PAPER *Surveys on Discovery Science*

Knowledge Discovery and Self-Organizing State Space Model

Tomoyuki HIGUCHI[†] *and* **Genshiro KITAGAWA**[†], *Nonmembers*

SUMMARY A hierarchical structure of the statistical models involving the parametric, state space, generalized state space, and self-organizing state space models is explained. It is shown that by considering higher level modeling, it is possible to develop models quite freely and then to extract essential information from data which has been difficult to obtain due to the use of restricted models. It is also shown that by rising the level of the model, the model selection procedure which has been realized with human expertise can be performed automatically and thus the automatic processing of huge time series data becomes realistic. In other words, the hierarchical statistical modeling facilitates both automatic processing of massive time series data and a new method for knowledge discovery.

key words: non-Gaussian non-linear time series model, generalized state space model, self-organizing state space model, hierarchical structure model, Bayesian Model

1. Knowledge Discovery and Statistical Modeling

1.1 Role of Statistical Model in Knowledge Discovery

The scientific theories developed so far by human being should be considered as reasonable approximations to the phenomena in the realm of the human recognition, rather than exact description of the truth. Therefore, there is a possibility of obtaining scientific discovery from the process of comparing existing theory with actual data [2]. Actually, the main objective of statistics before 20th century was the discovery of low based on massive observations. On the other hand, the modern statistics aimed at possibility of precise inference by small sample based on rigorously designed experiment or sample survey. However, the recent progress of information technologies required anew to establish a method for automatic processing of massive observations and a method of knowledge discovery based on it.

Data always contain some errors and thus only by proper processing of the errors, it becomes possible to separate the essential or universal part from the errors that occurred only for that data. The treatment of observation errors in scientific research has been extensively analyzed from the previous century. However, in the analysis of complex, massive and/or multivariate phenomena with nonstationarity or nonlinearity, it is almost impossible to express it from simple scientific theory and consider the difference from the actual data as observation noises. In such a case, by reasonably decomposing the data into a part that can be explained from the existing knowledge and other, the possibility of new scientific discovery emerges.

In the frontiers of sciences, unexpected phenomena appear at the very limit of the errors. Therefore, by simply considering the unknown part as the observation errors, it is almost impossible to find out the clue to the discovery. For the discovery in the frontiers of sciences, it is crucially important to express our expectation on the unknown part as a form of model, and perform the extraction of the information very actively.

Needless to say, in such a statistical modeling, use of "appropriate" model is crucially important. If the data did not contain any errors, or the objective of our analysis was just to describe the phenomena precisely, it is sufficient to use the model with highest ability of "description". However, in the actual analysis our objective is often to extract or discover a more "universal" knowledge. In the statistical science, this essential part is considered from the predictive point of view.

To obtain good models, it is necessary to develop appropriate model class and model evaluation criterion. Further, to make the modeling practical, it is also necessary to develop efficient computational method.

1.2 Computer Intensive Computational Methods

In general, flexible models with high ability of description inevitably contain increasingly many unknown parameters and require huge amount of computations for the estimation of them. In this article, we will consider the modeling of the time series which is the most important in the statistical analysis of massive data. For flexible modeling of time series, it is necessary to use a model with the number of parameters proportional to the data length N. For such models, the ordinary computational methods requires the computation with order $O(N^3)$, and are obviously unrealistic to apply. Without skillful computational methods which are based on the essential mathematical structure of the model, it is sometimes impossible to obtain reasonable estimates of the unknowns. Further, to treat massive data based on sophisticated models, it is essential to develop a computationally efficient method.

Manuscript received August 20, 1999.
Manuscript revised **, **.

[†] The authors are with the Institute of Statistical Mathematics

In the case of time series model, due to a mathematical structure, i.e., Markov property, a large class of models can be expressed by the **generalized state space model** [13], [19], [20]. The discrete valued process such as the {A,T,G,C} in DNA sequence can also be expressed with this model as well [7]. The discrete version of the generalized state space model is sometimes called **hidden Markov model** and is used for the analysis of DNA sequence and voice recognition as the model with high ability of description [4], [18].

This generalized state space model has another significant merit that it automatically realizes the estimation of its unknowns with $O(N)$ computations by the recursive filter and smoothing algorithms. Therefore, the generalized state space model is a useful platform for flexible modeling and at the same time is a base for generating efficient computation.

In this article, it will be shown that by considering higher level modeling, it is possible to develop models quite freely. By a numerical example, it will be shown that, as a result, it becomes possible to extract essential information from data which has been difficult due to the use of restricted models. It was also shown that by rising the level of the model, the model selection procedure which has been realized with human expertise can be performed automatically and thus the automatic processing of huge time series data becomes realistic. In other words, the high level statistical modeling facilitates both "automatic processing of massive time series data" and "a new method for knowledge discovery."

2. Hierarchical Statistical Models

2.1 Hierarchical Structure of Statistical Models

Figure 1 shows the hierarchical structure of time series modeling. On the top there is a **self-organizing state space model** [22] which is the main subject of this article. In the middle level there exists the **generalized state space model** [19], [24]. This generalized state space model contains the famous **state space model** [3] as a special case. Finally, in the lower level there are **parametric models**. Actually, this hierarchical structure of statistical models have not been well recognized so far. Rather, by mitigating the difficulties arisen in actual modeling process, this was gradually established by the "generalization of the models".

