SOME STATISTICAL TOPICS ON SEQUENTIAL DATA ASSIMILATION

by

LUI CHIU SING GILBERT

A thesis submitted in partial fulfilment of requirement for

the Degree of Doctor of Philosophy

at The University of Hong Kong

July 2008

Abstract of thesis entitled

SOME STATISTICAL TOPICS ON SEQUENTIAL DATA ASSIMILATION

Submitted by

LUI CHIU SING GILBERT

for the Degree of Doctor of Philosophy

at The University of Hong Kong

in July 2008

Among many techniques in sequential data assimilation, the ensemble Kalman filter (EnKF), proposed by Geir Evensen in 1994, had become popular in recent years. Nevertheless, EnKF still faces a number of pitfalls in some applications. In this study, two new variants of EnKF are proposed. The first one is initiated by the Goldberger-Theil's mixed estimation and the updating equation of the conventional EnKF under this estimation becomes no longer linear. This variant of the EnKF is called the ensemble Goldberger-Theil Kalman filter (EnGTKF). Due to the possible deviation of prediction and filtering densities from Gaussianity, the multivariate Gram-Charlier densities were suggested for the estimation of the error statistics on both measurements and states. As a by-product, an ensemble Goldberger-Theil Kalman smoother (EnGTKS) was derived in which only information of the prediction density and the likelihood of measurement conditional on the states are required.

The second variant of the EnKF was motivated by the Markov switching of

regimes in the nonlinear state space model. Under the switching between hidden regimes, both the nonlinearity of state space model and the multi-modal feature of the measurement and state errors are considered. The resulting variant is called the ensemble Markov switching Kalman filter (EnMSKF). Due to the increase in the number of components over time, an approximation was suggested on the filtering density to keep this filter operational. As a by-product, a recursive smoother called the ensemble Markov switching Kalman smoother (EnMSKS) was also derived.

For the parameter identification of both models, a hybrid strategy for maximum likelihood estimation was proposed. The algorithm consists of three features (1) an orthogonal transformation procedure was introduced in the optimization algorithm to ensure the non-negative definiteness of the variance-covariance matrix during estimation; (2) a localized stochastic search procedure was suggested to estimate the initial parameter values and (3) conventional local optimization was used to improve the estimation results of previous global procedure. In order to draw statistical inference of the estimated model parameters in the case of EnGTKF, the analytical formulae of standard errors were derived and the ensemble members of prediction density were used for estimation.

Finally, as empirical applications, the algal bloom data in Hong Kong was used to show the validity of the suggested data assimilation procedures.

SOME STATISTICAL TOPICS ON SEQUENTIAL DATA ASSIMILATION

by

LUI CHIU SING GILBERT

A thesis submitted in partial fulfilment of requirement for

the Degree of Doctor of Philosophy

at The University of Hong Kong

July 2008

Declaration

This thesis is submitted to The University of Hong Kong for the award of the degree of Doctor of Philosophy. The whole work has been undertaken after registration for the degree and has not been previously included in a thesis, dissertation or report submitted to this or any other institution for a degree, diploma or other qualifications.

Signed _____

Lui Chiu Sing Gilbert

Acknowledgements

I would like to express my most sincere appreciation to my supervisor Professor W.K. Li, Chair of Statistics in the Department of Statistics and Actuarial Science, The University of Hong Kong for his guidance, support, encouragement and discussion during the period of my PhD study. Another thanks must go to my co-supervisor Professor Joseph H.W. Lee, Redmond Chair of Civil Engineering in the Department of Civil Engineering for his guidance and support of research of water quality. Also, Dr. Kenneth M.Y. Leung in the Department of Ecology and Biodiversity taught me a lot about the ecological risk assessment. The support of the Area of Excellence Scheme under the University Grants Committee of the Hong Kong Special Administration Region, China (Project No. AoE/P-04/2004) and the partial support by Research Grants Council of Hong Kong Special Administration Region, China (Project No. HKU7110/04E) are acknowledged. Finally, I would like to thank the AoE members who have helped me during the period of PhD study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of Sequential Data Assimilation

Data assimilation has been applied to the areas of oceanography and meteorology for many years. Basically, data assimilation consists of two approaches, namely, variational and sequential. Apart from the variational approach which is based on the method of optimal control, sequential data assimilation is less computationally intensive and easy to implement. The sequential data assimilation is closely related to the state space modelling. In this chapter, the relationship between the sequential data assimilation and the state space modelling is explained in details. Indeed, Bertino et al. (2003) provided a detailed review of sequential data assimilation in oceanography and suggested the usage of Gaussian anamorphosis in sequential data assimilation. Therefore, only some applications of sequential data assimilation to oceanography are reviewed here.

In sequential data assimilation, observed measurement $\mathbf{y}_n$ and a set of differential equations which describe the evolutionary dynamics of unobserved model state $\mathbf{x}_n$ are the essential ingredients. In state space modelling, the measurement equation describes the relationship between the observed measurement and the unobserved model state and the transition equation represents the discretized version of the evolutionary dynamics of the model state. Typically, the model state in the state space model is estimated by the recursive prediction and filtering procedures and they are called the forecasting and analysis steps respectively

in the terminology of data assimilation. For example, at time n, given the value of the filtered model state, which is denoted as $\mathbf{x}_{n|n}$, the predicted model state for next period $\mathbf{x}_{n+1|n}$ is estimated by the transition equation. Once the measurement at time $n+1$, $\mathbf{y}_{n+1}$, becomes available, the predicted model state $\mathbf{x}_{n+1|n}$ is updated and the updated model state becomes $\mathbf{x}_{n+1|n+1}$. Therefore, the operation of sequential data assimilation is simple once the initial model state $\mathbf{x}_0$ and the model parameters are available.

Furthermore, smoothing is another direction of sequential data assimilation. Indeed, the model state $\mathbf{x}_n$ is estimated by using all observations in the sample. Suppose that the model state $\mathbf{x}_n$ is denoted as $\mathbf{x}_{n|N}$ where N is the number of observations in the sample. Here, $\mathbf{x}_{n|N}$ represents the mean of model state $\mathbf{x}_n$ conditional on all information at time N. In the terminology of Kalman filtering, the smoother for the estimation of $\mathbf{x}_{n|N}$ is called the fixed-interval smoother in which the model state $\mathbf{x}_{n-1|N}$ is estimated by $\mathbf{x}_{n|N}$ recursively for fixed N. Typically, the smoother is initialized by the filtered model state at the end point of the sample, that is, $\mathbf{x}_{N|N}$, and the estimation of the model state $\mathbf{x}_{n|N}$ is propagated backward in time. Actually, two additional types of smoothers are available in the literature and they are the fixed-point smoother and the fixed-lag smoother. For fixed-point smoother, the recursive estimation of the model state $\mathbf{x}_{n|m}$ by $\mathbf{x}_{n|m-1}$ for $m \geq n$ is obtained by fixing the time point n. Clearly, the fixed-point smoother is initialized by the predicted model state $\mathbf{x}_{n|n-1}$. For the smoother with fixed-lag m, the model state $\mathbf{x}_{n-m|n}$ is estimated by $\mathbf{x}_{n-m|n-1}$ recursively for $0 \leq m \leq N$ where this smoother is initialized by $\mathbf{x}_{0|-1} = \mathbf{x}_0$ and $\mathbf{x}_{-m|-1} = \mathbf{0}$.

The prototype of sequential data assimilation is the Kalman filter (KF) which was developed by Kalman (1960). Afterwards, many variants of KF were devel-

oped, such as, extended Kalman filter (EKF), reduced rank square-root (RRSQRT) filter (Verlaan and Heemink, 1995), singular evolutive extended Kalman (SEEK) filter (Pham et al., 1998) and ensemble Kalman filter (EnKF) (Evensen, 1994).

Early applications of data assimilation with KF are devoted to Talagrand and Courtier (1987) and Courtier and Talagrand (1987) although both of them used the variational approach to assimilate the meteorological data. Since the KF was developed for linear dynamical systems, the usage of EKF which is based on linearization of nonlinear dynamical system is a better alternative for data assimilation, for example, see Evensen (1991). Due to the high-dimensional problem of state space model in the area of meteorology, Verlaan and Heemink (1995) proposed the RRSQRT filter where the computational time and storage space were reduced by using the reduced rank of error covariance matrices with a square root factorization. Verlaan and Heemink (1997) applied the RRSQRT filter with a set of swallow water equations to model the tidal flow of North Sea and then to forecast the tides and storm surges. Their experimental results showed that the RRSQRT filter could reduce the number of computations and storage space without reducing the accuracy of estimated model state a lot. Brasseur et al. (1999) extended the SEEK filter in the directions of statistical learning approach and dynamic adjustment and assimilated the altimetric data in the mid-latitude oceans with the Miami Isopycnic Coordinate Ocean Model (MICOM). Verlaan and Heemink (2001) extended the RRSQRT filter to capture the second order approximation of nonlinear dynamical system and proposed a measure of nonlinearity of dynamical system. Their assimilation results showed that the measure could detect the failure of Kalman-type filtering. Also, they found that, for highly nonlinear dynamical system, EnKF is more appropriate for data assimila-

tion while RRSQRT is useful for weakly nonlinear dynamical system.

Pham (2001) proposed the second-order-exact sampling techniques that could reduce the ensemble size in the Monte Carlo filters. The experimental results demonstrated that the ensemble size was significantly reduced in EnKF and particle filter. Cãnizares et al. (2001) implemented EnKF with the so-called MIKE 21 model (Madsen and Cãnizares, 1999) and DYNOCS (Dynamics of connecting seas) model (Jensen, 1997) to assimilate the water levels of inner Danish waters. Their results demonstrated the possibility of formation of the prototype of storm surge prediction system. Wolf et al. (2001) adopted the RRSQRT filter with the hydrodynamical TRIM3D model (Casulli and Cattani, 1994) to reconstruct the water level data of Odra lagoon which was across the German-Polish border during the flood period in 1997. Their simulation results showed that spatial and temporal data could be assimilated efficiently with the TRIM3D model especially in the case of poorly spatially distributed measurement, for example, the water level around the whole lagoon area. In Bertino et al. (2002), estuarine applications by RRSQRT filter and EnKF with the input of geostatistics were examined. Two simulations, an 1-D ecological model and TRIM3D model on Orda lagoon, were studied. They showed that the data assimilation schemes were efficient for spatial modeling in the estuarine applications. Brusdal et al. (2003) compared the performance of data assimilation of SEEK filter, EnKF and the ensemble Kalman smoother (EnKS) of Evensen and van Leeuwen (2000) by using the nonlinear MICOM model. The merits of these three data assimilation procedures were also discussed in that paper. They found that the performance of these three assimilation methods were similar.

Apart from RRSQRT and SEEK filters which are referred to their original pa-

pers, the basic formulation of previously mentioned filters, together with Gaussian sum filter (GSF) and particle filter, are described in the following sections and some of the materials are followed from Arulampalam et al. (2002) and Chen (2003).

1.1.1 Prediction and Filtering

1.1.1.1 Kalman Filter

The Kalman filter (KF) was developed by Kalman (1960) and the state space model follows a linear dynamics:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{h}_n \mathbf{x}_n + \mathbf{v}_n, & \mathbf{v}_n &\sim N(\mathbf{0}, \mathbf{R}_n), \\ \mathbf{x}_n &= \mathbf{f}_n \mathbf{x}_{n-1} + \mathbf{u}_n, & \mathbf{u}_n &\sim N(\mathbf{0}, \mathbf{Q}_n), \end{aligned} \tag{1.1}$$

for $n = 1, \ldots, N$, where $\mathbf{x}_n \in \mathbb{R}^{m_x}$, $\mathbf{y}_n \in \mathbb{R}^{m_y}$, $\mathbf{h}_n$ and $\mathbf{f}_n$ are matrices of order $m_x \times m_y$ and $m_x \times m_y$ respectively, $\mathbb{E}(\mathbf{v}_n \mathbf{u}_m) = \mathbf{0}$, $\forall m, n$. Denote $\mathcal{Y}_n \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ as a set of measurements up to time n.

To derive the recursive estimation of the state vector $\mathbf{x}_n$, various approaches have appeared in the literature, least squares, mixed estimation, maximum a posteriori (MAP) estimation and maximum likelihood estimation (MLE). Among them, MAP, which indeed is the Bayesian approach, seems to be the most popular one because this approach can be extended to the nonlinear state space model easily. Therefore, the derivation of the state estimation is formulated on the basis of MAP and the derivation can be found in Chen (2003, p.11–12).

Denote $\mathbf{x}_{t|s}$ and $\mathbf{\Sigma}^{\mathbf{xx}}_{t|s}$ as the mean and variance of $\mathbf{x}_t$ conditional on information up to time s. Then, $\mathbf{x}_{t|s} \equiv \mathbb{E}(\mathbf{x}_t|\mathcal{Y}_s)$ and $\mathbf{\Sigma}^{\mathbf{xx}}_{t|s} \equiv \mathbb{E}((\mathbf{x}_t - \mathbf{x}_{t|s})(\mathbf{x}_t - \mathbf{x}_{t|s})^{\mathsf{T}}|\mathcal{Y}_s)$ and their estimates are represented by $\widehat{\mathbf{x}}_{t|s}$ and $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{t|s}$ respectively. Given the initial

state $\widehat{\mathbf{x}}_{0|0}$ and the variance matrix $\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{0|0}$, the recursive state estimation is given by

$$\begin{aligned}
\widehat{\mathbf{x}}_{n|n-1} &= \mathbf{f}_n \widehat{\mathbf{x}}_{n-1|n-1}, \\
\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1} &= \mathbf{f}_n \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n-1|n-1} \mathbf{f}_n^{\mathsf{T}} + \mathbf{Q}_{n-1}, \\
\widehat{\mathbf{x}}_{n|n} &= \widehat{\mathbf{x}}_{n|n-1} + \widehat{\mathbf{K}}_n (\mathbf{y}_n - \mathbf{h}_n \widehat{\mathbf{x}}_{n|n-1}), \\
\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n} &= \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1} - \widehat{\mathbf{K}}_n \mathbf{h}_n \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1},
\end{aligned}$$

for $n = 1, \ldots, N$, where $\widehat{\mathbf{K}}_n = \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1} \mathbf{h}_n^{\mathsf{T}} (\mathbf{h}_n \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1} \mathbf{h}_n^{\mathsf{T}} + \mathbf{R}_n)^{-1}$ is the Kalman gain matrix. When the recursive formulation is expressed in the form of density functions, they become

$$\begin{aligned}
p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_{n-1}; \widehat{\mathbf{x}}_{n-1|n-1}, \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n-1|n-1}), \\
p(\mathbf{x}_n|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n-1}, \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1}), \\
p(\mathbf{x}_n|\mathcal{Y}_n) &\approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n}, \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n}),
\end{aligned}$$

where

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{m_x}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

1.1.1.2 Extended Kalman Filter

For the EKF, the formulation is similar to that of KF, but a nonlinear state space model is considered instead.

$$\begin{aligned} \mathbf{y}_n &= \mathbf{h}_n(\mathbf{x}_n) + \mathbf{v}_n, \qquad & \mathbf{v}_n &\sim N(\mathbf{0}, \mathbf{R}_n), \\ \mathbf{x}_n &= \mathbf{f}_n(\mathbf{x}_{n-1}) + \mathbf{u}_n, \qquad & \mathbf{u}_n &\sim N(\mathbf{0}, \mathbf{Q}_n), \end{aligned} \tag{1.2}$$

for $n = 1, \ldots, N$ where $\mathbf{x}_n \in \mathbb{R}^{m_x}$, $\mathbf{y}_n \in \mathbb{R}^{m_y}$, $\mathbf{h}_n : \mathbb{R}^{m_x} \to \mathbb{R}^{m_y}$ and $\mathbf{f}_n : \mathbb{R}^{m_x} \to \mathbb{R}^{m_x}$ are nonlinear continuous and differentiable functions, $\mathbb{E}(\mathbf{v}_n \mathbf{u}_m) = \mathbf{0}$, $\forall m, n$. Given the initial state $\widehat{\mathbf{x}}_{0|0}$ and the variance matrix $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{0|0}$, the recursive state estimation is formulated by

$$\begin{aligned} \widehat{\mathbf{x}}_{n|n-1} &= \mathbf{f}_n(\widehat{\mathbf{x}}_{n-1|n-1}), \\ \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1} &= \widetilde{\mathbf{f}}_n \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|n-1} \widetilde{\mathbf{f}}_n^{\mathsf{T}} + \mathbf{Q}_{n-1}, \\ \widehat{\mathbf{x}}_{n|n} &= \widehat{\mathbf{x}}_{n|n-1} + \widehat{\mathbf{K}}_n(\mathbf{y}_n - \mathbf{h}_n(\widehat{\mathbf{x}}_{n|n-1})), \\ \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n} &= \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1} - \widehat{\mathbf{K}}_n \widetilde{\mathbf{h}}_n \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1}, \\ \widehat{\mathbf{K}}_n &= \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1} \widetilde{\mathbf{h}}_n^{\mathsf{T}} (\widetilde{\mathbf{h}}_n \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1} \widetilde{\mathbf{h}}_n^{\mathsf{T}} + \mathbf{R}_n)^{-1}, \end{aligned}$$

for $n = 1, \ldots, N$, where

$$\widetilde{\mathbf{f}}_n = \left. \frac{d\mathbf{f}_n(\mathbf{x})}{d\mathbf{x}} \right|_{\mathbf{x}=\widehat{\mathbf{x}}_{n-1|n-1}} \text{ and } \widetilde{\mathbf{h}}_n = \left. \frac{d\mathbf{h}_n(\mathbf{x})}{d\mathbf{x}} \right|_{\mathbf{x}=\widehat{\mathbf{x}}_{n|n-1}}.$$

Then, the prediction and filtering densities of the state vector $\mathbf{x}_n$ are estimated recursively by

$$\begin{aligned} p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_{n-1}; \widehat{\mathbf{x}}_{n-1|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|n-1}), \\ p(\mathbf{x}_n|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1}), \end{aligned}$$

$$p(\mathbf{x}_n|\mathcal{Y}_n) \approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n}).$$

From the above formulae, one can observe that the coefficient matrices in KF are replaced by the first order derivatives in EKF. Under EKF, both functions $\mathbf{h}_n(\cdot)$ and $\mathbf{f}_n(\cdot)$ are approximated by the respective first order term in their Taylor's expansion. One typical problem of EKF under nonlinear dynamical system is the possible divergence of the filter because the first order approximation may not be sufficient to capture the curvature of the functions $\mathbf{h}_n(\cdot)$ and $\mathbf{f}_n(\cdot)$ when the dynamical system is highly nonlinear. Although higher order approximation is feasible, the formulation becomes more sophisticated, for example, see Tanizaki (1996, p.52–55).

1.1.1.3 Gaussian Sum Filter

One suggestion to overcome the possible divergence of EKF is that the prediction and filtering densities be approximated by the sums of Gaussian densities. This initiates the development of the Gaussian sum filter (GSF). GSF was originated by Sorenson and Alspach (1971) and Alspach and Sorenson (1972). The nonlinear state space model considered is the same as that in EKF. The derivation of GSF by Bayesian estimation can be found in Anderson and Moore (1979, p.214–216), Sorenson and Alspach (1971) and Tanizaki (1996, p.73–77). The recursive estimation of the state vector $\mathbf{x}_n$ is given by the following formula: suppose that at time $n-1$, the filtering density of the state vector $\mathbf{x}_{n-1}$ is approximated by

$$p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) \approx \sum_{j=1}^{m} \widehat{\alpha}_{j,n-1} \mathcal{N}(\mathbf{x}_{n-1}; \widehat{\mathbf{x}}_{j,n-1|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n-1|n-1}),$$

for $j = 1, \ldots, m$ with $0 \leq \widehat{\alpha}_{j,n-1} \leq 1$ and $\sum_{j=1}^{m} \widehat{\alpha}_{j,n-1} = 1$. Then,

$$\begin{aligned}
\widehat{\mathbf{x}}_{j,n|n-1} &= \mathbf{f}_n(\widehat{\mathbf{x}}_{j,n-1|n-1}), \\
\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1} &= \widetilde{\mathbf{f}}_n(\widehat{\mathbf{x}}_{j,n-1|n-1})\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n-1|n-1}\widetilde{\mathbf{f}}_n^{\mathsf{T}}(\widehat{\mathbf{x}}_{j,n-1|n-1}) + \mathbf{Q}_{n-1},
\end{aligned}$$

for $j = 1, \ldots, m$ and $n = 1, \ldots, N$, where

$$\widetilde{\mathbf{f}}_n(\widehat{\mathbf{x}}_{j,n-1|n-1}) = \left.\frac{d\mathbf{f}_n(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\widehat{\mathbf{x}}_{j,n-1|n-1}}.$$

The overall prediction mean and variance of the state vector $\mathbf{x}_n$ are

$$\widehat{\mathbf{x}}_{n|n-1} = \sum_{j=1}^{m} \widehat{\alpha}_{j,n-1}\widehat{\mathbf{x}}_{j,n|n-1},$$

and

$$\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1} = \sum_{j=1}^{m} \widehat{\alpha}_{j,n-1}\left(\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1} + (\widehat{\mathbf{x}}_{j,n|n-1} - \widehat{\mathbf{x}}_{n|n-1})(\widehat{\mathbf{x}}_{j,n|n-1} - \widehat{\mathbf{x}}_{n|n-1})^{\mathsf{T}}\right),$$

respectively. Then, the prediction density of the state vector $\mathbf{x}_n$ is approximated by

$$p(\mathbf{x}_n|\mathcal{Y}_{n-1}) \approx \sum_{j=1}^{m} \widehat{\alpha}_{j,n-1}\mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{j,n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}),$$

During the filtering stage, the predicted state vector $\widehat{\mathbf{x}}_{j,n|n-1}$ is updated by

$$\widehat{\mathbf{x}}_{j,n|n} = \widehat{\mathbf{x}}_{j,n|n-1} + \widehat{\mathbf{K}}_{j,n}(\mathbf{y}_n - \mathbf{h}_n(\widehat{\mathbf{x}}_{j,n|n-1})),$$

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n} &= \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n-1} - \widehat{\mathbf{K}}_{j,n}\widetilde{\mathbf{h}}_n(\widehat{\mathbf{x}}_{j,n|n-1})\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}, \\
\widehat{\mathbf{K}}_{j,n} &= \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}\widetilde{\mathbf{h}}_n^{\mathsf{T}}(\widehat{\mathbf{x}}_{j,n|n-1})(\widetilde{\mathbf{h}}_n(\widehat{\mathbf{x}}_{j,n|n-1})\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}\widetilde{\mathbf{h}}_n^{\mathsf{T}}(\widehat{\mathbf{x}}_{j,n|n-1}) + \mathbf{R}_n)^{-1}, \\
\widehat{\alpha}_{j,n} &= \frac{\widehat{\alpha}_{j,n-1}\mathcal{N}(\mathbf{y}_n; \mathbf{h}_n(\widehat{\mathbf{x}}_{j,n|n-1}), \widetilde{\mathbf{h}}_n(\widehat{\mathbf{x}}_{j,n|n-1})\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}\widetilde{\mathbf{h}}_n^{\mathsf{T}}(\widehat{\mathbf{x}}_{j,n|n-1}) + \mathbf{R}_n)}{\sum_{i=1}^m \widehat{\alpha}_{i,n-1}\mathcal{N}(\mathbf{y}_n; \mathbf{h}_n(\widehat{\mathbf{x}}_{i,n|n-1}), \widetilde{\mathbf{h}}_n(\widehat{\mathbf{x}}_{i,n|n-1})\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{i,n|n-1}\widetilde{\mathbf{h}}_n^{\mathsf{T}}(\widehat{\mathbf{x}}_{i,n|n-1}) + \mathbf{R}_n)},
\end{aligned}
$$

for $j = 1, \ldots, m$, where

$$
\widetilde{\mathbf{h}}_n(\widehat{\mathbf{x}}_{j,n|n-1}) = \left.\frac{d\mathbf{h}_n(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\widehat{\mathbf{x}}_{j,n|n-1}}.
$$

The overall filtered mean and variance of the state vector $\mathbf{x}_n$ are

$$
\widehat{\mathbf{x}}_{n|n} = \sum_{j=1}^m \widehat{\alpha}_{j,n}\widehat{\mathbf{x}}_{j,n|n}
$$

and

$$
\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n} = \sum_{j=1}^m \widehat{\alpha}_{j,n}\left(\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n} + (\widehat{\mathbf{x}}_{j,n|n} - \widehat{\mathbf{x}}_{n|n})(\widehat{\mathbf{x}}_{j,n|n} - \widehat{\mathbf{x}}_{n|n})^{\mathsf{T}}\right),
$$

respectively. The filtering density of the state vector $\mathbf{x}_n$ is approximated by

$$
p(\mathbf{x}_n|\mathcal{Y}_n) \approx \sum_{j=1}^m \widehat{\alpha}_{j,n}\mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{j,n|n}, \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{j,n|n}).
$$

Therefore, the GSF can be considered as a parallel run of m EKF's at the same time and the recursive approximation of prediction and filtering densities by Gaussian sums. The component means and variances are then combined after each stage is completed.

1.1.1.4 Particle Filter

In previously described filters, the measurement and state errors are Gaussian. When they are non-Gaussian, Kalman-type filtering does not seem to be appropriate and this initiates the development of the particle filter. Particle filtering was originated from the sequential Monte Carlo method which was developed by various authors individually. Sometimes, it may be called bootstrap filtering (Gordon et al., 1993), condensation algorithm (MacCormick and Blake, 2000), particle filtering (Carpenter et al., 1999), interacting particle approximation (Crisan et al., 1999; del Moral, 1996) and survival of the fittest (Kanazawa et al., 1995). Indeed, Arulampalam et al. (2002), Doucet et al. (2001) and Chen (2003) provided the overview and practical details of the particle filter. Therefore, only the main feature of the particle filter is presented here.

Suppose that the nonlinear state space model is

$$\begin{aligned} \mathbf{y}_n &= \mathbf{h}_n(\mathbf{x}_n) + \mathbf{v}_n, \\ \mathbf{x}_n &= \mathbf{f}_n(\mathbf{x}_{n-1}) + \mathbf{u}_n. \end{aligned} \tag{1.3}$$

for $n = 1, \ldots, N$, where the density functions of $\mathbf{v}_n$ and $\mathbf{u}_n$ may not be Gaussian.

Basically, particle filtering can be considered as the estimation of the posterior density of the state vector by a set of particles which are generated by the Monte Carlo method. Indeed, a technique of sequential importance sampling (SIS) is used here. In addition to the notation of $\mathcal{Y}_n$, denote $\mathcal{X}_n \equiv \{\mathbf{x}_0, \ldots, \mathbf{x}_n\}$ as a set of state vectors up to time n. Note that the initial state vector is included in $\mathcal{X}_n$. Under the particle filter, the posterior density of $\mathcal{X}_n$ is approximated by

$$p(\mathcal{X}_n|\mathcal{Y}_n) \approx \sum_{k=1}^{K} w_n^{(k)} \delta(\mathcal{X}_n - \mathcal{X}_n^{(k)}),$$

where $\delta(\cdot)$ is the delta-Dirac mass, $\{\mathcal{X}_n^{(k)}, w_n^{(k)}\}_{k=1}^K$ represents a set of particles $\{\mathcal{X}_n^{(k)}\}_{k=1}^K$ which are weighted by $\{w_n^{(k)}\}_{k=1}^K$ respectively with $0 \leq w_n^{(k)} \leq 1$ and $\sum_{k=1}^K w_n^{(k)} = 1$.

However, in many cases, drawing random samples from $p(\mathcal{X}_n|\mathcal{Y}_n)$ is not an easy task. Thus, one may draw the particles $\{\mathcal{X}_n^{(k)}\}_{k=1}^K$ from an importance density $q(\mathcal{X}_n|\mathcal{Y}_n)$ and the weights are re-defined as

$$w_n^{(k)} \propto \frac{p(\mathcal{X}_n^{(k)}|\mathcal{Y}_n)}{q(\mathcal{X}_n^{(k)}|\mathcal{Y}_n)}.$$

Furthermore, assume that the importance density can be factorized as

$$q(\mathcal{X}_n|\mathcal{Y}_n) = q(\mathbf{x}_n|\mathcal{X}_{n-1}, \mathcal{Y}_n) q(\mathcal{X}_{n-1}|\mathcal{Y}_{n-1}).$$

Then, the set of particles $\{\mathcal{X}_n^{(k)}\}_{k=1}^K$ can be generated from $\{\mathcal{X}_{n-1}^{(k)}\}_{k=1}^K$ with the augmentation of $q(\mathbf{x}_n|\mathcal{X}_{n-1}, \mathcal{Y}_n)$ recursively. The recursive derivation of the importance weights in the posterior density can be obtained from

$$\begin{aligned}
p(\mathcal{X}_n|\mathcal{Y}_n) &= \frac{p(\mathbf{y}_n|\mathcal{X}_n, \mathcal{Y}_{n-1}) p(\mathcal{X}_n|\mathcal{Y}_{n-1})}{p(\mathbf{y}_n|\mathcal{Y}_{n-1})} \\
&= \frac{p(\mathbf{y}_n|\mathcal{X}_n, \mathcal{Y}_{n-1}) p(\mathbf{x}_n|\mathcal{X}_{n-1}, \mathcal{Y}_{n-1}) p(\mathcal{X}_{n-1}|\mathcal{Y}_{n-1})}{p(\mathbf{y}_n|\mathcal{Y}_{n-1})} \\
&= \frac{p(\mathbf{y}_n|\mathbf{x}_n) p(\mathbf{x}_n|\mathbf{x}_{n-1}) p(\mathcal{X}_{n-1}|\mathcal{Y}_{n-1})}{p(\mathbf{y}_n|\mathcal{Y}_{n-1})} \\
&\propto p(\mathbf{y}_n|\mathbf{x}_n) p(\mathbf{x}_n|\mathbf{x}_{n-1}) p(\mathcal{X}_{n-1}|\mathcal{Y}_{n-1}),
\end{aligned}$$

where the last equality follows from the assumed state space model. Followed from the re-defined weights, the importance weights are written as

$$
\begin{aligned}
w_n^{(k)} &\propto \frac{p(\mathbf{y}_n|\mathbf{x}_n^{(k)})p(\mathbf{x}_n^{(k)}|\mathbf{x}_{n-1}^{(k)})p(\mathcal{X}_{n-1}^{(k)}|\mathcal{Y}_{n-1})}{q(\mathbf{x}_n^{(k)}|\mathcal{X}_{n-1}^{(k)},\mathcal{Y}_n)q(\mathcal{X}_{n-1}^{(k)}|\mathcal{Y}_{n-1})} \\
&= w_{n-1}^{(k)}\frac{p(\mathbf{y}_n|\mathbf{x}_n^{(k)})p(\mathbf{x}_n^{(k)}|\mathbf{x}_{n-1}^{(k)})}{q(\mathbf{x}_n^{(k)}|\mathcal{X}_{n-1}^{(k)},\mathcal{Y}_n)}.
\end{aligned}
$$

If a more restrictive assumption $q(\mathbf{x}_n|\mathcal{X}_{n-1},\mathcal{Y}_n) = q(\mathbf{x}_n|\mathcal{X}_{n-1},\mathbf{y}_n)$ is imposed, the weights can be simplified further as

$$
w_n^{(k)} \propto w_{n-1}^{(k)}\frac{p(\mathbf{y}_n|\mathbf{x}_n^{(k)})p(\mathbf{x}_n^{(k)}|\mathbf{x}_{n-1}^{(k)})}{q(\mathbf{x}_n^{(k)}|\mathcal{X}_{n-1}^{(k)},\mathbf{y}_n)},
$$

or

$$
w_n^{(k)} = \frac{w_{n-1}^{(k)}p(\mathbf{y}_n|\mathbf{x}_n^{(k)})p(\mathbf{x}_n^{(k)}|\mathbf{x}_{n-1}^{(k)})/q(\mathbf{x}_n^{(k)}|\mathcal{X}_{n-1}^{(k)},\mathbf{y}_n)}{\sum_{j=1}^{K} w_{n-1}^{(j)}p(\mathbf{y}_n|\mathbf{x}_n^{(j)})p(\mathbf{x}_n^{(j)}|\mathbf{x}_{n-1}^{(j)})/q(\mathbf{x}_n^{(j)}|\mathcal{X}_{n-1}^{(j)},\mathbf{y}_n)}.
$$

The filtering density of the model state $\mathbf{x}_n$ becomes

$$
p(\mathbf{x}_n|\mathcal{Y}_n) \approx \sum_{k=1}^{K} w_n^{(k)}\delta(\mathbf{x}_n - \mathbf{x}_n^{(k)}).
$$

Due to the possible degeneration of weights $w_n^{(k)}$, a resampling step was suggested by Gordon et al. (1993) in which the K particles $\{\mathbf{x}_n^{(k)}\}_{k=1}^K$ are resampled with replacement according to the importance weights $\{w_n^{(k)}\}_{k=1}^K$. Nevertheless, the resampling step occurs when $\widehat{N}_{eff} \equiv \{\sum_{k=1}^K (w_n^{(k)})^2\}^{-1}$ is less than a threshold value under conventional SIS. Ever since, many variants of the particle filter were developed, for example, sequential importance resampling (SIR) particle filter (Gordon et al., 1993), auxiliary particle filter (Pitt and Shephard, 1999), rejection particle filter (Tanizaki, 1999) and regularized particle filter (Musso et al.,

2001).

1.1.1.5 Ensemble Kalman Filter

EnKF was suggested by Evensen (1994) and the comprehensive review and technical treatment can be found in Evensen (2003) and Evensen (2007). One advantage of this filter is that only the mean and variance of the state vector $\mathbf{x}_n$ are propagated over time in contrast to the particle filter where the whole posterior density function of the state vector $\mathbf{x}_n$ is propagated. Also, the mean and variance are approximated by a cloud of ensemble members and even in highly nonlinear dynamical systems this results in an improvement of estimation of mean and variance of the state vector $\mathbf{x}_n$ over EKF.

Suppose that the nonlinear state space model is

$$\begin{aligned} \mathbf{y}_n &= \mathbf{h}_n \mathbf{x}_n + \mathbf{v}_n, & \mathbf{v}_n &\sim N(\mathbf{0}, \mathbf{R}_n), \\ \mathbf{x}_n &= \mathbf{f}_n(\mathbf{x}_{n-1}) + \mathbf{u}_n, & \mathbf{u}_n &\sim N(\mathbf{0}, \mathbf{Q}_n), \end{aligned} \tag{1.4}$$

for $n = 1, \ldots, N$, where $\mathbf{x}_n \in \mathbb{R}^{m_x}$, $\mathbf{y}_n \in \mathbb{R}^{m_y}$, $\mathbf{f}_n : \mathbb{R}^{m_x} \to \mathbb{R}^{m_x}$ is a nonlinear function, $\mathbf{h}_n$ is a $m_x \times m_y$ matrix, $\mathbb{E}(\mathbf{v}_n \mathbf{u}_m) = \mathbf{0}$, $\forall m, n$. Given the mean and variance of the initial state, $\widehat{\mathbf{x}}_{0|0}$ and $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{0|0}$, the ensemble members of the state vector $\mathbf{x}_n$ are generated by

$$\widehat{\mathbf{x}}^{(k)}_{n|n-1} = \mathbf{f}_n(\widehat{\mathbf{x}}^{(k)}_{n-1|n-1}) + \mathbf{u}^{(k)}_n, \qquad \mathbf{u}^{(k)}_n \sim N(\mathbf{0}, \mathbf{Q}_n),$$

for $n = 1, \ldots, N$ and $k = 1, \ldots, K$ where $\widehat{\mathbf{x}}^{(k)}_{n-1|n-1}$ is the kth ensemble member in the previous filtering stage, $\mathbf{u}^{(k)}_n$ is drawn from the Gaussian density $\mathcal{N}(\mathbf{u}_n; \mathbf{0}, \mathbf{Q}_n)$. The ensemble mean and variance of the predicted state vector $\mathbf{x}_n$ are

$$\widehat{\mathbf{x}}_{n|n-1} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{x}}_{n|n-1}^{(k)},$$

and

$$\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}} = \frac{1}{K-1}\sum_{k=1}^{K}(\widehat{\mathbf{x}}_{n|n-1}^{(k)} - \widehat{\mathbf{x}}_{n|n-1})(\widehat{\mathbf{x}}_{n|n-1}^{(k)} - \widehat{\mathbf{x}}_{n|n-1})^{\mathsf{T}},$$

respectively. During the filtering stage, each ensemble member is updated by the following linear updating equation.

$$\widehat{\mathbf{x}}_{n|n}^{(k)} = \widehat{\mathbf{x}}_{n|n-1}^{(k)} + \widehat{\mathbf{K}}_n(\mathbf{y}_n^{(k)} - \mathbf{h}_n\widehat{\mathbf{x}}_{n|n-1}^{(k)}),$$

for $k = 1, \ldots, K$, where

$$\begin{aligned}
\mathbf{y}_n^{(k)} &= \mathbf{y}_n + \mathbf{v}_n^{(k)}, \qquad \mathbf{v}_n^{(k)} \sim N(\mathbf{0}, \mathbf{R}_n), \\
\widehat{\mathbf{K}}_n &= \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}}\mathbf{h}_n^{\mathsf{T}}(\mathbf{h}_n\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}}\mathbf{h}_n^{\mathsf{T}} + \mathbf{R}_n)^{-1}.
\end{aligned}$$

The ensemble filtered mean and variance can be obtained in a similar way, that is,

$$\widehat{\mathbf{x}}_{n|n} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{x}}_{n|n}^{(k)} \quad \text{and} \quad \widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}} = \frac{1}{K-1}\sum_{k=1}^{K}(\widehat{\mathbf{x}}_{n|n}^{(k)} - \widehat{\mathbf{x}}_{n|n})(\widehat{\mathbf{x}}_{n|n}^{(k)} - \widehat{\mathbf{x}}_{n|n})^{\mathsf{T}}.$$

Note that the measurement $\mathbf{y}_n$ is perturbed by the stochastic error $\mathbf{v}_n^{(k)}$ from Gaussian density $\mathcal{N}(\mathbf{v}_n; \mathbf{0}, \mathbf{R}_n)$. This is the special feature of EnKF and its validity was shown by Burgers et al. (1998). The recursive approximation of the

posterior density functions of the state vector $\mathbf{x}_n$ is given by

$$\begin{aligned} p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_{n-1}; \widehat{\mathbf{x}}_{n-1|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|n-1}), \\ p(\mathbf{x}_n|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1}), \\ p(\mathbf{x}_n|\mathcal{Y}_n) &\approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n}). \end{aligned}$$

Therefore, EnKF can be considered as the recursive Gaussian approximation of the prediction and filtering densities. This also explains why EnKF is an suboptimal solution of particle filter when $\mathbf{v}_n$ and $\mathbf{u}_n$ are not Gaussian.

1.1.2 Smoothing

1.1.2.1 Kalman Smoother

The Kalman smoother was derived on the basis of the linear state space model (1.1). Here, only the fixed-interval smoother is considered since the state vector $\mathbf{x}_n$ can be estimated by this smoother more accurately intuitively. For the fixed-point and fixed-lag smoothers, the derivation and technical details can be found in (Anderson and Moore, 1979, Chapter 7). For the fixed-interval smoother which estimates the model state $\mathbf{x}_n$ by using all information of the sample, given the filtered model state $\widehat{\mathbf{x}}_{N|N}$ and the variance matrix $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{N|N}$, the recursive estimation of $\mathbf{x}_{n|N}$ and its variance is given by

$$\begin{aligned} \widehat{\mathbf{x}}_{n-1|N} &= \widehat{\mathbf{x}}_{n-1|n-1} + \widehat{\mathbf{A}}_{n-1}(\widehat{\mathbf{x}}_{n|N} - \widehat{\mathbf{x}}_{n|n-1}), \\ \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|N} &= \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|n-1} + \widehat{\mathbf{A}}_{n-1}(\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|N} - \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1})(\widehat{\mathbf{A}}_{n-1})^{\mathsf{T}}, \end{aligned}$$

for $n = N, \ldots, 2$, where $\widehat{\mathbf{A}}_{n-1} = \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n-1|n-1}\mathbf{f}_{n-1}^{\mathsf{T}}(\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1})^{-1}$ is the gain matrix for the fixed-interval smoother. Clearly, the recursive formula requires both prediction and filtering results. In density form, the smoothing density of the state vector $\mathbf{x}_n$ is approximated recursively by

$$\begin{aligned} p(\mathbf{x}_{n-1}|\mathcal{Y}_N) &\approx \mathcal{N}(\mathbf{x}_{n-1}; \widehat{\mathbf{x}}_{n-1|N}, \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n-1|N}), \\ p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_{n-1}; \widehat{\mathbf{x}}_{n-1|n-1}, \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n-1|n-1}), \\ p(\mathbf{x}_n|\mathcal{Y}_{n-1}) &\approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n-1}, \widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1}). \end{aligned}$$

1.1.2.2 Gaussian Sum Smoother

As proposed by Kitagawa (1994), the Gaussian sum smoother was derived for the linear state space model

$$\begin{aligned} \mathbf{y}_n &= \mathbf{x}_n + \mathbf{v}_n, & \mathbf{v}_n &\sim N(\mathbf{0}, \mathbf{R}_n), \\ \mathbf{x}_n &= \mathbf{f}_n\mathbf{x}_{n-1} + \mathbf{u}_n, & \mathbf{u}_n &\sim N(\mathbf{0}, \mathbf{Q}_n), \end{aligned} \tag{1.5}$$

for $n = 1, \ldots, N$.

Basically, the Gaussian sum smoother consists of two components: the derivation of two-filter formula (which consists of the conventional filtering and the backward filtering algorithms) and the Gaussian sum approximation. Denote $\mathcal{Y}^n \equiv \{\mathbf{y}_n, \ldots, \mathbf{y}_N\}$ which contains the current and future information of the measurement. Then, by the Bayes' Theorem, the smoothing density $p(\mathbf{x}_n|\mathcal{Y}_N)$ is expressed as

$$
\begin{aligned}
p(\mathbf{x}_n|\mathcal{Y}_N) &= p(\mathbf{x}_n|\mathcal{Y}_{n-1}, \mathcal{Y}^n) \\
&\propto p(\mathcal{Y}^n|\mathbf{x}_n)p(\mathbf{x}_n|\mathcal{Y}_{n-1}).
\end{aligned}
$$

Clearly, the smoothing density $p(\mathbf{x}_n|\mathcal{Y}_N)$ is determined by the one-step ahead prediction density $p(\mathbf{x}_n|\mathcal{Y}_{n-1})$ and the density $p(\mathcal{Y}^n|\mathbf{x}_n)$. Once the recursive derivation of $p(\mathcal{Y}^n|\mathbf{x}_n)$ is available, the smoothing density $p(\mathbf{x}_n|\mathcal{Y}_N)$ can be derived recursively. Then, the recursive formula of $p(\mathcal{Y}^n|\mathbf{x}_n)$ is derived by the backward filtering algorithm. Specifically,

$$
\begin{aligned}
p(\mathcal{Y}^{n+1}|\mathbf{x}_n) &= \int p(\mathcal{Y}^{n+1}, \mathbf{x}_{n+1}|\mathbf{x}_n)d\mathbf{x}_{n+1} \\
&= \int p(\mathcal{Y}^{n+1}|\mathbf{x}_{n+1}, \mathbf{x}_n)p(\mathbf{x}_{n+1}|\mathbf{x}_n)d\mathbf{x}_{n+1} \\
&= \int p(\mathcal{Y}^{n+1}|\mathbf{x}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{x}_n)d\mathbf{x}_{n+1},
\end{aligned}
$$

and

$$
\begin{aligned}
p(\mathcal{Y}^n|\mathbf{x}_n) &= p(\mathcal{Y}^{n+1}, \mathbf{y}_n|\mathbf{x}_n) \\
&= p(\mathbf{y}_n|\mathbf{x}_n, \mathcal{Y}^{n+1})p(\mathcal{Y}^{n+1}|\mathbf{x}_n) \\
&= p(\mathbf{y}_n|\mathbf{x}_n)p(\mathcal{Y}^{n+1}|\mathbf{x}_n).
\end{aligned}
$$

for $n = N-1, \ldots, 1$. When $n = N$, $p(\mathcal{Y}^n|\mathbf{x}_n) = p(\mathbf{y}_N|\mathbf{x}_N)$.

Now, assume that the densities $p(\mathcal{Y}^n|\mathbf{x}_n)$ and $p(\mathbf{x}_n|\mathcal{Y}_{n-1})$ are approximated

by the sum of Gaussian densities, that is,

$$p(\mathcal{Y}^n|\mathbf{x}_n) \approx \sum_{i=1}^{\ell} \widehat{\beta}_{i,n} \mathcal{N}(\mathbf{y}_n; \widehat{\mathbf{y}}_{i,n|n}, \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{i,n|n}),$$

and

$$p(\mathbf{x}_n|\mathcal{Y}_{n-1}) \approx \sum_{j=1}^{m} \widehat{\gamma}_{j,n} \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{j,n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}),$$

with $0 \leq \widehat{\beta}_{i,n}, \widehat{\gamma}_{j,n} \leq 1$ for $i = 1, \ldots, \ell; j = 1, \ldots, m$ and $\sum_{i=1}^{\ell} \widehat{\beta}_{i,n} = \sum_{j=1}^{m} \widehat{\gamma}_{j,n} = 1$. Then, the Gaussian sum smoother is obtained by

$$\begin{aligned} p(\mathbf{x}_n|\mathcal{Y}_N) &\propto p(\mathcal{Y}^n|\mathbf{x}_n) p(\mathbf{x}_n|\mathcal{Y}_{n-1}) \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{m} \widehat{\beta}_{i,n} \widehat{\gamma}_{j,n} \mathcal{N}(\mathbf{y}_n; \widehat{\mathbf{y}}_{i,n|n}, \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{i,n|n}) \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{j,n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}) \\ &\equiv \sum_{i=1}^{\ell} \sum_{j=1}^{m} \widehat{\beta}_{i,n} \widehat{\gamma}_{j,n} \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{ij,n|N}, \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{ij,n|N}), \end{aligned}$$

where

$$\begin{aligned} \widehat{\mathbf{J}}_{ij,n} &= \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1} (\widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{i,n|n} + \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1})^{-1}, \\ \widehat{\mathbf{x}}_{ij,n|N} &= \widehat{\mathbf{x}}_{j,n|n-1} + \widehat{\mathbf{J}}_{ij,n} (\widehat{\mathbf{y}}_{i,n|n} - \widehat{\mathbf{x}}_{j,n|n-1}), \\ \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{ij,n|N} &= (\mathbf{I} - \widehat{\mathbf{J}}_{ij,n}) \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{j,n|n-1}, \end{aligned}$$

for $i = 1, \ldots, \ell$ and $j = 1, \ldots, m$.

After renumbering the double summation by a single summation, the smoothing density $p(\mathbf{x}_n|\mathcal{Y}_N)$ can also be expressed in the form of a Gaussian sum.

1.1.2.3 Particle Smoother

The particle smoother was derived based on the nonlinear state space model (1.3). At the beginning, the particle filter is run to propagate forward in time so that the results of the prediction and filtering densities of the state vector $\mathbf{x}_n$ are stored. Then, the density of $\mathcal{X}^{n-1} \equiv \{\mathbf{x}_{n-1}, \ldots, \mathbf{x}_N\}$ conditional on all measurements $\mathcal{Y}_N$ is obtained backward in time recursively.

$$\begin{aligned} p(\mathcal{X}^{n-1}|\mathcal{Y}_N) &= p(\mathbf{x}_{n-1}, \mathcal{X}^n|\mathcal{Y}_N) \\ &= p(\mathbf{x}_{n-1}|\mathcal{X}^n, \mathcal{Y}_N)p(\mathcal{X}^n|\mathcal{Y}_N) \\ &= p(\mathbf{x}_{n-1}|\mathcal{X}^n, \mathcal{Y}_{n-1})p(\mathcal{X}^n|\mathcal{Y}_N) \\ &= p(\mathcal{X}^n|\mathcal{Y}_N)\frac{p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathcal{Y}_{n-1})p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1})}{p(\mathbf{x}_n|\mathcal{Y}_{n-1})}, \end{aligned}$$

for $n = N-1, \ldots, 2$. The third equality is obtained by the Markovian property of $\mathbf{x}_n$ and the fourth equality is derived by the Bayes' Theorem. As a usual practice in particle filtering, the smoothing density $p(\mathcal{X}^{n-1}|\mathcal{Y}_N)$ can be approximated by a set of particles:

$$p(\mathcal{X}^{n-1}|\mathcal{Y}_N) \approx \sum_{k=1}^{K} w_N^{(k)} \delta(\mathcal{X}^{n-1} - \mathcal{X}^{n-1(k)}),$$

where $\delta(\cdot)$ is the delta-Dirac mass, $\{\mathcal{X}^{n-1(k)}, w_N^{(k)}\}_{k=1}^K$ represents a set of particles $\{\mathcal{X}^{n-1(k)}\}_{k=1}^K$ which are weighted by $\{w_N^{(k)}\}_{k=1}^K$ respectively with $0 \leq w_N^{(k)} \leq 1$ and $\sum_{k=1}^K w_N^{(k)} = 1$. One may see that the smoothing density $p(\mathbf{x}^{n-1}|\mathcal{Y}_N)$ may be obtained by marginalizing out $\mathcal{X}^n$. However, it is infeasible in the practical sense because of the degeneration of weights during resampling. On the other

hand, Doucet et al. (2000) proposed a fixed-interval particle smoother to tackle this problem. The smoothing density $p(\mathbf{x}_{n-1}|\mathcal{Y}_N)$ can be written in the following recursive form:

$$p(\mathbf{x}_{n-1}|\mathcal{Y}_N) = p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) \int \frac{p(\mathbf{x}_n|\mathcal{Y}_N)p(\mathbf{x}_n|\mathbf{x}_{n-1})}{p(\mathbf{x}_n|\mathcal{Y}_{n-1})} d\mathbf{x}_n.$$

Then, the smoothing density $p(\mathbf{x}_{n-1}|\mathcal{Y}_N)$ is approximated by

$$p(\mathbf{x}_{n-1}|\mathcal{Y}_N) \approx \sum_{k=1}^{K} w_{n-1|N}^{(k)} \delta(\mathbf{x}_n - \mathbf{x}_n^{(k)}),$$

where the smoothing weight $w_{n-1|N}^{(k)}$ is obtained by the following formula recursively.

$$w_{n-1|N}^{(k)} = \sum_{k'=1}^{K} w_{n|N}^{(k)} \frac{w_{n-1}^{(k)} p(\mathbf{x}_n^{(k')}|\mathbf{x}_{n-1}^{(k)})}{\sum_{k=1}^{K} w_{n-1}^{(k)} p(\mathbf{x}_n^{(k')}|\mathbf{x}_{n-1}^{(k)})},$$

for $n = N-1, \ldots, 2$ where $w_{n-1}^{(k)}$ is obtained in the filtering stage. Clearly, when $n = N$, the smoothing weight $w_{n|N}^{(k)}$ is initialized by $w_N^{(k)}$ and this setting is reasonable because the filtering density and the smoothing density are the same when $n = N$.

Furthermore, the particle smoother can also be derived from the rejection particle filters and the details can be found in Kitagawa (1996), Tanizaki and Mariano (1998) and Hüzeler and Künsch (1998).

1.2 Some Backgrounds of Algal Blooms

Generally, algal blooms can be considered as the dramatic growth of phytoplankton cells. In some cases, they are called "red tides" due to the red color pigments

in the phytoplankton cells. Usually, red tides occur during the spring season. One consequence of algal blooms is hypoxia because a huge amount of dissolved oxygen is consumed by the large population of phytoplankton cells. Another consequence is the release of toxin into water. This in turn leads to the massive kill of fish, shellfish and marine mammals by various toxins, such as, ciguatoxin (ciguatera poisoning), brevetoxin (neurotoixc poisoning) and saxitoxin (paralytic poisoning). The consumption of poisoned fish and shellfish is also harmful to the human health. On the other hand, the blooms of blue-green algae (cyanobacteria) are also typical in waters. Certain kinds of toxin by these algae are harmful to the liver and cause eye and skin irritation because of prolonged exposure in the waters. Therefore, the study of algal dynamics is essential to maintain the diversity of mariculture and the human health.

Mainly, phytoplankton which causes algal blooms consists of two types, namely, diatoms and dinoflagellates. The biological behavior of these two species varies a lot due to the difference in the biological structure. For diatom cells, they have a thick cell wall which is made of silicate and hence their biomass is typically larger than that of dinoflagellates. On the other hand, dinoflagellates do not possess the silicate cell walls, but most of them have two dissimilar flagella. They can move along the water column with diurnal variation of environmental factors, such as water temperature and solar radiation. Therefore, during daytime, dinoflagellates can move to the water surface for photosynthesis and produce oxygen and glucose. During nighttime, they can move downward to the sea bed for respiration and absorption of nutrients. However, diatoms do not have the same biological behavior. Due to its biomass, they can move along the water column by upwelling and downwelling with turbulence. Therefore, their production of oxygen

is relied on the upwelling effect of turbulence (Kamykowski and Yamazaki, 1997; Yamamoto and Okai, 2000). The decomposition of dinoflagellates and diatoms after death raises the level of nutrient concentration and then induces the growth of phytoplankton cells.

Generally, the occurrence of algal blooms is determined by many environmental factors, for example, concentration of chlorophyll-a, concentration of dissolved oxygen, concentration of nutrients (nitrates and phosphates mainly), water temperature, wind direction, wind speed, turbulence, solar radiation, salinity and so on (see, for example, Thomann and Mueller, 1987). However, the algal dynamics is not known clearly. Typically, the concentration of nitrates and phosphates is a crucial factor to cause algal blooms. Due to the biological structure of diatoms, the concentration of silicate is a factor for their growth as well. Although one can consider that the fertilization of nitrates and phosphates leads to the algal blooms, it need not be the case in certain scenarios, such as the seasonal variation of water temperature and the upwelling of nutrients by turbulence. One of the complexities could be due to various biological behavior across the species of phytoplankton and the mix of species in the waters. In addition, due to the stochastic behavior of environmental factors, for example, solar radiation, wind direction, wind speed and turbulence, these also enhance the difficulty of understanding of the algal bloom dynamics. These in turn lead to the outcome of many physical models which attempt to explain the algal dynamics. For example, Franks (1997) reviewed the models of harmful algal blooms and they were categorized into four types, (1) aggregated models; (2) multispecies models; (3) models with simple physics and (4) model with detailed physics. Recently, Chattopadhyay et al. (2002) and Chattopadhyay et al. (2004) suggested mathematical models to

explain the interaction behavior between phytoplankton and zooplankton where both toxin producing phytoplankton and non-toxic phytoplankton were included into the models. Then, the behavior of the models were verified by the field data and numerical simulations.

Due to the increasing complexity of mathematical models, sequential data assimilation with these models becomes more difficult than before. Therefore, statistical models could be another direction to explain the algal dynamics and they are attempted to use in this thesis.

1.3 Outline of the Thesis

In this thesis, two extensions of ensemble Kalman filter (EnKF) are proposed in Chapters 2 and 3 respectively. Specifically, the nonlinear updating equation in EnKF is shown in Chapter 2 while the EnKF with Markov switching structure is given in Chapter 3.

In Chapter 2, since the linear updating equation is implemented with the conventional version of EnKF, a nonlinear updating equation is suggested to improve EnKF and the existence of the nonlinear updating equation is derived mathematically. The estimation of ensemble states is derived using Goldberger-Theil's mixed estimation (Theil and Goldberger, 1961). The variant of EnKF under this kind of operation is called the ensemble Goldberger-Theil Kalman filter (EnGTKF). Due to the possible severe deviation of asymptotic Gaussianity of estimated ensemble mean and variances, a multivariate version of Gram-Charlier density is suggested to approximate the prediction and filtering densities of the model state. This suggestion can be extended to the derivation of the likeli-

hood function in parameter identification. Furthermore, as a natural extension of EnGTKF, a recursive ensemble smoother, ensemble Kalman Goldberger-Theil smoother (EnGTKS), is derived and the formulation is different from those by van Leeuwen and Evensen (1996) and Evensen and van Leeuwen (2000). Under this ensemble smoother, only information of the prediction density of the model state and the likelihood of measurement conditional on the predicted model state are required to be stored. Thus, it is computationally favorable and can save a lot of storage space.

