - E_{\phi_s^-} [\mathbf{H}(\mathbf{X}_0)] - \mathbf{a}_t, \mathbf{S}_{0i} \right)}{\sum_{s=1}^{\check{m}} \sum_{t=1}^m \alpha_t \phi \left(\mathbf{Z}_0 - E_{\phi_s^-} [\mathbf{H}(\mathbf{X}_0)] - \mathbf{a}_t, \mathbf{S}_{0i} \right)} \quad (108)$$

$$I_0 = \check{m} * m, \quad (109)$$

for $s = 1, \dots, \check{m}$ and $t = 1, \dots, m$.

In Proposition 7, we denote $E_{\phi_s^-}$ as the expectation of \mathbf{X}_0 under the s -th mixture in the approximation (101). $\check{\mathbf{a}}_s$ in (103)-(105), and $\check{\mathbf{P}}_s$ in (107) for $s = 1, \dots, \check{m}$ are the mean and variance from the approximation (101). Also, \mathbf{a}_t in (103), \mathbf{P}_t in (104), and α_t in (108) for $t = 1, \dots, m$, are the mean, variance and weight associated with the t -th mixture from the approximation (60).

3.4 Algorithm

We denoted the filter through Proposition 3-5 as $\hat{\mathfrak{F}}_\mu^{\text{ng}}$. As in Section 2, the expectations in Proposition 3-5 admit no closed form solution generally and would be evaluated based on Quasi Monte Carlo approach, and we call this filtering as Quasi Monte Carlo Kalman filter for nonlinear and non-Gaussian state space models, denoted as $\hat{\mathfrak{F}}_\mu^{\text{ng}}$ with $\mathbf{X}_0 \sim \mu$. When F and H are linear function, we have analytic solution for the mean and variance of the transformation associated with them, thus the implementation of Proposition 3-5 can be simplified. When they are nonlinear, Quasi Monte Carlo based numerical approach would be used for the implementation. In the following, we present two algorithms: Algorithm 2 is for linear models and Algorithm 3 is for nonlinear models.

4 Stability

4.1 Gaussian model

The problem of filter stability was raised as early as in Kalman's 1960 seminal paper. The proof of the stability of Kalman filter for linear and Gaussian model strongly relies on the fact that the Kalman filter consists of a linear recursion and a Riccati equation, which is not available for most of the existing nonlinear filtering methods. However, in our approach, since we transform the nonlinear system into linear system by linearization (7)-(8), the analysis of stability based on the linear recursion and Riccati equation can be borrowed to discuss the stability of our nonlinear filtering approach (33)-(42). We begin by reviewing the definitions of various types of stability for dynamic systems.

Definition 3 (Lyapunov stable)

The filter (33)-(42) is said to be Lyapunov stable if $\forall \delta > 0, \exists \delta' > 0$ s.t. $\|\bar{\mathbf{X}}_0 - \bar{\mathbf{X}}_0'\| < \delta$ implies $\forall t \geq 0 \|\bar{\mathbf{X}}_t - \bar{\mathbf{X}}_t'\| < \delta'$, where $\bar{\mathbf{X}}_t$ and $\bar{\mathbf{X}}_t'$ satisfy (33)-(42) with initial conditions $\bar{\mathbf{X}}_0 = \mathbf{X}$ and $\bar{\mathbf{X}}_0' = \mathbf{X}' \forall \mathbf{X}, \mathbf{X}'$.

Definition 4 ((asymptotically) stable)

The filter (33)-(42) is said to be (asymptotically) stable if $\|\bar{\mathbf{X}}_t - \bar{\mathbf{X}}_t'\| \rightarrow 0$ as $t \rightarrow \infty$, where $\bar{\mathbf{X}}_t$ and $\bar{\mathbf{X}}_t'$ satisfy (33)-(42) with initial conditions $\bar{\mathbf{X}}_0 = \mathbf{X}$ and $\bar{\mathbf{X}}_0' = \mathbf{X}' \forall \mathbf{X}, \mathbf{X}'$.

Definition 5 (exponentially stable)

The filter (33)-(42) is said to be exponentially stable if $\exists a > 0, b > 0, \delta > 0$ s.t. $\|\bar{\mathbf{X}}_0 - \bar{\mathbf{X}}_0'\| < \delta$ implies $\|\bar{\mathbf{X}}_t - \bar{\mathbf{X}}_t'\| < a \|\bar{\mathbf{X}}_0 - \bar{\mathbf{X}}_0'\| e^{-bt} \forall t > 0$, where $\bar{\mathbf{X}}_t$ and $\bar{\mathbf{X}}_t'$ satisfy (33)-(42) with initial conditions $\bar{\mathbf{X}}_0 = \mathbf{X}$ and $\bar{\mathbf{X}}_0' = \mathbf{X}' \forall \mathbf{X}, \mathbf{X}'$.

Algorithm 2: QMCKF for linear and non-Gaussian model

- Step 1: For given $p_{\mathbf{v}}(\cdot)$, compute (and store) α_i, \mathbf{a}_i and \mathbf{P}_i for $i = 1, 2, \dots, m$.
 - Step 2: Compute $\mathbf{b}_{0i}, \mathbf{P}_{0i}, \mathbf{K}_{0i}$ and β_{0i} according to Prop.2 for $i = 1, 2, \dots, m$.
 - Step 3: Compute $\mathbf{c}_{k+1i}, \mathbf{Q}_{k+1i}$ and γ_{k+1i} for $i = 1, 2, \dots, J_{k+1}$, where $J_{k+1} = I_k$, according to Proposition 4.
 - Step 4: Compute (and store) $\hat{\mathbf{X}}_{k+1|k}$ and $\hat{\mathbf{P}}_{k+1|k}$.
 - Step 5: Compute $\mathbf{b}_{k+1i}, \mathbf{P}_{k+1i}, \mathbf{K}_{k+1i}$ and β_{k+1i} for $s = 1, 2, \dots, J_{k+1}$, $t = 1, 2, \dots, m$ according to Proposition 5. Set $I_{k+1} = J_{k+1} * m$.
 - Step 6: Compute (and store) $\hat{\mathbf{X}}_{k+1|k+1}$ and $\hat{\mathbf{P}}_{k+1|k+1}$.
 - Step 7: Repeat Step 3-6 for $k = 0, 1, 2, \dots, N$.
-

Algorithm 3: QMCKF for nonlinear and non-Gaussian model

- Step 1: For given $p_{\mathbf{v}}(\cdot)$, compute (and store) α_i, \mathbf{a}_i and \mathbf{P}_i for $i = 1, 2, \dots, m$.
 - Step 2: Generate a Box-Muller transformed halton sequence $\left\{ \mathbf{X}_k^{(g)} \right\}_{g=1}^G$ from $\phi(\mathbf{X}_k - \mathbf{b}_{ki}, \mathbf{P}_{ki})$. Compute and store $\mathbf{c}_{k+1i} = \frac{1}{G} \sum_{g=1}^G \mathbf{F}\left(\mathbf{X}_k^{(g)}\right) + \mu_{\mathbf{u}}$ and $\mathbf{Q}_{k+1i} = \frac{1}{G} \sum_{g=1}^G \left(\mathbf{F}\left(\mathbf{X}_k^{(g)}\right) - \mathbf{c}_{k+1i} \right) \left(\mathbf{F}\left(\mathbf{X}_k^{(g)}\right) - \mathbf{c}_{k+1i} \right)^T + \Sigma_{\mathbf{u}}$.
 - Step 3: Repeat Step 2 for $i = 1, 2, \dots, J_{k+1}$, where $J_{k+1} = I_k$. Compute and store $\hat{\mathbf{X}}_{k+1|k}$ and $\hat{\mathbf{P}}_{k+1|k}$.
 - Step 4: Generate a Box-Muller transformed halton sequence $\left\{ \mathbf{X}_{k+1s}^{(g)} \right\}_{g=1}^G$ and $\left\{ \mathbf{v}_{k+1t}^{(g)} \right\}_{g=1}^G$ from $\phi(\mathbf{c}_{k+1s}, \mathbf{Q}_{k+1s})$ and $\phi(\mathbf{v} - \mathbf{a}_t, \mathbf{P}_t)$ as we did in Step 2. Then generate $\mathbf{Z}_{k+1i}^{(g)} = \mathbf{H}\left(\mathbf{X}_{k+1s}^{(g)}\right) + \mathbf{v}_{k+1t}^{(g)}$. Compute $\bar{\mathbf{Z}}_{k+1i} = \frac{1}{G} \sum_{g=1}^G \mathbf{Z}_{k+1i}^{(g)}$, $\mathbf{V}_{k+1i} = \frac{1}{G} \sum_{g=1}^G \left(\mathbf{Z}_{k+1i}^{(g)} - \bar{\mathbf{Z}}_{k+1i} \right) \left(\mathbf{Z}_{k+1i}^{(g)} - \bar{\mathbf{Z}}_{k+1i} \right)^T$, and $\mathbf{S}_{k+1i} = \frac{1}{G} \sum_{g=1}^G \left(\mathbf{X}_{k+1s}^{(g)} - \mathbf{c}_{k+1s} \right) \left(\mathbf{Z}_{k+1i}^{(g)} - \bar{\mathbf{Z}}_{k+1i} \right)^T$
 - Step 5: Compute $\mathbf{K}_{k+1i} = \mathbf{S}_{k+1i} \mathbf{V}_{k+1i}^{-1}$, $\mathbf{P}_{k+1i} = \mathbf{Q}_{k+1s} - \mathbf{K}_{k+1i} \mathbf{V}_{k+1i} \mathbf{K}_{k+1i}^T$, and $\mathbf{b}_{k+1i} = \mathbf{c}_{k+1s} + \mathbf{K}_{k+1i} (\mathbf{Z}_{k+1i} - \bar{\mathbf{Z}}_{k+1i})$.
 - Step 6: Repeat Step 4-5 for $s = 1, 2, \dots, J_{k+1}$, $t = 1, 2, \dots, m$. Set $I_{k+1} = J_{k+1} * m$. Compute and store $\hat{\mathbf{X}}_{k+1|k+1}$ and $\hat{\mathbf{P}}_{k+1|k+1}$.
 - Step 7: Repeat Step 2-6 for $k = 0, 1, 2, \dots, N$.
-

Here, Lyapunov stability is the weakest definition. It says, in a dynamic system, if two solutions starting from two different initials that are close to each other, then, the two solutions would be always close to each other. Asymptotic stability is a stronger definition, and it requires the distance between the two solutions would finally disappear. The strongest definition is the exponential stability, which guarantees a minimal rate of decay of the distance between the two solutions. There are other notions of stability for a dynamic system, but the above three are the most closely related to our interests. In the next, we introduce two more definitions: observable and uniformly observable. They are the most important concepts in the discussion of linear and nonlinear filtering. In their 1968 book, Bucy and Joseph pointed out that: "...suppose a nonlinear model is not observable, then it is rather wasteful to perform the computations necessary to determine the conditional distribution, because any estimate, including the optimal one, will perform poorly."

Definition 6 (Observability)

The system (31)-(32) is said to be observable if the observability Gram

$$\mathcal{O}_{l,n} = \sum_{i=n}^l \left(\prod_{j=n}^{i-1} \hat{\mathbf{B}}_j \right)^T \hat{\mathbf{A}}_{1i}^T \left(\hat{\Sigma}_{c_i} + \hat{\mathbf{A}}_{2i}^T \Sigma_{\eta} \hat{\mathbf{A}}_{2i} \right)^{-1} \hat{\mathbf{A}}_{1i} \left(\prod_{j=n}^{i-1} \hat{\mathbf{B}}_j \right) \quad (110)$$

for $l > n$ is positive-definite for some l, n .

Definition 7 (Uniform Observability)

The system (31)-(32) is uniformly observable if \exists integer L and $\alpha_1 > 0, \alpha_2 > 0$ s.t.

$$0 \leq \alpha_1 \mathbf{I} \leq \mathcal{O}_{l,l-L} \leq \alpha_2 \mathbf{I} \quad (111)$$

$\forall l \geq L$.

Observability means if $\mathbf{X} \neq \mathbf{X}'$, then $\bar{\mathbf{X}}_0 = \mathbf{X}$ and $\bar{\mathbf{X}}_0 = \mathbf{X}'$ will give rise to different observation sequence $\{\mathbf{Z}_t\}_{t \geq 1}$. The concepts of observability and uniformly observability are easy to understand, and the problem is how to verify whether the observability or uniformly observability hold for a specific nonlinear state space model (1)-(2). We describe a sufficient condition for the uniform observability in Lemma 1.

Assumption 1

$\{\mathbf{X}_t\}$ is Feller.⁴

⁴A Markov process $X = \{X_t\}$ is Feller if

- $\lim_{x \rightarrow y} p(X_t | X_0 = x) = p(X_t | X_0 = y)$
- $\text{plim}_{t \downarrow 0} X_t = x$ given $X_0 = x$

For both of our Gaussian state space model (1)-(2) and non-Gaussian model (58)-(59), if F is Lipschitz, then $\{X_t\}$

Lemma 1

Under Assumption 1, and if \mathbf{H}_1 has a uniformly continuous inverse, then the system (31)-(32) is uniformly observable.

Proof. See R. van Handel (2009) ■

According to Lemma 1, for the system (31)-(32) to be uniformly observable, we need both \mathbf{F} and \mathbf{H}_1 to be smooth enough. Thus, when \mathbf{F} or \mathbf{H}_1 is non-smooth, we do not have stability of nonlinear filtering because the filter would not be observable. Also, the assumptions needed in Lemma 1 can be relaxed if we restrict the space of state \mathcal{X} further. For example, when \mathcal{X} is finite, a sufficient and necessary condition for uniform observability can be derived (R. van Handel, 2009). The difficulty we have here when \mathcal{X} is non-compact is if \mathbf{F} or \mathbf{H}_1 is not smooth enough, it would be impossible to infer the value of state variable given the observations, and then observability and stability of filtering would collapse.

Now we are in a position to discuss the stability of the filtering (33)-(42). Based on (33) and (41), $\bar{\mathbf{X}}_t$ can be written as

$$\bar{\mathbf{X}}_t = (\mathbf{I} - \mathbf{K}_t \hat{\mathbf{A}}_{1t}) \hat{\mathbf{B}}_{t-1} \bar{\mathbf{X}}_{t-1} + \mathbf{K}_t \mathbf{Z}_t - \mathbf{K}_t \hat{\mathbf{A}}_{1t} \hat{\mathbf{b}}_{t-1} - \mathbf{K}_t \hat{\mathbf{a}}_t + \hat{\mathbf{b}}_{t-1} \quad (112)$$

This is the Riccati equation for $\bar{\mathbf{X}}_t$, and the discussion of the stability of (33)-(42) will be based on this equation.

The proof is based on the Lyapunov's direct method (also called Lyapunov's second method) and Lypaunov stability theorem. The idea of this method is to find a Lyapunov function $V(\mathbf{X})$ (also called "energy") associated with the state variable, and if the energy is strictly decreasing over time and bounded below by 0, the energy would go to 0, which means the state variable will converge to the zero state. We refer Kalman and Bertram (1960) and Kalman (1963) for the details of this method. We first, in Lemma 2 and Lemma 3, prove the error covariance in (42) is bounded from above and below. They are the first step in proving the stability of nonlinear filtering, and in Theorem 2, we would define a Lyapunov function, check whether it is indeed decreasing and complete the proof.

Lemma 2

If the system (31)-(32) is uniformly observable, then $\forall \tau \geq 1$, there exists a constant p_τ s.t.

$$\mathbf{P}_t \leq p_\tau \mathbf{I},$$

$$\forall t \geq \tau.$$

is Feller. However, if \mathcal{X} is discrete, we do not need Lipschitz continuity of F for $\{X_t\}$ to be Feller.

Lemma 3

If the system (31)-(32) is uniformly observable, and Σ_0 is positive-definite, then $\forall \tau \geq 1, \exists q_\tau$ s.t.

$$\mathbf{P}_t \geq q_\tau \mathbf{I},$$

$$\forall t \geq \tau.$$

Theorem 2

Under Assumption 1 and if the system (31)-(32) is uniformly observable, then \mathfrak{F}^g is exponentially stable in $\bar{\mathbf{X}}_t$, i.e.: $\exists \kappa_1 > 0, \vartheta_1 > 0, \delta > 0$ such that

$$\forall t > 0, \quad \|\bar{\mathbf{X}}_t^\mu - \bar{\mathbf{X}}_t^\nu\| < \kappa_1 \|\mu_0 - \nu_0\| e^{-\vartheta_1 t}, \quad (113)$$

whenever $\|\mu_0 - \nu_0\| < \delta$. μ and ν are well defined Gaussian initial measures on \mathbf{X}_0 with mean μ_0 and ν_0 , and $\|\cdot\|$ is some suitable norm.

Note that Theorem 2 is about the stability of \mathfrak{F}^g : (33)-(42), which is infeasible in practice since we do not know the true value of expectations in (33)-(42). In the next theorem, we would discuss the stability of QMCKF $\hat{\mathfrak{F}}_\mu^g$: (46)-(51).

Theorem 3

Under Assumption 1 and if the system (31)-(32) is uniformly observable, then $\hat{\mathfrak{F}}^g$ is exponentially stable in $\bar{\mathbf{X}}_t^G$, i.e.: $\exists \kappa_2 > 0, \vartheta_2 > 0, \delta > 0$ such that

$$\forall t > 0, \quad \|\bar{\mathbf{X}}_t^{G,\mu} - \bar{\mathbf{X}}_t^{G,\nu}\| < \kappa_2 \|\mu_0 - \nu_0\| e^{-\vartheta_2 t}, \quad (114)$$

whenever $\|\mu_0 - \nu_0\| < \delta$. μ and ν are well defined Gaussian initial measures on \mathbf{X}_0 with mean μ_0 and ν_0 , and $\|\cdot\|$ is some suitable norm.

Except the uniform observability, another assumption called uniform controllability⁵ is usually required for the stability of linear filtering. Controllability is to rule out the possibility of knowing the value of state variable exactly based on the observations, otherwise there is no filtering problem. However, in our case, it can be verified that if we assume the initial measure μ is a well defined

⁵The system (31)-(32) is said to be uniformly controllable if the controllability Gram $\mathcal{C}_{l,n}$ is such that there are positive constants α_3 and α_4 :

$$0 \leq \alpha_3 \mathbf{I} \leq \mathcal{C}_{l,l-L} \leq \alpha_4 \mathbf{I}$$

$\forall l \geq L$, where $\mathcal{C}_{l,n}$ is defined by:

$$\mathcal{C}_{l,n} \equiv \sum_{i=n}^l \left(\prod_{j=n}^{i-1} \hat{\mathbf{B}}_j \right)^T \left(\hat{\Sigma}_{e_i} + \Sigma_\epsilon \right)^{-1} \left(\prod_{j=n}^{i-1} \hat{\mathbf{B}}_j \right)$$

Gaussian distribution, than uniform controllability is satisfied automatically. More generally, if μ is non-Gaussian (for example, in the regime switching models), then controllability would be guaranteed by various assumptions on non-degeneracy of the noises ε and η in (1)-(2).⁶ For example, the assumption (A1) (a) and (A1) (b) in Douc et al. (2004) for Hamilton filter, and assumption (H1) in Del Moral and Guionnet (2001) for particle filter. Other similar examples can also be found in R. van Handel (2009, Assumption 3.1 and 3.2).

In the next theorem, we will discuss the convergence of filter $\hat{\mathfrak{F}}^g$ to \mathfrak{F}^g . Del Moral and Guionnet (2001) established the uniform convergence of the particle filter by writing the difference between the exact filter and approximate filter as a telescoping sum, and we will follow their idea. More specifically, we represent \mathfrak{F}^g as $\bar{\mathbf{X}}_t = \mathbb{F}(\mathbf{Z}_t, \bar{\mathbf{X}}_{t-1}) \equiv \mathbb{F}_t \bar{\mathbf{X}}_{t-1}$ and $\hat{\mathfrak{F}}^g$ as $\bar{\mathbf{X}}_t^G = \mathbb{F}^G(\mathbf{Z}_t, \bar{\mathbf{X}}_{t-1}^G) \equiv \mathbb{F}_t^G \bar{\mathbf{X}}_{t-1}^G$ for function \mathbb{F} and \mathbb{F}^G . Then $\bar{\mathbf{X}}_t^G - \bar{\mathbf{X}}_t$ can be written as

$$\bar{\mathbf{X}}_t^G - \bar{\mathbf{X}}_t = \sum_{i=1}^t \left\{ \mathbb{F}_t \cdots \mathbb{F}_{i+1} \mathbb{F}_i^G \bar{\mathbf{X}}_{i-1}^G - \mathbb{F}_t \cdots \mathbb{F}_{i+1} \mathbb{F}_i \bar{\mathbf{X}}_{i-1}^G \right\}. \quad (115)$$

The property of $\bar{\mathbf{X}}_t^G - \bar{\mathbf{X}}_t$ is summarized in Theorem 4.

Theorem 4

Under Assumption 1 and if the system (31)-(32) is uniformly observable, then there exists a finite positive Υ_t such that

$$\|\bar{\mathbf{X}}_t^G - \bar{\mathbf{X}}_t\| \leq \Upsilon_t D_G \quad (116)$$

where D_G is the discrepancy of the sequence used in QMCKF with simulation size G .

The value of Υ_t is shown in appendix, and it is independent to the types of sequence or the simulation size used in QMCKF. The implication of Theorem 4 is strong. It says for every time t , the smaller the discrepancy of the sequence we use in QMCKF (46)-(51) is, the smaller $\|\bar{\mathbf{X}}_t^G - \bar{\mathbf{X}}_t\|$ would be. Note that we have shown in Section 2 that $\bar{\mathbf{X}}_t$ from $\hat{\mathfrak{F}}^g$ is equivalent to the true mean of the state variable.

4.2 Non-Gaussian model

We first extend the idea of observability for non-Gaussian models.