In this article, by using a simple example, we will exemplify how the possibility of knowledge discovery is achieved by the adoption of higher level models. Consider the estimation of the mean structure, namely the mean value function t_n, of the time series shown in Fig. 2(a). y_n is generated by the following model:

file=IEICEFig1.eps,width=6cm,height=5cm

Fig. 1 Hierarchical structure of statistical models

$$y_n = \begin{cases} -0.8 & + & v_{1,n} & (n = 1, \dots, 50) \\ -1 & + & v_{1,n} & (n = 51, \dots, 100) \\ -1 & + & v_{2,n} & (n = 101, \dots, 150) \\ 1 & + & v_{2,n} & (n = 151, \dots, 200) \end{cases} \quad (1)$$

where $v_{1,n}$ and $v_{2,n}$ follow $v_{1,n} \sim N(0, 0.01)$ and $N(0,1)$, respectively. Here $N(0,\sigma^2)$ denotes the Gaussian distribution with mean 0 and the variance σ^2. In this example, the mean value function t_n has jumps at $n = 51$ and $n = 151$. Jump sizes are designed to be two times of σ, namely 2σ.

2.2 Parametric Models

In general, the observed data are inevitably subject to the observation noise. Unfortunately, the relationship between the observation noise and the mean value structure is unknown for us. To explore the mean value structure, we therefore give a mathematical functional form of this relationship as the observation model. For the case shown in Fig. 2(a), we assume that the observation noise w_n is added to t_n, namely

$$y_n = t_n + w_n, \quad w_n \sim N(0, \sigma^2). \quad (2)$$

It should be reminded that while the variance of the observation noise for actual data increases by one hundred times at $n = 101$, the variance for the observation model, σ^2, is assumed to be constant over time.

We fit a parametric model to y_n on a basis of the observation model (2), where the parametric model is defined such that t_n is described as an analytical function of n with a small number of parameters such as a polynomial function and combination of the Fourier series. A visual inspection on Fig. 2(a) gives us a conjecture that there exists a jump around $n = 150$. It motivated us to choose a sigmoid function given by

$$t_n = \frac{\gamma_1}{1 + \exp\left(-\gamma_2(n - \gamma_3)\right)} + \gamma_4, \quad (3)$$

which is frequently used as a nonlinear function for the artificial neural network. The parameter vector of this statistical model, composed of (2) and (3), is specified by $\boldsymbol{\theta} = [\gamma_1, \gamma_2, \gamma_3, \gamma_4, \sigma^2]'$. In Fig. 2(b), thick curve shows the result obtained by applying the least squares fit of the sigmoid function.

Under the assumption of (2), the least squares fitting is equivalent to the maximum likelihood estimation of $\boldsymbol{\theta}$. In this case, an apparent dependency of the variance of the observation noise on time suggests that the simple least squares fit, i.e., observation model (2) is inappropriate. A more sophisticated analysis with a parametric model is to estimate a time-varying variance of the observation noise, $\widehat{\sigma}_n^2$, by a practical method [24] and then perform a weighted least squares fit based on the estimated $\widehat{\sigma}_n^2$. This approach would extract much information than a simple least squares fit, but this kind of transaction composed of several steps in data analysis is likely to lose an opportunity to find small signals. The estimation on both $\boldsymbol{\sigma}_n$ and $\boldsymbol{t}_n$ should be made simultaneously to avoid such unsatisfactory information processing. This attempt can be realized with the self-organizing state space model explained in 2.5.

In a framework of estimating the mean value structure with a parametric model, it is important to choose an appropriate model based on the well-known theories behind the observations. This process cannot be automated and thus requires the interventions of researchers expertized with an analysis of the data. When there is no natural reason to choose any particular parametric model, a model selection with AIC [1], [28] should be performed for choosing the best parametric model among the competing models.

2.3 State Space Models

2.3.1 Linear Gaussian Trend Model

Consider a statistical model based on the first order stochastic difference equation

$$\boldsymbol{t}_n = \boldsymbol{t}_{n-1} + \boldsymbol{v}_n, \quad \boldsymbol{v}_n \sim N(0, \boldsymbol{\tau}^2). \tag{4}$$

This type of models are frequently used in time series analysis for the estimation of trend and are called the trend models or random walk models [23], [24].

2.3.2 State Space Model

Using the state vector $\boldsymbol{x}_n = [\boldsymbol{t}_n]$, the models (2) and (4) can be represented by a state space model [3]

$$\boldsymbol{x}_n = \boldsymbol{F}\boldsymbol{x}_{n-1} + \boldsymbol{G}\boldsymbol{v}_n \quad \text{(system model)} \tag{5}$$

$$\boldsymbol{y}_n = \boldsymbol{H}\boldsymbol{x}_n + \boldsymbol{w}_n \quad \text{(observation model)} \tag{6}$$

where $\boldsymbol{v}_n \sim N(0, \boldsymbol{Q})$ and $\boldsymbol{w}_n \sim N(0, \boldsymbol{R})$ are Gaussian white noises and are called the system noise and the observation noise, respectively. $\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{H}$ are matrices with appropriate dimensions.