For parameter identification, a hybrid approach of maximum likelihood estimation is adopted here. Traditionally, the likelihood function is obtained by the decomposition of the measurement prediction error and the unknown parameters are estimated by conventional local optimization algorithms, for example, Newton-Raphson and Quasi-Newton. Due to the high-dimensional feature in the state space model, the positive definiteness of error covariance matrices of the measurements and model state are difficult to maintain. As a result, a kind of orthogonal decomposition of the matrices is suggested together with a localized stochastic search algorithm, a kind of global optimization algorithm. The global search result is used as the initial parameter estimates of conventional local optimization algorithms. To draw statistical inference on the parameters, a recursive estimation method for the observed Fisher's information matrix is suggested. Then, asymptotic standard errors of estimated parameters can be obtained accordingly.

As an application of the newly derived filter, the algal bloom data in Hong Kong is assimilated. The complete sampling period is 2000-2004. The observations in 2000-2001 are selected as the in-sample period, that is, they are used

for parameter identification, and those in 2002-2004 are selected as the out-of-the-sample period. To capture the transitions between non-blooming and blooming periods, the time-varying vector smoothed transition autoregressive (TV-VSTAR) model is proposed to fit the observations. To validate the prediction performance of the TV-STAR model, the vector autoregressive (VAR) model is used as a benchmark. For the selection of lag parameters in the TV-VSTAR model, the one with the smallest Akaike information criterion (AIC) during the in-sample period is selected as the appropriate one.

In Chapter 3, although the assumption of Gaussian error in the measurement and model state is sufficient in many applications, this assumption is inappropriate in certain cases. As a result, the structure of Markov switching in unobserved regimes is introduced into the EnKF and this new filter is called the ensemble Markov switching Kalman filter (EnMSKF). This new filter serves two purposes: the switching between nonlinear dynamics and the use of mixtures of Gaussian densities. Unlike the dynamic linear state space model by Kim (1994), the model state is estimated by ensemble members. Hence, even under switching between nonlinear dynamic models, the ensemble mean and variance can be estimated in the usual way. The growth in the number of unobserved regimes over time is controlled by the marginalization of the filtering density with the same filtering mean and variance. As a by-product of this filter, a new version of the ensemble smoother, the ensemble Markov switching Kalman smoother (EnMSKS), is derived and this smoother can act as an alternative smoother for the non-Gaussian case.

For parameter identification, the hybrid optimization algorithm suggested in Chapter 2 is applied again. Furthermore, due to the complexity of the likelihood

function of the Markov switching model, a recursive method is suggested to estimate the observed Fisher's information matrix. Then, the standard errors of model parameters can be estimated for statistical inference. To choose the number of switching regimes, the Bayesian information criterion (BIC) is suggested.

For an application of EnMSKF, some assimilations of algal bloom data in Hong Kong are carried out. As in Chapter 2, the complete sampling period covers 2000-2004 is splitted into in-sample and out-of-sample periods by the same criterion. The underlying model used is a Markov switching vector autoregressive (MS-VAR) model. The two features of this model may be appropriate for the algal bloom data. The MS-VAR model can capture the sudden changes in algal bloom data during blooming periods and the drops in biomass of phytoplankton cells near the end of blooming periods. Furthermore, the prediction probability of algal blooms can also be produced and this is informative in constructing an algal bloom alarm system.

Finally, the conclusions of the thesis are drawn are drawn in Chapter 4.

Chapter 2

Ensemble Kalman Filter with Nonlinear Updating Equation

2.1 Introduction

Kalman (1960) developed the Kalman filter (KF) which has been applied to various aspects of engineering, atmospheric science, economics, finance for a long period of time. One advantage of KF is easy to implement. However, its optimality is maintained under the assumptions of linearity of model with Gaussian errors. Afterwards, many variants of KF have been proposed, such as the extended Kalman filter (EKF), iterative extended Kalman filter (IEKF), square root filter (SQKF) and so on. The details of these filters can be found in Jazwinski (1970) and Anderson and Moore (1979).

The ensemble Kalman filter (EnKF) was proposed by Evensen (1994) and has been used in the atmospheric science over 10 years. Indeed, EnKF could be considered as a suboptimal solution of the particle filter (for example, Carlin et al., 1992; Gordon et al., 1993; del Moral, 1996; Crisan et al., 1999; Gilks and Berzuini, 2001; and Arulampalam et al., 2002) in which the whole density function is estimated by simulations. Under the framework of EnKF, the means and variances of prediction and filtering densities are obtained by simulated measurements and states. Then, by the law of large numbers, the estimated means and variances are consistent. Unlike EKF where the linearization of the state space model is required, no linearization is required in EnKF and the nonlinearity of the

state space model is captured by a cloud of ensemble members. Evensen (2003) and Evensen (2007) gave an overview of development of EnKF and discussed the implementation issues of this filter.

Mainly, the applications of EnKF have been focused on atmospheric data assimilation due to its historical background. For example, Eknes and Evensen (2002) applied EnKF to assimilate biological data, that is, concentrations of nutrient, phytoplankton and zooplankton. They showed that the EnKF could handle the nonlinear instabilities of data assimilation during Spring algal bloom. In particular, the assimilation of concentration of phytoplankton seemed to denominate the concentrations of nutrient and zooplankton and this resulted in the control of the whole assimilation system. Haugen and Evensen (2002) implemented the EnKF to monitor and predict the variations of Indian Ocean by using remotely sensed observations of sea-level anomaly and sea-surface temperature with the Miami Isopycnic Coordinate Ocean Model (MICOM). They demonstrated that the EnKF can control the model evolution over time effectively. Also, the multivariate correlation between variables were highly anisotropic and dependent on location. Mitchell et al. (2002) studied the effects of ensemble size and the localization on EnKF with a global forecast model by the Canadian Meteorological Centre. Their results indicated that an increase in the ensemble size and more severe localization could improve the assimilation of atmospheric data although the computational cost could be expensive. Vrugt et al. (2005) applied a global optimization algorithm with EnKF to assimilate streamflow data of the Leaf River watershed which was located north of Collins, Mississippi. They demonstrated the possibility of simultaneous optimization, which indeed was a process of model calibration, and data assimilation.

In most cases, EnKF can produce reasonable and consistent predictions of states and measurements in the sense of minimum mean squared errors. Inevitably, it still faces the problems of state estimation and parameter estimation.

Firstly, EnKF faced a problem that its updating equation is linear in both new measurement and predicted states. Consequently, in the case of nonlinear measurements and/or states, the conditional variance of filtered states may not be minimized and this implies that measurement data during the updating step may not be used efficiently. This initiates the proposal of a new EnKF, which is called the ensemble Goldberger-Theil Kalman filter (EnGTKF), in this chapter. Under this new filter, the updating equation needs no longer be linear. The derivation of the nonlinear updating equation is shown as an extension of the linear updating equation of the conventional EnKF. Furthermore, since the prediction densities of states and measurements and the filtering density of states may deviate from that of asymptotic Gaussianity, the construction of error statistics by ensemble averages may not be appropriate. Consequently, the nonlinear updating equation is extended to adapt to the case of non-Gaussian density. Indeed, the multivariate Gram-Charlier densities by Perote and del Brío (2006) is suggested to approximate the densities so that the error statistics for states and measurements can then be calculated accordingly. This specification of multivariate Gram-Charlier density ensures its positive density over the support without specific restrictions on the parameter values.

As a by-product of EnGTKF, the ensemble Goldberger-Theil Kalman smoother (EnGTKS) is derived. Indeed, the smoothed estimates of mean and variance-covariance matrix of states and measurements are expressed in the forms of weighted ensemble averages where the weights are determined by the likelihood

of measurements conditional on states recursively.

Secondly, focused on the issue of parameter estimation, due to the large numbers of unknown parameters in multivariate state space model, maximization of the likelihood function by conventional estimation methods, such as the Newton-Raphson and Quasi-Newton methods, is typical, but they may not be suitable in the current situation because the positive definiteness of a high-dimensional variance-covariance matrix is hard to maintain. One may suggest that global optimization algorithm can find the estimation results easily. However, one pitfall of global optimization algorithm is its accuracy, except in the case of discrete parameter values. On the other hand, the local optimization algorithms can provide accurate parameter estimates. However, the optimal solutions can be trapped in local optimal when the objective function is not unimodal. To tackle the problems of parameter estimation, a hybrid optimization procedure is suggested to estimate the unknown parameters. Indeed, the proposed algorithm consists of three components, namely, (1) localized stochastic search algorithm; (2) evolutionary strategy, a kind of evolutionary algorithm; and (3) local optimization algorithm, for example Quasi-Newton algorithm. A brief description of this algorithm can be given as follows. A transformation procedure in evolutionary strategy is applied to decompose the variance-covariance matrices into products of orthogonal rotation matrices which are determined by rotation angles only. Then, a localized stochastic search algorithm is applied to search for the optimal parameter estimates. Finally, a local optimization algorithm is used to enhance the estimation results in the global optimization. In addition to the estimation of unknown parameters in the model, drawing inference on unknown parameters is also essential. Therefore, the derivatives of the objective function are essential

in estimation of standard errors and they are also derived in this chapter.

In the followings, the specification of the model and its assumptions are given in Section 2.2. The derivation of the ensemble Goldberger-Theil Kalman filter (EnGTKF) and the ensemble Goldberger-Theil Kalman smoother (EnGTKS) is presented in Section 2.3. Then, a detailed description of the estimation algorithm is provided in Section 2.4. The estimation of standard errors of estimated parameters is then discussed in Section 2.5. Since the prediction and filtering densities may deviate from the asymptotic Gaussianity, a method to approximate the non-Gaussian densities by the multivariate Gram-Charlier expansion is suggested in Section 2.6. In Section 2.7, numerical simulations and empirical application of EnGTKF are provided. Finally, conclusions are drawn in Section 2.8.

2.2 The Model

Consider the following general nonlinear state space model:

$$\mathbf{y}_n = \mathbf{h}\left(\mathbf{x}_n; \boldsymbol{\xi}_n\right) + \mathbf{v}_n, \tag{2.1}$$

$$\mathbf{x}_n = \mathbf{f}\left(\mathbf{x}_{n-1}; \boldsymbol{\xi}_n\right) + \mathbf{u}_n, \tag{2.2}$$

for $n = 1, \ldots, N$, where $\mathbf{y}_n \in \mathbb{R}^{m_y}$, $\mathbf{x}_n \in \mathbb{R}^{m_x}$, $\mathbf{h} : \mathbb{R}^{m_x} \to \mathbb{R}^{m_y}$ and $\mathbf{f} : \mathbb{R}^{m_x} \to \mathbb{R}^{m_x}$ are measurable functions. Also, $\boldsymbol{\xi}_n \in \mathbb{R}^p$ is a vector of model parameters. Both functions $\mathbf{h}(\cdot)$ and $\mathbf{f}(\cdot)$ are assumed to be smooth, that is, all derivatives of $\mathbf{h}(\cdot)$ and $\mathbf{f}(\cdot)$ exist and continuous. In the state space modelling, equations (2.1) and (2.2) are known as the measurement equation and the transition equation respectively. The specification of (2.1) and (2.2) is fairly general including many

parametric and non-parametric nonlinear time series models. Particularly non-parametric models such as smoothing splines and nonparametric regression can be considered as special cases of (2.1) and (2.2). Note that the time-varying parameters can be included in the state equation. Under this situation, the augmented state vector which contains the model state vector and the time-varying parameter. simultaneously. Indeed, Kitagawa (1998) investigated this problem carefully and considered the time-varying parameters to follow a random walks as an example.

The density functions of $\mathbf{v}_n$ and $\mathbf{u}_n$ are assumed to be

$$p\left(\mathbf{u}_n\right) = \mathcal{N}\left(\mathbf{u}_n; \mathbf{0}, \mathbf{Q}_n\right), \tag{2.3}$$

and

$$p\left(\mathbf{v}_n\right) = \mathcal{N}\left(\mathbf{v}_n; \mathbf{0}, \mathbf{R}_n\right), \tag{2.4}$$

respectively where

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{m_x}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Here, both $\mathbf{Q}_n$ and $\mathbf{R}_n$ are assumed to be non-negative definite. In the cases of singular $\mathbf{Q}_n$ and $\mathbf{R}_n$, their generalized inverses are used instead. Both m_x and m_y are assumed to be finite. The density function of the initial state $\mathbf{x}_0$ is defined by

$$p\left(\mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_0; \mathbf{x}_{0|0}, \boldsymbol{\Sigma}_{0|0}\right).$$

The assumptions for the conditional density functions of $\mathbf{y}_n$ and $\mathbf{x}_n$ are given below:

Assumption 2.1 Denote $\mathcal{X}_n \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{Y}_n \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. Then,

(i) Conditional on $\mathbf{x}_n$, current measurement $\mathbf{y}_n$ is independent of the past information of $\mathbf{x}_n$ and $\mathbf{y}_n$, $\mathcal{X}_{n-1}$ and $\mathcal{Y}_{n-1}$, that is,

$$p\left(\mathbf{y}_n|\mathcal{X}_n, \mathcal{Y}_{n-1}\right) = p\left(\mathbf{y}_n|\mathbf{x}_n, \mathcal{X}_{n-1}, \mathcal{Y}_{n-1}\right) = p\left(\mathbf{y}_n|\mathbf{x}_n\right) \quad \text{for } n = 2, \ldots, N,$$

and

$$p\left(\mathbf{y}_n|\mathcal{X}_n\right) = p(\mathbf{y}_n|\mathbf{x}_n) \quad \text{for } n = 1.$$

The conditional independence assumption is implied by the measurement equation (2.1).

(ii) The state process $\{\mathbf{x}_n\}_{n=1}^N$ is Markovian and homogenous. Also, it admits an invariant probability measure, that is,

$$p\left(\mathbf{x}_n|\mathcal{X}_{n-1}, \mathcal{Y}_{n-1}\right) = p\left(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathcal{X}_{n-2}, \mathcal{Y}_{n-1}\right) = p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) \quad \text{for } n = 1, \ldots, N.$$

(iii) The Markov chain $\{\mathbf{x}_n\}_{n=1}^N$ is irreducible and aperiodic.

(iv) The initial distribution of $\mathbf{x}_0$ is stationary.

Assumption 2.1 implies that given the values of $\mathbf{x}_n$, the density function of $\mathbf{y}_n$ can be derived directly. Furthermore, the density function of the state vector $\mathbf{x}_n$ can be derived only from the state vector $\mathbf{x}_n$ at time $n-1$. Indeed, condition (ii) implies that the Markov chain $\{\mathbf{x}_n\}_{n=1}^N$ is ergodic (Chan and Tong, 2001, p.34). In addition to the assumption of Harris recurrent Markov chain $\{\mathbf{x}_n\}_{n=1}^N$, by Meyn

and Tweedie (1993, Theorem 13.0.1), the state process $\{\mathbf{x}_n\}_{n=1}^N$ converges to a stationary distribution for every initial condition $\mathbf{x}_0 \in \mathbb{R}^{m_x}$. This theorem also ensures the stability of the state space model of (2.1) and (2.2) since the Foster's condition (Meyn and Tweedie, 1993, p.501) is satisfied automatically. To enable the EnKF, more assumptions on the independence of $\mathbf{u}_n$, $\mathbf{v}_n$ and $\mathbf{x}_0$ are necessary and given below:

Assumption 2.2

(i) The disturbances $\mathbf{u}_n$ and $\mathbf{v}_n$ are uncorrelated within themselves over n and with each other for all time periods, that is, $\mathbb{E}\left(\mathbf{u}_n\mathbf{v}_m^{\mathsf{T}}\right) = \mathbf{0}$ for all m and n.

(ii) The initial state vector $\mathbf{x}_0$ is uncorrelated with $\mathbf{u}_n$ and $\mathbf{v}_n$, that is, $\mathbb{E}\left(\mathbf{u}_n\mathbf{x}_0^{\mathsf{T}}\right) = \mathbb{E}\left(\mathbf{v}_n\mathbf{x}_0^{\mathsf{T}}\right) = \mathbf{0}$ for $n = 1, \ldots, N$.

This assumption can simplify the result of nonlinear updating equation.

2.3 Recursive Estimation of Model States

Now, a new version of EnKF for the general nonlinear state space model is derived in this section. The prediction step of the new filter follows from Evensen's EnKF and the nonlinear updating rule is provided in Section 2.3.2. For the state estimation, assume that parameters $\{\boldsymbol{\xi}_n, \mathbf{R}_n, \mathbf{Q}_n\}_{n=1}^N$, $\mathbf{x}_{0|0}$ and $\boldsymbol{\Sigma}_{0|0}$ are known. The prediction and filtering densities for the state vector $\mathbf{x}_n$ are given by

$$\begin{aligned} p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) &= \int p\left(\mathbf{x}_n, \mathbf{x}_{n-1}|\mathcal{Y}_{n-1}\right) d\mathbf{x}_{n-1} \\ &= \int p\left(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}\right) p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) d\mathbf{x}_{n-1}, \text{ for } n \geq 1 \end{aligned} \quad (2.5)$$

$$\text{and} \quad p(\mathbf{x}_n|\mathcal{Y}_n) = p(\mathbf{x}_n|\mathbf{y}_n, \mathcal{Y}_{n-1})$$
$$= \frac{p(\mathbf{x}_n|\mathcal{Y}_{n-1})\, p(\mathbf{y}_n|\mathbf{x}_n)}{p(\mathbf{y}_n|\mathcal{Y}_{n-1})}$$
$$= \frac{p(\mathbf{x}_n|\mathcal{Y}_{n-1})\, p(\mathbf{y}_n|\mathbf{x}_n)}{\int p(\mathbf{x}_n|\mathcal{Y}_{n-1})\, p(\mathbf{y}_n|\mathbf{x}_n)\, d\mathbf{x}_n}, \text{ for } n > 1 \tag{2.6}$$

respectively. The derivation of the above densities requires Assumptions 2.1(i) and (ii) and the Bayes' Theorem. To initiate EnKF, the initial density $p(\mathbf{x}_0|\mathbf{y}_0)$ is assumed to be

$$p(\mathbf{x}_0|\mathbf{y}_0) = \frac{p(\mathbf{y}_0|\mathbf{x}_0)\, p(\mathbf{x}_0)}{p(\mathbf{y}_0)} = p(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0; \mathbf{x}_{0|0}, \mathbf{\Sigma}^{\mathbf{xx}}_{0|0}), \tag{2.7}$$

Implicitly, $p(\mathbf{y}_0|\mathbf{x}_0) = p(\mathbf{y}_0)$ is assumed in the above expression.

2.3.1 Prediction

Due to a flaw in the literature of the conventional EnKF, it is desirable to investigate the statistical properties of ensemble mean and ensemble variance in the prediction stage.

Assume that the filtering density for the state vector $\mathbf{x}_n$ for $n \geq 1$ is

$$p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) = \mathcal{N}(\mathbf{x}_{n-1}; \mathbf{x}_{n-1|n-1}, \mathbf{\Sigma}^{\mathbf{xx}}_{n-1|n-1}), \tag{2.8}$$

where $\mathbf{x}_{n-1|n-1}$ and $\mathbf{\Sigma}^{\mathbf{xx}}_{n-1|n-1}$ denote the mean and variance of $\mathbf{x}_{n-1}$ conditional on $\mathcal{Y}_{n-1}$ respectively, that is,

$$\mathbf{x}_{n-1|n-1} \equiv \int \mathbf{x}_{n-1} p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1})\, d\mathbf{x}_{n-1}$$

and

$$\boldsymbol{\Sigma}^{\mathbf{xx}}_{n-1|n-1} \equiv \int \left(\mathbf{x}_{n-1} - \mathbf{x}_{n-1|n-1}\right)\left(\mathbf{x}_{n-1} - \mathbf{x}_{n-1|n-1}\right)^{\mathsf{T}} p\left(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}\right) d\mathbf{x}_{n-1}.$$

Although the conditional mean and variance are functions of parameter $\boldsymbol{\xi}_n$ implicitly, $\boldsymbol{\xi}_n$ is omitted in the formula for notation simplification.

Then, from (2.5), the conditional mean and variance of the one-step ahead prediction for the state vector $\mathbf{x}_n$ at time n given the measurement up to time $n-1$ are

$$\mathbf{x}_{n|n-1} \equiv \int \mathbf{x}_n p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) d\mathbf{x}_n$$

and

$$\boldsymbol{\Sigma}^{\mathbf{xx}}_{n|n-1} \equiv \int \left(\mathbf{x}_n - \mathbf{x}_{n|n-1}\right)\left(\mathbf{x}_n - \mathbf{x}_{n|n-1}\right)^{\mathsf{T}} p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) d\mathbf{x}_n,$$

where $\mathbf{x}_{n|n-1}$ and $\boldsymbol{\Sigma}^{\mathbf{xx}}_{n|n-1}$ denote the mean and variance of $\mathbf{x}_n$ conditional on $\mathcal{Y}_{n-1}$ respectively.

Under the criterion of minimum mean squared error (MMSE), the one-step ahead prediction mean $\mathbf{x}_{n|n-1}$ and variance $\boldsymbol{\Sigma}^{\mathbf{xx}}_{n|n-1}$ for state vector $\mathbf{x}_n$ are estimated when its conditional prediction variance is minimized:

$$\mathbf{x}_{n|n-1} = \underset{\widehat{\mathbf{x}}_n \in \mathbb{R}^{m_x}}{\operatorname{argmin}} \mathbb{E}\left[\left(\widehat{\mathbf{x}}_n - \mathbf{x}_n\right)^{\mathsf{T}}\left(\widehat{\mathbf{x}}_n - \mathbf{x}_n\right)\middle| \mathcal{Y}_{n-1}\right].$$

The one-step ahead prediction estimate of $\mathbf{x}_n$ is its conditional mean

$$\begin{aligned}
\mathbf{x}_{n|n-1} &= \mathbb{E}\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) \\
&= \mathbb{E}\left(\mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)+\mathbf{u}_n|\mathcal{Y}_{n-1}\right) \\
&= \int \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right) p\left(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}\right) d\mathbf{x}_{n-1} + \int \mathbf{u}_n p\left(\mathbf{u}_n\right) d\mathbf{u}_n,
\end{aligned}$$

in which the first term can be approximated by a cloud of ensemble members $\mathbf{f}(\mathbf{x}^{(k)}_{n-1|n-1};\boldsymbol{\xi}_n)$ where $\mathbf{x}^{(k)}_{n-1|n-1}$ is drawn from the Gaussian density $p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1})$ for $k=1,\ldots,K$. The second term can be approximated by an ensemble of $\mathbf{u}_n$ which is drawn from the density function $p\left(\mathbf{u}_n\right)=\mathcal{N}\left(\mathbf{u}_n;\mathbf{0},\mathbf{Q}_n\right)$. This results in the one-step ahead prediction for $\mathbf{x}_n$ by K ensemble members.

$$\widehat{\mathbf{x}}^{(k)}_{n|n-1} = \mathbf{f}(\widehat{\mathbf{x}}^{(k)}_{n-1|n-1};\boldsymbol{\xi}_n)+\mathbf{u}^{(k)}_n, \quad \mathbf{u}^{(k)}_n \sim N\left(\mathbf{0},\mathbf{Q}_n\right), \text{ for } k=1,\ldots,K, \tag{2.9}$$

$$\widehat{\mathbf{x}}_{n|n-1} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{x}}^{(k)}_{n|n-1}, \tag{2.10}$$

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n-1} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{x}}^{(k)}_{n|n-1}-\widehat{\mathbf{x}}_{n|n-1}\right)\left(\widehat{\mathbf{x}}^{(k)}_{n|n-1}-\widehat{\mathbf{x}}_{n|n-1}\right)^{\mathsf{T}}. \tag{2.11}$$

On the other hand, the one-step ahead prediction density for the measurement vector $\mathbf{y}_n$ is

$$\begin{aligned}
p\left(\mathbf{y}_n|\mathcal{Y}_{n-1}\right) &= \int p\left(\mathbf{y}_n,\mathbf{x}_n|\mathcal{Y}_{n-1}\right) d\mathbf{x}_n \\
&= \int p\left(\mathbf{y}_n|\mathbf{x}_n\right) p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) d\mathbf{x}_n.
\end{aligned} \tag{2.12}$$

Similar to the prediction of the state vector $\mathbf{x}_n$, the one-step ahead prediction

for the measurement vector $\mathbf{y}_n$ is estimated by minimizing its prediction variance

$$\mathbf{y}_{n|n-1} = \underset{\widehat{\mathbf{y}}_n \in \mathbb{R}^{my}}{\operatorname{argmin}} \mathbb{E}\left[\left(\widehat{\mathbf{y}}_n - \mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}_n)\right)^{\mathsf{T}} \left(\widehat{\mathbf{y}}_n - \mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}_n)\right) \middle| \mathcal{Y}_{n-1} \right].$$

The one-step ahead prediction estimate of $\mathbf{y}_n$ is given by its conditional mean,

$$\begin{aligned} \mathbf{y}_{n|n-1} &= \mathbb{E}\left(\mathbf{h}\left(\mathbf{x}_n; \boldsymbol{\xi}_n\right) | \mathcal{Y}_{n-1}\right) \\ &= \int \mathbf{h}\left(\mathbf{x}_n; \boldsymbol{\xi}_n\right) p\left(\mathbf{x}_n | \mathcal{Y}_{n-1}\right) d\mathbf{x}_n. \end{aligned}$$

Its conditional variance is

$$\boldsymbol{\Sigma}^{\mathbf{yy}}_{n|n-1} = \int \left(\mathbf{h}\left(\mathbf{x}_n; \boldsymbol{\xi}_n\right) - \mathbf{y}_{n|n-1}\right) \left(\mathbf{h}\left(\mathbf{x}_n; \boldsymbol{\xi}_n\right) - \mathbf{y}_{n|n-1}\right)^{\mathsf{T}} p\left(\mathbf{x}_n | \mathcal{Y}_{n-1}\right) d\mathbf{x}_n.$$

Then, their ensemble estimates are given by

$$\widehat{\mathbf{y}}_{n|n-1} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathbf{y}}^{(k)}_{n|n-1} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}(\widehat{\mathbf{x}}^{(k)}_{n|n-1}; \boldsymbol{\xi}_n), \tag{2.13}$$

and

$$\widehat{\boldsymbol{\Sigma}}^{\mathbf{yy}}_{n|n-1} = \frac{1}{K-1} \sum_{k=1}^{K} \left(\widehat{\mathbf{y}}^{(k)}_{n|n-1} - \widehat{\mathbf{y}}_{n|n-1}\right) \left(\widehat{\mathbf{y}}^{(k)}_{n|n-1} - \widehat{\mathbf{y}}_{n|n-1}\right)^{\mathsf{T}}, \tag{2.14}$$

respectively where the kth ensemble member $\widehat{\mathbf{y}}_{n|n-1}^{(k)}$ is generated by

$$\widehat{\mathbf{y}}_{n|n-1}^{(k)} = \mathbf{h}(\widehat{\mathbf{x}}_{n|n-1}^{(k)}; \boldsymbol{\xi}_n) \quad \text{for} \quad k = 1, \ldots, K.$$

From now on, $f_{i,n-1|n-1}^{(k)}$, $f_{i,n-1}$, $h_{i,n|n-1}^{(k)}$ and $h_{i,n}$ denote the ith elements of $\mathbf{f}(\widehat{\mathbf{x}}_{n-1|n-1}^{(k)}; \boldsymbol{\xi}_n)$, $\mathbf{f}(\mathbf{x}_{n-1}; \boldsymbol{\xi}_n)$, $\mathbf{h}(\widehat{\mathbf{x}}_{n|n-1}^{(k)}; \boldsymbol{\xi}_n)$ and $\mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}_n)$ respectively; $\widehat{x}_{i,n-1|n-1}^{(k)}$, $\widehat{x}_{i,n|n-1}^{(k)}$, $\widehat{y}_{i,n|n-1}^{(k)}$, $u_{i,n}^{(k)}$ and $v_{i,n}^{(k)}$ denote the ith elements of $\widehat{\mathbf{x}}_{n-1|n-1}^{(k)}$, $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$, $\widehat{\mathbf{y}}_{n|n-1}^{(k)}$, $\mathbf{u}_n^{(k)}$ and $\mathbf{v}_n^{(k)}$ respectively; $x_{i,n-1|n-1}$, $x_{i,n|n-1}$ and $y_{i,n|n-1}$ denote the ith elements of $\mathbf{x}_{n-1|n-1}$, $\mathbf{x}_{n|n-1}$ and $\mathbf{y}_{n|n-1}$ respectively; $\Sigma_{ij,n|n-1}^{\mathbf{xx}}$, $\Sigma_{ij,n|n}^{\mathbf{xx}}$ and $\Sigma_{ij,n|n-1}^{\mathbf{yy}}$ denote the (i,j)th elements of $\boldsymbol{\Sigma}_{n|n-1}^{\mathbf{xx}}$, $\boldsymbol{\Sigma}_{n|n}^{\mathbf{xx}}$ and $\boldsymbol{\Sigma}_{n|n-1}^{\mathbf{yy}}$ respectively. Then, the following theorem is provided.

Theorem 2.1. ***Asymptotics of One-step Ahead Predicted States*** *Assume that the filtering density for the state vector* $\mathbf{x}_n$ *for* $n \geq 1$ *is*

$$p\left(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}\right) = \mathcal{N}(\mathbf{x}_{n-1}; \mathbf{x}_{n-1|n-1}, \boldsymbol{\Sigma}_{n-1|n-1}^{\mathbf{xx}}). \tag{2.15}$$

Then, the ensemble mean and variance of the state vector $\mathbf{x}_n$ *under the prediction density* $p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right)$ *can be estimated by (2.10) and (2.11) respectively. Furthermore, assume that*

(i) $\mathbb{E}\left|f_{i,n-1|n-1}^{(k)}\right| < \infty$ *for* $i = 1, \ldots, m_x, k = 1, \ldots, K$,

(ii) $\mathbb{E}\left(f_{i,n-1|n-1}^{(k)}\right) = \mathbb{E}\left(f_{i,n-1}|\mathcal{Y}_{n-1}\right)$ *for* $i = 1, \ldots, m_x, k = 1, \ldots, K$,

(iii) $\mathbb{E}\left|f_{i,n-1|n-1}^{(k)} f_{j,n-1|n-1}^{(k)}\right| < \infty, \mathbb{E}\left|f_{i,n-1|n-1}^{(k)} u_{j,n}^{(k)}\right| < \infty$ *for* $i, j = 1, \ldots, m_x, k = 1, \ldots, K$,

(iv) $\mathbb{E}\left(\widehat{x}_{i,n|n-1}^{(k)} \widehat{x}_{j,n|n-1}^{(k)}\right) = \boldsymbol{\Sigma}_{ij,n|n-1}^{\mathbf{xx}} + x_{i,n|n-1} x_{j,n|n-1}$ *for* $i, j = 1, \ldots, m_x, k = 1, \ldots, K$.

The strong consistency of $\widehat{\mathbf{x}}_{n|n-1}$ *can then be ensured, that is,*

$$\widehat{\mathbf{x}}_{n|n-1} \xrightarrow{a.s.} \mathbf{x}_{n|n-1} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1} \xrightarrow{a.s.} \mathbf{\Sigma}^{\mathbf{xx}}_{n|n-1} \textit{ when } K \to \infty,$$

Furthermore, if condition (i) is replaced by $\mathbb{E}\left|f^{(k)}_{i,n-1|n-1}\right|^2 < \infty$ *and positive definite* $\mathbf{\Sigma}^{\mathbf{xx}}_{n|n-1}$ *is assumed, the asymptotic normality of* $\widehat{\mathbf{x}}_{n|n-1}$ *is ensured as well, that is,*

$$\sqrt{K}\left(\widehat{\mathbf{x}}_{n|n-1} - \mathbf{x}_{n|n-1}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}^{\mathbf{xx}}_{n|n-1}) \textit{ when } K \to \infty.$$

Theorem 2.2. ***Asymptotics of One-step Ahead Predicted Measurements*** *Assume that the ensemble prediction mean and variance for the measurement vector* $\mathbf{y}_n$ *are estimated by (2.13) and (2.14) respectively. In addition to the assumptions of Theorem 2.1, assume that*

(v) $\mathbb{E}\left|h^{(k)}_{i,n|n-1}\right| < \infty$ *for* $i = 1, \ldots, m_y, k = 1, \ldots, K$,

(vi) $\mathbb{E}\left(h^{(k)}_{i,n|n-1}\right) = \mathbb{E}\left(h_{i,n}|\mathcal{Y}_{n-1}\right)$ *for* $i = 1, \ldots, m_y, k = 1, \ldots, K$,

(vii) $\mathbb{E}\left|h^{(k)}_{i,n|n-1}h^{(k)}_{j,n|n-1}\right| < \infty$ *for* $i, j = 1, \ldots, m_y, k = 1, \ldots, K$,

(viii) $\mathbb{E}\left(h^{(k)}_{i,n|n-1}h^{(k)}_{j,n|n-1}\right) = \mathbb{E}\left(h_{i,n}h_{j,n}|\mathcal{Y}_{n-1}\right)$ *for* $i, j = 1, \ldots, m_y, k = 1, \ldots, K$.

Their asymptotic properties are similar to those of the state vector, that is,

$$\widehat{\mathbf{y}}_{n|n-1} \xrightarrow{a.s.} \mathbf{y}_{n|n-1} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{n|n-1} \xrightarrow{a.s.} \mathbf{\Sigma}^{\mathbf{yy}}_{n|n-1} \textit{ when } K \to \infty,$$

As in the case of ensemble states, if condition (v) is replaced by $\mathbb{E}\left|h^{(k)}_{i,n|n-1}\right|^2 < \infty$ *and* $\mathbf{\Sigma}^{\mathbf{yy}}_{n|n-1}$ *is positive definite, the asymptotic normality of* $\widehat{\mathbf{y}}_{n|n-1}$ *is ensured, that is,*

$$\sqrt{K}\left(\widehat{\mathbf{y}}_{n|n-1} - \mathbf{y}_{n|n-1}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}^{\mathbf{yy}}_{n|n-1}) \textit{ when } K \to \infty.$$

Proof of Theorems 2.1 and 2.2. Since the estimators are in the forms of sample averages and all ensemble members are considered as i.i.d. within an ensemble, their consistencies can be shown easily by White (2001, Proposition 3.2).

To show the consistency of $\widehat{\mathbf{x}}_{n|n-1}$, one need to show that

$$\mathbb{E}\left|\widehat{x}^{(k)}_{i,n|n-1}\right| < \infty \quad \text{for} \quad i = 1, \ldots, m_x.$$

Now,

$$\begin{aligned}
\mathbb{E}\left|\widehat{x}^{(k)}_{i,n|n-1}\right| &= \mathbb{E}\left|f^{(k)}_{i,n-1|n-1} + u^{(k)}_{i,n}\right| \\
&\leq \mathbb{E}\left|f^{(k)}_{i,n-1|n-1}\right| + \mathbb{E}\left|u^{(k)}_{i,n}\right| \\
&< \infty,
\end{aligned}$$

for $i = 1, \ldots, m_x$. The first inequality is obtained by the triangle inequality and the second one is obtained by condition (i) above and the normality of $u^{(k)}_{i,n}$. In addition to condition (ii) above which is satisfied automatically, by Komolgorov's SLLN (Rao, 1973, p.114),

$$\widehat{\mathbf{x}}_{n|n-1} \overset{a.s.}{\to} \mathbf{x}_{n|n-1} \text{ when } K \to \infty.$$

The consistency of $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1}$ can be shown similarly because

$$\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{x}}^{(k)}_{n|n-1}\right)\left(\widehat{\mathbf{x}}^{(k)}_{n|n-1}\right)^{\mathsf{T}} - \frac{K}{K-1}\left(\widehat{\mathbf{x}}_{n|n-1}\right)\left(\widehat{\mathbf{x}}_{n|n-1}\right)^{\mathsf{T}}.$$

Then, it can be shown that

$$\frac{1}{K-1}\sum_{k=1}^{K}\widehat{x}_{i,n|n-1}^{(k)}\widehat{x}_{j,n|n-1}^{(k)} \overset{a.s.}{\rightarrow} \mathbf{\Sigma}_{ij,n|n-1}^{\mathbf{xx}} + x_{i,n|n-1}x_{j,n|n-1} \text{ when } K \rightarrow \infty$$

for $i, j = 1, \ldots, m_x$ by conditions (iii) and (iv) above. The consistency result can be implied directly because the sequence $\{\widehat{x}_{i,n|n-1}^{(k)}\widehat{x}_{j,n|n-1}^{(k)}\}_{k=1}^{K}$ is i.i.d..

The consistency result for the prediction of the measurement vector $\mathbf{y}_n$ is shown by similar method, but the conditions (v) to (viii) above are used instead.

The asymptotic normality of $\widehat{\mathbf{x}}_{n|n-1}$ and $\widehat{\mathbf{y}}_{n|n-1}$ can be proved by two approaches, namely: (1) by using the Cramér-Wold theorem (Rao, 1973, p.123) and the univariate version of Lindeberg-Lévy's central limit theorem (CLT); (2) by deriving the joint characteristic functions of $\widehat{\mathbf{x}}_{n|n-1}$ and $\widehat{\mathbf{y}}_{n|n-1}$ accordingly (Ferguson, 1996, p.26–27; White, 2001, p.114–115). Indeed, the Cramér-Wold theorem can show the multivariate Gaussianity of a random vector $\mathbf{Z}$ by showing that the linear combinations of elements of z in $\mathbf{Z}$ are Gaussian where the weights of linear combinations are normalized to a unit circle. ■

Theorems 2.1 and 2.2 imply that the prediction densities (2.5) and (2.12) can be approximated by Gaussian density. Stronger assumptions on $\mathbf{f}(\widehat{\mathbf{x}}_{n-1|n-1}^{(k)}; \boldsymbol{\xi}_n)$ and $\mathbf{h}(\widehat{\mathbf{x}}_{n|n-1}^{(k)}; \boldsymbol{\xi}_n)$ are used here because no concavity is assumed on $\mathbf{h}(\cdot)$ and $\mathbf{f}(\cdot)$. As a result, Jensen's inequality cannot be used. Next, the case of multi-step ahead prediction of measurement and state is considered.

For the q-step ahead prediction of the state vector $\mathbf{x}_n$ and $q \geq 1$, the conditional mean and variance are obtained by (2.9), (2.10) and (2.11) iteratively. Specifically,

$$
\begin{aligned}
\widehat{\mathbf{x}}^{(k)}_{n+q-1|n-1} &= \mathbf{f}(\widehat{\mathbf{x}}^{(k)}_{n+q-2|n-1}; \boldsymbol{\xi}_n), \quad k = 1, \ldots, K, \\
\widehat{\mathbf{x}}_{n+q-1|n-1} &= \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathbf{x}}^{(k)}_{n+q-1|n-1}, \\
\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n+q-1|n-1} &= \frac{1}{K-1} \sum_{k=1}^{K} \left(\widehat{\mathbf{x}}^{(k)}_{n+q-1|n-1} - \widehat{\mathbf{x}}_{n+q-1|n-1} \right) \left(\widehat{\mathbf{x}}^{(k)}_{n+q-1|n-1} - \widehat{\mathbf{x}}_{n+q-1|n-1} \right)^{\mathsf{T}}.
\end{aligned}
$$

Similarly, the q-step ahead prediction means and variances of measurement vector $\mathbf{y}_n$ under the conditional Gaussian density function for $q \geq 1$ are given by

$$
\begin{aligned}
\widehat{\mathbf{y}}^{(k)}_{n+q-1|n-1} &= \mathbf{h}(\widehat{\mathbf{x}}^{(k)}_{n+q-1|n-1}; \boldsymbol{\xi}_n), \quad k = 1, \ldots, K, \\
\widehat{\mathbf{y}}_{n+q-1|n-1} &= \frac{1}{K} \sum_{k=1}^{K} \widehat{\mathbf{y}}^{(k)}_{n+q-1|n-1}, \\
\widehat{\boldsymbol{\Sigma}}^{\mathbf{yy}}_{n+q-1|n-1} &= \frac{1}{K-1} \sum_{k=1}^{K} \left(\widehat{\mathbf{y}}^{(k)}_{n+q-1|n-1} - \widehat{\mathbf{y}}_{n+q-1|n-1} \right) \left(\widehat{\mathbf{y}}^{(k)}_{n+q-1|n-1} - \widehat{\mathbf{y}}_{n+q-1|n-1} \right)^{\mathsf{T}},
\end{aligned}
$$

respectively.

Actually, the q-step ahead prediction is preceded by the previous $(q-1)$-step ahead predictions recursively. One can expect that the prediction variances of the measurement and state variables increase with the prediction horizon because no new information of measurements is provided for updating previous predictions.

2.3.2 Filtering

Assume that the prediction density for the state vector $\mathbf{x}_n$ is

$$
p\left(\mathbf{x}_n | \mathcal{Y}_{n-1}\right) = \mathcal{N}(\mathbf{x}_n; \mathbf{x}_{n|n-1}, \boldsymbol{\Sigma}_{n|n-1}),
$$

and the conditional mean and variance of the filtered state vector $\mathbf{x}_n$ for $n > 1$ are denoted as $\mathbf{x}_{n|n}$ and $\mathbf{\Sigma}_{n|n}$ respectively, that is,

$$\mathbf{x}_{n|n} \equiv \int \mathbf{x}_n p\left(\mathbf{x}_n|\mathcal{Y}_n\right) d\mathbf{x}_n$$

and

$$\mathbf{\Sigma}_{n|n} \equiv \int \left(\mathbf{x}_n - \mathbf{x}_{n|n}\right)\left(\mathbf{x}_n - \mathbf{x}_{n|n}\right)^{\mathsf{T}} p\left(\mathbf{x}_n|\mathcal{Y}_n\right) d\mathbf{x}_n.$$

Rather than generating ensemble members of $\mathbf{x}_{n|n}$ from the posterior density $p\left(\mathbf{x}_n|\mathcal{Y}_n\right)$ directly, an estimation procedure is suggested here instead. Following Evensen (1994) and Burgers et al. (1998), $\mathbf{y}_n^{(k)}$ denotes the kth ensemble member which is constructed by the perturbation of measurement $\mathbf{y}_n$ at time n, that is,

$$\mathbf{y}_n^{(k)} = \mathbf{y}_n + \mathbf{v}_n^{(k)}, \qquad \mathbf{v}_n^{(k)} \sim N\left(\mathbf{0}, \mathbf{R}_n\right), \qquad \text{for } k = 1, \ldots, K,$$

Furthermore, the relationship between the true and the predicted state vector is

$$\widehat{\mathbf{x}}_{n|n-1}^{(k)} = \mathbf{x}_n - \widehat{\mathbf{u}}_n^{(k)} \qquad \text{for } k = 1, \ldots, K,$$

where $\widehat{\mathbf{u}}_n^{(k)}$ denotes the prediction error for the the kth ensemble member of the state vector $\mathbf{x}_n$.

The above two expressions in the conventional EnKF are essential for the derivation of the nonlinear updating equation by Goldberger-Theil's mixed estimation which was proposed by Theil and Goldberger (1961). Their original formulation considered how extraneous information could be incorporated into the

generalized least squares estimation. Indeed, their idea can be regarded as the Bayesian estimation because the extraneous statistical information and generalized least squares estimation considered in their paper were the prior information and the likelihood respectively, if Gaussian densities were assumed for the disturbance terms and regression parameters. Hence, the mixed estimator of regression parameters was the posterior mean.

Now, their idea of mixed estimation is extended to the case where the measurement equation (2.1) at time n is considered as the extraneous nonlinear statistical information and the estimated ensemble states have the conditional mean of the state vector $\mathbf{x}_n$ up to the information at time $n-1$. Hence, the filtering process in terms of ensemble members above and the one-step ahead prediction ensemble member $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ can be reformulated as

$$\begin{bmatrix} \mathbf{y}_n^{(k)} \\ \widehat{\mathbf{x}}_{n|n-1}^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}_n) \\ \mathbf{x}_n \end{bmatrix} + \begin{bmatrix} \mathbf{v}_n^{(k)} \\ -\widehat{\mathbf{u}}_n^{(k)} \end{bmatrix}. \tag{2.16}$$

Note that $\mathbf{v}_n^{(k)}$ and $\widehat{\mathbf{u}}_n^{(k)}$ are uncorrelated because $\mathbf{v}_n^{(k)}$ is uncorrelated with $\mathbf{x}_n$ and $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$. Then, the error covariance structure of (2.16) is

$$\mathrm{Var} \begin{bmatrix} \mathbf{v}_n^{(k)} \\ \widehat{\mathbf{u}}_n^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_n & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}} \end{bmatrix}.$$

By Goldberger-Theil's mixed estimation, the unknown state $\mathbf{x}_n$ can be obtained by minimizing the weighted sum of squared residuals

$$\widehat{\mathbf{x}}_{n|n}^{(k)} = \underset{\widehat{\mathbf{x}}_n^{(k)} \in \mathbb{R}^{m_x}}{\operatorname{argmin}} \, B(\widehat{\mathbf{x}}_n^{(k)}), \tag{2.17}$$

where $B(\widehat{\mathbf{x}}_n^{(k)}) \equiv \left(\mathbf{y}_n^{(k)} - \mathbf{h}(\widehat{\mathbf{x}}_n^{(k)}; \boldsymbol{\xi}_n)\right)^{\mathsf{T}} \mathbf{R}_n^{-1} \left(\mathbf{y}_n^{(k)} - \mathbf{h}(\widehat{\mathbf{x}}_n^{(k)}; \boldsymbol{\xi}_n)\right)$
$+ \left(\widehat{\mathbf{x}}_n^{(k)} - \widehat{\mathbf{x}}_{n|n-1}^{(k)}\right)^{\mathsf{T}} (\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1} \left(\widehat{\mathbf{x}}_n^{(k)} - \widehat{\mathbf{x}}_{n|n-1}^{(k)}\right)$, $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ denotes the kth ensemble member of $\widehat{\mathbf{x}}_{n|n-1}$ in the previous prediction step. The formulation of (2.17) is sensible since the minimization function can be interpreted as the maximization of the posterior density in (2.6):

$$
\begin{aligned}
& \underset{\mathbf{x}_n \in \mathbb{R}^{m_x}}{\operatorname{argmax}}\, p\left(\mathbf{x}_n | \mathcal{Y}_n\right) \\
= & \underset{\mathbf{x}_n \in \mathbb{R}^{m_x}}{\operatorname{argmax}}\, p\left(\mathbf{y}_n | \mathbf{x}_n\right) p\left(\mathbf{x}_n | \mathcal{Y}_{n-1}\right) \\
= & \underset{\mathbf{x}_n \in \mathbb{R}^{m_x}}{\operatorname{argmax}} \left[\log p\left(\mathbf{y}_n | \mathbf{x}_n\right) + \log p\left(\mathbf{x}_n | \mathcal{Y}_{n-1}\right)\right] \\
= & \underset{\widehat{\mathbf{x}}_n^{(k)} \in \mathbb{R}^{m_x}}{\operatorname{argmin}} \Big[\left(\mathbf{y}_n^{(k)} - \mathbf{h}(\widehat{\mathbf{x}}_n^{(k)}; \boldsymbol{\xi}_n)\right)^{\mathsf{T}} \mathbf{R}_n^{-1} \left(\mathbf{y}_n^{(k)} - \mathbf{h}(\widehat{\mathbf{x}}_n^{(k)}; \boldsymbol{\xi}_n)\right) \\
& \qquad + \left(\widehat{\mathbf{x}}_n^{(k)} - \widehat{\mathbf{x}}_{n|n-1}^{(k)}\right)^{\mathsf{T}} (\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1} \left(\widehat{\mathbf{x}}_n^{(k)} - \widehat{\mathbf{x}}_{n|n-1}^{(k)}\right) \Big],
\end{aligned}
$$

where the conditional mean and variance are replaced by the ensemble estimates in previous prediction step. Note that the first equality is obtained by the Bayes' Theorem and the denominator term $p\left(\mathbf{y}_n | \mathcal{Y}_{n-1}\right)$ does not depend on $\widehat{\mathbf{x}}_n$. The second equality is trivial as a monotonic transformation is applied. The third equality is trivial under Gaussian density.

Typically, the optimal $\widehat{\mathbf{x}}_{n|n}^{(k)}$ of (2.17) is obtained by solving the first order condition, that is,

$$
\left.\frac{\partial \mathbf{h}(\widehat{\mathbf{x}}_n^{(k)}; \boldsymbol{\xi}_n)^{\mathsf{T}}}{\partial \widehat{\mathbf{x}}_n^{(k)}}\right|_{\widehat{\mathbf{x}}_n^{(k)} = \widehat{\mathbf{x}}_{n|n}^{(k)}} \mathbf{R}_n^{-1} \left(\mathbf{y}_n^{(k)} - \mathbf{h}(\widehat{\mathbf{x}}_{n|n}^{(k)}; \boldsymbol{\xi}_n)\right) + (\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1} \left(\widehat{\mathbf{x}}_{n|n}^{(k)} - \widehat{\mathbf{x}}_{n|n-1}^{(k)}\right) = \mathbf{0}. \tag{2.18}
$$

Now, we assume that the Jacobian matrix of the left hand side of (2.18) with respect to $\mathbf{y}_n^{(k)}$ and $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ is invertible, then, by the Implicit Function Theorem, there exists a function $\mathbf{g} : \mathbb{R}^{m_y} \times \mathbb{R}^{m_x} \to \mathbb{R}^{m_x}$ such that

$$\widehat{\mathbf{x}}_{n|n}^{(k)} = \mathbf{g}(\mathbf{y}_n^{(k)}, \widehat{\mathbf{x}}_{n|n-1}^{(k)}). \tag{2.19}$$

Obviously, the updating equation (2.19) for the predicted ensemble state $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ is more general than the one suggested by Evensen (1994) in which the conventional Kalman filter updating equation was used. Although the ensemble predicted state $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ is updated by the perturbed measurement $\mathbf{y}_n^{(k)}$, no Kalman filter gain matrix is computed. Indeed, the explicit formulation of the updating equation can be found in a few special cases only. This nonlinear updating equation is related to the single-stage iteration filter by Wishner et al. (1969) where the updating process was performed by an iterative Newton-Raphson algorithm. Since the Goldberger-Theil's mixed estimation has been incorporated into EnKF, this new filter is called the ensemble Goldberger-Theil Kalman filter (EnGTKF).

For each ensemble member $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$, an estimate $\widehat{\mathbf{x}}_{n|n}^{(k)}$ is obtained from (2.19). In most cases, the analytical solution of $\widehat{\mathbf{x}}_{n|n}^{(k)}$ cannot be found easily because the perturbed measurement $\mathbf{y}_n^{(k)}$ is nonlinear in the state $\widehat{\mathbf{x}}_{n|n}^{(k)}$. This in turn results in a suggestion of a hybrid optimization of $\widehat{\mathbf{x}}_{n|n}^{(k)}$ which consists of two parts: (1) select the initial value for filtered state vector by stochastic search algorithm (Spall, 2003, p.38) and (2) improve the estimation of the filtered state vector further by local optimization procedure, for example Quasi-Newton. The procedure is described below:

1. Simulate an initial value for the state vector $\widehat{\mathbf{x}}_n^{(k)}$ in (2.17) from a specified

distribution, for example, uniform distribution or Gaussian distribution and denote its value as $\widehat{\mathbf{x}}_n^{(k)}(0)$. Set $j = 0$. Compute the objective function $B(\widehat{\mathbf{x}}_n^{(k)}(0))$.

2. At iteration j $(j \geq 0)$, generate a new value for the state vector $\widehat{\mathbf{x}}_n^{(k)}$ according to the specified distribution. If $B(\widehat{\mathbf{x}}_n^{(k)}(j+1)) < B(\widehat{\mathbf{x}}_n^{(k)}(j))$, then $\widehat{\mathbf{x}}_{n|n}^{(k)} = \widehat{\mathbf{x}}_n^{(k)}(j+1)$. Otherwise, $\widehat{\mathbf{x}}_{n|n}^{(k)} = \widehat{\mathbf{x}}_n^{(k)}(j)$.

3. Repeat Step 2 if the maximum number of iterations is not exceeded or the recent iterations of $\widehat{\mathbf{x}}_n^{(k)}(j)$ yield an improvement in the estimate $\widehat{\mathbf{x}}_{n|n}^{(k)}$.

4. Once $\widehat{\mathbf{x}}_{n|n}^{(k)}$ is selected by stochastic search algorithm, it is then used as the initial value in the Quasi-Newton method.

The convergence result of the stochastic search optimization procedure above was shown with an enormous number of iterations and the existence of a unique solution (Spall, 2003, Theorem 2.1). However, it is infeasible to use a huge number of iterations to estimate the filtered state. The local optimization procedure, for example Quasi-Newton method, is used to improve the estimated filtered state further.

Over the K ensemble members, the filtered mean and variance of $\widehat{\mathbf{x}}_{n|n}^{(k)}$ are estimated by

$$\widehat{\mathbf{x}}_{n|n} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{x}}_{n|n}^{(k)}, \tag{2.20}$$

$$\text{and} \quad \widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{x}}_{n|n}^{(k)} - \widehat{\mathbf{x}}_{n|n}\right)\left(\widehat{\mathbf{x}}_{n|n}^{(k)} - \widehat{\mathbf{x}}_{n|n}\right)^{\mathsf{T}}, \tag{2.21}$$

respectively. The asymptotic properties of (2.20) and (2.21) are given by the following theorem.

Theorem 2.3. ***Nonlinear Filtering*** *Given that the prediction density for the state vector* $\mathbf{x}_n$ *is*

$$p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) = \mathcal{N}\left(\mathbf{x}_n; \mathbf{x}_{n|n-1}, \mathbf{\Sigma}_{n|n-1}\right),$$

and assume that

(i) $\mathbb{E}\left|\widehat{x}_{i,n|n}^{(k)}\right| < \infty$, *for* $i = 1, \ldots, m_x, k = 1, \ldots, K$,

(ii) $\mathbb{E}\left(\widehat{x}_{i,n|n}^{(k)}\right) = \mathbb{E}\left(x_{i,n}|\mathcal{Y}_n\right)$, *for* $i = 1, \ldots, m_x, k = 1, \ldots, K$,

(iii) $\mathbb{E}\left|\widehat{x}_{i,n|n}^{(k)}\widehat{x}_{j,n|n}^{(k)}\right| < \infty$ *for* $i, j = 1, \ldots, m_x, k = 1, \ldots, K$,

(iv) $\mathbb{E}\left(\widehat{x}_{i,n|n}^{(k)}\widehat{x}_{j,n|n}^{(k)}\right) = \mathbb{E}\left(x_{i,n}x_{j,n}|\mathcal{Y}_n\right)$ *for* $i, j = 1, \ldots, m_x, k = 1, \ldots, K$.

Then,

$$\widehat{\mathbf{x}}_{n|n} \xrightarrow{a.s.} \mathbf{x}_{n|n} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}} \xrightarrow{a.s.} \mathbf{\Sigma}_{n|n}^{\mathbf{xx}} \textit{ when } K \to \infty.$$

If condition (i) is replaced by $\mathbb{E}\left|\widehat{x}_{i,n|n}^{(k)}\right|^2 < \infty$ *and* $\mathbf{\Sigma}_{n|n}^{\mathbf{xx}}$ *is positive definite, the asymptotic normality is also ensured, that is,*

$$\sqrt{K}\left(\widehat{\mathbf{x}}_{n|n} - \mathbf{x}_{n|n}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Sigma}_{n|n}^{\mathbf{xx}}\right) \textit{ when } K \to \infty.$$

Proof. The asymptotic properties of $\widehat{\mathbf{x}}_{n|n}$ and $\widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}}$ can be derived from the same reference as in Theorem 2.1. ■

This theorem has a similar implication as in Theorem 2.1 and the filtering density (2.6) can be approximated by the Gaussian density. The combined results of Theorems 2.1 and 2.3 imply that the density functions of the state vector $\mathbf{x}_n$ in EnGTKF can be approximated by the Gaussian density recursively and this also holds for Evensen's EnKF.

The method of nonlinear state updating has been described already. This

proposed method can also be reduced to the linear updating equation of EnKF and the asymptotic properties of ensemble filtered state are summarized by the following corollary.

Corollary 2.4. *If* $\mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}_n)$ *is linear in* $\mathbf{x}_n$*, that is,* $\mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}_n)=\mathbf{h}_n\mathbf{x}_n$*, the ensemble members* $\widehat{\mathbf{x}}_{n|n}^{(k)}$ *can be derived from the case of EnKF with linear measurement, that is,*

$$\widehat{\mathbf{x}}_{n|n}^{(k)}=\widehat{\mathbf{x}}_{n|n-1}^{(k)}+\mathbf{K}_n(\mathbf{y}_n^{(k)}-\mathbf{h}_n\widehat{\mathbf{x}}_{n|n-1}^{(k)}), \tag{2.22}$$

where $\mathbf{K}_n=(\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{h}_n+(\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1})^{-1}\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}$.

