Self-organizing Time Series Model

Tomoyuki Higuchi
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

1 Introduction

1.1 Generalized state space model

The generalized state space model (GSSM) that we deal with in this study is defined by a set of two equations

$$\begin{aligned} \textit{system model} \qquad \mathbf{x}_t &= f(\mathbf{x}_{t-1}, \mathbf{v}_t), \quad \text{and} && (1.1)\\ \textit{observation model} \qquad \mathbf{y}_t &\sim r(\ \cdot\ |\mathbf{x}_t, \boldsymbol{\theta}_{obs}), && (1.2) \end{aligned}$$

where $\mathbf{x}_t$ is an $n_x \times 1$ vector of unobserved sate variables, and $\mathbf{y}_t$ is an n_y dimensional vector observation. $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_v} \to \mathbb{R}^{n_x}$ is a given function. $\{\mathbf{v}_t\}$ is an independent and identically distributed (i.i.d.) random process with $\mathbf{v}_t \sim q(\mathbf{v}|\boldsymbol{\theta}_{sys})$. r is the conditional distribution of $\mathbf{y}_t$ given $\mathbf{x}_t$. $q(\cdot|\cdot)$ and $r(\cdot|\cdot)$ are, in general, non-Gaussian densities specified by the unknown parameter vectors, $\boldsymbol{\theta}_{sys}$ and $\boldsymbol{\theta}_{obs}$, respectively. In this study, we set $\boldsymbol{\theta} = [\boldsymbol{\theta}'_{sys}, \boldsymbol{\theta}'_{obs}]'$. The initial state $\mathbf{x}_0$ is distributed according to the density $p_0(\mathbf{x})$.

The GSSM includes the nonlinear non-Gaussian state space model (Kitagawa 1987, Kitagawa 1991, Tanizaki 1993, Gordon, Salmond and Smith 1993) as a special case

$$\begin{aligned} \textit{system model} \qquad \mathbf{x}_t &= f(\mathbf{x}_{t-1}, \mathbf{v}_t) && (1.3)\\ \textit{observation model} \qquad \mathbf{y}_t &= h(\mathbf{x}_t, \mathbf{e}_t), && (1.4) \end{aligned}$$

where $\{\mathbf{e}_t\}$ is an i.i.d. random process that follows $\mathbf{e}_t \sim r(\mathbf{e}|\boldsymbol{\theta}_{obs})$. $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_e} \to \mathbb{R}^{n_y}$ is a given function. The simplest version described by the nonlinear non-Gaussian state space model is reduced to the well-known state space model (Anderson and Moore 1979) given by:

$$\begin{aligned} \textit{system model} \qquad \mathbf{x}_t &= F\mathbf{x}_{t-1} + G\mathbf{v}_t, \quad \mathbf{v}_t \sim N(0, Q) && (1.5)\\ \textit{observation model} \qquad \mathbf{y}_t &= H\mathbf{x}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(0, R) && (1.6) \end{aligned}$$

where F, G, and H are $n_x \times n_x$, $n_x \times n_v$, and $n_y \times n_x$ matrices, respectively. Q and R are the variance covariance matrix of $\mathbf{v}$ and $\mathbf{e}$, respectively.

The GSSM also includes a model which is frequently used for the analysis of discrete valued time series (West, Harrison and Migon 1985, Kitagawa

1987, Kitagawa 1991, Kitagawa and Gersch 1996, West and Harrison 1997, Higuchi 1999):

$$\textit{system model} \qquad \mathbf{x}_t = F\mathbf{x}_{t-1} + G\mathbf{v}_t, \quad \mathbf{v}_t \sim N(0, Q) \tag{1.7}$$

$$\textit{observation model} \qquad \mathbf{y}_t \sim \exp(\alpha_t'\mathbf{y}_t - b(\alpha_t) + c(\mathbf{y}_t)), \quad \alpha_t = H\mathbf{x}_t, \tag{1.8}$$

where F, G, and H are properly defined matrices, Q is the variance covariance matrix, and $b(\cdot)$ and $c(\cdot)$ are properly defined functions. This type of distribution, called the exponential family of distributions, can cover a broad class of distributions frequently used in statistical analysis, such as the Poisson distribution and the binomial distribution. The model specified by (1.7) and (1.8) is called the Dynamic Generalized Linear Model (DGLM) (West et al. 1985, West and Harrison 1997). For DGLM, several computationally efficient and precise methods for state estimation were proposed. This type of model has the favourable 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 (West et al. 1985, Fahrmeir 1992, Schnatter 1986, Frühwirth-Schnatter 1994, Durbin and Koopman 1997).

1.2 Monte Carlo filter

The GSSM is very flexible and suitable for a wide variety of time series, and its usefulness is revealed by the fact that it allows the use of recursive formulas for estimating the state vector given observations (Kitagawa 1987). However we have to solve computational problems due to the repeated necessary integrations over a state space, which increases enormously with respect to the state dimension n_x (Carlin, Polson and Stoffer 1992, Fahrmeir 1992, Frühwirth-Schnatter 1994, Schnatter 1986). A Monte Carlo method for filtering and smoothing, called Monte Carlo Filter (MCF) has been proposed (Kitagawa 1996) to overcome this numerical problem. The "bootstrap filter" (Gordon et al. 1993, Doucet, Barat and Duvaut 1995) is a similar algorithm.

To review MCF, suppose that $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ and $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ are approximated by the N realizations,

$$X_{t|t-1} \equiv \{\mathbf{x}_{t|t-1}^{(i)} | i = 1, \ldots, N\} \qquad \text{and} \tag{1.9}$$

$$X_{t|t} \equiv \{\mathbf{x}_{t|t}^{(i)} | i = 1, \ldots, N\}, \tag{1.10}$$

respectively. It can be shown that these particles can be generated recursively by the following algorithm:

1. For $i = 1, \ldots, N$ generate n_x-dimensional random number $\mathbf{x}_{0|0}^{(i)} \sim p_0(\mathbf{x})$.

2. Repeat the following steps for $t = 1, \ldots, T$. For (a), (b) and (c), repeat N times independently for $i = 1, \ldots, N$.

(a) Generate n_v-dimensional system random number $\mathbf{v}_t^{(i)} \sim q(\mathbf{v}|\boldsymbol{\theta}_{sys})$.