For the above simple example, $\boldsymbol{F} = \boldsymbol{G} = \boldsymbol{H} = 1$ and $\boldsymbol{Q} = \boldsymbol{\tau}^2$ and $\boldsymbol{R} = \boldsymbol{\sigma}^2$. Most of the models used in time series analysis can be uniformly expressed in and treated by the state space model [12], [13], [23], [24].

2.3.3 Kalman Filter and Smoothing Algorithms

Once a time series model is expressed in state space model form, the estimate of $\boldsymbol{t}_n$, $\widehat{\boldsymbol{t}}_n$, is obtained as a component of the estimate of the state vector $\boldsymbol{x}_n$. Given $\boldsymbol{\sigma}^2$ and $\boldsymbol{\tau}^2$, the estimate of the state is obtained by the Kalman filter and the smoothing algorithms (KFS for short, [3]). For small $\boldsymbol{\tau}^2$ (precisely, for small $\boldsymbol{\tau}^2/\boldsymbol{\sigma}^2$), $\widehat{\boldsymbol{t}}_n$ becomes very close to a constant. On the other hand, for large $\boldsymbol{\tau}^2$, $\boldsymbol{t}_n$ resembles to the data.

In this trend model, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\tau}^2$ are usually called the parameters of the model. However, emphasizing the hierarchical structure of the model, in Bayesian modeling framework, we will call the **hyper-parameter** [26]. Note that in this modeling, the hyper-parameter is assumed to be a constant. Estimation of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\tau}^2$ is performed by the maximum likelihood method [23] and each estimate is denoted by $\widehat{\boldsymbol{\sigma}}^2$ and $\widehat{\boldsymbol{\tau}}^2$, respectively.

In Fig. 2(c), the estimate $\widehat{\boldsymbol{t}}_n$ by the trend model ($\widehat{\boldsymbol{\tau}}^2 = 0.021$) is shown by thick curve. In contrast to the sigmoid function, the trend model can freely express various shape. Therefore, by using the trend model, it is possible to discover gradually changing mean structure from data. However, even this trend model has a limitation in extracting the characteristics of the data. Actually in the current example, the location and the shape of the jump is not so clearly identified. Further, in responding to the presence of jumps, a large value of $\widehat{\boldsymbol{\tau}}^2$ is selected and as a result, the estimate of the trend becomes inappropriately very wiggly.

2.4 Generalized State Space Models

2.4.1 Linear Non-Gaussian Trend Model

As an extension of the linear-Gaussian trend model, we consider the case when the noise $\boldsymbol{v}_n$ in (4) is distributed as a heavy-tailed non-Gaussian distribution, such as the Cauchy distribution. The $\boldsymbol{v}_n$ distributed as the Cauchy distribution is concentrated around 0. However, it may also take a very large value, such as the $\pm 5\boldsymbol{\sigma}$ values of the Gaussian distribution, with a relatively high probability. With this noise distribution, the realization of the trend, $\boldsymbol{t}_n$, generally becomes very close to a constant. However, at the same time, it may occasionally have very large jumps. This model is geared to automatically detect the jumps in the mean value structure.

On the other hand, by using a non-Gaussian distribution for observation noise, automatic detection of the outliers becomes possible [2], [25]. As a non-Gaussian observation noise distribution, symmetric heavy-tailed distribution, such as the type VII Pearson family of

Fig. 2 Estimation of the mean structure. (a) test data, (b) estimated sigmoid function, (c) linear Gaussian trend model, (d) linear non-Gaussian (Cauchy) trend Model, (e) estimate by the self-organizing state space model, (f) estimate $\log_{10}\sigma_n^2$.

distributions

$$p(v;0,\tau^2,b)=\frac{\Gamma(b)\tau^{2b-1}}{\Gamma(1/2)\Gamma(b-1/2)}\frac{1}{(v^2+\tau^2)^b}, \qquad (7)$$

can be used. Here, b is called the shape parameter ($b > 1/2$). For $b = +\infty$ and 1, this distribution becomes the Gaussian and the Cauchy distributions, respectively. τ^2 is called the dispersion parameter and characterizes the spread of the distribution. Although this family of distribution has one more parameters than the Gaussian model, the complexity of the model is not significantly increased in the sense that they can be estimated by the maximum likelihood method. Compared with the Gaussian model, this family of distributions has a flexibility in the shape of the noise distribution which enables automatic detection of the location of the jumps in the mean value structure. On the other hand, these models share the common characteristics that the hyper-parameters, $\boldsymbol{\lambda} = [\sigma^2, b, \tau^2]'$ in the current case, does not depend on time n. This point is the main difference from the one in the **self-organizing state space model** introduced in the next section.