Definition 8 (Uniform Observability for non-Gaussian models)

The system (58)-(59) is uniformly observable if \exists integer L and $\alpha_5 > 0, \alpha_6 > 0$ s.t. $\forall i = 1, \dots, m$

⁶Non-degeneracy of the noises means (1) the transition density of state variable is bounded from below and above (this is equivalent to assume the state variable is uniformly ergodic), and (2) the conditional law of observation has a positive density.

$$\forall l \geq L, \quad 0 \leq \alpha_5 \mathbf{I} \leq \mathcal{O}_{l,l-L,i} \leq \alpha_6 \mathbf{I} \quad (117)$$

where

$$\mathcal{O}_{l,n,i} = \sum_{k=n}^l \left(\prod_{j=n}^{k-1} \hat{\mathbf{B}}_j \right)^T \hat{\mathbf{A}}_k^T \left(\hat{\Sigma}_{\mathbf{c}_k} + \mathbf{P}_i \right)^{-1} \hat{\mathbf{A}}_k \left(\prod_{j=n}^{k-1} \hat{\mathbf{B}}_j \right) \quad (118)$$

such that \mathbf{P}_i is the variance matrix of i -th mixture in (60), and $\hat{\mathbf{B}}_j$ and $\hat{\mathbf{A}}_i$ are defined as those in Gaussian model ⁷.

It's easy to see that uniform observability holds for non-Gaussian models if the process of the state variable is Feller, the nonlinear function \mathbf{H} is smooth enough and \mathbf{P}_i is well-defined for all $i = 1, \dots, m$.

We also present an assumption on the non-Gaussian density $p_{\mathbf{v}}$. Suppose $\phi_{\theta}(x)$ is Gaussian probability measure on x with θ as parameters, and denote

$$\mathcal{G} \equiv \left\{ \phi_{\theta}(x), x \in \mathcal{X} : \theta \in \Theta \subset \mathbb{R}^{\dim(\theta)} \right\}$$

as the location-scale family of Gaussian density parameterized by θ . Also, define $\mathcal{C}_{\mathcal{G}, \mathbb{P}}$ as a class of density functions, that is

$$\mathcal{C}_{\mathcal{G}, \mathbb{P}} \equiv \left\{ f_{\mathbb{P}}(x) : f_{\mathbb{P}}(x) = \int_{\Theta} \phi_{\theta}(x) \mathbb{P}(d\theta) \right\}$$

and \mathbb{P} is a probability measure on Θ .

Assumption 2

We assume $p_{\mathbf{v}} \in \mathcal{C}_{\mathcal{G}, \mathbb{P}}$ ⁸.

We are interested in the error of finite Gaussian mixture approximation of a given density function. Under Assumption 2, Li and Barron (1999) quantify this approximation error in relative entropy sense and show that the rate of convergence of this approximation is the inverse of the number of mixtures in the approximation. The bound of the approximation error can be found in

⁷ $\mathbf{H}_2(\mathbf{X}_t; \theta)$ in Gaussian model is assumed to be $\mathbf{1}$ in non-Gaussian model, so $\hat{\mathbf{A}}_{2i}$ in (110) degenerates to $\mathbf{1}$ and we denote $\hat{\mathbf{A}}_i$ for $\hat{\mathbf{A}}_{1i}$

⁸Examples of distribution in $\mathcal{C}_{\mathcal{G}, \mathbb{P}}$:

- Generalized Hyperbolic distribution: Student's t , normal-inverse Gaussian, normal-inverse Gamma, variance-gamma and others
- Normal-exponential-gamma distribution: Laplace and others
- Exponentially modified Gaussian

their paper, and more general results are given in Genovese and Wasserman (2000), Ghosal (2001), Ghosal and van der Vaart (2001, 2007).

Theorem 5

Under Assumption 1 and 2, and if the system (58)-(59) is uniformly observable, then the filtered distribution of filter \mathfrak{F}^{ng} is asymptotically stable in relative entropy sense, i.e.,

$$\mathcal{H}(p^\mu(\mathbf{X}_k|\mathbf{Z}_k)|p^\nu(\mathbf{X}_k|\mathbf{Z}_k)) \rightarrow 0 \quad \text{as } k, m \rightarrow \infty \quad (119)$$

where μ and ν are well defined Gaussian initial measure on \mathbf{X}_0 .

Theorem 6

Under Assumption 1 and 2, and if the system (58)-(59) is uniformly observable, then the filtered distribution of filter $\hat{\mathfrak{F}}^{\text{ng}}$ is asymptotically stable in relative entropy sense, i.e.,

$$\mathcal{H}(p_G^\mu(\mathbf{X}_k|\mathbf{Z}_k)|p_G^\nu(\mathbf{X}_k|\mathbf{Z}_k)) \rightarrow 0 \quad \text{as } k, m, G \rightarrow \infty \quad (120)$$

where μ and ν are well defined Gaussian initial measure on \mathbf{X}_0 , and G is simulation size in QMCKF.

5 Parameter estimation

The classical state space problem involves not only determining the mean and distribution of unobservable state variable, but also learning the value of unknown parameters in the state space models given the training set of observations. In this section, we develop and present two approaches for the estimation of unknown parameters in state space models, and we also discuss the asymptotic properties of these estimators.

5.1 Off-line estimation

5.1.1 Likelihood-based estimator

The off-line estimation method is based on the likelihood of the observation. For the sake of simplicity, we consider the following state space model:

$$\mathbf{X}_{k+1} = \mathbf{F}(\mathbf{X}_k; \theta) + \mathbf{u}_k \quad (121)$$

$$\mathbf{Z}_k = \mathbf{H}(\mathbf{X}_k; \theta) + \mathbf{v}_k \quad (122)$$

In the model (121)-(122), \mathbf{X} is the unobservable state variable, \mathbf{Z} is the observation, and $\theta \in \Theta$ is the unknown parameter in the model. Without loss of generality, we assume $\mathbf{u}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}}, \Sigma_{\mathbf{u}})$, $\mathbf{v}_k \sim p_{\mathbf{v}}(\cdot)$, and $\mathbf{X}_0 \sim \mu$ which is Gaussian with mean μ_0 and Σ_0 .

If the model (121)-(122) is Gaussian, that is, if \mathbf{v}_k is a Gaussian noise with $\mathcal{N}(\mu_{\mathbf{v}}, \Sigma_{\mathbf{v}})$, then given the observation of $\mathfrak{Z}_N = \{\mathbf{Z}_k\}_{k=1}^N$ and based on the filtering approach QMCKF (Algorithm 1), the approximated log-likelihood function, $L_{\mathfrak{g},N}^G(\theta)$, for parameter θ can be written as:

$$L_{\mathfrak{g},N}^G(\theta) \equiv \log f_{\mathfrak{g}}^G(\theta; \mathfrak{Z}_N) = \sum_{k=1}^N \log f_{\mathfrak{g}}^G(\theta; \mathbf{Z}_k) \quad (123)$$

$$= -\frac{1}{2} \sum_{k=1}^N \left[\log(2\pi) + \log |\mathbf{S}_k^G| + (\mathbf{Z}_k - \bar{\mathbf{Z}}_k^G)^T (\mathbf{S}_k^G)^{-1} (\mathbf{Z}_k - \bar{\mathbf{Z}}_k^G) \right], \quad (124)$$

where

$$\mathbf{S}_k^G = \frac{1}{G} \sum_{i=1}^G \left(H(\mathbf{X}_k^{(g)-}) - \bar{\mathbf{Z}}_k^G \right) \left(H(\mathbf{X}_k^{(g)-}) - \bar{\mathbf{Z}}_k^G \right)^T + \boldsymbol{\Sigma}_{\mathbf{v}}, \quad (125)$$

and

$$\bar{\mathbf{Z}}_k^G = \frac{1}{G} \sum_{i=1}^G H(\mathbf{X}_t^{(g)-}), \quad (126)$$

with G being the simulation size in QMCKF. Then for the Gaussian model, the maximum simulated likelihood estimator is defined as:

$$\hat{\theta}_{\mathfrak{g},MSLE} \equiv \operatorname{argmax}_{\theta \in \Theta} L_{\mathfrak{g},N}^G(\theta). \quad (127)$$

While if the model (121)-(122) is non-Gaussian, that is, if \mathbf{v}_k is a non-Gaussian noise, then given the observation of $\mathfrak{Z}_N = \{\mathbf{Z}_k\}_{k=1}^N$ and based on our proposed filtering approach (Algorithm 3), the approximated log-likelihood function, $L_{\mathfrak{ng},N}^G(\theta)$, for parameter θ can be written as:

$$L_{\mathfrak{ng},N}^G(\theta) \equiv \log f_{\mathfrak{ng}}^G(\theta; \mathfrak{Z}_N) = \sum_{k=1}^N \log f_{\mathfrak{ng}}^G(\theta; \mathbf{Z}_k) = \\ \sum_{k=1}^N \log \left[\sum_{s,t}^{I_k} \frac{\alpha_t}{\sqrt{2\pi |\mathbf{S}_{kt}^G|}} \exp \left(-(\mathbf{Z}_k - \bar{\mathbf{Z}}_{ks})^T (\mathbf{S}_{kt}^G)^{-1} (\mathbf{Z}_k - \bar{\mathbf{Z}}_{ks}) \right) \right], \quad (128)$$

where \mathbf{S}_{kt}^G and $\bar{\mathbf{Z}}_{ks}$ are based on Algorithm 3. Then the maximum simulated likelihood estimator is selected to maximize $L_{\mathfrak{ng},N}^G(\theta)$ for non-Gaussian model:

$$\hat{\theta}_{\mathfrak{ng},MSLE} \equiv \operatorname{argmax}_{\theta \in \Theta} L_{\mathfrak{ng},N}^G(\theta). \quad (129)$$

Various numerical optimization method, including EM algorithm can be used to compute the MSLE.

5.1.2 Asymptotic property of likelihood-based estimator

We describe the asymptotic property of $\hat{\theta}_{g,MSLE}$ and $\hat{\theta}_{ng,MSLE}$ by first presenting one more assumption. Define $L_{g,N}(\theta) = \log f_g(\theta; \mathbf{z}_N)$, and $L_{ng,N}(\theta) = \log f_{ng}(\theta; \mathbf{z}_N)$ as the log-likelihood function for Gaussian model 1-2 and non-Gaussian model with observation \mathbf{z}_N .

Assumption 3

This assumption is based on the assumptions of Theorem 2.1 of Newey and McFadden (1994). We assume

- (i) $\theta \in \Theta$, a compact subset of $\mathbb{R}^{\dim(\theta)}$
- (ii) $E[\log f_g(\theta; \mathbf{z}_N)], \frac{1}{N}L_{g,N}(\theta), E[\log f_{ng}(\theta; \mathbf{z}_N)], \frac{1}{N}L_{ng,N}(\theta)$ are continuous in θ
- (iii) $\theta_0 \equiv \operatorname{argmax}_{\theta \in \Theta} E[\log f(\theta; \mathbf{z}_N)]$ is unique for $f(\theta; \mathbf{z}_N) = f_g(\theta; \mathbf{z}_N)$ or $f(\theta; \mathbf{z}_N) = f_{ng}(\theta; \mathbf{z}_N)$
- (iv) $\frac{1}{N}L_{g,N}(\theta) \rightarrow E[\log f_g(\theta; \mathbf{z}_N)], \frac{1}{N}L_{ng,N}(\theta) \rightarrow E[\log f_{ng}(\theta; \mathbf{z}_N)]$ in probability uniformly in $\theta \in \Theta$ as $N \rightarrow \infty$

Theorem 7

Under Assumption 1 and 3 (and Assumption 2 for the non-Gaussian model), and if the system (121)-(122) is uniformly observable, then $\hat{\theta}_{g,MSLE}$ (or $\hat{\theta}_{ng,MSLE}$), is consistent if $G \rightarrow \infty$ as $N \rightarrow \infty$ (and $m \rightarrow \infty$ for the non-Gaussian model).

5.2 On-line estimation

As an on-line estimation method, the augmented approach (or joint estimation) is generally used in other filtering for state space models, such as EKF, UKF and PF. More specifically, for the model (121)-(122), we first assume the unknown parameter is time-varying and follows a random walk process:

$$\theta_{k+1} = \theta_k + \xi_k$$

where $\xi_k \sim \mathcal{N}(0, \iota)$ for a small positive number ι . Then we construct an augmented matrix for the state variable and unknown parameter:

$$\vec{\mathbf{X}}_t = \begin{bmatrix} \mathbf{X}_t \\ \theta_t \end{bmatrix},$$

and the original state space model (121)-(122) can be rewritten as

$$\vec{\mathbf{X}}_{k+1} = \vec{\mathbf{F}}(\vec{\mathbf{X}}_k) + \vec{\mathbf{u}}_k \quad (130)$$

$$\mathbf{Z}_k = \vec{\mathbf{H}}(\vec{\mathbf{X}}_k) + \vec{\mathbf{v}}_k, \quad (131)$$

where

$$\begin{aligned}\vec{\mathbf{F}}(\vec{\mathbf{X}}_t) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{F}(\mathbf{X}_k; \theta_k) \\ \theta_k \end{bmatrix}, \\ \vec{\mathbf{u}}_k &= \begin{bmatrix} \mathbf{u}_k \\ \xi_k \end{bmatrix}, \\ \vec{\mathbf{H}}(\vec{\mathbf{X}}_k) &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{H}(\mathbf{X}_k; \theta_k) \\ \theta_k \end{bmatrix}, \\ \vec{\mathbf{v}}_k &= [\mathbf{v}_k].\end{aligned}$$

Then QMCKF can be applied to this augmented state space model (130)-(131). The filtered estimation of the mean of θ_k would be the (time-varying) point estimation of the unknown parameters, and the filtered distribution on θ_k can be used for inference.

6 Examples and discussions

In this section, we provide several examples to illustrate the performance of QMCKF for various state space model. Sections 6.1-6.3 are for the simulated data and Section 6.4 is an example on real data.

6.1 Linear and Gaussian model

In this example, we will examine the performance of QMCKF and compare it with Kalman filter and Particle filter (Gordon et al, 1993; Kitagawa, 1993; detail about the algorithm of Particle filter is in the Appendix B). Since this model is linear, we know in this case KF is the optimal filter, and it would be the benchmark when we compare the QMCKF and PF for this model.

The model is

$$X_{t+1} = 0.99 * X_t + \varepsilon_t \quad (132)$$

$$Z_t = X_t + \eta_t \quad (133)$$

where $X_0 \sim N(0.1, 0.001)$, $\varepsilon_t \sim N(0, 0.01)$, and $\eta_t \sim N(0, 0.01)$. The sample size is 250. Although this model is linear, in QMCKF (Algorithm 1, we will treat it as nonlinear model and the simulation size is $G = 1000$). For Particle filter, the simulation size is $N = 50000$. It takes 0.11 second for KF, 0.97 second for QMCKF and 3.31 seconds for PF.

Figure 1 is the plot of posteriori estimation of mean of the state variable X_t in (132)-(133). As we can see, both of QMCKF and PF perform well in estimating the mean of the state variable. But

the performance of Particle filter in estimating the posteriori distribution is worse than QMCKF (the pdf's of posteriori distribution for $t = 50, 100, 150, 200$ are shown in Figure 2, and the cdf's are shown in Figure 3). This problem of PF is not alleviated even the simulation size is increased. The estimation of posteriori distribution based on PF is dependent on the way the pdf is smoothed and it differs when different bandwidth is chosen.

6.2 Nonlinear and Gaussian model

In this example, we will compare the performance of QMCKF and PF for the following nonlinear and Gaussian model:

$$X_{t+1} = 0.99 * X_t + \frac{1}{300} * X_t^2 + 0.01 + \varepsilon_t \quad (134)$$

$$Z_t = \exp(X_t) + \eta_t \quad (135)$$

where $X_0 \sim N(0.1, 0.001)$, $\varepsilon_t \sim N(0, 0.05)$, and $\eta_t \sim N(0, 0.05)$. The sample size is 250.

In the above model, both of the functions associated with the measurement equation and dynamic equation on state variable. The simulation size for QMCKF (Algorithm 1) is $G = 1000$ and $N = 50000$ for PF. It takes 0.98 second for QMCKF and 3.89 seconds for PF.

In Figure 4, we report the result of filtered estimation of X for both QMCKF and PF, and the true value of state variable is also plotted in Figure 4. Similar to the linear and Gaussian example in Section 6.1, both the QMCKF and PF can estimate the mean of the state variable accurately. However, the filtered distribution based on PF is problematic. The density and cdf of the filtered distribution of QMCKF and PF are given in Figure 5 and Figure 6 respectively. As shown in these two figures, compared with QMCKF, PF gives more nonsmooth estimation of the density and cdf. We also run the PF on the same simulated data set for 1000 different times and we report the MSE in estimating the mean of the state variable in Figure 7, the estimated density of the filtered distribution in Figure 8 and the estimated cdf in Figure 9. In Figure 7, the horizontal line in red is the result of QMCKF and the circles are the results of 1000 trials of PF. We can find the MSE of PF is volatile and QMCKF is better than PF with smaller MSE for almost all of the trials. In Figure 8, the result of QMCKF is plotted in black solid line and the 5 of the 1000 trials of PF are plotted in dashed line with 5 different colors. In Figure 9, the red solid line is for the cdf based on QMCKF and dots in 5 different colors are the cdf based on 5 of the 1000 trials of PF. It is clear that for the same data set and same algorithm, PF will lead to very different estimations on the density and cdf of the filtered distribution. This implies that likelihood-based inference method according to PF could be unrobust in practice.

Table 1: Gaussian Mixture Approximation of Standard Normal Distribution: APPROX1

i	α_i	a_i	P_i
1	0.0005	-0.01	1.84
2	0.999	0	1
3	0.0005	0.01	1.84

Table 2: Gaussian Mixture Approximation of Standard Normal Distribution: APPROX2

i	α_i	a_i	P_i
1	0.0067	-1.75	0.62
2	0.8667	0	0.8
3	0.0067	1.75	0.62

6.3 Linear and non-Gaussian model.

6.3.1 Revisit Linear and Gaussian model

In this example, we revisit the linear and Gaussian model in Section 6.1. Specifically, we consider :

$$X_{t+1} = X_t + \varepsilon_t \quad (136)$$

$$Z_t = X_t + \eta_t \quad (137)$$

where $X_0 \sim \mathcal{N}(10.5, 0.02)$, $\varepsilon_t \sim \mathcal{N}(0, 1)$, and $\eta_t \sim \mathcal{N}(0, 0.01)$.

In the above model, ε_t is assumed to be Gaussian, but in QMCKF, we treat it as non-Gaussian model and approximate its density using Gaussian mixtures, that is

$$\frac{1}{\sqrt{2}} \exp\left(-\frac{1}{2}\varepsilon_t^2\right) \approx \sum_{i=1}^m \alpha_i \phi(\varepsilon_t - a_i, P_i).$$

In Table 1 and 2, we provide two approximations for $m = 3$ and the resulting approximated densities with the true density are plotted in Figure 10. It's obvious that the first approximation is better in minimizing the distance between the approximated density and the true. We use these two approximation to examine the performance of QMCKF (Algorithm 2) and check the effect of accuracy of Gaussian mixture approximation to the estimation of filtered density.

The estimation of the mean of state variable, the density and the cdf of the filtered distribution based on the first Gaussian mixture approximation are given in Figure 11, Figure 12, and Figure 13, and those based on the second approximation are given in Figure 14-16. For comparison, the results based on Kalman filter and PF are also given in these figures. By comparing these figures we

Table 3: Gaussian Mixture Approximation of Students' t_3

i	α_i	a_i	P_i
1	0.0033	-4.50	2500
2	0.008	-3.49	6.6049
3	0.012	-2.68	0.64
4	0.0273	-1.80	0.3025
5	0.0853	-1.18	4.2849
6	0.73	0	0.7407
7	0.0853	1.18	4.2849
8	0.0273	1.80	0.3025
9	0.012	2.68	0.64
10	0.008	3.49	6.6049
11	0.0033	4.50	2500

can find the accuracy of estimation on the filtering distribution is strongly dependent on wellness in the Gaussian mixture approximation.

6.3.2 Linear and non-Gaussian model

In this example, we consider a linear and non-Gaussian model:

$$X_{t+1} = 0.99 * X_t + 0.1 + \varepsilon_t \quad (138)$$

$$Z_t = \varphi + X_t + \eta_t \quad (139)$$

where $X_0 \sim \mathcal{N}(5, 0.01)$, $\varepsilon_t \sim t_3$, and $\eta_t \sim \mathcal{N}(0, 0.001)$.

In this model, the noise ε_t follows Student's t distribution with degree of freedom 3. We approximate the density of t_3 using Gaussian mixture model:

$$\frac{2}{\pi\sqrt{3}} \left(1 + \frac{\varepsilon_t^2}{3}\right)^{-2} \approx \sum_{i=1}^m \alpha_i \phi(\varepsilon_t - a_i, P_i).$$

We set $m = 11$, and provide the values of parameters in approximation in Table 3, and the resulting density is plotted in Figure 17.

The estimation of the mean of the state variable is given in Figure 18, and the density and cdf of the filtered distribution are given in Figure 19 and Figure 20 respectively. As we can see from this figures, the difference of the results between both QMCKF and PF, in estimating either the

mean or the distribution, is small, but QMCKF is much faster than PF (0.60 second for QMCKF with simulation size 1000, and 25.67 seconds for PF with simulation size 50000 or 270.73 seconds for PF with simulation size 500000).

7 Applications

In this section, we discuss two applications of our proposed QMCKF to two popular models in economics and finance, and use actual data to illustrate the approach. In the first example, we investigate a stochastic volatility model, and use finite Gaussian mixture model to approximate the noise to the model and run QMCKF to estimate the unobservable volatility of asset returns. We apply our proposed approach to the foreign exchange rate data between Sterling and Dollar from 10/01/1981 to 06/28/1985, and compare our results with Kim, Shephard, and Chib (1998). In the second example, we discuss a discrete time jump model. We show the jump size has a Gaussian mixture representation and can be well applied by QMCKF. We use a simulating example to illustrate performance of our approach. In addition, we use the 3-month Treasury bill rate data from 1991 to 1993 to estimate the jump probability and jump size, and connect the dates of jumps with major macroeconomic events.

7.1 Stochastic Volatility

7.1.1 The Model

This example is based on Kim, Shephard and Chib (1998). To model the time dependent variance of return on assets, we consider the following stochastic volatility (SV) model:

$$y_t = \beta e^{h_t/2} \varepsilon_t, \quad t \geq 1 \quad (140)$$

$$h_{t+1} = \mu + \varphi (h_t - \mu) + \sigma_\eta * \eta_t \quad (141)$$

$$h_1 \sim N\left(\mu, \frac{\sigma_\eta^2}{1 - \varphi^2}\right). \quad (142)$$

In the above model, y_t is the return on asset at time t , h_t is the unobservable log volatility at time t which follows a stationary mean-reverting process with the Gaussian initial measure on h_1 being the stationary distribution. ε_t and η_t are uncorrelated standard normal noise. $e^{h_t/2}$ is the volatility of return y_t at time t , σ_η is the volatility of the log volatility (h_t), and β is a constant scaling factor of the volatility of y_t . This model was first proposed by Taylor (1986) to model the price of financial assets, and was discussed by Hull and White (1987) and Jacquier, Polson and Rossi (1994), extended by Jacquier, Polson and Rossi (2004) for fat-tailed SV model.