(b) Compute $\mathbf{x}_{t|t-1}^{(i)} = f(\mathbf{x}_{t-1|t-1}^{(i)}, \mathbf{v}_t^{(i)})$.

(c) Compute $\widetilde{w}_t^{(i)} = r(\mathbf{y}_t|\mathbf{x}_t = \mathbf{x}_{t|t-1}^{(i)}, \boldsymbol{\theta}_{obs})$.

(d) Obtain $X_{t|t}$ by sampling with replacement from $X_{t|t-1}$ with sampling probabilities proportional to $\widetilde{w}_t^{(1)}, \cdots, \widetilde{w}_t^{(N)}$.

A significant advantage 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 smoothing by storing the past particles and resample the vector of particles $(\mathbf{x}_{t|t-1}^{(i)}, \mathbf{x}_{t-1|t-1}^{(i)}, \cdots, \mathbf{x}_{t-L|t-L}^{(i)})$ rather than the single particle $\mathbf{x}_{t|t-1}^{(i)}$. Other smoothing algorithms can be refferred to (Clapp and Godsill 1999).

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

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) = \int r(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_{obs}) p(\mathbf{x}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) d\mathbf{x}_t \quad \cong \quad \frac{1}{N}\sum_{i=1}^{N} \widetilde{w}_t^{(i)}. \quad (1.11)$$

Then the log likelihood in the MCF, $l^{\mathrm{MCF}}(\boldsymbol{\theta})$, is given by

$$\begin{aligned} \log p(\mathbf{y}_{1:T}|\boldsymbol{\theta}) &= \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) \\ &\cong \sum_{t=1}^{T} \log \left(\sum_{i=1}^{N} \widetilde{w}_t^{(i)} \right) - T \log N = l^{\mathrm{MCF}}(\boldsymbol{\theta}). \end{aligned} \quad (1.12)$$

The log-likelihood calculated by MCF, l^{MCF}, suffers from an error inherent in the Monte Carlo approximation, giving rise to difficulties in parameter estimation which is usually accomplished via maximization of the log-likelihood. This error problem somewhat reduces the applicability of MCF in practical data analysis, and an alternative procedure for estimating a parameter vector is now required unless a vary large number of particles are used or an average of the approximated log-likelihoods is computed by the massive parallel use of many Monte Carlo filters.

The organization of this paper is as follows. In Section 2, we explain two types of self-organizing time series model which can diminish the difficulties associated with parameter estimation. One is a Genetic algorithm filter which is based on a strong parallelism of the Monte Carlo filter to a Genetic algorithm (GA). The other one is a self-organizing state space model that is an extension of the GSSM. In Section 3, we discuss a resampling procedure which plays an important role in the filtering of the MCF. It is shown in Section 2 that the resampling is identical to the selection procedure in GA that has a variety of methods to implement it on a computer. The typical three selection

methods are realized in the filtering of the MCF, and then each performance is discussed by applying them to a smoothing problem of a time series. In Section 4, we demonstrate an example of applying the self-organizing state space model. In this study, we focus on estimation of a time-dependent frequency which changes rapidly over a relatively short interval. In particular, we deal with an analysis of the small count time series for which a time-dependent mean exhibits a wavy behaviour whose frequency evolves in time. For this case, conventional methods of identifying a time-varying frequency are incapable of estimating it precisely. A benefit of considering the self-organizing state space model for this problem is illustrated by applying it to actual data. Finally, we conclude in Section 5.

2 Self–organizing time series model

2.1 Genetic algorithm filter

2.1.1 Parallelism to genetic algorithm

The Genetic Algorithm (GA) is the adaptive search procedure inspired by evolution and one of the most popular amongst the population-based optimization techniques (Holland 1975, Goldberg 1989, Davis 1991, Whitley 1994). GA is characterized by keeping the N candidates for optimal solution at each iteration composed of three steps: crossover, mutation, and selection (or reproduction). Each candidate at the t–th iteration is called a string at t–th generation. It has been pointed out that there exists a close relationship between MCF and GA and that an essential structure involved in MCF is quite similar to that in GA (Higuchi 1997). The correspondence between MCF's and GA's terminology is summarized in Table 1. Here we give a brief explanation of the strong parallelism of MCF to GA.

$\mathbf{x}_{t|t-1}^{(i)}$ and $\mathbf{x}_{t|t}^{(i)}$ identified as particles in the MCF are considered as the strings in GA. The filtering procedure of MCF exactly corresponds to the selection of GA by regarding $r(\cdot|\boldsymbol{\theta}_{obs})/N$ as the evaluation function in GA. Accordingly in the GA, the importance weight $w_t^{(i)} = r(\mathbf{y}_t|\mathbf{x}_{t|t-1}^{(i)}, \boldsymbol{\theta}_{obs}) / \sum_{j=1}^{N} r(\mathbf{y}_t|\mathbf{x}_{t|t-1}^{(j)}, \boldsymbol{\theta}_{obs})$ can be called the fitness of the string $\mathbf{x}_{t|t-1}^{(i)}$. It should be noticed that whereas $r(\mathbf{y}_t|\mathbf{x}_{t|t-1}^{(i)}, \boldsymbol{\theta}_{obs})/N$ is obviously dependent on time due to the presence of data $\mathbf{y}_t$, the evaluation function in GA without any modification such as a scaling (stretching) (e.g., (Goldberg 1989)) is independent of t. $\mathbf{x}_{t-1|t-1}^{(i)}$ and $\mathbf{x}_{t|t-1}^{(i)}$ are regarded as the parent and offspring, respectively, because $\mathbf{x}_{t-1|t-1}^{(i)}$ creates $\mathbf{x}_{t|t-1}^{(i)}$ which undergoes selection. The maximum likelihood principle is interpreted as a rule to choose the model that maximizes the family fitness under a circumstance of given data $\mathbf{y}_{1:T}$.