2.4.2 Generalized State Space Model

The above trend model can be generally expressed in nonlinear non-Gaussian state space model form [11], [19], [20], [31]

$$\begin{aligned} \boldsymbol{x}_n &= \boldsymbol{f}(\boldsymbol{x}_{n-1}, \boldsymbol{v}_n) && (8)\\ \boldsymbol{y}_n &= \boldsymbol{h}(\boldsymbol{x}_n, \boldsymbol{w}_n), && (9) \end{aligned}$$

where the system noise $\boldsymbol{v}_n$ and the observation noise $\boldsymbol{w}_n$ follow the density functions $q(\boldsymbol{v})$ and $r(\boldsymbol{w})$, respectively. The initial state $\boldsymbol{x}_0$ is distributed according to the density $p_0(\boldsymbol{x})$.

Although in the above example, $\boldsymbol{f}(\boldsymbol{x}_{n-1}, \boldsymbol{v}_n) = \boldsymbol{x}_{n-1} + \boldsymbol{v}_n$,$\boldsymbol{h}(\boldsymbol{x}_n, \boldsymbol{w}_n) = \boldsymbol{x}_n + \boldsymbol{w}_n$ are linear, in general, $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{v})$ and $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{w})$ are nonlinear functions. $\boldsymbol{q}(\boldsymbol{v})$ and $\boldsymbol{r}(\boldsymbol{w})$ are, in general, non-Gaussian densities specified by hyper-parameters. As in 2.4.1, the vector consisting of these unknown hyper-parameters is denoted by $\boldsymbol{\lambda}$.

(8) and (9) can be considered as a special case of the following **generalized state space model**:

$$\begin{aligned} \boldsymbol{x}_n &\sim \boldsymbol{Q}(\ \cdot\ |\boldsymbol{x}_{n-1}) && (10)\\ \boldsymbol{y}_n &\sim \boldsymbol{R}(\ \cdot\ |\boldsymbol{x}_n). && (11) \end{aligned}$$

Here, $\boldsymbol{Q}$ and $\boldsymbol{R}$ are the conditional distributions given the states $\boldsymbol{x}_{n-1}$ and $\boldsymbol{x}_n$, respectively. As an example, Dynamic Generalized Linear Model (DGLM) [32],[33] is frequently used for the analysis of discrete valued time series [16],[19],[20],[24],[32],[33]:

$$\begin{aligned} \boldsymbol{x}_n &= \boldsymbol{F}\boldsymbol{x}_{n-1} + \boldsymbol{G}\boldsymbol{v}_n, \quad \boldsymbol{v}_n \sim N(0, \boldsymbol{Q}) && (12)\\ \boldsymbol{\alpha}_n &= \boldsymbol{H}\boldsymbol{x}_n && \\ \boldsymbol{y}_n &\sim \exp(\boldsymbol{\alpha}_n' \boldsymbol{y}_n - \boldsymbol{b}(\boldsymbol{\alpha}_n) + \boldsymbol{c}(\boldsymbol{y}_n)). && (13) \end{aligned}$$

where $\boldsymbol{F}, \boldsymbol{G}$, and $\boldsymbol{H}$ are properly defined matrices, $\boldsymbol{Q}$ is the variance covariance matrix, and $\boldsymbol{b}(\cdot)$ and $\boldsymbol{c}(\cdot)$ are properly defined functions. This type of distributions, called exponential family of distributions, can cover broad class of distributions frequently used in statistical analysis, such as the Poisson distribution and the binomial distribution.

2.4.3 Automatic Detection of Jump Points

With a non-Gaussian observation noise distribution or system noise distribution, the KFS cannot yield good estimate of $\widehat{\boldsymbol{t}_n}$, and for an efficient estimation of $\boldsymbol{t}_n$, it is necessary to use the non-Gaussian filter [19]. For simplicity, we here assume that $\boldsymbol{b} = 1$, namely, that the noise distribution is Cauchy. Thick curve in Fig. 2(d) shows the result of the non-Gaussian filter for $\widehat{\boldsymbol{t}_n}$ with the maximum likelihood estimate of the hyper-parameter $\boldsymbol{\tau}^2 = 1.6 \times 10^{-5}$. Compared with the Gaussian case, the jump of the trend is clearly detected. It should be emphasized here again that the location of the jump is not pre-specified and the jump was automatically identified by "the maximum likelihood method and the non-Gaussian filter".

So far we have shown that the non-Gaussian model can automatically detect the jump at $\boldsymbol{n} = 151$. However, unfortunately, it failed to detect small jump at $\boldsymbol{n} = 51$. This is due to the assumption that the variance of the observation noise $\boldsymbol{\sigma}^2$ is a constant over time, because the maximum likelihood estimates of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\tau}^2$ yield the large confidence interval of $\boldsymbol{p}(\boldsymbol{t}_n|Y_N)$ between $\boldsymbol{n} = 1 \sim 100$ compared with the size of the small jump. This example clearly shows that the identification of the mean value structure may be deteriorated by restricting the flexibility of the model, namely by assuming that the hyper-parameter $\boldsymbol{\sigma}^2$ is constant over time.