Furthermore, its asymptotic properties in Theorem 2.3 are maintained.

Proof. When $\mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}_n)=\mathbf{h}_n\mathbf{x}_n$, the optimization problem becomes

$$\begin{aligned}\widehat{\mathbf{x}}_{n|n}^{(k)} &= \underset{\widehat{\mathbf{x}}_n\in\mathbb{R}^{m_x}}{\operatorname{argmin}}\Big[\left(\mathbf{y}_n^{(k)}-\mathbf{h}_n\widehat{\mathbf{x}}_n\right)^{\mathsf{T}}\mathbf{R}_n^{-1}\left(\mathbf{y}_n^{(k)}-\mathbf{h}_n\widehat{\mathbf{x}}_n\right)\\ &\quad+\left(\widehat{\mathbf{x}}_{n|n-1}^{(k)}-\widehat{\mathbf{x}}_n\right)^{\mathsf{T}}(\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1}\left(\widehat{\mathbf{x}}_{n|n-1}^{(k)}-\widehat{\mathbf{x}}_n\right)\Big].\end{aligned}$$

From (2.18), we have

$$\begin{aligned}\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\left(\mathbf{y}_n^{(k)}-\mathbf{h}_n\widehat{\mathbf{x}}_{n|n}^{(k)}\right)+(\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1}\left(\widehat{\mathbf{x}}_{n|n-1}-\widehat{\mathbf{x}}_{n|n}^{(k)}\right)&=\mathbf{0}\\ \mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{y}_n^{(k)}+(\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1}\widehat{\mathbf{x}}_{n|n-1}^{(k)}&=\left(\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{h}_n+(\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1}\right)\widehat{\mathbf{x}}_{n|n}^{(k)}.\end{aligned}$$

Hence,

$$\widehat{\mathbf{x}}_{n|n}^{(k)} = \left(\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{h}_n + (\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1}\right)^{-1}\left(\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{y}_n^{(k)} + (\widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1}\widehat{\mathbf{x}}_{n|n-1}^{(k)}\right).$$

By using the results in Meditch (1969, p.190) or Lemma in Diderrich (1985), letting $\mathbf{P}_0 = \widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{xx}}$, $\mathbf{P}_1 = \left(\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{h}_n + \mathbf{P}_0^{-1}\right)^{-1}$ and $\mathbf{K}_n = \mathbf{P}_1\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}$. Then,

$$\begin{aligned}
\widehat{\mathbf{x}}_{n|n}^{(k)} &= \mathbf{P}_1(\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{y}_n^{(k)} + \mathbf{P}_0^{-1}\widehat{\mathbf{x}}_{n|n-1}^{(k)}) \\
&= \mathbf{K}_n\mathbf{y}_n^{(k)} + \mathbf{P}_1\mathbf{P}_0^{-1}\widehat{\mathbf{x}}_{n|n-1}^{(k)} \\
&= \mathbf{K}_n\mathbf{y}_n^{(k)} + (\mathbf{I}_{m_x} - \mathbf{K}_n\mathbf{h}_n)\,\widehat{\mathbf{x}}_{n|n-1}^{(k)} \\
&= \widehat{\mathbf{x}}_{n|n-1}^{(k)} + \mathbf{K}_n(\mathbf{y}_n^{(k)} - \mathbf{h}_n\widehat{\mathbf{x}}_{n|n-1}^{(k)}),
\end{aligned}$$

where $\mathbf{P}_1\mathbf{P}_0^{-1} = \mathbf{I}_{m_x} - \mathbf{K}_n\mathbf{h}_n$ in the third equality is shown below:

$$\begin{aligned}
\mathbf{P}_1 &= \left(\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{h}_n + \mathbf{P}_0^{-1}\right)^{-1} \\
\mathbf{P}_1^{-1} &= \mathbf{P}_0^{-1} + \mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{h}_n \\
\mathbf{I}_{m_x} &= \mathbf{P}_1\mathbf{P}_0^{-1} + \mathbf{P}_1\mathbf{h}_n^{\mathsf{T}}\mathbf{R}_n^{-1}\mathbf{h}_n,
\end{aligned}$$

and the result is shown.

Alternatively, the same result can be achieved by considering $\widehat{\mathbf{x}}_{n|n}^{(k)}$ as a generalized least squares estimator. Equation (2.16) in this linear case can be rewritten as

$$\begin{bmatrix} \mathbf{y}_n^{(k)} \\ \widehat{\mathbf{x}}_{n|n-1}^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_n \\ \mathbf{I}_{m_x} \end{bmatrix} \mathbf{x}_n + \begin{bmatrix} \mathbf{v}_n^{(k)} \\ \widehat{\mathbf{u}}_n^{(k)} \end{bmatrix},$$

where $\mathbb{E}\begin{bmatrix} \mathbf{v}_n^{(k)} \\ \widehat{\mathbf{u}}_n^{(k)} \end{bmatrix} = \mathbf{0}$ and $\mathrm{Var}\begin{bmatrix} \mathbf{v}_n^{(k)} \\ \widehat{\mathbf{u}}_n^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_n & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}} \end{bmatrix}$. Then, by generalized least squares, an ensemble member of the estimated filtered state is

$$\begin{aligned} \widehat{\mathbf{x}}_{n|n}^{(k)} &= \left(\begin{bmatrix} \mathbf{h}_n & \mathbf{I}_{m_x} \end{bmatrix} \begin{bmatrix} \mathbf{R}_n & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_n \\ \mathbf{I}_{m_x} \end{bmatrix} \right)^{-1} \\ &\quad \times \left(\begin{bmatrix} \mathbf{h}_n & \mathbf{I}_{m_x} \end{bmatrix} \begin{bmatrix} \mathbf{R}_n & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_n^{(k)} \\ \widehat{\mathbf{x}}_{n|n-1}^{(k)} \end{bmatrix} \right) \\ &= \left(\mathbf{h}_n \mathbf{R}_n^{-1} \mathbf{h}_n + (\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1} \right)^{-1} \left(\mathbf{h}_n \mathbf{R}_n^{-1} \mathbf{y}_n^{(k)} + (\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1} \widehat{\mathbf{x}}_{n|n-1}^{(k)} \right), \end{aligned}$$

which indeed is the same as the one derived previously. ■

An analogous proof was given in Harvey (1981, p.108-109), Diderrich (1985) and Tanizaki (1996) for classical Kalman filter.

From Theorems 2.1, 2.2 and 2.3, one can construct the confidence regions for the prediction state vector $\mathbf{x}_{n|n-1}$, the measurement vector $\mathbf{y}_{n|n-1}$ and the filtered state vector $\mathbf{x}_{n|n}$ easily due to its recursive Gaussian properties. Specifically, their $(1-\alpha)\%$ confidence regions can be obtained by

$$\begin{aligned} \mathcal{S}_{n|n-1}^{\mathbf{y}} &= \left\{ \mathbf{y} : \left(\mathbf{y} - \widehat{\mathbf{y}}_{n|n-1} \right)^{\mathsf{T}} (\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}})^{-1} \left(\mathbf{y} - \widehat{\mathbf{y}}_{n|n-1} \right) \leq z_{1-\alpha/2} \right\}, \\ \mathcal{S}_{n|n-1}^{\mathbf{x}} &= \left\{ \mathbf{x} : \left(\mathbf{x} - \widehat{\mathbf{x}}_{n|n-1} \right)^{\mathsf{T}} (\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}})^{-1} \left(\mathbf{x} - \widehat{\mathbf{x}}_{n|n-1} \right) \leq z_{1-\alpha/2} \right\}, \end{aligned}$$

$$\text{and} \qquad \mathcal{S}^{\mathbf{x}}_{n|n} \;=\; \left\{\mathbf{x} : \left(\mathbf{x} - \widehat{\mathbf{x}}_{n|n}\right)^{\mathsf{T}} \left(\widehat{\boldsymbol{\Sigma}}^{\mathbf{xx}}_{n|n}\right)^{-1} \left(\mathbf{x} - \widehat{\mathbf{x}}_{n|n}\right) \leq z_{1-\alpha/2}\right\},$$

respectively where $z_{1-\alpha/2}$ is the $100\,(1-\alpha/2)$th quantile of the Gaussian density.

Recursive state estimation procedures are summarized below:

1. Generate the initial ensemble states. When $n = 0$, given the values of $\mathbf{x}_{0|0}$ and $\boldsymbol{\Sigma}_{0|0}$, draw K ensembles from the density function $\mathcal{N}\left(\mathbf{x}_0; \mathbf{x}_{0|0}, \boldsymbol{\Sigma}_{0|0}\right)$ and denote them as $\{\widehat{\mathbf{x}}_0^{(k)}\}_{k=1}^{K}$.

2. Predict the ensemble states. When $n = 1$, draw an ensemble of K members from the density function $\mathcal{N}\left(\mathbf{u}_n; \mathbf{0}, \mathbf{Q}_n\right)$ and denote them as $\{\mathbf{u}_n^{(k)}\}_{k=1}^{K}$. Then, generate ensemble predictions with stochastic forces by

$$\widehat{\mathbf{x}}^{(k)}_{n|n-1} = \mathbf{f}(\widehat{\mathbf{x}}^{(k)}_{n-1|n-1}, \boldsymbol{\xi}_n) + \mathbf{u}_n^{(k)} \quad \text{for } k = 1, \ldots, K.$$

3. Generate the ensemble measurements. When $n = 1$, draw an ensemble of K members from the density function $\mathcal{N}\left(\mathbf{v}_n; \mathbf{0}, \mathbf{R}_n\right)$ and denote them as $\{\mathbf{v}_n^{(k)}\}_{k=1}^{K}$. Then, generate ensemble measurements by

$$\mathbf{y}_n^{(k)} = \mathbf{y}_n + \mathbf{v}_n^{(k)} \quad \text{for } k = 1, \ldots, K.$$

4. Update the ensemble states. The kth ensemble member of the state vector $\mathbf{x}_n$ is updated by minimizing (2.17).

5. Repeat Steps 2, 3 and 4 for $n = 2, \ldots, N$.

6. For $n = 1, \ldots, N$, the prediction and filtered mean and variance of the state vector $\mathbf{x}_n$ are estimated by (2.10), (2.11), (2.20) and (2.21) respectively.

Also, the one-step prediction mean and variance of the measurement vector $\mathbf{y}_n$ are given by (2.13) and (2.14) respectively.

2.3.3 Smoothing

After the derivation of prediction and filtered states which progress forward in time, a fixed-interval smoother for EnGTKF is derived in this section. The smoothing density for the state vector $\mathbf{x}_n$ is

$$\begin{aligned}
p\left(\mathbf{x}_{n-1}|\mathcal{Y}_N\right) &= \int p\left(\mathbf{x}_n, \mathbf{x}_{n-1}|\mathcal{Y}_N\right) d\mathbf{x}_n \\
&= \int p\left(\mathbf{x}_n|\mathcal{Y}_N\right) p\left(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathcal{Y}_N\right) d\mathbf{x}_n \\
&= \int p\left(\mathbf{x}_n|\mathcal{Y}_N\right) p\left(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathcal{Y}_{n-1}\right) d\mathbf{x}_n \\
&= \int p\left(\mathbf{x}_n|\mathcal{Y}_N\right) \frac{p\left(\mathbf{x}_{n-1}, \mathbf{x}_n|\mathcal{Y}_{n-1}\right)}{p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right)} d\mathbf{x}_n \\
&= \int p\left(\mathbf{x}_n|\mathcal{Y}_N\right) \frac{p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) p\left(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}\right)}{p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right)} d\mathbf{x}_n \\
&= p\left(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}\right) \int \frac{p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) p\left(\mathbf{x}_n|\mathcal{Y}_N\right)}{p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right)} d\mathbf{x}_n,
\end{aligned}$$

for $n = N, N-1, \ldots, 2$. Note that Assumptions 2.1(i) and (ii) is used to show the third equality. Specifically,

$$\begin{aligned}
& p\left(\mathbf{x}_{n-1}|\mathbf{x}_n, \mathcal{Y}_N\right) \\
= & \frac{p(\mathbf{x}_{n-1}, \mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_N)}{p(\mathbf{x}_n, \mathbf{y}_1, \ldots, \mathbf{y}_N)} \\
= & \frac{p(\mathbf{y}_N|\mathcal{Y}_{N-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) \ldots p(\mathbf{y}_{n+1}|\mathcal{Y}_n, \mathbf{x}_n, \mathbf{x}_{n-1}) p(\mathbf{y}_n|\mathcal{Y}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) p(\mathcal{Y}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1})}{p(\mathbf{y}_N|\mathcal{Y}_{N-1}, \mathbf{x}_n) \ldots p(\mathbf{y}_{n+1}|\mathcal{Y}_n, \mathbf{x}_n) p(\mathbf{y}_n|\mathcal{Y}_{n-1}, \mathbf{x}_n) p(\mathcal{Y}_{n-1}, \mathbf{x}_n)} \\
= & \frac{p(\mathbf{y}_N|\mathcal{Y}_{N-1}) \ldots p(\mathbf{y}_{n+1}|\mathcal{Y}_n) p(\mathbf{y}_n|\mathcal{Y}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1}) p(\mathcal{Y}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n-1})}{p(\mathbf{y}_N|\mathcal{Y}_{N-1}) \ldots p(\mathbf{y}_{n+1}|\mathcal{Y}_n) p(\mathbf{y}_n|\mathcal{Y}_{n-1}, \mathbf{x}_n) p(\mathcal{Y}_{n-1}, \mathbf{x}_n)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{p(\mathbf{y}_n|\mathcal{Y}_{n-1},\mathbf{x}_n,\mathbf{x}_{n-1})p(\mathcal{Y}_{n-1},\mathbf{x}_n,\mathbf{x}_{n-1})}{p(\mathbf{y}_n|\mathcal{Y}_{n-1},\mathbf{x}_n)p(\mathcal{Y}_{n-1},\mathbf{x}_n)} \\
&= \frac{p(\mathbf{y}_n|\mathbf{x}_n)p(\mathcal{Y}_{n-1},\mathbf{x}_n,\mathbf{x}_{n-1})}{p(\mathbf{y}_n|\mathbf{x}_n)p(\mathcal{Y}_{n-1},\mathbf{x}_n)} \\
&= \frac{p(\mathcal{Y}_{n-1},\mathbf{x}_n,\mathbf{x}_{n-1})}{p(\mathcal{Y}_{n-1},\mathbf{x}_n)} \\
&= p(\mathbf{x}_{n-1}|\mathbf{x}_n,\mathcal{Y}_{n-1}).
\end{aligned}
$$

The posterior density $p\left(\mathbf{x}_{n-1}|\mathcal{Y}_N\right)$ for the state vector $\mathbf{x}_n$ follows from Harvey (1989) and Kitagawa (1987). Although the smoothing density can be derived by backward recursion in this way, both the prediction density and the filtering density of the state vector need to be stored. Typically, the smoothed mean and variance of $\mathbf{x}_{n-1}$ can be estimated by the Bayesian method. Nevertheless, following Tanizaki (1996, p.217–220), the estimates can be derived by the MMSE criterion.

$$
\mathbf{x}_{n-1|N} = \underset{\widehat{\mathbf{x}}_n \in \mathbb{R}^{m_x}}{\operatorname{argmin}} \mathbb{E}\left[\left(\widehat{\mathbf{x}}_{n-1} - \mathbf{x}_{n-1}\right)^{\mathsf{T}} \left(\widehat{\mathbf{x}}_{n-1} - \mathbf{x}_{n-1}\right) \middle| \mathcal{Y}_N \right],
$$

for $n = N, N-1, \ldots, 2$.

The optimal value should be its conditional mean which is given by

$$
\begin{aligned}
\mathbf{x}_{n-1|N} &= \mathbb{E}\left(\mathbf{x}_{n-1}|\mathcal{Y}_N\right) \\
&= \int \mathbf{x}_{n-1} p\left(\mathbf{x}_{n-1}|\mathcal{Y}_N\right) d\mathbf{x}_{n-1} \\
&= \iint \mathbf{x}_{n-1} p\left(\mathbf{x}_{n-1}, \mathcal{X}_{N,-(n-1)}|\mathcal{Y}_N\right) d\mathcal{X}_{N,-(n-1)} d\mathbf{x}_{n-1} \\
&= \int \mathbf{x}_{n-1} p\left(\mathcal{X}_N|\mathcal{Y}_N\right) d\mathcal{X}_N \\
&= \frac{\int \mathbf{x}_{n-1} p\left(\mathcal{Y}_N|\mathcal{X}_N\right) p\left(\mathcal{X}_N\right) d\mathcal{X}_N}{\int p\left(\mathcal{Y}_N|\mathcal{X}_N\right) p\left(\mathcal{X}_N\right) d\mathcal{X}_N}.
\end{aligned} \tag{2.23}
$$

for $n = N, N-1, \ldots, 2$, where $\mathcal{X}_{N,-(n-1)} \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_{n-2}, \mathbf{x}_n, \ldots \mathbf{x}_N\}$ indicates the set of all state vectors excluding the one at time $n-1$.

From Assumptions 2.1(i) and (ii),

$$\begin{aligned} p\left(\mathcal{X}_N\right) &= p\left(\mathbf{x}_0\right) \prod_{n=1}^{N} p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right), \\ \text{and} \qquad p\left(\mathcal{Y}_N|\mathcal{X}_N\right) &= \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{x}_n\right). \end{aligned}$$

Then, (2.23) becomes

$$\mathbf{x}_{n-1|N} = \frac{\int \mathbf{x}_{n-1} \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{x}_n\right) p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) d\mathcal{X}_N}{\int \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{x}_n\right) p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) d\mathcal{X}_N}, \tag{2.24}$$

for $n = N, N-1, \ldots, 2$. Similarly, the conditional variance can be obtained by

$$\mathbf{\Sigma}^{\mathbf{xx}}_{n-1|N} = \frac{\int \left(\mathbf{x}_{n-1} - \mathbf{x}_{n-1|N}\right)\left(\mathbf{x}_{n-1} - \mathbf{x}_{n-1|N}\right)^{\mathsf{T}} \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{x}_n\right) p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) d\mathcal{X}_N}{\int \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{x}_n\right) p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) d\mathcal{X}_N},$$

for $n = N, N-1, \ldots, 2$.

From (2.24), the computation of the smoothed state vector $\mathbf{x}_{n-1|N}$ requires (i) the ensemble members $\{\widehat{\mathbf{x}}^{(k)}_{n|n-1}\}_{k=1}^{K}$ from the prediction density $p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)$ and (ii) the density of $\mathbf{y}_n$ conditional on $\mathbf{x}_n$ $p(\mathbf{y}_n|\mathbf{x}_n = \widehat{\mathbf{x}}^{(k)}_{n|n-1})$ for $k = 1, \ldots, K$. The ensemble mean and variance of the smoothed state vector $\mathbf{x}_{n-1|N}$ are estimated by

$$\widehat{\mathbf{x}}_{n-1|N} = \frac{\sum_{k=1}^{K} \widehat{w}^{(k)}_N \widehat{\mathbf{x}}^{(k)}_{n|n-1}}{\sum_{k=1}^{K} \widehat{w}^{(k)}_N}, \tag{2.25}$$

and

$$\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|N} = \frac{\sum_{k=1}^{K} \widehat{w}_N^{(k)} \left(\widehat{\mathbf{x}}^{(k)}_{n|n-1} - \widehat{\mathbf{x}}_{n-1|N}\right)\left(\widehat{\mathbf{x}}^{(k)}_{n|n-1} - \widehat{\mathbf{x}}_{n-1|N}\right)^{\mathsf{T}}}{\sum_{k=1}^{K} \widehat{w}_N^{(k)}}, \tag{2.26}$$

respectively for $n = N, N-1, \ldots, 2$. Indeed, a more accurate estimator of variance can be obtained by

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|N} &= \frac{\sum_{k=1}^{K} \widehat{w}_N^{(k)}}{(\sum_{k=1}^{K} \widehat{w}_N^{(k)})^2 - \sum_{k=1}^{K} (\widehat{w}_N^{(k)})^2} \\ &\quad \times \sum_{k=1}^{K} \widehat{w}_N^{(k)} \left(\widehat{\mathbf{x}}^{(k)}_{n|n-1} - \widehat{\mathbf{x}}_{n-1|N}\right)\left(\widehat{\mathbf{x}}^{(k)}_{n|n-1} - \widehat{\mathbf{x}}_{n-1|N}\right)^{\mathsf{T}}.\end{aligned}$$

The smoothing weight $\widehat{w}_N^{(k)}$ is computed by the following recursive relation:

$$\widehat{w}_n^{(k)} = \mathcal{N}(\widehat{\mathbf{y}}^{(k)}_{n|n-1}; \widehat{\mathbf{y}}_{n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{n|n-1}) \widehat{w}_{n-1}^{(k)} \quad \text{for } n = 1, \ldots, N, \tag{2.27}$$

and $\widehat{w}_0^{(k)}$ is initialized with $\widehat{w}_0^{(k)} = K^{-1}$ for $k = 1, \ldots, K$. Indeed, the initialization of weight $\widehat{w}_0^{(k)}$ is somewhat arbitrary because

$$\begin{aligned}\widehat{\mathbf{x}}_{n-1|N} &= \frac{\sum_{k=1}^{K} \prod_{n=1}^{N} \mathcal{N}(\widehat{\mathbf{y}}^{(k)}_{n|n-1}; \widehat{\mathbf{y}}_{n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{n|n-1}) \widehat{w}_0^{(k)} \widehat{\mathbf{x}}^{(k)}_{n|n-1}}{\sum_{k=1}^{K} \prod_{n=1}^{N} \mathcal{N}(\widehat{\mathbf{y}}^{(k)}_{n|n-1}; \widehat{\mathbf{y}}_{n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{n|n-1}) \widehat{w}_0^{(k)}} \\ &= \frac{\sum_{k=1}^{K} \prod_{n=1}^{N} \mathcal{N}(\widehat{\mathbf{y}}^{(k)}_{n|n-1}; \widehat{\mathbf{y}}_{n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{n|n-1}) \widehat{\mathbf{x}}^{(k)}_{n|n-1}}{\sum_{k=1}^{K} \prod_{n=1}^{N} \mathcal{N}(\widehat{\mathbf{y}}^{(k)}_{n|n-1}; \widehat{\mathbf{y}}_{n|n-1}, \widehat{\mathbf{\Sigma}}^{\mathbf{yy}}_{n|n-1})},\end{aligned}$$

is independent of $\widehat{w}_0^{(k)}$ for fixed $\widehat{w}_0^{(k)}$. Similar property can also applied to $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}}_{n-1|N}$.

On the other hand, the measurement vector $\mathbf{y}_{n-1}$ can also be smoothed, that is,

$$
\begin{aligned}
\mathbf{y}_{n-1|N} &= \mathbb{E}\left(\mathbf{y}_{n-1}|\mathcal{Y}_N\right) \\
&= \mathbb{E}\left(\mathbf{h}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_{n-1}\right)|\mathcal{Y}_N\right) \\
&= \int \mathbf{h}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_{n-1}\right) p\left(\mathbf{x}_{n-1}|\mathcal{Y}_N\right) d\mathbf{x}_{n-1} \\
&= \iint \mathbf{h}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_{n-1}\right) p\left(\mathbf{x}_{n-1},\mathcal{X}_{N,-(n-1)}|\mathcal{Y}_N\right) d\mathcal{X}_{N,-(n-1)} d\mathbf{x}_{n-1} \\
&= \int \mathbf{h}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_{n-1}\right) p\left(\mathcal{X}_N|\mathcal{Y}_N\right) d\mathcal{X}_N \\
&= \frac{\int \mathbf{h}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_{n-1}\right) p\left(\mathcal{Y}_N|\mathcal{X}_N\right) p\left(\mathcal{X}_N\right) d\mathcal{X}_N}{\int p\left(\mathcal{Y}_N|\mathcal{X}_N\right) p\left(\mathcal{X}_N\right) d\mathcal{X}_N},
\end{aligned}
$$

for $n = N, N-1, \ldots, 2$.

By using the same smoothing method as for the state vector $\mathbf{x}_{n-1}$, the ensemble mean and variance of the smoothed measurement vector $\mathbf{y}_{n-1|N}$ are estimated recursively by

$$
\widehat{\mathbf{y}}_{n-1|N} = \frac{\sum_{k=1}^{K} \widehat{w}_N^{(k)} \widehat{\mathbf{y}}_{n|n-1}^{(k)}}{\sum_{k=1}^{K} \widehat{w}_N^{(k)}}, \tag{2.28}
$$

and

$$
\widehat{\boldsymbol{\Sigma}}_{n-1|N}^{\mathbf{yy}} = \frac{\sum_{k=1}^{K} \widehat{w}_N^{(k)} \left(\widehat{\mathbf{y}}_{n|n-1}^{(k)} - \widehat{\mathbf{y}}_{n-1|N}\right)\left(\widehat{\mathbf{y}}_{n|n-1}^{(k)} - \widehat{\mathbf{y}}_{n-1|N}\right)^{\mathsf{T}}}{\sum_{k=1}^{K} \widehat{w}_N^{(k)}}, \tag{2.29}
$$

for $n = N, N-1, \ldots, 2$.

The estimation of smoothed states and measurements by (2.25), (2.26), (2.28) and (2.29) is called the ensemble Goldberger-Theil Kalman smoother (EnGTKS). It seems that the derived smoothers used information of the prediction density

$p(\mathbf{y}_n|\mathcal{Y}_{n-1})$ only. Indeed, information from the filtering density is also used in the recursive derivation of the prediction density. However, only the ensemble members of $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ and $\widehat{\mathbf{y}}_{n|n-1}^{(k)}$ are required and the computation time of the smoothed estimates is reduced. Indeed, this smoother is analogous to those by van Leeuwen and Evensen (1996) and Evensen and van Leeuwen (2000) in which both of them were derived by the representer method (Bennett, 2004, p.19). Unlike the EnGTKF, both of their smoothers have used linear updating equations during the filtering stage. The asymptotic properties of smoothed mean and variance are given by the following theorem.

Theorem 2.5. ***Fixed-interval Smoothing*** *Assume that the one-step ahead prediction densities of* $\mathbf{x}_n$ *and* $\mathbf{y}_n$ *are given by Theorems 2.1 and 2.2. Furthermore, assume that for* $n = N, N-1, \ldots, 2$.

(i) $\mathbb{E}\left|f_{i,n-1|n-1}^{(k)}\right|^{2+\delta} < \infty$ *for some* $\delta > 0$ *with* $i = 1, \ldots, m_x, k = 1, \ldots, K$,

(ii) $\mathbb{E}\left|f_{i,n-1|n-1}^{(k)} f_{j,n-1|n-1}^{(k)}\right|^{1+\delta} < \infty$ *for some* $\delta > 0$ *with* $i, j = 1, \ldots, m_x, k = 1, \ldots, K$,

(iii) $\mathbf{\Sigma}_{n-1|N}^{\mathbf{xx}}$ *is positive definite.*

Then, by the Markov's SLLN, the estimated mean and variance of the smoothed state vector by (2.25) and (2.26) are consistent when the ensemble size increases infinitely, that is,

$$\widehat{\mathbf{x}}_{n-1|N} \xrightarrow{a.s.} \mathbf{x}_{n-1|N} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{xx}} \xrightarrow{a.s.} \mathbf{\Sigma}_{n-1|N}^{\mathbf{xx}} \textit{ when } K \to \infty.$$

Furthermore, by CLT,

$$\sqrt{K}\left(\widehat{\mathbf{x}}_{n-1|N} - \mathbf{x}_{n-1|N}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_{n-1|N}^{\mathbf{xx}}) \textit{ when } K \to \infty.$$

On the other hand, for the measurement vector $\mathbf{y}_{n-1}$*, assume that for* $n = N, N-1, \ldots, 2$.

(iv) $\mathbb{E}\left|h_{i,n|n-1}^{(k)}\right|^{2+\delta} < \infty$ *for some* $\delta > 0$ *with* $i = 1, \ldots, m_y, k = 1, \ldots, K$,

(v) $\mathbb{E}\left|h_{i,n|n-1}^{(k)} h_{j,n|n-1}^{(k)}\right|^{1+\delta} < \infty$ *for some* $\delta > 0$ *with* $i, j = 1, \ldots, m_y, k = 1, \ldots, K$,

(vi) $\mathbf{\Sigma}_{n|N}^{\mathbf{yy}}$ *is positive definite.*

The estimated mean and variance of smoothed measurement vector $\mathbf{y}_{n-1}$ *by (2.28) and (2.29) share the similar properties as the state vector, that is,*

$$\widehat{\mathbf{y}}_{n-1|N} \xrightarrow{a.s.} \mathbf{y}_{n-1|N} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{yy}} \xrightarrow{a.s.} \mathbf{\Sigma}_{n-1|N}^{\mathbf{yy}} \textit{ when } K \to \infty,$$

by Markov's SLLN and

$$\sqrt{K}\left(\widehat{\mathbf{y}}_{n-1|N} - \mathbf{y}_{n-1|N}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}_{n-1|N}^{\mathbf{yy}}) \textit{ when } K \to \infty,$$

by CLT.

Proof. From (2.25),

$$\widehat{\mathbf{x}}_{n-1|N} = \frac{K^{-1}\sum_{k=1}^{K} \widehat{w}_N^{(k)} \widehat{\mathbf{x}}_{n|n-1}^{(k)}}{K^{-1}\sum_{k=1}^{K} \widehat{w}_N^{(k)}}.$$

Under the regularity conditions for Markov's SLLN, the denominator term converges almost surely, that is,

$$\frac{1}{K}\sum_{k=1}^{K} \widehat{w}_N^{(k)} \xrightarrow{a.s.} \int \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{x}_n\right) p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) d\mathcal{X}_N.$$

By Minkowski's inequality,

$$
\begin{aligned}
& \mathbb{E}\left|\widehat{w}_{N-1}^{(k)}\widehat{x}_{i,n|n-1}^{(k)}\right|^{2+\delta} \\
\leq\ & \left[\left(\mathbb{E}\left|f_{i,n-1|n-1}^{(k)}\right|^{2+\delta}\right)^{1/(2+\delta)} + \left(\mathbb{E}\left|u_{i,n}^{(k)}\right|^{2+\delta}\right)^{1/(2+\delta)}\right]^{2+\delta} \\
<\ & \infty,
\end{aligned}
$$

for $i = 1, \ldots, m_x$. The first term should be finite by condition (i) above and the second term is also finite due to the Gaussianity of $\mathbf{u}_n$. Then, by Markov's SLLN (White, 2001, p.35), the consistency of $\widehat{\mathbf{x}}_{n-1|N}$ is shown directly. Similar procedure can be applied to show the consistency of $\widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{xx}}$. The asymptotic Gaussianity of $\widehat{\mathbf{x}}_{n-1|N}$ can be derived by using Liapounov's CLT and Cramér-Wold theorem. The details of proofs are referred to White (2001, p.114, 118). Similar arguments can be applied to prove for the smoothed measurement vector $\widehat{\mathbf{y}}_{n-1|N}$ as well. ■

Unlike previous prediction and filtering steps, the slightly higher order moments of $\mathbf{x}_n$ are controlled in the smoothing stage because of the heterogeneities of $\widehat{w}_N^{(k)}\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ and $\widehat{w}_N^{(k)}\widehat{\mathbf{y}}_{n|n-1}^{(k)}$ within an ensemble. Conditions (i) and (iv) in this theorem are stronger than needed. Indeed, they are referred to as a simplified version of Markov's conditions (White, 2001, p.35). To show the consistencies of $\widehat{\mathbf{x}}_{n-1|N}$ and $\widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{xx}}$, only the moments of order slightly greater than one are necessary. Stronger assumptions are specified here to satisfy the Lindeberg condition and hence to induce the asymptotic Gaussianity of $\widehat{\mathbf{x}}_{n-1|N}$.

2.4 Estimation of Model Parameters

In this section, the likelihood function the state space model (2.1) and (2.2) is constructed and the estimation of unknown parameters is considered. Following Dee (1995) and Dee and Silva (1995), the likelihood function for this EnGTKF problem is expressed in terms of prediction errors of measurement and approximated by

$$\begin{aligned}
& \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right) \\
= \; & p\left(\mathbf{y}_N|\mathcal{Y}_{N-1}\right) p\left(\mathbf{y}_{N-1}|\mathcal{Y}_{N-2}\right) \ldots p\left(\mathbf{y}_2|\mathcal{Y}_1\right) p\left(\mathbf{y}_1\right) \\
= \; & \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathcal{Y}_{n-1}\right) \\
= \; & (2\pi)^{-\frac{Nm_y}{2}} \prod_{n=1}^{N} \left|\boldsymbol{\Sigma}^{\mathbf{vv}}_{n|n-1}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{v}^{\mathsf{T}}_{n|n-1}\left(\boldsymbol{\Sigma}^{\mathbf{vv}}_{n|n-1}\right)^{-1}\mathbf{v}_{n|n-1}\right),
\end{aligned}$$

with $p\left(\mathbf{y}_1\right) = p\left(\mathbf{y}_1|\mathcal{Y}_0\right)$ where $\mathbf{v}_{n|n-1} = \mathbf{y}_n - \mathbf{y}_{n|n-1}$, $\boldsymbol{\Sigma}^{\mathbf{vv}}_{n|n-1} = \boldsymbol{\Sigma}^{\mathbf{yy}}_{n|n-1} + \mathbf{R}_n$, $\boldsymbol{\theta} = \left[\mathbf{x}^{\mathsf{T}}_{0|0}, \mathrm{vech}(\boldsymbol{\Sigma}_{0|0})^{\mathsf{T}}, \boldsymbol{\psi}^{\mathsf{T}}_1, \ldots, \boldsymbol{\psi}^{\mathsf{T}}_N\right]^{\mathsf{T}}$, $\boldsymbol{\psi}_n = \left[\boldsymbol{\xi}^{\mathsf{T}}_n, \mathrm{vech}(\mathbf{Q}_n)^{\mathsf{T}}, \mathrm{vech}(\mathbf{R}_n)^{\mathsf{T}}\right]^{\mathsf{T}}$. The Gaussian prediction density of the measurement vector $\mathbf{y}_n$ conditional on $\mathcal{Y}_{n-1}$ is derived from Theorem 2.2.

The log-likelihood is then given by

$$\begin{aligned}
\log \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right) = \; & -\frac{Nm_y}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{n=1}^{N}\log\left|\boldsymbol{\Sigma}^{\mathbf{vv}}_{n|n-1}\right| \\
& -\frac{1}{2}\sum_{n=1}^{N}\mathbf{v}^{\mathsf{T}}_{n|n-1}\left(\boldsymbol{\Sigma}^{\mathbf{vv}}_{n|n-1}\right)^{-1}\mathbf{v}_{n|n-1}. \qquad (2.30)
\end{aligned}$$

The unknown parameter estimate is the solution of the following optimization

problem:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right). \tag{2.31}$$

Under the regularity conditions of Jensen and Petersen (1999), the estimate $\widehat{\boldsymbol{\theta}}$ is strongly consistent and asymptotically normal (Jensen and Petersen, 1999, Theorem 3.3).

Practically, the maximization of (2.30) can be performed by conventional local optimization algorithms, such as Newton-Raphson, Quasi-Newton and so on. However, in many cases, the log-likelihood may not be unimodal even if the measurement and state errors are Gaussian. In addition, the initial values are essential for the convergence of local optimization algorithm. Consequently, a hybrid procedure of optimization algorithm is suggested to estimate unknown parameters in (2.30).

2.4.1 Orthogonal Decomposition

Due to the multivariate nature of the state space model, the positive definiteness of variance-covariance matrices are difficult to ensure during the process of estimation especially the high-dimensional cases. Nevertheless, a special transformation technique is suggested here. Specifically, a variance-covariance matrix can be expressed as a product of rotation matrices and column vector of variances (Schwefel, 1981; Rudolph, 1992). This technique has been used in evolutionary strategy which is a kind of evolutionary algorithms (Bäck, 1996), for a long period of time.

Let $\mathbf{C} = (c_{ij})$ be an $s \times s$ variance-covariance matrix which is positive definite

and symmetric. Then, it can be decomposed into elementary rotation matrices and diagonal matrices (Faddeev and Faddeeva, 1963, p.27). Specifically,

$$\mathbf{C} = \left(\prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \widetilde{\mathbf{R}}(\alpha_{ij}) \right)^{\mathsf{T}} \mathbf{D}^{\mathsf{T}} \mathbf{D} \left(\prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \widetilde{\mathbf{R}}(\alpha_{ij}) \right), \tag{2.32}$$

where $\alpha_{ij} \in (-\pi, \pi]$, $\widetilde{\mathbf{R}}(\alpha_{ij}) = \widetilde{\mathbf{R}}^{\mathsf{T}}(-\alpha_{ij})$ is an $s \times s$ elementary rotation matrix which indeed is a unit matrix except elements $r_{ii} = r_{jj} = \cos\alpha_{ij}$, $r_{ij} = -r_{ji} = -\sin\alpha_{ij}$ and $\mathbf{D} = (\mathrm{diag}(\mathbf{C}))^{1/2}$ denotes a diagonal matrix whose diagonal elements are square roots of diagonal elements of $\mathbf{C}$. For example, $\widetilde{\mathbf{R}}(\alpha_{24})$ is given by

$$\widetilde{\mathbf{R}}\left(\alpha_{24}\right) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cos\alpha_{24} & 0 & -\sin\alpha_{24} & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & \sin\alpha_{24} & 0 & \cos\alpha_{24} & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

By using this orthogonal decomposition, the total number of unknown parameters of variance matrix $\mathbf{C}$ remains the same, that is, $s(s-1)/2+s$ or $s(s+1)/2$. Nevertheless, the positive definiteness of matrix $\mathbf{C}$ during optimization can be maintained easily. Rudolph (1992) showed the usefulness of this orthogonal decomposition with the mutation operator in evolutionary strategy.

In the current situation, the matrices $\mathbf{R}$'s, $\mathbf{Q}$'s and $\mathbf{\Sigma}_{0|0}$ are the unknown matrices being estimated.

2.4.2 Stochastic Search Optimization

To find the optimal values of parameters in the case of (2.31), one can draw a set of random parameter values for $\boldsymbol{\theta}$ from a specified distribution, typically the uniform distribution, and substitute the values into the objective function in (2.31). Then, the set of parameter values with the largest value of the objective function will be the optimal parameter values for the problem. If the number of random draws is sufficiently large, it can be shown that the estimated optimal parameter value converges to the true optimal parameter value (Spall, 2003, Theorem 2.1). This result is intuitive because a random draw of parameter values can be considered as a stochastic search over the parameter space globally. Then, a sufficiently large number of random draws can cover a wide range of parameter values over the parameter space and hence the optimal parameter values can be located. To improve the searching results, a localized stochastic search method can be adapted here (Spall, 2003, p.44). Indeed, this approach can incorporate priori information of parameters into the stochastic search. Rather than searching for the whole spectrum of parameter space, the priori information of parameters is used to construct the initial parameter estimates and the range of parameter estimates. Then, the stochastic search begins around the initial parameter estimates with specified range. Theoretical convergence results of the search were proved by many researchers, as mentioned in Spall (2003, p.44).

However, as shown in Spall (2003), it is difficult to ensure this kind of global convergence practically because a huge number of iterations is required. As in the case of state estimation, local optimization is adapted to improve the estimation results. Then, the optimization algorithm for (2.31) can be summarized below:

1. Determine the initial parameter values by using information of the structure

of the state space model.

2. Decide the ranges of unknown parameters except variance-covariance matrices where the localized stochastic search occurs.

3. By using (2.32), each variance-covariance matrices of dimension d should have $d(d+1)/2$ parameters. The first d ones are corresponding to the standard deviations and the remaining $d(d-1)/2$ ones are those rotation angles over ranges $(-\pi, \pi]$.

4. Generate a sufficiently large number of possible parameter values by uniform distribution according to the intervals defined previously.

5. Locate the set of parameter values with the largest value in (2.31) and the result is obtained directly.

6. The set of parameter values in Step 5 is considered as the initial values of a local optimization algorithm, for example, Quasi-Newton.

2.5 Estimation of Standard Errors

In this section, all parameter estimates by (2.31) are indicated by $\widehat{\ }$. In order to estimate the standard errors of estimated parameters in Section 2.4, the expected Fisher's information matrix is required and defined by

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left(\frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}} \frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right). \tag{2.33}$$

Apart from Cavanaugh and Shumway (1996) which computed the expected Fisher's information matrix (2.33) for linear state space model, a typical and

consistent approximation of Fisher's information matrix, the observed Fisher's information matrix, is used here, that is,

$$\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \log \mathcal{L}_n}{\partial \boldsymbol{\theta}}\frac{\partial \log \mathcal{L}_n}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \tag{2.34}$$

where $\mathcal{L}_n$ stands for the likelihood of the nth observation. Note that $(\mathcal{I}(\widehat{\boldsymbol{\theta}}))^{-1}$ is the variance of the estimated parameter $\widehat{\boldsymbol{\theta}}$ by Cramér-Rao lower bound under suitable regularity conditions (Casella and Berger, 2001, p.335). In order to simplify the derivation of the observed Fisher information matrix (2.34), the likelihood and hence the log-likelihood functions are reformulated as

$$\mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right) = \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathcal{Y}_{n-1}\right) = \prod_{n=1}^{N}\mathcal{L}_n,$$

and

$$\log \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right) = \sum_{n=1}^{N}\log \mathcal{L}_n, \tag{2.35}$$

respectively where

$$\mathcal{L}_n \equiv \begin{cases} \iint p\left(\mathbf{y}_1|\mathbf{x}_1\right) p\left(\mathbf{x}_1|\mathbf{x}_0\right) p\left(\mathbf{x}_0\right) d\mathbf{x}_0 d\mathbf{x}_1 & \text{for } n = 1, \\ \int p\left(\mathbf{y}_n|\mathbf{x}_n\right) p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) d\mathbf{x}_n & \text{for } n > 1. \end{cases}$$

Note that $\boldsymbol{\theta}$ is omitted in $\mathcal{L}_n$ for notation simplicity. Also, $\mathcal{L}_n$ is complicated for $n = 1$ because the stochastic initial state vector $\mathbf{x}_0$ is assumed. The representation of the likelihood function here follows from Assumptions 2.1(i) and (ii). Indeed, the parameter vector $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta} = \left[\mathbf{x}_{0|0}^{\mathsf{T}}, \operatorname{vech}\left(\boldsymbol{\Sigma}_{0|0}\right)^{\mathsf{T}}, \boldsymbol{\psi}_1^{\mathsf{T}}, \ldots, \boldsymbol{\psi}_N^{\mathsf{T}}\right]^{\mathsf{T}},$$

where $\boldsymbol{\psi}_n = \left[\boldsymbol{\xi}_n^{\mathsf{T}}, \text{vech}\left(\mathbf{Q}_n\right)^{\mathsf{T}}, \text{vech}\left(\mathbf{R}_n\right)^{\mathsf{T}}\right]^{\mathsf{T}}$.

When $n = 1$, the derivative of $\mathcal{L}_n$ with respect to $\boldsymbol{\theta}$ is

$$\begin{aligned}\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\theta}} &= \iint \frac{\partial p\left(\mathbf{y}_1|\mathbf{x}_1\right)}{\partial \boldsymbol{\theta}} p\left(\mathbf{x}_1|\mathbf{x}_0\right) p\left(\mathbf{x}_0\right) d\mathbf{x}_0 d\mathbf{x}_1 \\ &+ \iint p\left(\mathbf{y}_1|\mathbf{x}_1\right) \frac{\partial p\left(\mathbf{x}_1|\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}} p\left(\mathbf{x}_0\right) d\mathbf{x}_0 d\mathbf{x}_1 \\ &+ \iint p\left(\mathbf{y}_1|\mathbf{x}_1\right) p\left(\mathbf{x}_1|\mathbf{x}_0\right) \frac{\partial p\left(\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}} d\mathbf{x}_0 d\mathbf{x}_1. \end{aligned} \tag{2.36}$$

When $n > 1$, the derivative of $\mathcal{L}_n$ with respect to $\boldsymbol{\theta}$ is

$$\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}} = \int \frac{\partial p(\mathbf{y}_n|\mathbf{x}_n)}{\partial \boldsymbol{\theta}} p(\mathbf{x}_n|\mathcal{Y}_{n-1}) d\mathbf{x}_n + \int p(\mathbf{y}_n|\mathbf{x}_n) \frac{\partial p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}} d\mathbf{x}_n. \tag{2.37}$$

Recall that the recursive relationship of prediction and filtering densities of the model state is

$$\begin{aligned} p\left(\mathbf{x}_n|\mathcal{Y}_{n-1}\right) &= \int p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) d\mathbf{x}_{n-1}, \\ p(\mathbf{x}_n|\mathcal{Y}_n) &= \frac{p(\mathbf{y}_n|\mathbf{x}_n) p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\mathcal{L}_n}, \end{aligned}$$

for $n > 1$. Then, the recursive relationship of their derivatives with respect to $\boldsymbol{\theta}$ is given by

$$\frac{\partial p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}} = \int \frac{\partial p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)}{\partial \boldsymbol{\theta}} p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) d\mathbf{x}_{n-1}$$

$$+\int p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)\frac{\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1})}{\partial\boldsymbol{\theta}}d\mathbf{x}_{n-1}, \tag{2.38a}$$

$$\frac{\partial p(\mathbf{x}_n|\mathcal{Y}_n)}{\partial\boldsymbol{\theta}} = \frac{1}{\mathcal{L}_n}\left\{\frac{\partial p(\mathbf{y}_n|\mathbf{x}_n)}{\partial\boldsymbol{\theta}}p(\mathbf{x}_n|\mathcal{Y}_{n-1}) + p(\mathbf{y}_n|\mathbf{x}_n)\frac{\partial p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial\boldsymbol{\theta}}\right\} - \frac{p(\mathbf{x}_n|\mathcal{Y}_n)}{\mathcal{L}_n}\frac{\partial\mathcal{L}_n}{\partial\boldsymbol{\theta}}. \tag{2.38b}$$

At this stage, $\partial p(\mathbf{x}_1|\mathcal{Y}_1)/\partial\boldsymbol{\theta}$ is considered as the initial value of recursive derivatives in (2.37) and it is obtained by

$$\begin{aligned} p(\mathbf{x}_1|\mathcal{Y}_1) &= \frac{p(\mathbf{y}_1|\mathbf{x}_1)}{\mathcal{L}_1}\int p(\mathbf{x}_1|\mathbf{x}_0)p(\mathbf{x}_0)d\mathbf{x}_0 \\ \frac{\partial p(\mathbf{x}_1|\mathcal{Y}_1)}{\partial\boldsymbol{\theta}} &= \frac{1}{\mathcal{L}_1}\left\{\frac{\partial p(\mathbf{y}_1|\mathbf{x}_1)}{\partial\boldsymbol{\theta}}\int p(\mathbf{x}_1|\mathbf{x}_0)p(\mathbf{x}_0)d\mathbf{x}_0\right. \\ &\quad + p(\mathbf{y}_1|\mathbf{x}_1)\int\frac{\partial p\left(\mathbf{x}_1|\mathbf{x}_0\right)}{\partial\boldsymbol{\theta}}p(\mathbf{x}_0)d\mathbf{x}_0 \\ &\quad \left.+ p(\mathbf{y}_1|\mathbf{x}_1)\int p(\mathbf{x}_1|\mathbf{x}_0)\frac{\partial p\left(\mathbf{x}_0\right)}{\partial\boldsymbol{\theta}}d\mathbf{x}_0\right\} - \frac{p(\mathbf{x}_1|\mathcal{Y}_1)}{\mathcal{L}_1}\frac{\partial\mathcal{L}_1}{\partial\boldsymbol{\theta}}. \end{aligned} \tag{2.39}$$

Furthermore, the derivatives in (2.36), (2.37), (2.38) and (2.39) are rewritten in the form of

$$\frac{\partial p(\mathbf{z})}{\partial\boldsymbol{\theta}} = \frac{\partial\log p(\mathbf{z})}{\partial\boldsymbol{\theta}}p(\mathbf{z}),$$

where $p(\mathbf{z})$ represents the densities $p(\mathbf{x}_n|\mathcal{Y}_{n-1})$, $p(\mathbf{x}_n|\mathcal{Y}_n)$, $p(\mathbf{x}_1|\mathbf{x}_0)$ and $p(\mathbf{x}_0)$. After the above expression is applied to (2.36), (2.37), (2.38) and (2.39), the recursive relationship of derivatives of log-density is obtained directly. More explicitly, they become

$$\frac{\partial\log p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial\boldsymbol{\theta}}$$

$$
\begin{aligned}
&= \frac{1}{p(\mathbf{x}_n|\mathcal{Y}_{n-1})}\left\{\int \frac{\partial \log p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)}{\partial \boldsymbol{\theta}} p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) d\mathbf{x}_{n-1}\right. \\
&\quad \left. + \int \frac{\partial \log p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}} p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}) d\mathbf{x}_{n-1}\right\}, \qquad (2.40a) \\
&\frac{\partial \log p(\mathbf{x}_n|\mathcal{Y}_n)}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n)}{\partial \boldsymbol{\theta}} + \frac{\partial \log p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}} - \frac{1}{\mathcal{L}_n}\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}}, \qquad (2.40b) \\
&\frac{\partial \log p(\mathbf{x}_1|\mathcal{Y}_1)}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \log p(\mathbf{y}_1|\mathbf{x}_1)}{\partial \boldsymbol{\theta}} + \frac{p(\mathbf{y}_1|\mathbf{x}_1)}{p(\mathbf{x}_1|\mathcal{Y}_1)\mathcal{L}_1} \\
&\quad \times \int \left\{\frac{\partial \log p\left(\mathbf{x}_1|\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}} + \frac{\partial \log p\left(\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}}\right\} p\left(\mathbf{x}_1|\mathbf{x}_0\right) p(\mathbf{x}_0) d\mathbf{x}_0 \\
&\quad - \frac{1}{\mathcal{L}_1}\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\theta}}. \qquad (2.40c)
\end{aligned}
$$

and the derivative of the likelihood with respect to $\boldsymbol{\theta}$ in terms of the derivatives of log-density becomes

$$
\begin{aligned}
&\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\theta}} \\
&= \iint \left\{\frac{\partial \log p\left(\mathbf{y}_1|\mathbf{x}_1\right)}{\partial \boldsymbol{\theta}} + \frac{\partial \log p\left(\mathbf{x}_1|\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}} + \frac{\partial \log p\left(\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}}\right\} \\
&\quad \times p\left(\mathbf{x}_1|\mathbf{x}_0\right) p\left(\mathbf{x}_0\right) d\mathbf{x}_0 d\mathbf{x}_1. \qquad (2.41a) \\
&\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}} \\
&= \int \left\{\frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n)}{\partial \boldsymbol{\theta}} + \frac{\partial \log p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}\right\} p(\mathbf{y}_n|\mathbf{x}_n) p(\mathbf{x}_n|\mathcal{Y}_{n-1}) d\mathbf{x}_n \\
&\qquad (2.41b)
\end{aligned}
$$

for $n = 2, \ldots, N$.

Indeed, the log-density functions used in (2.40) and (2.41) are specified as

$$
\begin{aligned}
& \log p\left(\mathbf{x}_0\right) \\
= & -\frac{m_x}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right| - \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_{0|0}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\left(\mathbf{x}_0 - \mathbf{x}_{0|0}\right), \\
& \log p\left(\mathbf{y}_n|\mathbf{x}_n\right) \\
= & -\frac{m_y}{2}\log(2\pi) - \frac{1}{2}\log\left|\mathbf{R}_n\right| - \frac{1}{2}\left(\mathbf{y}_n - \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}}\mathbf{R}_n^{-1}\left(\mathbf{y}_n - \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)\right), \\
\text{and} \quad & \log p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right) \\
= & -\frac{m_x}{2}\log(2\pi) - \frac{1}{2}\log\left|\mathbf{Q}_n\right| - \frac{1}{2}\left(\mathbf{x}_n - \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}}\mathbf{Q}_n^{-1}\left(\mathbf{x}_n - \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)\right),
\end{aligned}
$$

respectively. Since the estimation of the model state and hence the measurement can be computed in Section 2.2, iterative estimations between parameter $\boldsymbol{\theta}$ and the model state $\mathbf{x}_n$ by the EnGTKF imply the joint estimation of parameter and model state by maximization of (2.35).

In order to derive the derivatives of log-density with respect to $\boldsymbol{\theta}$, the following expressions (Schott, 1997, p.336) are required:

$$
\begin{aligned}
\frac{\partial \log|\mathbf{A}|}{\partial \mathrm{vech}\left(\mathbf{A}\right)^{\mathsf{T}}} &= \frac{\partial \log|\mathbf{A}|}{\partial|\mathbf{A}|}\frac{\partial|\mathbf{A}|}{\partial \mathrm{vech}\left(\mathbf{A}\right)^{\mathsf{T}}} = |\mathbf{A}|^{-1}\mathrm{vec}\left(\mathbf{A}_{\#}^{\mathsf{T}}\right)^{\mathsf{T}}\mathbf{D}_m \\
&= \mathrm{vec}\left(\mathbf{A}^{-1}\right)^{\mathsf{T}}\mathbf{D}_m, \qquad (2.42) \\
\text{and} \quad \frac{\partial \mathrm{vec}\left(\mathbf{A}^{-1}\right)}{\partial \mathrm{vech}\left(\mathbf{A}\right)^{\mathsf{T}}} &= -\left(\mathbf{A}^{-1}\otimes\mathbf{A}^{-1}\right)\mathbf{D}_m, \qquad (2.43)
\end{aligned}
$$

with $\mathbf{A}^{-1} = |\mathbf{A}|^{-1}\mathbf{A}_{\#}$ where $\mathbf{D}_m$ denotes the duplication matrix of order m such that $\mathbf{D}_m\mathrm{vech}(\mathbf{A}) = \mathrm{vec}(\mathbf{A})$, $\mathbf{A}_{\#}$ is the adjoint of $\mathbf{A}$ and $\mathbf{A}$ is a $m\times m$ symmetric matrix.