The crossover and mutation in GA may be interpreted from the Bayesian

Monte Carlo Filter (MCF)	$\Longleftrightarrow$	Genetic Algorithm (GA)
t: time	—	generation
$\mathbf{x}_{t\|t-1}^{(i)}$ and $\mathbf{x}_{t\|t}^{(i)}$: particle	—	string
$f(\cdot, \mathbf{v}_t = 0)$	—	genetic drift
system noise	—	crossover, mutation
filtering	—	selection
$r(\mathbf{y}_t\|\mathbf{x}_t, \boldsymbol{\theta}_{obs})/N$	—	evaluation function
$w_t^{(i)}$	—	fitness
$p(\mathbf{y}_t\|\mathbf{y}_{1:t-1}, \boldsymbol{\theta})$	—	population fitness
$p(\mathbf{y}_{1:T}\|\boldsymbol{\theta})$	—	family fitness
$\mathbf{y}_t$: observation	—	environment
$\mathbf{y}_{1:T}$: data	—	history of environment

Table 1. Analogy between the MCF and the GA

point of view. In particular, a fluctuation caused by a mutation is exactly regarded as the system noise which stems from a non–Gaussian probability density function. When we represent a model using a binary code and use the canonical GA, a closed form of this density function can be obtained easily. In the MCF there exists an apparent movement of the population at each step which results from the systematic behaviour of the particle driven by a nonlinear function f with $\mathbf{v}_t \equiv 0$ in the system model. The GA penalises any such drift in model space that is independent of any performance of optimization, called genetic drift.

2.1.2 GA filter

Based on the close relationship between MCF and GA, a new algorithm, in which prediction step in MCF is replaced by the mutation and crossover operators in GA, has been proposed to avoid an estimation of $\boldsymbol{\theta}_{sys}$ involved in the system model via maximization of $l^{\mathrm{MCF}}(\boldsymbol{\theta})$. While within MCF a stochastic behaviour of particles is determined by an outer force independent of particle, the crossover yields strong interactions between particles and a description of the resulting stochastic behaviour cannot take a simple form as in MCF. Namely, there may occur a gradual change in inertial force driving the population through the crossover. This means that the time series model involving the genetic operators is no longer the same model as one expressed in terms of GSSM. We therefore call the GSSM involving these genetic operators GA filter. Examples of applications of the GA filter including a seasonal adjustment

can be found (Higuchi 1997).

2.2 Self–organizing state space model

As explained in Section 1.2, the parameter estimation of $\boldsymbol{\theta}$ via maximization of the likelihood happens to be made impractical by the sampling error of $l^{\mathrm{MCF}}(\boldsymbol{\theta})$. To mitigate such difficulty, a self-organizing state space model has been proposed (Kitagawa 1998). The general form of the self-organizing state space model is obtained by augmenting the state vector $\mathbf{x}_t$ with the parameter vector $\boldsymbol{\theta}$ as $\mathbf{z}_t = [\mathbf{x}_t', \boldsymbol{\theta}_t']'$. The state space model for this augmented state vector $\mathbf{z}_t$ is given by

$$\begin{aligned} \textit{system model} \qquad \mathbf{z}_t &= F^*(\mathbf{z}_{t-1}, \mathbf{v}_t) && (2.1) \\ \textit{observation model} \qquad \mathbf{y}_t &\sim r(\cdot|\mathbf{z}_t), && (2.2) \end{aligned}$$

where the nonlinear function F^* is defined by

$$F^*(\mathbf{z}, \mathbf{v}) = \begin{bmatrix} f(\mathbf{x}, \mathbf{v}) \\ \boldsymbol{\theta} \end{bmatrix}. \tag{2.3}$$

f and r are given in (1.1) and (1.2), respectively. Another approach to deal with such a problem of learning about time-varying state vectors and fixed model parameters is discussed (Liu and West 2000).

Here we consider Bayesian estimation of $\boldsymbol{\theta}$ instead of using the maximum likelihood method. Once Monte Carlo smoothing is performed for this self-organizing state space model, the posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$, is simply approximated by $p(\boldsymbol{\theta}|\mathbf{y}_{1:T}) \cong \int p(\mathbf{z}_T|\mathbf{y}_{1:T})d\mathbf{x}_T$. Similarly, the posterior distribution of the state vector $\mathbf{x}_t$ is defined by a marginal posterior distribution given by $p(\mathbf{x}_t|\mathbf{y}_{1:T}) = \int p(\mathbf{z}_t|\mathbf{y}_{1:T})d\boldsymbol{\theta}_t$.

When we need to define an optimal $\boldsymbol{\theta}$, several methods are available based on $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$. In this study, the optimal $\boldsymbol{\theta}$ is determined by calculating a median of the marginal distribution of $p(\boldsymbol{\theta}|\mathbf{y}_{1:T})$ for each variable. A successful usage of this parameter estimation procedure relies on how to set its initial distribution, $p_0(\boldsymbol{\theta})$. Usually, it is recommended to adopt a uniform distribution over a range which covers the possible values of an optimal $\boldsymbol{\theta}$. For the cases with available prior information, the distribution can be tailored to each application.

For the model with time-varying parameter, $\boldsymbol{\theta} = \boldsymbol{\theta}_t$, a model for time-changes of the parameter $\boldsymbol{\theta}_t$ is necessary. For example, we may use the random walk model $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t$ is a white noise with density function $\phi(\boldsymbol{\varepsilon}|\boldsymbol{\xi})$. For this case, the nonlinear function F^* is defined by

$$F^*(\mathbf{z}_{t-1}, \mathbf{u}_t) = \begin{bmatrix} f(\mathbf{x}_{t-1}, \mathbf{v}_t) \\ \boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t \end{bmatrix}, \tag{2.4}$$

where the system noise is defined by $\mathbf{u}_t = [\mathbf{v}_t', \boldsymbol{\varepsilon}_t']'$. 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 parameter in the state vector. In the self-organizing state space models, the parameter vector, $\boldsymbol{\theta}_t$, is automatically determined as the estimate of the state vector. The parameter to be estimated by the maximum likelihood method, $\boldsymbol{\xi}$, is anticipated to be a one or two dimensional vector.