By taking the logarithm of y_t^2 , we can transform the above model into a linear and non-Gaussian model:

$$z_t \equiv \log y_t^2 = h_t + \log \varepsilon_t^2 \quad (143)$$

Let $\tilde{\varepsilon}_t \equiv \log \varepsilon_t^2$, then (143) can be written as:

$$z_t = h_t + \tilde{\varepsilon}_t \quad (144)$$

where $\tilde{\varepsilon}_t \sim \log \chi_1^2$ with $E(\tilde{\varepsilon}_t) = -1.2704$ and $\text{Var}(\tilde{\varepsilon}_t) = 4.93$. Equation (144) and (141)-(142) construct a linear and non-Gaussian state space model for the return of asset. To estimate this model, Harvey, Ruiz and Shephard (1994) assume $\tilde{\varepsilon}_t \sim \mathcal{N}(-1.2704, 4.93)$ and apply a Kalman filter to the linearized SV model. A QMLE method is used in that paper to estimate the unknown parameters in the model. As pointed out by Kim, Shephard and Chib (1994), although the QMLE by Harvey, Ruiz and Shephard (1994) is consistent and asymptotically normally distributed, but the deviation of $\log \chi_1^2$ from Gaussian makes that QMLE perform poorly in small sample. Jacquier, Polson and Rossi (1994) develop a Bayesian MCMC method for the analysis of this model. Kim, Shephard and Chib (1994) design an offset mixture of normal distribution to approximate the likelihood and they argue this can improve the efficiency of MCMC procedure.

7.1.2 FX: Sterling vs. Dollar

Following Kim, Shephard and Chib (1994), in this example, we apply QMCKF to (144) and (141), and examine its performance in read data. The dataset is the daily observations of weekday close exchange rates for Sterling/Dollar from 10/01/1981 to 06/28/1985, downloaded from the website of FED. The sample size is $n = 937$ (946 in Kim, Shephard and Chib (1994)). We denote r_t as the Sterling/Dollar exchange rate at time t , and the mean-corrected returns is given by:

$$y_t = 100 * \left\{ \frac{r_t - r_{t-1}}{r_{t-1}} - \frac{1}{n} \sum_{i=1}^n \left(\frac{r_i - r_{i-1}}{r_{i-1}} \right) \right\}.$$

Plot of $|y_t|$ is given in Figure 22.

To initiate QMCKF, we need first to approximate the density of $\log \chi_1^2$ by Gaussian mixtures. More specifically, for $\tilde{\varepsilon}_t$ in (144), we approximate its density by following model:

$$\tilde{p}_{\tilde{\varepsilon}}(\tilde{\varepsilon}_t) = \sum_{i=1}^m \alpha_i \phi(\tilde{\varepsilon}_t - a_i, P_i) \quad (145)$$

where m is the number of mixtures in the approximation, while the true pdf of $\log \chi_1^2$ is:

$$p_{\tilde{\varepsilon}}(\tilde{\varepsilon}_t) = \frac{1}{\sqrt{2\Gamma(\frac{1}{2})}} \exp\left(\frac{1}{2}\tilde{\varepsilon}_t - \frac{1}{2}\exp(\tilde{\varepsilon}_t)\right). \quad (146)$$

Table 4: Approximation of the distribution of $\log \chi_1^2$ using 7 Gaussian Mixtures

i	α_i	a_i	P_i
1	0.00730	-11.40039	5.79596
2	0.10556	-5.24321	2.61369
3	0.00002	-9.83726	5.17950
4	0.04395	1.50746	0.16735
5	0.34001	-0.65098	0.64009
6	0.24566	0.52478	0.34023
7	0.25750	-2.35859	1.26261

For a preset m , the parameters $\{\alpha_i, a_i, P_i\}_{i=1}^m$ in 145 are selected to minimize the distance between $\tilde{p}_{\tilde{\varepsilon}}(\tilde{\varepsilon}_t)$ and $p_{\tilde{\varepsilon}}(\tilde{\varepsilon}_t)$. In Table 4, we present the values of parameters for this approximation based on 7 mixtures. This is taken from Kim, Shephard and Chib (1994) paper. Also, in Table 5, we provide another approximation based on 9 mixtures, and the approximated densities of these two approximation with the true density are plotted in Figure 21. It's obvious that this 9-mixture approximation is better than the 7-mixture approximation of Kim, Shephard and Chib (1994) in minimizing the approximation error. In the sequel, we will use this 9-mixture approximation to analysis the above SV model.

We report the filtered estimate of the volatility ($\exp(\frac{h_t}{2})$) in Figure 23, and the filtered estimate of the long-run mean of the volatility ($\exp(\frac{\mu}{2})$) in Figure 24. For comparison, we also plot the estimate of $\exp(\frac{\mu}{2})$ from Kim, Shephard and Chib (1994) as a horizontal line in red. It is easy to see that our estimate of $\exp(\frac{\mu}{2})$ is time-varying, and become close to Kim, Shephard and Chib's estimate very quickly.

7.2 Jumps

7.2.1 The Model

It is well-known that the financial report, macroeconomic announcement, release of new products or wars will heavily impact the financial market and change the price of stocks and commodities, and interest rates. This demonstrates the importance of including jumps into the dynamic models on financial assets. A large body of literature has documented compelling empirical evidence for jumps. See Babbs and Webber (1997), Das and Sundaram (1999), Johannes (2004), Le, Singleton and Dai (2010) for researches on interest rates, and Engle and Ng (1993), Bakshi, Cao, and Chen (1997), Bates (2000), Duffie, Pan, and Singleton (2000), and Pan (2002) for researches on stock

Table 5: Approximation of the distribution of $\log \chi_1^2$ using 9 Gaussian Mixtures

i	α_i	a_i	P_i
1	0.01013	-9.25636	3.14958
2	0.00807	-7.79639	0.80596
3	0.09743	-4.38321	1.36069
4	0.01305	-6.51726	0.70950
5	0.03295	1.35546	0.11235
6	0.32671	-0.58298	0.62139
7	0.24546	0.55478	0.39823
8	0.25512	-2.16859	1.05116
9	0.01008	1.92825	0.10648

prices and returns. To provide a framework that allows for general specifications and maximum likelihood based estimation of jump models, we consider the following jump model:

$$y_t = h_t + \tilde{J}_t \times Z_t \quad (147)$$

$$h_{t+1} = \phi h_t + \eta_t \quad (148)$$

$$\eta_t \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2) \quad (149)$$

$$\tilde{J}_t = \mathbb{J}(J_t) \quad (150)$$

$$J_t \sim Ber(\lambda) \quad (151)$$

$$\lambda \sim Beta(\alpha, \beta) \quad (152)$$

$$Z_t | J_t = j \sim \mathcal{N}(\mu_j, \sigma_j^2) \quad (153)$$

In the above model, we consider y_t as the observation, and h_t as the unobservable state variable with an *i.i.d* noise η . We assume the jump component \tilde{J} is a simple linear function of J , where J follows a Beta-Bernoulli process. That is, we assume J follows a Bernoulli distribution with λ as the jump arrival rate, and λ is based on a Beta distribution with parameters α and β . In (153), we assume the jump size Z , conditioning on J , is normally distributed. Depends on different function form of \mathbb{J} , the model (147)-(153) can be interpreted differently, and applied to model various jump process. We discuss this later with more details.

By combining (151) and (152), we can compute the possibility of $J_t = j$ for $j = 0, 1$ as following:

$$P(J_t = j) = \frac{\Gamma(2)\Gamma(\alpha+j)\Gamma(1-j+\beta)\Gamma(\alpha+\beta)}{\Gamma(j+1)\Gamma(2-j)\Gamma(1+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)}, \quad (154)$$

Table 6: Parameter values of a simulating example for the jump model

α	β	ϕ	μ_η	σ_η^2	μ_0	σ_0^2	μ_1	σ_1^2
2	80	0.90	0	0.1	0	0.5	0	20

where Γ is denoted as the gamma function. It can be easily verified that if we plug (154) into (153), the jump term $\tilde{J}_t \times Z_t$ has a Gaussian mixture representation:

$$\varepsilon_t \equiv \tilde{J}_t \times Z_t \sim P(J_t = 0) \times \mathcal{N}(\mu_0, \sigma_0^2) + P(J_t = 1) \times \mathcal{N}(\mu_1, \sigma_1^2), \quad (155)$$

where $P(J_t = 0) + P(J_t = 1) = 1$.

We have a straightforward interpretation for (155). We can consider J as a hidden state variable that describes the regime of the financial market, and $P(J_t = 0)$ and $P(J_t = 1)$ as the possibilities of the different regimes. With the help of \mathbb{J} function, \tilde{J}_t can be used to model “jump happens or not” if we let $\mathbb{J}(J) = J$, or “ordinary noises v.s. extreme events” if we let $\mathbb{J}(J) = 1$. Given the financial market is in different regimes, the noise to the observation y is drawn from normal distributions with different means and variances. That is, this model can describe the jumps in the mean and jumps in the volatility of the price and return of financial assets. In the remainder of this section, we adopt $\mathbb{J}(J) = 1$.

This model can also be considered as a simple discrete time stochastic volatility jump (SVJ) model suggested by Bates (1996), and it can be further generalized to capture other statistical properties of prices and return of financial assets. For example, nonlinear transformation can be applied in (147)-(148) to develop a nonlinear term structure models. See Glasserman and Kou (2003), Le, Singleton and Dai (2010), and Chernov and Muller (2012) for examples. Also, asymmetric jumps can be incorporated into this model to capture the leptokurtic feature of the return distribution of assets. See Eraker (2001), Kou (2002), Eraker, Johannes and Polson (2003), Kou and Wang (2004), and Aït-Sahalia, Cacho-Diaz, and Laeven (2015) for examples.

7.2.2 A Simulating Example

We consider model (147)-(153), and adopt the parameter values as given in Table 6. Based on the values of α and β given in this table and (154), we can compute the possibilities of different jumps. That is:

$$\begin{aligned} P(J_t = 0) &= 0.9756, \\ P(J_t = 1) &= 0.0244. \end{aligned}$$

We simulate a sample of $\{J_t, Z_t, h_t, y_t\}_{t=1}^T$, where $T = 500$, and display them in Figure 25. Specifically, Figure 25 presents the observations (y , the top panel), the true values of the state

variable (h , the second panel from above), the true values of the jump component (\tilde{J} or J , the third panel from above), and the true values of the jump term ($\tilde{J} \times Z$, the bottom panel). Please note that even when $J_t = 1$, the jump size, $\tilde{J}_t \times Z_t$, may not be significant so that we cannot observe sharp kinks in the jump size in this case. For example, $t = 300$ in Figure 25.

Our interest is to estimate the jump size and the jump probability given the observations using QMCKF. Figure 26 presents the filtered estimation of the jump size ($E(\tilde{J}_t \times Z_t | y_t, y_{t-1}, \dots)$, the top panel), the estimated jump probability ($P(J_t = 1 | y_t, y_{t-1}, \dots)$, the middle panel), and compares it with the true jump component (the bottom panel). As shown in Figure 26, we can correctly identify 11 ones out of the 15 jumps, where significant change of the data happens.

7.2.3 Short Rates: Jumps and Macroeconomic Events

There is substantial evidence to support the finding that the dynamics of interest rates can be affected by macroeconomic events, such as Federal Reserve monetary policy announcements, wars and political events, release of job data. See Piazzesi (2000), Eraker, Johannes and Polson (2003), Johannes (2004), Dubinsky et al.(2006) and Kim and Wright (2015). Following Johannes (2004), we use 3-month Treasury bill (T-bill) rate from 1991-1993 to investigate the connection between the occurrence of jumps and macroeconomic events. The data set is downloaded from the website of FED, and it has 479 daily observations. We apply the model (147)-(153) and QMCKF to estimate the jump size and jump probability, and identify the events that generated these jumps.

Figure 27 shows the time series of short rate changes (in percentage, the top panel), and filtered estimation of the jump probability (the bottom panel). From Figure 27 , we can identify approximately 10 jumps during the period from 1991 to 1993, which gives an arrival rate of 0.0209. Also, when these jump happens, the short rate will change by more than 10 percents in all the cases.

Following Johannes (2004), we connect the dates of jumps with the following major macroeconomic events: (A) 01/09/1991, the outbreak of the Gulf War; (B) 02/01/1991, U.S. unemployment announcement and comments by the Federal Reserve; (C) 08/19/1991, the collapse of the Soviet Union; (D) 08/21/1991, the emergence of Boris Yeltsin as leader of the remnants of the Soviet Union, (E) 12/20/1991, the Federal Reserve lowers the discount rate; (F) 4/9/1992, large Japanese equity market decline; (G) 7/2/1992, the Federal Reserve lowers the discount rate; (H) 9/4/92, a U.S. unemployment announcement; and, (I) October 1992, the Bush-Clinton presidential debates. We also identify a jump at 11/06/1991, but fail to connect it with any macroeconomic events. The results is reported in Figure 28

8 Conclusion

State space model has numerous applications in economics. In this paper, we propose a unified set of tools for nonlinear and non-Gaussian state space models, including filtering, likelihood evaluation, parameter estimation, and the asymptotic properties. This approach is simulation-based, but provides a simple and tractable estimation on the filtering distribution. The key features of our new approach are the least square based linearization and Gaussian mixture approximation of the non-Gaussian density of the noise in the model. This, combined with the efficient integration method of Quasi Monte Carlo, enables us to approximate the filtering distribution of complex models using the well-established Kalman filter. Simulation experiments show that our approach can achieve significant improvement not only in terms of computation speed and ease of implementation, but also in terms of accuracy and asymptotic property.

For the parameter estimation of the state space model. We present off-line method and on-line method to estimate the unknown parameters in the model. In practice, EM algorithm would be an efficient approach to compute the estimate. Especially, given the structure of the log-likelihood function, on-line version of the EM algorithm for the augmented state variable would be an efficient method to estimate the parameters, and we will leave this for future research.

Also, in the paper we prove our least square approximation can match the true first two moments of the state variable, but matching the true higher moments can improve the estimation of the filtered distribution further. Work continues on how to match the higher moments while keep the linear structure of the approximation.

References

- Aït-Sahalia, Y., J. Cacho-Diaz, and R.J.A. Laeven (2015): Modeling financial contagion using mutually exciting jump processes, *Journal of Financial Economics*, Vol. 117, No. 3
- Ahn, D.H., R.F. Dittmar, and A.R. Gallant (2002): Quadratic Term Structure Models: Theory and Evidence, *Review of Financial Studies*, 15(1), 243-288.
- Babbs, S., and N. Webber (1997): Term Structure Modelling under Alternative Official Regimes, in *Mathematics of Derivative Securities*, eds., M.A.H. Dempster and S.R. Pliska. Cambridge University Press, Cambridge, UK
- Bakshi, G., C. Cao, and Z. Chen (2012): Empirical Performance of Alternative Option Pricing Models, *The Journal of Finance*, Vol. 51, No. 5
- Bates, D.S. (1996): Jumps and Stochastic volatility: Exchange Rate Processes Implicity in Deutsche Mark Options, *The Review of Financial Studies*. volume 9, number 1, pages 69–107.
- Bates, D.S. (2000): Post-'87 crash fears in the S&P 500 futures option market, *Journal of Econometrics*, Vol. 94, 181-238
- Caffisch, R.E. (1998): Monte Carlo and quasi-Monte Carlo methods, *Acta Numerica*, pp. 1-49
- Chernov, M. and P. Mueller (2012): The term structure of inflation expectations, *Journal of Financial Economics* 106, 367-394.
- Creal, D., S.J. Koopman, and A. Lucas (2013): Generalized Autoregressive Score Models with Applications, *Journal of Applied Econometrics*, Vol. 28, No. 5, pp. 777-795, 2013.
- Das, S.R., and R.K. Sundaram (1999): Of Smiles and Smirks: A Term Structure Perspective, *The Journal of Financial and Quantitative Analysis* Vol. 34, No. 2 pp. 211-239
- Doucet, A., N. de Freitas, K.P. Murphy and S.J. Russell (2000): Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks, *Proc. UAI*
- Del Moral, P. (1996): Non Linear Filtering: Interacting Particle Solution, *Markov Processes and Related Fields*. 2 (4): 555–580.
- Del Moral, P. (1998): Measure Valued Processes and Interacting Particle Systems. Application to Non Linear Filtering Problems, *Annals of Applied Probability* (Publications du Laboratoire de Statistique et Probabilités, 96-15 ed.). 8 (2): 438–495.

Del Moral, P., and A. Guionnet (1999): Central limit theorem for nonlinear filtering and interacting particle systems, *The Annals of Applied Probability*. 9 (2): 275–297.

Del Moral, P., and A. Guionnet (1999): On the stability of Measure Valued Processes with Applications to filtering, *C. R. Acad. Sci. Paris*. 39 (1): 429–434.

Del Moral, P., and A. Guionnet (2001): On the stability of interacting processes with applications to filtering and genetic algorithms, *Annales de l’Institut Henri Poincaré*. 37 (2): 155–194.

Douc, R., E. Moulines, and T. Rydén (2004): Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime, *The Annals of statistics* 32 (5), 2254-2304

Dubinsky, A., M. Johannes, A. Kaeck, and N.J. Seeger (2019): Option pricing of earnings announcement risks, *The Review of Financial Studies*, Vol. 32, No. 2

Duffie, D., J. Pan, K. Singleton (2000): Transform Analysis and Asset Pricing for Affine Jump-Diffusions, *Econometrica*, Vol. 68, No. 6

Engle, R.F., and V.K. Ng (1993): Measuring and Testing the Impact of News on Volatility, *The Journal of Finance*, Vol. 48, No. 5

Eraker, B. (2001): Do stock prices and volatility jump? Reconciling evidence from spot and option prices. Unpublished working paper, Duke University.

Eraker, B., M. Johannes, and N. Polson (2003): The impact of jumps in equityindex volatilityand returns. *Journal of Finance* 58, 1269–1300.

Fernández-Villaverde, J., and J.F. Rubio-Ramírez (2007): Estimating macroeconomic models: a likelihood approach, *Review of Economic Studies* 74: 1059–1087.

Fernandez-Villaverde, J., P. Guerron-Quintana, J.F. Rubio-Ramirez, and M. Uribe (2011): Risk Matters: The Real Effects of Volatility Shocks, *American Economic Review*, American Economic Association, vol. 101(6), pages 2530-61.

Flury, T., and N. Shepard (2011): Bayesian inference based only on a simulated likelihood, *Econometric Theory* 27: 933-956.

Genovese, C. and L. Wasserman (2000): Rates of convergence for the Gaussian mixture sieve, *The Annals of Statistics*, Volume 28, Number 4, 1105-1127.

Glasserman, P., and S.G. Kou (2003): Measuring and Testing the Impact of News on Volatility, *Mathematical Finance*, Vol. 13, No. 3, 383-410.

Ghosal, S. and A.W. van der Vaart (2001): Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities, *The Annals of Statistics*. 29 1233–1263.

Ghosal, S. (2001): Convergence rates for density estimation with Bernstein polynomials, *The Annals of Statistics*. 29 1264–1280. MR1873330

Ghosal, S. and A. van der Vaart (2007): Convergence rates of posterior distributions for noniid observations, *The Annals of Statistics*, Vol. 35, No. 1, 192–223

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

Hamilton, J.D. (1989): A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, *Econometrica* Vol. 57, No. 2, pp. 357-384

Harvey, A.C., E. Ruiz, and N. Shephard (1994): Multivariate stochastic variance models, *Reviews of Economic Studies*, 61, 247-264.

Hull, J., and A. White (1987): The Pricing of options on assets with stochastic volatilities, *Journal of Finance*, 42, 281-300

Jacquier, E., N.G. Polson, and P.E. Rossi (1994): Bayesian analysis of stochastic volatility models, *Journal of Business and Economic Statistics*, 12, 371-417

Jacquier, E., N.G. Polson, and P.E. Rossi (2004): Bayesian analysis of stochastic volatility models with fat-tails and correlated errors, *Journal of Econometrics* Volume 122, Issue 1, Pages 185-212

Jazwinski, A.H. (1970): *Stochastic Processes and Filtering Theory*, New York: Academic Press

Johannes, M. (2004): The Statistical and Economic Role of Jumps in Continuous-Time Interest Rate Models, *The Journal of Finance*, Vol. 59, No. 1

Julier, S.J., and J.K. Uhlmann (1997): A New Extension of the Kalman Filter to Nonlinear Systems, in *Proceedings of AeroSense: The 11th International Symposium on Aerospace/Defence Sensing, Simulation and Controls*.

Kalman, R.E., J.F. Bertram (1960): Control System Analysis and Design via the Second Method of Lyapunov, *Journal of Basic Engrg.* 88: 371, 394.

Kalman, R.E. (1963): Lyapunov functions for the problem of Lur'e in automatic control, *Proceedings of the National Academy of Sciences of the United States of America*. 49 (2): 201–205.

Kim, D.H. and J.H. Wright (2015): Jumps in Bond Yields at Known Times, NBER Working Paper No. 20711

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

Kitagawa, G. (1996): Monte carlo filter and smoother for non-Gaussian nonlinear state space models, *Journal of Computational and Graphical Statistics*. 5 (1): 1–25

Kou, S.G. (2002): A Jump-Diffusion Model for Option Pricing, *Management Science*, Vol. 48, No. 8

Kou, S.G., and H. Wang (2004): Option pricing under a double exponential jump diffusion model, *Management Science*, Vol. 50, No. 9

Le, A., K.J. Singleton, and Q. Dai (2010): Discrete-Time Affine \mathbb{Q} Term Structure Models with Generalized Market Prices of Risk, *The Review of Financial Studies* , Vol. 23, No. 5 pp. 2184- 2227

Li, J.Q., and A.R. Barron (1999): Mixture Density Estimation. Conference on Neural Information Processing Systems

Lopes, H.F., and R.S. Tsay (2011): Particle Filters and Bayesian Inference in Financial Econometrics, *Journal of Forecast*. 30, 168–209

Maybeck, P.S. (1982): Stochastic models, estimation and control, Volume 3, New York: Academic Press

Merwe, R., A. Doucet, N. de Freitas, and E. Wan (2000): The unscented particle filter, *Proceedings of the Conference on Neural Information Processing Systems*

Morokoff, W.J., and R.E. Caflisch (1994): Quasi-random sequences and their discrepancies, *SIAM Journal of Scientific Computing*, 15, no. 6, 1251–1279

Newey, W. K., and D. McFadden (1994): Large sample estimation and hypothesis testing, *Handbook of Econometrics* Volume 4, Pages 2111-2245

Niederreiter, H. G. (1978): Quasi-Monte Carlo methods and pseudo-random numbers, *Bull. Amer. Math. Soc.* 84, no. 6, 957–1041

Niederreiter, H. G. (1992): Random Number Generation and Quasi-Monte Carlo Methods, Society for Industrial and Applied Mathematics.