Now, consider the derivatives of log-density with respect to $\boldsymbol{\theta}$.

$$
\begin{aligned}
&\frac{\partial \log p\left(\mathbf{x}_0\right)}{\partial \mathbf{x}_{0|0}^{\mathsf{T}}} \\
= & -\frac{1}{2}\frac{\partial}{\partial \mathbf{x}_{0|0}^{\mathsf{T}}}\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)\right) \\
= & \left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}, \qquad (2.44)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \log p\left(\mathbf{x}_0\right)}{\partial \mathrm{vech}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}} \\
= & -\frac{1}{2}\frac{\partial \log\left|\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right|}{\partial \mathrm{vech}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}}-\frac{1}{2}\frac{\partial \mathrm{tr}\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)\right)}{\partial \mathrm{vech}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}} \\
= & -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}\mathbf{D}_{m_x} \\
& -\frac{1}{2}\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\otimes\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\right)\frac{\partial \mathrm{vec}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}}{\partial \mathrm{vech}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)} \\
= & -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}\mathbf{D}_{m_x} \\
& +\frac{1}{2}\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\otimes\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\right)\left(\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\otimes\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\right)\mathbf{D}_{m_x} \\
= & -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}\mathbf{D}_{m_x} \\
& +\frac{1}{2}\left(\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\right)\otimes\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\right)\right)\mathbf{D}_{m_x}, \qquad (2.45)
\end{aligned}
$$

where $\mathbf{D}_{m_x}\mathrm{vech}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right) = \mathrm{vec}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)$. The first term in the second equality of $\partial \log p\left(\mathbf{x}_0\right)/\partial \mathrm{vech}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}$ is obtained by (2.42) while the second term is obtained by first using the following identity (Schott, 1997, p.263):

$$
\mathrm{tr}\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)\right)=\left(\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)^{\mathsf{T}}\otimes\left(\mathbf{x}_0-\mathbf{x}_{0|0}\right)\right)\mathrm{vec}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{-1}.
$$

and then by (2.43). Note that the derivatives corresponding to the components

other than $\mathbf{x}_{0|0}$ and $\mathrm{vech}\left(\mathbf{\Sigma}_{0|0}^{\mathbf{xx}}\right)$ are zero. Note that

$$\begin{aligned}
&\frac{\partial \log p\left(\mathbf{y}_n|\mathbf{x}_n\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}} \\
= \;& \left(\mathbf{y}_n - \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \mathbf{R}_n^{-1} \frac{\partial \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}} \qquad (2.46) \\
&\frac{\partial \log p\left(\mathbf{y}_n|\mathbf{x}_n\right)}{\partial \mathrm{vech}\left(\mathbf{R}_n\right)^{\mathsf{T}}} \\
= \;& -\frac{1}{2}\mathrm{vec}\left(\mathbf{R}_n\right)^{\mathsf{T}} \mathbf{D}_{m_y} \\
&+\frac{1}{2}\left(\left(\mathbf{y}_n - \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \otimes \left(\mathbf{y}_n - \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}}\right)\left(\mathbf{R}_n^{-1} \otimes \mathbf{R}_n^{-1}\right)\mathbf{D}_{m_y} \\
= \;& -\frac{1}{2}\mathrm{vec}\left(\mathbf{R}_n\right)^{\mathsf{T}} \mathbf{D}_{m_y} \\
&+\frac{1}{2}\left(\left(\left(\mathbf{y}_n - \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \mathbf{R}_n^{-1}\right) \otimes \left(\left(\mathbf{y}_n - \mathbf{h}\left(\mathbf{x}_n;\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \mathbf{R}_n^{-1}\right)\right)\mathbf{D}_{m_y}, \qquad (2.47)
\end{aligned}$$

where $\mathbf{D}_{m_y}\mathrm{vech}(\mathbf{R}_n) = \mathrm{vec}(\mathbf{R}_n)$. Note that the derivatives corresponding to the components other than $\boldsymbol{\xi}_n$ and $\mathrm{vech}(\mathbf{R}_n)$ are zero. Then,

$$\begin{aligned}
&\frac{\partial \log p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}} \\
= \;& \left(\mathbf{x}_n - \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \mathbf{Q}_n^{-1} \frac{\partial \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}}, \qquad (2.48) \\
&\frac{\partial \log p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)}{\partial \mathrm{vech}\left(\mathbf{Q}_n\right)^{\mathsf{T}}} \\
= \;& -\frac{1}{2}\mathrm{vec}\left(\mathbf{Q}_n\right)^{\mathsf{T}} \mathbf{D}_{m_x} \\
&+\frac{1}{2}\left(\left(\mathbf{x}_n - \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \otimes \left(\mathbf{x}_n - \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}}\right)\left(\mathbf{Q}_n^{-1} \otimes \mathbf{Q}_n^{-1}\right)\mathbf{D}_{m_x} \\
= \;& -\frac{1}{2}\mathrm{vec}\left(\mathbf{Q}_n\right)^{\mathsf{T}} \mathbf{D}_{m_x} \\
&+\frac{1}{2}\left(\left(\left(\mathbf{x}_n - \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \mathbf{Q}_n^{-1}\right) \otimes \left(\left(\mathbf{x}_n - \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}_n\right)\right)^{\mathsf{T}} \mathbf{Q}_n^{-1}\right)\right)\mathbf{D}_{m_x},
\end{aligned}$$

(2.49)

where $\mathbf{D}_{m_x}\text{vech}(\mathbf{Q}_n) = \text{vec}(\mathbf{Q}_n)$. Note that the derivatives corresponding to the components other than $\boldsymbol{\xi}_n$ and $\text{vech}(\mathbf{Q}_n)$ are zero.

Indeed, the derivatives of the log-densities and hence those of the likelihood can be estimated by the Monte Carlo method because they can be considered as the expectation terms. Specifically,

$$\begin{aligned}
&\widehat{\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\theta}}} \\
&= \sum_{k=1}^{K}\sum_{j=1}^{K}\left\{\left[\frac{\partial \log p(\mathbf{y}_1|\mathbf{x}_1)}{\partial \boldsymbol{\theta}}\right]^{(j)} + \left[\frac{\partial \log p(\mathbf{x}_1|\mathbf{x}_0)}{\partial \boldsymbol{\theta}}\right]^{(j,k)} + \left[\frac{\partial \log p(\mathbf{x}_0)}{\partial \boldsymbol{\theta}}\right]^{(k)}\right\}, \qquad (2.50a)
\end{aligned}$$

$$\begin{aligned}
&\widehat{\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}}} \\
&= \sum_{k=1}^{K}\mathcal{C}_n^{(k)}\left\{\left[\frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n)}{\partial \boldsymbol{\theta}}\right]^{(k)} + \left[\frac{\partial \log p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}\right]^{(k)}\right\}, \qquad (2.50b)
\end{aligned}$$

and the subsequent derivatives are

$$\begin{aligned}
&\left[\frac{\partial \log p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}\right]^{(j)} \\
&= \frac{1}{p(\mathbf{x}_n = \widehat{\mathbf{x}}_{n|n-1}^{(j)}|\mathcal{Y}_{n-1})}\sum_{k=1}^{K}\mathcal{B}_n^{(j,k)}\left\{\left[\frac{\partial \log p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)}{\partial \boldsymbol{\theta}}\right]^{(j,k)}\right. \\
&\quad \left. + \left[\frac{\partial \log p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}\right]^{(k)}\right\}, \qquad (2.51a)
\end{aligned}$$

$$\begin{aligned}
&\left[\frac{\partial \log p(\mathbf{x}_n|\mathcal{Y}_n)}{\partial \boldsymbol{\theta}}\right]^{(j)} \\
&= \left[\frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n)}{\partial \boldsymbol{\theta}}\right]^{(j)} + \left[\frac{\partial \log p(\mathbf{x}_n|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}\right]^{(j)} - \frac{1}{\widehat{\mathcal{L}}_n}\widehat{\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}}}, \qquad (2.51b)
\end{aligned}$$

$$
\begin{aligned}
&\left[\frac{\partial \log p(\mathbf{x}_1|\mathcal{Y}_1)}{\partial \boldsymbol{\theta}}\right]^{(j)} \\
= &\left[\frac{\partial \log p(\mathbf{y}_1|\mathbf{x}_1)}{\partial \boldsymbol{\theta}}\right]^{(j)} + \frac{1}{\sum_{k=1}^{K} B_1^{(j,k)}} \\
&\times \sum_{k=1}^{K} B_1^{(j,k)} \left\{ \left[\frac{\partial \log p\left(\mathbf{x}_1|\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}}\right]^{(k)} + \left[\frac{\partial \log p\left(\mathbf{x}_0\right)}{\partial \boldsymbol{\theta}}\right]^{(k)} \right\} \\
&- \frac{1}{\widehat{\mathcal{L}}_1} \frac{\partial \widehat{\mathcal{L}}_1}{\partial \boldsymbol{\theta}},
\end{aligned} \tag{2.51c}
$$

and

$$
\left[\frac{\partial \log p\left(\mathbf{x}_0\right)}{\partial \mathbf{x}_{0|0}^{\mathsf{T}}}\right]^{(k)} = \widehat{\mathbf{u}}_{0|0}^{\mathsf{T}(k)} (\widehat{\boldsymbol{\Sigma}}_{0|0}^{\mathbf{xx}})^{-1},
$$

$$
\begin{aligned}
\left[\frac{\partial \log p\left(\mathbf{x}_0\right)}{\partial \mathrm{vech}\left(\boldsymbol{\Sigma}_{0|0}^{\mathbf{xx}}\right)^{\mathsf{T}}}\right]^{(k)} &= -\frac{1}{2} \mathrm{vec}(\widehat{\boldsymbol{\Sigma}}_{0|0}^{\mathbf{xx}})^{\mathsf{T}} \mathbf{D}_{m_x} \\
&\quad + \frac{1}{2} \left(\left(\widehat{\mathbf{u}}_{0|0}^{\mathsf{T}(k)} (\widehat{\boldsymbol{\Sigma}}_{0|0}^{\mathbf{xx}})^{-1} \right) \otimes \left(\widehat{\mathbf{u}}_{0|0}^{\mathsf{T}(k)} (\widehat{\boldsymbol{\Sigma}}_{0|0}^{\mathbf{xx}})^{-1} \right) \right) \mathbf{D}_{m_x},
\end{aligned} \tag{2.52a}
$$

(2.52b)

$$
\left[\frac{\partial \log p\left(\mathbf{y}_n|\mathbf{x}_n\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}}\right]^{(j)} = \widehat{\mathbf{v}}_{n|n-1}^{\mathsf{T}(j)} \widehat{\mathbf{R}}_n^{-1} \left[\frac{\partial \mathbf{h}\left(\mathbf{x}_n; \boldsymbol{\xi}_n\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}}\right]^{(j)}_{\mathbf{x}_n = \widehat{\mathbf{x}}_{n|n-1}^{(j)}}, \tag{2.52c}
$$

$$
\begin{aligned}
\left[\frac{\partial \log p\left(\mathbf{y}_n|\mathbf{x}_n\right)}{\partial \mathrm{vech}\left(\mathbf{R}_n\right)^{\mathsf{T}}}\right]^{(j)} &= -\frac{1}{2} \mathrm{vec}(\widehat{\mathbf{R}}_n)^{\mathsf{T}} \mathbf{D}_{m_y} \\
&\quad + \frac{1}{2} \left(\left(\widehat{\mathbf{v}}_{n|n-1}^{\mathsf{T}(j)} \widehat{\mathbf{R}}_n^{-1} \right) \otimes \left(\widehat{\mathbf{v}}_{n|n-1}^{\mathsf{T}(j)} \widehat{\mathbf{R}}_n^{-1} \right) \right) \mathbf{D}_{m_y},
\end{aligned} \tag{2.52d}
$$

$$
\left[\frac{\partial \log p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}}\right]^{(j,k)} = \widehat{\mathbf{u}}_{n|n-1}^{\mathsf{T}(j,k)} \widehat{\mathbf{Q}}_n^{-1} \left[\frac{\partial \mathbf{f}\left(\mathbf{x}_{n-1}; \boldsymbol{\xi}_n\right)}{\partial \boldsymbol{\xi}_n^{\mathsf{T}}}\right]^{(k)}_{\mathbf{x}_{n-1} = \widehat{\mathbf{x}}_{n-1|n-1}^{(k)}}, \tag{2.52e}
$$

$$
\begin{aligned}
\left[\frac{\partial \log p\left(\mathbf{x}_n|\mathbf{x}_{n-1}\right)}{\partial \mathrm{vech}\left(\mathbf{Q}_n\right)^{\mathsf{T}}}\right]^{(j,k)} &= -\frac{1}{2} \mathrm{vec}(\widehat{\mathbf{Q}}_n)^{\mathsf{T}} \mathbf{D}_{m_x} \\
&\quad + \frac{1}{2} \left(\left(\widehat{\mathbf{u}}_{n|n-1}^{\mathsf{T}(j,k)} \widehat{\mathbf{Q}}_n^{-1} \right) \otimes \left(\widehat{\mathbf{u}}_{n|n-1}^{\mathsf{T}(j,k)} \widehat{\mathbf{Q}}_n^{-1} \right) \right) \mathbf{D}_{m_x},
\end{aligned} \tag{2.52f}
$$

for $j, k = 1, \ldots, N$ where $\widehat{\mathbf{u}}_{0|0}^{(k)} = \widehat{\mathbf{x}}_{0|0}^{(k)} - \widehat{\mathbf{x}}_{0|0}$, $\widehat{\mathbf{u}}_{n|n-1}^{(j,k)} = \widehat{\mathbf{x}}_{n|n-1}^{(j)} - \mathbf{f}(\widehat{\mathbf{x}}_{n-1|n-1}^{(k)}; \widehat{\boldsymbol{\xi}}_n)$,

$\widehat{\mathbf{v}}_{n|n-1}^{(k)} = \mathbf{y}_n - \mathbf{h}(\widehat{\mathbf{x}}_{n|n-1}^{(k)}; \widehat{\boldsymbol{\xi}}_n)$, $\mathcal{B}_n^{(j,k)} = p(\mathbf{x}_n = \widehat{\mathbf{x}}_{n|n-1}^{(j)} | \mathbf{x}_{n-1} = \widehat{\mathbf{x}}_{n-1|n-1}^{(k)})$, $\mathcal{C}_n^{(k)} = p(\mathbf{y}_n | \mathbf{x}_n = \widehat{\mathbf{x}}_{n|n-1}^{(k)})$ for $n \geq 1$, $\widehat{\mathbf{x}}_{0|0}^{(k)}$ is drawn from the density $p(\mathbf{x}_0)$, $\widehat{\mathcal{L}}_n$ is the likelihood of the nth observation evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

In the case of time-invariant parameters, that is, $\boldsymbol{\psi}_1 = \cdots = \boldsymbol{\psi}_N = \boldsymbol{\psi}$, the estimated parameter vector $\widehat{\boldsymbol{\theta}}$ is reduced to

$$\widehat{\boldsymbol{\theta}} = \left[\widehat{\mathbf{x}}_{0|0}^{\mathsf{T}}, \mathrm{vech}(\widehat{\boldsymbol{\Sigma}}_{0|0})^{\mathsf{T}}, \widehat{\boldsymbol{\psi}}^{\mathsf{T}} \right]^{\mathsf{T}},$$

and the derivatives of the likelihood function are followed by the same procedures easily. Thus, they are not reproduced here for convenience.

Then, the components in the observed Fisher's information matrix $\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}})$ in (2.34) can be constructed from the outer product of (2.50). Specifically, $\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}})$ is estimated by

$$\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{n=1}^{N} \widehat{\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}}} \widehat{\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}^{\mathsf{T}}}}.$$

Once all estimates of $\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}})$ are obtained, the standard errors of $\widehat{\boldsymbol{\theta}}$ are just the square roots of the diagonal elements of $(\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}))^{-1}$. Although the first derivatives given in (2.50) show the feasibility of the estimation of standard errors, it may be impractical as the number of parameters increases faster than that of observations resulting an identification problem. Nevertheless, these first derivatives can provide the general framework for models in the reduced form of (2.1) and (2.2), for example, models with time-invariant parameters or the model in Section 2.7.

2.6 Extension to Non-Gaussian Prediction Densities

In previous sections, the prediction densities for the state vector $\mathbf{x}_n$ and the measurement vector $\mathbf{y}_n$, that is, $p(\mathbf{x}_n|\mathcal{Y}_{n-1})$ and $p(\mathbf{y}_n|\mathcal{Y}_{n-1})$, are asymptotically Gaussian. However, practically, they may deviate from the Gaussian density even if the measurement and state errors are Gaussian. This in turn leads to suggest that the prediction densities $p(\mathbf{x}_n|\mathcal{Y}_{n-1})$ and $p(\mathbf{y}_n|\mathcal{Y}_{n-1})$ are approximated by the Gram-Charlier densities. Indeed, they are obtained from the Gram-Charlier type-A series expansion around Gaussian densities. Specifically, for a random variable X with mean zero and variance σ^2, the Gram-Charlier expansion of a density function $f(x)$ around a Gaussian density $\mathcal{N}(x; 0, \sigma^2)$ is given by

$$f(x) = \mathcal{N}\left(x; 0, \sigma^2\right)\left[1 + \sum_{s=1}^{b} d_s \mathcal{H}_s(x)\right], \tag{2.53}$$

where

$$\mathcal{N}\left(x; \mu, \sigma^2\right) \equiv \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

and $\mathcal{H}_s(x)$ denotes the Hermite polynomials of order s. Following Kendall et al. (1977),

$$\mathcal{H}_s(x) = \begin{cases} \sum_{i=0}^{s/2} (-1)^i x^{s-2i} \frac{s!}{2^i i!(s-2i)!} & \text{if } s \text{ is even,} \\ \sum_{i=0}^{(s-1)/2} (-1)^i x^{s-2i} \frac{s!}{2^i i!(s-2i)!} & \text{if } s \text{ is odd.} \end{cases}$$

As mentioned in Cramér (1946, p.223), the convergence of the series in (2.53) is ensured when $\int_{-\infty}^{\infty} \exp(x^2/4)\, dF(x) < \infty$ for $b \to \infty$ where $F(x)$ denotes

the cumulative distribution function of x. Practically, a finite and small number of terms, b, is always used to approximate the density function $f(x)$. On the other hand, the definition of the Gram-Charlier density in (2.53) faced a problem that the positive estimated density is not guaranteed without certain restrictions imposed on the parameters d_s's. In the univariate case, Jondeau and Rockinger (2001) showed the positivity conditions on the parameters d_s's and developed a two-step procedure for parameter estimation. Alternatively, Perote and del Brío (2006) suggested a restricted form of the Gram-Charlier density for a random variable X with mean zero and variance σ^2, that is,

$$f(x) = \mathcal{N}\left(x; 0, \sigma^2\right) \left[1 + \sum_{s=1}^{b} d_s \mathcal{H}_s(x)\right]^2 \frac{1}{\kappa}, \tag{2.54}$$

where $\kappa = 1 + \sum_{s=1}^{b} d_s^2 s!$. Indeed, this specification of the Gram-Charlier density can lead to a separable form of the log-likelihood function, which can be computed more efficiently,

$$\log f(x) = \log \mathcal{N}\left(x; 0, \sigma^2\right) + \log \left[1 + \sum_{s=1}^{b} d_s \mathcal{H}_s(x)\right]^2 - \log \kappa.$$

However, in the multivariate case, the situation becomes more complicated because of the correlations among variables and more complicated function of Hermite polynomials. Perote and del Brío (2006) specified a restricted form of multivariate Gram-Charlier density to ensure positive density rather than imposing restrictions on the parameter d_s's. In fact, an m-variate Gram-Charlier density for a random vector $\mathbf{X}$ can be specified in the following two forms:

$$\begin{aligned}
&\mathcal{F}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{d}) \\
= \;& \frac{1}{m+1}\left\{\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) + \prod_{i=1}^{m}\mathcal{N}\left(x_i;\mu_i,\sigma_i^2\right)\sum_{i=1}^{m}\frac{1}{\kappa_i}\left[1+\sum_{s=1}^{b}d_{si}\mathcal{H}_s\left(x_i-\mu_i\right)\right]^2\right\},
\end{aligned} \tag{2.55}$$

or

$$\begin{aligned}
&\mathcal{F}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{d}) \\
= \;& \frac{1}{m+1}\left\{\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) + \prod_{i=1}^{m}\mathcal{N}\left(x_i;\mu_i,\sigma_i^2\right)\sum_{i=1}^{m}\frac{1}{c_i}\left[\sum_{s=1}^{b}d_{si}\mathcal{H}_s\left(x_i-\mu_i\right)\right]^2\right\},
\end{aligned} \tag{2.56}$$

where x_i, μ_i and σ_i^2 represent the ith element of $\mathbf{x}$, the ith element of $\boldsymbol{\mu}$ and the ith diagonal element of $\boldsymbol{\Sigma}$ respectively, $\kappa_i = 1 + \sum_{s=1}^{b} d_{si}^2 s!$ and $c_i = \kappa_i - 1$. The specifications (2.55) and (2.56) are different from those specified by Perote and del Brío (2006) in which $\mathcal{F}(\mathbf{x})$ was a density of random vector $\mathbf{X}$ with mean zero while the density with non-zero mean random vector $\mathbf{X}$ is specified here. Note that both specifications (2.55) and (2.56) cannot be reduced to (2.54) by putting $m = 1$ directly. Nevertheless, as shown in Perote and del Brío (2006), one sound feature of (2.55) and (2.56) is that the marginal density of a variate in $\mathbf{X}$ in (2.55) and (2.56) can also be expressed in the univariate versions of (2.55) and (2.56) respectively.

Returning to the current situation, the ensemble mean and variance for possible non-Gassianities of $p(\mathbf{x}_n|\mathcal{Y}_{n-1})$ and $p(\mathbf{y}_n|\mathcal{Y}_{n-1})$ are estimated by the maxi-

mization of the likelihood for (2.55) or (2.56) rather than using the simple averages in (2.10), (2.11), (2.13) and (2.14). Focusing on the multivariate Gram-Charlier density (2.55), the ensemble mean and variance of state vector $\mathbf{x}_n$ given $\mathcal{Y}_{n-1}$ are estimated by the maximization of the function

$$\left(\widetilde{\mathbf{x}}_{n|n-1}, \widetilde{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1}, \widetilde{\mathbf{d}}_{\mathbf{x},n|n-1}\right) = \underset{\mathbf{x},\mathbf{\Sigma},\mathbf{d}}{\operatorname{argmax}} \prod_{k=1}^{K} \mathcal{F}(\widehat{\mathbf{x}}^{(k)}_{n|n-1}; \mathbf{x}, \mathbf{\Sigma}, \mathbf{d}),$$

where $\mathbf{d} = (d_{11}, \ldots, d_{1m_x}, \ldots, d_{s1}, \ldots, d_{sm_x}, \ldots, d_{bm_x}, \ldots, d_{bm_x})$. Note that the orders of Hermite polynomials are restricted to b in all cases for simplicity although this restriction is not essential.

Similar operation can be applied to $p(\mathbf{y}_n|\mathcal{Y}_{n-1})$ as well, that is, the ensemble mean and variance of measurement vector $\mathbf{y}_n$ conditional on $\mathcal{Y}_{n-1}$ are obtained by the maximization of the following likelihood function:

$$\left(\widetilde{\mathbf{y}}_{n|n-1}, \widetilde{\mathbf{\Sigma}}^{\mathbf{yy}}_{n|n-1}, \widetilde{\mathbf{d}}_{\mathbf{y},n|n-1}\right) = \underset{\mathbf{y},\mathbf{\Sigma},\mathbf{d}}{\operatorname{argmax}} \prod_{k=1}^{K} \mathcal{F}(\widehat{\mathbf{y}}^{(k)}_{n|n-1}; \mathbf{y}, \mathbf{\Sigma}, \mathbf{d}),$$

where the dimension of $\mathbf{d}$ is conformable with that of $\mathbf{y}$.

In the case of Gaussian density, the ensemble members of filtering density $p(\mathbf{x}_n|\mathcal{Y}_n)$ can be obtained by the generalized least squares method. Now, the ensemble members of filtering density $p(\mathbf{x}_n|\mathcal{Y}_n)$ in the current situation are obtained by maximum a posteriori (MAP) (Sorenson, 1980, p.199).

$$\begin{aligned}
\widehat{\mathbf{x}}^{(k)}_{n|n} &= \underset{\mathbf{x}_n \in \mathbb{R}^{m_x}}{\operatorname{argmax}} p(\mathbf{x}_n|\mathcal{Y}_n) \\
&= \underset{\mathbf{x}_n \in \mathbb{R}^{m_x}}{\operatorname{argmax}} \left[p(\mathbf{y}_n|\mathbf{x}_n)\, p(\mathbf{x}_n|\mathcal{Y}_{n-1})\right] \\
&= \underset{\widehat{\mathbf{x}}^{(k)}_n \in \mathbb{R}^{m_x}}{\operatorname{argmax}} \mathcal{N}(\mathbf{y}^{(k)}_n - \mathbf{h}(\widehat{\mathbf{x}}^{(k)}_n; \boldsymbol{\xi}_n); \mathbf{0}, \mathbf{R}_n) \mathcal{F}(\widehat{\mathbf{x}}^{(k)}_n; \widetilde{\mathbf{x}}^{(k)}_{n|n-1}, \widetilde{\mathbf{\Sigma}}^{\mathbf{xx}}_{n|n-1}, \widetilde{\mathbf{d}}_{\mathbf{x},n|n-1}).
\end{aligned}$$

Similar to the Gaussian case, the ensemble of filtered state $\widehat{\mathbf{x}}_{n|n}^{(k)}$ can be expressed as a function of $\mathbf{y}_n^{(k)}$ and $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ although the formulation becomes more complicated. Then, the ensemble mean and variance for filtering density can be found by the maximization of

$$\left(\widetilde{\mathbf{x}}_{n|n}, \widetilde{\boldsymbol{\Sigma}}_{n|n}^{\mathbf{xx}}, \widetilde{\mathbf{d}}_{\mathbf{x},n|n}\right) = \underset{\mathbf{x},\boldsymbol{\Sigma},\mathbf{d}}{\operatorname{argmax}} \prod_{k=1}^{K} \mathcal{F}(\widehat{\mathbf{x}}_{n|n}^{(k)}; \mathbf{x}, \boldsymbol{\Sigma}, \mathbf{d}).$$

Furthermore, the smoothed state vector $\mathbf{x}_n$ and measurement vector $\mathbf{y}_n$ can be derived in a similar sense because they relied on the derivation of the prediction densities of the state vector $\mathbf{x}_n$ and the measurement vectors $\mathbf{y}_n, \ldots, \mathbf{y}_N$. Together with the conditional density $p(\mathbf{y}_n|\mathbf{x}_n)$, the smoothed estimates $\widehat{\mathbf{y}}_{n|N}$ and $\widehat{\mathbf{x}}_{n|N}$ can be computed easily.

For parameter estimation, the likelihood function is no longer expressed in terms of approximated Gaussian densities. Instead, the unknown parameters are estimated by maximizing the following likelihood function:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmax}} \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right) \\
&= \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathcal{Y}_{n-1}\right) \\
&= \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{n=1}^{N} \mathcal{F}(\mathbf{y}_n; \widetilde{\mathbf{y}}_{n|n-1}, \widetilde{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{yy}}, \widetilde{\mathbf{d}}_{\mathbf{y},n|n-1}).
\end{aligned}$$

2.7 Empirical Applications

In this section, the assimilation of data by EnGTKF is focused on the time-varying vector smooth transition autoregressive model (TV-VSTAR) which was

a multivariate extension of Lundbergh et al. (2003) but a simplified version of He et al. (2005). Here, the case of a two-regime transition function and the same transition function for all measurements in $\mathbf{y}_n$ is considered. Specifically, we can consider the following p-order two-regime TV-VSTAR model for the measurement vector $\mathbf{y}_n$.

$$\begin{aligned}\mathbf{y}_n &= \left[\left(1-G_1\left(\mathbf{y}_{n-d}\right)\right)\sum_{j=1}^{p}\mathbf{\Phi}_{1,j}\mathbf{y}_{n-j}+G_1\left(\mathbf{y}_{n-d}\right)\sum_{j=1}^{p}\mathbf{\Phi}_{2,j}\mathbf{y}_{n-j}\right]\left(1-G_2\left(n\right)\right)\\ &\quad\left[\left(1-G_1\left(\mathbf{y}_{n-d}\right)\right)\sum_{j=1}^{p}\mathbf{\Phi}_{3,j}\mathbf{y}_{n-j}+G_1\left(\mathbf{y}_{n-d}\right)\sum_{j=1}^{p}\mathbf{\Phi}_{4,j}\mathbf{y}_{n-j}\right]G_2\left(n\right)\\ &\quad+\mathbf{v}_n, \end{aligned} \tag{2.57}$$

for $n=1,\ldots,N$, where $\mathbf{v}_n \sim N\left(\mathbf{0},\mathbf{R}\right)$, $\mathbf{y}_n = \left(y_{1,n},\ldots,y_{m_y,n}\right)^{\mathsf{T}}$ is a $(m_y \times 1)$ column vector, $\mathbf{\Phi}_{r,s}$'s are $(m_y \times m_y)$ coefficient matrices for $r=1,\ldots,4$ and $s=1,\ldots,p$, transition functions $G_1\left(\mathbf{y}_{n-d}\right)$ and $G_2\left(n\right)$ are defined as $G_1\left(\mathbf{y}_{n-d}\right) = \left[1+\exp\left\{-\gamma_1\prod_{i=1}^{m_y}\left(y_{i,n-d}-c_{1,i}\right)\right\}\right]^{-1}$ and $G_2\left(n\right) = \left[1+\exp\left\{-\gamma_2\left(n-c_2\right)\right\}\right]^{-1}$ respectively with $\gamma_1,\gamma_2 > 0$. Note that both of them satisfy the condition $0 \leq G_1\left(\cdot\right), G_2\left(\cdot\right) \leq 1$. Here, the model (2.57) is denoted as VSTAR(p,d). Without loss of generality, assume that $1 \leq d \leq p$. In state space form, the measurement and transition equations for (2.57) can be rewritten as

$$\mathbf{y}_n = \left[\begin{array}{cccc}\mathbf{I}_{m_y} & \mathbf{0} & \cdots & \mathbf{0}\end{array}\right]\mathbf{x}_n + \mathbf{v}_n, \tag{2.58a}$$

$$\mathbf{x}_n = \begin{bmatrix} \mathbf{y}_n \\ \mathbf{y}_{n-1} \\ \vdots \\ \mathbf{y}_{n-p+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi}_1 & \cdots & \boldsymbol{\Psi}_{p-1} & \boldsymbol{\Psi}_p \\ \mathbf{I}_{m_y} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{I}_{m_y} & \mathbf{0} \end{bmatrix} \mathbf{x}_{n-1} + \begin{bmatrix} \boldsymbol{\eta}_n \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \quad (2.58b)$$

for $n = 1, \ldots, N$, where $\boldsymbol{\eta}_n \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{\eta})$,

$$\boldsymbol{\Psi}_i = (1 - g_{2,n}) \left[(1 - g_{1,n}) \boldsymbol{\Phi}_{1,i} + g_{1,n} \boldsymbol{\Phi}_{2,i} \right] + g_{2,n} \left[(1 - g_{1,n}) \boldsymbol{\Phi}_{3,i} + g_{1,n} \boldsymbol{\Phi}_{4,i} \right],$$

for $i = 1, \ldots, p$, $g_{1,n} = G_1(\mathbf{y}_{n-d})$ and $g_{2,n} = G_2(n)$. After reparameterization, the vector autoregressive coefficient $\boldsymbol{\Psi}_i$ consists of the time-invariant component

$$(1 - g_{1,n}) \boldsymbol{\Phi}_{1,i} + g_{1,n} \boldsymbol{\Phi}_{2,i},$$

and the time-varying component

$$g_{2,n} \left[(1 - g_{1,n}) (\boldsymbol{\Phi}_{3,i} - \boldsymbol{\Phi}_{1,i}) + g_{1,n} (\boldsymbol{\Phi}_{4,i} - \boldsymbol{\Phi}_{2,i}) \right].$$

According to (2.57) and (2.58), the nonlinearity of the model arises from the randomness of coefficient matrices $\boldsymbol{\Psi}_i$'s which are determined by the transition functions $G_1(\mathbf{y}_{n-d})$ and $G_2(n)$. These two functions share the same property that they are a type of logistic curve and always transit between zero and one. Focusing on $G_1(\mathbf{y}_{n-d})$, when the level of measurement $\mathbf{y}_{n-d}$ falls below the threshold level $\mathbf{c}_1 \equiv (c_{1,1}, c_{1,2})$, the function $G_1(\mathbf{y}_{n-d})$ moves between 0 and 0.5. This implies that the measurement $\mathbf{y}_n$ moves between the regimes with coefficients $\boldsymbol{\Phi}_{1,i}$'s and $\boldsymbol{\Phi}_{3,i}$'s more probably. On the other hand, if the measurement $\mathbf{y}_{n-d}$ jumps above

the threshold level $\mathbf{c}_1$, the function $G_1(\mathbf{y}_{n-d})$ moves between 0.5 and 1. the measurement $\mathbf{y}_n$ moves between the regimes with coefficients $\mathbf{\Phi}_{2,i}$'s and $\mathbf{\Phi}_{4,i}$'s more probably. Similar interpretation is applied to the function $G_2(n)$. It allows the measurement $\mathbf{y}_n$ to move between the regimes with the coefficients $\mathbf{\Phi}_{1,i}$'s and $\mathbf{\Phi}_{2,i}$'s more probably when $n < c_2$. When $c > c_2$, the measurement $\mathbf{y}_n$ moves between the regimes with the coefficients $\mathbf{\Phi}_{3,i}$'s and $\mathbf{\Phi}_{4,i}$'s more probably. The purpose of the logistic function for $G_1(\cdot)$ and $G_2(\cdot)$ is to allow the smooth transition between regimes.

When both $G_1(\cdot)$ and $G_2(\cdot)$ are considered together, the situation becomes more complicated. At the beginning of the time series, the measurement $\mathbf{y}_n$ transits between the regimes with coefficients $\mathbf{\Phi}_{1,i}$'s and $\mathbf{\Phi}_{2,i}$'s more probably and the level of $\mathbf{y}_{n-d}$ determine whether the coefficients $\mathbf{\Phi}_{1,i}$'s or $\mathbf{\Phi}_{2,i}$'s play the role of movement. As the time goes by and after time $n = c_2$, the movement of measurement $\mathbf{y}_n$ is determined by $\mathbf{\Phi}_{3,i}$'s and $\mathbf{\Phi}_{4,i}$'s more probably. The smooth transition between coefficient $\mathbf{\Phi}_{3,i}$'s and $\mathbf{\Phi}_{4,i}$'s determined by the level of $\mathbf{y}_{n-d}$. The speed of transition is determined by the parameters γ_1 and γ_2. When both of them are large, the model coefficients can transit between regimes very quickly. In extreme cases, when γ_1 and γ_2 become infinite, the model coefficients switch between regimes instantaneously and the model (2.57) is reduced to the case of self-excited threshold autoregressive (SETAR) model by Tong (1983). On the other hand, when both $G_1(\cdot)$ and $G_2(\cdot)$ are constant, (2.57) is reduced to standard case of vector autoregressive model.

In vector form, the unknown parameters of (2.58) in the corresponding likelihood function are represented as

$$\boldsymbol{\theta} = \left[\mathbf{x}_{0|0}^{\mathsf{T}}, \mathrm{vech}(\boldsymbol{\Sigma}_{0|0})^{\mathsf{T}}, \boldsymbol{\xi}^{\mathsf{T}}, \mathrm{vech}\,(\mathbf{R})^{\mathsf{T}}, \mathrm{vech}(\boldsymbol{\Sigma}^{\eta})^{\mathsf{T}}\right],$$

where $\boldsymbol{\xi}^{\mathsf{T}} = \left[\mathrm{vec}(\boldsymbol{\Phi}_{1,1})^{\mathsf{T}}, \ldots, \mathrm{vec}(\boldsymbol{\Phi}_{4,p})^{\mathsf{T}}, \gamma_1, \gamma_2, c_{1,1}, \ldots, c_{1,m_y}, c_2\right]$. These unknown parameters are estimated by maximization of (2.30) while their standard errors are computed by (2.50) in which the derivatives of $\mathbf{h}\,(\cdot)$ and $\mathbf{f}\,(\cdot)$ with respect to $\boldsymbol{\xi}^{\mathsf{T}}$ are required. Obviously,

$$\frac{\partial \mathbf{h}\,(\mathbf{x}_n; \boldsymbol{\xi})}{\partial \boldsymbol{\xi}^{\mathsf{T}}} = \mathbf{0},$$

for $n = 1, \ldots, N$. On the other hand, the function $\mathbf{f}\,(\mathbf{x}_{n-1}; \boldsymbol{\xi})$ for the transition equation in (2.58) can be written as

$$\mathbf{f}\,(\mathbf{x}_{n-1}; \boldsymbol{\xi}) = \begin{bmatrix} \sum_{i=1}^{p} \boldsymbol{\Psi}_i \mathbf{y}_{n-i} \\ \mathbf{y}_{n-2} \\ \vdots \\ \mathbf{y}_{n-p} \end{bmatrix}.$$

Then, the derivative of $\mathbf{f}\,(\mathbf{x}_{n-1}; \boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}^{\mathsf{T}}$ is partitioned into two parts, namely,

$$\frac{\partial \mathbf{f}\,(\mathbf{x}_{n-1}; \boldsymbol{\xi})}{\partial \left[\mathrm{vec}(\boldsymbol{\Phi}_{1,1})^{\mathsf{T}}, \ldots, \mathrm{vec}(\boldsymbol{\Phi}_{1,p})^{\mathsf{T}}, \ldots, \mathrm{vec}(\boldsymbol{\Phi}_{4,1})^{\mathsf{T}}, \ldots, \mathrm{vec}(\boldsymbol{\Phi}_{4,p})^{\mathsf{T}}\right]}$$

$$= \begin{bmatrix} \frac{\partial \boldsymbol{\Psi}_1 \mathbf{y}_{n-1}}{\partial \mathrm{vec}(\boldsymbol{\Phi}_{1,1})^{\mathsf{T}}} & \cdots & \frac{\partial \boldsymbol{\Psi}_p \mathbf{y}_{n-p}}{\partial \mathrm{vec}(\boldsymbol{\Phi}_{1,p})^{\mathsf{T}}} & \cdots & \frac{\partial \boldsymbol{\Psi}_1 \mathbf{y}_{n-1}}{\partial \mathrm{vec}(\boldsymbol{\Phi}_{4,1})^{\mathsf{T}}} & \cdots & \frac{\partial \boldsymbol{\Psi}_p \mathbf{y}_{n-p}}{\partial \mathrm{vec}(\boldsymbol{\Phi}_{4,p})^{\mathsf{T}}} \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \vdots & & \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix},$$

and

$$\frac{\partial \mathbf{f}\left(\mathbf{x}_{n-1};\boldsymbol{\xi}\right)}{\partial\left[\gamma_1,\gamma_2,c_{1,1},\ldots,c_{1,m_y},c_2\right]}$$

$$=\begin{bmatrix} \sum_{i=1}^{p}\frac{\partial \boldsymbol{\Psi}_i}{\partial \gamma_1}\mathbf{y}_{n-i} & \sum_{i=1}^{p}\frac{\partial \boldsymbol{\Psi}_i}{\partial \gamma_2}\mathbf{y}_{n-i} & \sum_{i=1}^{p}\frac{\partial \boldsymbol{\Psi}_i}{\partial c_{1,1}}\mathbf{y}_{n-i} & \cdots & \sum_{i=1}^{p}\frac{\partial \boldsymbol{\Psi}_i}{\partial c_{1,m_y}}\mathbf{y}_{n-i} & \sum_{i=1}^{p}\frac{\partial \boldsymbol{\Psi}_i}{\partial c_2}\mathbf{y}_{n-i} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where for $i = 1, \ldots, p$,

$$\begin{aligned}
\frac{\partial \boldsymbol{\Psi}_i \mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\Phi_{1,i}\right)^{\mathsf{T}}} &= \left(1-g_{1,n}\right)\left(1-g_{2,n}\right)\frac{\partial \boldsymbol{\Phi}_{1,i}\mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\boldsymbol{\Phi}_{1,i}\right)^{\mathsf{T}}}, \\
\frac{\partial \boldsymbol{\Psi}_i \mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\Phi_{2,i}\right)^{\mathsf{T}}} &= g_{1,n}\left(1-g_{2,n}\right)\frac{\partial \boldsymbol{\Phi}_{2,i}\mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\boldsymbol{\Phi}_{2,i}\right)^{\mathsf{T}}}, \\
\frac{\partial \boldsymbol{\Psi}_i \mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\Phi_{3,i}\right)^{\mathsf{T}}} &= \left(1-g_{1,n}\right)g_{2,n}\frac{\partial \boldsymbol{\Phi}_{3,i}\mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\boldsymbol{\Phi}_{3,i}\right)^{\mathsf{T}}}, \\
\frac{\partial \boldsymbol{\Psi}_i \mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\Phi_{4,i}\right)^{\mathsf{T}}} &= g_{1,n}g_{2,n}\frac{\partial \boldsymbol{\Phi}_{4,i}\mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\boldsymbol{\Phi}_{4,i}\right)^{\mathsf{T}}}, \\
\frac{\partial \boldsymbol{\Phi}_{j,i}\mathbf{y}_{n-i}}{\partial \mathrm{vec}\left(\boldsymbol{\Phi}_{j,i}\right)^{\mathsf{T}}} &= \mathbf{y}_{n-i}^{\mathsf{T}} \otimes \mathbf{I}_{m_y} \text{ for } j = 1,2,3,4, \\
\frac{\partial \boldsymbol{\Psi}_i}{\partial \gamma_1} &= \left[\left(1-g_{2,n}\right)\left(\boldsymbol{\Phi}_{2,i}-\boldsymbol{\Phi}_{1,i}\right)+g_{2,n}\left(\boldsymbol{\Phi}_{4,i}-\boldsymbol{\Phi}_{3,i}\right)\right]\left(\frac{\partial g_{1,n}}{\partial \gamma_1}\right), \\
\frac{\partial \boldsymbol{\Psi}_i}{\partial c_{1,j}} &= \left[\left(1-g_{2,n}\right)\left(\boldsymbol{\Phi}_{2,i}-\boldsymbol{\Phi}_{1,i}\right)+g_{2,n}\left(\boldsymbol{\Phi}_{4,i}-\boldsymbol{\Phi}_{3,i}\right)\right]\left(\frac{\partial g_{1,n}}{\partial c_{1,j}}\right) \\
&\qquad\qquad \text{for } j = 1,\ldots,m_y, \\
\frac{\partial \boldsymbol{\Psi}_i}{\partial \gamma_2} &= \left[\left(1-g_{1,n}\right)\left(\boldsymbol{\Phi}_{3,i}-\boldsymbol{\Phi}_{1,i}\right)+g_{1,n}\left(\boldsymbol{\Phi}_{4,i}-\boldsymbol{\Phi}_{2,i}\right)\right]\left(\frac{\partial g_{2,n}}{\partial \gamma_2}\right), \\
\frac{\partial \boldsymbol{\Psi}_i}{\partial c_2} &= \left[\left(1-g_{1,n}\right)\left(\boldsymbol{\Phi}_{3,i}-\boldsymbol{\Phi}_{1,i}\right)+g_{1,n}\left(\boldsymbol{\Phi}_{4,i}-\boldsymbol{\Phi}_{2,i}\right)\right]\left(\frac{\partial g_{2,n}}{\partial c_2}\right), \\
\frac{\partial g_{1,n}}{\partial \gamma_1} &= \frac{\prod_{k=1}^{m_y}\left(y_{k,n-d}-c_{1,k}\right)\exp\left\{-\gamma_1\prod_{k=1}^{m_y}\left(y_{k,n-d}-c_{1,k}\right)\right\}}{\left(1+\exp\left\{-\gamma_1\prod_{k=1}^{m_y}\left(y_{k,n-d}-c_{1,k}\right)\right\}\right)^2}, \\
\frac{\partial g_{1,n}}{\partial c_{1,j}} &= -\frac{\gamma_1\prod_{k\neq j}\left(y_{k,n-d}-c_{1,k}\right)\exp\left\{-\gamma_1\prod_{k=1}^{m_y}\left(y_{k,n-d}-c_{1,k}\right)\right\}}{\left(1+\exp\left\{-\gamma_1\prod_{k=1}^{m_y}\left(y_{k,n-d}-c_{1,k}\right)\right\}\right)^2}
\end{aligned}$$

for $j = 1, \ldots, m_y$,

$$\begin{aligned}
\frac{\partial g_{2,n}}{\partial \gamma_2} &= \frac{(n - c_2) \exp\left\{-\gamma_2 (n - c_2)\right\}}{\left(1 + \exp\left\{-\gamma_2 (n - c_2)\right\}\right)^2}, \\
\frac{\partial g_{2,n}}{\partial c_2} &= -\frac{\gamma_2 \exp\left\{-\gamma_2 (n - c_2)\right\}}{\left(1 + \exp\left\{-\gamma_2 (n - c_2)\right\}\right)^2}.
\end{aligned}$$

Before the illustration of the assimilation of empirical data by the TV-VSTAR model, numerical simulations, which show the validity of the EnGTKF filter, are given in the following sections.

2.7.1 Numerical Simulations

2.7.1.1 Vector Autoregressive Model

In state space form, the observations in the vector autoregressive (VAR) model of order 1 are generated in the following form:

$$\begin{aligned}
\mathbf{y}_n &= \mathbf{x}_n + \mathbf{v}_n, \\
\mathbf{x}_n &= \mathbf{\Phi} \mathbf{x}_{n-1} + \boldsymbol{\eta}_n.
\end{aligned}$$

for $n = 1, \ldots, 100$, where $\mathbf{y}_n = (y_{1,n}, y_{2,n})^{\mathsf{T}}$ is a (2×1) column vector, $\mathbf{\Phi}$ is a coefficient matrix of order (2×2). The values of parameters are specified below:

$$\begin{aligned}
\mathbf{\Phi} &= \begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.7 \end{bmatrix}, \ \mathbf{x}_0 \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right), \\
\mathbf{v}_n &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.5 \end{bmatrix}\right), \ \boldsymbol{\eta}_n \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.6 & -0.2 \\ -0.2 & 1.8 \end{bmatrix}\right).
\end{aligned}$$

From Figure 2.1, one can see that the prediction, filtering and smoothing by the EnGTKF perform well over the sample. The main reason is attributed to the linear feature of the state space model. Then, the prediction and filtering densities of state and measurement variables do not deviate from the Gaussian density. Hence, consistent estimates of the ensemble mean and variance can be achieved easily. From Figure 2.2, the interval estimates of measurement variables are displayed as well. The confidence limits are approximated by the asymptotic Gaussianity due to the ensemble size. Obviously, due to the asymptotic normality of estimated ensemble means and stable estimates of ensemble variances, the widths of confidence intervals do not change over the sample dramatically. Furthermore, the confidence limits of predicted model states are wider than those of filtered and smoothed model states. Intuitively, the filtered and smoothed estimates are obtained by using more information and then more accurate estimates are obtained.

2.7.1.2 Time-varying Vector Smooth Transition Autoregressive Model

The lags chosen for simulation of model (2.57) are $m_y = 2$, $p = 1$ and $d = 1$, that is, a bivariate VSTAR(1,1) model. The detailed specification of parameters are given below:

$$\begin{aligned}
\boldsymbol{\Phi}_{1,1} &= \begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.7 \end{bmatrix}, \ \boldsymbol{\Phi}_{2,1} = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.4 \end{bmatrix}, \\
\boldsymbol{\Phi}_{3,1} &= \begin{bmatrix} 0.6 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}, \ \boldsymbol{\Phi}_{4,1} = \begin{bmatrix} 0.2 & -0.01 \\ -0.01 & 0.6 \end{bmatrix},
\end{aligned}$$

$$\mathbf{x}_0 \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right), \ \gamma_1 = 8.7, \gamma_2 = 2.7, \mathbf{c}_1 = (-0.3, -0.5), c_2 = 20,$$

$$\mathbf{v}_n \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.5 \end{bmatrix}\right), \ \boldsymbol{\eta}_n \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.6 & -0.2 \\ -0.2 & 1.8 \end{bmatrix}\right).$$

for $n = 1, \ldots, 100$.

Unlike the previous example, although the measurement equation is still linear in this case, the transition equation is not linear any more. Indeed, the coefficients in the transition equation vary with the level of measurement $\mathbf{y}_n$. The prediction and filtering results in Figure 2.3 show that the predicted and filtered measurement variables move along with actual values of measurement. Indeed, the state vector $\mathbf{x}_n$ is updated efficiently as shown in Figures 2.3(b) and (e). This can be explained by the similar reason as in previous example. Furthermore, the confidence intervals in Figure 2.4 show the performance of EnGTKF under this model is similar to that under previous model.

2.7.1.3 A Nonlinear Gaussian Simulation Model by Kitagawa

Here, Gaussian means that the measurement and state errors are Gaussian. The model being studied in this section was originated in Netto et al. (1978) and then studied in Kitagawa (1987, 1991, 1996, 1998) for smoothers. Indeed, the model is given by

$$\begin{aligned} y_n &= \frac{x_n^2}{50} + v_n, \\ x_n &= 0.5x_{n-1} + \frac{25x_{n-1}}{1 + x_{n-1}^2} + 8\cos(1.2n) + u_n, \end{aligned}$$

where $v_n \sim N(0,1)$, $u_n \sim N(0,1)$, $x_0 \sim N(0,1)$ for $n = 1, \ldots, 200$.

According to Figure 2.5(a), the simulated measurement y_n fluctuates between 0 and 20 and below 0 occasionally because of the term x_n^2 and the noise v_n in the measurement equation. For the simulated state x_n in Figure 2.5(b), the values always fall within the range -20 and 20. Due to the presence of the deterministic cosine function in the transition equation, together with the stochastic component of the equation, the state x_n proceeds periodically over a number of time points above zero and continues the oscillations below zero.

The prediction and filtering results by the conventional EnKF are shown in Figures 2.5(d) to (f). Several unexpected spikes of prediction and filtering of measurement and state variables are observed in the diagrams. These estimates can be explained by the fact that the prediction and filtering densities of state x_n deviate from Gaussian. The same reason can also be applied to the case of measurement y_n. A pick of observation at $n = 50$ shows their multimodal conditional densities. In Figures 2.6(a) to (c), the prediction density of y_n is skewed to the right. On the other hand, when the prediction and filtering densities of measurement and state variables at $n = 100$ are investigated in Figures 2.6(d) and (e), the situation is much better than the previous case. The prediction density of y_n is not skewed to the right although certain bimodal feature is observed near the tail of the density. As a result, the conventional EnKF cannot provide a satisfactory result of the predicted mean of x_n and y_n and the method discussed in Section 2.6 is attempted here, that is, the Gram-Charlier density is used to approximate these densities.

To estimate the ensemble mean and variance Gram-Charlier density, one should choose the order of the Hermite polynomial. Here, the order is chosen

by the one with minimum RMSE in the prediction stage and the results are shown in Table 2.1.

Table 2.1: RMSEs of predicted x_n and y_n under various orders of Hermite polynomial, b

Order	y_n	x_n
1	5.2239	9.3915
2	5.6118	9.1596
3	6.4111	10.6886
4	6.9949	11.4211
5	7.8988	12.2895
6	7.6440	11.5733

From Table 2.1, the RMSE of y_n increases with the order of Hermite polynomial while that of x_n is minimized at order $b = 2$. Hence, the diagrams in Figures 2.5(a) to (c) are referred to the case of order $b = 2$. From Figure 2.5, the spikes in the predictions of the measurement y_n and the state variable x_n disappear. However, the prediction and filtering results of the state variable x_n are not totally satisfactory. Two possible reasons are suggested here. Firstly, the periodical term $8\cos(1.2n)$ induces the process x_n to be periodic and hence non-ergodic. Then, Assumption 2.1(iii) is violated and the performance of EnKF is accountable in Figure 2.5. Secondly, the method discussed in Section 2.6 can solve the problem in the conditional densities of y_n and x_n partially. From Table 2.1, the RMSEs of predicted y_n and x_n increase quickly when the order of Hermite polynomial increases. This is possibly due to the divergence feature of the Gram-Charlier density with high order of Hermite polynomial, as demonstrated by Blinnikov and Moessner (1998). Nevertheless, since the prediction of the measurement variables is more important than that of the state variables in many real applications, under this simulation model, the EnGTKF still outperforms

the conventional EnKF marginally.

2.7.2 Algal Bloom Data

To demonstrate the use of EnGTKF with empirical application, a bivariate model of algal bloom data is assimilated by the TV-VSTAR model in (2.57). The TV-VSTAR model is chosen due to its capability of capturing the asymmetric cyclical fluctuations during bloom and non-bloom periods. In this application, the bivariate model contains dependent variables of standardized chlorophyll fluorescence (CHL) and standardized dissolved oxygen concentration (DO) which were obtained from the algal bloom dynamics field monitoring station of the University of Hong Kong at Kat O, Hong Kong. The full sampling period covered between 2000 and 2004. The detailed data description was provided by Lee et al. (2003, 2004).

In this empirical application, the daily observations during 2000-2001 are selected to perform the in-sample estimation. Before the estimation of the model parameters, the lag parameters p and d in the TV-VSTAR model should be identified firstly. A common approach is the usage of the Akaike information criterion (AIC) and the lag parameters p and d are selected with the minimum AIC. Indeed, the AIC is defined by

$$AIC = -2\log \mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_N) + 2\dim(\widehat{\boldsymbol{\theta}}),$$

where $\dim(\widehat{\boldsymbol{\theta}})$ represents the number of parameters in the model.

The results of AICs of the models are presented in Table 2.2. Clearly, TV-VSTAR(1,1) model is chosen among the model candidates, with AIC=14318. On

the other hand, VAR(1) model is chosen with AIC=14368. When the RMSEs of both models are compared, the prediction of CHL by the TV-VSTAR model is outerperformed by the VAR model except when $p = 1$ and 3. Furthermore, the prediction of DO by the TV-VSTAR model improves that by the VAR model generally. To highlight the results of VSTAR(1,1) model, the RMSEs of CHL and DO are 0.7174 and 0.5488 respectively and they are smaller than those of VAR(1) model. From Table 2.3, the in-sample predictions of both measurements by the univariate versions of TV-VSTAR and VAR models demonstrate the domination of their results over those by multivariate models.

Table 2.2: In-sample performance of TV-VSTAR and VAR models where the columns of CHL and represent the RMSEs of CHL and DO respectively. The columns of AIC represent the AICs of corresponding models.

		TV-VSTAR			VAR		
p	d	CHL	DO	AIC	CHL	DO	AIC
1	1	0.7174	0.5488	14318	0.7713	0.6036	14368
2	1	1.5319	0.5466	47245	0.6952	0.6127	14761
2	2	0.7110	0.5921	51778	0.6952	0.6127	14761
3	1	0.7614	0.6332	34521	0.8117	0.7173	42484
3	2	0.8070	0.7277	45472	0.8117	0.7173	42484
3	3	0.8168	0.7323	47360	0.8117	0.7173	42484

After the determination of the lag parameters of TV-VSTAR model, the parameter estimates and its standard errors are given in Table 2.4.

Table 2.3: In-sample performance of TV-VSTAR and VAR models (univariate) where the columns of CHL and DO represent the RMSEs of CHL and DO respectively.

		TV-VSTAR		VAR	
p	d	CHL	DO	CHL	DO
1	1	0.7044	0.5473	0.7329	0.5595
2	1	0.6982	0.5796	0.7116	0.5410
2	2	0.6982	0.5910	0.7116	0.5410
3	1	0.7015	0.5621	0.7016	0.6180
3	2	0.7015	0.5621	0.7016	0.6180
3	3	0.7016	0.5620	0.7016	0.6180

Table 2.4: Parameter estimates of VSTAR(1,1) model by EnGTKF for CHL and DO where the numbers insides the brackets represent the corresponding elements within the matrices or vectors. The p-values are approximated by the asymptotic Gaussianity.

Parameters	Estimates	Standard Errors	p-values
$\mathbf{x}_{0\|0}(1)$	0.0710	0.0332	0.0323
$\mathbf{x}_{0\|0}(2)$	−0.1093	0.0650	0.0925
$\mathbf{\Sigma}_{0\|0}(1,1)$	0.2212	0.1167	0.0581
$\mathbf{\Sigma}_{0\|0}(2,1)$	−0.3935	0.2582	0.1275
$\mathbf{\Sigma}_{0\|0}(2,2)$	0.7435	0.4353	0.0877
$\mathbf{\Phi}_{1,1}(1,1)$	0.6976	0.3297	0.0343
$\mathbf{\Phi}_{1,1}(1,2)$	0.1716	0.1107	0.1211
$\mathbf{\Phi}_{1,1}(2,1)$	−0.1011	0.0839	0.2282
$\mathbf{\Phi}_{1,1}(2,2)$	0.8898	0.3725	0.0169
$\mathbf{\Phi}_{2,1}(1,1)$	0.7102	0.3241	0.0285

Continued on next page

Continued from previous page

Parameters	Estimates	Standard Errors	p-values
$\mathbf{\Phi}_{2,1}(1,2)$	-0.0929	0.0633	0.1420
$\mathbf{\Phi}_{2,1}(2,1)$	-0.2196	0.0948	0.0205
$\mathbf{\Phi}_{2,1}(2,2)$	0.9011	0.4418	0.0414
$\mathbf{\Phi}_{3,1}(1,1)$	0.7861	0.3277	0.0164
$\mathbf{\Phi}_{3,1}(1,2)$	0.0603	0.0434	0.1650
$\mathbf{\Phi}_{3,1}(2,1)$	0.0579	0.0319	0.0699
$\mathbf{\Phi}_{3,1}(2,2)$	0.9999	0.5415	0.0648
$\mathbf{\Phi}_{4,1}(1,1)$	0.6165	0.2850	0.0305
$\mathbf{\Phi}_{4,1}(1,2)$	-0.0333	0.0217	0.1246
$\mathbf{\Phi}_{4,1}(2,1)$	0.3788	0.2175	0.0816
$\mathbf{\Phi}_{4,1}(2,2)$	0.7682	0.3298	0.0198
$\mathbf{R}(1,1)$	0.1732	0.1006	0.0849
$\mathbf{R}(2,1)$	-0.3283	0.2170	0.1303
$\mathbf{R}(2,2)$	0.6498	0.5080	0.2009
$\mathbf{\Sigma}^{\eta}(1,1)$	0.2999	0.1357	0.0271
$\mathbf{\Sigma}^{\eta}(2,1)$	-0.6075	0.2912	0.0369
$\mathbf{\Sigma}^{\eta}(2,2)$	1.2306	0.9330	0.1872
γ_1	19.8845	11.2258	0.0765
γ_2	20.1574	11.8266	0.0883
$\mathbf{c}_1(1)$	0.0098	0.0075	0.1898
$\mathbf{c}_1(2)$	-0.0236	0.0141	0.0936
c_2	10.1231	4.5180	0.0251

According to Table 2.4, almost all parameters are significantly different from zero. When the parameter estimate of $\mathbf{c}_1$ is investigated, the values of 0.0098 and -0.0236 correspond to the threshold level of standardized CHL and DO at which the transition between bloom and non-bloom periods occurred on average respectively. Furthermore, the time-varying transition function indicates that a 10-day cycle is observed for CHL and DO to change coefficients on average.