2.5 Self-Organizing State Space Models

2.5.1 Time-Changes of Hyper-Parameters

To mitigate the problem discussed in 2.4.3, we consider a model whose hyper-parameter of the noise distribution in the observation model (2) changes with time:

$$\begin{aligned} \boldsymbol{y}_n &= \boldsymbol{t}_n + \boldsymbol{w}_n, \quad \boldsymbol{w}_n \sim N(0, \boldsymbol{\sigma}_n^2) && (14)\\ \log \boldsymbol{\sigma}_n^2 &= \log \boldsymbol{\sigma}_{n-1}^2 + \varepsilon_n, \quad \varepsilon_n \sim \boldsymbol{C}(0, \boldsymbol{d}^2), && (15) \end{aligned}$$

where $\boldsymbol{C}(0, \boldsymbol{d}^2)$ is the Cauchy distribution with the center at 0. With this extension, it is expected that the observation noise with suddenly changing variance can properly modeled. Here the logarithm of $\boldsymbol{\sigma}_n^2$ is modeled to assure the positivity of $\boldsymbol{\sigma}^2$. Of course, we can use the Pearson family of distributions for ε_n. As the system model for $\boldsymbol{t}_n$, we assume the linear non-Gaussian (Cauchy distribution $\boldsymbol{C}(0, \boldsymbol{\tau}^2)$) model as 2.4.1.

These models can be expressed in state space model form:

$$\begin{bmatrix} \boldsymbol{t}_n \\ \log_{10} \boldsymbol{\sigma}_n^2 \\ \log_{10} \boldsymbol{\tau}^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{t}_{n-1} \\ \log_{10} \boldsymbol{\sigma}_{n-1}^2 \\ \log_{10} \boldsymbol{\tau}^2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{v}_n \\ \varepsilon_n \\ 0 \end{bmatrix}, \quad (16)$$

$$\boldsymbol{y}_n = [\,1, 0, 0\,] \begin{bmatrix} \boldsymbol{t}_n \\ \log_{10} \boldsymbol{\sigma}_n^2 \\ \log_{10} \boldsymbol{\tau}^2 \end{bmatrix} + \boldsymbol{w}_n, \quad \boldsymbol{w}_n \sim N(0, \boldsymbol{\sigma}_n^2). (17)$$

It should be noted that in the hyper-parameter vector $\boldsymbol{\lambda}_n = [\log_{10} \boldsymbol{\sigma}_n^2, \log_{10} \boldsymbol{\tau}^2]'$, only $\log_{10} \boldsymbol{\sigma}_n^2$ depends on time $\boldsymbol{n}$. For the initial state $\boldsymbol{z}_0 = [\boldsymbol{t}_0, \log_{10} \theta_0, \log_{10} \boldsymbol{\tau}^2]'$, it is assumed, for example, that $\boldsymbol{t}_0 \sim N(0, 4)$, $\log_{10} \boldsymbol{\sigma}_0^2 \sim U([-4, 4])$, $\log_{10} \boldsymbol{\tau}^2 \sim U([-4, 2])$, where $U([\ ,\ ])$ denotes the uniform distribution. The "parameter" of this model is the variance (or dispersion) of the ε_n, $\boldsymbol{\xi} = [\boldsymbol{d}^2]$. In this article, we call $\boldsymbol{\xi}$ the **hyper-hyper-parameter**.

2.5.2 Self-Organizing State Space Models

The general form of the self-organizing state space model [22] is obtained by augmenting the state vector $\boldsymbol{x}_n$ with the hyper-parameter vector $\boldsymbol{\lambda}$ as $\boldsymbol{z}_n = [\boldsymbol{x}_n', \boldsymbol{\lambda}']'$. The state space model for this augmented state vector $\boldsymbol{z}_n$ is given by

$$\boldsymbol{z}_n = \boldsymbol{F}(\boldsymbol{z}_{n-1}, \boldsymbol{v}_n), \quad \boldsymbol{y}_n = \boldsymbol{H}(\boldsymbol{z}_n, \boldsymbol{w}_n), \quad (18)$$

where

$$\boldsymbol{F}(\boldsymbol{z}, \boldsymbol{v}) = \begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{v}) \\ \boldsymbol{\lambda} \end{bmatrix}, \quad \boldsymbol{H}(\boldsymbol{z}, \boldsymbol{w}) = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{w}). \quad (19)$$

For the model with time-varying hyper-parameter, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_n$, a model for time-changes of the hyper-parameter $\boldsymbol{\lambda}_n$ is necessary. For example, we may use the random walk model $\boldsymbol{\lambda}_n = \boldsymbol{\lambda}_{n-1} + \varepsilon_n$. For the estimation of the time-varying hyper-parameter, we define the augmented state vector $\boldsymbol{z}_n$ by $\boldsymbol{z}_n = [\boldsymbol{x}_n', \boldsymbol{\lambda}_n']'$, where ε_n is a white noise with density function $\boldsymbol{\phi}(\varepsilon|\boldsymbol{\xi})$. The nonlinear function $\boldsymbol{F}$ is defined by

$$F(z_{n-1}, u_n) = [f'(x_{n-1}, v_n), (\lambda_{n-1} + \varepsilon_n)']', \quad (20)$$

where the system noise is defined by $u_n = [v_n', \varepsilon_n']'$. Therefore, there is no formal difference between the generalized state space model and the self-organizing state space model. The essential difference is that the self-organizing state space model contains the hyperparameter in the state vector.