Pan, J. (2002): The jump-risk premia implicit in options: evidence from an integrated time-series study, *Journal of Financial Economics*, 63, 3-50

Piazzesi, M. (2000): Bond Yields and the Federal Reserve, *Journal of Political Economy*, vol. 113, no. 2

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

Taylor, S.J. (1986): *Modeling Financial Time Series*, Chichester: John Wiley

Titterington, D.M., A.F.M. Smith, and U.E. Makov (1985): *Statistical Analysis of finite mixture distributions* Chichester: John Wiley

van der Merwe, R., and E. Wan (2001): The Square-Root Unscented Kalman Filter for State and Parameter Estimation, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6, (Salt Lake City, UT), pp. 3461–3464

van Handel, R. (2009): Uniform observability of hidden Markov models and filter stability for unstable signals, *The Annals of Applied Probability*, vol. 19, pp. 1172–1199

Wan, E.A., R. van der Merwe, and A.T. Nelson (2000): Dual Estimation and the Unscented Transformation, in *Advances in Neural Information Processing Systems*, ed. by S. Solla, T. Leen, and K.R. Muller, MIT Press, pages 666–672

A Proof

A.1 Proof of lemma

A.1.1 Proof of Lemma 2

For $t \geq \tau$, define a state estimate

$$\widetilde{\bar{X}}_t = \tilde{\mathcal{O}}_{t,t-\tau}^{-1} \sum_{i=t-\tau}^t \left(\prod_{j=t}^{i-1} \hat{B}_j \right)^T \hat{A}_{1i}^T \left(\hat{\Sigma}_{c_i} + \hat{A}_{2i} \Sigma_\eta \hat{A}_{2i}^T \right)^{-1} Z_i \quad (156)$$

where $\mathcal{O}_{l,n} = \left(\prod_{j=n}^{l-1} \hat{B}_j \right)^T \tilde{\mathcal{O}}_{l,n} \left(\prod_{j=n}^{l-1} \hat{B}_j \right)$.

Since this estimate is not the optimal one and not of minimum variance, we have the inequality

$$P_t \leq \text{var} \left(\bar{X}_t - \widetilde{\bar{X}}_t \right)$$

It can be verified that

$$\begin{aligned} \text{var} \left(\bar{X}_t - \widetilde{\bar{X}}_t \right) &= \tilde{\mathcal{O}}_{t,t-\tau}^{-1} + \\ &\quad \tilde{\mathcal{O}}_{t,t-\tau}^{-1} \sum_{j=t-\tau}^{t-1} \left(\sum_{i=t-\tau}^j \left(\prod_{k=t}^{i-1} \hat{B}_k \right)^T \hat{A}_{1i}^T \left(\hat{\Sigma}_{c_i} + \hat{A}_{2i} \Sigma_\eta \hat{A}_{2i}^T \right)^{-1} \hat{A}_{1i} \left(\prod_{k=t}^{i-1} \hat{B}_k \right) \right) \times \\ &\quad \left(\prod_{k=j+1}^{t-1} \hat{B}_k \right) \left(\hat{\Sigma}_{e_j} + \Sigma_\epsilon \right) \left(\prod_{k=j+1}^{t-1} \hat{B}_k \right)^T \times \\ &\quad \left(\sum_{i=t-\tau}^j \left(\prod_{k=t}^{i-1} \hat{B}_k \right)^T \hat{A}_{1i}^T \left(\hat{\Sigma}_{c_i} + \hat{A}_{2i} \Sigma_\eta \hat{A}_{2i}^T \right)^{-1} \hat{A}_{1i} \left(\prod_{k=t}^{i-1} \hat{B}_k \right) \right) \tilde{\mathcal{O}}_{t,t-\tau}^{-1} \end{aligned}$$

An upper bound of $\text{var} \left(\bar{X}_t - \widetilde{\bar{X}}_t \right)$ could be obtained based on the assumption of uniform observability (Definition 7).

A.1.2 Proof of Lemma 3

Define $Q_t = P_t^{-1} = \left[(P_t^{-1})^{-1} + \hat{A}_{1t}^T \left(\hat{\Sigma}_{c_t} + \hat{A}_{2t} \Sigma_\eta \hat{A}_{2t}^T \right)^{-1} \hat{A}_{1t} \right]$ and let

$$\begin{aligned} \check{Q}_t &= Q_t - \hat{A}_{1t}^T \left(\hat{\Sigma}_{c_t} + \hat{A}_{2t} \Sigma_\eta \hat{A}_{2t}^T \right)^{-1} \hat{A}_{1t} = (P_t^-)^{-1} \\ \check{Q}_{t+1}^- &= \left(\hat{B}_{t+1}^T \right)^{-1} Q_t \hat{B}_{t+1}^{-1} \end{aligned} \quad (157)$$

From (33)-(42), we have

$$\check{Q}_t = \left(\left(\check{Q}_t^- \right)^{-1} + \hat{\Sigma}_{e_{t-1}} + \Sigma_\epsilon \right)^{-1} \quad (158)$$

and

$$\check{\mathbf{Q}}_{t+1}^- = \left(\hat{\mathbf{B}}_{t+1}^T \right)^{-1} \check{\mathbf{Q}}_t \hat{\mathbf{B}}_{t+1}^{-1} + \left(\hat{\mathbf{B}}_{t+1}^T \right)^{-1} \hat{\mathbf{A}}_{1t}^T \left(\hat{\Sigma}_{c_t} + \hat{\mathbf{A}}_{2t} \Sigma_\eta \hat{\mathbf{A}}_{2t}^T \right)^{-1} \hat{\mathbf{A}}_{1t} \hat{\mathbf{B}}_{t+1}^{-1} \quad (159)$$

Based on (158)-(159), $\check{\mathbf{Q}}_t$ and $\check{\mathbf{Q}}_{t+1}^-$ can be interpreted as posteriori and priori estimation error covariance matrix for the following linear and Gaussian system:

$$\check{\mathbf{X}}_{t+1} = \left(\hat{\mathbf{B}}_t^T \right)^{-1} \check{\mathbf{X}}_t + \left(\hat{\mathbf{B}}_t^T \right)^{-1} \hat{\mathbf{A}}_{1t}^T \check{\boldsymbol{\varepsilon}}_t \quad (160)$$

$$\check{\mathbf{Z}}_t = \check{\mathbf{X}}_t + \check{\boldsymbol{\eta}}_t \quad (161)$$

where

$$\check{\boldsymbol{\varepsilon}}_t \sim N \left(0, \left(\hat{\Sigma}_{c_t} + \hat{\mathbf{A}}_{2t} \Sigma_\eta \hat{\mathbf{A}}_{2t}^T \right)^{-1} \right) \quad (162)$$

$$\check{\boldsymbol{\eta}}_t \sim N \left(0, \left(\hat{\Sigma}_{e_{t-1}} + \Sigma_\epsilon \right)^{-1} \right) \quad (163)$$

From the stability analysis of Kalman filter for linear and Gaussian system, we know

$$\begin{aligned} \check{\mathbf{Q}}_t &\leq \left[\sum_{i=t-\tau-1}^{t-1} \left(\prod_{j=i+1}^t \hat{\mathbf{B}}_j \right) \left(\hat{\Sigma}_{e_i} + \Sigma_\epsilon \right) \left(\prod_{j=i+1}^t \hat{\mathbf{B}}_j^T \right) \right]^{-1} \\ &+ \sum_{i=t-\tau}^{t-1} \left(\prod_{j=t}^i \hat{\mathbf{B}}_j^T \right) \hat{\mathbf{A}}_{1i}^T \left(\hat{\Sigma}_{c_i} + \hat{\mathbf{A}}_{2i} \Sigma_\eta \hat{\mathbf{A}}_{2i}^T \right)^{-1} \hat{\mathbf{A}}_{1i} \left(\prod_{j=t}^i \hat{\mathbf{B}}_j \right) \end{aligned}$$

According to (157), there is a positive constant \tilde{q}_N such that

$$\begin{aligned} \check{\mathbf{Q}}_t &\leq \left[\sum_{i=t-\tau-1}^{t-1} \left(\prod_{j=i+1}^t \hat{\mathbf{B}}_j \right) \left(\hat{\Sigma}_{e_i} + \Sigma_\epsilon \right) \left(\prod_{j=i+1}^t \hat{\mathbf{B}}_j^T \right) \right]^{-1} \\ &+ \sum_{i=t-\tau}^{t-1} \left(\prod_{j=t}^i \hat{\mathbf{B}}_j^T \right) \hat{\mathbf{A}}_{1i}^T \left(\hat{\Sigma}_{c_i} + \hat{\mathbf{A}}_{2i} \Sigma_\eta \hat{\mathbf{A}}_{2i}^T \right)^{-1} \hat{\mathbf{A}}_{1i} \left(\prod_{j=t}^i \hat{\mathbf{B}}_j \right) \\ &\leq \tilde{q}_\tau \mathbf{I} \end{aligned}$$

Then $\mathbf{P}_t = \mathbf{Q}_t^{-1} \geq \tilde{q}_\tau^{-1} \mathbf{I} \equiv q_\tau \mathbf{I}$

A.2 Proof of propositions

A.2.1 Proof of Proposition 1

We can plug (7)-(8) into (3)-(6), and we would have:

$$\begin{aligned} \tilde{p}(\mathbf{X}_t | \mathcal{Z}_t) &= \tilde{c}_t \tilde{p}(\mathbf{X}_t | \mathcal{Z}_{t-1}) \tilde{p}(\mathbf{Z}_t | \mathbf{X}_t) \\ &= \tilde{c}_t \tilde{p}(\mathbf{X}_t | \mathcal{Z}_{t-1}) \phi(\mathbf{Z}_t - \mathbf{A}_{1t} \mathbf{X}_t - \mathbf{a}_t, \mathbf{A}_{2t} \Sigma_\eta \mathbf{A}_{2t}^T + \Sigma_{c_t}) \end{aligned}$$

where $1/\bar{c}_t = \int \tilde{p}(\mathbf{X}_t|\mathcal{Z}_{t-1}) \tilde{p}(\mathbf{Z}_t|\mathbf{X}_t) d\mathbf{X}_t$. And

$$\begin{aligned}\tilde{p}(\mathbf{X}_t|\mathcal{Z}_{t-1}) &= \int \tilde{p}(\mathbf{X}_{t-1}|\mathcal{Z}_{t-1}) \tilde{p}(\mathbf{X}_t|\mathbf{X}_{t-1}) d\mathbf{X}_{t-1} \\ &= \int \tilde{p}(\mathbf{X}_{t-1}|\mathcal{Z}_{t-1}) \phi(\mathbf{X}_t - \mathbf{B}_{t-1}\mathbf{X}_{t-1} - \mathbf{b}_{t-1}, \Sigma_\varepsilon + \Sigma_{e_{t-1}}) d\mathbf{X}_{t-1}\end{aligned}$$

If we begin the filter with Gaussian μ , i.e., if we let $\tilde{p}(\mathbf{X}_0|\mathcal{Z}_{-1}) = \mu$, then it's easy to verify that $\tilde{p}(\mathbf{X}_t|\mathcal{Z}_t)$ and $\tilde{p}(\mathbf{X}_t|\mathcal{Z}_{t-1})$ have the Gaussian kernel.

A.2.2 Proof of Proposition 2

Elemental.

A.2.3 Proof of Proposition 3-7

We only provide the proof for Proposition 3 because the proofs for Proposition 4-7 are similar as this proof.

$$\begin{aligned}p(\mathbf{X}_0|\mathcal{Z}_0) &= \sum_{i=1}^m c_{0i} p(\mathbf{X}_0) \phi(\mathbf{Z}_0 - \mathbf{H}(\mathbf{X}_0) - \mathbf{a}_i, \mathbf{P}_i) \\ &= \sum_{i=1}^m c_{0i} \phi(\mathbf{X}_0 - \mu_0, \Sigma_0) \phi(\mathbf{Z}_0 - \mathbf{A}_0\mathbf{X}_0 - \mathbf{a}_0 - \mathbf{a}_i, \mathbf{P}_i + \Sigma_{c_0})\end{aligned}$$

where we make use of $p(\mathbf{X}_0) = \mu \equiv \phi(\mathbf{X}_0 - \mu_0, \Sigma_0)$, and apply the linearization:

$$\mathbf{H}(\mathbf{X}_0) \approx \mathbf{A}_0\mathbf{X}_0 + \mathbf{a}_0 + \mathbf{c}_0$$

such that $c_0 \sim \mathcal{N}(0, \Sigma_{c_0})$. Notice that in the above formula, $p(\mathbf{X}_0|\mathcal{Z}_0)$ is a weighed sum of product of two Gaussian densities. It's easy to verify that $\phi(\mathbf{X}_0 - \mu_0, \Sigma_0) \phi(\mathbf{Z}_0 - \mathbf{A}_0\mathbf{X}_0 - \mathbf{a}_0 - \mathbf{a}_i, \mathbf{P}_i)$ can be represented as another Gaussian density. Note that unlike the linear case, here $p(\mathbf{X}_0|\mathcal{Z}_0)$ is a Gaussian mixture, and may not be Gaussian.

A.3 Proof of theorems

A.3.1 Proof of Theorem 2

Define a Lyapunov function

$$V(\bar{\mathbf{X}}_t, t) = \bar{\mathbf{X}}_t^T \mathbf{P}_t^{-1} \bar{\mathbf{X}}_t \quad (164)$$

Based on Lemma 2 and Lemma 3, we can construct $\gamma_1(\|\bar{\mathbf{X}}_t\|) = p_N \|\bar{\mathbf{X}}_t\|^2$ and $\gamma_2(\|\bar{\mathbf{X}}_t\|) = q_N \|\bar{\mathbf{X}}_t\|^2$ and it is easy to verify that

$$0 \leq \gamma_1(\|\bar{\mathbf{X}}_t\|) \leq V(\bar{\mathbf{X}}_t, t) \leq \gamma_2(\|\bar{\mathbf{X}}_t\|) \quad (165)$$

From (33)-(42), we have

$$\bar{\mathbf{X}}_t = (\mathbf{I} - \mathbf{K}_t \hat{\mathbf{A}}_{1t}) \hat{\mathbf{B}}_{t-1} \bar{\mathbf{X}}_{t-1} + \mathbf{K}_t \mathbf{Z}_t - \mathbf{K}_t \hat{\mathbf{A}}_{1t} \hat{\mathbf{b}}_{t-1} - \mathbf{K}_t \hat{\mathbf{a}}_t + \hat{\mathbf{b}}_{t-1} \quad (166)$$

We only consider the homogeneous part of above equation, since the remaining part is bounded. Then

$$\begin{aligned} V(\bar{\mathbf{X}}_t, t) &\leq \bar{\mathbf{X}}_{t-1}^T \mathbf{P}_{t-1}^{-1} \mathbf{X}_{t-1} - \bar{\mathbf{X}}_t^T \hat{\mathbf{A}}_{1t}^T \left(\hat{\Sigma}_{c_t} + \hat{\mathbf{A}}_{2t} \Sigma_\eta \hat{\mathbf{A}}_{2t}^T \right)^{-1} \hat{\mathbf{A}}_{1t} \bar{\mathbf{X}}_t \\ &\quad - \bar{\mathbf{X}}_{t-1}^T \hat{\mathbf{B}}_{t-1}^T \hat{\mathbf{A}}_{1t}^T \mathbf{K}_t^T (\mathbf{P}_t^-)^{-1} \mathbf{K}_t \hat{\mathbf{A}}_{1t} \hat{\mathbf{B}}_{t-1} \bar{\mathbf{X}}_{t-1} \end{aligned}$$

By arranging the terms, we get

$$\begin{aligned} V(\bar{\mathbf{X}}_t, t) - V(\bar{\mathbf{X}}_{t-1}, t-1) &\leq -\bar{\mathbf{X}}_t^T \hat{\mathbf{A}}_{1t}^T \left(\hat{\Sigma}_{c_t} + \hat{\mathbf{A}}_{2t} \Sigma_\eta \hat{\mathbf{A}}_{2t}^T \right)^{-1} \hat{\mathbf{A}}_{1t} \bar{\mathbf{X}}_t \\ &\quad - \bar{\mathbf{X}}_{t-1}^T \hat{\mathbf{B}}_{t-1}^T \hat{\mathbf{A}}_{1t}^T \mathbf{K}_t^T (\mathbf{P}_t^-)^{-1} \mathbf{K}_t \hat{\mathbf{A}}_{1t} \hat{\mathbf{B}}_{t-1} \bar{\mathbf{X}}_{t-1} \end{aligned}$$

It follows that

$$\begin{aligned} V(\bar{\mathbf{X}}_t, t) - V(\bar{\mathbf{X}}_{t-\tau}, t-\tau) &\leq -\sum_{i=t-\tau}^t \left(\bar{\mathbf{X}}_i^T \hat{\mathbf{A}}_{1i}^T \left(\hat{\Sigma}_{c_i} + \hat{\mathbf{A}}_{2i} \Sigma_\eta \hat{\mathbf{A}}_{2i}^T \right)^{-1} \hat{\mathbf{A}}_{1i} \bar{\mathbf{X}}_i \right. \\ &\quad \left. + \bar{\mathbf{X}}_{i-1}^T \hat{\mathbf{B}}_{i-1}^T \hat{\mathbf{A}}_{1i}^T \mathbf{K}_i^T (\mathbf{P}_i^-)^{-1} \mathbf{K}_i \hat{\mathbf{A}}_{1i} \hat{\mathbf{B}}_{i-1} \bar{\mathbf{X}}_{i-1} \right) \\ &\leq -\nu_1 \|\bar{\mathbf{X}}_{t-\tau-1}\|^2 \leq -\nu_2 \|\bar{\mathbf{X}}_t\|^2 \equiv \gamma_3(\|\bar{\mathbf{X}}_t\|) \end{aligned}$$

A.3.2 Proof of Theorem 3

By Koksma-Hlawka inequality, it can be verified that for large enough simulation size, Lemma 2 and Lemma 3 hold for \mathbf{P}_t^G of 51. Define a Lyapunov function for non-Gaussian model

$$V(\bar{\mathbf{X}}_t^G, t) = (\bar{\mathbf{X}}_t^G)^T (\mathbf{P}_t^G)^{-1} \bar{\mathbf{X}}_t^G. \quad (167)$$

And the Riccati equation for non-Gaussian model is

$$\bar{\mathbf{X}}_t^G = (\mathbf{I} - \mathbf{K}_t^G \hat{\mathbf{A}}_{1t}^G) \hat{\mathbf{B}}_{t-1}^G \bar{\mathbf{X}}_{t-1}^G + \mathbf{K}_t^G \mathbf{Z}_t - \mathbf{K}_t^G \hat{\mathbf{A}}_{1t}^G \hat{\mathbf{b}}_{t-1}^G - \mathbf{K}_t^G \hat{\mathbf{a}}_t^G + \hat{\mathbf{b}}_{t-1}^G \quad (168)$$

The result can be established based on this equation.

A.3.3 Proof of Theorem 4

Based on Theorem 2, $\|\mathbb{F}_t \cdots \mathbb{F}_{i+1} \bar{\mathbf{X}}_i^\mu - \mathbb{F}_t \cdots \mathbb{F}_{i+1} \bar{\mathbf{X}}_i^\nu\| \leq \kappa_1 e^{\vartheta_1(i-t)} \|\mu_0 - \nu_0\|$, then

$$\|\bar{\mathbf{X}}_t^G - \bar{\mathbf{X}}_t\| \leq \sum_{i=1}^t \kappa_1 e^{\vartheta_1(i-t)} \|\mathbb{F}_i^G \bar{\mathbf{X}}_{i-1}^G - \mathbb{F}_i \bar{\mathbf{X}}_{i-1}^G\| \quad (169)$$

Based on Koksma-Hlawka inequality (Theorem 1), $\| \mathbb{F}_i^G \bar{\mathbf{X}}_{i-1}^G - \mathbb{F}_i \bar{\mathbf{X}}_{i-1}^G \| \leq W(\mathbb{F}_i) D_G$, where $W(\mathbb{F}_i)$ is defined in (45), and D_G is the discrepancy defined in (43). Since the discrepancy D_G is only dependent on the type of low-discrepancy sequence used in QMCKF and the simulation size G , we can further simplify the above formula as:

$$\| \bar{\mathbf{X}}_t^G - \bar{\mathbf{X}}_t \| \leq \left(\sum_{i=1}^t \kappa_1 e^{\vartheta_1(i-t)} W(\mathbb{F}_i) \right) D_G = \Upsilon_t D_G \quad (170)$$

where $\Upsilon_t \equiv \sum_{i=1}^t \kappa_1 e^{\vartheta_1(i-t)} W(\mathbb{F}_i)$. It's easy to see that Υ_t is bounded since $W(\mathbb{F}_i)$ is bounded for every \mathbb{F}_i .

A.3.4 Proof of Theorem 5

Step 1: We denote $\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k)$ as the true filtering distribution for non-Gaussian model (58)-(59) such that μ is a well defined Gaussian initial measure on \mathbf{X}_0 . Suppose we initiate the filtering from Gaussian $\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_{k-1})$ and since $p_\nu \in \mathcal{C}_{\mathcal{G}, \mathbb{P}}$, we have

$$\begin{aligned} \bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k) &= c_k \bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_{k-1}) p_\nu(\mathbf{Z}_k | \mathbf{X}_k) \\ &= c_k \bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_{k-1}) \int_{\Theta} \phi_\theta(\mathbf{Z}_k | \mathbf{X}_k) \mathbb{P}(d\theta) \\ &= \int_{\Theta} c_k \bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_{k-1}) \phi_\theta(\mathbf{Z}_k | \mathbf{X}_k) \mathbb{P}(d\theta) \end{aligned}$$

It's clear that $\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_{k-1}) \phi_\theta(\mathbf{Z}_k | \mathbf{X}_k)$ has a Gaussian kernel because $\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_{k-1})$ is a Gaussian. This implies $\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k) \in \mathcal{C}_{\mathcal{G}, \mathbb{P}}$.