To see the in-sample prediction performance of the TV-VSTAR model, the first 100 in-sample estimation of CHL and DO are presented in Figure 2.7. Refer to Figure 2.7(a), the prediction of CHL is reasonably well including the peaks at $n = 6$, $n = 33$ and $n = 66$. When the filtering result of CHL is considered, the filtered values of CHL are sufficiently close to the actual values. From Figure 2.7(c), the smoothers of CHL and DO perform well in a sense that the actual movements of both variables are reflected in their smoothed values. In Figure 2.8, the estimated confidence intervals in prediction, filtering and smoothing stages shows that the performance of EnGTKF is fairly stable for both CHL and DO.

As mentioned in Lee et al. (2003, 2004) and Muttil et al. (2004), algal blooms occurred frequently during July 2001 and October 2001. This also initiates the study of the predictability and filtering of TV-VSTAR model by EnGTKF for this period. The estimated CHL and DO and their interval estimates are displayed in Figures 2.9 and 2.10 respectively. The RMSEs of standardized CHL and DO within this period are 0.6259 and 0.6385 respectively and they are comparable with the results in Lee et al. (2004) and Muttil et al. (2004). In fact, the RMSE of CHL was 0.5345 in the paper of Lee et al. (2004) after standardization where the artificial neural network was used to perform the prediction. On the other hand, Muttil et al. (2004) produced the RMSE of 0.4793 for CHL after

standardization in which the genetic programming was applied to predict CHL. Nevertheless, one should emphasize that only a bivariate model with 2 lags for prediction is considered.

From these figures, the in-sample prediction by TV-VSTAR model seems to capture the algal blooms at $n = 421$ quite well. When the results of filtering are investigated, it seems that the observations can be assimilated very well even in the cases of spikes. Therefore, it is interesting to investigate the prediction and filtering densities of CHL and DO and they are shown in Figure 2.11 where the prediction and filtering density of CHL and DO at $n = 421$ and $n = 431$ are produced. From the figures, even if the prediction and filtering densities of CHL and DO deviate from the Gaussian density a bit, the assimilations of measurements CHL and DO are still successful. This also implies that the nonlinear updating equation is not very sensitive to the assumption of Gaussian density under current empirical application.

To assess the out-of-sample prediction performance of CHL and DO in 2002-2004, their RMSEs and AICs under various orders of the TV-VSTAR models are reported in Table 2.5.

Table 2.5: Out-of-sample performance of TV-VSTAR and VAR models in 2002-2004 where the columns of CHL and DO represent the RMSEs of CHL and DO respectively. The columns of AIC represent the AICs of corresponding models.

		TV-VSTAR			VAR		
p	d	CHL	DO	AIC	CHL	DO	AIC
1	1	0.3061	0.4638	4336	0.4846	0.4979	5787
2	1	0.3743	0.4605	25533	0.2447	0.5672	3675
2	2	0.3832	0.4587	35194	0.2447	0.5672	3675
3	1	0.4201	0.4859	22477	0.4573	0.4920	25661
3	2	0.4670	0.4993	31160	0.4573	0.4920	25661
3	3	0.4800	0.4991	33207	0.4573	0.4920	25661

Table 2.6: Out-of-sample RMSEs of TV-VSTAR and VAR models in 2002-2004 (univariate) where the columns of CHL and DO represent the RMSEs of CHL and DO respectively.

		TV-VSTAR		VAR	
p	d	CHL	DO	CHL	DO
1	1	0.5964	0.4617	0.4990	0.4690
2	1	0.3309	0.5060	0.4683	0.4627
2	2	0.3314	0.5222	0.4683	0.4627
3	1	0.3387	0.4567	0.3386	0.5021
3	2	0.3392	0.4567	0.3386	0.5021
3	3	0.3392	0.4569	0.3386	0.5021

From Table 2.5, the performance of the VSTAR(1,1) model is still better than that of the other TV-VSTAR models and this is consistent with the in-sample prediction performance. Surprisingly, the out-of-sample prediction of CHL by the VAR(2) model is better than the other models and an improvement of accuracy over the TV-VSTAR models is nearly 20%. Nevertheless, the out-of-sample prediction of DO by the TV-VSTAR model is better and this is different from its in-sample prediction performance. The univariate prediction in Table 2.6 demonstrates the consistency of model performance during in-sample prediction. In general, the RMSEs in Table 2.6 are smaller than those in Table 2.5 except the VSTAR(1,1) model.

The out-of-sample predictions of CHL and DO by the VSTAR(1,1) model are shown in Figure 2.12. The prediction performance of CHL during the start and the end of the sampling period is reasonably well even when the variation of CHL is larger relative to the remaining observations during the out-of-sample period. Furthermore, the prediction of DO is satisfactory over the full out-of-sample period.

2.8 Conclusion

In this chapter, a new filter EnGTKF is proposed in which the conventional EnKF is extended in two directions. Firstly, the nonlinear updating equation is derived based on the Gaussian density assumption of the measurement and state errors. Under the Gaussian assumption, together with the use of Goldberger-Theil's mixed estimation, the filtered state can be obtained by the method of generalized least squares. Then, the ensemble mean and variance of state vector $\mathbf{x}_n$ are estimated by the ensemble averages. Secondly, when the prediction densities of the measurement and state vectors, $\mathbf{y}_n$ and $\mathbf{x}_n$, deviate from that of Gaussian, the multivariate Gram-Charlier density is suggested to approximate the non-Gaussian densities. The ensemble mean and variances are obtained by the maximization of the corresponding likelihood functions. Under the non-Gaussian prediction densities, the generalized least squares method becomes not robust enough to estimate the filtered state. As a result, the MAP approach is used to derive the nonlinear updating equation. As in the case of prediction densities, the ensemble filtering mean and filtering variance are estimated by the maximization of the likelihood function.

To estimate the model parameters in the multivariate state space model, an orthogonal transformation technique is suggested to ensure the positive definiteness of a variance-covariance matrix. In addition, a hybrid approach of optimization procedure is suggested where the localized stochastic search algorithm is incorporated with the local optimization procedure, quasi-Newton say, to estimate the parameters. To draw statistical inference on the parameters, a recursive estimation method of the derivative of the likelihood with respect to the parameters is provided and hence the observed information matrix can then be estimated.

In the numerical simulations of the VAR and TV-VSTAR models, the estimation of the state variables shows a satisfactory result in prediction, filtering and smoothing. In the case of nonlinear simulation model by Kitagawa, the prediction of measurement variable y_n is satisfactory after choosing the appropriate order of Hermite polynomial while that of state variable x_n is not completely satisfactory due to the presence of the sinusoidal component.

Furthermore, an empirical application of EnGTKF is illustrated for the algal bloom data in Hong Kong. However, since its measurement equation is linear in the state variables under TV-VSTAR model, its performance should be equivalent to the conventional EnKF. In this study, the amount of standardized chlorophyll fluorescence and standardized concentration of dissolved oxygen in water are the measurement variables in the TV-VSTAR and VAR models. Their in-sample prediction RMSEs show that the TV-VSTAR model can outperform the VAR model at the optimal lag chosen by the AIC. On the other hand, the in-sample prediction of CHL in the selected period is comparable with those by Lee et al. (2004) and Muttil et al. (2004). However, one should emphasize that only a bivariate model with 2 lags is considered for prediction while they used multivariate prediction equations with longer lags. For the out-of-sample predictability of the model, the prediction RMSEs of various orders of p and d show that the out-of-sample prediction performance is consistent with the in-sample prediction performance of the models generally. Finally, the specified state space model can be extended to the non-Gaussian model and state errors, for example Gaussian mixtures, while more theoretical and empirical works should be done in the future.

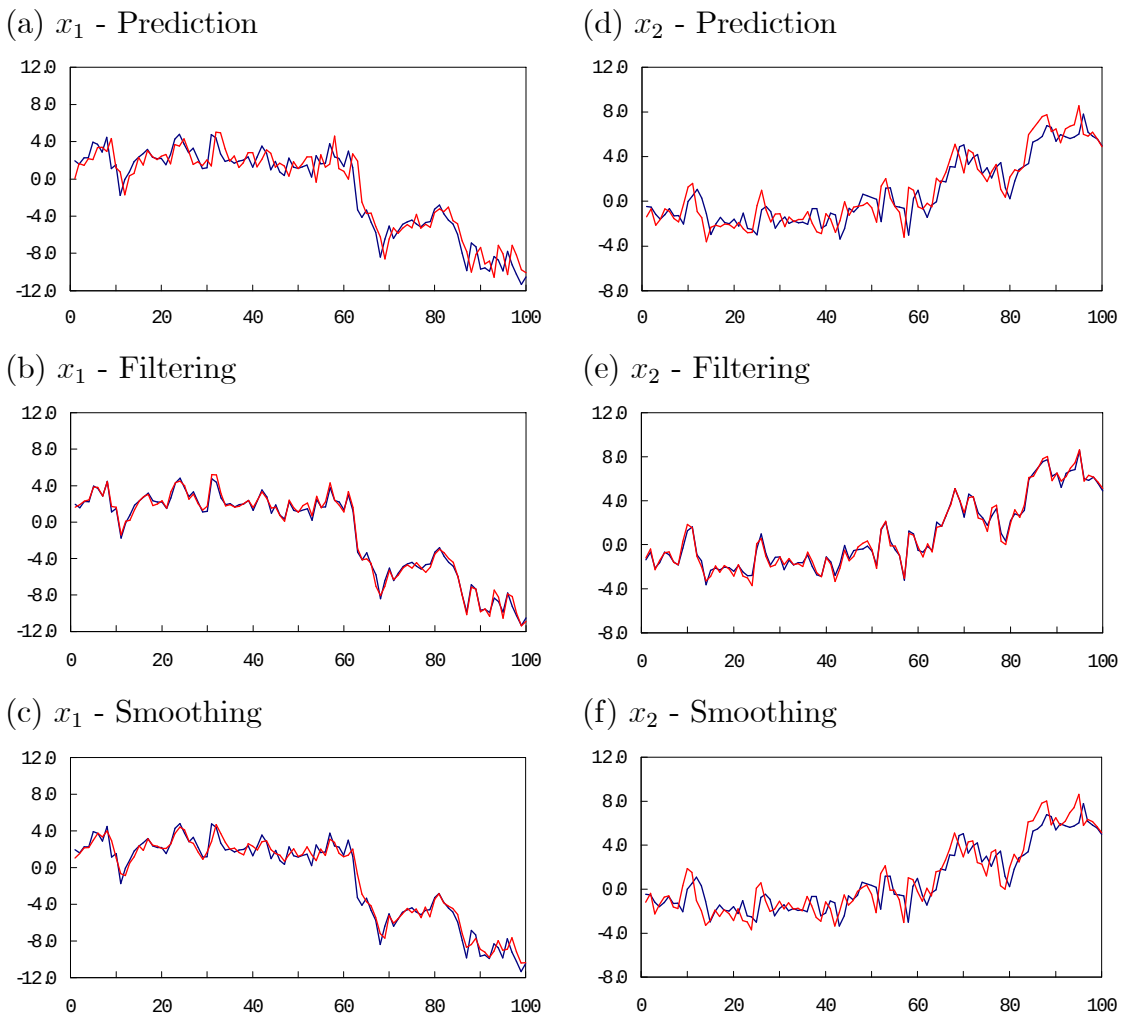


Figure 2.1: Prediction, filtering and smoothing of simulated VAR(1) model by EnGTKF where the left and right panels represent the results of prediction, filtering and smoothing of x_1 and x_2 respectively. The red lines indicate the estimated quantities while the blue lines indicate the simulated quantities.

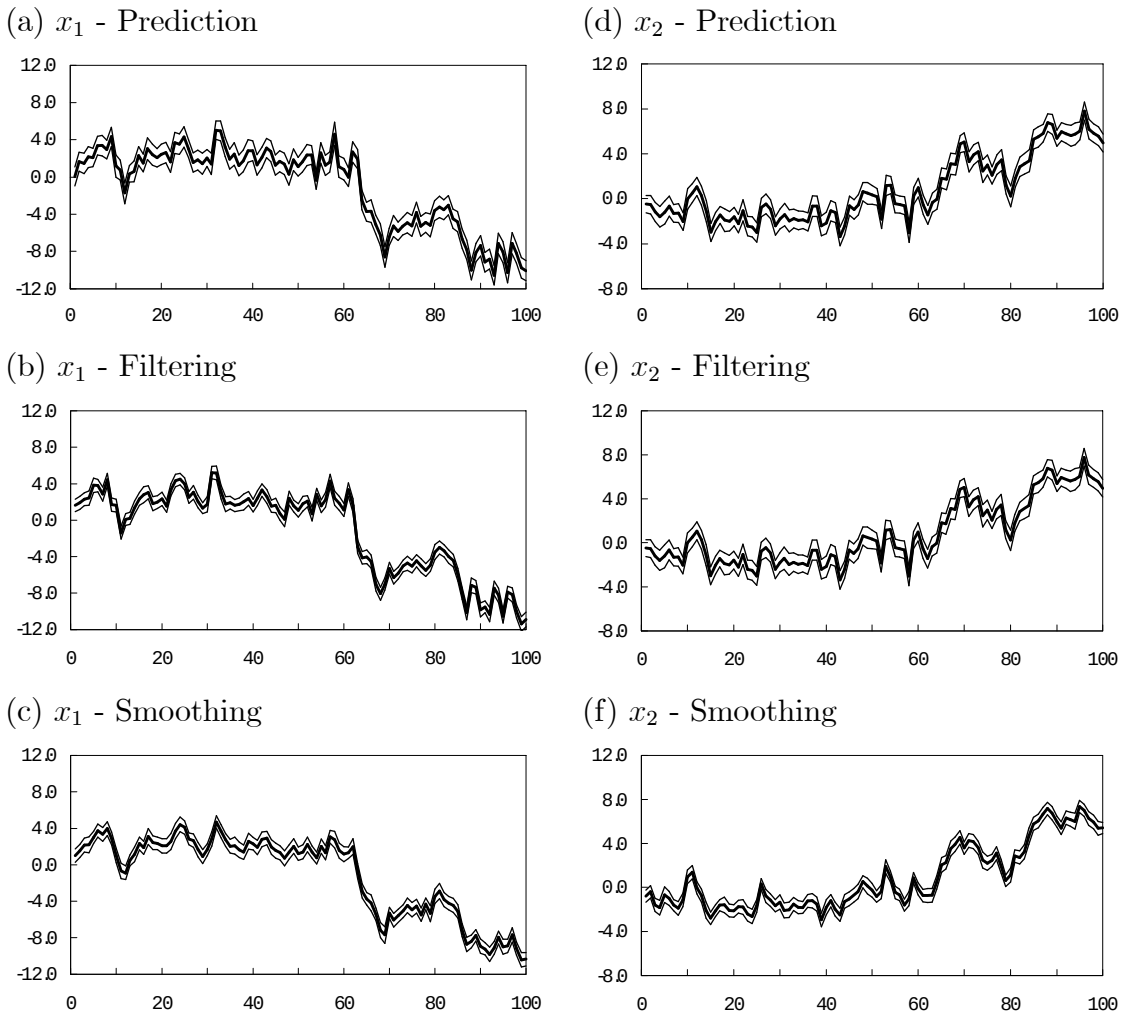


Figure 2.2: Estimation of confidence intervals of prediction, filtering and smoothing of simulated VAR(1) model by EnGTKF where the left and right panels represent the results of prediction, filtering and smoothing of x_1 and x_2 respectively. The middle thick lines indicate the estimated quantities while the thin lines indicate the lower and upper confidence limits.

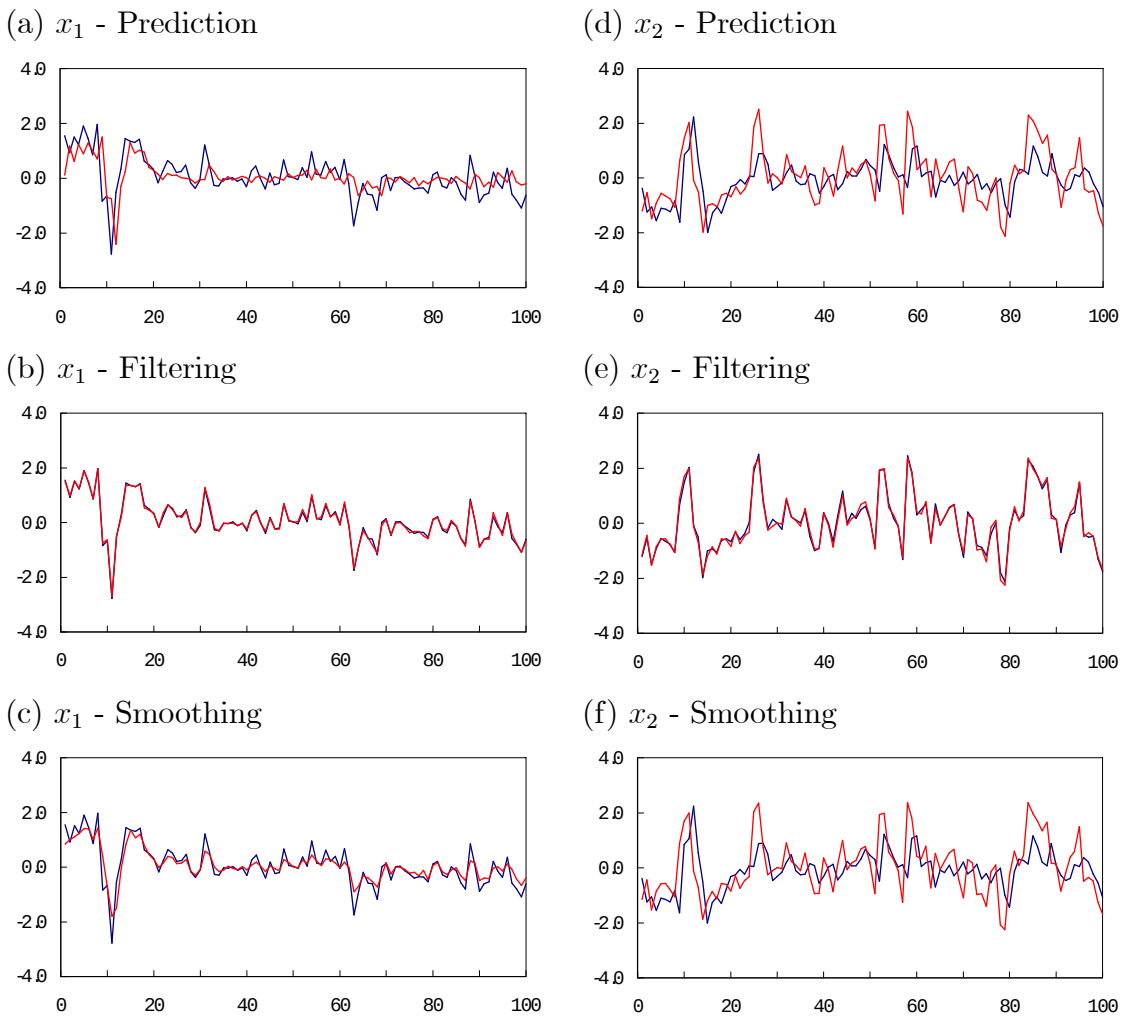


Figure 2.3: Prediction, filtering and smoothing of simulated TV-VSTAR(1,1) model by EnGTKF where the left and right panels represent the results of prediction, filtering and smoothing of x_1 and x_2 respectively. The red lines indicate the estimated quantities while the blue lines indicate the simulated quantities.

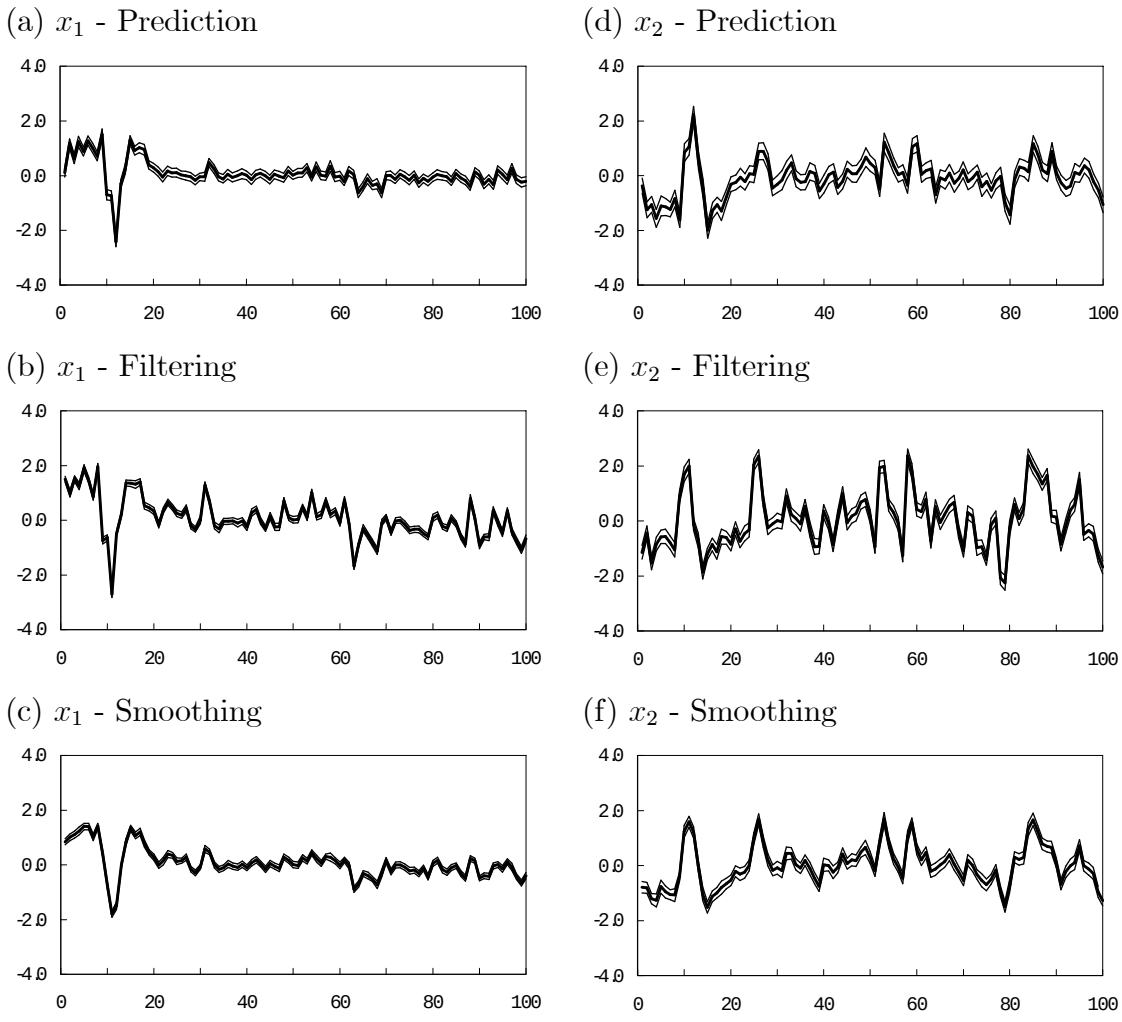


Figure 2.4: Estimation of confidence intervals of prediction, filtering and smoothing of simulated VSTAR(1,1) model by EnGTKF where the left and right panels represent the results of prediction, filtering and smoothing of x_1 and x_2 respectively. The middle thick lines indicate the estimated quantities while the thin lines indicate the lower and upper confidence limits.

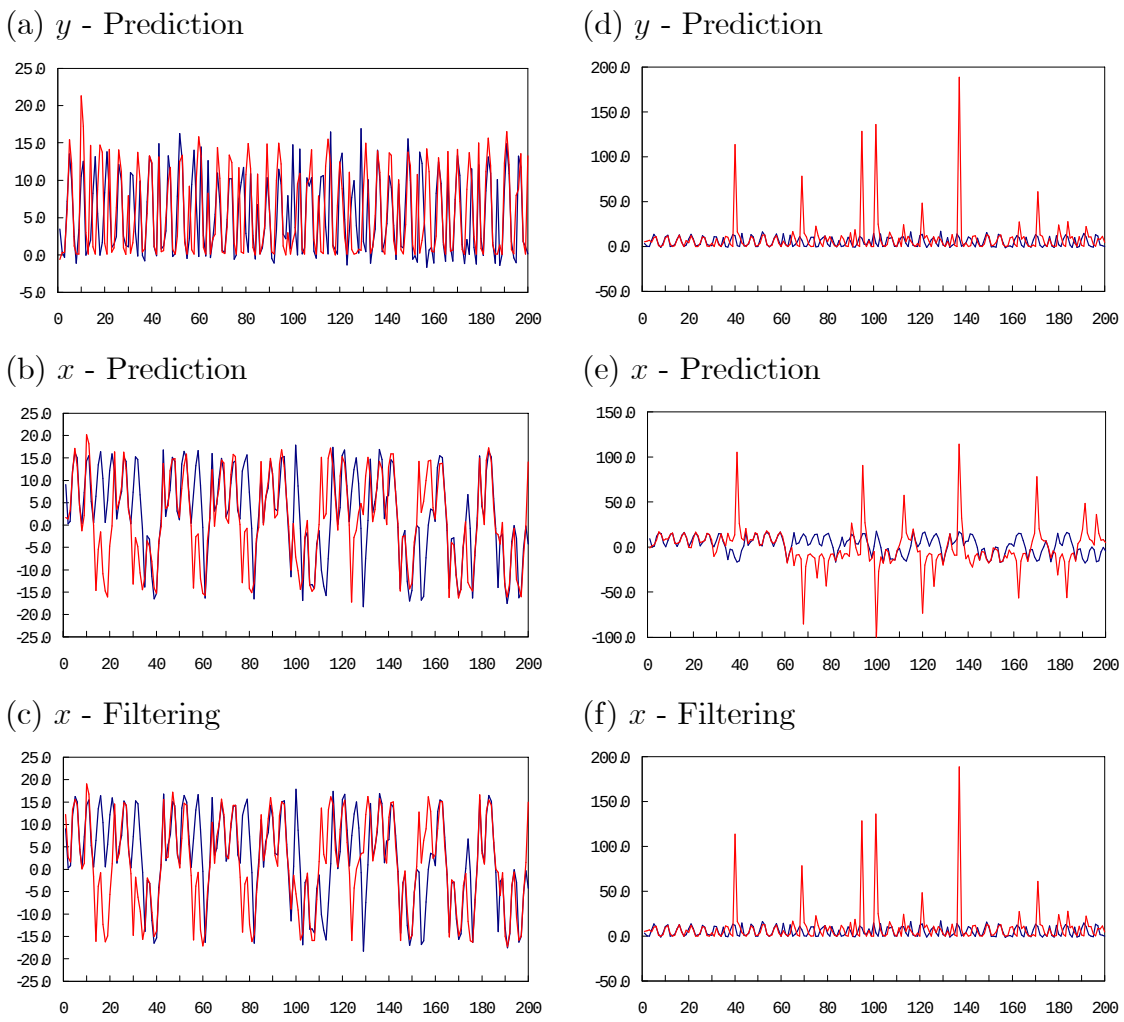


Figure 2.5: Prediction and filtering of Kitagawa's simulation model by EnGTKF and EnKF where the left and right panels represent the nonlinear and linear updating equations respectively. The red lines indicate the estimated quantities while the blue lines indicate the simulated quantities.

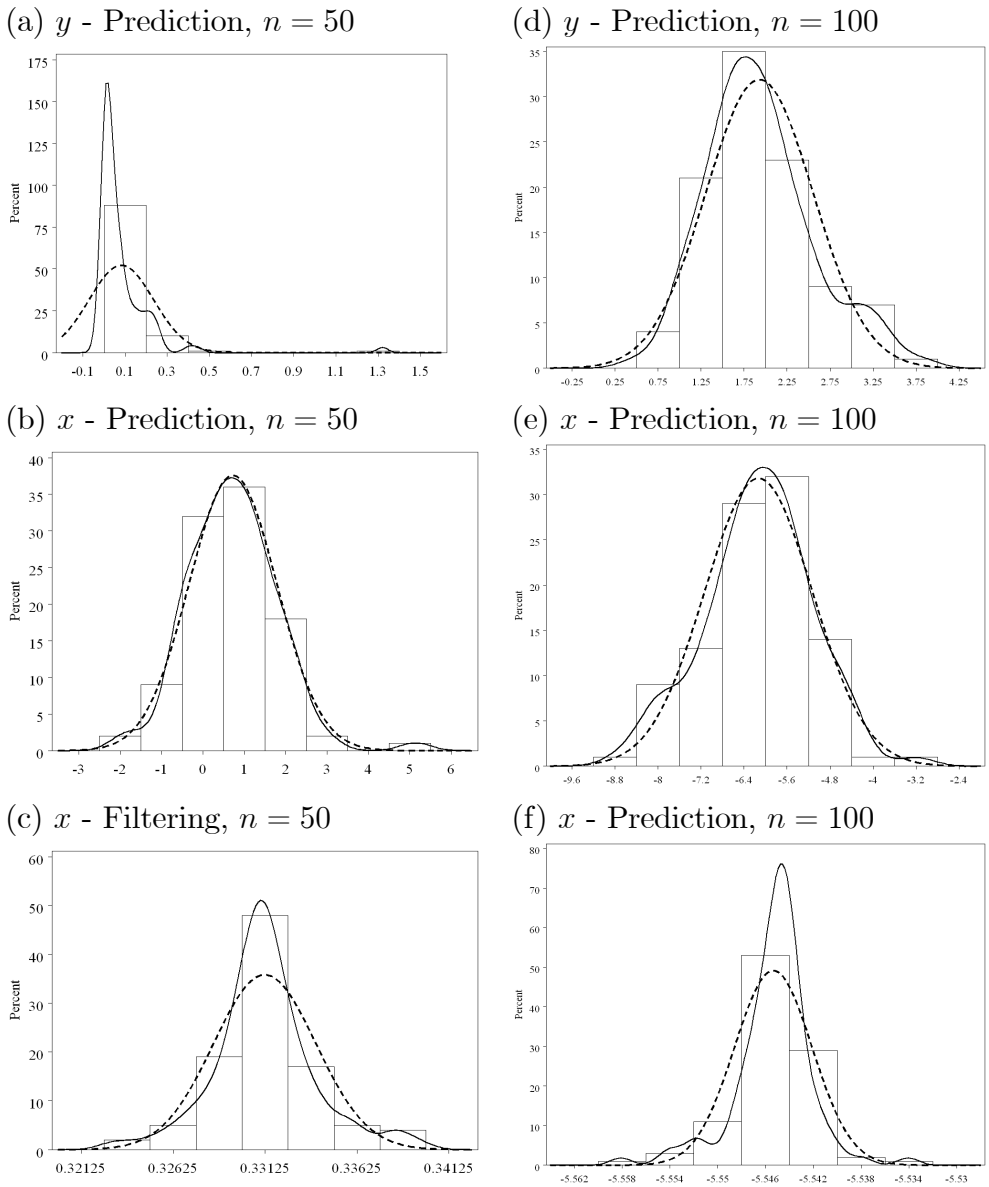


Figure 2.6: Prediction density of y_n, prediction and filtering densities of x_n when $n = 50$ and $n = 100$ where the solid lines denote the estimated density and the dotted lines denote the Gaussian density. The order of Hermite polynomial used in the Gram-Charlier density is 2.

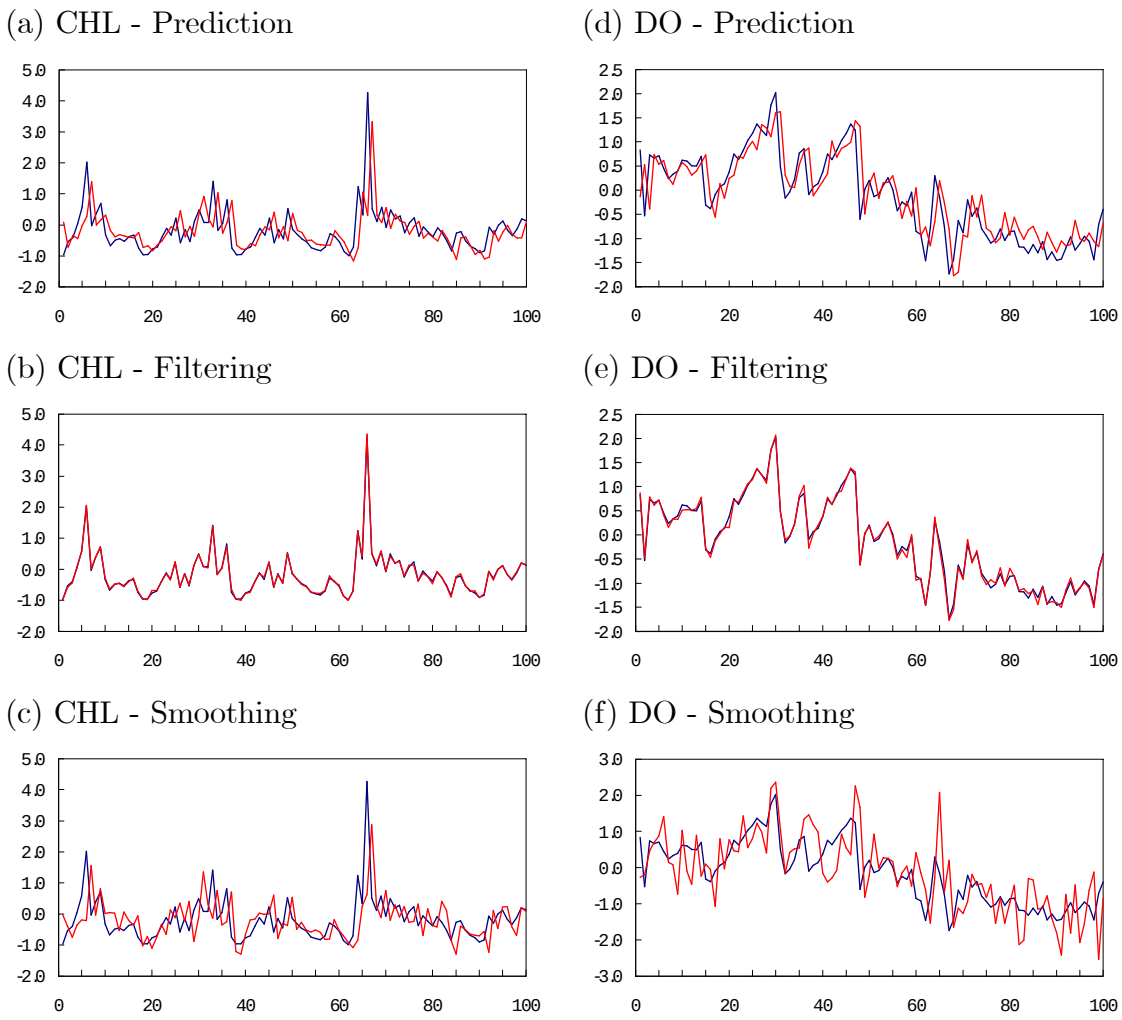


Figure 2.7: In-sample estimation of CHL and DO by EnGTKF with VSTAR(1,1) model for the first 100 observations where the left and right panels indicate the prediction, filtering and smoothing of CHL and DO respectively. Red lines represent the estimated quantities while the blue lines represent the actual qnanitites

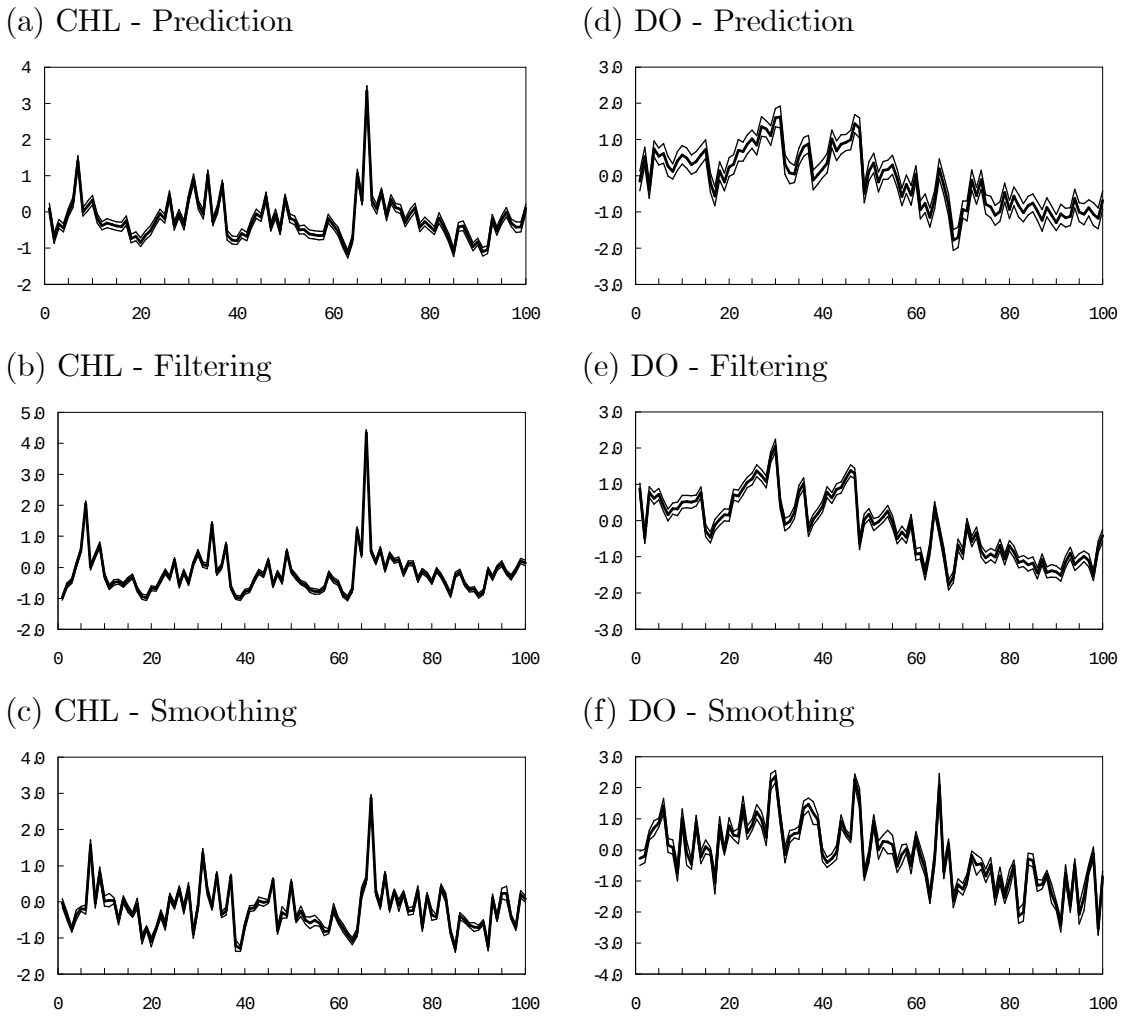


Figure 2.8: In-sample estimated confidence intervals of CHL and DO with VSTAR(1,1) model for the first 100 observations where the left and right panels indicate the prediction, filtering and smoothing of CHL and DO respectively. The middle thick lines represent the estimated quantities while the thin lines represent the lower and upper confidence limits.

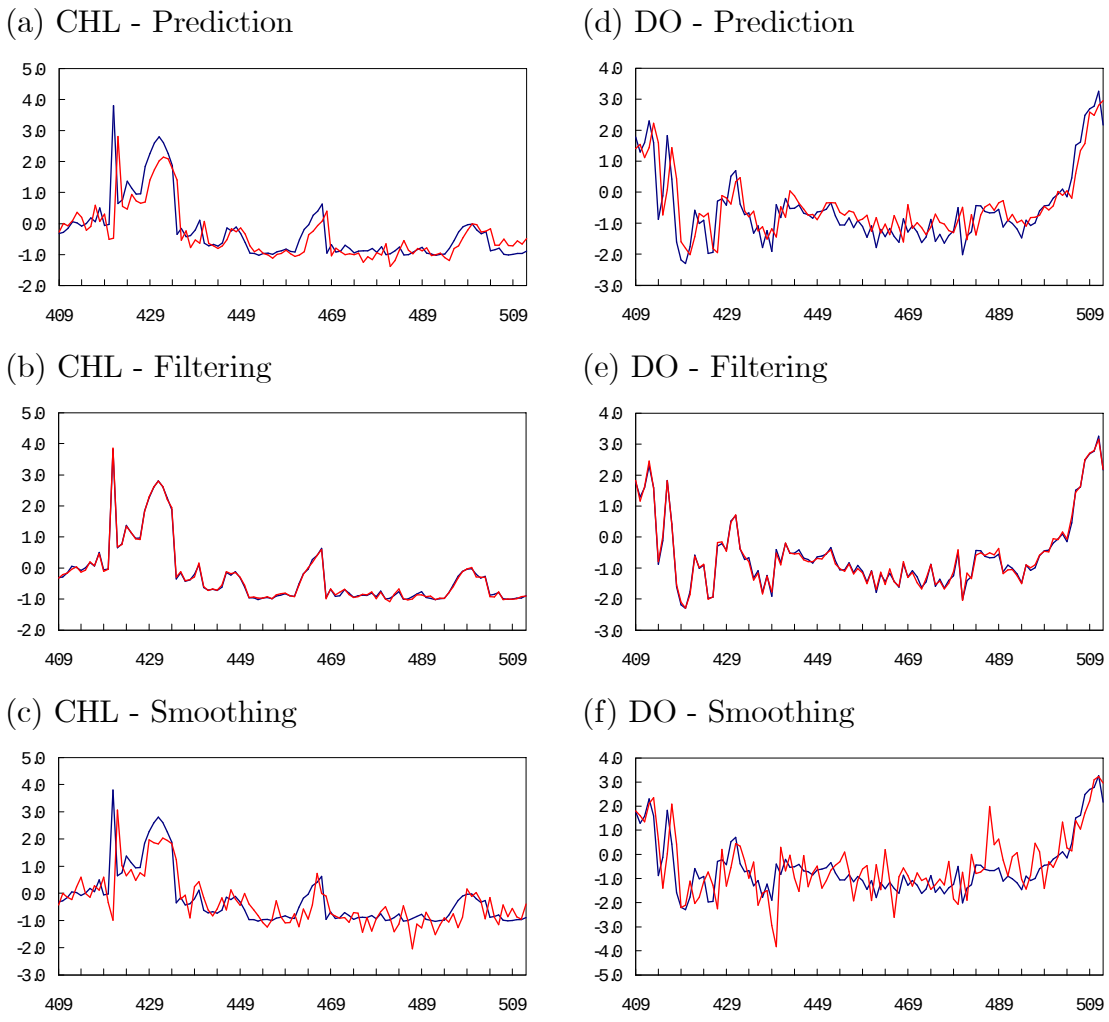


Figure 2.9: In-sample estimation of CHL and DO by EnGTKF with VSTAR(1,1) model during 1 July, 2001 and 31 October, 2001 where the left and right panels indicate the prediction, filtering and smoothing of CHL and DO respectively. Red lines represent the estimated quantities while the blue lines represent the actual qnanitites.

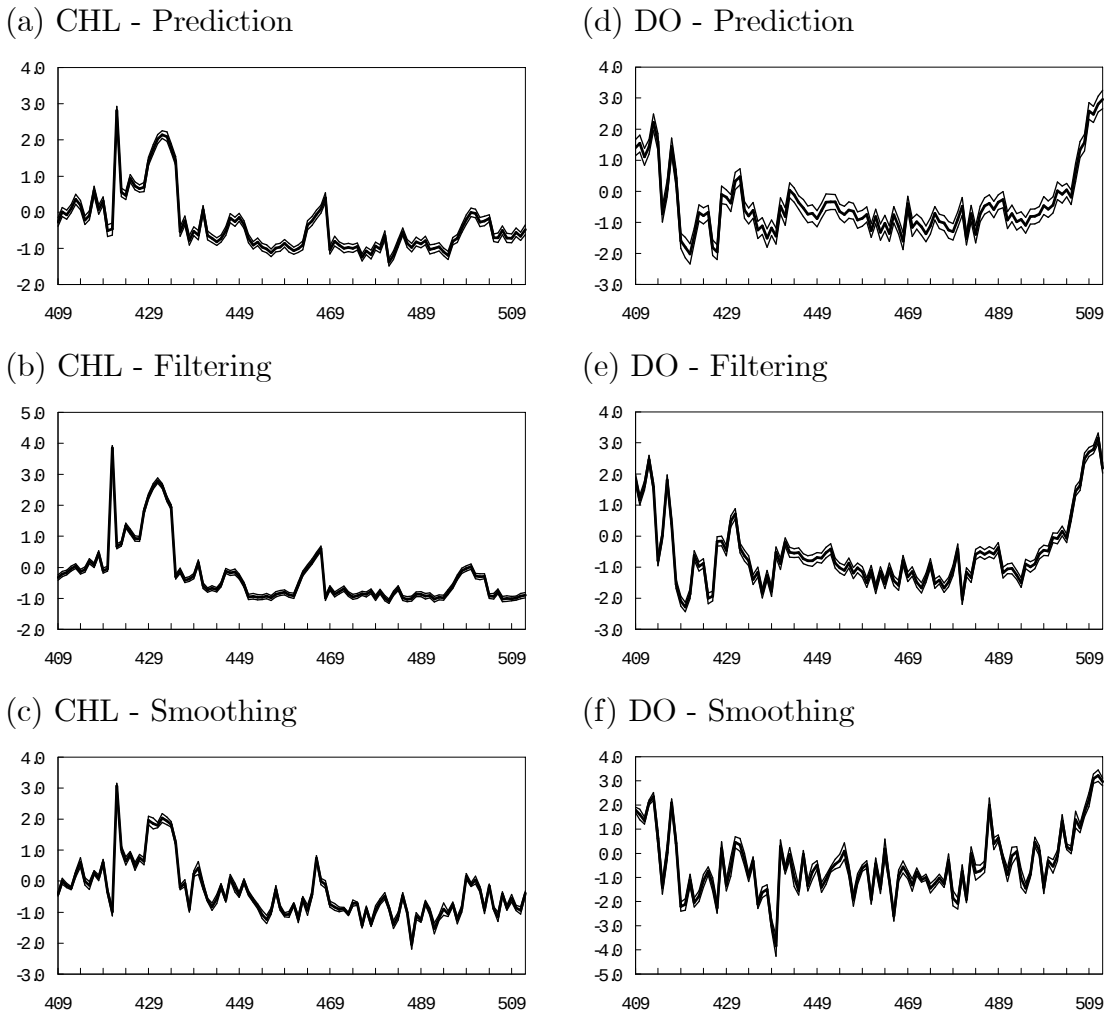
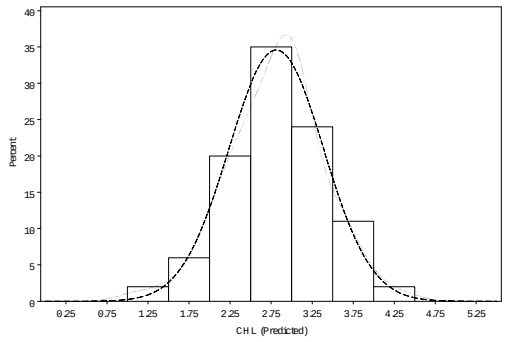
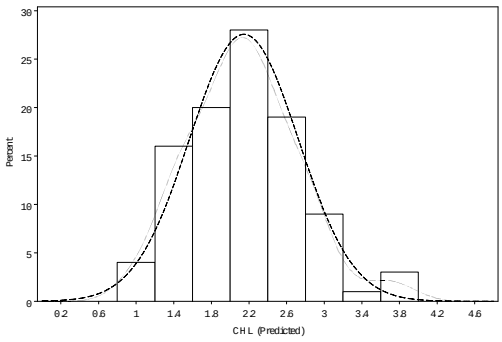


Figure 2.10: In-sample estimated confidence intervals of CHL and DO with VSTAR(1,1) model during 1 July, 2001 and 31 October, 2001 where the left and right panels indicate the prediction, filtering and smoothing of CHL and DO respectively. The middle thick lines represent the estimated quantities while the thin lines represent the lower and upper confidence limits.

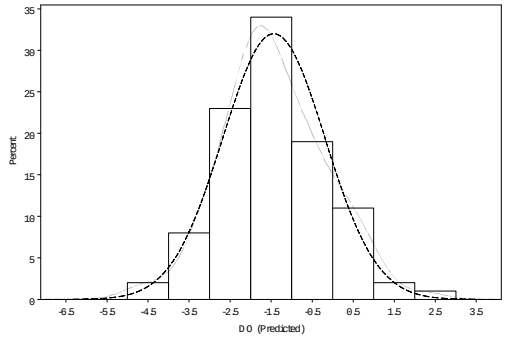
(a) CHL (Predicted), $n = 421$



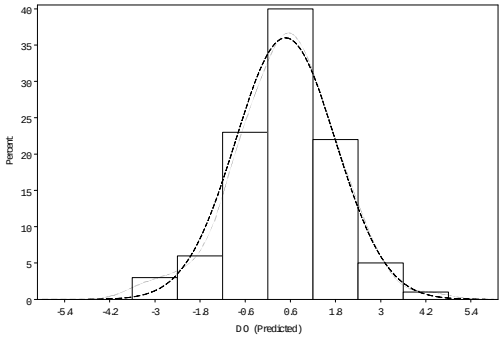
(e) CHL (Predicted), $n = 431$



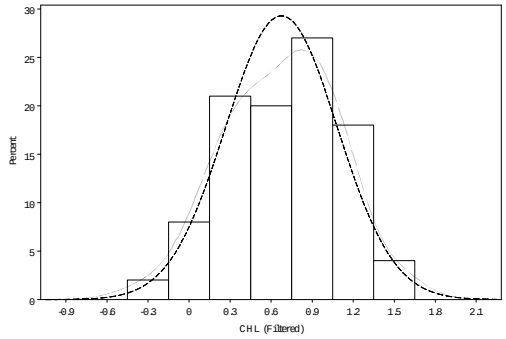
(b) DO (Predicted), $n = 421$



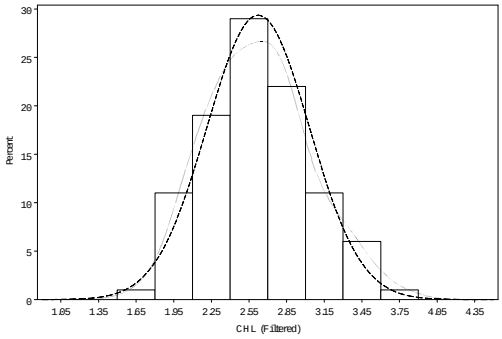
(f) DO (Predicted), $n = 431$



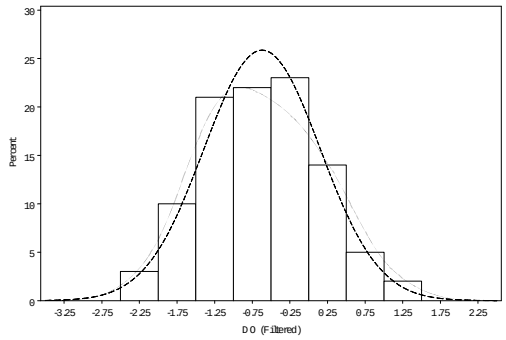
(c) CHL (Filtered), $n = 421$



(g) CHL (Filtered), $n = 431$



(d) DO (Filtered), $n = 421$



(h) DO (Filtered), $n = 431$

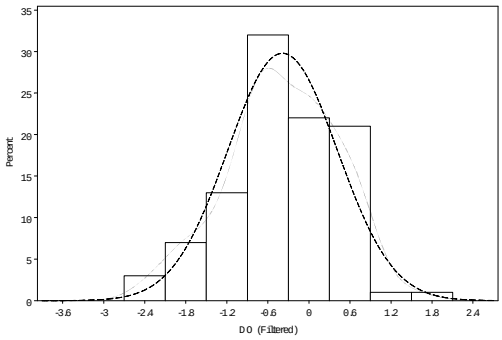


Figure 2.11: Prediction and filtering densities of CHL and DO when $n = 421$ and $n = 431$ where the solid lines denote the estimated density and the dotted lines denote the Gaussian density

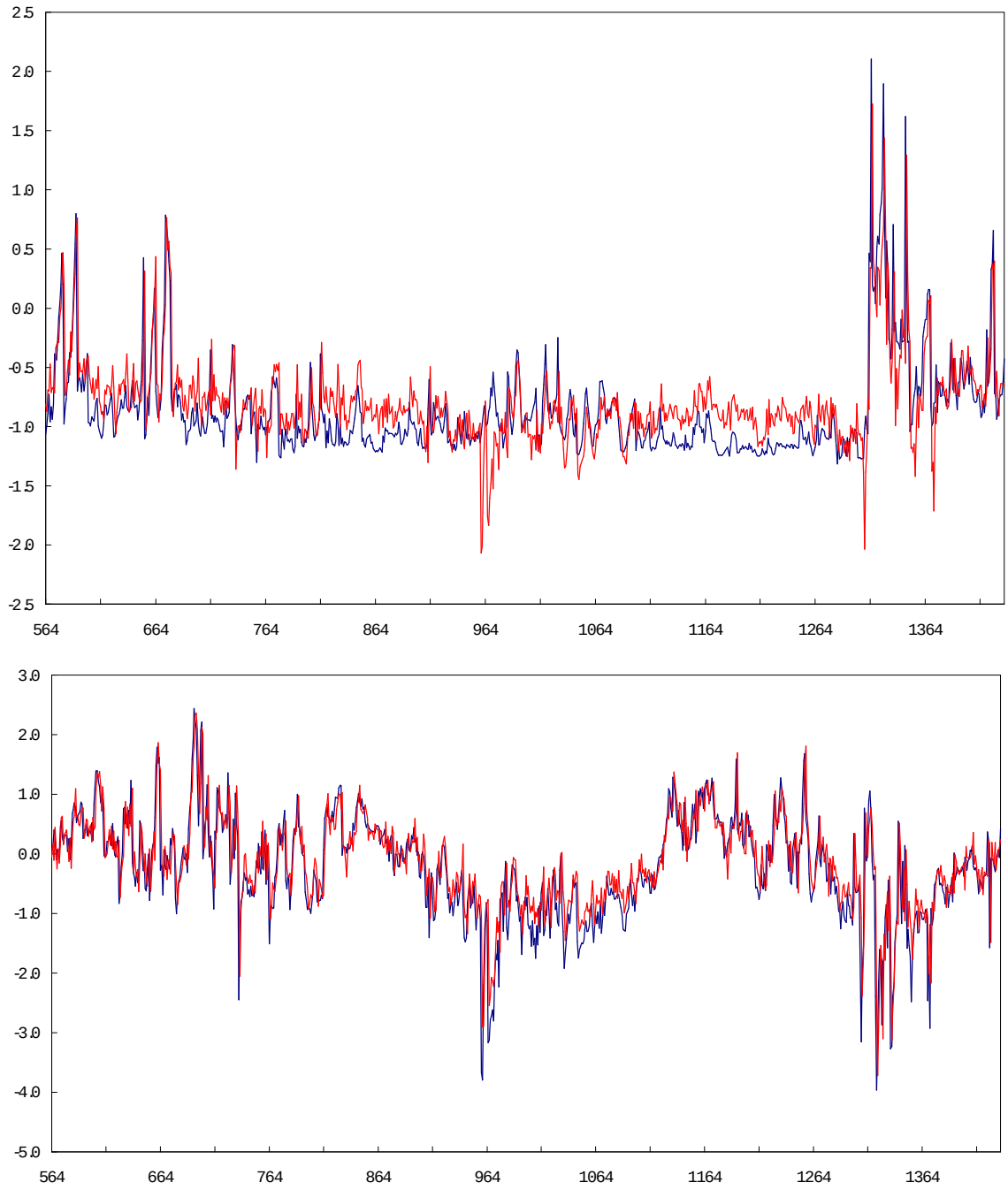


Figure 2.12: Out-of-sample prediction of CHL and DO by EnGTKF during 2002-2004 where the upper and lower panels denote CHL and DO respectively. The red lines represent the predicted quantities and the blue lines represent the actual quantities.