3 Resampling scheme of the filtering

3.1 Selection scheme

As mentioned above, the resampling form appearing in the filtering of the MCF is identical to one in the selection procedure of the GA. Consequently we can choose the resampling scheme best suited for estimating the log–likelihood, from the several different ways to do selection (Goldberg 1989, Whitley 1994). Here we compare three typical selection methods among selection alternatives classified by Brindle (Brindle 1981):

1. Stochastic sampling with replacement (Roulette wheel selection)
2. Remainder stochastic sampling with replacement
3. Deterministic sampling

To explain them, we imagine a roulette wheel where each particle $\mathbf{x}_{t|t-1}^{(i)}$ is represented by a space that proportionally corresponds to $w_t^{(i)}$. The stochastic sampling is very simple such that we choose N particles by repeatedly spinning this roulette wheel. The remainder stochastic sampling begins with calculating the integer and fractional portions of $Nw_t^{(i)}$, $k_t^{(i)}$ and $\xi_t^{(i)}$, respectively. We denote the sum of $k_t^{(i)}$ by $N_{int} = \sum_{i=1}^{N} k_t^{(i)}$. Secondly, we get $k_t^{(i)}$ copies of $\mathbf{x}_{t|t-1}^{(i)}$ for each particle. Finally, we choose the $N_{frac} = N - N_{int}$ particles by repeatedly spinning the "remainder" roulette wheel where each particle has a space in proportion to $\xi_t^{(i)}$. This sampling method can be realized in an efficient way by adopting the resampling method known as stochastic universal sampling (Baker 1987). Deterministic sampling does not employ this roulette wheel but instead chooses the N_{frac} particles in a deterministic way such that we order $\xi_t^{(i)}$ and take the N_{frac} largest particles.

3.2 Comparison of performance: Simulation study

We discuss three resampling methods by considering a problem of smoothing a time series. Figure 1(a) shows the artificially generated data which is obtained

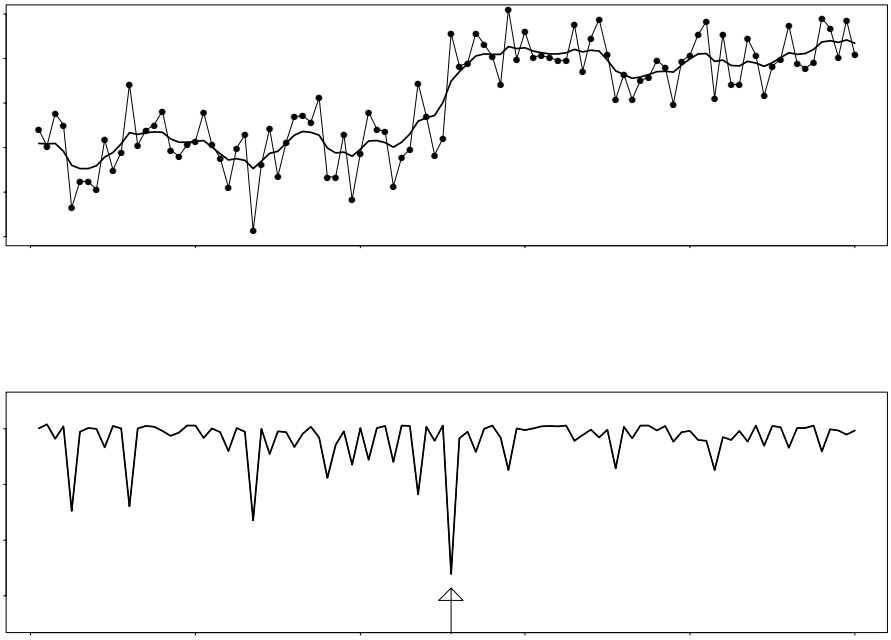

Figure 1. (a) Data and smoothed data by KFS. (b) log likelihood $p(y_t|y_{1:t-1})$.

by adding i.i.d. Gaussian white noise with a variance of 0.1 to the step function with a jump of 1 at $t = 51$, where $T = 100$. To smooth the given data, we consider the simplest model defined by

$$\begin{aligned} \textit{system model} \quad \mu_t &= \mu_{t-1} + v_t, \qquad v_t \sim N(0, \tau^2) \\ \textit{observation model} \quad y_t &= \mu_t + e_t, \qquad e_t \sim N(0, \sigma^2) \end{aligned} \tag{3.1}$$

where μ_t is a trend component at time t, thereby setting the 1–dimensional state vector $\mathbf{x}_t = [\mu_t]$, and $\mathbf{v}_t$ is the 1– dimensional white noise sequences. For this case, the Kalman filter yields a "true" value of the likelihood. The true value of the log-likelihood based on the Kalman filter is specified by l^{true} henceforth in which we omit $\boldsymbol{\theta}$ for simplicity. In this case, the maximum likelihood estimate $\boldsymbol{\theta}^* = [\sigma^{2*}, \tau^{2*}]'$ can be easily obtained, because the likelihood function has a very simple form as a function of $\lambda^2 = \tau^2/\sigma^2$, the ratio of the variance of the system noise to that of the observation noise. The solid curve is obtained by applying the fixed interval smoother with $\lambda^{2*} = 0.25$.

Figure 1(b) shows the log–likelihood of $\mathbf{y}_t$ for the model with λ^{2*}, $\log p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}^*)$, which is denoted by l_t^{true}. It is seen that l_t^{true} takes a minimum value at $t = 51$, indicated by the arrow, where there is a discontinuous change

in the given trend. An approximation by MCF to l_t^{true} is denoted by $l_t^{\text{MCF},[j]}$, where $[j]$ indicates the jth trial.

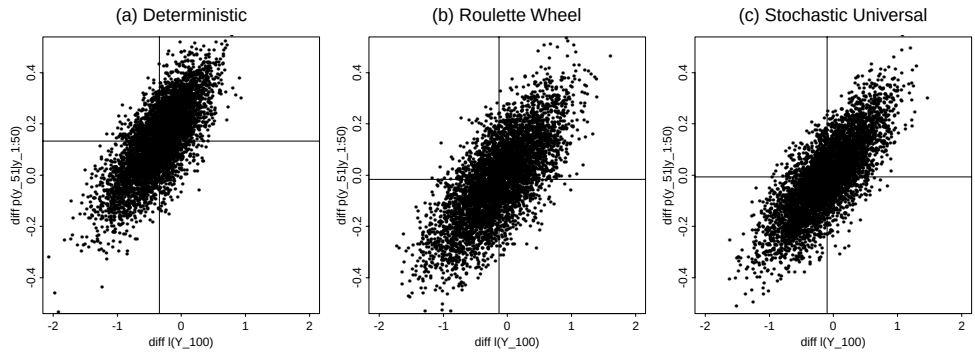

Figure 2. Plots of $N_p = 5,000$ trials of $\Delta l^{[j]}$ versus $\Delta l_{51}^{[j]}$ for (a) deterministic, (b) roulette wheel, and (c) stochastic universal sampling methods.