Fig. 2(e) and (f) show the estimates $\widehat{t}_n$ and $\log_{10} \widehat{\sigma}_n^2$ obtained by the self-organizing state space model, respectively. Increase of the variance of the observation noise by about 100 times (2 in log-scale) at $n = 101$ is detected automatically. Due to this self-adjustment, not only the junp at $n = 101$, but also the small one at $n = 51$ was clearly detected.

3. Recursive Techniques for State Estimation

3.1 Conventional Methods

Self-organizing state space is formally a special case of the generalized state space model, it suffices to develop practical recursive methods for filtering and smoothing for the generalized state space model. In engineering, the extended Kalman filter has been used for simultaneous estimation of the state and the parameters [3], [27], [30]. However, the extended Kalman filter essentially approximates a non-Gaussian distribution by one Gaussian distribution, it has been considered that it did not work well in practice [3], [20]. In this article, we will show two numerical methods for recursive estimation that can yield more precise state estimate than the extended Kalman filter. Since they can yield precise conditional state distributions, it can be applied to self-organizing state space models.

3.2 Conditional State Distributions

The generalized state space model specifies the following two conditional distributions: $Q(x_n|x_{n-1}, \lambda)$, the distribution of x_n given the previous state x_{n-1}, and $R(y_n|x_n, \lambda)$, the distribution of y_n given the state x_n. For the self-organizing state space model, they correspond to $Q(z_n|z_{n-1}, \xi)$ and $R(y_n|z_n, \xi)$. The set of the observations up to time point j is denoted by $Y_j = \{y_1, \cdots, y_j\}$. Then $p(x_n|Y_{n-1}, \lambda)$ defines the one-step-ahead predictive density. For the recursive estimation of the state vector, we consider the following three conditional distributions:

$p(x_n\|Y_{n-1})$	predictive density
$p(x_n\|Y_n)$	filter density
$p(x_n\|Y_N)$	smoother density.

Given the time series $Y_N = \{y_1, \ldots, y_N\}$, the likelihood of the model specified by the hyper-parameter λ can be decomposed as

$$L(\lambda) = p(y_1, \cdots, y_N|\lambda) = \prod_{n=1}^{N} p(y_n|Y_{n-1}, \lambda), \quad (21)$$

where $p(y_n|Y_{n-1}, \lambda)$ is obtained by

$$p(y_n|Y_{n-1}, \lambda) = \int p(y_n|x_n, \lambda) p(x_n|Y_{n-1}, \lambda)\, dx_n. \quad (22)$$

Therefore, the maximum likelihood estimate of the parameter λ is obtained by maximizing the log-likelihood.

3.3 Numerical Methods

In the following subsections, we shall show two methods of computing conditional distributions.

3.3.1 Implementation by Non-Gaussian Smoothing

A non-Gaussian filter for the nonlinear and non-Gaussian state space model (10) and (11) is shown in [19], [20]:

$$p(x_n|Y_{n-1}) = \int p(x_n|x_{n-1}) p(x_{n-1}|Y_{n-1})\, dx_{n-1} \quad (23)$$

$$p(x_n|Y_n) = \frac{p(y_n|x_n) p(x_n|Y_{n-1})}{p(y_n|Y_{n-1})}, \quad (24)$$

where $p(y_n|Y_{n-1})$ appears in (22). Note that, for simplicity, the appearance of the parameter λ is suppressed.

The final estimate of the state x_n is obtained by the smoothing algorithm [19], [20]

$$\begin{aligned} p(x_n|Y_N) &= p(x_n|Y_n) \\ &\quad \times \int \frac{p(x_{n+1}|x_n) p(x_{n+1}|Y_N)}{p(x_{n+1}|Y_n)}\, dx_{n+1}. \end{aligned} \quad (25)$$

Since this non-Gaussian filter and smoother are realized by computationally intensive numerical integration, it can be applied to only low order state models.

3.3.2 Implementation by Monte Carlo Smoothing

Kitagawa [21] developed a Monte Carlo filter and smoother. The "bootstrap filter" [5], [11] is a similar algorithm. In this method, each density function is approximated by many (say $m = 10^4 \sim 10^5$) particles, that can be considered as independent realizations from that distribution. It can be shown that these particles can be generated recursively by the following algorithm:

1. For $j = 1, \ldots, m$, generate k-dimensional random number $f_0^{(j)} \sim p_0(x)$.
2. Repeat the following steps for $n = 1, \ldots, N$. For (a), (b) and (c), repeat m times independently for $j = 1, \ldots, m$.
   a. Generate ℓ-dimensional random number $v_n^{(j)} \sim q(v)$ for system noise.

b. Compute $\boldsymbol{p}_n^{(j)} = F(\boldsymbol{f}_{n-1}^{(j)}, \boldsymbol{v}_n^{(j)})$.

c. Compute $\boldsymbol{\alpha}_n^{(j)} = p(\boldsymbol{y}_n|\boldsymbol{x}_n = \boldsymbol{p}_n^{(j)})$.

d. Obtain $\boldsymbol{f}_n^{(1)}, \cdots, \boldsymbol{f}_n^{(m)}$ by the sampling with replacement from $\boldsymbol{p}_n^{(1)}, \cdots, \boldsymbol{p}_n^{(m)}$ with sampling probabilities proportional to $\boldsymbol{\alpha}_n^{(1)}, \cdots, \boldsymbol{\alpha}_n^{(m)}$.