Chapter 3

An Ensemble-based Dynamic Switching Kalman Filter

3.1 Introduction

In atmospheric science, Evensen (1994)'s ensemble Kalman filter (EnKF) has been applied to data assimilation over ten years. The main feature of this filter is to improve over the estimation of the first two moments of prediction and filtering densities under nonlinear state space models although the Gaussian densities of disturbance terms are still assumed. Unlike the particle filter (Carlin et al., 1992; Gordon et al., 1993; del Moral, 1996; Crisan et al., 1999; Gilks and Berzuini, 2001; and Arulampalam et al., 2002), the computational burden is reduced by the estimation of only the first two moments and their propagation over time. Since the EnKF generates stochastic distribution to brute force the propagation of state variables over time, it needs to know the functional forms of the density functions. Although Gaussian density is sufficient in many applications, large deviation from the Gaussian density may lead to divergence of propagated means and variances over time due to incorrect estimation of the mean and variance by the ensemble members.

A simple scenario is that the density functions of measurement and state errors may not be unimodal, but multimodal. Some researchers have suggested that the multimodal densities can be approximated by a Gaussian mixture or Gaussian densities with Markov switching. Early studies of Gaussian mixture in Kalman

filter was due to Sorenson and Alspach (1971) and Alspach and Sorenson (1972). Afterwards, some researchers suggested the usage of Gaussian mixtures in the EnKF although this results in the estimation of more parameters. For example, Anderson and Anderson (1999) was the earliest one to consider the application of EnKF with Gaussian mixtures in which the unknown density function is approximated by a sum of two Gaussian densities with different means and same variance matrix by different scalars. Miller and Ehret (2002) suggested that if the covariances of state variables in the state equation was large, even if the initial density function is Gaussian, bimodal density may be generated by certain highly dynamical systems. Also, Chen and Liu (2000) suggested the mixture Kalman filter (MKF) which included the sequential Monte Carlo and resampling features of particle filter with Kalman filter prediction and updating rules. Bengtsson et al. (2003) proposed the mixture ensemble Kalman filter (XEnsF) and local-local ensemble filter (LLEnsF) where the prior density was assumed to be a Gaussian mixture and the number of Gaussian components was re-estimated by clustering techniques. Concerning with Markov switching, Hamiltion (1989, 1990) was a pioneer to introduce Markov switching into time series model. Furthermore, Kim (1994), Krolzig (1997), Chen and Liu (2000) and Higuchi and Fukuda (2003) specified various dynamic linear models in the form of state space models with Markov switching structure and different approximations of densities were used during the operation of the filters. It is well known that the structure of Markov switching can include the mixture of densities as a special case.

In this chapter, a new variant of EnKF called the ensemble Markov switching Kalman filter (EnMSKF) is proposed in which the feature of Markov switching is incorporated into EnKF such that the dynamic nonlinear time series model

can switch between unobserved regimes over time. Indeed, the proposed filter is closely related to that by Kim (1994). However, a different sampling method of ensemble members is considered in this new nonlinear filter. Therefore, the setting of Gaussian mixture in Anderson and Anderson (1999) is extended to Markov switching by a formal statistical treatment. To allow for Markov switching of the measurement and state errors, the number of Gaussian components and hence the number of parameters increases with the sample size exponentially. This results in the infeasible operation of the filter. An approximation is suggested in Kim (1994) to control the number of Gaussian components.

In many atmospheric applications, little emphases were placed on the estimation of unknown parameters. Typically, unknown parameters are estimated by some ad hoc methods, for example, the method of moment estimation. In fact, a vast literature in statistics focused on the issue of parameter estimation, for example, see McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) for an overview. Typically, local optimization algorithm is applied to find the unknown parameters. However, a typical difficulty in the estimation of model with mixture errors is the multimodal feature of likelihood. As a result, in this chapter, a localized stochastic search algorithm, is implemented to improve the quality of the initial values. As well known in many local optimization algorithms, good quality of initial values can lead to the fast convergence of estimation results. One advantage of this algorithm is that it does not require the objective function to be differentiable with respect to the unknown parameters. A feature in evolutionary strategy is introduced into this algorithm to ensure the positive definiteness of variance-covariance matrices across generations. The variance-covariance matrices are decomposed into products of orthogonal rotation matrices which are

determined by rotation angles only. After the global optimization of the likelihood function, a local optimization algorithm, such as Newton-Raphson and Quasi-Newton algorithms, is used to improve the estimation results further.

In Section 3.2, the model and assumptions are presented. Then, in Section 3.3, EnMSKF is derived in which the ensemble predictions of measurement and state variables and prediction and filtered probabilities of the switching regimes are included. The asymptotic properties of estimated mean and variances are studied afterwards. A natural extension of the ensemble Kalman smoother (EnKS) is given in deriving an smoother which is called the ensemble Markov switching Kalman smoother (EnMSKS). In Section 3.4, the estimation of model parameters is investigated. The estimation of standard errors of model parameters and the determination of the number of switching regimes are given in Sections 3.5 and 3.6 respectively. An empirical application of this filter is given in Section 3.7. Finally, the conclusions are drawn in Section 3.8.

3.2 The Model

Consider the nonlinear state space model:

$$\mathbf{y}_n = \mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}^{[s_n]}) + \mathbf{v}_n, \tag{3.1}$$

$$\mathbf{x}_n = \mathbf{f}(\mathbf{x}_{n-1}; \boldsymbol{\xi}^{[s_n]}) + \mathbf{u}_n, \tag{3.2}$$

for $n = 1, 2, \ldots, N$, where $\mathbf{y}_n \in \mathbb{R}^{m_y}$, $\mathbf{x}_n \in \mathbb{R}^{m_x}$, $\mathbf{h} : \mathbb{R}^{m_x} \to \mathbb{R}^{m_y}$ and $\mathbf{f} : \mathbb{R}^{m_x} \to \mathbb{R}^{m_x}$ are vector-valued functions. Also, $\boldsymbol{\xi}^{[s_n]} \in \mathbb{R}^z$ is a vector of model parameters which depend on the hidden state variable s_n. Both functions $\mathbf{h}(\cdot)$ and $\mathbf{f}(\cdot)$

are assumed to be smooth, that is, all derivatives of $\mathbf{h}(\cdot)$ and $\mathbf{f}(\cdot)$ exist and continuous. In state space modelling, equations (3.1) and (3.2) are known as measurement equation and transition equation respectively. The specification of (3.1) and (3.2) can include many time series models, such as, Markov switching vector autoregressive (MS-VAR) model, multivariate version of self-excited threshold autoregressive (SETAR) model and vector smooth transition autoregressive (VSTAR) model. The details are referred to Krolzig (1997, p.36).

For notation simplicity, we denote $\mathcal{X}_n \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathcal{Y}_n \equiv \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ and $\mathcal{S}_n \equiv \{s_1, \ldots, s_n\}$. The conditional density functions of $\mathbf{v}_n$ and $\mathbf{u}_n$ are assumed to be

$$p(\mathbf{v}_n|\mathcal{Y}_{n-1}, s_n) = \mathcal{N}(\mathbf{v}_n; \mathbf{0}, \mathbf{Q}^{[s_n]}) \quad \text{and} \quad p(\mathbf{u}_n|\mathcal{Y}_{n-1}, s_n) = \mathcal{N}(\mathbf{u}_n; \mathbf{0}, \mathbf{R}^{[s_n]}) \quad (3.3)$$

respectively where

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{m_x}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

$\mathbf{Q}^{[s_n]}$ and $\mathbf{R}^{[s_n]}$ are non-negative definite for all n and s_n. The hidden state variable s_n possesses a Markovian structure and takes a value from $\{1, \ldots, M\}$ where M is finite. The transition probability for s_n is defined by $p_{ij} = \Pr(s_n = j|s_{n-1} = i) > 0$ for $i, j = 1, \ldots, M$. Then, it follows that $\sum_{j=1}^{M} p_{ij} = 1$ for $i = 1, \ldots, M$. Note that the Markov switching model can be reduced to the mixture model when $\Pr(s_n = j|s_{n-1} = i) = \Pr(s_n = j)$ or $p_{ij} = p_{i'j}$ for $i \neq i'$ is assumed.

The density function of the initial state $\mathbf{x}_0$ is defined by

$$p\left(\mathbf{x}_0|s_0=j\right)=\mathcal{N}(\mathbf{x}_0;\mathbf{x}_{0|0}^{[j]},\boldsymbol{\Sigma}_{0|0}^{[j]}),\quad \text{for } j=1,\ldots,M.$$

For identifiability of the mixture components, many authors have proposed various constraints on the model parameters. For example, Titterington et al. (1985, Corollary 3.11) showed a generic identifiability constraint on a finite mixture in which the transformed component densities are linearly independent. Frühwirth-Schnatter (2006, p.19–20) suggested that in a multi-parameter case, two mixture components can be identified when at least one of the parameter elements are distinct from the other components and the mixing proportions are arranged in ascending order. This can avoid the problem of label switching. In this chapter, the constraints by Frühwirth-Schnatter (2006) are imposed on the model parameters, $\mathbf{x}_{0|0}^{[s_n]}$, $\boldsymbol{\Sigma}_{0|0}^{[s_n]}$, $\boldsymbol{\xi}^{[s_n]}$, $\mathbf{R}^{[s_n]}$, and $\mathbf{Q}^{[s_n]}$.

The assumptions for conditional density functions of $\mathbf{y}_n$, $\mathbf{x}_n$ and s_n are given below:

Assumption 3.1

(i) Conditional on $\mathbf{x}_n$, the current measurements $\mathbf{y}_n$ are independent of $\mathcal{X}_{n-1}$, $\mathcal{Y}_{n-1}$ and $\mathcal{S}_{n-1}$, that is,

$$p\left(\mathbf{y}_n|\mathcal{X}_n,\mathcal{Y}_{n-1},\mathcal{S}_n\right)=p\left(\mathbf{y}_n|\mathbf{x}_n,s_n\right),\quad \text{for}\quad n=2,\ldots,N,$$

and

$$p\left(\mathbf{y}_n|\mathcal{X}_n,\mathcal{S}_n\right)=p\left(\mathbf{y}_n|\mathbf{x}_n,s_n\right),\quad \text{for}\quad n=1.$$

The conditional independence assumption follows from the measurement equation (3.1).

(ii) The state vector $\mathbf{x}_n$ depends on its one-period lagged value $\mathbf{x}_{n-1}$ and the current hidden state variable s_n only, that is,

$$p\left(\mathbf{x}_n|\mathcal{X}_{n-1},\mathcal{Y}_{n-1},\mathcal{S}_n\right)=p\left(\mathbf{x}_n|\mathbf{x}_{n-1},s_n\right),\quad \text{for}\quad n=1,\ldots,N.$$

(iii) The hidden state variable s_n is Markovian and the Markov chain $\{s_n\}_{n=1}^N$ is homogenous. Then,

$$\Pr(s_n|\mathcal{S}_{n-1},\mathcal{Y}_{n-1})=\Pr\left(s_n|s_{n-1}\right),\quad \text{for}\quad n=1,\ldots,N.$$

(iv) The Markov chain $\{\mathbf{x}_n,s_n\}_{n=1}^N$ is irreducible, aperiodic and admits an invariant probability measure.

Assumption 3.1 implies that given the values of $\mathbf{x}_n$ and current hidden state s_n, the density function of $\mathbf{y}_n$ can be derived directly. Furthermore, the density function of the state vector $\mathbf{x}_n$ can be derived only from the state vector $\mathbf{x}_n$ at time $n-1$ and the current hidden state s_n. Indeed, conditions (ii), (iii) and (iv) imply that the Markov chain $\{\mathbf{x}_n,s_n\}_{n=1}^N$ is ergodic (Chan and Tong, 2001, p.34). Furthermore, when the Markov chain $\{\mathbf{x}_n,s_n\}_{n=1}^N$ is assumed to be Harris recurrent by Meyn and Tweedie (1993, Theorem 13.0.1), the state process $\{\mathbf{x}_n,s_n\}_{n=1}^N$ converges to a stationary distribution for every initial condition $(\mathbf{x}_0,s_0)\in\mathbb{R}^{m_x}\times\{1,\ldots,M\}$. This theorem also ensures the stability of the state space model of (3.1) and (3.2) since the Foster's condition (Meyn and Tweedie, 1993, p.501) is satisfied automatically. One may suggest that $\{\mathbf{x}_n,s_n\}_{n=1}^N$ can be considered as an augmented state space and the state space model can be expressed in a simpler form. Note that $\{\mathbf{x}_n\}_{n=1}^N$ and $\{s_n\}_{n=1}^N$ are

defined in continuous and discrete state spaces respectively. One feature of the proposed filter is that no simulation related to $\{s_n\}_{n=1}^N$ is required and the operation of this filter is in parallel with the conventional EnKF. To enable the new filter, more assumptions on $\mathbf{u}_n$, $\mathbf{v}_n$ and $\mathbf{x}_0$ are necessary and given below:

Assumption 3.2

(i) The disturbances $\mathbf{u}_n$ and $\mathbf{v}_n$ are uncorrelated with each other for all time periods, that is, $\mathbb{E}\left(\mathbf{u}_n\mathbf{v}_m^\mathsf{T}\right) = \mathbf{0}$ for all m and n.

(ii) The initial state vector $\mathbf{x}_0$ is uncorrelated with $\mathbf{u}_n$ and $\mathbf{v}_n$, that is, $\mathbb{E}\left(\mathbf{u}_n\mathbf{x}_0^\mathsf{T}\right) = \mathbb{E}\left(\mathbf{v}_n\mathbf{x}_0^\mathsf{T}\right) = \mathbf{0}$ for $n = 1, \ldots, N$.

3.3 Recursive Estimation of Model States

In this section, the recursive estimations of the state vector $\mathbf{x}_n$ and the probability mass of the hidden state variable s_n are derived. As a by-product, the prediction of the measurement vector $\mathbf{y}_n$ is derived as well. To simplify the notation for the probability of the hidden state s_n, $s_n^{[j]}$ and $s_{n-1,n}^{[i,j]}$ denote $(s_n = j)$ and $(s_n = j, s_{n-1} = i)$ respectively. Hence, $\Pr(s_{n-1,n}^{[i,j]}) = \Pr(s_n^{[j]}|s_{n-1}^{[i]}) \times \Pr(s_{n-1}^{[i]})$. The parameters $\{\Pr(s_0^{[j]})\}_{j=1}^M$, $\{\mathbf{x}_{0|0}^{[s_n]}, \mathbf{\Sigma}_{0|0}^{[s_n]}, \boldsymbol{\xi}^{[s_n]}, \mathbf{R}^{[s_n]}, \mathbf{Q}^{[s_n]}\}_{n=1}^N$ and $\{p_{ij}\}_{i,j=1}^M$ are assumed to be known in this section. Then, the recursive estimation of the model states and measurements in the forms of density functions is given below.

Given $s_n^{[j]}$, the prediction and filtering densities for the state vector $\mathbf{x}_n$ are given by

$$p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) = \int p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]}) p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) d\mathbf{x}_{n-1}, \tag{3.4}$$

and

$$p(\mathbf{x}_n|\mathcal{Y}_n, s_n^{[j]}) = \frac{p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})}{\int p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})d\mathbf{x}_n}, \tag{3.5}$$

respectively for $n \geq 1$ where the filtering density is initialized by

$$p(\mathbf{x}_0|\mathbf{y}_0, s_0^{[j]}) = p(\mathbf{x}_0|s_0^{[j]}) = \mathcal{N}(\mathbf{x}_0; \mathbf{x}_{0|0}^{[j]}, \mathbf{\Sigma}_{0|0}^{\mathbf{xx}[j]}).$$

The one-step ahead prediction density for the measurement vector $\mathbf{y}_n$ is

$$p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) = \int p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})d\mathbf{x}_n. \tag{3.6}$$

The recursive estimation of filtering and prediction probabilities of the hidden state variable s_n, $\Pr(s_n^{[j]}|\mathcal{Y}_n)$ and $\Pr(s_n^{[j]}|\mathcal{Y}_{n-1})$, is

$$\begin{aligned}
\Pr(s_n^{[j]}|\mathcal{Y}_n) &= \Pr(s_n^{[j]}|\mathbf{y}_n, \mathcal{Y}_{n-1}) \\
&= \frac{p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) \Pr(s_n^{[j]}|\mathcal{Y}_{n-1})}{p\left(\mathbf{y}_n|\mathcal{Y}_{n-1}\right)} \\
&= \frac{p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) \Pr(s_n^{[j]}|\mathcal{Y}_{n-1})}{\sum_{i=1}^{M} p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[i]}) \Pr(s_n^{[i]}|\mathcal{Y}_{n-1})},
\end{aligned} \tag{3.7}$$

with $\Pr(s_0^{[j]}|\mathcal{Y}_0) = \Pr(s_0^{[j]})$ and

$$\begin{aligned}
\Pr(s_n^{[j]}|\mathcal{Y}_{n-1}) &= \sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1}) \\
&= \sum_{i=1}^{M} \Pr(s_n^{[j]}|s_{n-1}^{[i]}, \mathcal{Y}_{n-1}) \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{M} \Pr(s_n^{[j]}|s_{n-1}^{[i]}) \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \\
&= \sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}), \qquad (3.8)
\end{aligned}
$$

where the third equality in (3.8) is deduced by the homogeneity of Markov chain $\{s_n\}_{n=1}^{N}$. The probability $\Pr(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[j]})$ in (3.7) in fact is the denominator term in (3.5). Therefore, the conditional probabilities (3.7) and (3.8) are derived by using the transition probabilities p_{ij} and the normalizing constant in the filtering density (3.5).

Now, the notations of conditional mean and variance of the measurement and state vectors are denoted in the following manner:

$$
\begin{aligned}
\mathbf{z}_{t|s}^{[j]} &\equiv \mathbb{E}\left(\mathbf{z}_t|\mathcal{Y}_s, s_t^{[j]}\right), \\
\mathbf{\Sigma}_{t|s}^{\mathbf{zz}[j]} &\equiv \mathbb{E}\left((\mathbf{z}_t - \mathbf{z}_{t|s}^{[j]})(\mathbf{z}_t - \mathbf{z}_{t|s}^{[j]})^{\mathsf{T}}|\mathcal{Y}_s, s_t^{[j]}\right), \\
\mathbf{z}_{t|s}^{[i,j]} &\equiv \mathbb{E}\left(\mathbf{z}_t|\mathcal{Y}_s, s_{s,t}^{[i,j]}\right), \\
\mathbf{\Sigma}_{t|s}^{\mathbf{zz}[i,j]} &\equiv \mathbb{E}\left((\mathbf{z}_t - \mathbf{z}_{t|s}^{[i,j]})(\mathbf{z}_t - \mathbf{z}_{t|s}^{[i,j]})^{\mathsf{T}}|\mathcal{Y}_s, s_{s,t}^{[i,j]}\right),
\end{aligned}
$$

where the variable $\mathbf{z}$ is replaced by $\mathbf{x}$ and $\mathbf{y}$ correspondingly.

Furthermore, $\left(f_{n-1|n-1}^{[i,j](k)}\right)_p$, $\left(f_{n-1}^{[i,j]}\right)_p$, $\left(h_{n|n-1}^{[i,j](k)}\right)_p$ and $\left(h_n^{[i,j]}\right)_p$ denote the pth elements of $\mathbf{f}(\widehat{\mathbf{x}}_{n-1|n-1}^{[i,j](k)}; \boldsymbol{\xi}^{[i]})$, $\mathbb{E}\left(\mathbf{f}(\mathbf{x}_{n-1}; \boldsymbol{\xi}^{[i]})|\mathcal{Y}_{n-1}, s_{n-1,n}^{[i,j]}\right)$, $\mathbf{h}(\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)}; \boldsymbol{\xi}^{[i]})$ and $\mathbb{E}\left(\mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}^{[i]})|\mathcal{Y}_{n-1}, s_{n-1,n}^{[i,j]}\right)$ respectively; $\left(\widehat{x}_{n-1|n-1}^{[i,j](k)}\right)_p$, $\left(\widehat{x}_{n|n-1}^{[i,j](k)}\right)_p$, $\left(\widehat{y}_{n|n-1}^{[i,j](k)}\right)_p$, $\left(u_n^{[j](k)}\right)_p$ and $\left(v_n^{[j](k)}\right)_p$ denote the pth elements of $\widehat{\mathbf{x}}_{n-1|n-1}^{[i,j](k)}$, $\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)}$, $\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)}$, $\mathbf{u}_n^{[j](k)}$ and $\mathbf{v}_n^{[j](k)}$ respectively for $i, j = 1, \ldots, M$; $\left(x_{n-1|n-1}^{[i,j]}\right)_p$, $\left(x_{n|n-1}^{[i,j]}\right)_p$ and $\left(y_{n|n-1}^{[i,j]}\right)_p$ denote the pth elements of $\mathbf{x}_{n-1|n-1}^{[i,j]}$, $\mathbf{x}_{n|n-1}^{[i,j]}$ and $\mathbf{y}_{n|n-1}^{[i,j]}$ respectively for $i, j = 1, \ldots, M$;

$\left(\widehat{\Sigma}^{\mathbf{xx}[i,j]}_{n|n-1}\right)_{pq}$, $\left(\widehat{\Sigma}^{\mathbf{xx}[i,j]}_{n|n}\right)_{pq}$, $\left(\widehat{\Sigma}^{\mathbf{yy}[i,j]}_{n|n-1}\right)_{pq}$, $\left(\Sigma^{\mathbf{xx}[i,j]}_{n|n-1}\right)_{pq}$, $\left(\Sigma^{\mathbf{xx}[i,j]}_{n|n}\right)_{pq}$ and $\left(\Sigma^{\mathbf{yy}[i,j]}_{n|n-1}\right)_{pq}$ denote the (p,q)th elements of $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}[i,j]}_{n|n-1}$, $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}[i,j]}_{n|n}$, $\widehat{\mathbf{\Sigma}}^{\mathbf{yy}[i,j]}_{n|n-1}$, $\mathbf{\Sigma}^{\mathbf{xx}[i,j]}_{n|n-1}$, $\mathbf{\Sigma}^{\mathbf{xx}[i,j]}_{n|n}$ and $\mathbf{\Sigma}^{\mathbf{yy}[i,j]}_{n|n-1}$ respectively for $i, j = 1, \ldots, M$.

The propagation of the whole conditional density function of the state vector $\mathbf{x}_n$ is clearly computationally unfavorable and it is more desirable to consider propagating its mean and variance over time.

3.3.1 Prediction

Assume that the filtering density for the state vector $\mathbf{x}_n$ for $n \geq 1$ is

$$p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s^{[i]}_{n-1}) = \mathcal{N}(\mathbf{x}_{n-1}; \mathbf{x}^{[i]}_{n-1|n-1}, \mathbf{\Sigma}^{\mathbf{xx}[i]}_{n-1|n-1}), \tag{3.9}$$

for $i = 1, \ldots, M$.

From (3.4) and (3.5), one may consider to generate ensemble members of the state vector $\mathbf{x}_{n-1|n-1}$ by a jumping of the hidden state from $s^{[i]}_{n-1}$ to $s^{[j]}_n$, that is, generating ensemble members of density $p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s^{[j]}_n)$ from $p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s^{[i]}_{n-1})$. The derivation of density $p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s^{[j]}_n)$ from $p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s^{[i]}_{n-1})$ is then given below:

$$p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s^{[j]}_n) = \frac{p(\mathbf{x}_{n-1}, s^{[j]}_n|\mathcal{Y}_{n-1})}{\Pr(s^{[j]}_n|\mathcal{Y}_{n-1})}. \tag{3.10}$$

Consider the numerator term $p(\mathbf{x}_{n-1}, s^{[j]}_n|\mathcal{Y}_{n-1})$,

$$p(\mathbf{x}_{n-1}, s^{[j]}_n|\mathcal{Y}_{n-1}) \quad = \quad \sum_{i=1}^{M} p(\mathbf{x}_{n-1}, s^{[i,j]}_{n-1,n}|\mathcal{Y}_{n-1})$$

$$
\begin{aligned}
&= \sum_{i=1}^{M} p(s_n^{[j]}|s_{n-1}^{[i]}, \mathcal{Y}_{n-1}, \mathbf{x}_{n-1}) p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \\
&= \sum_{i=1}^{M} p_{ij} p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]}) \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}),
\end{aligned}
$$

where the third equality is obtained by the homogeneity of Markov chain $\{s_n\}_{n=1}^{N}$.

Then, together with (3.8), (3.10) becomes

$$
p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) = \frac{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]})}{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}. \tag{3.11}
$$

for $j = 1, \ldots, M$.

Obviously, $p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]})$ is a mixture of Gaussian densities due to (3.9). The derivation here is slightly different from that in Frühwirth-Schnatter (2006, p.408). Furthermore, the mixing proportions in (3.11) are generated from (3.7).

Then, following from (3.4), (3.6) and (3.11), the kth ensemble members of the predicted state and measurement vectors are generated by

$$
\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)} = \mathbf{f}(\widehat{\mathbf{x}}_{n-1|n-1}^{[i,j](k)}; \boldsymbol{\xi}^{[j]}) + \mathbf{u}_n^{[j](k)}, \quad \mathbf{u}_n^{[j](k)} \sim N(\mathbf{0}, \mathbf{Q}^{[j]}), \quad \text{for} \quad i, j = 1, \ldots, M, \tag{3.12}
$$

and

$$
\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)} = \mathbf{h}(\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)}; \boldsymbol{\xi}^{[j]}), \quad \text{for} \quad i, j = 1, \ldots, M, \tag{3.13}
$$

respectively for $k = 1, \ldots, K$. The ensemble member $\widehat{\mathbf{x}}_{n-1|n-1}^{[i,j](k)}$ is generated from $\widehat{\mathbf{x}}_{n-1|n-1}^{[i](k)}$ with density (3.11). Although this approach is similar to that of Kim

(1994), the Gaussian mixture of the filtering density at time $n-1$ arises from the jump in the hidden state s_n which was not pointed out explicitly by Kim (1994).

The conditional mean and variance of the (i,j)th component of the predicted state vector $\mathbf{x}_{n|n-1}$ are approximated by

$$\widehat{\mathbf{x}}_{n|n-1}^{[i,j]} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)}, \tag{3.14}$$

and

$$\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}[i,j]} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)} - \widehat{\mathbf{x}}_{n|n-1}^{[i,j]}\right)\left(\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)} - \widehat{\mathbf{x}}_{n|n-1}^{[i,j]}\right)^{\mathsf{T}}, \tag{3.15}$$

respectively for $i,j = 1,\ldots,M$.

Similar formulae can be applied to the measurement vector $\mathbf{y}_n$. Hence,

$$\widehat{\mathbf{y}}_{n|n-1}^{[i,j]} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)}, \tag{3.16}$$

and

$$\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)} - \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}\right)\left(\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)} - \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}\right)^{\mathsf{T}}, \tag{3.17}$$

for $i,j = 1,\ldots,M$.

The overall one-step ahead prediction mean and variance for the state vector $\mathbf{x}_n$ are

$$\widehat{\mathbf{x}}_{n|n-1}^{[j]} = \frac{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1})\widehat{\mathbf{x}}_{n|n-1}^{[i,j]}}{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1})}, \tag{3.18}$$

and

$$\begin{aligned} & \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}[j]} \\ = & \frac{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1}) \left(\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}[i,j]} + \left(\widehat{\mathbf{x}}_{n|n-1}^{[i,j]} - \widehat{\mathbf{x}}_{n|n-1}^{[j]} \right) \left(\widehat{\mathbf{x}}_{n|n-1}^{[i,j]} - \widehat{\mathbf{x}}_{n|n-1}^{[j]} \right)^{\mathsf{T}} \right)}{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1})}, \end{aligned} \tag{3.19}$$

respectively for $j = 1, \ldots, M$ where the discrete probabilities in (3.18) and (3.19) are estimated by the following formulae recursively:

$$\begin{aligned} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1}) &= \Pr(s_n^{[j]}|s_{n-1}^{[i]}) \sum_{i'=1}^{M} \Pr(s_{n-2,n-1}^{[i',i]}|\mathcal{Y}_{n-1}) \\ &= p_{ij} \sum_{i'=1}^{M} \Pr(s_{n-2,n-1}^{[i',i]}|\mathcal{Y}_{n-1}), \qquad (3.20a) \\ \Pr(s_{n-2,n-1}^{[i,j]}|\mathcal{Y}_{n-1}) &= \frac{p(s_{n-2,n-1}^{[i,j]}, \mathbf{y}_{n-1}|\mathcal{Y}_{n-2})}{\sum_{i=1}^{M}\sum_{j=1}^{M} p(s_{n-2,n-1}^{[i,j]}, \mathbf{y}_{n-1}|\mathcal{Y}_{n-2})}, \qquad (3.20b) \\ p(s_{n-2,n-1}^{[i,j]}, \mathbf{y}_{n-1}|\mathcal{Y}_{n-2}) &= p(\mathbf{y}_{n-1}|\mathcal{Y}_{n-2}, s_{n-2,n-1}^{[i,j]}) \Pr(s_{n-2,n-1}^{[i,j]}|\mathcal{Y}_{n-2}) \\ &\approx \mathcal{N}(\mathbf{y}_{n-1}; \widehat{\mathbf{y}}_{n-1|n-2}^{[i,j]}, \widehat{\mathbf{\Sigma}}_{n-1|n-2}^{\mathbf{yy}[i,j]}) \Pr(s_{n-2,n-1}^{[i,j]}|\mathcal{Y}_{n-2}). \qquad (3.20c) \end{aligned}$$

The derivation of probabilities in (3.20) above follows from that of Kim (1994) where the estimates of prediction mean and variance are replaced by the ensemble estimates (3.16) and (3.17) respectively. The approximation of Gaussian density

arises from the estimated mean and variance and this approximation also holds for the dynamic linear state space model of Kim (1994).

Similar procedures can be applied to obtain the overall mean and variance of the measurement vector $\mathbf{y}_n$, that is,

$$\widehat{\mathbf{y}}_{n|n-1}^{[j]} = \frac{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1})\widehat{\mathbf{y}}_{n|n-1}^{[i,j]}}{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1})}, \tag{3.21}$$

and

$$\begin{aligned} &\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[j]} \\ &= \frac{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1}) \left(\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]} + \left(\widehat{\mathbf{y}}_{n|n-1}^{[i,j]} - \widehat{\mathbf{y}}_{n|n-1}^{[j]} \right) \left(\widehat{\mathbf{y}}_{n|n-1}^{[i,j]} - \widehat{\mathbf{y}}_{n|n-1}^{[j]} \right)^{\mathsf{T}} \right)}{\sum_{i=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_{n-1})}, \end{aligned} \tag{3.22}$$

for $j = 1, \ldots, M$.

Actually, the overall mean and variance of the predicted state and measurement vectors are derived based on the marginalization of prediction densities of $\mathbf{x}_n$ and $\mathbf{y}_n$ respectively (Harrison and Stevens, 1976). The asymptotic properties are shown easily once those of individual components are derived. The asymptotic properties of ensemble estimates are given by the following theorem.

Theorem 3.1. ***Asymptotics of One-step Ahead Predicted States*** *Assume that the filtering density for the state vector* $\mathbf{x}_{n-1}$ *given* $s_{n-1}^{[i]}$ *for* $i \in \{1, \ldots, M\}$, $n \geq 1$ *is*

$$p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]}) = \mathcal{N}(\mathbf{x}_{n-1}; \mathbf{x}_{n-1|n-1}^{[i]}, \mathbf{\Sigma}_{n-1|n-1}^{\mathbf{xx}[i]}).$$

Then, the ensemble mean and variance of the state vector $\mathbf{x}_n$ *by prediction den-*

sity $p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})$ *are estimated by (3.18) and (3.19) respectively. Furthermore, given* $(i,j) \in \{1,\ldots,M\} \times \{1,\ldots,M\}$*, assume that*

(i) $\mathbb{E}\left|\left(f_{n-1|n-1}^{[i,j](k)}\right)_p\right|^{2+\delta} < \infty$ *for some* $\delta > 0$ *with* $p = 1,\ldots,m_x, k = 1,\ldots,K,$

(ii) $\mathbb{E}\left|\left(f_{n-1|n-1}^{[i,j][(k)}\right)_p \left(f_{n-1|n-1}^{[i,j][(k)}\right)_q\right|^{1+\delta} < \infty$ *for some* $\delta > 0$ *with* $p,q = 1,\ldots,m_x;$ $k = 1,\ldots,K,$

Then,

$$\widehat{\mathbf{x}}_{n|n-1}^{[j]} \xrightarrow{a.s.} \mathbf{x}_{n|n-1}^{[j]} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xx}[j]} \xrightarrow{a.s.} \mathbf{\Sigma}_{n|n-1}^{\mathbf{xx}[j]} \textit{ when } K \to \infty,$$

and $\widehat{\mathbf{x}}_{n|n-1}^{[j]}$ *converges in distribution to a mixture of Gaussian densities, that is,*

$$\sqrt{K}\left(\widehat{\mathbf{x}}_{n|n-1}^{[j]} - \mathbf{x}_{n|n-1}^{[j]}\right) \xrightarrow{d} \sum_{i=1}^{M} \alpha_n^{[i,j]} \mathcal{N}(\mathbf{x}_n; \mathbf{0}, \mathbf{\Sigma}_{n|n-1}^{\mathbf{xx}[i,j]}), \textit{ when } K \to \infty,$$

where $\alpha_n^{[i,j]} = p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) / \sum_{\ell=1}^{M} p_{\ell j} \Pr(s_{n-1}^{[\ell]}|\mathcal{Y}_{n-1})$.

Theorem 3.2. ***Asymptotics of One-step Ahead Predicted Measurements*** *Assume that the ensemble prediction mean and variance for the measurement vector* $\mathbf{y}_n$ *are estimated by (3.21) and (3.22) respectively. In addition to the assumptions for ensemble states, given* $(i,j) \in \{1,\ldots,M\} \times \{1,\ldots,M\}$*, assume that*

(iii) $\mathbb{E}\left|\left(h_{n|n-1}^{[i,j](k)}\right)_p\right|^{2+\delta} < \infty$ *for some* $\delta > 0$ *with* $p = 1,\ldots,m_y,\ k = 1,\ldots,K,$

(iv) $\mathbb{E}\left|\left(h_{n|n-1}^{[i,j](k)}\right)_p \left(h_{n|n-1}^{[i,j](k)}\right)_q\right|^{1+\delta} < \infty$ *for some* $\delta > 0$ *with* $p,q = 1,\ldots,m_y,$ $k = 1,\ldots,K,$

Then,

$$\widehat{\mathbf{y}}_{n|n-1}^{[j]} \xrightarrow{a.s.} \mathbf{y}_{n|n-1}^{[j]} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[j]} \xrightarrow{a.s.} \mathbf{\Sigma}_{n|n-1}^{\mathbf{yy}[j]} \textit{ when } K \to \infty,$$

and $\widehat{\mathbf{y}}_{n|n-1}^{[j]}$ *converges in distribution to a mixture of Gaussian densities, that is,*

$$\sqrt{K}\left(\widehat{\mathbf{y}}_{n|n-1}^{[j]} - \mathbf{y}_{n|n-1}^{[j]}\right) \xrightarrow{d} \sum_{i=1}^{M} \alpha_n^{[i,j]} \mathcal{N}(\mathbf{y}_n; \mathbf{0}, \mathbf{\Sigma}_{n|n-1}^{\mathbf{yy}[i,j]}), \text{ when } K \to \infty.$$

where $\alpha_n^{[i,j]} = p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) / \sum_{\ell=1}^{M} p_{\ell j} \Pr(s_{n-1}^{[\ell]}|\mathcal{Y}_{n-1})$.

Proof of Theorems 3.1 and 3.2. Note that the estimators are in the form of sample averages. All ensemble members are considered as heterogeneous but independent since they are generated from a Gaussian mixture where their variances are dependent on the hidden state variable.

To show the consistency of $\widehat{\mathbf{x}}_{n|n-1}$, one need to show that

$$\mathbb{E}\left|\left(\widehat{x}_{n|n-1}^{[i,j](k)}\right)_p\right|^{1+\delta} \leq \mathbb{E}\left|\left(\widehat{x}_{n|n-1}^{[i,j](k)}\right)_p\right|^{2+\delta} < \infty \text{ for } p = 1, \ldots, m_x.$$

By Minkowski's inequality,

$$\begin{aligned}
\mathbb{E}\left|\left(\widehat{x}_{n|n-1}^{[i,j](k)}\right)_p\right|^{2+\delta} &= \mathbb{E}\left|\left(f_{n-1|n-1}^{[i,j](k)}\right)_p + u_n^{[j](k)}\right|^{2+\delta} \\
&\leq \left[\left(\mathbb{E}\left|\left(f_{n-1|n-1}^{[i,j](k)}\right)_p\right|^{2+\delta}\right)^{1/(2+\delta)} + \left(\mathbb{E}\left|\left(u_n^{[j](k)}\right)_p\right|^{2+\delta}\right)^{1/(2+\delta)}\right]^{2+\delta} \\
&\leq \mathbb{E}\left|\left(f_{n-1|n-1}^{[i,j](k)}\right)_p\right| + \mathbb{E}\left|u_n^{[j](k)}\right| \\
&< \infty,
\end{aligned}$$

for $p = 1, \ldots, m_x$. The first term should be finite by condition (i) above and the second term is also finite due to the Gaussianity of $\mathbf{u}_n$. Then, by Markov's SLLN (White, 2001, p.35), the consistency of $\widehat{\mathbf{x}}_{n|n-1}^{[i,j]}$ is shown directly. This also induces the strong consistency of $\widehat{\mathbf{x}}_{n|n-1}^{[j]}$ from (3.18). Similar procedure can be applied to

show the consistency of $\widehat{\mathbf{\Sigma}}^{\mathbf{xx}[i,j]}_{n|n-1}$ by condition (ii) because

$$\widehat{\mathbf{\Sigma}}^{\mathbf{xx}[i,j]}_{n|n-1} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{x}}^{[i,j](k)}_{n|n-1}\right)\left(\widehat{\mathbf{x}}^{[i,j](k)}_{n|n-1}\right)^{\mathsf{T}} - \frac{K}{K-1}\left(\widehat{\mathbf{x}}^{[i,j]}_{n|n-1}\right)\left(\widehat{\mathbf{x}}^{[i,j]}_{n|n-1}\right)^{\mathsf{T}}.$$

Now, the asymptotic distribution of $\widehat{\mathbf{x}}^{[i,j]}_{n|n-1}$ is $\mathcal{N}(\mathbf{x}_n; \mathbf{x}^{[i,j]}_{n|n-1}, \mathbf{\Sigma}^{\mathbf{xx}[i,j]}_{n|n-1})$ with order $O_p(K^{-1/2})$ by using Liapounov's CLT and Cramér-Wold theorem. The details of proofs are referred to White (2001, p.114, 118). Furthermore,

$$\begin{aligned} p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s^{[j]}_n) &= \sum_{i=1}^{M} p(\mathbf{x}_n, s^{[i]}_{n-1}|\mathcal{Y}_{n-1}, s^{[j]}_n) \\ &= \sum_{i=1}^{M} \Pr(s^{[i]}_{n-1}|\mathcal{Y}_{n-1}, s^{[j]}_n) p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s^{[i,j]}_{n-1,n}) \\ &= \frac{\sum_{i=1}^{M} p_{ij} \Pr(s^{[i]}_{n-1}|\mathcal{Y}_{n-1}) p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s^{[i,j]}_{n-1,n})}{\sum_{i=1}^{M} p_{ij} \Pr(s^{[i]}_{n-1}|\mathcal{Y}_{n-1})}, \end{aligned}$$

which is a weighted average of $p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s^{[i,j]}_{n-1,n})$. Then, by the Continuous Mapping Theorem (Pollard, 1984, p.68), the $\widehat{\mathbf{x}}^{[j]}_{n|n-1}$ converges in distribution to a mixture of Gaussian densities, that is,

$$\sqrt{K}\left(\widehat{\mathbf{x}}^{[j]}_{n|n-1} - \mathbf{x}^{[j]}_{n|n-1}\right) \xrightarrow{d} \sum_{i=1}^{M} \alpha^{[i,j]}_n \mathcal{N}(\mathbf{x}_n; \mathbf{0}, \mathbf{\Sigma}^{\mathbf{xx}[i,j]}_{n|n-1}).$$

Similar arguments can be applied to prove for the predicted measurement vector $\widehat{\mathbf{y}}^{[i,j]}_{n|n-1}$ as well while the conditions (iii) and (iv) are used instead. ■

Theorem 3.1 indicates that the one-step ahead prediction of the state vector $\mathbf{x}_n$ can be considered as a combined prediction of the state vector $\mathbf{x}_n$ for each Gaussian error. Hence, its prediction density can be described by Gaussian mix-

ture in the best due to the jump in the hidden state from $s_{n-1}^{[j]}$ to $s_n^{[j]}$. Similar phenomenon can also be found for the measurement vector $\mathbf{y}_n$. Consequently, Bayesian forecasting can be performed in the way of classical forecasting while the forecasting results are combined linearly although they are weighted by discrete probabilities which are determined by ensemble members of the state vector $\mathbf{x}_n$ in previous period.

Following the result of Theorem 3.2, the prediction density of the measurement $\mathbf{y}_n$ is approximated by

$$p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) \approx \frac{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \mathcal{N}(\mathbf{y}_n; \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}, \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]})}{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}. \tag{3.23}$$

The formula is useful in approximating the likelihood function for parameter estimation.

3.3.2 Filtering

In this section, the filtering problem of the state vector $\mathbf{x}_n$ is studied once the measurement data $\mathbf{y}_n$ becomes available. One can expect that the prediction density of the state vector $\mathbf{x}_n$ is a mixture of Gaussian densities. Then, following Kim (1994) and Evensen (2003), the (i,j)th component of the predicted state vector $\widehat{\mathbf{x}}_{n|n-1}$ is updated by the following equation:

$$\widehat{\mathbf{x}}_{n|n}^{[i,j](k)} = \widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)} + \widehat{\mathbf{K}}_n^{[i,j]}(\mathbf{y}_n^{[j](k)} - \widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)}), \quad \text{for } i,j,=1,\ldots,M; k=1,\ldots,K, \tag{3.24}$$

where $\widehat{\mathbf{K}}_n^{[i,j]} = \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xy}[i,j]}(\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]} + \mathbf{R}^{[j]})^{-1}$, $\mathbf{y}_n^{[j](k)} = \mathbf{y}_n + \mathbf{v}_n$, $\mathbf{v}_n \sim N(\mathbf{0}, \mathbf{R}^{[j]})$ and the estimate $\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xy}[i,j]}$ in $\widehat{\mathbf{K}}_n^{[i,j]}$ is given by

$$\widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{xy}[i,j]} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)} - \widehat{\mathbf{x}}_{n|n-1}^{[i,j]}\right)\left(\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)} - \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}\right)^{\mathsf{T}}.$$

Similar to the prediction of the state vector, the conditional mean and variance of the updated state vector $\mathbf{x}_{n|n}^{[i,j]}$ are estimated by

$$\widehat{\mathbf{x}}_{n|n}^{[i,j]} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbf{x}}_{n|n}^{[i,j](k)}, \tag{3.25}$$

and

$$\widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}[i,j]} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{\mathbf{x}}_{n|n}^{[i,j](k)} - \widehat{\mathbf{x}}_{n|n}^{[i,j]}\right)\left(\widehat{\mathbf{x}}_{n|n}^{[i,j](k)} - \widehat{\mathbf{x}}_{n|n}^{[i,j]}\right)^{\mathsf{T}}, \tag{3.26}$$

respectively for $i, j = 1, \ldots, M$.

The overall conditional mean and variance are then estimated by

$$\widehat{\mathbf{x}}_{n|n}^{[j]} = \frac{\sum_{i=1}^{M}\Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_n)\widehat{\mathbf{x}}_{n|n}^{[i,j]}}{\sum_{i=1}^{M}\Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_n)}, \tag{3.27}$$

and

$$\widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}[j]} = \frac{\sum_{i=1}^{M}\Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_n)\left(\widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}[i,j]} + \left(\widehat{\mathbf{x}}_{n|n}^{[i,j]} - \widehat{\mathbf{x}}_{n|n}^{[j]}\right)\left(\widehat{\mathbf{x}}_{n|n}^{[i,j]} - \widehat{\mathbf{x}}_{n|n}^{[j]}\right)^{\mathsf{T}}\right)}{\sum_{i=1}^{M}\Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_n)}, \tag{3.28}$$

respectively $j = 1, \ldots, M$.

Alternatively, one may consider the filtered ensemble members to be generated by the method of maximum a posteriori (MAP) (Sorenson, 1980, p.199) from

(3.5). However, this approach is computationally unfavorable as it can be shown as follows. The posterior density $p(\mathbf{x}_n|\mathcal{Y}_n, s_{n-1,n}^{[i,j]})$ is given by

$$\begin{aligned} p(\mathbf{x}_n|\mathcal{Y}_n, s_{n-1,n}^{[i,j]}) &\propto p(\mathbf{y}_n|\mathbf{x}_n, s_{n-1,n}^{[i,j]}) p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_{n-1,n}^{[i,j]}) \\ &= p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]}) p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_{n-1,n}^{[i,j]}), \end{aligned}$$

where the equality is obtained from Assumption 3.1(i).

Note that both densities are approximately Gaussian from (3.1) and the result of Theorem 3.1. Nevertheless, the maximum of this product of densities may not be easy to achieve due to the nonlinear feature of measurement equation (3.1). Furthermore, M^2 optimizations are required to carry out although more precise estimates of updated state vector can be obtained. Therefore, the method of MAP is not preferred due to its computational complexity.

As pointed out by Kim (1994), the updating equation by (3.24) is not the best estimate under the mixture of Gaussian densities even in the case of linear state space. Nevertheless, under certain regularity conditions, the approximation of updating equation (3.24) can still be reasonable.

Theorem 3.3. ***Filtering*** *In addition to the assumptions in Theorem 3.1, given* $(i,j) \in \{1,\ldots,M\} \times \{1,\ldots,M\}$ *assume that*
(i) $\mathbb{E}\left|\left(\widehat{x}_{n|n}^{[i,j](k)}\right)_p\right|^{2+\delta} < \infty$ *for some* $\delta > 0$ *with* $p = 1,\ldots,m_x, k = 1,\ldots,K$,
(ii) $\mathbb{E}\left|\left(\widehat{x}_{n|n}^{[i,j][(k)}\right)_p \left(\widehat{x}_{n|n}^{[i,j][(k)}\right)_q\right|^{1+\delta} < \infty$ *for some* $\delta > 0$ *with* $p,q = 1,\ldots,m_x$; $k = 1,\ldots,K$,

Then, the ensemble estimates by (3.27) and (3.28) have the following prop-

erties:

$$\widehat{\mathbf{x}}_{n|n}^{[j]} \xrightarrow{a.s.} \mathbf{x}_{n|n}^{[j]} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}[j]} \xrightarrow{a.s.} \mathbf{\Sigma}_{n|n}^{\mathbf{xx}[j]} \ \textit{when } K \to \infty,$$

and $\widehat{\mathbf{x}}_{n|n}^{[j]}$ *converges in distribution to a mixture of Gaussian densities, that is,*

$$\sqrt{K}\left(\widehat{\mathbf{x}}_{n|n}^{[j]} - \mathbf{x}_{n|n}^{[j]}\right) \xrightarrow{d} \sum_{i=1}^{M} \alpha_n^{[i,j]} \mathcal{N}(\mathbf{x}_n; \mathbf{0}, \mathbf{\Sigma}_{n|n}^{\mathbf{xx}[i,j]}), \ \textit{when } K \to \infty,$$

where $\alpha_n^{[i,j]} = p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) / \sum_{\ell=1}^{M} p_{\ell j} \Pr(s_{n-1}^{[\ell]}|\mathcal{Y}_{n-1})$.

Proof. The strong consistency of the mean and variance of filtered ensemble members can be shown by the similar method as in the case of Theorem 3.1. Now, the asymptotic distribution of the mean of filtered ensemble members is shown below. As mentioned in Frühwirth-Schnatter (2006, p.409), the filtering density of the state vector $\mathbf{x}_n$ is

$$\begin{aligned} p(\mathbf{x}_n|\mathcal{Y}_n, s_n^{[j]}) &= \sum_{i=1}^{M} p(\mathbf{x}_n, s_{n-1}^{[i]}|\mathcal{Y}_n, s_n^{[j]}) \\ &= \sum_{i=1}^{M} p(\mathbf{x}_n|\mathcal{Y}_n, s_{n-1,n}^{[i,j]}) \Pr(s_{n-1}^{[i]}|\mathcal{Y}_n, s_n^{[j]}). \end{aligned}$$

As in the proof of Theorem 3.1, the first density is asymptotically Gaussian $\mathcal{N}(\mathbf{x}_n; \mathbf{x}_{n|n}^{[i,j]}, \mathbf{\Sigma}_{n|n}^{[i,j]})$ by condition (i). By the Bayes' Theorem, the conditional probability $\Pr(s_{n-1}^{[i]}|\mathcal{Y}_n, s_n^{[j]})$ is

$$\Pr(s_{n-1}^{[i]}|\mathcal{Y}_n, s_n^{[j]}) = \frac{p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_n)}{\sum_{\ell=1}^{M} p_{\ell j} \Pr(s_{n-1}^{[\ell]}|\mathcal{Y}_n)}.$$

Then, by the Continuous Mapping Theorem (Pollard, 1984, p.68),

$$\sqrt{K}\left(\widehat{\mathbf{x}}_{n|n}^{[j]} - \mathbf{x}_{n|n}^{[j]}\right) \rightarrow_d \sum_{i=1}^{M} \alpha_n^{[i,j]} \mathcal{N}(\mathbf{x}_n; \mathbf{0}, \mathbf{\Sigma}_{n|n}^{[i,j]}).$$

■

From the filtering density of the state vector $\mathbf{x}_n$, the measurement vector $\mathbf{y}_n$ is brute force by Gaussian component $\mathbf{v}_n^{[j]} \sim N(\mathbf{0}, \mathbf{R}^{[j]})$ for $j = 1, \ldots, M$ and this follows from the conventional feature of EnKF. On the other hand, the number of Gaussian components in the mixture is increased by M times. However, this problem was not found in the conventional EnKF of Evensen (1994, 2003) and the mixture ensemble Kalman filter of Bengtsson et al. (2003). Specifically, at time n, the number of Gaussian components becomes M^n.

Therefore, the filtering density should be approximated in an appropriate way such that the number of Gaussian components is controlled over time. In the literature, many authors have suggested various methods to tackle this problem. For example, Kitagawa (1989, 1994) proposed that the Gaussian components in the Gaussian mixture be combined after each filtering stage based on the Kullback-Leibler divergence criterion. Intuitively, two Gaussian components with similar mean and variance were combined and approximated by one Gaussian component with the same combined mean and variance of these two components. On the other hand, Kim (1994) collapsed the mixture filtering density of $\widehat{\mathbf{x}}_{n|n}^{[j]}$ to a single Gaussian by keeping the same filtered mean and variance after each filtering stage. In this chapter, this approach is adopted where the estimated mean and variance are replaced by ensemble mean and ensemble variance respectively, that is,

$$p(\mathbf{x}_n|\mathcal{Y}_n, s_n^{[j]}) \approx \mathcal{N}(\mathbf{x}_n; \widehat{\mathbf{x}}_{n|n}^{[j]}, \widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}[j]}), \tag{3.29}$$

for $j = 1, \ldots, M$ where $\widehat{\mathbf{x}}_{n|n}^{[j]}$ and $\widehat{\mathbf{\Sigma}}_{n|n}^{\mathbf{xx}[j]}$ are the ensemble estimates given by (3.27) and (3.28) respectively.

The recursive operation of prediction and filtering on the state vector $\mathbf{x}_n$ provides the estimates of ensemble mean and variance in (3.18), (3.19), (3.27) and (3.28). The Markov switching state space model operated by EnKF in this way is called the ensemble Markov switching Kalman filter (EnMSKF). Actually, the ensemble mean and variances are adjusted by the marginalization of prediction and filtering densities of the state vector $\mathbf{x}_n$. The idea of the estimation of mean and variance is originated from Kim (1994) in which the linear state space model with Markov switching structure was considered, but now the method is extended to the nonlinear state space model and incorporated into EnKF.

The recursive state estimation procedure is given below:

1. Generate the initial ensemble states. When $n = 0$, given the values of $\mathbf{x}_{0|0}^{[j]}$, $\mathbf{\Sigma}_{0|0}^{[j]}$ and j, draw K ensembles from the density function $\mathcal{N}(\mathbf{x}_0; \mathbf{x}_{0|0}^{[j]}, \mathbf{\Sigma}_{0|0}^{[j]})$ and denote them as $\{\widehat{\mathbf{x}}_0^{[j](k)}\}_{k=1}^{K}$ for $j = 1, \ldots, M$.

2. Generate the modified ensemble states. When $n = 0$, generate the ensemble member $\widehat{\mathbf{x}}_{0|0}^{[i,j](k)}$ for $i, j = 1, \ldots, M$; $k = 1, \ldots, K$ from the Gaussian mixture (3.11) with mixing proportions $p_{ij} \Pr(s_0^{[j]}) / \sum_{i=1}^{M} p_{ij} \Pr(s_0^{[j]}) = p_{ij} / \sum_{i=1}^{M} p_{ij}$.

3. Predict the ensemble states. When $n = 1$, draw an ensemble of K members from the density function $\mathcal{N}(\mathbf{u}_n; \mathbf{0}, \mathbf{Q}^{[j]})$ and denote them as $\{\mathbf{u}_n^{[j](k)}\}_{k=1}^{K}$ for $j = 1, \ldots, M$. Then, generate the ensemble predictions for the (i, j)th Gaussian component with stochastic forces by

$$\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)} = \mathbf{f}(\widehat{\mathbf{x}}_{n-1|n-1}^{[i,j](k)}; \boldsymbol{\xi}^{[j]}) + \mathbf{u}_n^{[j](k)} \quad \text{for } k = 1, \ldots, K.$$

4. Generate the ensemble measurements. When $n = 1$, draw an ensemble of K members from the density function $\mathcal{N}(\mathbf{v}_n; \mathbf{0}, \mathbf{R}^{[j]})$ and denote them as $\{\mathbf{v}_n^{[j](k)}\}_{k=1}^{K}$ for $j = 1, \ldots, M$. Then, generate the ensemble measurements by

$$\mathbf{y}_n^{[j](k)} = \mathbf{y}_n + \mathbf{v}_n^{[j](k)} \quad \text{for } k = 1, \ldots, K.$$

5. Obtain the updated ensemble states by (3.24).

6. Compute the prediction and filtering probabilities of the hidden state s_n by (3.7) and (3.8) respectively, that is, $\Pr(s_1|\mathcal{Y}_0)$ and $\Pr(s_1|\mathcal{Y}_1)$.

7. Collapse the number of components of the filtering density $p(\mathbf{x}_n|\mathcal{Y}_n, s_n^{[j]})$ from M to 1 by using (3.29).

8. Repeat Steps 1 to 7 for time $n = 2, \ldots, N$.

9. At any time $n = 1, \ldots, N$, the prediction and filtered mean and variance of the state vector $\mathbf{x}_n$ are given by (3.18), (3.19), (3.27) and (3.28). Also, the one-step prediction mean and variance of the measurement vector $\mathbf{y}_n$ are given by (3.21) and (3.22) respectively.