Step 2: From (61), we have the approximated filtering distribution based on Gaussian mixture approximation as

$$p^\mu(\mathbf{X}_k | \mathcal{Z}_k) \approx \sum_{i=1}^m c_{ki} p^\mu(\mathbf{X}_k | \mathcal{Z}_{k-1}) \phi(\mathbf{Z}_k - \mathbf{H}(\mathbf{X}_k) - \mathbf{a}_i, \mathbf{P}_i) \quad (171)$$

This means $p^\mu(\mathbf{X}_k | \mathcal{Z}_k)$ is a Gaussian mixture approximation of $\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k)$. According to Genovese and Wasserman (2000)⁹, we have the bound for the error of this approximation:

$$\inf_{p^\mu} \mathcal{H}(\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k) | p^\mu(\mathbf{X}_k | \mathcal{Z}_k)) = O(\log m / m) \quad (172)$$

Thus we have $\mathcal{H}(\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k) | p^\mu(\mathbf{X}_k | \mathcal{Z}_k)) \rightarrow 0$ as $m \rightarrow \infty$.

Step 3: For two well defined Gaussian initial measures μ and ν ,

$$\mathcal{H}(p^\mu(\mathbf{X}_k | \mathcal{Z}_k) | p^\nu(\mathbf{X}_k | \mathcal{Z}_k)) \leq \mathcal{H}(\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k) | p^\mu(\mathbf{X}_k | \mathcal{Z}_k))$$

⁹Ghosal and van der Vaart (2007) obtain an improved bound: $\inf_{p^\mu} \mathcal{H}(\bar{p}^\mu(\mathbf{X}_k | \mathcal{Z}_k) | p^\mu(\mathbf{X}_k | \mathcal{Z}_k)) = O(\log m / \exp(m))$

$$+ \mathcal{H}(\bar{p}^\nu(\mathbf{X}_k|\mathfrak{Z}_k) | p^\nu(\mathbf{X}_k|\mathfrak{Z}_k)) + \mathcal{H}(\bar{p}^\mu(\mathbf{X}_k|\mathfrak{Z}_k) | \bar{p}^\nu(\mathbf{X}_k|\mathfrak{Z}_k)) \quad (173)$$

We assume the true filtering distribution is stable, then $\mathcal{H}(\bar{p}^\mu(\mathbf{X}_k|\mathfrak{Z}_k) | \bar{p}^\nu(\mathbf{X}_k|\mathfrak{Z}_k)) \rightarrow 0$ as $k \rightarrow \infty$. Then the stability of approximated filtered distribution of \mathfrak{F}^{ng} is established:

$$\mathcal{H}(p^\mu(\mathbf{X}_k|\mathfrak{Z}_k) | p^\nu(\mathbf{X}_k|\mathfrak{Z}_k)) \rightarrow 0 \quad \text{as } k, m \rightarrow \infty. \quad (174)$$

A.3.5 Proof of Theorem 6

Similarly to the proof of Theorem 4, we define two functionals \mathbb{J} and \mathbb{J}^G such that $p^\mu(\mathbf{X}_k|\mathfrak{Z}_k) = \mathbb{J}(p^\mu(\mathbf{X}_{k-1}|\mathfrak{Z}_{k-1}), \mathbf{Z}_k) = \mathbb{J}_k p_{k-1}^\mu$, and $p_G^\mu(\mathbf{X}_k|\mathfrak{Z}_k) = \mathbb{J}^G(p_G^\mu(\mathbf{X}_{k-1}|\mathfrak{Z}_{k-1}), \mathbf{Z}_k) = \mathbb{J}_k^G p_{k-1,G}^\mu$, where we let $p_{k-1}^\mu \equiv p^\mu(\mathbf{X}_k|\mathfrak{Z}_k)$, $p_{k-1,G}^\mu \equiv p_G^\mu(\mathbf{X}_{k-1}|\mathfrak{Z}_{k-1})$ to simplify the notation. Then it follows:

$$\mathcal{H}\left(p_k^\mu | p_{k,G}^\mu\right) \leq \sum_{i=1}^k \mathcal{H}\left(\mathbb{J}_k \cdots \mathbb{J}_{i+1} \mathbb{J}_i^G p_{i-1,G}^\mu | \mathbb{J}_k \cdots \mathbb{J}_{i+1} \mathbb{J}_i p_{i-1,G}^\mu\right). \quad (175)$$

Based on the construction of QMCKF and Koksma-Hlawka inequality, we conclude $\mathcal{H}\left(p_k^\mu | p_{k,G}^\mu\right) \rightarrow 0$ as $G \rightarrow \infty$. The result of Theorem 6 follows from Theorem 5.

A.3.6 Proof of Theorem 7

From the construction of quasi Monte-Carlo integration and the Koksma-Hlawka inequality, it is easy to verify that $\sup_{\theta} \frac{1}{N} \left| L_{g,N}^G(\theta) - L_{g,N}(\theta) \right| \rightarrow 0$ as $G \rightarrow \infty$, $N \rightarrow \infty$, and $\sup_{\theta} \frac{1}{N} \left| L_{ng,N}^G(\theta) - L_{ng,N}(\theta) \right| \rightarrow 0$ as $G \rightarrow \infty$, $N \rightarrow \infty$, and $m \rightarrow \infty$.

From (iv) of Assumption 3, we can conclude that $\frac{1}{N} L_{g,N}^G(\theta)$ converges uniformly to $E[\log f_g(\theta; \mathfrak{Z}_N)]$ and $\frac{1}{N} L_{ng,N}^G(\theta)$ converges uniformly to $E[\log f_{ng}(\theta; \mathfrak{Z}_N)]$. Then consistency of the MSLEs follows by Theorem 2.1 of Newey and McFadden (1994).

B Algorithm of Bootstrap Particle filter

The Particle filter used in this paper for comparison with QMCKF is the Bootstrap Particle filter based on Gordon et al. (1993) with multinomial resampling approach. Suppose we have the following state space model:

$$\begin{aligned} X_{t+1} &= F(X_t, \varepsilon_t; \theta_F) \\ Z_t &= H(X_t, \eta_t; \theta_H) \end{aligned}$$

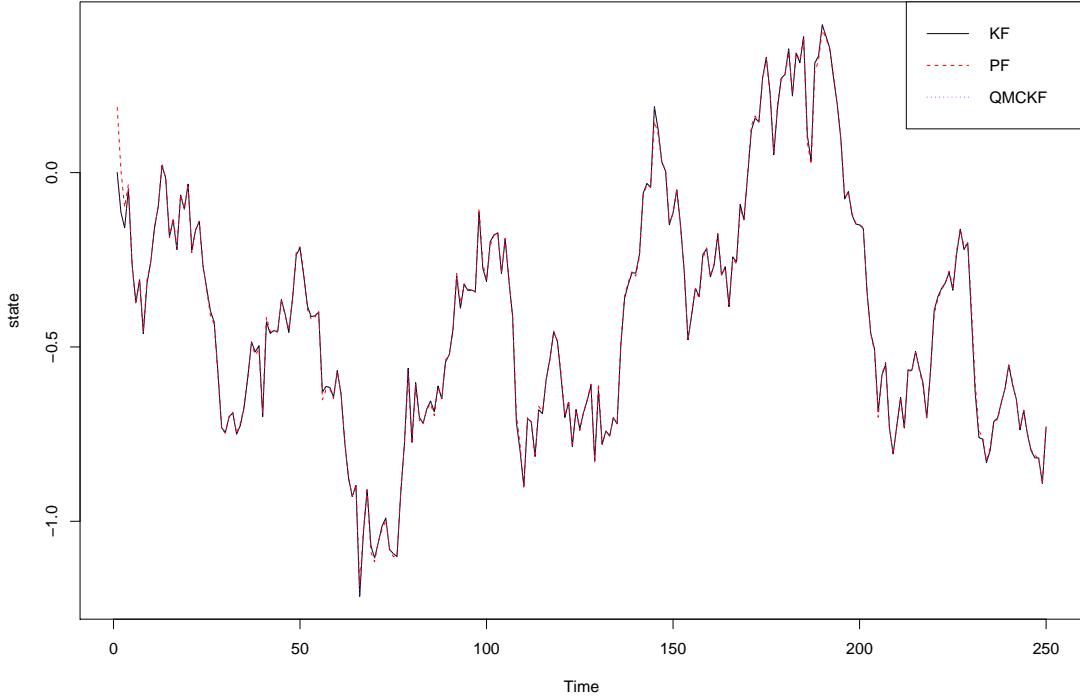
Let $\theta = \{\theta_F, \theta_H\}$ be the set of parameters in the above model. Also, we assume $\varepsilon \sim p_\varepsilon(\cdot)$, $\eta \sim p_\eta(\cdot)$, and $X_0 \sim \mu(\cdot)$. The algorithm of Bootstrap Particle filter is summarized in Algorithm 4.

Algorithm 4: Bootstrap Particle filter

- Step 1: At $t = 0$, for $i = 1, \dots, N$, sample $X_0^{(i)} \sim \mu$ and set $w_0^{(i)} = \frac{1}{N}$.
 - Step 2: At time t , for $i = 1, \dots, N$, propagate $X_t^{(i)}$ by sampling $\tilde{X}_t^{(i)}$ from $p(X_t^{(i)} | X_{t-1}^{(i)}, \varepsilon_{t-1})$
 - Step 3: At time t , for $i = 1, \dots, N$, evaluate $w_t^{(i)} = p(Z_t | \tilde{X}_t^{(i)}, \eta_t)$ and normalize this importance weights.
 - Step 4: At time t , for $i = 1, \dots, N$, resample $X_t^{(i)}$ from $(\tilde{X}_t^{(i)}, w_t^{(i)})$
 - Step 5: Repeat Step 2-4 for $t = 1, \dots, T$.
-

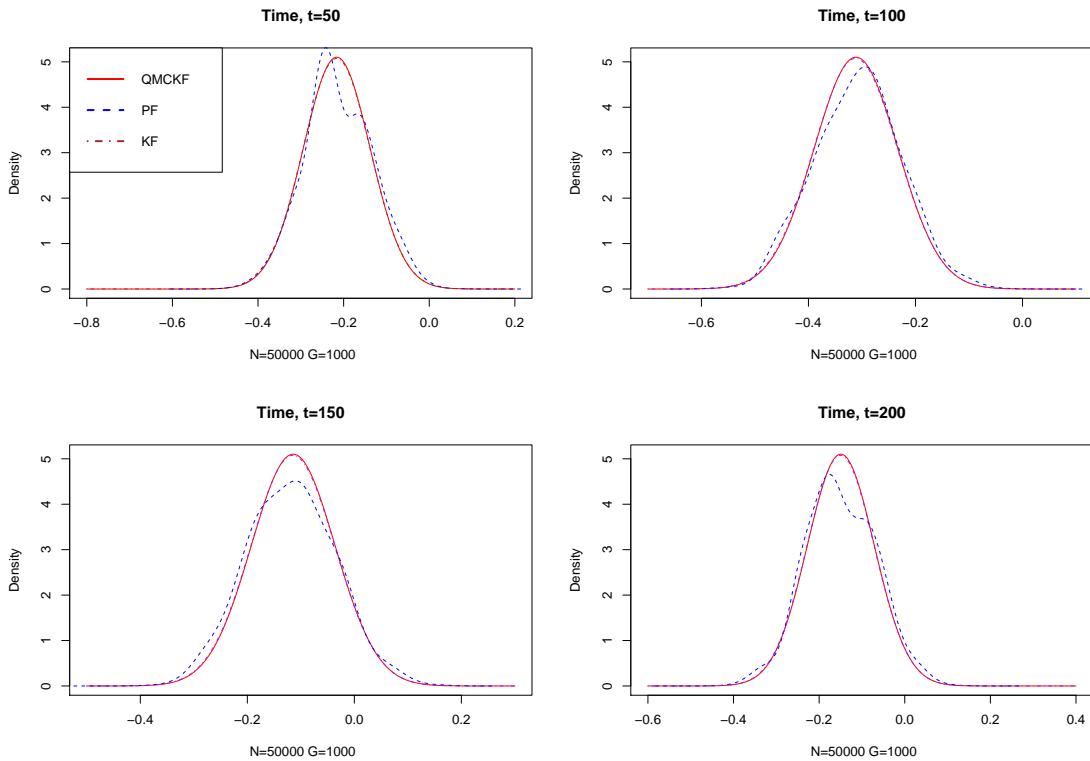
C Figures

Figure 1: Estimation of the mean of state variable for linear and Gaussian model.



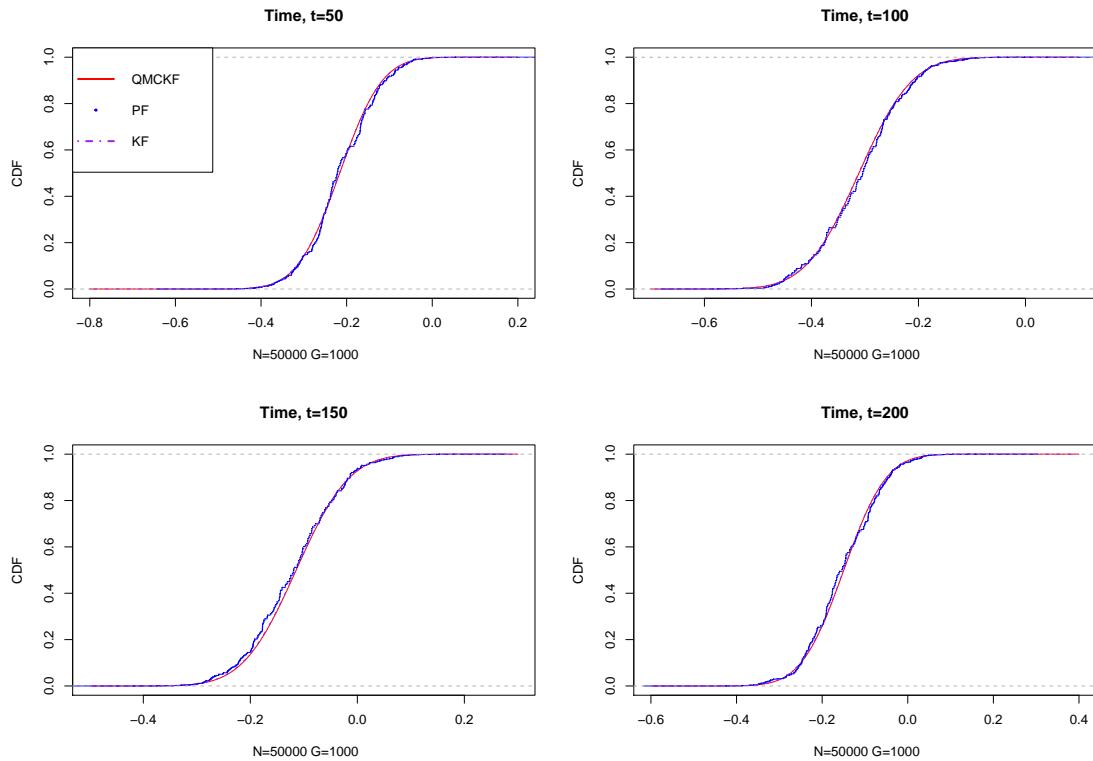
Note: The solid line in black is the estimation of KF, the dashed line in red is the estimation of PF and the dotted line in purple is the estimation of QMCKF. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$.

Figure 2: Plot of the density of filtered distribution at $t = 50, 100, 150, 200$.



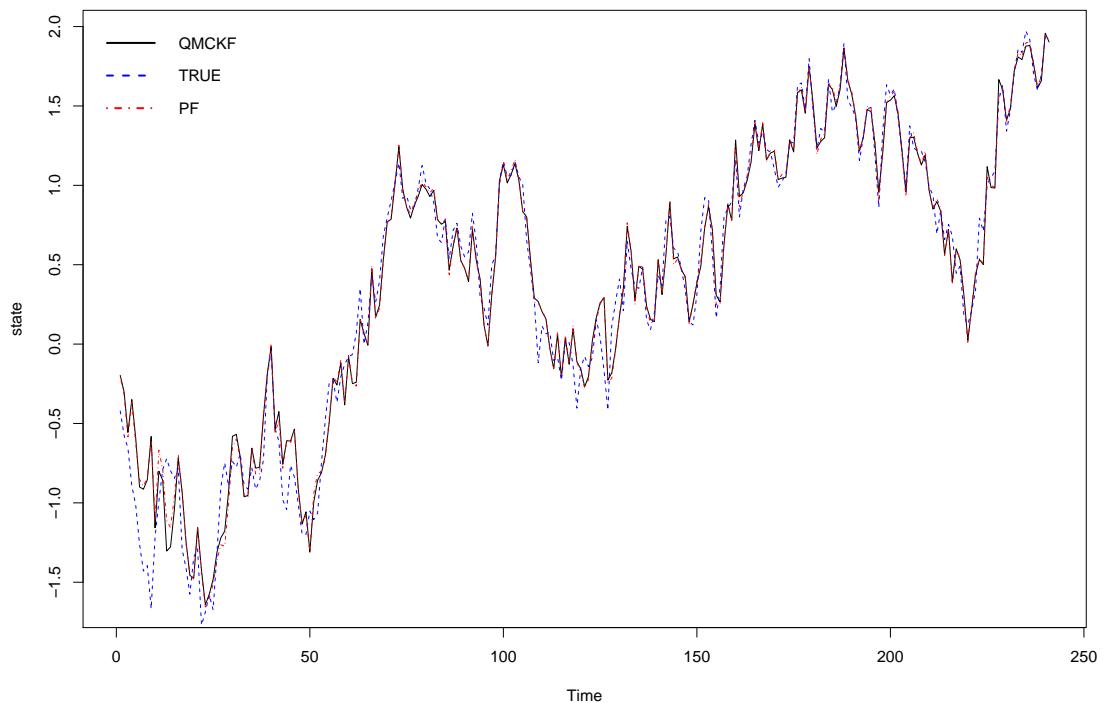
Note: The solid line in red is the estimation of the density based on QMCKF, the dashed line in blue is the kernel estimation based on PF and the dash-dotted line in purple is the estimation based on KF. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$.

Figure 3: Plot of the CDF of filtered distribution at $t = 50, 100, 150, 200$.



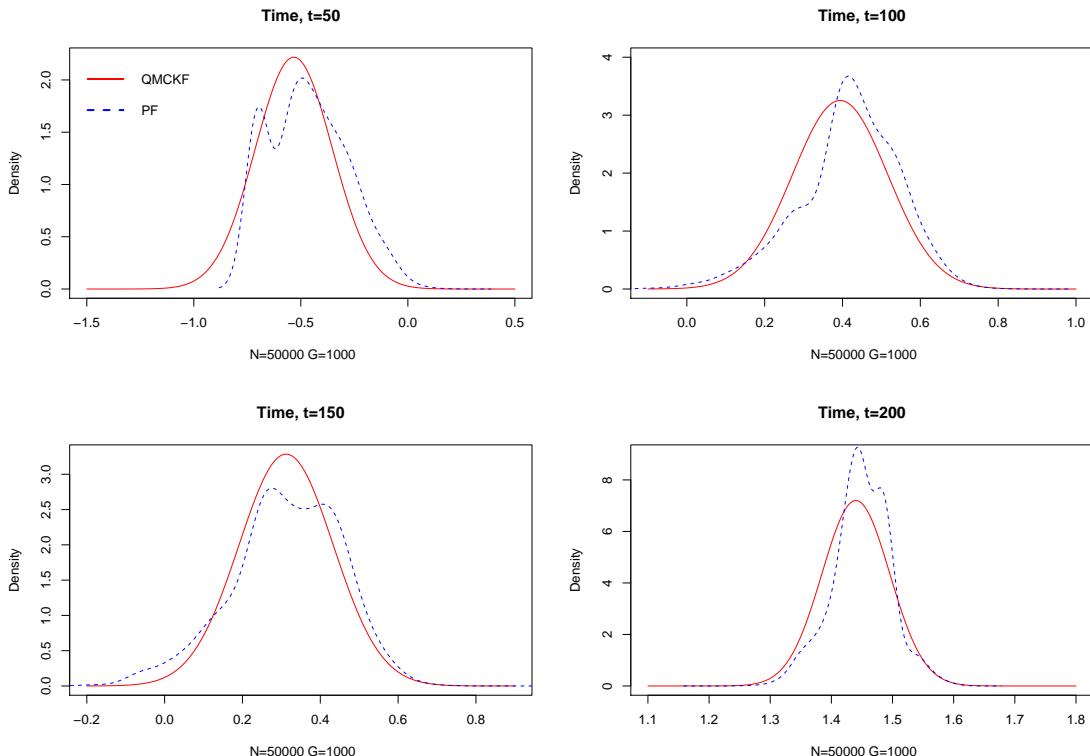
Note: The solid line in red is the estimation of the CDF based on QMCKF, the dots in blue is the empirical CDF based on PF and the dash-dotted line in purple is the estimation based on KF. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$.

Figure 4: Estimation of the mean of state variable for nonlinear and Gaussian model .