We concentrate on the discrepancy of $l^{\text{MCF},[j]}$ from l^{true}, $\Delta l^{[j]}$, as well as $\Delta l_{51}^{[j]} = l_{51}^{\text{MCF},[j]} - l_{51}^{\text{true}}$. The reason why we pay special attention to $\Delta l_{51}^{[j]}$ is that a sparse distribution of $\mathbf{x}_{51|50}^{(i)}$ around y_{51} could result in poor approximation of $l_{51}^{\text{MCF},[j]}$ to l_{51}^{true}. Figure 2(a) shows a distribution of $(\Delta l^{[j]}, \Delta l_{51}^{[j]})$ for a case where the deterministic sampling is adopted to implement the resampling in the filtering procedure of MCF. The number of trials is $N_p = 5,000$ and the particle number is fixed $N = 1,000$ for each trial. The vertical and horizontal lines indicate the mean values of $\Delta l^{[j]}$ and $\Delta l_{51}^{[j]}$, $\overline{\Delta l^{[j]}}$ and $\overline{\Delta l_{51}^{[j]}}$, respectively. We call $\overline{\Delta l^{[j]}}$ the bias. An apparent positive correlation seen in this figure (correlation coefficient $\rho = 0.737$) supports our conjecture that the poor approximation of the MCF to l_{51}^{true} is significant in controlling $\Delta l^{[j]}$. The straight line obtained by the least squares fit to these plots has a slope of 0.265. ¿From the positive value of slope smaller than 1, it suggests that there remains a systematic error in approximation of $l_{t'}^{\text{MCF}}$ for $51 < t'$. Namely $X_{51|50}$ with negative $\Delta l_{51}^{[j]}$ may lead $X_{52|51}$ to give negative $\Delta l_{52}^{[j]}$ through $X_{51|51}$.

Figures 2(b) and (c) demonstrate a distribution of $(\Delta l^{[j]}, \Delta l_{51}^{[j]})$ for the roulette wheel selection and for the stochastic universal sampling, respectively. As in Figure 2(a), we can see the same tendency that $\Delta l^{[j]}$ depends mainly on $\Delta l_{51}^{[j]}$. Whereas the sign of the bias is common to three cases ($\overline{\Delta l^{[j]}}$ <0), its magnitude for the deterministic sampling is apparently larger than others. It is clearly seen that plots of $(\Delta l^{[j]}, \Delta l_{51}^{[j]})$ for the roulette wheel selection show a broader distribution compared with plots for the stochastic universal sampling. Table 2 summarizes the bias, standard deviation of $\Delta l^{[j]}$, correlation coefficient, and slope of the line using a least squares fit, for each

resampling scheme. Based on these statistics, we suggest that the stochastic universal sampling is the most suitable for the resampling in the filtering procedure of the MCF.

Sampling Scheme	$\overline{\Delta l^{[j]}}$	s.d. of $\Delta l^{[j]}$	ρ	slope
Deterministic	-0.344	0.392	0.737	0.265
Roulette wheel	-0.127	0.496	0.723	0.234
stochastic universal	-0.095	0.455	0.776	0.249

Table 2. Comparison of selection schemes: $N_p = 5,000$ trials

A reduction of standard deviation can be achieved by using the resampling suitable for the filtering, but Figure 2 demonstrates that there still remains a larger variance unfavourable for choosing an optimal parameter value. Of course, we can make a drastic reduction of variance by increasing the particle number N as much as the computer memory space permits. Even an increase in N only for the predictive distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ would lead to a significant reduction of variance (Gordon et al. 1993). There are interesting resampling schemes designed to reduce a variance (Crisan, Moral and Lyons 1999, Crisan and Grunwald 1999).

4 Application

4.1 Time-varying frequency wave in small count data

In this article we are concerned with an analysis of the nearly cyclic (wave–like) signal which can be characterized by repeating a similar pattern successively with a gradual change in its wave form in the time domain. Most of the attention in actual data analysis is usually paid to an estimation of a time–dependence of its frequency (or equivalently, wavelength) and/or amplitude among quantities specifying its time–dependent structure. In particular, when the wave–like signal seems to consist of a sinusoid corrupted by observation noise, a good estimation of time–dependence of frequency may lead to further insights into a system which generates the cyclic behaviour. For example, such kind of a task to investigate the time-varying structure of a power spectrum with a few peaks plays an important role in a destructive test of the wing of the airplane in the laboratory, because an eigen oscillation of the wing shows a different behaviour as the wind velocity increases towards making a wing destructive (Ikoma 1996). In this study we deal with such kind of a monotonic sinusoidal wave whose frequency changes as time goes.

In general, an attempt to estimate the time-varying frequency of the time-varying frequency wave is achieved by searching a significant peak over an

instantaneous power spectra (IPS) of which the concept can be realized in different ways. One of the useful methods for calculating IPS is to fit the time-varying coefficient AR model (TVCAR) to data, giving IPS that is based on the relationship between a stationary AR process and the theoretical spectrum of the process (Kitagawa and Gersch 1985, Gersch and Kitagawa 1988). One of the motivations to develop a method for an analysis of the time-varying frequency wave is that when an identification of the time-varying frequency wave is difficult due to significant observation noise, TVCAR is likely to provide no peak in IPS. This occurs even in a case where a signature of the time-varying frequency wave can be obviously detected by visual inspection of plots of data. Another motivation comes from a demand such that we treat the time–varying frequency as the state variable (Ikoma 1996), because a linear increase or decrease in the time–varying frequency is very frequently discussed in data analysis and theory.

Our attention here is focused on the time series involving small counts which are frequently obtained in many fields such as the biomedical statistics (for example, monthly numbers of Polio incidences (Zeger 1988)) and astrophysics. Recent progress in instruments for measuring faint light from star brightness such as CCD camera has led to a drastic change in the obtained data form from light intensity, which is usually allowed to take a continuous value, to a photon count number. When we want to investigate a behaviour of targets with a small time scale, the sampling time within which photons arriving at the instrument are counted is shortened and results in giving us the small count data time series. Thus it is important to model a small count data in astrophysics.