3.3.3 Smoothing

Now, a fixed-interval smoother for EnMSKF is derived. The joint smoothing density for the state vector $\mathbf{x}_n$ and the hidden state variable s_n is

$$\begin{aligned} & p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathcal{Y}_N) \\ = & \sum_{j=1}^{M} \int p(\mathbf{x}_n, \mathbf{x}_{n-1}, s_{n-1,n}^{[i,j]}|\mathcal{Y}_N) d\mathbf{x}_n \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{M} \int p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_N) p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_N) d\mathbf{x}_n \\
&= \sum_{j=1}^{M} \int p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_N) p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1}) d\mathbf{x}_n \\
&= \sum_{j=1}^{M} \int p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_N) \frac{p(\mathbf{x}_n, s_n^{[j]}|\mathbf{x}_{n-1}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1}) p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_{n-1})} d\mathbf{x}_n \\
&= p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \sum_{j=1}^{M} \int \frac{p(\mathbf{x}_n, s_n^{[j]}|\mathbf{x}_{n-1}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1}) p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_N)}{p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_{n-1})} d\mathbf{x}_n \\
&= p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \sum_{j=1}^{M} \int p_{ij} \frac{p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]}) p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_N)}{p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) \Pr(s_n^{[j]}|\mathcal{Y}_{n-1})} d\mathbf{x}_n, \qquad (3.30)
\end{aligned}
$$

for $i, j = 1, \ldots, M$ and $n = N, N-1, \ldots, 2$. Note that $\{\mathbf{x}_n, s_n\}_{n=1}^{N}$ is considered as a Markov chain and Assumptions 3.1(i), (ii) and (iii) are used to show the third equality. Specifically,

$$
\begin{aligned}
& p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_N) \\
= \; & \frac{p(\mathbf{x}_{n-1}, \mathbf{x}_n, s_{n-1}^{[i]}, s_n^{[j]}, \mathbf{y}_1, \ldots, \mathbf{y}_N)}{p(\mathbf{x}_n, s_n^{[j]}, \mathbf{y}_1, \ldots, \mathbf{y}_N)} \\
= \; & \frac{p(\mathbf{y}_N|\mathcal{Y}_{N-1}) \ldots p(\mathbf{y}_{n+1}|\mathcal{Y}_n) p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{x}_{n-1}, s_n^{[j]}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1}) p(\mathbf{x}_n, \mathbf{x}_{n-1}, s_n^{[j]}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1})}{p(\mathbf{y}_N|\mathcal{Y}_{N-1}) \ldots p(\mathbf{y}_{n+1}|\mathcal{Y}_n) p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1}) p(\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1})} \\
= \; & \frac{p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{x}_{n-1}, s_n^{[j]}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1}) p(\mathbf{x}_n, \mathbf{x}_{n-1}, s_n^{[j]}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1})}{p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1}) p(\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1})} \\
= \; & \frac{p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]}) p(\mathbf{x}_n, \mathbf{x}_{n-1}, s_n^{[j]}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1})}{p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]}) p(\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1})} \\
= \; & \frac{p(\mathbf{x}_n, \mathbf{x}_{n-1}, s_n^{[j]}, s_{n-1}^{[i]}, \mathcal{Y}_{n-1})}{p(\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1})} \\
= \; & p(\mathbf{x}_{n-1}, s_{n-1}^{[i]}|\mathbf{x}_n, s_n^{[j]}, \mathcal{Y}_{n-1}).
\end{aligned}
$$

In fact, the derivation of posterior density of the smoothed state vector in

(3.30) follows from Harvey (1989) and Kitagawa (1987). Although the smoothing density can be derived by backward recursion, both prediction and filtering densities of the state vector need to be stored. Moreover, in order to obtain the smoothed mean and variance of the state vector, the hidden state variable s_{n-1} should be integrated out. Typically, the smoothed mean and variance of $\mathbf{x}_{n-1}$ and smoothed probability of $s_{n-1}^{[i]}$ are estimated by the Bayesian method. Unlike the fixed-interval smoothers derived by Koopman (1993) and deJong and Shephard (1995) which were based on the minimum mean squared error (MMSE) criterion in which a linear state space model was considered, the joint density (3.30) is non-Gaussian and intractable. As a result, an alternative derivation of smoothed mean of the state vector $\mathbf{x}_{n-1}$ based on the MMSE criterion is provided here.

Following Tanizaki (1996, p.217–220), the smoothed mean of $\mathbf{x}_{n-1}$ given $s_{n-1}^{[i]}$ for $i \in \{1, \ldots, M\}$ is estimated by the MMSE criterion, that is,

$$
\begin{aligned}
& \mathbf{x}_{n-1|N}^{[i]} \\
= \; & \mathbb{E}(\mathbf{x}_{n-1}|\mathcal{Y}_N, s_{n-1}^{[i]}) \\
= \; & \int \mathbf{x}_{n-1} p(\mathbf{x}_{n-1}|\mathcal{Y}_N, s_{n-1}^{[i]}) d\mathbf{x}_{n-1} \\
= \; & \sum_{\mathcal{S}_N} \iint \mathbf{x}_{n-1} p(\mathbf{x}_{n-1}, \mathcal{X}_{N,-(n-1)}, \mathcal{S}_N|\mathcal{Y}_N, s_{n-1}^{[i]}) d\mathcal{X}_{N,-(n-1)} d\mathbf{x}_{n-1} \\
= \; & \sum_{\mathcal{S}_N} \int \mathbf{x}_{n-1} p(\mathcal{X}_N, \mathcal{S}_N|\mathcal{Y}_N, s_{n-1}^{[i]}) d\mathcal{X}_N \\
= \; & \int \mathbf{x}_{n-1} \frac{\sum_{\mathcal{S}_N} p(\mathcal{X}_N, \mathcal{Y}_N, \mathcal{S}_N)}{p(s_{n-1}^{[i]}, \mathcal{Y}_N)} d\mathcal{X}_N \\
= \; & \int \mathbf{x}_{n-1} \frac{\sum_{\mathcal{S}_N} p(\mathcal{Y}_N|\mathcal{X}_N, \mathcal{S}_N) p(\mathcal{X}_N|\mathcal{S}_N) \Pr(\mathcal{S}_N)}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N) p(\mathcal{Y}_N)} d\mathcal{X}_N \\
= \; & \frac{\sum_{\mathcal{S}_N} \int \mathbf{x}_{n-1} p(\mathcal{Y}_N|\mathcal{X}_N, \mathcal{S}_N) p(\mathcal{X}_N|\mathcal{S}_N) \Pr(\mathcal{S}_N) d\mathcal{X}_N}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N) \sum_{\mathcal{S}_N} \int p(\mathcal{Y}_N|\mathcal{X}_N, \mathcal{S}_N) p(\mathcal{X}_N|\mathcal{S}_N) \Pr(\mathcal{S}_N) d\mathcal{X}_N}, \qquad (3.31)
\end{aligned}
$$

for $n = N, N-1, \ldots, 2$, where $\mathcal{X}_{N,-(n-1)} \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_{n-2}, \mathbf{x}_n, \ldots \mathbf{x}_N\}$ indicates a set of all state vectors except the one at time $n-1$ and the sum $\sum_{\mathcal{S}_N}$ denotes $\sum_{s_0=1}^{M} \sum_{s_1=1}^{M} \cdots \sum_{s_N=1}^{M}$.

From Assumption 3.1,

$$\begin{aligned} p\left(\mathcal{X}_N|\mathcal{S}_N\right) &= p\left(\mathbf{x}_0|s_0\right) \prod_{n=1}^{N} p\left(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n\right), \\ p\left(\mathcal{Y}_N|\mathcal{X}_N, \mathcal{S}_N\right) &= \prod_{n=1}^{N} p\left(\mathbf{y}_n|\mathbf{x}_n, s_n\right) \\ \text{and} \qquad \Pr\left(\mathcal{S}_N\right) &= \Pr(s_0) \prod_{n=1}^{N} \Pr\left(s_n|s_{n-1}\right). \end{aligned}$$

From (3.31), the computation of the smoothed state vector $\mathbf{x}_{n-1|N}$ requires (i) the ensemble members $\{\widehat{\mathbf{x}}_{n|n-1}^{[s_{n-1},s_n](k)}\}_{k=1}^{K}$ from the prediction density $p\left(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n\right)$ and (ii) $p(\mathbf{y}_n|\mathbf{x}_n = \widehat{\mathbf{x}}_{n|n-1}^{[s_n](k)})$ for $k = 1, \ldots, K$ and (iii) the smoothed probability $\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N)$. The main difficulty of this smoother is integrating out the hidden state variable s_n for all n.

From (3.18) and (3.20a), one can observe that

$$\mathbf{x}_{n|n-1}^{[j]} = \frac{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \mathbf{x}_{n|n-1}^{[i,j]}}{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}.$$

To integrate out the hidden state $s_n^{[j]}$, the following relation is used:

$$p(\mathbf{x}_n|\mathcal{Y}_{n-1}) = \sum_{j=1}^{M} p(\mathbf{x}_n, s_n^{[j]}|\mathcal{Y}_{n-1}) = \sum_{j=1}^{M} p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) \Pr(s_n^{[j]}|\mathcal{Y}_{n-1}).$$

Hence,

$$\mathbf{x}_{n|n-1} = \sum_{j=1}^{M} \Pr(s_n^{[j]}|\mathcal{Y}_{n-1})\mathbf{x}_{n|n-1}^{[j]}.$$

Then, the ensemble mean and variance of the smoothed state vector given $s_{n-1}^{[i]}$ for $i \in \{1, \ldots, M\}$ are estimated by

$$\widehat{\mathbf{x}}_{n-1|N}^{[i]} = \frac{1}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N)} \sum_{k=1}^{K} \widehat{w}_N^{(k)} \widehat{\mathbf{x}}_{n|n-1}^{(k)}, \tag{3.32}$$

and

$$\widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{xx}[i]} = \frac{1}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N)} \sum_{k=1}^{K} \widehat{w}_N^{(k)} \left(\widehat{\mathbf{x}}_{n|n-1}^{(k)} - \widehat{\mathbf{x}}_{n-1|N}^{[i]}\right) \left(\widehat{\mathbf{x}}_{n|n-1}^{(k)} - \widehat{\mathbf{x}}_{n-1|N}^{[i]}\right)^{\mathsf{T}}, \tag{3.33}$$

respectively for $n = N, N-1, \ldots, 2$, where

$$\begin{aligned}
\widehat{w}_n^{[i,j](k)} &= \frac{p_{ij}\mathcal{N}(\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)}; \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}, \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]}) w_{n-1}^{[i](k)}}{\sum_{k=1}^{K} p_{ij}\mathcal{N}(\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)}; \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}, \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]}) w_{n-1}^{[i](k)}} \quad \text{for } n = 2, \ldots, N, \\
\widehat{w}_1^{[i,j](k)} &= \frac{\Pr(s_0^{[i]}) p_{ij}\mathcal{N}(\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)}; \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}, \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]})}{\sum_{k=1}^{K} \Pr(s_0^{[i]}) p_{ij}\mathcal{N}(\widehat{\mathbf{y}}_{n|n-1}^{[i,j](k)}; \widehat{\mathbf{y}}_{n|n-1}^{[i,j]}, \widehat{\mathbf{\Sigma}}_{n|n-1}^{\mathbf{yy}[i,j]})} \quad \text{for } \; j = 1, \ldots, M, \\
\widehat{w}_n^{[j](k)} &= \sum_{i=1}^{M} \widehat{w}_n^{[i,j](k)} \quad \text{for } \; n = 1, \ldots, N, \\
\widehat{w}_n^{(k)} &= \sum_{j=1}^{M} \widehat{w}_n^{[j](k)}, \\
\widehat{\mathbf{x}}_{n|n-1}^{[j](k)} &= \frac{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)}}{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}, \\
\widehat{\mathbf{x}}_{n|n-1}^{(k)} &= \sum_{j=1}^{M} \Pr(s_n^{[j]}|\mathcal{Y}_{n-1}) \widehat{\mathbf{x}}_{n|n-1}^{[j](k)}
\end{aligned}$$

The measurement vector $\mathbf{y}_{n-1}$ can also be smoothed in a similar way, that is,

$$\begin{aligned}
& \mathbf{y}_{n-1|N}^{[i]} \\
= \; & \mathbb{E}(\mathbf{y}_{n-1}|\mathcal{Y}_N, s_{n-1}^{[i]}) \\
= \; & \sum_{\mathcal{S}_N} \int \mathbf{h}(\mathbf{x}_{n-1}; \xi^{[i]}) \frac{p\left(\mathcal{Y}_N|\mathcal{X}_N, \mathcal{S}_N\right) p\left(\mathcal{X}_N|\mathcal{S}_N\right)}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N) p(\mathcal{Y}_N)} d\mathcal{X}_N \\
= \; & \frac{\sum_{\mathcal{S}_N} \mathbf{h}(\mathbf{x}_{n-1}; \xi^{[i]}) \int p\left(\mathcal{Y}_N|\mathcal{X}_N, \mathcal{S}_N\right) p\left(\mathcal{X}_N|\mathcal{S}_N\right) \Pr(\mathcal{S}_N) d\mathcal{X}_N}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N) \sum_{\mathcal{S}_N} \int p\left(\mathcal{Y}_N|\mathcal{X}_N, \mathcal{S}_N\right) p\left(\mathcal{X}_N|\mathcal{S}_N\right) \Pr(\mathcal{S}_N) d\mathcal{X}_N},
\end{aligned} \tag{3.34}$$

for $n = N, N-1, \ldots, 2$.

By using the same smoothing method as in the state vector $\mathbf{x}_{n-1}$, the ensemble mean and variance of the smoothed measurement vector $\mathbf{y}_{n-1}$ given $s_{n-1}^{[i]}$ for $i \in \{1, \ldots, M\}$ are estimated by

$$\widehat{\mathbf{y}}_{n-1|N}^{[i]} = \frac{1}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N)} \sum_{k=1}^{K} \widehat{w}_N^{(k)} \widehat{\mathbf{y}}_{n|n-1}^{(k)}, \tag{3.35}$$

and

$$\widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{yy}[i]} = \frac{1}{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N)} \sum_{k=1}^{K} \widehat{w}_N^{(k)} \left(\widehat{\mathbf{y}}_{n|n-1}^{(k)} - \widehat{\mathbf{y}}_{n-1|N}^{[i]}\right) \left(\widehat{\mathbf{y}}_{n|n-1}^{(k)} - \widehat{\mathbf{y}}_{n-1|N}^{[i]}\right)^{\mathsf{T}}, \tag{3.36}$$

respectively for $n = N, N-1, \ldots, 2$ where the notation of $\widehat{\mathbf{y}}_{n|n-1}^{(k)}$ follows from $\widehat{\mathbf{x}}_{n|n-1}^{(k)}$ easily.

To compute the smoothed mean and variance of the measurement and state vectors in (3.32), (3.33), (3.35) and (3.36), it is necessary to derive the recursive es-

timation of smoothed probability $\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N)$. Given the value of $i \in \{1, \ldots, M\}$, the smoothed probability for the hidden state variable s_{n-1} is given by

$$
\begin{aligned}
\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N) &= \sum_{j=1}^{M} \Pr(s_{n-1,n}^{[i,j]}|\mathcal{Y}_N) \\
&= \sum_{j=1}^{M} \Pr(s_{n-1}^{[i]}|s_n^{[j]}, \mathcal{Y}_N) \Pr(s_n^{[j]}|\mathcal{Y}_N) \\
&= \sum_{j=1}^{M} \Pr(s_{n-1}^{[i]}|s_n^{[j]}, \mathcal{Y}_{n-1}, \mathcal{Y}_{n:N}) \Pr(s_n^{[j]}|\mathcal{Y}_N) \\
&= \sum_{j=1}^{M} \frac{p(\mathcal{Y}_{n:N}|s_{n-1,n}^{[i,j]}, \mathcal{Y}_{n-1}) \Pr(s_{n-1}^{[i]}|s_n^{[j]}, \mathcal{Y}_{n-1})}{p(\mathcal{Y}_{n:N}|s_n^{[j]}, \mathcal{Y}_{n-1})} \Pr(s_n^{[j]}|\mathcal{Y}_N) \\
&= \sum_{j=1}^{M} \Pr(s_{n-1}^{[i]}|s_n^{[j]}, \mathcal{Y}_{n-1}) \Pr(s_n^{[j]}|\mathcal{Y}_N) \\
&= \frac{\sum_{j=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \Pr(s_n^{[j]}|\mathcal{Y}_N)}{\sum_{j=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})},
\end{aligned}
$$

where $\mathcal{Y}_{n:N} \equiv \{\mathbf{y}_n, \ldots, \mathbf{y}_N\}$. The fifth equality follows from Assumption 3.1(i). Indeed, the recursion of smoothed probability follows from Kim (1994) although the derivation is somehow related to Tanizaki (1996). It is interesting that the smoothed probability of the hidden state s_{n-1} is a mixture of the filtered probability of the hidden state s_n. The mixing proportions are computed by using the results in the prediction and filtering stages, that is, equations (3.7) and (3.8).

The estimation of smoothed states and measurements by (3.32), (3.33), (3.35) and (3.36) is called the ensemble Markov switching Kalman smoother (EnMSKS). At first glance of (3.32) and (3.33), the estimation of mean and variance of the smoothed model state, $\widehat{\mathbf{x}}_{n-1|N}^{[i]}$ and $\widehat{\Sigma}_{n-1|N}^{\mathbf{xx}[i]}$, requires information of the prediction density $p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_{n-1,n}^{[i,j]})$ and the smoothed probability of the hidden state s_{n-1}

only. Implicitly, the smoothed probability $\Pr(s_{n-1}^{[i]}|\mathcal{Y}_N)$ is also dependent on the state vector $\mathbf{x}_n$ from (3.20), Indeed, information from the filtering density is used in the recursive derivation of the prediction density of the model state implicitly. Furthermore, the recursion of smoothed probability of the hidden state variable requires the likelihood $p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n)$ and the transition probabilities $\{p_{ij}\}_{i,j=1}^{M}$. Therefore, the ensemble members of $\widehat{\mathbf{x}}_{n|n-1}^{[i,j](k)}$ and $\widehat{\mathbf{y}}_{n|n-1}^{[i.j](k)}$ and the predicted and filtered probabilities of the hidden state s_n are needed for the computation of the smoothed estimates. The strong consistency of the ensemble estimates is given by the following theorem.

Theorem 3.4. ***Smoothing*** *Assume that the one-step ahead prediction densities of* $\mathbf{x}_n$ *and* $\mathbf{y}_n$ *are given by Theorems 3.1 and 3.2. Given that* $i \in \{1, \ldots, M\}$*, the ensemble estimates (3.32) and (3.33) have the following strong consistency.*

$$\widehat{\mathbf{x}}_{n-1|N}^{[i]} \overset{a.s.}{\to} \mathbf{x}_{n-1|N}^{[i]} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{xx}[i]} \overset{a.s.}{\to} \mathbf{\Sigma}_{n-1|N}^{\mathbf{xx}[i]} \textit{ when } K \to \infty.$$

Similarly, for measurement vector $\mathbf{y}_{n-1}$*,*

$$\widehat{\mathbf{y}}_{n-1|N}^{[i]} \overset{a.s.}{\to} \mathbf{y}_{n-1|N}^{[i]} \quad \textit{and} \quad \widehat{\mathbf{\Sigma}}_{n-1|N}^{\mathbf{yy}[i]} \overset{a.s.}{\to} \mathbf{\Sigma}_{n-1|N}^{\mathbf{yy}[i]} \textit{ when } K \to \infty,$$

Proof. Since the estimates of the smoothed model states are derived from the predicted model states and the number of switching states M is finite, the strong consistency of smoothed ensemble mean and variance follows from the result of Theorems 3.1 and 3.2 directly. ■

3.4 Hybrid Estimation of Model Parameters

From the result of Theorems 3.1 and 3.2, the joint density of observations is approximated by

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_N) = \left\{ \prod_{n=2}^{N} \sum_{j=1}^{M} p(\mathbf{y}_n | \mathcal{Y}_{n-1}, s_n^{[j]}) \Pr(s_n^{[j]} | \mathcal{Y}_{n-1}) \right\} \left\{ \sum_{j=1}^{M} p(\mathbf{y}_1 | s_1^{[j]}) \Pr(s_1^{[j]}) \right\}, \tag{3.37}$$

where $\boldsymbol{\theta} = \left[\{p_{ij}\}_{i,j=1}^{M}, \boldsymbol{\psi}^{[1]\mathsf{T}}, \ldots, \boldsymbol{\psi}^{[M]\mathsf{T}} \right]^{\mathsf{T}}$, $\boldsymbol{\psi}^{[j]\mathsf{T}} = \Big[\Pr(s_0^{[j]}), (\mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}, \mathrm{vech}(\boldsymbol{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}, \boldsymbol{\xi}^{[j]\mathsf{T}}, \mathrm{vech}(\mathbf{R}^{[j]})^{\mathsf{T}}, \mathrm{vech}(\mathbf{Q}^{[j]})^{\mathsf{T}} \Big]$ for $j = 1, \ldots, M$ and the calculations of densities follow from (3.7), (3.8) and (3.23).

The log-likelihood is then given by

$$\log \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right) = \sum_{n=1}^{N} \log \mathcal{L}_n, \tag{3.38}$$

where $\mathcal{L}_n$ stands for the likelihood of the nth observation, that is,

$$\mathcal{L}_n = \sum_{j=1}^{M} \mathcal{N}(\mathbf{y}_n; \widehat{\mathbf{y}}_{n|n-1}^{[j]}, \widehat{\boldsymbol{\Sigma}}_{n|n-1}^{\mathbf{yy}[j]}) \Pr(s_n^{[j]} | \mathcal{Y}_{n-1}),$$

for $n > 1$. When $n = 1$,

$$\mathcal{L}_n = \sum_{j=1}^{M} p(\mathbf{y}_1 | s_1^{[j]}) \Pr(s_1^{[j]}).$$

The prediction probability of the hidden state variable s_n is estimated by (3.7) and (3.8) recursively.

The unknown parameter estimate is the solution of following optimization problem:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathcal{L}\left(\boldsymbol{\theta}; \mathcal{Y}_N\right). \tag{3.39}$$

Under the regularity conditions of Fuh (2006, p.2042–2043), it can be shown that estimate $\widehat{\boldsymbol{\theta}}$ should be strongly consistent and asymptotically normal (Fuh, 2006, Theorems 5 and 6).

Although the maximization problem of (3.38) can be performed by standard optimization algorithm, such as Newton-Raphson, Quasi-Newton and so on. In particular, the expectation-maximum (EM) algorithm is a popular method in the estimation of Markov switching and mixture models. However, this algorithm still faces the similar problem as in conventional optimization algorithm, for example, traps in multiple modes of likelihood functions. Several authors have proposed various variants of EM algorithms to tackle this problem, such as, Monte Carlo EM by Wei and Tanner (1990), stochastic EM by Celeux and Diebolt (1985), supplementary EM by Meng and Rubin (1991) and simulated EM by Ruud (1991). Nevertheless, the EM algorithm faces a problem in the state space model where the model states of full sample are smoothed for conditional expectation of the model states (Shumway and Stoffer, 2006, Chapter 6). Due to the computation burden of the smoother in the nonlinear state space model, the EM algorithm does not seem to be a favorable tool in maximizing the likelihood (3.38).

To avoid the traps in local optimal and multiple optimal solutions and ensure the numerical stability of the estimation process, the global optimization algorithm suggested in Chapter 2 is then implemented in this chapter again to estimate the unknown parameters in (3.38). Indeed, this algorithm sheds lights on both localized stochastic search algorithm and evolutionary strategy and is suitable to tackle the current estimation problem.

3.5 Estimation of Standard Errors

In this section, all parameter estimates by (3.39) are indicated by $\widehat{\ }$. In order to estimate the approximate standard errors of estimated parameters in Section 3.3, the expected Fisher's information matrix is required and defined by

$$\mathcal{I}\left(\boldsymbol{\theta}\right) = \mathbb{E}\left(\frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}} \frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right), \tag{3.40}$$

Unlike Cavanaugh and Shumway (1996) which computed the expected Fisher's information matrix for linear state space model, a typical and consistent approximation of Fisher's information matrix, the observed Fisher's information matrix, is used here, that is,

$$\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \log \mathcal{L}_n}{\partial \boldsymbol{\theta}} \frac{\partial \log \mathcal{L}_n}{\partial \boldsymbol{\theta}^{\mathsf{T}}}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{3.41}$$

Once all estimates of $\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}})$ are obtained, the standard errors of $\widehat{\boldsymbol{\theta}}$ are just the square roots of the diagonal elements of $(\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}))^{-1}$.

To obtain the derivative of the likelihood $\mathcal{L}_n$ with respect to the parameter $\boldsymbol{\theta}$, a recursive procedure is proposed here. From (3.37), we have

$$\begin{aligned}
& \mathcal{L}(\boldsymbol{\theta}; \mathcal{Y}_N) \\
= & \prod_{n=1}^{N} \mathcal{L}_n \\
= & \left\{\prod_{n=2}^{N}\sum_{i=1}^{M}\sum_{j=1}^{M} p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) \Pr(s_n^{[j]}, s_{n-1}^{[i]}|\mathcal{Y}_{n-1})\right\}\left\{\sum_{i=1}^{M}\sum_{j=1}^{M} p_{ij} p(\mathbf{y}_1|s_1^{[j]}) \Pr(s_1^{[j]}, s_0^{[i]})\right\} \\
= & \left\{\prod_{n=2}^{N}\sum_{i=1}^{M}\sum_{j=1}^{M} p_{ij} p(\mathbf{y}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})\right\}\left\{\sum_{i=1}^{M}\sum_{j=1}^{M} p_{ij} p(\mathbf{y}_1|s_1^{[j]}) \Pr(s_0^{[i]})\right\}.
\end{aligned}$$

Then,

$$\mathcal{L}_n \equiv \begin{cases} \sum_{i=1}^{M} \sum_{j=1}^{M} p_{ij} \mathcal{L}_n^{[j]} \Pr(s_0^{[i]}) & \text{for } n = 1, \\ \sum_{i=1}^{M} \sum_{j=1}^{M} p_{ij} \mathcal{L}_n^{[j]} \Pr(s_{n-1}^{[i]} | \mathcal{Y}_{n-1}) & \text{for } n = 2, \ldots, N, \end{cases}$$

where

$$\mathcal{L}_n^{[j]} \equiv \begin{cases} p(\mathbf{y}_1 | s_1^{[j]}) & \text{for } n = 1, \\ p(\mathbf{y}_n | \mathcal{Y}_{n-1}, s_n^{[j]}) & \text{for } n = 2, \ldots, N. \end{cases}$$

Note that $\mathcal{L}_n^{[j]}$ can be expanded to

$$\mathcal{L}_n^{[j]} \equiv \begin{cases} \iint p(\mathbf{y}_1 | \mathbf{x}_1, s_1^{[j]}) p(\mathbf{x}_1 | \mathbf{x}_0, s_1^{[j]}) p(\mathbf{x}_0 | s_1^{[j]}) d\mathbf{x}_0 d\mathbf{x}_1 & \text{for } n = 1, \\ \int p(\mathbf{y}_n | \mathbf{x}_n, s_n^{[j]}) p(\mathbf{x}_n | \mathcal{Y}_{n-1}, s_n^{[j]}) d\mathbf{x}_n & \text{for } n = 2, \ldots, N. \end{cases}$$

When $n = 1$, the derivative of $\mathcal{L}_n$ with respect to $\boldsymbol{\theta}$ is

$$\begin{aligned} \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{M} \sum_{j=1}^{M} p_{ij} \mathcal{L}_1^{[j]} \Pr(s_0^{[i]}) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{M} \mathcal{L}_1^{[j]} \frac{\partial p_{ij} \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}} + \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{\partial \mathcal{L}_1^{[j]}}{\partial \boldsymbol{\theta}} p_{ij} \Pr(s_0^{[i]}) \end{aligned} \tag{3.42}$$

where

$$\frac{\partial p_{ij} \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}} = \frac{\partial p_{ij}}{\partial \boldsymbol{\theta}} \Pr(s_0^{[i]}) + p_{ij} \frac{\partial \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}},$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_1^{[j]}}{\partial \boldsymbol{\theta}} &= \iint \frac{\partial p(\mathbf{y}_1|\mathbf{x}_1, s_1^{[j]})}{\partial \boldsymbol{\theta}} p(\mathbf{x}_1|\mathbf{x}_0, s_1^{[j]}) p(\mathbf{x}_0|s_1^{[j]}) d\mathbf{x}_0 d\mathbf{x}_1 \\
&\quad + \iint p(\mathbf{y}_1|\mathbf{x}_1, s_1^{[j]}) \frac{\partial p(\mathbf{x}_1|\mathbf{x}_0, s_1^{[j]})}{\partial \boldsymbol{\theta}} p(\mathbf{x}_0|s_1^{[j]}) d\mathbf{x}_0 d\mathbf{x}_1 \\
&\quad + \iint p(\mathbf{y}_1|\mathbf{x}_1, s_1^{[j]}) p(\mathbf{x}_1|\mathbf{x}_0, s_1^{[j]}) \frac{\partial p(\mathbf{x}_0|s_1^{[j]})}{\partial \boldsymbol{\theta}} d\mathbf{x}_0 d\mathbf{x}_1.
\end{aligned}$$

The derivative $\partial p(\mathbf{x}_0|s_1^{[j]})/\partial \boldsymbol{\theta}$ is obtained by rearranging the expression for $p(\mathbf{x}_0|s_1^{[j]})$ and differentiating both sides with respect to $\boldsymbol{\theta}$. Then,

$$\begin{aligned}
p(\mathbf{x}_0|s_1^{[j]}) &= \frac{\sum_{i=1}^{M} p_{ij} \Pr(s_0^{[i]}) p(\mathbf{x}_0|s_0^{[i]})}{\sum_{i=1}^{M} p_{ij} \Pr(s_0^{[i]})} \\
\Longrightarrow \frac{\partial p(\mathbf{x}_0|s_1^{[j]})}{\partial \boldsymbol{\theta}} &= \left\{ \sum_{i=1}^{M} \frac{\partial p_{ij} \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}} p(\mathbf{x}_0|s_0^{[i]}) + \sum_{i=1}^{M} p_{ij} \Pr(s_0^{[i]}) \frac{p(\mathbf{x}_0|s_0^{[i]})}{\partial \boldsymbol{\theta}} \right. \\
&\quad \left. - \sum_{i=1}^{M} \frac{\partial p_{ij} \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}} p(\mathbf{x}_0|s_1^{[j]}) \right\} \frac{1}{\sum_{i=1}^{M} p_{ij} \Pr(s_0^{[i]})}
\end{aligned}$$

where

$$\frac{\partial p_{ij} \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}} = \frac{\partial p_{ij}}{\partial \boldsymbol{\theta}} \Pr(s_0^{[i]}) + p_{ij} \frac{\partial \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}}.$$

When $n > 1$, the derivative of $\mathcal{L}_n$ with respect $\boldsymbol{\theta}$ is

$$\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{M} \sum_{j=1}^{M} \mathcal{L}_n^{[j]} \frac{\partial p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}} + \sum_{i=1}^{M} \sum_{j=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) \frac{\partial \mathcal{L}_n^{[j]}}{\partial \boldsymbol{\theta}}, \quad (3.43)$$

where

$$
\begin{aligned}
\frac{\partial \mathcal{L}_n^{[j]}}{\partial \boldsymbol{\theta}} &= \int \frac{\partial p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})}{\partial \boldsymbol{\theta}} d\mathbf{x}_n + \int p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]}) \frac{\partial p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} d\mathbf{x}_n, \\
\frac{\partial p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} &= \int \frac{\partial p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) d\mathbf{x}_{n-1} \\
&\quad + \int p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]}) \frac{\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} d\mathbf{x}_{n-1}.
\end{aligned}
$$

Recall that the recursive relationships of conditional densities of state variables and hidden variables are given by

$$
\begin{aligned}
\Pr(s_n^{[j]}|\mathcal{Y}_{n-1}) &= \sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}), \\
\Pr(s_n^{[j]}|\mathcal{Y}_n) &= \frac{\mathcal{L}_n^{[j]} \Pr(s_n^{[j]}|\mathcal{Y}_{n-1})}{\mathcal{L}_n}, \\
p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]}) &= \int p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]}) p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) d\mathbf{x}_{n-1}, \\
p(\mathbf{x}_n|\mathcal{Y}_n, s_n^{[j]}) &= \frac{p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]}) p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})}{\mathcal{L}_n^{[j]}}, \\
p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) &= \frac{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]})}{\sum_{i=1}^{M} p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})},
\end{aligned}
$$

for $n > 1$. Then, the corresponding derivatives for the first three expressions are

$$
\frac{\partial \Pr(s_n^{[j]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{M} \frac{\partial p_{ij}}{\partial \boldsymbol{\theta}} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1}) + p_{ij} \frac{\partial \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}, \tag{3.44}
$$

$$
\begin{aligned}
\frac{\partial \Pr(s_n^{[j]}|\mathcal{Y}_n)}{\partial \boldsymbol{\theta}} &= \frac{1}{\mathcal{L}_n} \left\{ \frac{\partial \mathcal{L}_n^{[j]}}{\partial \boldsymbol{\theta}} \Pr(s_n^{[j]}|\mathcal{Y}_{n-1}) + \mathcal{L}_n^{[j]} \frac{\partial \Pr(s_n^{[j]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}} \right\} \\
&\quad - \frac{\Pr(s_n^{[j]}|\mathcal{Y}_n)}{\mathcal{L}_n} \frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}},
\end{aligned} \tag{3.45}
$$

$$
\frac{\partial p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} = \int \frac{\partial p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) d\mathbf{x}_{n-1}
$$

$$+\int p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]})\frac{\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}}d\mathbf{x}_{n-1}, \tag{3.46}$$

$$\begin{aligned}\frac{\partial p(\mathbf{x}_n|\mathcal{Y}_n, s_n^{[j]})}{\partial \boldsymbol{\theta}} &= \frac{1}{\mathcal{L}_n^{[j]}}\left\{\frac{\partial p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})}{\partial \boldsymbol{\theta}}p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})\right. \\ &\left. +p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})\frac{\partial p(\mathbf{x}_n|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}}\right\} - \frac{p(\mathbf{x}_n|\mathcal{Y}_n, s_n^{[j]})}{\mathcal{L}_n^{[j]}}\frac{\partial \mathcal{L}_n^{[j]}}{\partial \boldsymbol{\theta}}.\end{aligned} \tag{3.47}$$

respectively.

To consider the derivative $\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]})/\partial \boldsymbol{\theta}$, the expression for $p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]})$ is rearranged.

$$\sum_{i=1}^{M} p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) = \sum_{i=1}^{M} p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]}).$$

Differentiation of both sides with respect to $\boldsymbol{\theta}$ yields

$$\begin{aligned}&\sum_{i=1}^{M}\frac{\partial p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]}) + \sum_{i=1}^{M} p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})\frac{\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} \\ = &\sum_{i=1}^{M}\frac{\partial p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]}) + \sum_{i=1}^{M} p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})\frac{\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]})}{\partial \boldsymbol{\theta}}.\end{aligned}$$

This implies that

$$\begin{aligned}&\frac{\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\theta}} \\ = &\left\{\sum_{i=1}^{M}\frac{\partial p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1}, s_{n-1}^{[i]})\right.\end{aligned}$$

$$+\sum_{i=1}^{M} p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})\frac{\partial p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1},s_{n-1}^{[i]})}{\partial\boldsymbol{\theta}}$$
$$-\sum_{i=1}^{M}\frac{\partial p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial\boldsymbol{\theta}}p(\mathbf{x}_{n-1}|\mathcal{Y}_{n-1},s_n^{[j]})\Bigg\}\frac{1}{\sum_{i=1}^{M}p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}, \tag{3.48}$$

where

$$\frac{\partial p_{ij}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial\boldsymbol{\theta}}=\frac{\partial p_{ij}}{\partial\boldsymbol{\theta}}\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})+p_{ij}\frac{\partial\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial\boldsymbol{\theta}}.$$

Clearly, $\partial p(\mathbf{x}_1|\mathcal{Y}_1,s_1^{[j]})/\partial\boldsymbol{\theta}$ and $\partial\Pr(s_1^{[j]}|\mathcal{Y}_1)/\partial\boldsymbol{\theta}$ are the initial values of the recursive derivations of $\partial\mathcal{L}_n/\partial\boldsymbol{\theta}$ for $n>1$. Consider the derivative $\partial p(\mathbf{x}_1|\mathcal{Y}_1,s_1^{[j]})/\partial\boldsymbol{\theta}$,

$$p(\mathbf{x}_1|\mathcal{Y}_1,s_1^{[j]})=\frac{p(\mathbf{y}_1|\mathbf{x}_1,s_1^{[j]})p(\mathbf{x}_1|s_1^{[j]})}{\mathcal{L}_1^{[j]}}.$$

This implies that

$$\begin{aligned}\frac{\partial p(\mathbf{x}_1|\mathcal{Y}_1,s_1^{[j]})}{\partial\boldsymbol{\theta}} &= \frac{1}{\mathcal{L}_1^{[j]}}\left\{\frac{\partial p(\mathbf{y}_1|\mathbf{x}_1,s_1^{[j]})}{\partial\boldsymbol{\theta}}p(\mathbf{x}_1|s_1^{[j]})+p(\mathbf{y}_1|\mathbf{x}_1,s_1^{[j]})\frac{\partial p(\mathbf{x}_1|s_1^{[j]})}{\partial\boldsymbol{\theta}}\right\}\\ &\quad -\frac{p(\mathbf{x}_1|\mathcal{Y}_1,s_1^{[j]})}{\mathcal{L}_1^{[j]}}\frac{\partial\mathcal{L}_1^{[j]}}{\partial\boldsymbol{\theta}}\end{aligned} \tag{3.49}$$

where

$$\begin{aligned}\frac{\partial p(\mathbf{x}_1|s_1^{[j]})}{\partial\boldsymbol{\theta}} &= \frac{\partial}{\partial\boldsymbol{\theta}}\int p(\mathbf{x}_1|\mathbf{x}_0,s_1^{[j]})p(x_0|s_1^{[j]})d\mathbf{x}_0\\ &= \int\frac{\partial p(\mathbf{x}_1|\mathbf{x}_0,s_1^{[j]})}{\partial\boldsymbol{\theta}}p(x_0|s_1^{[j]})+p(\mathbf{x}_1|\mathbf{x}_0,s_1^{[j]})\frac{\partial p(x_0|s_1^{[j]})}{\partial\boldsymbol{\theta}}d\mathbf{x}_0.\end{aligned}$$

Consider the derivative $\partial \Pr(s_1^{[j]}|\mathcal{Y}_1)/\partial\boldsymbol{\theta}$,

$$\begin{aligned} \Pr(s_1^{[j]}|\mathcal{Y}_1) &= \frac{\mathcal{L}_1^{[j]} \Pr(s_1^{[j]})}{\mathcal{L}_1} \\ \Longrightarrow \frac{\partial \Pr(s_1^{[j]}|\mathcal{Y}_1)}{\partial\boldsymbol{\theta}} &= \frac{1}{\mathcal{L}_1}\left\{\frac{\partial \mathcal{L}_1^{[j]}}{\partial\boldsymbol{\theta}} \Pr(s_1^{[j]}) + \mathcal{L}_1^{[j]} \frac{\partial \Pr(s_1^{[j]})}{\partial\boldsymbol{\theta}}\right\} - \frac{\Pr(s_1^{[j]}|\mathcal{Y}_1)}{\mathcal{L}_1}\frac{\partial \mathcal{L}_1}{\partial\boldsymbol{\theta}} \end{aligned} \tag{3.50}$$

where

$$\begin{aligned} \frac{\partial \Pr(s_1^{[j]})}{\partial\boldsymbol{\theta}} &= \frac{\partial}{\partial\boldsymbol{\theta}} \sum_{i=1}^{M} p_{ij} \Pr(s_0^{[i]}) \\ &= \sum_{i=1}^{M} \frac{\partial p_{ij}}{\partial\boldsymbol{\theta}} \Pr(s_0^{[i]}) + p_{ij} \frac{\partial \Pr(s_0^{[i]})}{\partial\boldsymbol{\theta}}. \end{aligned}$$

Since the densities $p(\mathbf{x}_0|s_0^{[j]})$, $p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})$ and $p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]})$ are Gaussian, the recursive estimation of derivatives of density can be transformed to the estimation of derivatives of log-density easily by rewriting the expressions in the form of

$$\frac{\partial p(\mathbf{z})}{\partial\boldsymbol{\theta}} = \frac{\partial \log p(\mathbf{z})}{\partial\boldsymbol{\theta}} p(\mathbf{z}),$$

where $p(\mathbf{z})$ represents the densities $p(\mathbf{x}_n|\mathcal{Y}_{n-1})$, $p(\mathbf{x}_n|\mathcal{Y}_n)$, $p(\mathbf{x}_1|\mathbf{x}_0)$ and $p(\mathbf{x}_0)$.

Indeed, the log-density functions used in $\partial\mathcal{L}_n/\partial\boldsymbol{\theta}$ are specified as

$$\log p(\mathbf{x}_0|s_0^{[j]})$$

$$
\begin{aligned}
&= -\frac{m_x}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{0|0}^{[j]}| - \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}(\boldsymbol{\Sigma}_{0|0}^{[j]})^{-1}(\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]}),\\
&\log p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})\\
&= -\frac{m_y}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{R}^{[j]}| - \frac{1}{2}\left(\mathbf{y}_n - \mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{R}^{[j]})^{-1}\left(\mathbf{y}_n - \mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}^{[j]})\right),\\
&\log p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]})\\
&= -\frac{m_x}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{Q}^{[j]}| - \frac{1}{2}\left(\mathbf{x}_n - \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{Q}^{[j]})^{-1}\left(\mathbf{x}_n - \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})\right),
\end{aligned}
$$

respectively.

In order to derive the derivatives of log-density with respect to $\boldsymbol{\theta}$, the following expressions (Schott, 1997, p.336) are required:

$$
\begin{aligned}
\frac{\partial\log|\mathbf{A}|}{\partial\mathrm{vech}(\mathbf{A})^{\mathsf{T}}} &= \frac{\partial\log|\mathbf{A}|}{\partial|\mathbf{A}|}\frac{\partial|\mathbf{A}|}{\partial\mathrm{vech}(\mathbf{A})^{\mathsf{T}}} = |\mathbf{A}|^{-1}\mathrm{vec}\left(\mathbf{A}_{\#}^{\mathsf{T}}\right)^{\mathsf{T}}\mathbf{D}_m\\
&= \mathrm{vec}\left(\mathbf{A}^{-1}\right)^{\mathsf{T}}\mathbf{D}_m, \qquad (3.51)\\
\text{and}\quad \frac{\partial\mathrm{vec}(\mathbf{A}^{-1})}{\partial\mathrm{vech}(\mathbf{A})^{\mathsf{T}}} &= -\left(\mathbf{A}^{-1}\otimes\mathbf{A}^{-1}\right)\mathbf{D}_m, \qquad (3.52)
\end{aligned}
$$

with $\mathbf{A}^{-1} = |\mathbf{A}|^{-1}\mathbf{A}_{\#}$ where $\mathbf{D}_m$ denotes the duplication matrix of order m such that $\mathbf{D}_m\mathrm{vech}(\mathbf{A}) = \mathrm{vec}(\mathbf{A})$, $\mathbf{A}_{\#}$ is the adjoint of $\mathbf{A}$ and $\mathbf{A}$ is a $m\times m$ symmetric matrix.

Now, consider the derivatives of log-density with respect to $\boldsymbol{\theta}$.

$$
\begin{aligned}
&\frac{\partial\log p(\mathbf{x}_0|s_0^{[j]})}{\partial\mathbf{x}_{0|0}^{[j]\mathsf{T}}}\\
&= -\frac{1}{2}\frac{\partial}{\partial\mathbf{x}_{0|0}^{[j]\mathsf{T}}}\left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}(\boldsymbol{\Sigma}_{0|0}^{[j]})^{-1}(\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})\right)\\
&= (\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}(\boldsymbol{\Sigma}_{0|0}^{[j]})^{-1}, \qquad (3.53)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \log p(\mathbf{x}_0|s_0^{[j]})}{\partial \mathrm{vech}(\mathbf{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}} \\
= & -\frac{1}{2}\frac{\partial \log \left|\mathbf{\Sigma}_{0|0}^{[j]}\right|}{\partial \mathrm{vech}(\mathbf{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}} - \frac{1}{2}\frac{\partial \mathrm{tr}\left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}(\mathbf{\Sigma}_{0|0}^{[j]})^{-1}(\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})\right)}{\partial \mathrm{vech}(\mathbf{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}} \\
= & -\frac{1}{2}\mathrm{vec}(\mathbf{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}\mathbf{D}_{m_x} - \frac{1}{2}\left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}} \otimes (\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}\right)\frac{\partial \mathrm{vec}(\mathbf{\Sigma}_{0|0}^{[j]})^{-1}}{\partial \mathrm{vech}(\mathbf{\Sigma}_{0|0}^{[j]})} \\
= & -\frac{1}{2}\mathrm{vec}(\mathbf{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}\mathbf{D}_{m_x} \\
& +\frac{1}{2}\left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}} \otimes (\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}\right)\left((\mathbf{\Sigma}_{0|0}^{[j]})^{-1} \otimes (\mathbf{\Sigma}_{0|0}^{[j]})^{-1}\right)\mathbf{D}_{m_x} \\
= & -\frac{1}{2}\mathrm{vec}(\mathbf{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}\mathbf{D}_{m_x} \\
& +\frac{1}{2}\left(\left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}(\mathbf{\Sigma}_{0|0}^{[j]})^{-1}\right) \otimes \left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}(\mathbf{\Sigma}_{0|0}^{[j]})^{-1}\right)\right)\mathbf{D}_{m_x} \qquad (3.54)
\end{aligned}
$$

where $\mathbf{D}_{m_x}\mathrm{vech}(\mathbf{\Sigma}_{0|0}^{[j]}) = \mathrm{vec}(\mathbf{\Sigma}_{0|0}^{[j]})$. The first term in the second equality of $\partial \log p(\mathbf{x}_0|s_0^{[j]})/\partial \mathrm{vech}(\mathbf{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}$ is obtained by (3.51) while the second term is obtained by first using the following identity (Schott, 1997, p.263):

$$
\mathrm{tr}\left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}(\mathbf{\Sigma}_{0|0}^{[j]})^{-1}(\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})\right) = \left((\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})^{\mathsf{T}} \otimes (\mathbf{x}_0 - \mathbf{x}_{0|0}^{[j]})\right)\mathrm{vec}(\mathbf{\Sigma}_{0|0}^{[j]})^{-1}.
$$

and then by (3.52). Note that the derivatives corresponding to the components other than $\mathbf{x}_{0|0}^{[j]}$ and $\mathrm{vech}(\mathbf{\Sigma}_{0|0}^{[j]})$ are zero. Note that

$$
\begin{aligned}
&\frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})}{\partial \boldsymbol{\xi}^{[j]\mathsf{T}}} \\
= & \left(\mathbf{y}_n - \mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{R}^{[j]})^{-1}\frac{\partial \mathbf{h}(\mathbf{x}_n; \boldsymbol{\xi}^{[j]})}{\partial \boldsymbol{\xi}^{[j]\mathsf{T}}}, \\
&\frac{\partial \log p(\mathbf{y}_n|\mathbf{x}_n, s_n^{[j]})}{\partial \mathrm{vech}(\mathbf{R}^{[j]})^{\mathsf{T}}}
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{2}\mathrm{vec}(\mathbf{R}^{[j]})^{\mathsf{T}}\mathbf{D}_{m_y} + \frac{1}{2}\left(\left(\mathbf{y}_n - \mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}} \otimes \left(\mathbf{y}_n - \mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}\right) \\
&\quad \times \left((\mathbf{R}^{[j]})^{-1} \otimes (\mathbf{R}^{[j]})^{-1}\right)\mathbf{D}_{m_y} \\
&= -\frac{1}{2}\mathrm{vec}(\mathbf{R}^{[j]})^{\mathsf{T}}\mathbf{D}_{m_y} + \frac{1}{2}\left(\left(\left(\mathbf{y}_n - \mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{R}^{[j]})^{-1}\right)\right. \\
&\quad \left. \otimes \left(\left(\mathbf{y}_n - \mathbf{h}(\mathbf{x}_n;\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{R}^{[j]})^{-1}\right)\right)\mathbf{D}_{m_y}
\end{aligned}
$$

where $\mathbf{D}_{m_y}\mathrm{vech}(\mathbf{R}^{[j]}) = \mathrm{vec}(\mathbf{R}^{[j]})$. Note that the derivatives corresponding to the components other than $\boldsymbol{\xi}^{[j]}$ and $\mathrm{vech}(\mathbf{R}^{[j]})$ are zero. Then,

$$
\begin{aligned}
&\frac{\partial \log p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]})}{\partial \boldsymbol{\xi}^{[j]\mathsf{T}}} \\
&= \left(\mathbf{x}_n - \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{Q}^{[j]})^{-1}\frac{\partial \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})}{\partial \boldsymbol{\xi}^{[j]\mathsf{T}}}, \qquad (3.55) \\
&\frac{\partial \log p(\mathbf{x}_n|\mathbf{x}_{n-1}, s_n^{[j]})}{\partial \mathrm{vech}(\mathbf{Q}^{[j]})^{\mathsf{T}}} \\
&= -\frac{1}{2}\mathrm{vec}(\mathbf{Q}^{[j]})^{\mathsf{T}}\mathbf{D}_{m_x} + \frac{1}{2}\left(\left(\mathbf{x}_n - \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}} \otimes \left(\mathbf{x}_n - \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}\right) \\
&\quad \times \left((\mathbf{Q}^{[j]})^{-1} \otimes (\mathbf{Q}^{[j]})^{-1}\right)\mathbf{D}_{m_x} \\
&= -\frac{1}{2}\mathrm{vec}(\mathbf{Q}^{[j]})^{\mathsf{T}}\mathbf{D}_{m_x} + \frac{1}{2}\left(\left(\left(\mathbf{x}_n - \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{Q}^{[j]})^{-1}\right)\right. \\
&\quad \left. \otimes \left(\left(\mathbf{x}_n - \mathbf{f}(\mathbf{x}_{n-1};\boldsymbol{\xi}^{[j]})\right)^{\mathsf{T}}(\mathbf{Q}^{[j]})^{-1}\right)\right)\mathbf{D}_{m_x} \qquad (3.56)
\end{aligned}
$$

where $\mathbf{D}_{m_x}\mathrm{vech}(\mathbf{Q}^{[j]}) = \mathrm{vec}(\mathbf{Q}^{[j]})$. Note that the derivatives corresponding to the components other than $\boldsymbol{\xi}^{[j]}$ and $\mathrm{vech}(\mathbf{Q}^{[j]})$ are zero.

By replacing the unknown parameter $\boldsymbol{\theta}$ with the parameter estimate $\widehat{\boldsymbol{\theta}}$ and following the Monte Carlo approach in Chapter 2, the derivative $\partial \mathcal{L}_n/\partial \boldsymbol{\theta}$ can be estimated by

$$\widehat{\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\theta}}} = \sum_{i=1}^{M}\sum_{j=1}^{M} \widehat{\mathcal{L}}_1^{[j]} \widehat{\frac{\partial p_{ij} \Pr(s_0^{[i]})}{\partial \boldsymbol{\theta}}} + \sum_{i=1}^{M}\sum_{j=1}^{M} \widehat{p_{ij}} \widehat{\Pr(s_0^{[i]})} \widehat{\frac{\partial \mathcal{L}_1^{[j]}}{\partial \boldsymbol{\theta}}} \tag{3.57}$$

and

$$\widehat{\frac{\partial \mathcal{L}_n}{\partial \boldsymbol{\theta}}} = \sum_{i=1}^{M}\sum_{j=1}^{M} \widehat{\mathcal{L}}_n^{[j]} \widehat{\frac{\partial p_{ij} \Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})}{\partial \boldsymbol{\theta}}} + \sum_{i=1}^{M}\sum_{j=1}^{M} \widehat{p_{ij}} \widehat{\Pr(s_{n-1}^{[i]}|\mathcal{Y}_{n-1})} \widehat{\frac{\partial \mathcal{L}_n^{[j]}}{\partial \boldsymbol{\theta}}} \tag{3.58}$$

for $n = 2, \ldots, N$.

Note that only the sampling from the conditional density of the state variables $\mathbf{x}_n$ is required during the computation of derivatives of conditional density.

3.6 Determination of M

Up to now, the number of switching regimes M is treated as given. Under certain circumstances, the value of M can be determined by prior information. However, if no prior information is provided for M, some criteria should be applied to determine its value. Otherwise, the number of model parameters becomes extremely large and the estimation procedure becomes more complicated when M is large.

Here, the Bayesian information criterion (BIC) is suggested (for example, Frühwirth-Schnatter, 2006, p.422) to determine the value of M.

$$BIC = -2 \log \mathcal{L}(\widehat{\boldsymbol{\theta}}; \mathcal{Y}_N) + \log(N) \dim(\widehat{\boldsymbol{\theta}}),$$

where $\log \mathcal{L}(\widehat{\boldsymbol{\theta}}; \mathcal{Y}_N)$ is the approximated log-likelihood (3.38) evaluated at $\widehat{\boldsymbol{\theta}}$ and $\dim(\widehat{\boldsymbol{\theta}})$ is the number of parameters in the model. Implicitly, this criterion varies with the value of M. The appropriate value of M is chosen with the smallest

BIC value.

3.7 Empirical Applications

In this section, the assimilation of environmental data of Markov switching state space model is considered. Specifically, the Markov switching vector autoregressive (MS-VAR) model is attempted in this study. Here, the case of a two-regime switching function for all measurements in $\mathbf{y}_n$ is considered, that is, $M = 2$. Specifically, one can consider the following p-order two-regime MS-VAR model for the measurement vector $\mathbf{y}_n$.

$$\mathbf{y}_n = \sum_{j=1}^{p} \mathbf{\Phi}_j^{[s_n]} \mathbf{y}_{n-j} + \mathbf{v}_n, \tag{3.59}$$

for $n = 1, \ldots, N$, where $\mathbf{v}_n \sim N(\mathbf{0}, \mathbf{R}^{[s_n]})$, $\mathbf{y}_n = \left(y_{1,n}, \ldots, y_{m_y,n}\right)^{\mathsf{T}}$ is a $(m_y \times 1)$ column vector, $\mathbf{\Phi}_r$'s are $(m_y \times m_y)$ coefficient matrices for $r = 1, \ldots, p$, the hidden state variable s_n takes values within $\{1, 2\}$. Here, model (3.59) is denoted as MS-VAR(p). In state space form, the measurement and transition equations for (3.59) can be rewritten as

$$\mathbf{y}_n = \begin{bmatrix} \mathbf{I}_{m_y} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \mathbf{x}_n + \mathbf{v}_n, \tag{3.60a}$$

$$\mathbf{x}_n = \begin{bmatrix} \mathbf{y}_n \\ \mathbf{y}_{n-1} \\ \vdots \\ \mathbf{y}_{n-p+1} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_1^{[s_n]} & \cdots & \mathbf{\Phi}_{p-1}^{[s_n]} & \mathbf{\Phi}_p^{[s_n]} \\ \mathbf{I}_{m_y} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{I}_{m_y} & \mathbf{0} \end{bmatrix} \mathbf{x}_{n-1} + \begin{bmatrix} \boldsymbol{\eta}_n \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{3.60b}$$

for $n = 1, \ldots, N$, where $\boldsymbol{\eta}_n \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{\eta[s_n]})$.

In particular, when $M = 1$, models (3.59) and (3.60) are reduced to the specification of conventional vector autoregressive (VAR) model. Therefore, in the current application, the value of M should be greater than 1. According to (3.59) and (3.60), the operation of the model is described as follows: suppose that at time $n-1$, the hidden state variable $s_{n-1} = 1$. Then, two possible outcomes are chosen. The hidden state variable may stay at the same state at time n, that is, $s_n = 1$ or it may jump to the state $s_n = 2$. All the above decisions are made with probabilities p_{11} and p_{12} respectively. Similar explanation can be given for the case of $s_{n-1} = 2$, while the transition probabilities p_{22} and p_{21} are used instead. Afterwards, the vector autoregressive process propagates over time with coefficients $\Phi_j^{[1]}$'s or $\Phi_j^{[2]}$'s correspondingly.

The operation of the state space model (3.60) becomes more complicated because the prediction and filtering of the density of the state vector $\mathbf{x}_n$ and the probability mass of the hidden state s_n are estimated recursively. These features have been described in previous sections, so the repeated description of the operation is not provided here.

In vector form, the unknown parameters of (3.60) in the corresponding likelihood function are represented as

$$\boldsymbol{\theta} = \left[p_{11}, p_{12}, p_{21}, p_{22}, \boldsymbol{\psi}^{[1]\mathsf{T}}, \boldsymbol{\psi}^{[2]\mathsf{T}} \right],$$

where $\boldsymbol{\psi}^{[j]\mathsf{T}} = \left[\Pr(s_0^{[j]}), (\mathbf{x}_{0|0}^{[j]})^{\mathsf{T}}, \mathrm{vech}(\boldsymbol{\Sigma}_{0|0}^{[j]})^{\mathsf{T}}, \boldsymbol{\xi}^{[j]\mathsf{T}}, \mathrm{vech}(\mathbf{R}^{[j]})^{\mathsf{T}}, \mathrm{vech}(\boldsymbol{\Sigma}^{\eta[j]})^{\mathsf{T}} \right]$ and $\boldsymbol{\xi}^{[j]\mathsf{T}} = \left[\mathrm{vec}(\boldsymbol{\Phi}_1^{[j]})^{\mathsf{T}}, \ldots, \mathrm{vec}(\boldsymbol{\Phi}_p^{[j]})^{\mathsf{T}} \right]$ for $j = 1, 2$. These unknown parameters are estimated by the maximization of (3.38).