A significant merit of this Monte Carlo filter is that it can be applied to almost any type of high dimensional nonlinear and non-Gaussian state space models. This filtering algorithm can be extended to the smoothing by storing the past particles and resample the vector of particles $(\boldsymbol{p}_n^{(j)}, \boldsymbol{p}_{n-1}^{(j)}, \cdots, \boldsymbol{p}_{n-\ell}^{(j)})$ rather than the single particle $\boldsymbol{p}_n^{(j)}$.

Incidentally, in this Monte Carlo filter the likelihood is computed by

$$p(\boldsymbol{y}_n|Y_{n-1}) \cong \frac{1}{\boldsymbol{m}} \sum_{j=1}^{m} \boldsymbol{\alpha}_n^{(j)}. \tag{26}$$

Therefore in the estimation of the parameter of the model, it is difficult to obtain arbitrarily close approximations to the maximum likelihood estimator unless a very large number of particles are used or an average of the approximated log-likelihoods is computed by the parallel use of many Monte Carlo filters. On the other hand, in the self-organizing state space models, the number of hyper-hyper-parameters which need to be estimated by the maximum likelihood methods is usually one or two, and the hyper-parameters are automatically determined as the estimate of the state vector. The difficulty in the generalized state space modeling was thus solved by the use of high level modeling and intensive use of computers.

3.3.3 Other Methods

The usefulness of the numerical methods for generalized sate space model presented above is confirmed by the applications to various actual modeling. However, for some special class of models, several computationally efficient and precise methods for state estimation were proposed. Typical examples of such a class is the DGLM explained in subsection 2.4.2. This type of models has a favorable property that the filter distribution is unimodal and close to symmetric and by utilizing these properties and by properly treating the difference from the normal distribution, computationally efficient and precise estimators have been proposed [6], [8], [9], [29], [33].

3.4 Genetic Algorithm Filter

Before closing this section, we would like to introduce a model which allows the hyper-parameter to be time-dependent like the self-organizing state space model. A similar structure to the algorithm of the Monte Carlo filter (MCF) appears in the Genetic algorithm (GA) that is a population-based search procedure developed in analogy to genetic laws and natural selection [17]. In general GA is characterized by keeping the $\boldsymbol{m}$ candidates for optimal solution at each iteration composed of three steps: crossover, mutation, and reproduction (or selection) [10]. It has been pointed out that the filtering procedure composed of (c) and (d) is identical to the reproduction procedure by regarding $p(\boldsymbol{y}_n|\boldsymbol{p}_n^{(j)})/\boldsymbol{m}$ as the evaluation function in GA and that the prediction plays a similar role to mutation and crossover operators in giving wide variety among population [14], [15]. Using an analogy between MCF and GA, an interpretation of MCF from the viewpoint of GA has been presented and several practical issues concerning its implementation has been investigated [15].

This strong parallelism leads a new procedure, called the GA filter, which is based on the replacement of (8) by the genetic operator such as the crossover and mutation. The distribution function of stochastic fluctuations given to the population in the system model (8) is independent of time because of the fixed values of the hyper-parameter vector. Meanwhile, the statistical characteristics of fluctuations produced by the crossover can evolves on time, because the crossover is an interaction between each particle at each time, resulting in the fact that the stochastic fluctuations among particles are determined by population itself at each time.

4. Conclusion

In this article, it was shown that by considering a hierarchical structure of the statistical models, very flexible modeling becomes possible, resulting in the development of innovative information extraction procedure. It was also shown that by rising the level of the model, the model selection procedure which has been realized with human expertise can be performed automatically and thus the automatic processing of huge time series data becomes realistic. In other words, the higher-order statistical modeling facilitates both "automatic processing of massive time series data" and "a new method for knowledge discovery". In this respect, the self-organizing state space model is a promising tool for data mining. Obviously, by the higher order modeling, the required amount of computation increases considerably. However, by considering the rapid progress of the computing ability and the laborious human intervention becoming unnecessary by the development of automatic procedure, it clearly reveals the direction of the future development of time series modeling.

References

[1] H. Akaike, "A new look at the statistical model identification," IEEE Transactions on Automatic Control, AC-19,

716–723, 1974.

[2] H. Akaike, and G. Kitagawa ed., "The Practice of Time Series Analysis," Springer-Verlag, New York, 1999.

[3] B.D.O. Anderson, and J.B. Moore, "Optimal Filtering," Prentice-Hall, New Jersey, 1979.

[4] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Stat., 41, 164–171, 1970.