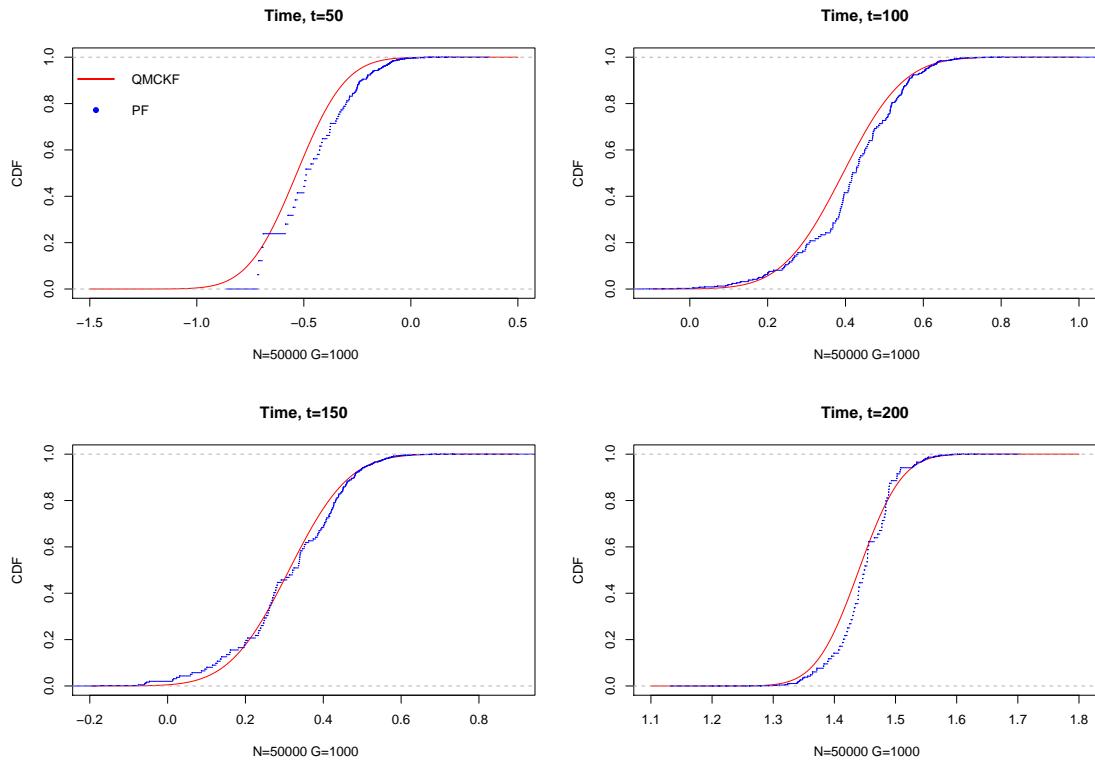
Note: The solid line in black is the estimation of QMCF, the dashed line in blue is the true value of the state variable and the dash-dotted line in red is the estimation of PF. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$.

Figure 5: Plot of the density of filtered distribution at $t = 50, 100, 150, 200$.



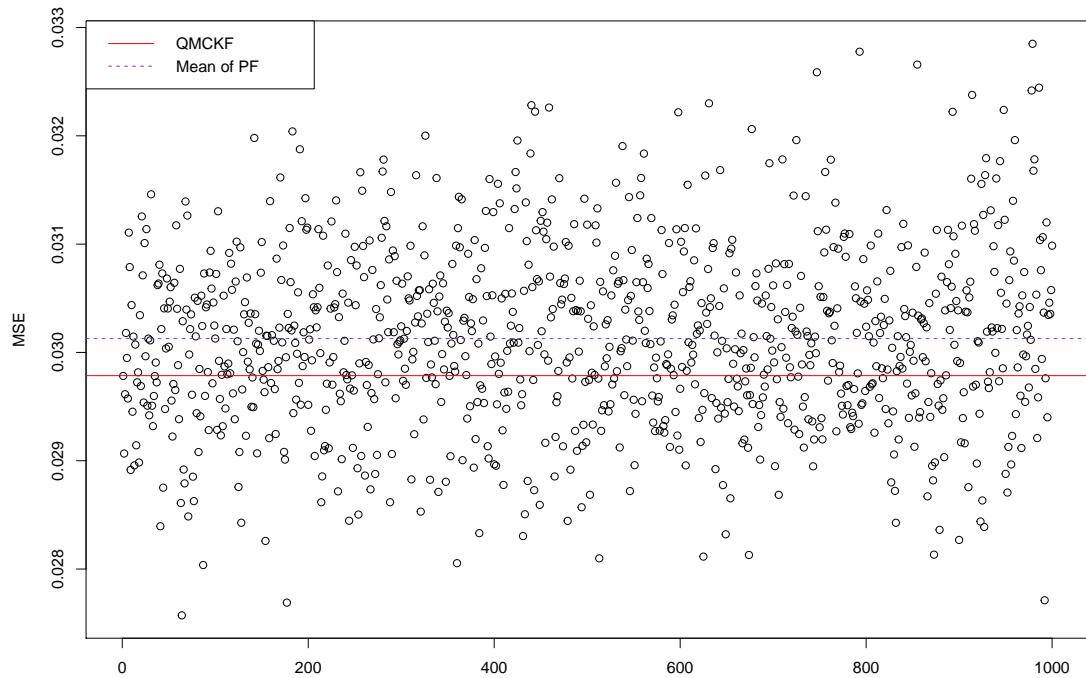
Note: The solid line in red is the estimation of the density based on QMCKF, and the dashed line in blue is the kernel estimation based on PF. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$.

Figure 6: Plot of the CDF of filtered distribution at $t = 50, 100, 150, 200$.



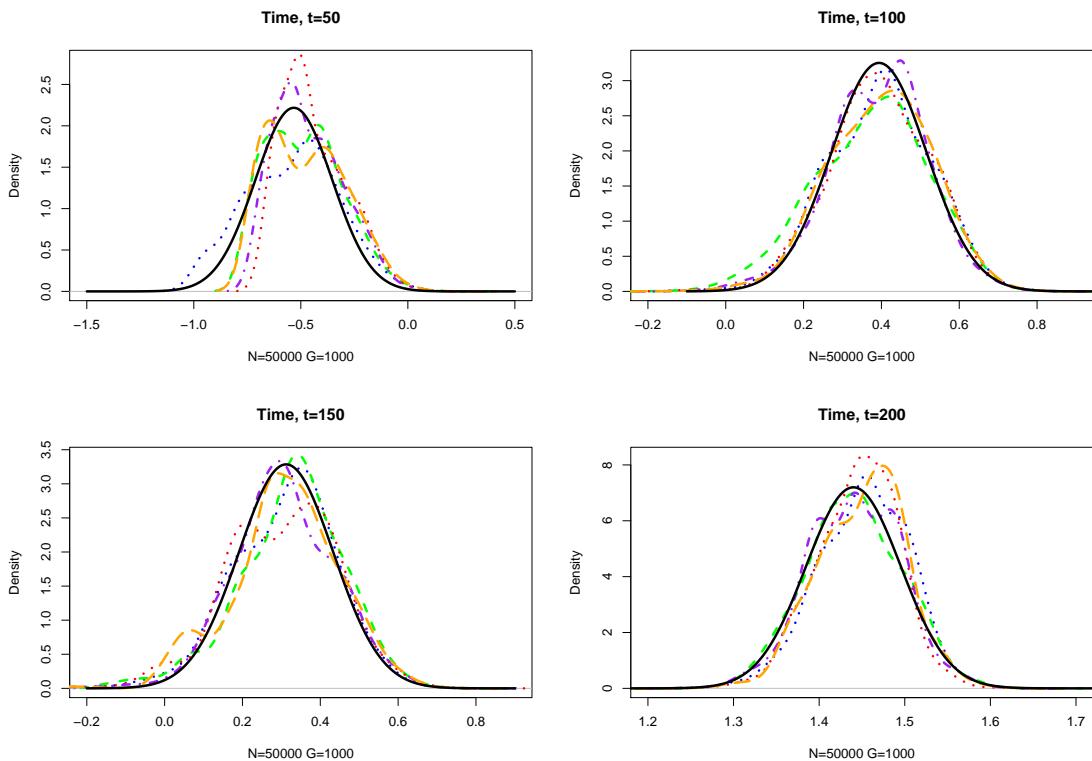
Note: The solid line in red is the estimation of the CDF based on QMCKF, and the dots in blue is the empirical CDF based on PF. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$.

Figure 7: MSE in estimating the mean of the state variable.



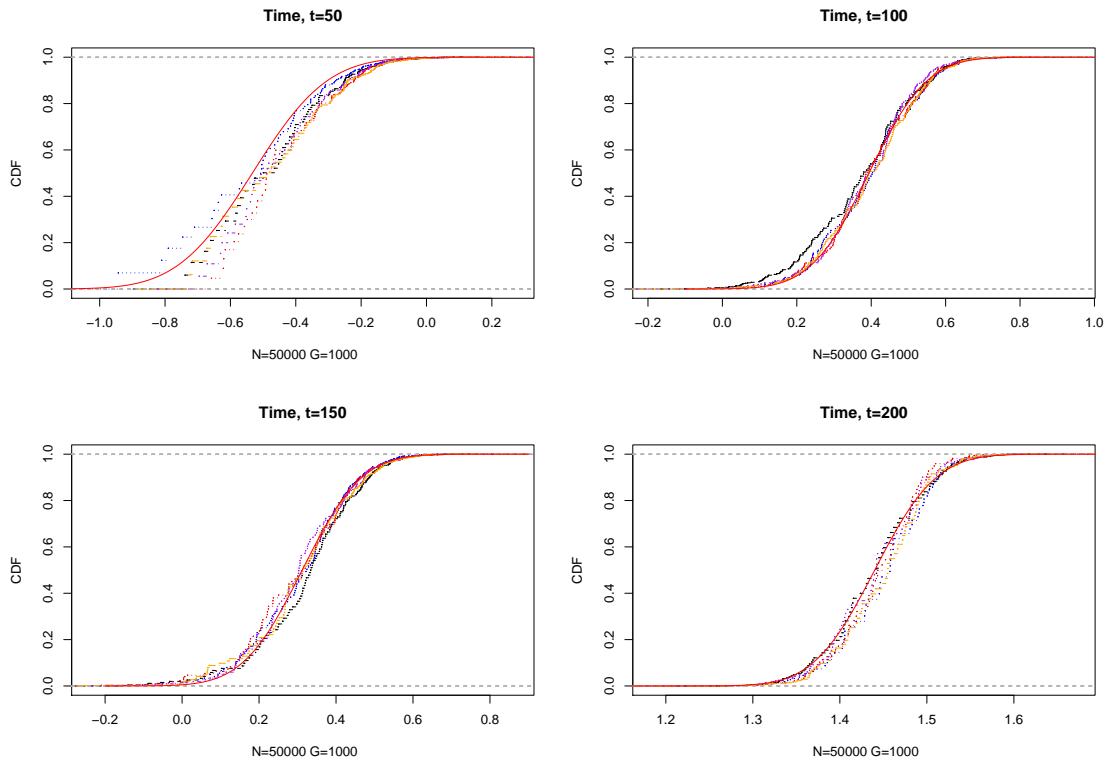
Note: The solid line in red is based on QMCKF and the 1000 circles are the results of 1000 different trials of PF.

Figure 8: Plot of the density of filtered distribution at $t = 50, 100, 150, 200$.



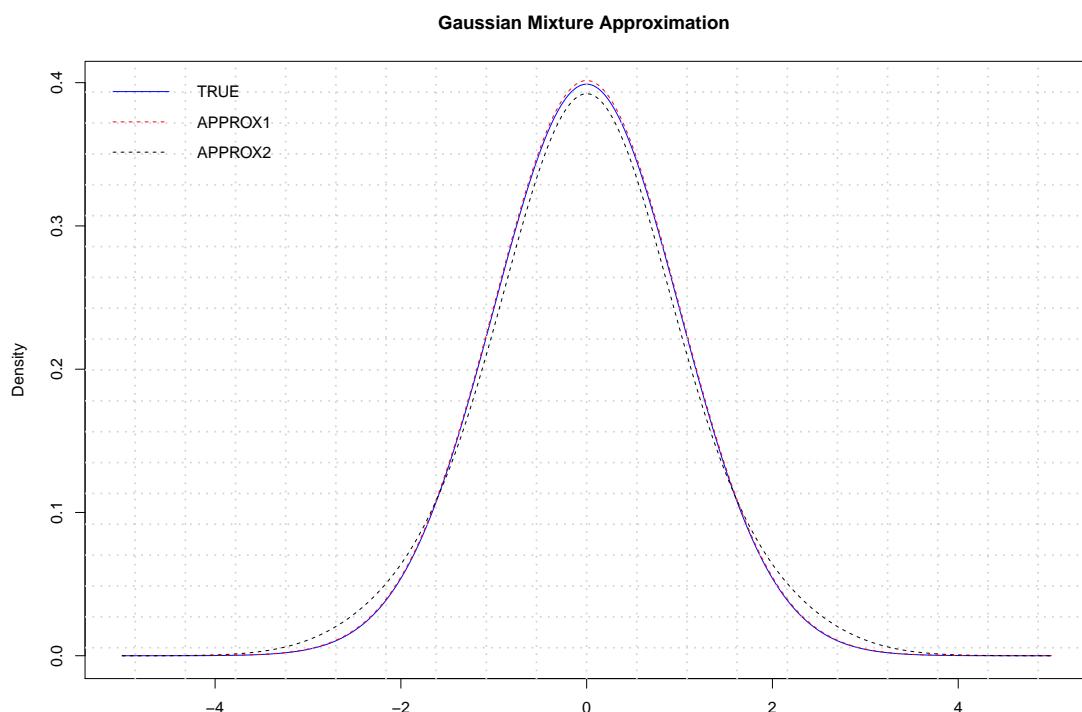
Note: The solid line in black is the estimation of the density based on QMCKF, and the dashed lines in other colors are the kernel estimation based on 10 different trials (only 5 of them are plotted) of PF on the same data set. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$ for every trial.

Figure 9: Plot of the CDF of filtered distribution at $t = 50, 100, 150, 200$.



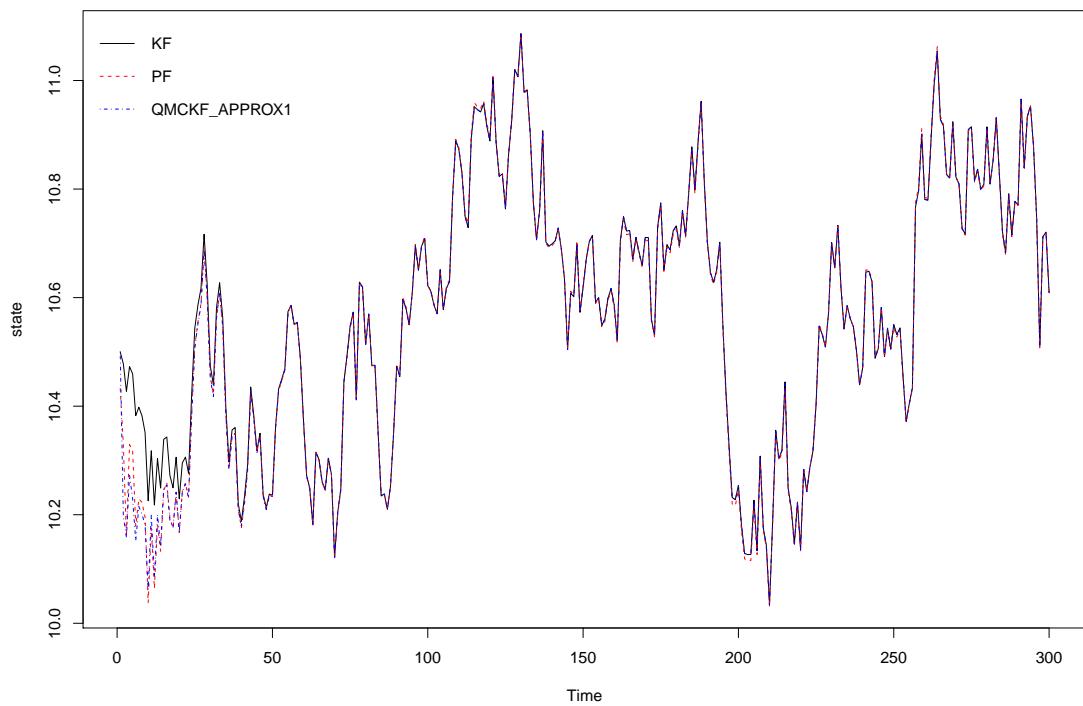
Note: The solid line in red is the estimation of the CDF based on QMCKF, and the dots in other colors are the empirical CDF based on 1000 different trails (only 5 of them are plotted) of PF on the same data set. The simulation size in QMCKF is $G = 1000$, and the simulation size in PF is $N = 50000$ for every trail.

Figure 10: Gaussian mixture approximation of the standard normal distribution.



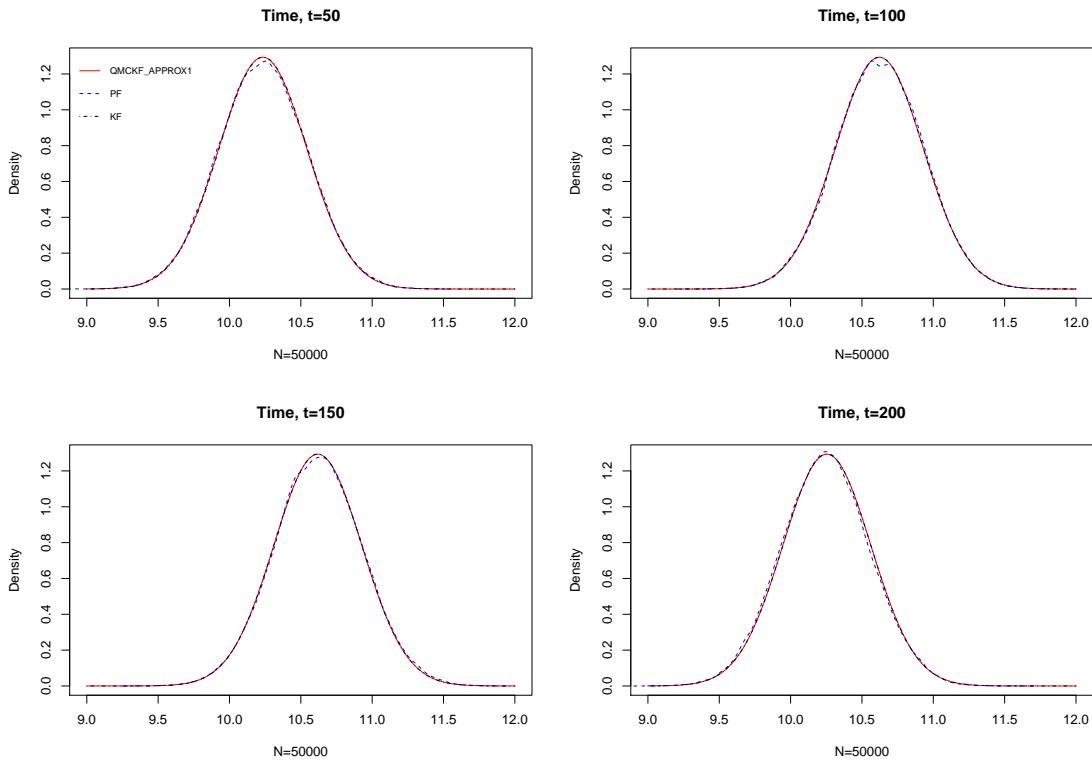
Note: The solid line in blue is the true density, the dashed line in red is the approximation based on the values of parameters selected according to the Table (1), and the dashed line in black is the approximation based on the values of parameters selected according to the Table 2.

Figure 11: Estimation of the mean of state variable for linear and Gaussian model.



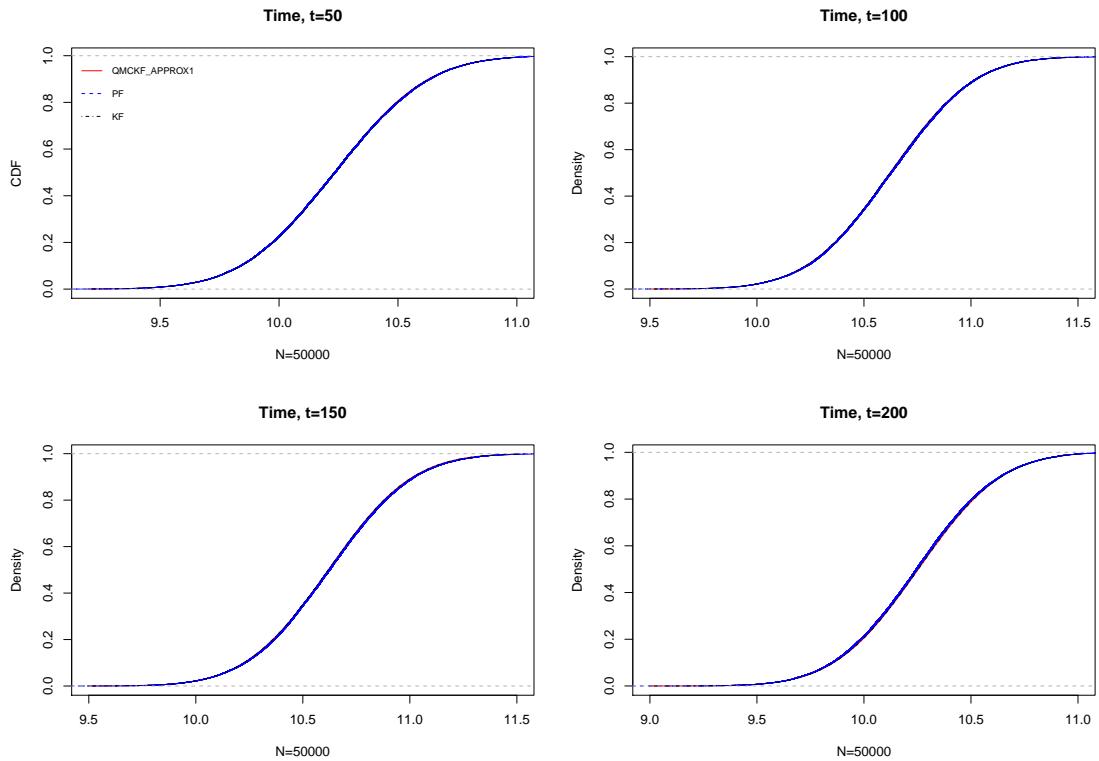
Note: We are using our approach for non-Gaussian model to approximate the filtering distribution for this Gaussian model. The solid line in black is the estimation of KF the dashed line in red is the estimation of PF, and the dash-dotted line in purple is the estimation of QMCKF. The simulation size in PF is $N = 50000$. The approximation of the density of the noise is according to the Table 1

Figure 12: Plot of the density of filtered distribution at $t = 50, 100, 150, 200$.