3.7.1 Numerical Simulation

The behavior of MS-VAR model under the EnMSKF is studied by numerical simulation. For simplicity, the simulated data is generated from the two-regime MS-VAR(1) model which is expressed as

$$\begin{aligned} \mathbf{y}_n &= \mathbf{x}_n + \mathbf{v}_n, \\ \mathbf{x}_n &= \mathbf{\Phi}^{[s_n]}\mathbf{x}_{n-1} + \boldsymbol{\eta}_n, \end{aligned}$$

for $n = 1, \ldots, 100$, where $\mathbf{y}_n = (y_{1,n}, y_{2,n})^{\mathsf{T}}$ is a (2×1) column vector, given $s_n \in \{1, 2\}$, $\mathbf{\Phi}^{[s_n]}$ is a coefficient matrix of order (2×2). The detailed specification of parameters is given below:

$$\mathbf{\Phi}^{[1]} = \begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.7 \end{bmatrix}, \ \mathbf{\Phi}^{[2]} = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.4 \end{bmatrix},$$

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix},$$

$$\Pr(s_0^{[1]}) = 0.3, \Pr(s_0^{[2]}) = 0.7,$$

$$\mathbf{x}_0|s_0^{[1]} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\right), \mathbf{x}_0|s_0^{[2]} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}\right),$$

$$\mathbf{v}_n|s_n^{[1]} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.5 \end{bmatrix}\right), \mathbf{v}_n|s_n^{[2]} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.6 & 0 \\ 0 & 0.3 \end{bmatrix}\right),$$

$$\boldsymbol{\eta}_n|s_n^{[1]} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.6 & -0.2 \\ -0.2 & 1.8 \end{bmatrix}\right), \boldsymbol{\eta}_n|s_n^{[2]} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 1.2 \end{bmatrix}\right).$$

From Figure 3.1, the predicted $y_{1,n}$ and $y_{2,n}$ seem to follow the movement of simulated measurements. When the results of filtering is investigated, the predicted $y_{1,n}$ and $y_{2,n}$ are updated by the new measurements reasonably. This implies that the current linear updating equation can use information of the new measurements efficiently even though the filtering density is a Gaussian mixture. Furthermore, the results in the diagrams show that the lines of the actual and filtered measurements are almost overlapped while the numerical results show their minor discrepancy in values.

3.7.2 Algal Bloom Data

In this section, the daily algal bloom data are extracted from the real time monitoring station at Kat O Island, Hong Kong. In this study, the variables standardized chlorophyll fluorescence (CHL) and standardized dissolved oxygen concentration (DO) are considered as the measurements, that is, $m_y = 2$. Furthermore, the full sampling period of 2000-2004 is splitted into 2000-2001 and 2002-2004 individually. The first period is regarded as the in-sample period and the observations are used for parameter estimation while the second period is the out-of-sample period used to test the prediction performance of the MS-VAR model under the EnMSKF filter. For the numerical stability during estimation, the measurements of standardized CHL and DO is obtained by the subtraction of the measurements by their respective means and divided by their respective standard deviations. Detailed data description was provided by Lee et al. (2003, 2004). Furthermore, the interpretation of the hidden state variable becomes more specific. Suppose that $s_n = 1$ and $s_n = 2$ stand for the non-blooming and blooming status at time n respectively. Then, given the current non-blooming status,

p_{11} and p_{12} represent the transition probabilities of moving to non-blooming and blooming status at next time period respectively. Similar interpretation can be applied to p_{21} and p_{22}.

One feature of the MS-VAR model is that the hidden state s_n can jump instantaneously to the other state. Therefore, the switching between non-blooming and blooming status can occur as a sharp movement of CHL and DO without any transition process. It seems that this is suitable to explain the sharp changes in CHL and DO during algal blooms and also the sudden drops in CHL after algal blooms are finished. In addition, the MS-VAR model under EnMSKF shares a feature of providing prediction probability that an algal bloom occurs given the current condition of water quality. Furthermore, the prediction probability that an algal bloom collapses can also be provided. They are informative in constructing an algal bloom alarm system.

Since the primary objective is the prediction of the measurements CHL and DO by the EnMSKF filter, the lag parameter p of the MS-VAR(p) model is chosen by the one with the smallest root mean squared error (RMSE) in the prediction stage. The in-sample prediction results are presented in Table 3.1.

Table 3.1: In-sample RMSEs of CHL and DO by MS-VAR and VAR models where the columns of CHL and DO represent the RMSEs of CHL and DO respectively. The columns of BOTH represent the RMSEs of both CHL and DO and they are obtained by the square roots of the sums of squared RMSEs of CHL and DO.

	MS-VAR			VAR		
p	CHL	DO	BOTH	CHL	DO	BOTH
1	0.7343	0.6222	0.9625	0.7713	0.6036	0.9794
2	0.7172	0.6012	0.9359	0.6952	0.6127	0.9266
3	0.7834	0.6495	1.0176	0.8117	0.7173	1.0832

For benchmark purpose, the vector autoregressive (VAR) model is chosen.

Table 3.2: In-sample RMSEs of CHL and DO by MS-VAR and VAR models (univariate) where the columns of CHL and represent the RMSEs of CHL and DO respectively.

	MS-VAR		VAR	
p	CHL	DO	CHL	DO
1	0.6975	0.5911	0.7329	0.5595
2	0.6742	0.5772	0.7116	0.5410
3	0.7520	0.6690	0.7016	0.6180

From Table 3.1, the in-sample prediction performance of the MS-VAR model for CHL is generally better than that of the VAR model although it is not true for the dissolved oxygen concentration DO. To identify the lag parameter p of the MS-VAR model, minimum RMSE of BOTH is shown in Table 3.1 when $p = 2$. When the RMSEs of CHL and DO are investigated in details, the RMSE of CHL is minimized when $p = 2$ under both models although this does not hold for DO exactly. Therefore, the lag parameter is chosen to be 2 and the MS-VAR(2) model is used for the interpretation of algal dynamics.

To compare the prediction results, the predictions of CHL and DO by the univariate version of the MS-VAR and VAR models are presented in Table 3.2. Generally, the univariate predictions of CHL and DO improve over the results of multivariate models. This may be caused by the introduction of variation of DO during the prediction of CHL and vice versa and the prediction task will become complicated.

Furthermore, the gain in prediction performance of MS-VAR model over VAR model increases with the lag parameter p. For example, when $p = 2$, the RMSEs of CHL and DO of the MS-VAR model are 0.7172 and 0.6012 respectively while those of the VAR model are 0.6952 and 0.6127 respectively.

The result of estimated parameters of the MS-VAR(2) model is given in the Table 3.3.

Table 3.3: Parameter estimates of MS-VAR(2) model by EnMSKF for CHL and DO where the numbers insides the brackets represent the corresponding elements within the matrices or vectors. The p-values are approximated by the asymptotic Gaussianity.

Parameters	Estimates	Standard Errors	p-values
p_{11}	0.0989	0.0471	0.0359
p_{12}	0.3734	0.2682	0.1639
p_{21}	0.9011	0.6763	0.1828
p_{22}	0.6266	0.2868	0.0289
$\Pr(s_0^{[1]})$	0.0720	0.0383	0.0602
$\mathbf{x}_{0\|0}^{[1]}(1)$	0.0082	0.0046	0.0708
$\mathbf{x}_{0\|0}^{[1]}(2)$	0.0109	0.0074	0.1383
$\mathbf{x}_{0\|0}^{[1]}(3)$	0.0206	0.0115	0.0737
$\mathbf{x}_{0\|0}^{[1]}(4)$	0.0100	0.0050	0.0437
$\mathbf{\Sigma}_{0\|0}^{[1]}(1,1)$	0.2849	0.1839	0.1214
$\mathbf{\Sigma}_{0\|0}^{[1]}(2,1)$	-0.2384	0.1239	0.0544
$\mathbf{\Sigma}_{0\|0}^{[1]}(3,1)$	0.2038	0.1542	0.1862
$\mathbf{\Sigma}_{0\|0}^{[1]}(4,1)$	-0.3098	0.2185	0.1562
$\mathbf{\Sigma}_{0\|0}^{[1]}(2,2)$	0.1995	0.1053	0.0582
$\mathbf{\Sigma}_{0\|0}^{[1]}(3,2)$	-0.1706	0.0801	0.0333
$\mathbf{\Sigma}_{0\|0}^{[1]}(4,2)$	0.2593	0.2157	0.2293

Continued on next page

Continued from previous page

Parameters	Estimates	Standard Errors	p-values
$\mathbf{\Sigma}_{0\|0}^{[1]}(3,3)$	0.1458	0.0698	0.0368
$\mathbf{\Sigma}_{0\|0}^{[1]}(4,3)$	−0.2216	0.1043	0.0336
$\mathbf{\Sigma}_{0\|0}^{[1]}(4,4)$	0.3370	0.2233	0.1314
$\mathbf{\Phi}_{1}^{[1]}(1,1)$	0.6912	0.3744	0.0649
$\mathbf{\Phi}_{1}^{[1]}(2,1)$	−0.0368	0.0194	0.0577
$\mathbf{\Phi}_{1}^{[1]}(1,2)$	−0.0170	0.0088	0.0526
$\mathbf{\Phi}_{1}^{[1]}(2,2)$	0.8562	0.5561	0.1236
$\mathbf{\Phi}_{2}^{[1]}(1,1)$	0.0612	0.0438	0.1625
$\mathbf{\Phi}_{2}^{[1]}(2,1)$	0.0088	0.0056	0.1164
$\mathbf{\Phi}_{2}^{[1]}(1,2)$	0.0026	0.0016	0.0988
$\mathbf{\Phi}_{2}^{[1]}(2,2)$	−0.0121	0.0057	0.0340
$\mathbf{R}^{[1]}(1,1)$	0.2902	0.2322	0.2114
$\mathbf{R}^{[1]}(2,1)$	−0.4526	0.2281	0.0472
$\mathbf{R}^{[1]}(2,2)$	0.7058	0.5678	0.2139
$\mathbf{\Sigma}^{\eta[1]}(1,1)$	0.2837	0.1894	0.1342
$\mathbf{\Sigma}^{\eta[1]}(2,1)$	−0.4447	0.2559	0.0822
$\mathbf{\Sigma}^{\eta[1]}(2,2)$	0.6979	0.4958	0.1593
$\Pr(s_0^{[2]})$	0.9280	0.5403	0.0859
$\mathbf{x}_{0\|0}^{[2]}(1)$	−0.0221	0.0148	0.1358
$\mathbf{x}_{0\|0}^{[2]}(2)$	−0.0175	0.0093	0.0597
$\mathbf{x}_{0\|0}^{[2]}(3)$	−0.0239	0.0163	0.1438
$\mathbf{x}_{0\|0}^{[2]}(4)$	−0.0199	0.0104	0.0566

Continued on next page

Continued from previous page

Parameters	Estimates	Standard Errors	p-values
$\mathbf{\Sigma}_{0\|0}^{[2]}(1,1)$	0.2836	0.1738	0.1026
$\mathbf{\Sigma}_{0\|0}^{[2]}(2,1)$	−0.2404	0.1511	0.1116
$\mathbf{\Sigma}_{0\|0}^{[2]}(3,1)$	0.2064	0.1253	0.0996
$\mathbf{\Sigma}_{0\|0}^{[2]}(4,1)$	−0.3250	0.2172	0.1346
$\mathbf{\Sigma}_{0\|0}^{[2]}(2,2)$	0.2039	0.1001	0.0418
$\mathbf{\Sigma}_{0\|0}^{[2]}(3,2)$	−0.1750	0.0812	0.0312
$\mathbf{\Sigma}_{0\|0}^{[2]}(4,2)$	0.2756	0.1448	0.0570
$\mathbf{\Sigma}_{0\|0}^{[2]}(3,3)$	0.1502	0.0924	0.1041
$\mathbf{\Sigma}_{0\|0}^{[2]}(4,3)$	−0.2365	0.1120	0.0347
$\mathbf{\Sigma}_{0\|0}^{[2]}(4,4)$	0.3726	0.2254	0.0984
$\mathbf{\Phi}_{1}^{[2]}(1,1)$	0.6811	0.4416	0.1230
$\mathbf{\Phi}_{1}^{[2]}(2,1)$	−0.0291	0.0164	0.0765
$\mathbf{\Phi}_{1}^{[2]}(1,2)$	−0.0268	0.0149	0.0725
$\mathbf{\Phi}_{1}^{[2]}(2,2)$	0.8492	0.5477	0.1210
$\mathbf{\Phi}_{2}^{[2]}(1,1)$	0.0699	0.0565	0.2161
$\mathbf{\Phi}_{2}^{[2]}(2,1)$	0.0008	0.0004	0.0384
$\mathbf{\Phi}_{2}^{[2]}(1,2)$	0.0025	0.0019	0.1761
$\mathbf{\Phi}_{2}^{[2]}(2,2)$	−0.0041	0.0023	0.0759
$\mathbf{R}^{[2]}(1,1)$	0.2860	0.1429	0.0453
$\mathbf{R}^{[2]}(2,1)$	−0.4623	0.2579	0.0730
$\mathbf{R}^{[2]}(2,2)$	0.7473	0.5343	0.1619
$\mathbf{\Sigma}^{\eta[1]}(1,1)$	0.2929	0.1366	0.0320

Continued on next page

Continued from previous page

Parameters	Estimates	Standard Errors	p-values
$\mathbf{\Sigma}^{\eta[1]}(2,1)$	-0.4449	0.2074	0.0319
$\mathbf{\Sigma}^{\eta[1]}(2,2)$	0.6759	0.3894	0.0826

As before, the non-blooming and blooming statuses are denoted as $s_n = 1$ and $s_n = 2$ respectively. From Table 3.3, the high estimated transition probability p_{12}, 0.9, indicates that the algal blooms are highly probable to occur within the in-sample period. Furthermore, the estimated p_{22} is over 0.5 showing that once the algal blooms occur, they will last for a period of time. The probability for the initial blooming status, $\Pr(s_0^{[2]})$, is rather high and this is consistent with the frequent occurrence of the algal blooms during 2000-2001.

To investigate the in-sample prediction performance of the MS-VAR model, the prediction and filtering results of the first 100 observations are presented in Figure 3.2. The peaks of CHL at $n = 6$, $n = 33$ and $n = 66$ are predicted reasonably well. Furthermore, the filtering result of CHL is sufficiently close the actual CHL and the similar result is also found in DO. This reflects the updating power of predicted CHL and DO with the EnMSKF. When the prediction and filtering densities of CHL and DO at $n = 66$ in Figures 3.4(a) to (d) are studied, the prediction density of CHL is highly concentrated at its mean with thin tails while that of DO has relatively thick tails. On the other hand, four peaks are found in their filtering densities although they are not shown in the diagrams very clearly.

The frequent occurrence of algal bloom during July 2001 and October 2001 as mentioned in Lee et al. (2003, 2004) and Muttil et al. (2004) encourages

the comparison of prediction performance of the MS-VAR model. Although the RMSEs of CHL and DO are 0.6532 and 0.6012 respectively, it is comparable with the results of Lee et al. (2004) and Muttil et al. (2004). Specifically, the RMSE of CHL was 0.5345 in the paper of Lee et al. (2004) where the artificial neural network was adopted for the prediction. On the other hand, Muttil et al. (2004) provided the RMSE of 0.4793 with the assistance of genetic programming. One should emphasize that a bivariate 2-lag model is used here while the multivariate model with longer lags were used in both cases of Lee et al. (2004) and Muttil et al. (2004).

The in-sample prediction results of CHL and DO in Figure 3.3 show that the movements of CHL and DO are captured reasonably well although the predictions seem to be smoothed after the peak levels of CHL occurred. When the prediction and filtering density of CHL and DO at $n = 431$, which falls within the period of July and October of 2001, are studied, similar observations as in the case of $n = 66$ are found in Figures 3.4(e) to (h). The multimodal filtering densities of CHL and DO represent that both measurements switch between regimes suddenly during the blooming period is provided. On the other hand, the highly concentrated prediction densities of CHL and DO shows that both measurements persist at certain levels for a while before the algal blooms occur.

For the assessment of the out-of-sample prediction of the MS-VAR model, the prediction RMSEs under various lags are compared with those of VAR model and they are given in Table 3.4.

From Table 3.4, the out-of-sample prediction performance of the MS-VAR(2) model is better than that of MS-VAR(1) and MS-VAR(3) models when both measurements are investigated. For the prediction of CHL, the performance of

Table 3.4: Out-of-sample RMSEs of CHL and DO by MS-VAR and VAR models in 2002-2004 where the columns of CHL and DO represent the RMSEs of CHL and DO respectively. The columns of BOTH represent the RMSEs of both CHL and DO and they are obtained by the square roots of the sums of squared RMSEs of CHL and DO.

	MS-VAR			VAR		
p	CHL	DO	BOTH	CHL	DO	BOTH
1	0.5391	0.5415	0.7641	0.4846	0.4979	0.6948
2	0.4830	0.5162	0.7070	0.2447	0.5672	0.6177
3	0.5848	0.5638	0.8123	0.4573	0.4920	0.6717

Table 3.5: Out-of-sample RMSEs of CHL and DO by MS-VAR and VAR models in 2002-2004 (univariate) where the columns of CHL and DO represent the RMSEs of CHL and DO respectively.

	MS-VAR		VAR	
p	CHL	DO	CHL	DO
1	0.4636	0.5091	0.4990	0.4690
2	0.4106	0.4905	0.4683	0.4647
3	0.5030	0.5356	0.3386	0.5021

VAR model denominates that of MS-VAR model generally. However, it may not be the case for the prediction of DO. Specifically, although the prediction performance of VAR(2) model for CHL is better, the MS-VAR(2) model can predict the DO better over the VAR(2) model. To consider the out-of-sample prediction of CHL and DO by the univariate version of both models in Table 3.5, the prediction of the MS-VAR model is dominated by its univariate results while this does not hold for the VAR model, especially the DO concentration.

To illustrate the out-of-sample prediction results, the predictions of CHL and DO during 25 July 2004 and 3 December, 2004 are presented in Figure 3.5 as the variation of both variables changes dramatically within this sampling period. Without the update of model parameters, the prediction of CHL is still quite

reasonable and the pattern of actual measurement of CHL is followed even though the variation of predicted values seems to be reduced. On the other hand, the prediction of DO is more close to the actual measurement of DO despite several sudden drops in DO concentration.

3.8 Conclusion

In this chapter, the ensemble Markov switching Kalman filter (EnMSKF) is derived and this filter is considered as an extension of the dynamic linear model by Kim (1994). Indeed, the Markov switching between nonlinear state space models is considered with the ensemble estimation of mean and variance. Furthermore, the mixture of Gaussian densities of the prediction density arises from the jump in switching regimes at the beginning of the prediction stage. Nevertheless, in order to control the growth in the number of components and hence the number of model parameters over time, the approximation of Gaussian mixture of filtering density by a single Gaussian density as in the case of Kim (1994) and Harrison and Stevens (1976) is maintained. The asymptotic properties of ensemble estimation of mean and variance is also justified by the theorems in this chapter. Once the regularity conditions are satisfied, the ensemble mean and variables are consistent and the asymptotic distribution of ensemble mean is shown to be Gaussian mixture.

As a result of EnMSKF, the ensemble Markov switching Kalman smoother (EnMSKS) is also derived. The complexity of this recursive ensemble smoother is caused by the marginalization of the hidden state s_n. When the hidden state variable s_n is removed, the resulting smoother can also be considered as an im-

provement over the ensemble Kalman smoother (EnKS) by Evensen and van Leeuwen (2000) due to its recursive and computationally efficient feature.

The numerical simulation result of MS-VAR model shows that the predicted measurements under the EnMSKF can capture the movements of simulated measurements. Surprisingly, the filtered measurements are sufficiently close to the simulated measurement and this implies the efficiency of linear updating equation with new coming measurements. For the empirical application of EnMSKF with the algal bloom data in Hong Kong, the filtering performance of EnMSKF again shows its capability of updating prediction results and this is consistent with the numerical simulation result. The in-sample prediction performance of CHL by the MS-VAR model can improve over the benchmark VAR model. Furthermore, the MS-VAR model can provide the insights of transition probabilities between blooming and non-blooming periods. As a comparison of the in-sample prediction, the results are comparable with those of Lee et al. (2004) and Muttil et al. (2004) within the same selection period. However, it should be noted that the results in this chapter are derived from a bivariate 2-lag model while their results relied on the multivariate model with longer lags. From the out-of-sample prediction results, it seems that the MS-VAR model under EnMSKF is more appropriate to provide some insights for the understandings of the algal dynamics while the VAR model is useful for the prediction of CHL and DO. Nevertheless, a nice prediction result of the MS-VAR(2) model is found for DO as compared with the VAR(2) model. Finally, further modelling work can be done on the prediction of chlorophyll fluorescence and dissolved oxygen concentration, for example, allowing the heteroskedasticity of variance over time or using the nonparametric approach.

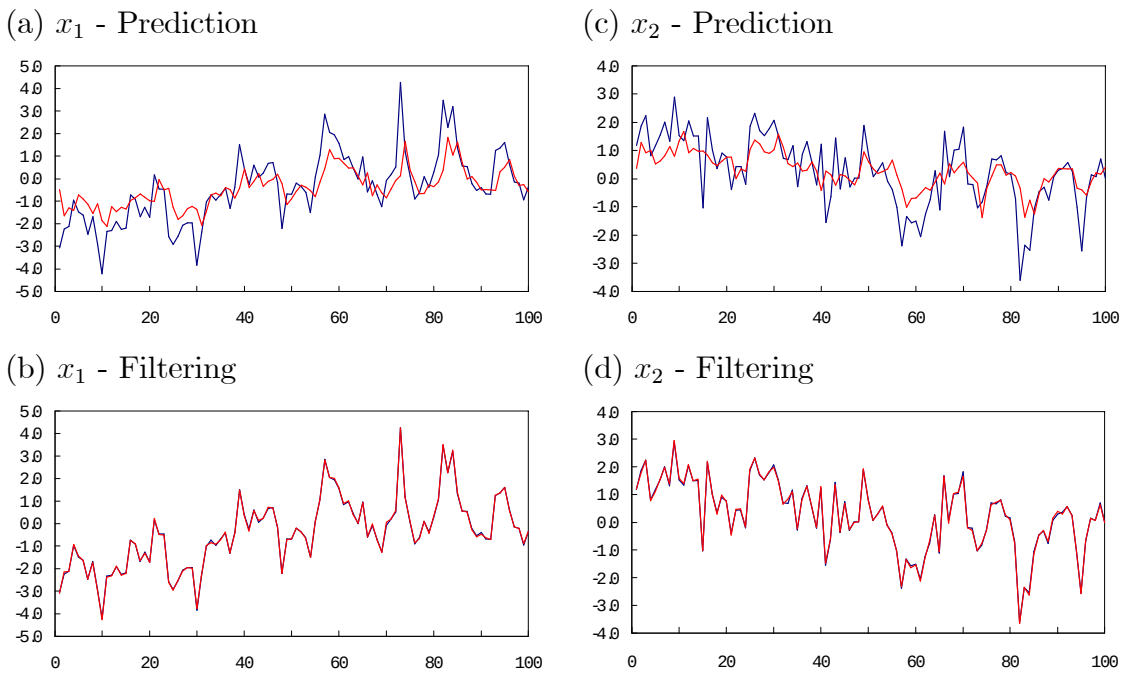


Figure 3.1: Prediction and filtering of simulated MS-VAR(1) model by EnMSKF where the left and right panels represent the results of prediction and filtering of x_1 and x_2 respectively. The red lines indicate the estimated quantities while the blue lines indicate the simulated quantities.

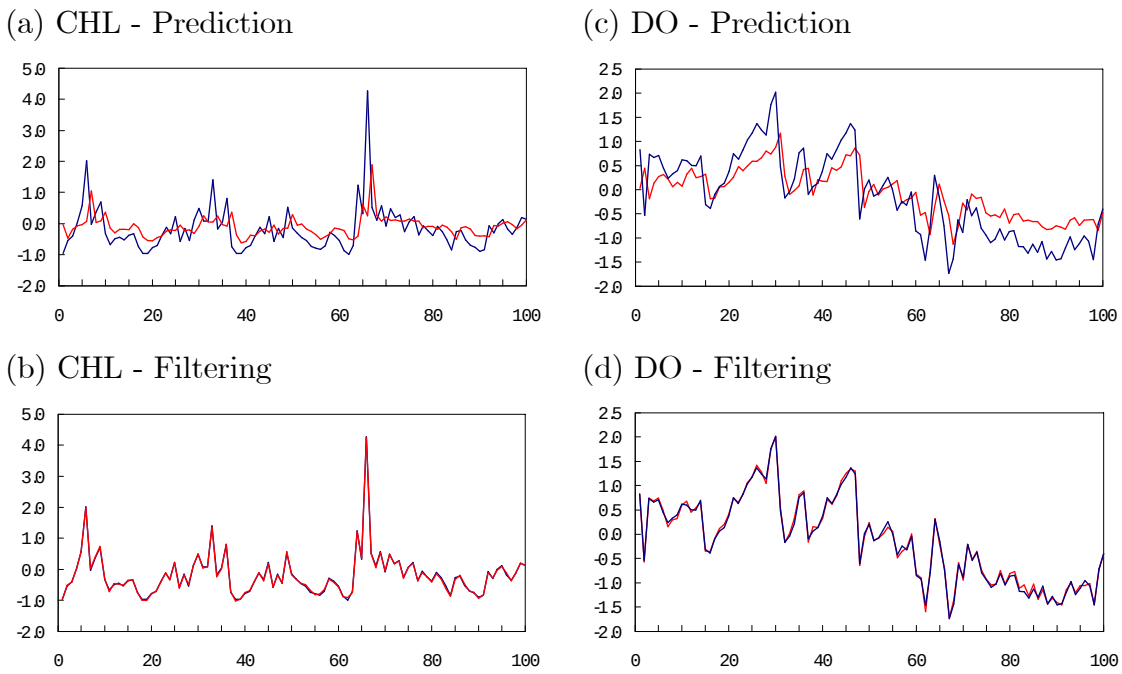


Figure 3.2: In-sample estimation of CHL and DO by EnMSKF with MS-VAR(2) model for the first 100 observations where the left and right panels indicate the prediction and filtering of CHL and DO respectively. Red lines represent the estimated quantities while the blue lines represent the actual qnanitites

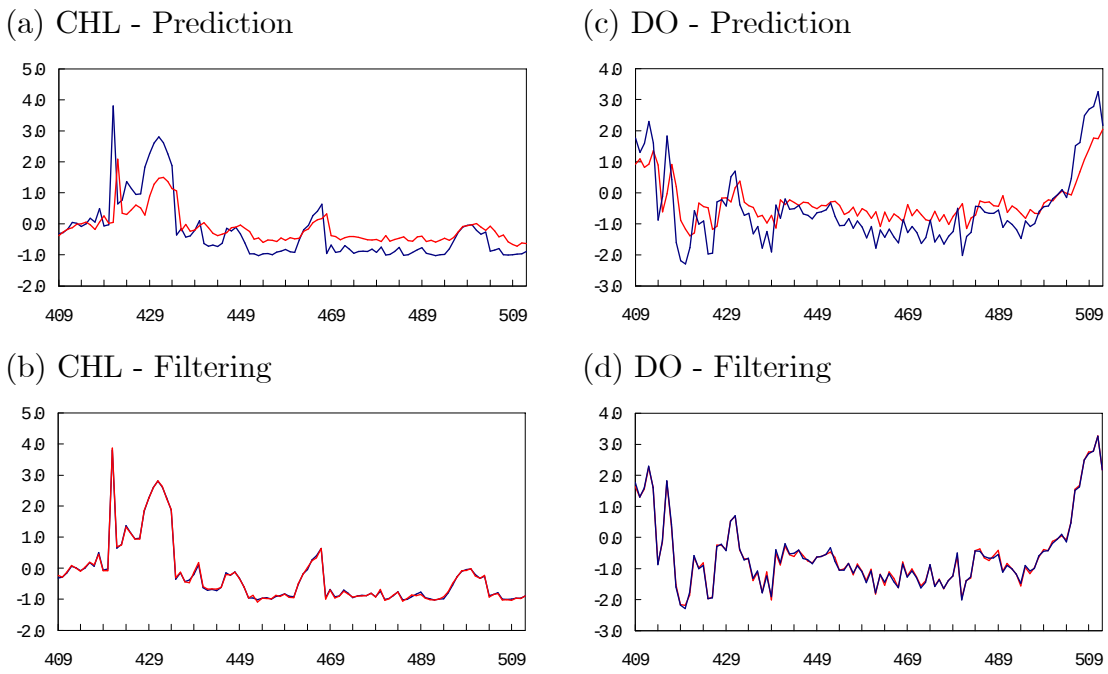
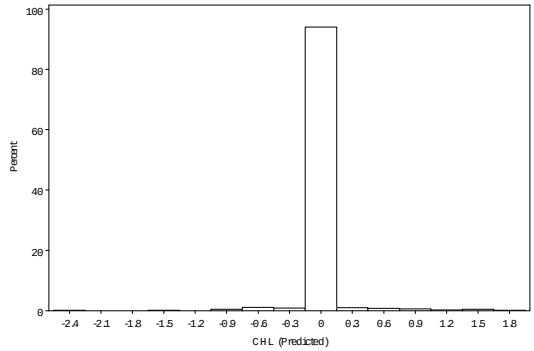
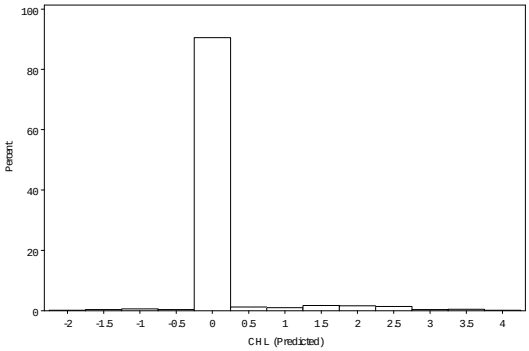


Figure 3.3: In-sample estimation of CHL and DO by EnMSKF with MS-VAR(2) model during 1 July, 2001 and 31 October, 2001 where the left and right panels indicate the prediction and filtering of CHL and DO respectively. Red lines represent the estimated quantities while the blue lines represent the actual qnanitites

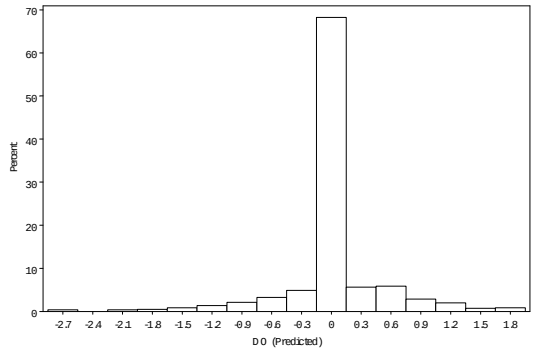
(a) CHL (Predicted), $n = 66$



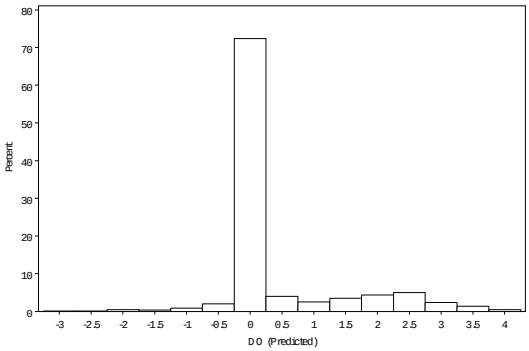
(e) CHL (Predicted), $n = 431$



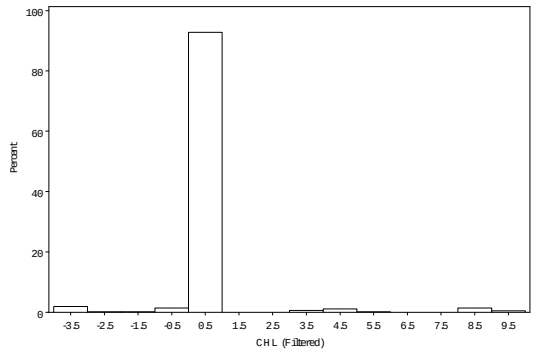
(b) DO (Predicted), $n = 66$



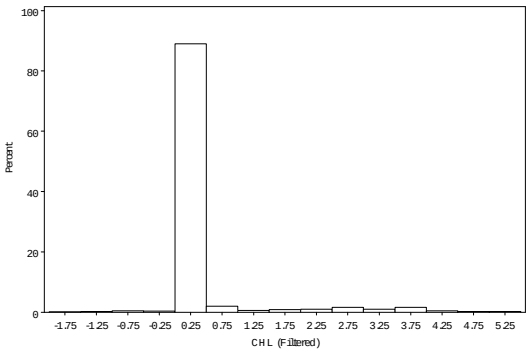
(f) DO (Predicted), $n = 431$



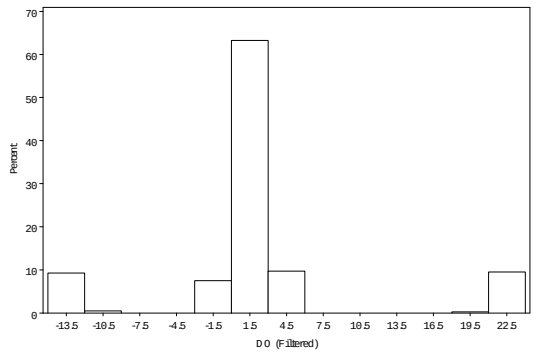
(c) CHL (Filtered), $n = 66$



(g) CHL (Filtered), $n = 431$



(d) DO (Filtered), $n = 66$



(h) DO (Filtered), $n = 431$

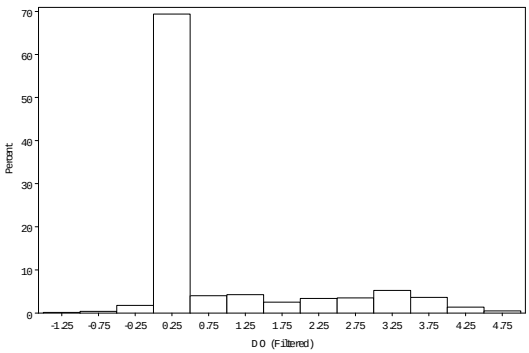


Figure 3.4: Prediction and filtering densities of CHL and DO when $n = 66$ and $n = 431$

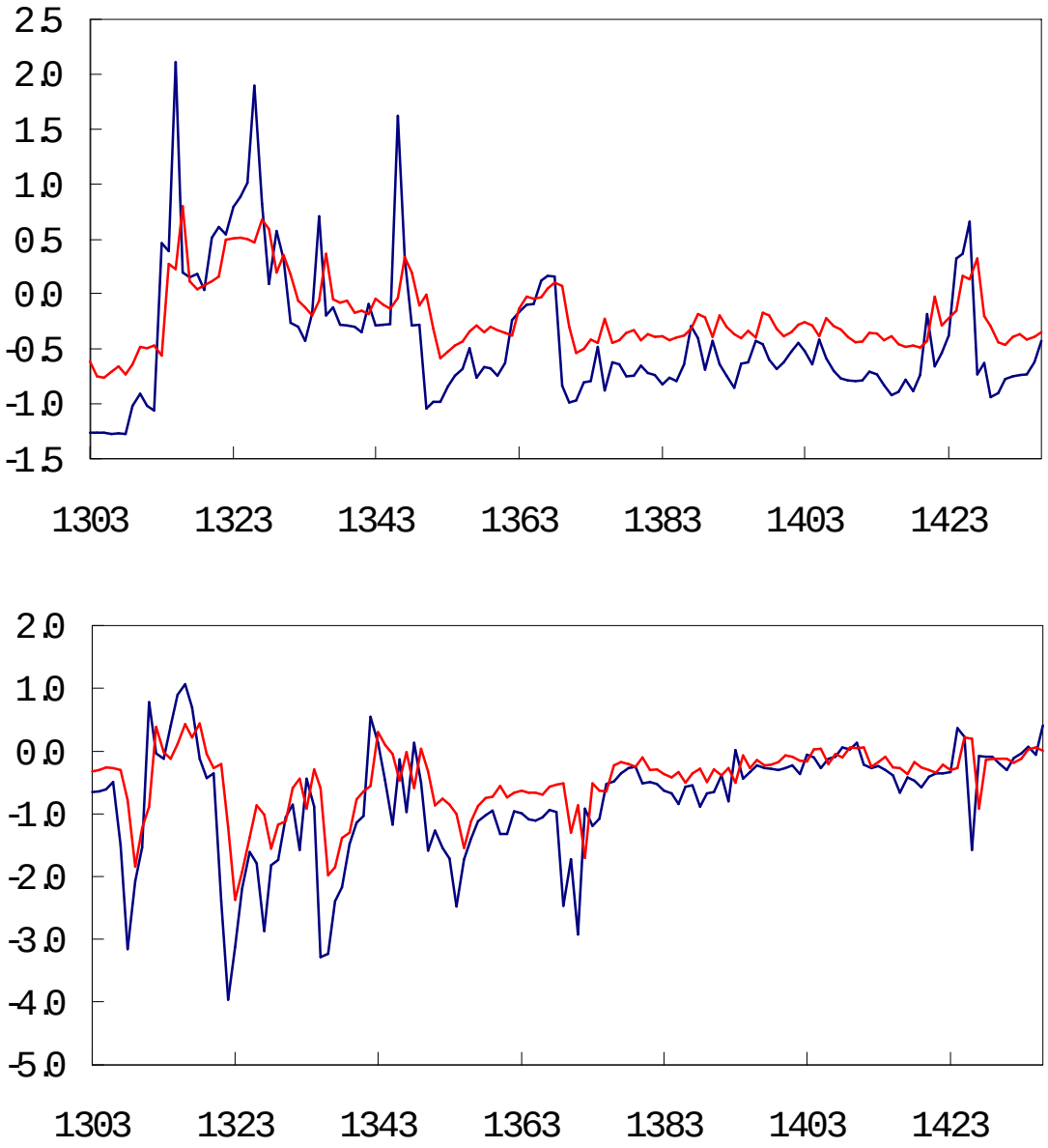


Figure 3.5: Out-of-sample prediction of CHL and DO by EnMSKF during 25th July, 2004 and 31st December, 2004 where the upper and lower panels denote CHL and DO respectively. The red lines represent the predicted quantities and the blue lines represent the actual quantities.

Chapter 4

Conclusion

In this thesis, two new variants of conventional EnKF are derived. Firstly, the ensemble Goldberger-Theil Kalman filter (EnGTKF) has incorporated the feature of Goldberger-Theil's mixed estimation into the conventional EnKF. This results in the nonlinear updating equation which is an extension of conventional linear updating equation. Since the deviation of asymptotic Gaussianity of the ensemble mean and variance may lead to the divergence of the filter, the multivariate Gram-Charlier density is suggested to re-estimate the ensemble mean and variance. This suggestion can also be applied to extend the nonlinear updating equation in EnGTKF although the computation becomes more sophisticated.

Secondly, the Markov switching structure of unobserved regimes is introduced into the EnKF and this results in the derivation of the ensemble Markov switching Kalman filter (EnMSKF). This suggested nonlinear filter can encompass the dynamic linear model by Kim (1994). The mixture of Gaussian densities in the prediction and filtering densities arises from the jump in the unobserved regimes during the prediction stage and this in turn leads to an alternative sampling method for ensembles under the framework of EnKF. The computation of ensemble mean and variances in the prediction stage basically follows from the marginalization of the prediction density as in the case of Harrison and Stevens (1976) and Kim (1994) and this also holds for the computation in the filtering stage. In parallel to the extensions of EnKF, two recursive ensemble smoothers are derived. Specifically, the EnGTKS and EnMSKS are derived from the EnGTKF and EnMSKF respectively.

To estimate the unknown model parameters in the nonlinear state space models, the likelihood for each observation is expressed as a function of prediction error of measurement. Then, it is approximated by the Gaussian density in the case of EnGTKF while it is approximated by the Gaussian density conditional on the value of hidden state variable in the case of EnMSKF. Due to the complexity of the likelihood function, a hybrid estimation procedure is suggested here. This algorithm consists of three components, namely, (1) orthogonal decomposition of error variance matrices; (2) localized stochastic search over parameter space and (3) conventional local optimization algorithm. This procedure can ensure the positive definiteness of error variance matrices and escape from the traps in local optimal during maximization of the likelihood function. The final component is just used to improve the results in previous localized stochastic search. Since the number of iterations is enormously large during the stochastic search, the local optimization can reduce the computational burden.

Focusing on the empirical application of derived filters, the algal bloom data in Hong Kong is assimilated to demonstrate their usefulness. Specifically, only two variables, the standardized chlorophyll fluorescence and standardized dissolved oxygen concentration, are being assimilated. The TV-VSTAR and MS-VAR models are formulated in the form of state space model and adopted with the filters respectively. The TV-VSTAR and MS-VAR models can capture the movements of measurements between blooming and non-blooming periods. By using the VAR model as a benchmark, the in-sample prediction results show that the TV-VSTAR models and MS-VAR models outerperform the VAR model. Surprisingly, the out-of-sample prediction performance of the TV-VSTAR model is still better than the VAR model at the optimal lag chosen by AIC. Furthermore,

the TV-VSTAR model can provide some insights of algal bloom dynamics where the measurements transit between the blooming and non-blooming periods. Furthermore, the MS-VAR model can be used to explain the sudden movements of chlorophyll fluorescence and dissolved oxygen during the algal blooms. The estimation results of both models can be informative to the physical modelling of algal bloom dynamics.

References

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

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. New York: Prentice Hall.

Anderson, J. L. and S. L. Anderson (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilation and forecasts. *Monthly Weather Review 127*, 2741–2758.

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing 50*, 174–188.

Bäck, T. (1996). *Evolutionary Algorithms in Theory and Practice : Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. New York: Oxford University Press.

Bengtsson, T., C. Snyder, and D. Nychka (2003). Toward a nonlinear ensemble filter for high-dimensional systems. *Journal of Geophysical Research 108(STS)*, 2,1–2,10.

Bennett, A. F. (2004). *Inverse Modeling of the Ocean and Atmosphere*. Cambridge: Cambridge University Press.

Bertino, L., G. Evensen, and H. Wackernagel (2002). Combining geostatistics and Kalman filtering for data assimilation in an estuarine system. *Inverse Problems 18*, 1–23.

Bertino, L., G. Evensen, and H. Wackernagel (2003). Sequential data assimilation techniques in oceanography. *International Statistical Review 71*, 223–241.

Blinnikov, S. and R. Moessner (1998). Expansions for nearly Gaussian distributions. *Astronomy & Astrophysics, Supplement Series 130*, 193–205.

Brasseur, P., J. Ballabrera-Poy, and J. Verron (1999). Assimilation of altimetric data in the mid-latitude oceans using the Singular Evolutive Extended Kalman filter with an eddy-resolving, primitive equation model. *Journal of Marine Systems 22*, 269–294.

Brusdal, K., J. Brankart, G. Halberstadt, G. Evensen, P. Brasseur, P. J. van Leeuwen, E. Dombrowsky, and J. Verron (2003). A demonstration of ensemble based assimilation methods with a layered OGCM. *Journal of Marine Systems 40-41*, 253–289.

Burgers, G., P. J. van Leeuwen, and G. Evensen (1998). Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review 126*, 1719–1724.

Cãnizares, R., H. Madsen, H. R. Jensen, and H. J. Vested (2001). Development in operational shelf sea modelling in danish waters. *Estuarine, Coastal and Shelf Science 53*, 595–605.

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

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

Casella, G. and R. L. Berger (2001). *Statistical Inference*. California: Brooks/Cole Pub. Co.

Casulli, V. and E. Cattani (1994). Stability, accuracy and efficiency of a semi-implicit method for three-dimensional shallow water flow. *Computers and Mathematics with Applications 27*, 99–112.

Cavanaugh, J. E. and R. H. Shumway (1996). On computing the expected Fisher information matrix for state-space model parameters. *Statistics and Probability Letters 26*, 347–355.

Celeux, G. and J. Diebolt (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly 2*, 73–82.

Chan, K. S. and H. Tong (2001). *Chaos : A Statistical Perspective*. New York: Springer.

Chattopadhyay, J., R. R. Sarkar, and A. E. Abdllaoui (2002). A delay differential equation model on harmful algal bloom in the presence of toxic substances. *IMA Journal of Mathematics Applied in Medicine and Biology 19*, 137–161.

Chattopadhyay, J., R. R. Sarkar, and S. Pal (2004). Mathematical modelling of harmful algal blooms supported by experimental findings. *Ecological Complexity 1*, 225–235.

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

Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. Technical report, Adaptive Systems Lab., McMaster University, Hamilton, Ontario, Canada.

Courtier, P. and O. Talagrand (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation ii: Numerical results. *Quarterly Journal of the Royal Meteorological Society 113*, 1329–1347.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Crisan, D., P. del Moral, and T. J. Lyons (1999). Non-linear filtering using branching and interacting particle systems. *Markov Processes Related Fields 5*, 293–319.

Dee, D. P. (1995). On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review 123*, 1128–1145.

Dee, D. P. and A. M. D. Silva (1995). Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Monthly Weather Review 127*, 1822–1834.

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

del Moral, P. (1996). Non-linear filtering: Interacting particle solution. *Markov Processes Related Fields 2*, 555–580.

Diderrich, G. T. (1985). The Kalman filter from the perspective of Goldberger-Theil estimators. *American Statistician 39*, 193–198.

Doucet, A., N. de Freitas, and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer.

Doucet, A., J. F. G. deFreitas, K. Murphy, and S. Russell (2000). Rao-Blackwellised particle filtering for dynamic Bayesian networks. In C. Boutilier and M. Goldszmidt (Eds.), *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pp. 176–183. San Francisco: Morgan Kaufmann Publishers Inc.

Eknes, M. and G. Evensen (2002). An ensemble Kalman filter with a 1-D marine ecosystem model. *Journal of Geophysical Research 97(C1)*, 17 905–17 924.

Evensen, G. (1991). Using extended Kalman filter with a multi-layer quasi-geostrophic ocean model. *Journal of Geophysical Research 99 (C5)*, 10143–10162.

Evensen, G. (1994). Sequential data association with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research 99(C5)*, 10 143–10 162.

Evensen, G. (2003). The ensemble Kalman filter : Theoretical formulation and practical implementation. *Ocean Dynamics 53*, 343–367.

Evensen, G. (2007). *Data Assimilation : The Ensemble Kalman Filter*. Berlin: Springer-Verlag.

Evensen, G. and P. J. van Leeuwen (2000). An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review 128*, 1852–1867.

Faddeev, D. K. and V. N. Faddeeva (1963). *Computational Methods of Linear Algebra*. San Francisco: W.H. Freeman.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. London: Chapman & Hall.

Franks, P. (1997). Models of harmful algal blooms. *Limnology and Oceanography 42*, 1273–1282.

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

Fuh, C. (2006). Efficient likelihood estimation in state space models. *The Annals of Statistics 34*, 2026–2068.

Gilks, W. R. and C. Berzuini (2001). Following a moving target-Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society, Series B 63*, 127–146.

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

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

Hamiltion, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics 45*, 39–70.

Harrison, P. J. and C. F. Stevens (1976). Bayesian forecasting (with discussion). *Journal of the Royal Statistical Society, Series B 38*, 205–247.

Harvey, A. C. (1981). *Time Series Models*. London: Philip Allen.

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

Haugen, V. E. and G. Evensen (2002). Assimilation of SLA and SST data into an OGCM for the Indian Ocean. *Ocean Dynamics 52*, 133–151.

He, C., T. Teräsvirta, and A. González (2005). Testing parameters constancy in stationary vector autoregressive models against continuous change. Working Paper Series in Economics and Finance 507, Stockholm School of Economics, Sweden.

Higuchi, T. and J. Fukuda (2003). Monte Carlo mixture Kalman filter and its application to GPS data analysis for space-time inversion. In *Proceedings of the 13th IFAC Symposium on System Identification*, pp. 1299–1304. Oxford: Elsevier Science.

Hüzeler, M. and H. R. Künsch (1998). Monte Carlo approximations for general state-space models. *Journal of Computational and Graphical Statistics 7*, 175–193.

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

Jensen, H. R. (1997). DYNOCS technical report. Regional Model MAS2-CT94-0088.

Jensen, J. L. and N. V. Petersen (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *The Annals of Statistics 27*, 514–535.

Jondeau, E. and M. Rockinger (2001). Gram-Charlier densities. *Journal of Economic Dynamic and Controls 25*, 1457–1483.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions ASME, Series D 82*, 35–45.

Kamykowski, D. and H. Yamazaki (1997). A study of metabolism-influenced orientation in the diel vertical migration of marine dinoflagellates. *Limnology and Oceanography 42*, 1189–1202.

Kanazawa, K., D. Koller, and S. J. Russell (1995). Stochastic simulation algorithms for dynamic probabilistic networks. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, pp. 346–35. San Francisco, CA: Morgan Kaufmann.

Kendall, M. G., A. Stuart, and J. K. Ord (1977). *The Advanced Theory of Statistics, Volume 1: Distribution Theory*. London: Charles Griffin & Co. Ltd.

Kim, C. J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics 60*, 1–22.

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

Kitagawa, G. (1989). Non-Gaussian seasonal adjustment. *Computers and Mathematics with Applications 18*, 503–514.

Kitagawa, G. (1991). A nonlinear smoothing method for time series analysis. *Statistica Sinica 1*, 371–388.

Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics 46*, 605–623.

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

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association 93*, 1203–1215.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–126.

Krolzig, H. M. (1997). *Markov-Switching Vector Autoregressive: Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Lecture Notes in Economics and Mathematical Systems. New York: Springer.

Lee, J. H. W., T. M. K. G. Fernando, and K. T. M. Wong (2004). Real time prediction of coastal algal blooms using artificial neural networks. In S. Y. Liong, K. K. Phoon, and V. Babovic (Eds.), *Proceedings of the 6th International Conference on Hydroinformatics*, pp. 1465–1472. Singapore: World Scientific.

Lee, J. H. W., Y. Huang, M. Dickman, and A. W. Jayawardena (2003). Neural network modelling of coastal algal blooms. *Ecological Modelling 159*, 179–201.

Lundbergh, S., T. Teräsvirta, and D. van Dijk (2003). Time-varying smooth transition autoregressive models. *Journal of Business and Economic Statistics 21*, 104–121.

MacCormick, J. and A. Blake (2000). A probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision 39*, 572–578.

Madsen, D. and R. Cãnizares (1999). Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling. *International Journal of Numerical Methods in Fluids 31*, 961–981.

Mariano, R. S. and H. Tanizaki (1995). Prediction of final data with use of preliminary and/or revised data. *Journal of Forecasting 14*, 351–380.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons Inc.

Meditch, J. S. (1969). *Stochastic Optimal Linear Estimation and Control*. New York: McGraw-Hill.

Meng, X. L. and D. B. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association 86*, 899–909.

Meyn, S. P. and R. L. Tweedie (1993). *Markov Chains and Stochastic Stability*. London: Springer-Verlag.

Miller, R. N. and L. L. Ehret (2002). Ensemble generation for models of multimodal system. *Monthly Weather Review 130*, 2313–2333.

Mitchell, H. L., P. L. Houtekamer, and G. Pellerin (2002). Ensemble size and model-error representation into an ensemble Kalman filter. *Monthly Weather Review 130*, 2791–2808.

Musso, C., N. Oudjane, and F. LeGland (2001). Improving regularised particle filters. In A. Doucet, N. de Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 247–271. New York: Springer.

Muttil, N., J. H. W. Lee, and A. W. Jayawardena (2004). Real-time prediction of coastal algal blooms using genetic programming. In S. Y. Liong, K. K. Phoon, and V. Babovic (Eds.), *Proceedings of the 6th International Conference on Hydroinformatics*, pp. 890–897. Singapore: World Scientific.

Netto, M. L. A., L. Gimeno, and M. J. Mendes (1978). On the optimial and suboptimal nonlinear filtering problem for discrete-time systems. *IEEE Transactions on Automatic Control 23*, 1062–1067.

Perote, J. and E. B. del Brío (2006). Positive definiteness of multivariate densities based on Hermite polynomials. *International Advances in Economic Research 12*, 425–425.

Pham, D. T. (2001). Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review 129*, 1194–1207.

Pham, D. T., J. Verron, and M. C. Roubaud (1998). A singular evolutive extended Kalman filter with EOF initialization for data assimilation in oceanography. *Journal of Marine Systems 16*, 323–340.

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

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons Inc.

Rudolph, G. (1992). On correlated mutations in evolution strategies. In R. Männer and B. Manderick (Eds.), *Parallel Problem Solving from Nature, 2 : Proceedings of the Second Conference on Parallel Problem Solving from Nature*, pp. 105–114. Amsterdam: North-Holland.

Ruud, P. A. (1991). Extension of estimation methods using the EM algorithm. *Journal of Econometrics 49*, 305–341.

Schott, J. R. (1997). *Matrix Analysis for Statistics*. New York: John Wiley & Sons Inc.

Schwefel, H. P. (1981). *Numerical Optimization of Computer Models*. New York: John Wiley & Sons Inc.

Shumway, R. H. and D. S. Stoffer (2006). *Time Series Analysis and its Applications : with R Examples*. New York: Springer.

Sorenson, H. W. (1980). *Parameter Estimation : Principle and Problems*. New York: Marcel Dekker.

Sorenson, H. W. and D. L. Alspach (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica 7*, 465–479.

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. New Jersey: Wiley-Interscience.

Talagrand, O. and P. Courtier (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation i: Theory. *Quarterly Journal of the Royal Meteorological Society 113*, 1311–1328.

Tanizaki, H. (1996). *Nonlinear Filters : Estimation and Applications*. Berlin: Springer-Verlag.

Tanizaki, H. (1999). On the nonlinear and non-normal filter using rejection sampling. *IEEE Transactions on Automatic Control 44*, 314–319.

Tanizaki, H. and R. S. Mariano (1994). Prediction, filtering and smoothing in nonlinear and nonnormal cases using Monte-Carlo integration. *Journal of Applied Econometrics 9*, 163–179.

Tanizaki, H. and R. S. Mariano (1998). Nonlinear adn non-Gaussian state-space modeling with Monte Carlo integration. *Journal of Econometrics 83*, 263–290.

Theil, H. and A. S. Goldberger (1961). On pure and mixed statistical estimation in economics. *International Economic Review 2*, 65–78.

Thomann, R. V. and J. A. Mueller (1987). *Principles of Surface Water Quality Modeling and Control*. New York: Harper & Row.

Titterington, D. M., A. M. F. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. New York: Springer-Verlag.

Treebushny, D. and H. Madsen (2003). A new reduced rank square root Kalman filter for data assimilation in mathematical models. In *Lecture Notes in Computer Science*, Volume 2657, pp. 482–491. Heidelberg: Springer Berlin.

van Leeuwen, P. J. and G. Evensen (1996). Data assimilation and inverse methods in terms of a probabilistic formulation. *Monthly Weather Review 124*, 2898–2913.

Verlaan, M. (1998). *Efficient Kalman Filtering Algorithms for hydrodynamic Models*. Ph. D. thesis, Delft University of Technology, The Netherlands.

Verlaan, M. and A. Heemink (1995). Reduced rank square root filters for large scale data assimilation problems. In M. Ghil, K. Ide, A. Bennett, and P. Courtier (Eds.), *2nd International Symposium on Assimilation of Observations in Meteorology and Oceanography*, pp. 247–252. Tokyo: Meteorological Society of Japan.

Verlaan, M. and A. W. Heemink (1997). Tidal flow forecasting using reduced rank square root filters. *Stochastic Hydrology and Hydraulics 11*, 349–368.

Verlaan, M. and A. W. Heemink (2001). Nonlinearity in data assimilation applications: A practical method for analysis. *Monthly Weather Review 129*, 1578–1589.

Vose, M. D. (1992). Modeling simple genetic algorithms. *Foundations of Genetic Algorithms 2*, 63–73.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005). Improved treatment of uncertainty in hydrologic modeling : Combining the strengths of global optimization and data assimilation. *Water Resources Research 41*, doi:10.1029/2004WR003059.

Wei, G. G. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and poor man's and data augmentation algorithm. *Journal of the American Statistical Association 85*, 699–704.

White, H. (2001). *Asymptotic Theory for Econometricians*. California: Academic Press.

Wishner, R. P., J. A. Tabaczynski, and M. Athans (1969). A comparison of three non-linear filters. *Automatica 5*, 487–496.

Wolf, T., J. Sénégas, L. Bertino, and H. Wackernagel (2001). Application of data assimilation to three-dimensional hydrodynamics: the case of the Orda lagoon. In P. Monestiez, D. Allard, and R. Froidevaux (Eds.), *GeoENV III: Geostatistics for Environmental Applications*, pp. 157–168. Amsterdam: Kluwer Academic.

Yamamoto, T. and M. Okai (2000). Effects of diffusion of upwelling on the formation of red tides. *Journal of Plankton Research 22*, 363–380.