[5] A. Doucet, E. Barat, and P. Duvaut, "A Monte Carlo approach to recursive Bayesian state estimation," Proc. IEEE Signal Processing/Athos Workshop on Higher Order Statistics, Girona, Spain, 1995.

[6] J. Durbin, and S.J. Koopman, "Monte Carlo maximum likelihood estimation for non–Gaussian state space models," Biometrika, vol.84, 669–684, 1997.

[7] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, "Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acides," Cambridge University Press, Cambridge and New York, 1998.

[8] L. Fahrmeir, "Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear model," Journal of the American Statistical Association, vol.87, no.418, 501–509, 1992.

[9] S. Frühwirth-Schnatter, "Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering," Statistics and Computing, vol.4, 259–269, 1994a.

[10] D.E. Goldberg, "Genetic Algorithm in Search, Optimization and Machine Learning," Addison-Wesley Publishing Company, Massachusetts, 1989.

[11] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," IEE Proceedings-F, 140, no.2, 107–113, 1993.

[12] P.J. Harrison, and C.F. Stevens, "Bayesian Forecasting (with discussion)," Journal of the Royal Statistical Society, Ser. B, vol.34, 1–41, 1976.

[13] A.C. Harvey, "Forcasting, structural Time Series Models and the Kalman Filter," Cambridge University Press, 1989.

[14] T. Higuchi, "Genetic algorithm and Monte Carlo filter (in Japanese with English Abstract)," Proceedings of the Institute of Statistical Mathematics, vol.44, no.1, 19–30, 1996.

[15] T. Higuchi, "Monte Carlo filter using the Genetic algorithm operators," Journal of Statistical Computation and Simulation, vol.59, no.1, 1–23, 1997.

[16] T. Higuchi, "Applications of quasi-periodic oscillation models to seasonal small count time series," Computational Statistics & Data Analysis, vol.30, 281–301, 1999.

[17] J.H. Holland, "Adaption in natural and artificial systems," The University of Michigan Press, Ann Arbor, 1975.

[18] X.D. Huang, Y. Ariki, and M.A. Jack, "Hidden Markov models for speech recognition," Edinburgh University Press, Edinburgh, 1990.

[19] G. Kitagawa, "Non-Gaussian state-space modeling of nonstationary time series (with discussion)," Journal of the American Statistical Association, vol.82, 1032–1063, 1987.

[20] G. Kitagawa, "A nonlinear smoothing method for time series analysis," Statistica Sinica, vol.1, no.2, 371–388, 1991.

[21] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space model," Journal of Computational and Graphical Statistics, vol.5, no.1, 1–25, 1996.

[22] G. Kitagawa, "Self-organizing state space model," Journal of the American Statistical Association, vol.93, no.443, 1203–1215, 1998.

[23] G. Kitagawa, and W. Gersch, "A smoothness priors-state space modeling of time series with trend and seasonality," Journal of the American Statistical Association, vol.79, no.386, 378–389, 1984.

[24] G. Kitagawa, and W. Gersch, "Smoothness Priors Analysis of Time Series," Springer-Verlag, New York, 1996.

[25] G. Kitagawa, and N. Matsumoto, "Detection of coseismic changes of underground water level," Journal of the American Statistical Association, vol.91, no.434, 521–528, 1996.

[26] D.V. Lindley, and A.F.M. Smith, "Bayes estimates of the linear model (with discussion)," Journal of the Royal Statistical Society Ser. B, vol.34, 1–41, 1972.

[27] L. Ljung, "Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems," IEEE Transactions on Automatic Control, AC-24, no.1, 36–50, 1979.

[28] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, "Akaike Information Criterion Statistics," D. Reidel Publishing Company, 1986.

[29] S. Schnatter, "Integration-based Kalman-filtering for a dynamic generalized linear trend model," Computational Statistics & Data Analysis, vol.13, 447–459, 1992.

[30] V. Solo, "Adaptive spectral factorization," IEEE Transactions on Automatic Control, AC-34, no.10, 1047–1051, 1989.

[31] H. Tanizaki, "Nonlinear Filters," Springer-Verlag, New York, 1993.

[32] M. West, and P.J. Harrison, "Bayesian Forecasting and Dynamic Models," 2^{nd} ed., Springer-Verlag, New York, 1997.

[33] M. West, P.J. Harrison, and H.S. Migon, "Dynamic generalized linear models and Bayesian forecasting (with discussion)," Journal of the American Statistical Association, 80, 73–97, 1985.

Tomoyuki HIGUCHI Associate Professor at the Institute of Statistical Mathematics and Associate Professor of Statistical Science at the Graduate University for Advanced Study. His primary research interests are in statistical modeling of space-time data, stochastic optimization techniques, and data mining.

Genshiro KITAGAWA Professor at the Institute of Statistical Mathematics. He is currently Deputy Director of the Institute of Statistical Mathematics and Professor of Statistical Science at the Graduate University for Advanced Study. His primary research interests are in time series analysis, non-Gaussian nonlinear filtering, and statistical modeling. He was awarded the 2nd Japan Statistical Society Prize in 1997.