The reason why we deal with small count data in terms of a statistical view point is that the time series involving relatively larger counts can be well analyzed by means of a time series model without regarding the fact that an observed data follows a discrete distribution, and does not require more sophisticated models in practice. Actually recent Bayesian approaches for a time series model have paid considerable interest to an analysis of small count data as an example to illustrate their applications (West et al. 1985, Fahrmeir 1992, Frühwirth-Schnatter 1994, Chan and Ledolter 1995, Kashiwagi and Yanagimoto 1992, Higuchi 1999). A practical example of the applications of the Monte Calro filter to various non-linear problems have been found in (Carpenter, Clifford and Fearnhead 1999).

4.2 Self-organizing state space model for time-varying frequency wave

We assume that the observation is generated from a Poisson distribution with time–varying mean λ_t: $\mathbf{y}_t \sim Poisson(\lambda_t)$. $\log \lambda_t$ is in this study decomposed into two factors: a trend component μ_t, and time–varying frequency

wave component s_t, $\log \lambda_t = \mu_t + s_t$. In other words, we deal with the non–stationary Poisson model in which the time–varying mean is expressed as the multiplicative form given by $\lambda_t = \exp(\mu_t)\exp(s_t)$.

We assume that μ_t follows a first order trend model given by $\mu_t = \mu_{t-1} + v_{t,\mu},\ v_{t,\mu} \sim C(0, \tau_\mu^2)$, where $C(0, \tau_\mu^2)$ represents the Cauchy distribution concentrated around 0 with dispersion parameter τ_μ^2. As mentioned above, we want to estimate the time–varying frequency. Accordingly we denote the time-varying frequency at time t by f_t and treat it as the state variable. The time–varying frequency wave component, s_t, with a time-varying frequency f_t, is described by the second order difference equation

$$s_t = 2\cos(2\pi f_t)\, s_{t-1} - s_{t-2}. \tag{4.1}$$

Since f_t is allowed to take a value between 0 to 1/2, a nonlinear transformation based on a sigmoid function, $f_t = 0.5/(1+\exp(-\beta_t))$ $(-\infty < \beta_t < +\infty)$, is performed and β_t is used as a state variable instead of f_t. We assume that β_t follows a first order trend model given by $\beta_t = \beta_{t-1} + v_{t,\beta},\ v_{t,\beta} \sim N(0, \tau_\beta^2)$. The parameter vector, $[\log_{10}\tau_\mu^2, \log_{10}\tau_\beta^2]'$, is also included in the state vector to optimize it by means of the self-organizing state space model. Then these models can be represented by the state space model form:

system model

$$\mathbf{x}_t = \begin{bmatrix} \mu_t \\ s_t \\ s_{t-1} \\ \beta_t \\ \log_{10}\tau_{t,\mu}^2 \\ \log_{10}\tau_{t,\beta}^2 \end{bmatrix} = \begin{bmatrix} \mu_{t-1} + v_{t,\mu} \\ 2\cos\left(\frac{\pi}{1+\exp(-\beta_{t-1})}\right) s_{t-1} - s_{t-2} \\ s_{t-1} \\ \beta_{t-1} + v_{t,\beta} \\ \log_{10}\tau_{t-1,\mu}^2 \\ \log_{10}\tau_{t-1,\beta}^2 \end{bmatrix}, \tag{4.2}$$

where $v_{t,\mu} \sim C(0, \tau_{t,\mu}^2)$ and $v_{t,\beta} \sim N(0, \tau_{t,\beta}^2)$.

4.3 Result

To demonstrate the applicability of our method to the small count time-varying frequency wave data, a simple test is performed on a simulated data set. f_t is beforehand given to show a linear increase and decrease during a relatively short interval. A dotted line in Figure 3(b) indicates this dependency of given f_t on time. The next step to generate simulated data is to define an amplitude and phase in s_t. A phase is determined by integrating given f_t. The time-varying amplitude is taken to present a gradual increase and decrease as indicated by the the dotted line in Figure 3(a). The given fixed value of μ_t $(\exp(\mu_t) = 20)$ together with given s_t produces λ_t by $\lambda_t = \exp(\mu_t + s_t)$. Finally $\mathbf{y}_t$ is obtained by drawing a random number from $Poisson(\lambda_t)$. The

total data number is $T = 100$. The solid straight line in Figure 3(a) shows simulated data obtained by the aforementioned procedure.

To illustrate good performance of our model in terms of an accuracy of estimation of the time-varying frequency, the TVCAR procedure is first of all applied to attempt to get a spectral peak frequency within a frequency range of $f = 0 \sim 0.2$ from IPS. The TVCAR with varying $M_{\rm AR}$ in a range of $M_{\rm AR} = 2, 4, \ldots, 10$ for $N_{obs} = 1, 2, 5, 10, 20$ is applied in this study, where $M_{\rm AR}$ and N_{obs} are the numbers of the AR coefficients and of successive data points for which the time-varying AR coefficients are set to be constant (for detail, see (Kitagawa and Gersch 1985, Gersch and Kitagawa 1988)). In fact, any combination of $(N_{obs}, M_{\rm AR})$ fails to define any peak frequency due to absence of peak in IPS.

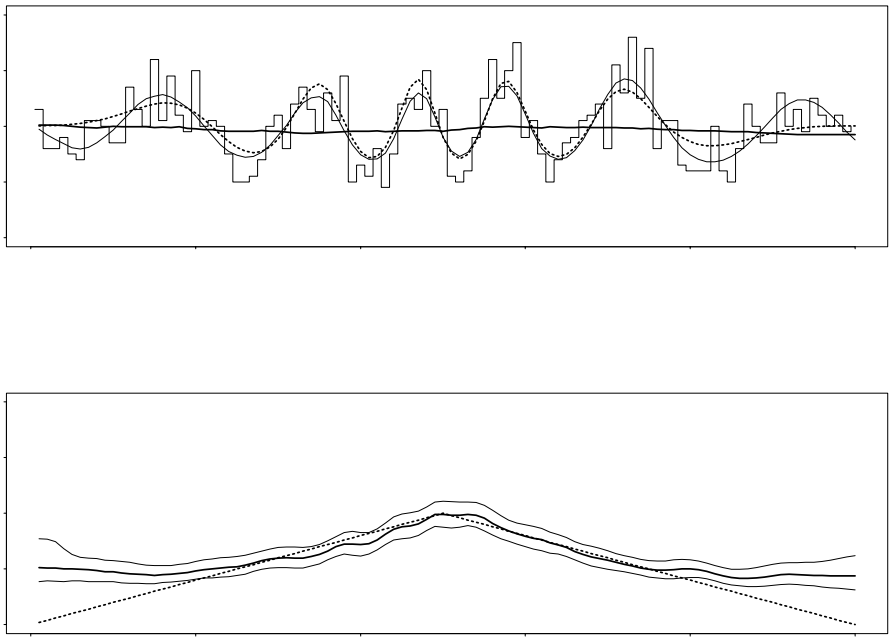

Figure 3. Simulation results. (a) Given data and estimated $\lambda_t = \exp(\mu_t + s_t)$. The thick line shows the estimated $\exp(\mu_t)$. (b) Estimated f_t and its confidence interval $\pm\sigma$.