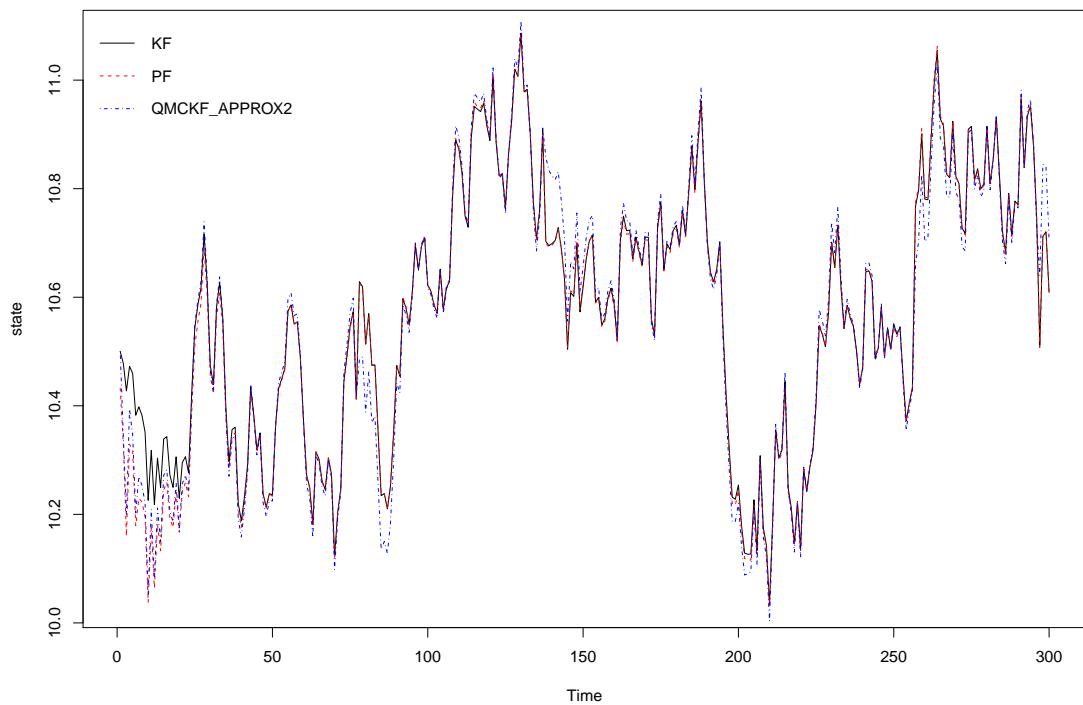
Note: The solid line in red is the estimation of the density based on QMCKF, and the dashed line in blue is the kernel estimation based on PF, and the dash-dotted line in purple is the estimation of KF. The simulation size in PF is $N = 50000$. The approximation of the density of the noise is according to the Table 1

Figure 13: Plot of the CDF of filtered distribution at $t = 50, 100, 150, 200$.



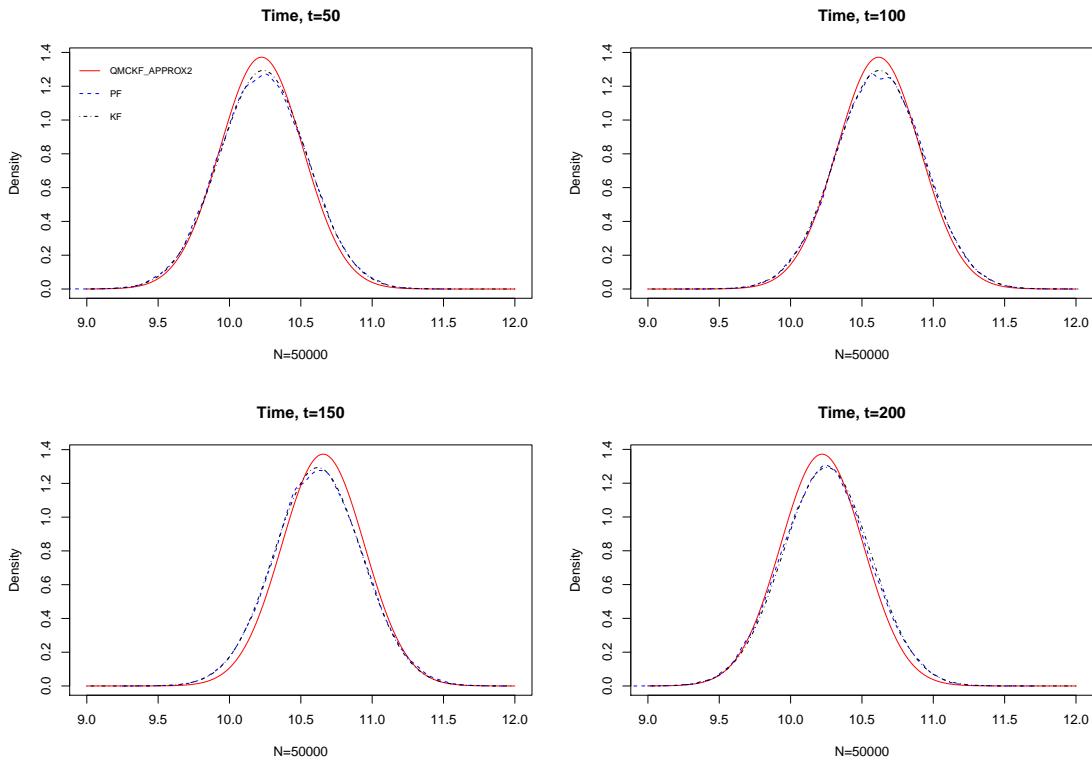
Note: The solid line in red is the estimation of the CDF based on QMCKF, and the dots in blue is the empirical CDF based on PF, and the dash-dotted line in purple is the estimation based on KF. The simulation size in PF is $N = 50000$. The approximation of the density of the noise is according to the Table 1

Figure 14: Estimation of the mean of state variable for linear and Gaussian model.



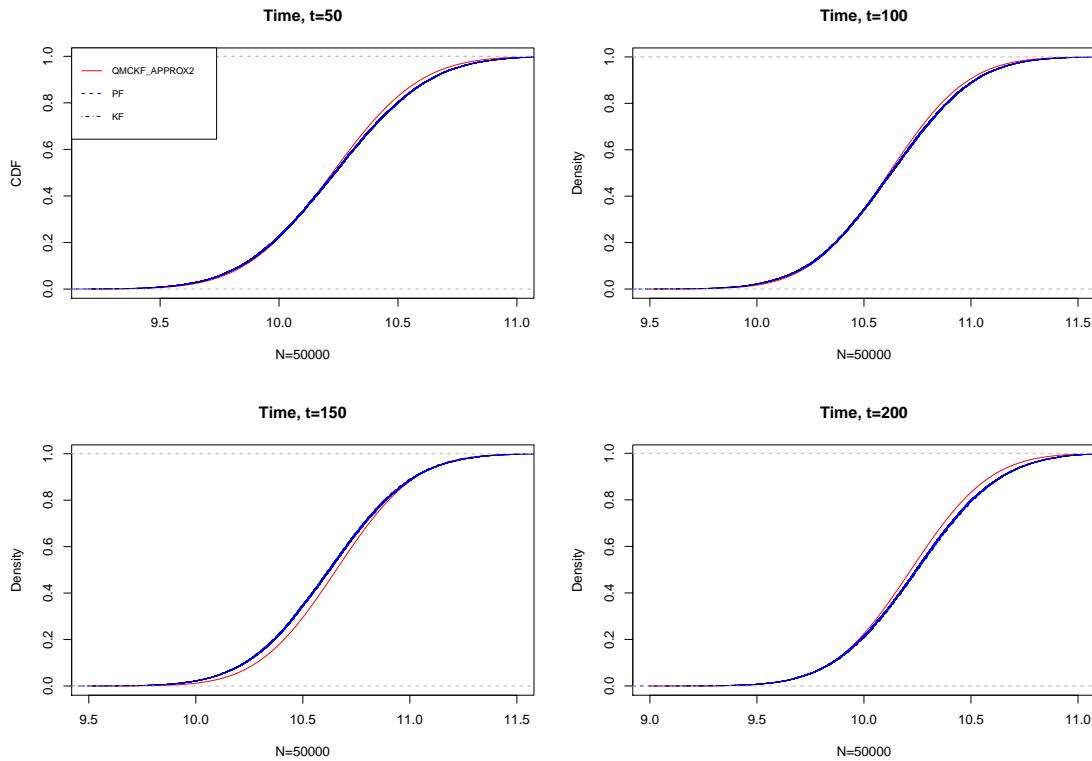
Note: The solid line in black is the estimation of KF the dashed line in red is the estimation of PF, and the dash-dotted line in purple is the estimation of QMCKF. The simulation size in PF is $N = 50000$. The approximation of the density of the noise is according to the Table 2

Figure 15: Plot of the density of filtered distribution at $t = 50, 100, 150, 200$.



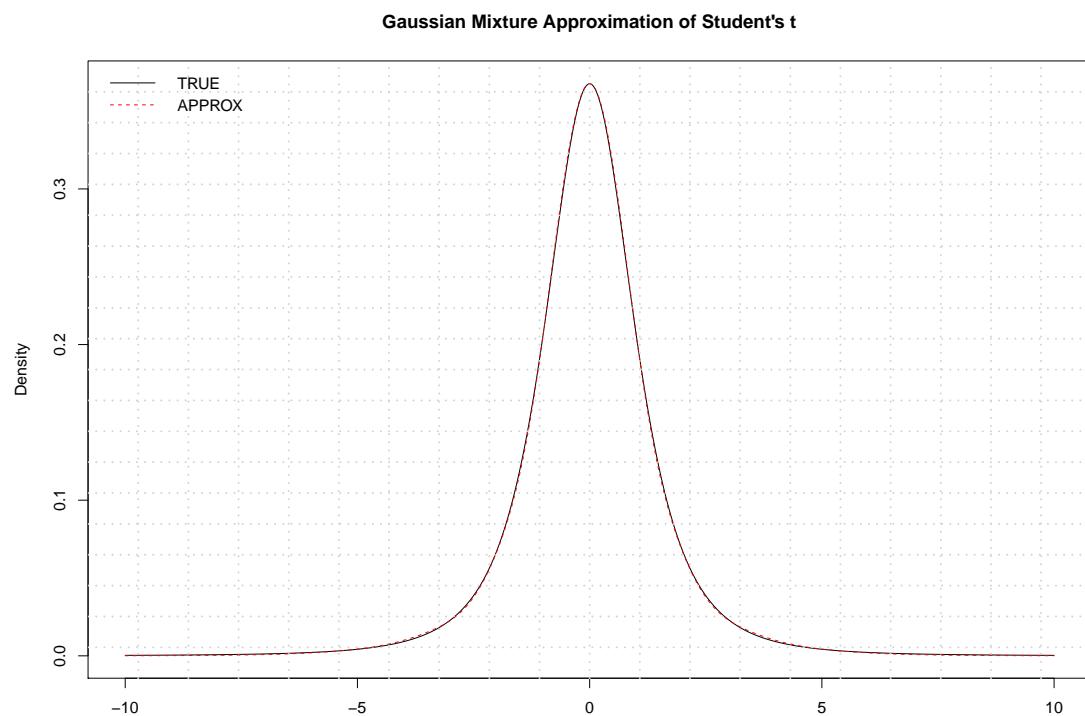
Note: The solid line in red is the estimation of the density based on QMCKF, the dashed line in blue is the kernel estimation based on PF, and the dash-dotted line in purple is the estimation of KF. The simulation size in PF is $N = 50000$. The approximation of the density of the noise is according to the Table 2

Figure 16: Plot of the CDF of filtered distribution at $t = 50, 100, 150, 200$.



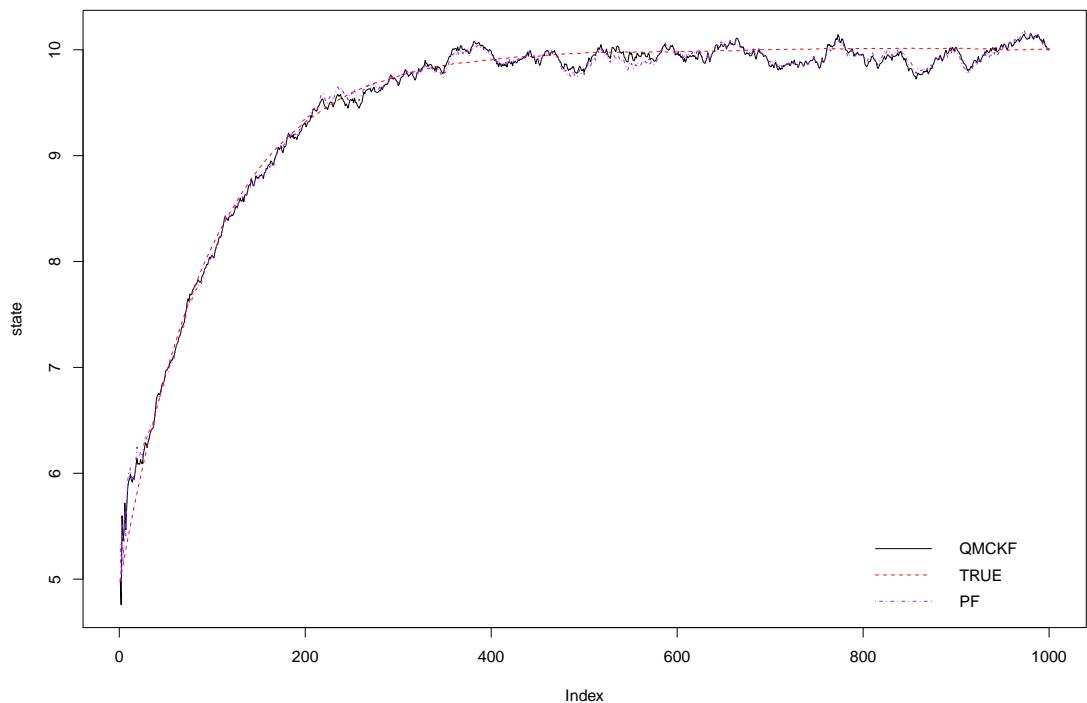
Note: The solid line in red is the estimation of the CDF based on QMCKF, the dots in blue is the empirical CDF based on PF, and the dash-dotted line in purple is the estimation based on KF. The simulation size in PF is $N = 50000$. The approximation of the density of the noise is according to the Table 2

Figure 17: Gaussian mixture approximation of Student's t distribution with degree of freedom 3.



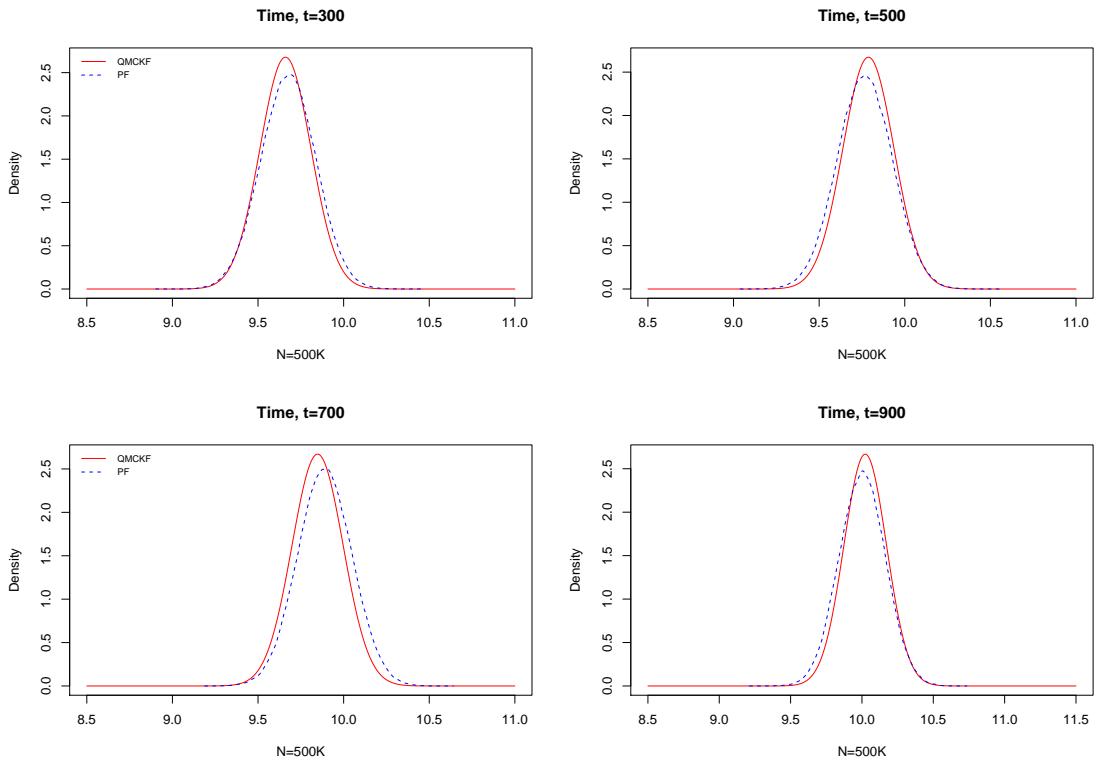
Note: The solid line in black is the true density and the dashed line in red is the approximated density according to Table 3.

Figure 18: Estimation of the mean of state variable for linear and non-Gaussian model.



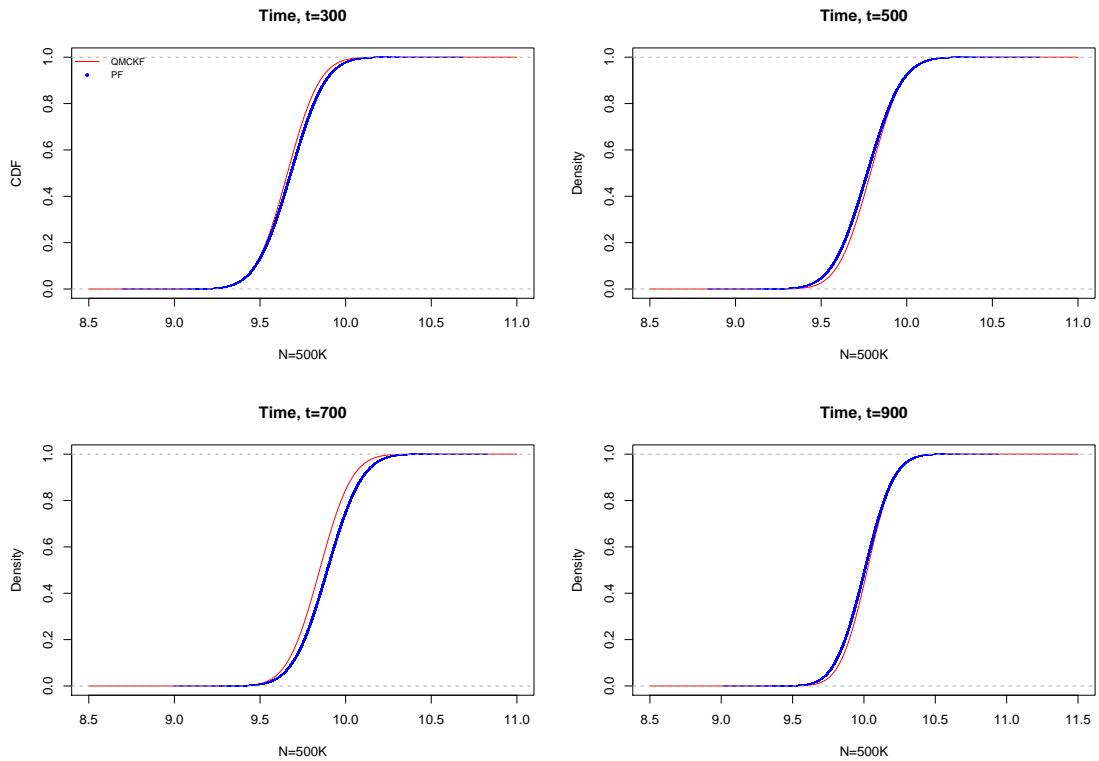
Note: The solid line in black is the estimation of QMCKF, the dashed line in red is the true value of state variable, and the dash-dotted line in purple is the estimation of PF. The simulation size in PF is $N = 500000$. The approximation of the density of the noise in QMCKF is according to the Table 3.

Figure 19: Plot of the density of filtered distribution at $t = 300, 500, 700, 900$.



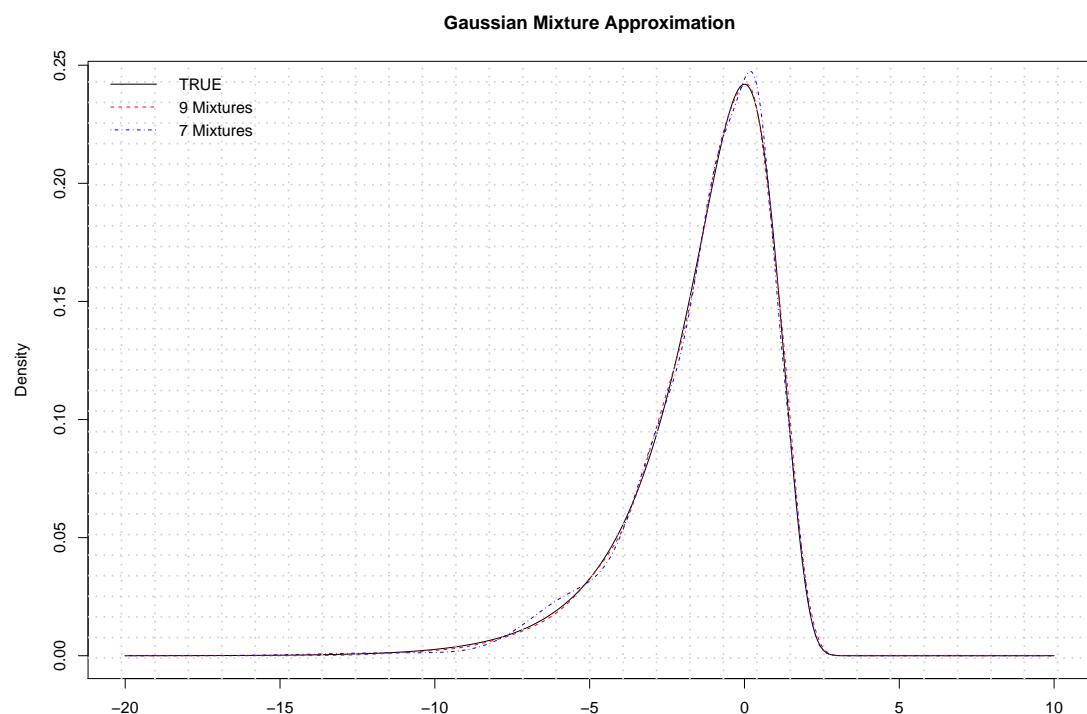
Note: The solid line in red is the estimation of the density based on QMCKF, and the dashed line in blue is the kernel estimation based on PF . The simulation size in PF is $N = 500000$. The approximation of the density of the noise is according to the Table 3.

Figure 20: Plot of the CDF of filtered distribution at $t = 300, 500, 700, 900$.