We apply the self-organizing state space model explained in Section 4.2 to this data set. For the initial state $\mathbf{x}_0$, it is assumed, for example, that $\mu_0 \sim N(2.56, 1)$, $s_{-1} \sim N(0, 0.5)$, $s_0 \sim (0, 0.5)$, $\beta_0 \sim U([-2.60, -0.81])$, $\log_{10}\tau^2_{0,\mu} = U([-7, -4])$, $\log_{10}\tau^2_{0,\beta} = U([-3, 0])$, where $U([\ ,\ ])$ denotes the uniform distribution. The initial range of β_0 corresponds to the range of

f_0 between 0.0346 and 0.154, which is inferred from a visual inspection of the given data. The self-organizing state space model estimates the values of $\log_{10}\tau^2_{T,\mu}$ and $\log_{10}\tau^2_{T,\beta}$: $\log_{10}\widehat{\tau}^2_{T,\mu} = -5.36(-5.99, -4.48)$, $\log_{10}\widehat{\tau}^2_{T,\beta} = -1.83(-2.14, -1.48)$. The two values in the parentheses indicate the confidence interval $\pm\sigma$.

The obtained optimal values for $\widehat{\tau}^2_{\mu}$ and $\widehat{\tau}^2_{\beta}$ allow us to conduct the MCF for the generalized state space model which is simply defined by setting $\tau^2_{\mu} = 10^{-5.36}$ and $\tau^2_{\beta} = 10^{-1.83}$ and taking $\mathbf{x}_t = [\mu_t, s_t, s_{t-1}, \beta_t]'$ as the state vector. The marginal posterior distribution of f_t, $p(f_t|\mathbf{y}_{1:T})$, is determined through $p(\mathbf{x}_t|\mathbf{y}_{1:T})$, and its median is shown by a thick line in Figure 3(b) as well as its confidence interval $\pm\sigma$ (two thin lines). Of course, the self-organizing state space model can produce the posterior distribution of f_t, $p(f_t|\mathbf{y}_{1:T})$, which can be obtained from the marginal posterior distribution of β_t, $p(\beta_t|\mathbf{y}_{1:T})$, but we are interested in examining how much of a difference arises between the two estimates. In fact, the estimate obtained by the generalized state space model is only shown in this study, because the difference between the two estimates is invisible.

The estimated $\lambda_t = \exp(\mu_t + s_t)$ and $\exp(\mu_t)$ are in Figure 3(a) denoted by the thin and thick lines, respectively. While a good agreement of the estimate with the given true curve of λ_t can be seen for the interval between $t = 20$ and 80, a significant discrepancy is found for the outside intervals where the given amplitude is smaller than that for the middle interval. A poor agreement of the estimated λ_t to the given one for those intervals turns out to be made apparent by a comparison of the estimated f_t to the given one (dotted straight line) in Figure 3(b). Such poor performance is explained by two reasons. One is due to the fact that the given amplitude for those intervals is smaller than that for the middle interval. The other one is inherent in the system model (4.1) which becomes close to the second order trend model as f_t decreases. Therefore, an estimation of the time-varying frequency is intrinsically difficult by means of the system model (4.1) for a case with $\cos(2\pi f_t) \sim 1$.

Here we show one example of an application of the above mentioned procedure to an actual data set. The data set which will be used for an illustration of our method is the signature of the spiral density waves obtained by the Voyager photopolarimeter (PPS) during a stellar occultation of Saturn's rings. The original Voyager PPS data is also a small count time series. The discussion of instruments and operation of Voyager PPS data can be seen in (Horn, Showalter and Russell 1996). In fact, we use the simulated data, shown in Figure 4(a), which is generated to resemble the typical density wave, named Prometheus 7:6. The reason why we have not used the original data is that the data set we have at hand has already undergone the detrending procedure, and then is no longer a count data. The simulated data is obtained by taking an average over every three points of the original data and quantizing it in an attempt to generate a count data. The behaviour displayed in this figure is one of the strongest density waves and the signature of the time-varying

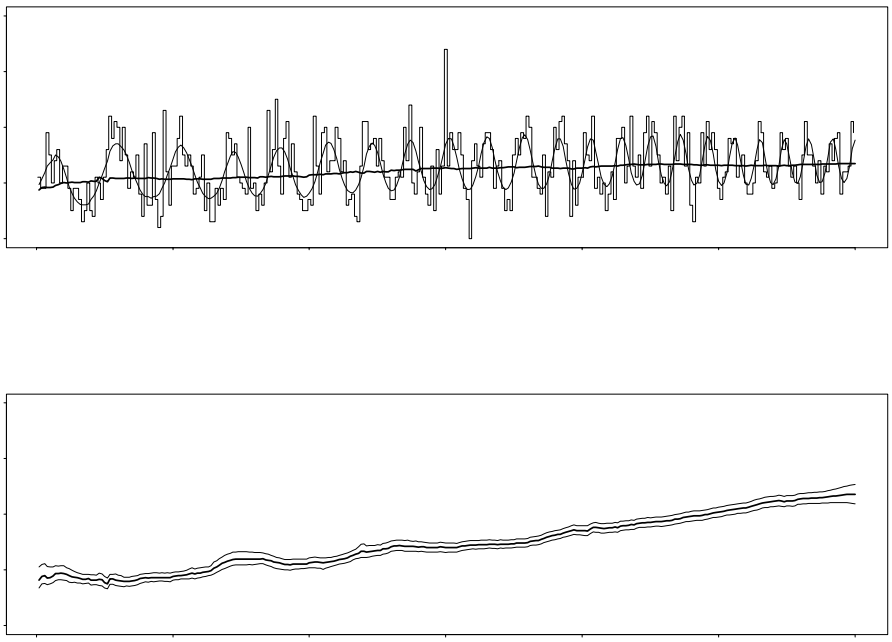

Figure 4. Analysis of the Voyager PPS data. (a) Given data and estimated $\lambda_t = \exp(\mu_t + s_t)$. The thick line shows the estimated $\exp(\mu_t)$. (b) Estimated f_t and its confidence interval $\pm\sigma$. The dotted line indicates the given f_t.

frequency wave is clearly visible in the data. Voyager data has revealed a wide variety of ring structure of Saturn; waves and wakes. Understanding of this structure plays an important role in studying a generation of planetary ring structure. A detail explanation of the spiral density waves is referred to e.g., (Horn et al. 1996).