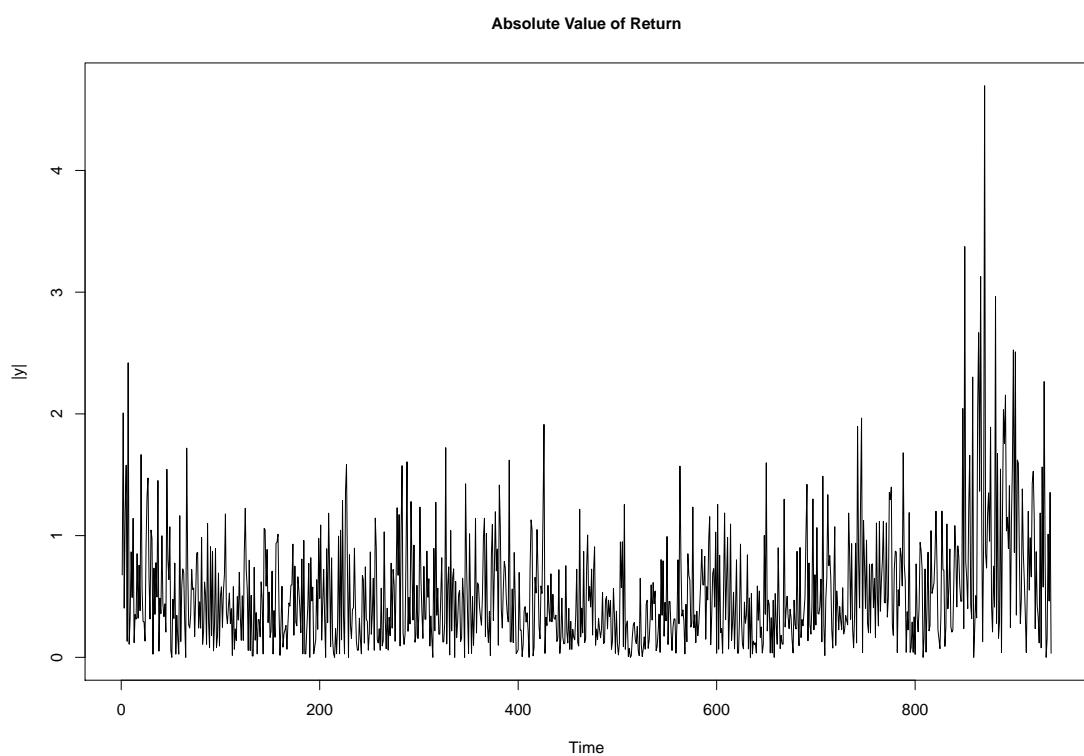
Note: The solid line in red is the estimation of the CDF based on QMCKF, and the dots in blue is the empirical CDF based on PF. The simulation size in PF is $N = 500000$. The approximation of the density of the noise is according to the Table 3.

Figure 21: Gaussian mixture approximation of $\log \chi_1^2$ distribution.



Note: The solid line in black is the true density, the dashed line in red is the approximated density according to the Table (5) with 9 mixtures and the dash-dotted line in blue is the approximated density according to the Table 4.

Figure 22: Plot of the absolute value of return of Sterling/Dollar.



Note: The dataset is the daily observations of weekday close exchange rates for Sterling/Dollar from 10/01/1981 to 06/28/1985, with 937 observations.

Figure 23: Filtered estimate of the volatility $\exp\left(\frac{h_t}{2}\right)$ based on QMCKF.

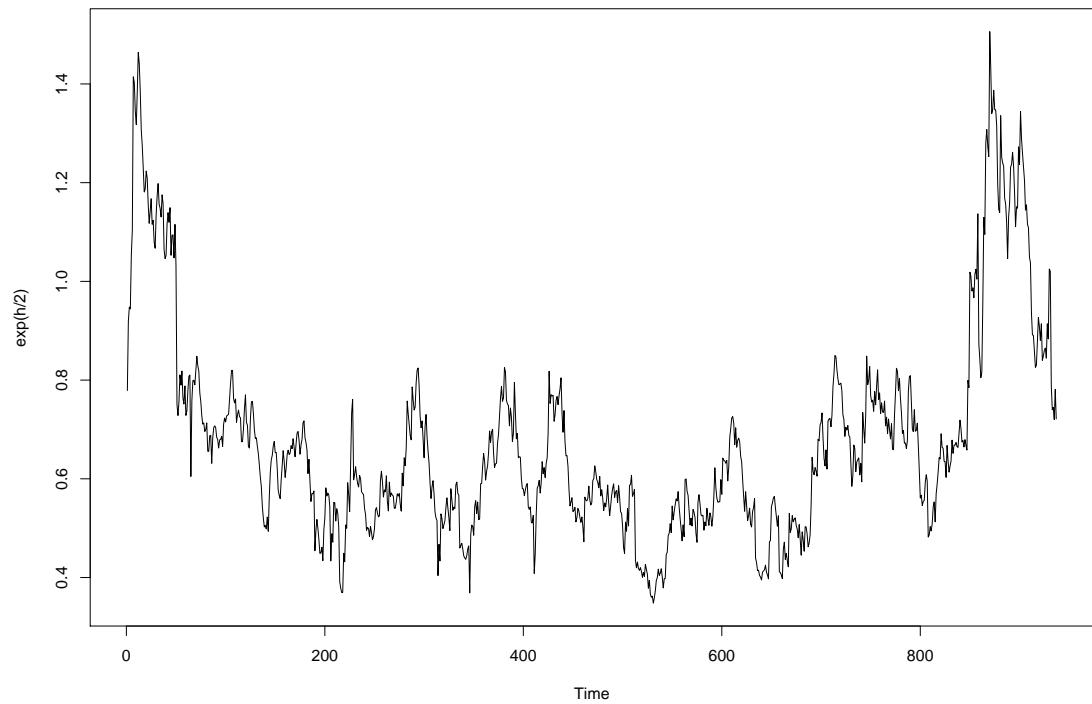


Figure 24: Estimate of the long-run mean of the volatility $\exp\left(\frac{\mu}{2}\right)$ based on QMCKF.

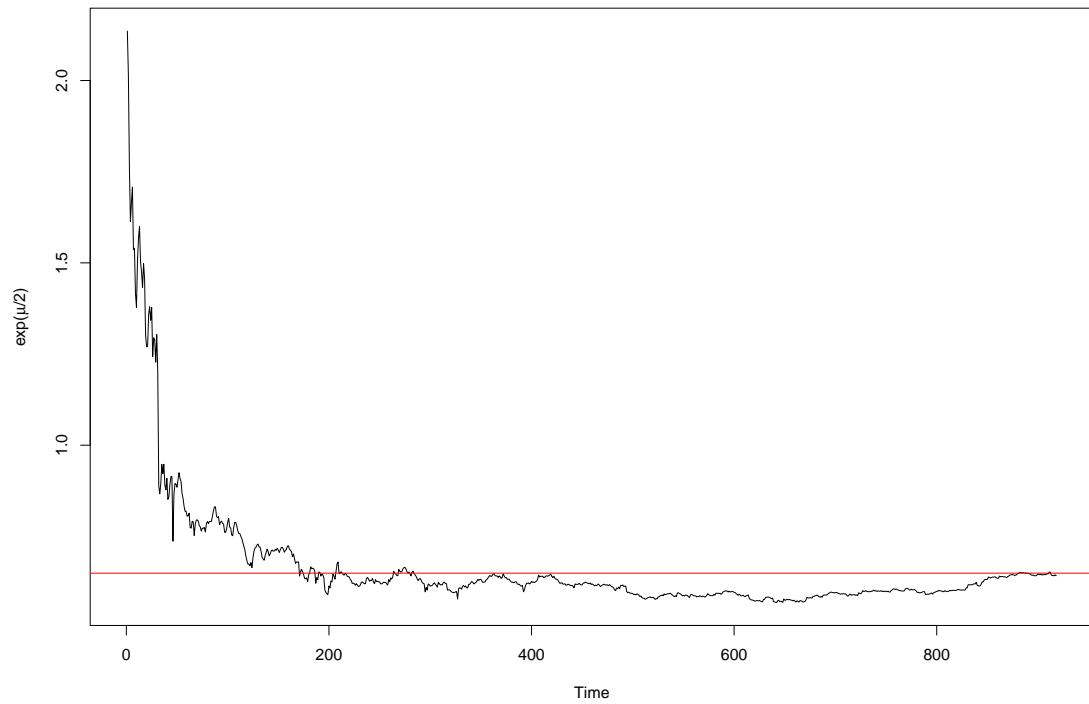
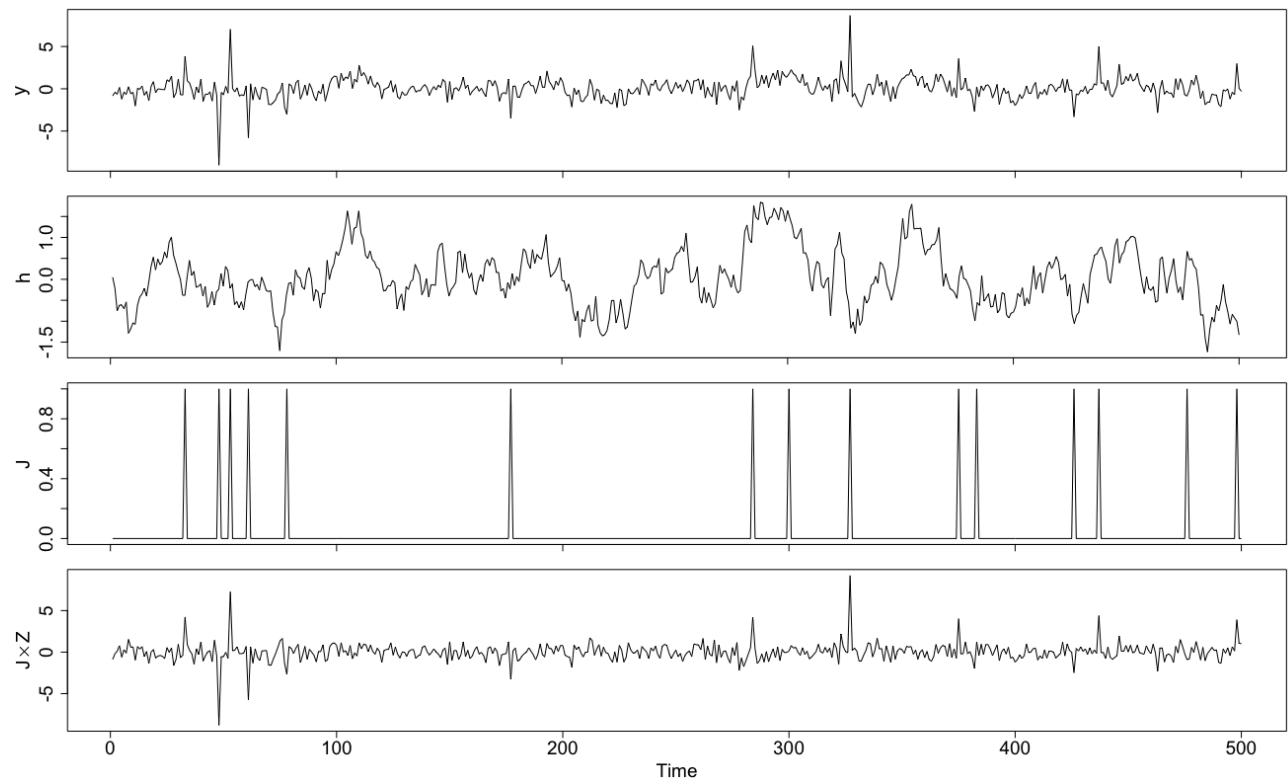
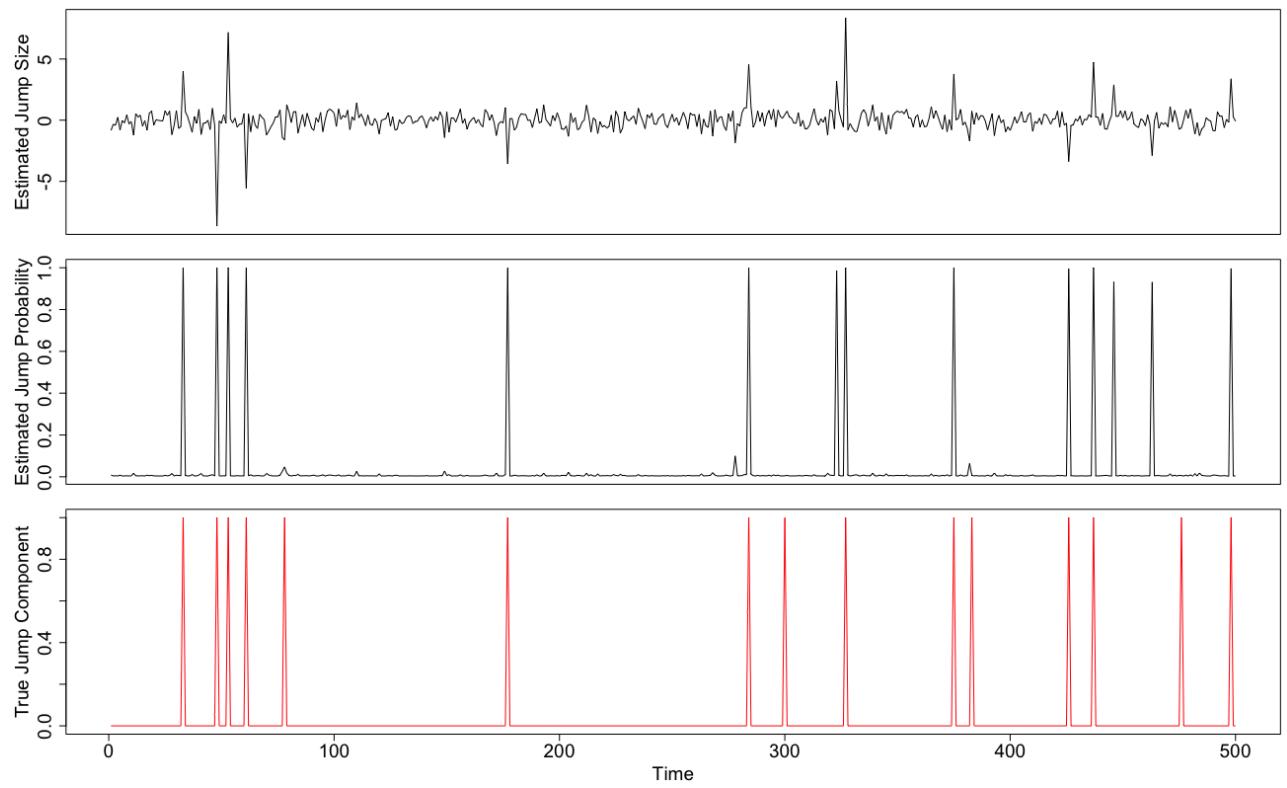


Figure 25: The time series of the observation, state variable, jump component and jump size in simulation.



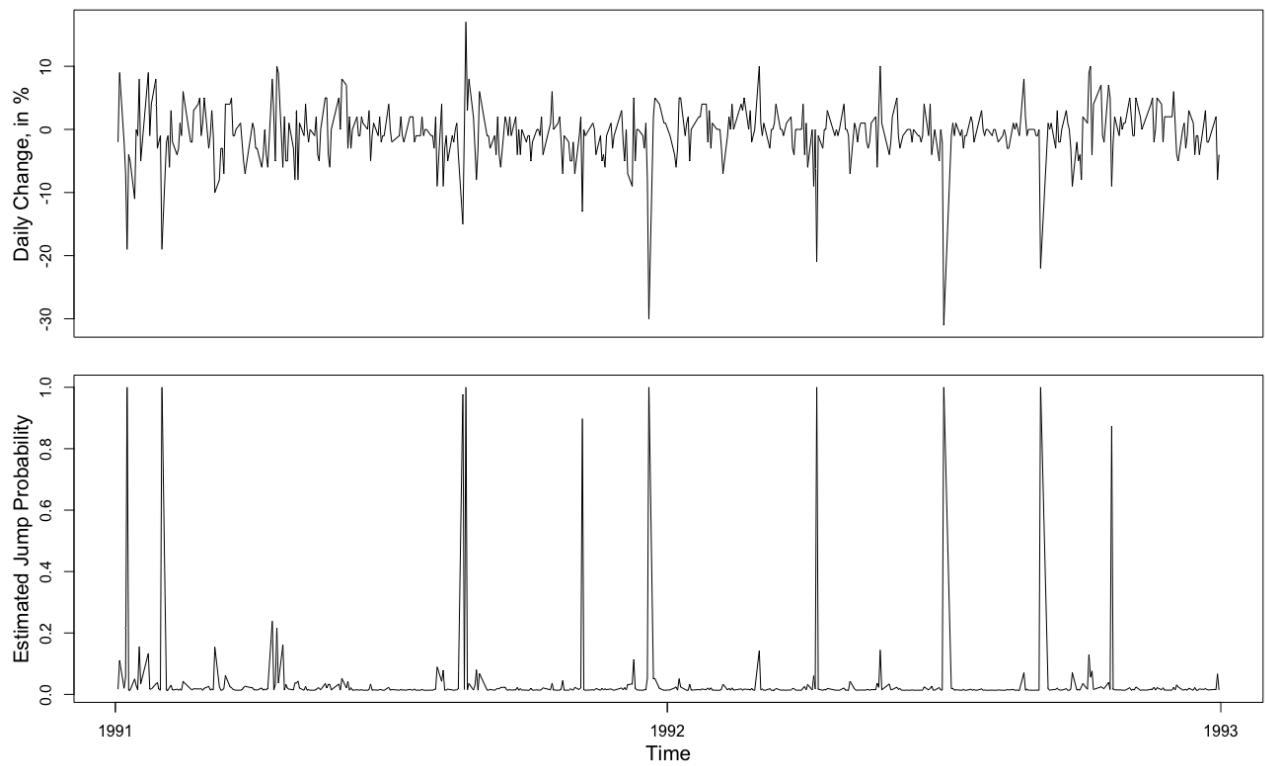
Note: The top panel shows the the observations (y); the second panel from above shows the true values of the state variable (h), the third panel from above shows the true vaules of the jump component (\tilde{J} or J), and the bottom panel shows the true values of the jump term ($\tilde{J} \times Z$).

Figure 26: The time series of the estimated jump size and jump probability, 1991-1993.



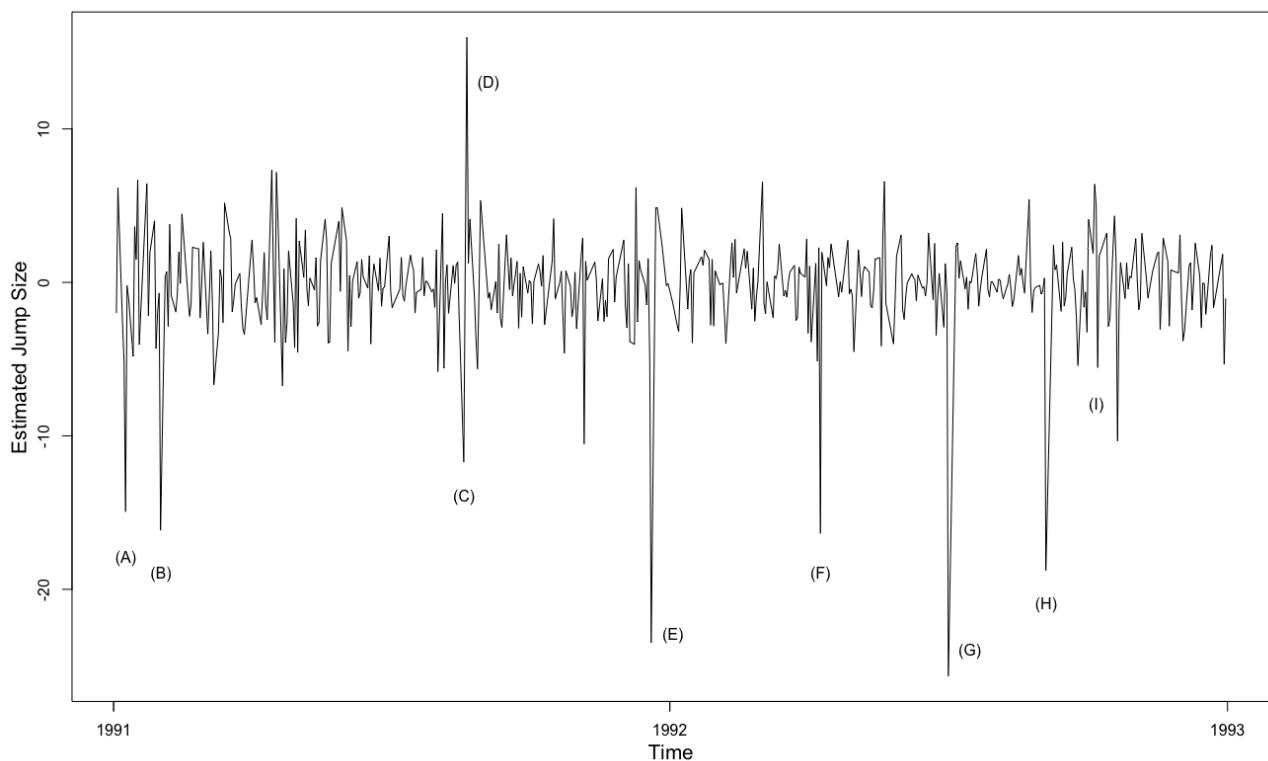
Note: The top panel shows the filtered estimation of the jump size ($E(\tilde{J}_t \times Z_t | y_t, y_{t-1}, \dots)$); the middle panel shows the estimated jump probability ($P(J_t = 1 | y_t, y_{t-1}, \dots)$); and the bottom panel shows the true jump component for comparison

Figure 27: The time series of the short rate changes and filtered jump probability, 1991-1993.



Note: The top panel shows the time series of short rate changes in percentage; and the bottom panel shows the filtered estimation of the jump probability

Figure 28: The time series of the estimate jump size and the related macroeconomic events, 1991-1993.



Note: The annotations (A)-(I) represent the following dates and events: (A) 01/09/1991, the outbreak of the Gulf War; (B) 02/01/1991, U.S. unemployment announcement and comments by the Federal Reserve; (C) 08/19/1991, the collapse of the Soviet Union; (D) 08/21/1991, the emergence of Boris Yeltsin as leader of the remnants of the Soviet Union; (E) 12/20/1991, the Federal Reserve lowers the discount rate; (F) 4/9/1992, large Japanese equity market decline; (G) 7/2/1992, the Federal Reserve lowers the discount rate; (H) 9/4/92, a U.S. unemployment announcement; and, (I) October 1992, the Bush-Clinton presidential debates.