It has been reported that a theory of the spiral density wave suggests a linear increase in f_t (Horn et al. 1996). We apply the self-organizing state space model as used in the simulated data set. The solid and thick lines in Figure 4(a) indicate the estimated λ_t and $\exp(\mu_t)$, respectively, which are based on the marginal posterior distribution $p(\mathbf{x}_t|\mathbf{y}_{1:T})$. Figure 4(b) shows an estimated f_t and its confidence interval $\pm\sigma$. The tendency of a linear increase in f_t, which is expected from a theoretical model based approach, is clearly identified.

5 Conclusions

In this study, we introduced two types of self-organizing time series models: the Genetic algorithm filter and the self-organizing state space model. The former has been developed on a basis of the strong parallelism between the Genetic algorithm and the MCF. The latter is an extension of the GSSM which is formed by augmenting the state vector with a parameter vector. Both methods mitigate the difficulty of parameter estimation via maximization of the likelihood associated with the sampling error. Namely, the self-organizing time series model makes a model selection procedure practical and facilitates the automatic processing of time series data. Obviously, the required amount of computation to realize the self-organizing time series model 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

Anderson, B. D. and Moore, J. B. (1979). *Optimal Filtering*, Prentice-Hall, New Jersey.

Baker, J. (1987). Reducing bias and inefficiency in the selection algorithm, *Proceedings of the Second International Conference on Genetic Algorithms and Their Applications*, Lawrence Erlbaum.

Brindle, A. (1981). Genetic algorithm for function optimization, *(Unpublished doctorial dissertation)*, University of Alberta, Edomonton.

Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992). Monte Carlo approach to nonnormal and nonlinear state–space modeling, *Journal of the American Statistical Association* **87**(418): 493–500.

Carpenter, J., Clifford, P. and Fearnhead, P. (1999). An improved particle filter for non-linear problems, *IEE Proceedings - F: Radar, Sonar and Navigation* **146**: 2–7.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts, *Journal of the American Statistical Association* **90**(429): 242–252.

Clapp, T. C. and Godsill, S. J. (1999). Fixed-lag smoothing using sequential importance sampling, *Bayesian Statistics 6*, Oxford University Press.

Crisan, D. and Grunwald, M. (1999). Large deviation comparison of branching algorithms versus resampling algorithms: Application to discrete

time stochastic filtering, *Cambridge University Statistical Laboratory, Research Report 9*.

Crisan, D., Moral, P. D. and Lyons, T. (1999). Discrete filtering using branching and interacting particle systems, *Markov Processes and Related Fields* **5**(3): 293–319.

Davis, L. D. (1991). *Handbook of Genetic Algorithm*, Van Nostrand Reinhold, New York.

Doucet, A., Barat, E. and Duvaut, P. (1995). Monte Carlo approach to recursive Bayesian state estimation, *Proc. IEEE Signal Processing/Athos Workshop on Higher Order Statistics*, Girona, Spain, pp. 12–14.

Durbin, J. and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models, *Biometrika* **84**: 1403–1412.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear model, *Journal of the American Statistical Association* **87**(418): 501–509.

Frühwirth-Schnatter, S. (1994). Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering, *Statistics and Computing* **4**: 259–269.

Gersch, W. and Kitagawa, G. (1988). Smoothness priors in time series, *Bayesian Analysis of Time Series and Dynamic Models*, New York, pp. 431–476.

Goldberg, D. E. (1989). *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison-Wesley, Massachusetts.

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

Higuchi, T. (1997). Monte Carlo filter using the genetic algorithm operators, *Journal of Statistical Computation and Simulation* **59**(1): 1–23.

Higuchi, T. (1999). Applications of quasi-periodic oscillation models to seasonal small count time series, *Computational Statistics & Data Analysis* **30**: 281–301.

Holland, J. H. (1975). *Adoption in natural and artificial systems*, The University of Michigan Press, Ann Arbor.

Horn, L. J., Showalter, M. R. and Russell, C. T. (1996). Detection and behavior of pan wakes in Saturn's a ring, *Icarus* **124**: 643–676.

Ikoma, N. (1996). Estimation of time varying peak of power spectrum based on non-Gaussian nonlinear state space modeling, *Journal of Signal Processing* **49**: 85–95.

Kashiwagi, N. and Yanagimoto, T. (1992). Smoothing serial count data through a state-space model, *Biometrics* **48**: 1187–1194.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion), *Journal of the American Statistical Association* **82**(400): 1032–1063.

Kitagawa, G. (1991). A nonlinear smoothing method for time series analysis, *Statistica Sinica* **1**(2): 1032–1063.

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

Kitagawa, G. (1998). Self-organizing state space model, *Journal of the American Statistical Association* **93**(443): 1203–1215.

Kitagawa, G. and Gersch, W. (1985). A smoothness priors time varying AR coefficient modeling of nonstationary time series, *IEEE Trans. on Automatic Control* **AC-30**: 48–56.

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*, Springer-Verlag, New York.

Liu, J. and West, M. (2000). Combined parameter and state estimation in simulation-based filtering, *Sequential Monte Carlo methods in practice*, Springer-Verlag, New-York.

Schnatter, S. (1986). Integration-based Kalman-filtering for a dynamic generalized linear trend model, *Computational Statistics & Data Analysis* **48**(13): 447–459.

Tanizaki, H. (1993). *Nonlinear Filters*, Springer-Verlag, New York.

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models, 2^{nd} ed.*, Springer-Verlag, New York.

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

Whitley, D. (1994). A genetic algorithm tutorial, *Statistics and Computing* **4**: 63–85.

Zeger, S. L. (1988). A regression model for time series of counts, *Biometrika* **75**: 621–629.