ABSTRACT

In this paper, we propose a new forecasting method which we call Bayesian forecasting with multiple state space model. This method increases the robustness of forecasting, that is, it has the ability to adapt to the various situations. A multiple state space model is composed of several state space models called sub-models. Its forecasting value is given by the weighted summation of each sub-model's forecasting value, which is calculated individually by means of a Kalman filter. The weight of each sub-model takes a value proportional to Bayesian posterior probability, which is calculated from the likelihood. A good fitting sub-model's posterior probability increases as the number of observations increases. Prior to forecasting, the initial state and noise variances of each sub-model are needed in order to use the Kalman filter. In this paper these parameters are estimated by numerical maximization of the likelihood concerning these parameters. Examples of how this method may be applied to monthly telephone revenue data and trunk group load data are given, demonstrating the ability of the method for one-step ahead forecasting. The transitions of posterior probabilities of sub-models are also shown.
2. FORECASTING METHOD

Let us consider a multiple state space model composed of n kinds of state space models as sub-model.

2.1 State Space Model

\( M_j \) is the \( j \)-th state space model (\( j=1,2, \ldots, n \)) and is described by a transition equation

\[
x_j(t+1) = F_j x_j(t) + G_j u_j(t)
\]

where,

\( x_j(t) \) is a vector of the state at time \( t \).
\( F_j \) is a system matrix.
\( G_j \) is a driving matrix.
\( u_j(t) \) is system noise at time \( t \).
\( v_j(t) \) is observation noise at time \( t \).

\( y(t) \) denotes an observation value at time \( t \) (\( t=1,2, \ldots \)) and \( y(t) \) is assumed to be represented by an observation equation

\[
y(t) = H_j x_j(t) + v_j(t)
\]

where,

\( H_j \) is an observation matrix.
\( v_j(t) \) is observation noise at time \( t \).

\( F_j, G_j \) and \( H_j \) are assumed to be known matrices. \( H_j^T \) is the transpose matrix of \( H_j \).
\( u_j(t) \) and \( v_j(t) \) are assumed to be mutually independent Gaussian white noise distributed with mean zero and variance \( Q_j \) and \( R_j \), respectively.

Next the observation data series \( Y(1, \cdot) = \{y(t) ; t=1,2, \ldots, i \} \) is divided into several time segments.

\[
Y(1, t_i) = \{y(t) ; t=t_i, t_i+1, t_i+2, \ldots, t_i+\cdot \}
\]

where \( t_i=0 \).

In Ref. [7] the conditional likelihood of model \( M_i \) for the \( i \)-th segment \( Y(1, t_{i-1}) \) is defined as

\[
L_{i,j}( Y(t_{i-1}+1, t_i) | Y(1, t_{i-1}), \Theta_j, x_{j}(0) )
\]

\[
= \frac{1}{(2\pi)^{n/2} |V_{j}(t,t-l)|^{1/2}} \exp \left( - \frac{1}{2} (y(t)-H_j x_{j}(t))^T V_{j}(t,t-l)^{-1} (y(t)-H_j x_{j}(t)) \right)
\]

where \( f_j(y(t) | Y(1,t-l)) \) is the conditional probability density function of \( y(t) \) given the past history \( Y(1,t-l), \Theta_j \) and \( x_j(0) \).

\( \Theta_j = (Q_j, R_j) \) is a vector composed of the system noise and observation noise variances of model \( M_j \). \( x_j(0) \) is the initial state of model \( M_j \). \( L_{i,j} \) is considered the local likelihood of model \( M_j \) for the \( i \)-th segment.

Assuming that \( x_j(0), u_j(t) \) and \( v_j(t) \) are Gaussian, the probability density function \( f_j \) given in the above expression is also Gaussian and Eq. (3) takes the form

\[
L_{i,j}( Y(t_{i-1}+1, t_i) | Y(1, t_{i-1}), \Theta_j, x_{j}(0) )
\]

\[
= \frac{1}{(2\pi)^{n/2} |V_{j}(t,t-l)|^{1/2}} \exp \left( - \frac{1}{2} (y(t)-H_j x_{j}(t))^T V_{j}(t,t-l)^{-1} (y(t)-H_j x_{j}(t)) \right)
\]

where, \( e_j(t) \) is defined by

\[
e_j(t) = y(t) - \hat{\theta}_j(t,t-l)
\]

\( \hat{\theta}_j(t,t-l) \) is a one-step ahead forecasting error when adopting the model \( M_j \) and is called innovation process [9]. \( e_j(t) \) is considered the part of the observation \( y(t) \) containing new information not carried \( y(t-1), y(t-2), \ldots \).

\[
\gamma_j(t,t-l) \text{ and } \nu_j(t,t-l) \text{ are, respectively, the conditional mean and its error variance of } y(t) \text{ given } y(1), y(2), \ldots, y(t-l), \Theta_j \text{ and } x_j(0).
\]

In the following section, we will show that \( \gamma_j(t,t-l) \) and \( \nu_j(t,t-l) \) are easily obtained by using Kalman filter recursive formula given \( \Theta_j \) and \( x_j(0) \). Originally \( \Theta_j \) and \( x_j(0) \) are unknown parameters and are estimated from numerical maximization of the conditional likelihood [7],[8] of the training data.

2.2 Recursive Filter [9]

As shown in the previous section, the likelihood of innovation process \( e_j(t) \) of model \( M_j \) is defined by Eq. (4). From our model in Eqs. (1) and (2), \( [\gamma_j(t,t-l)] \) and \( [\nu_j(t,t-l)] \) are obtained as

\[
\gamma_j(t,t-l) = H_j^T [\nu_j(t,t-l)]
\]

\[
\nu_j(t,t-l) = H_j^T P_{j}(t,t-l)H_j + R_j
\]

where \( \nu_j(t,t-l) \) and \( P_{j}(t,t-l) \) are, respectively, the conditional mean and covariance of the state vector \( x(t) \) given the observations up to time \( t-l \).

Given information on the initial state \( x_j(0) \) and an estimate \( \hat{\theta}_j \) of the noise parameters \( \Theta_j \), the conditional mean and covariance of the state \( x_j(t) \) of model \( M_j \) are obtained by a Kalman filter algorithm, as follows.

\[
\hat{\gamma}_j(t,t-l) = P_j \gamma_j(t,t-l)
\]

\[
\hat{\nu}_j(t,t-l) = K_j \left( y(t) - H_j^T \hat{\gamma}_j(t,t-l) \right)
\]

where \( K_j(t) = P_j(t,t-l)H_j + R_j \) is the Kalman gain.

\[
P_j(t,t-l) = (I - K_j(t) H_j) P_j(t,t-l) F_j + G_j Q_j G_j^T
\]

\( \hat{\gamma}_j(t,t-l) \) is the filtering of state \( x(t) \) and \( P_j(t,t-l) \) is its error variance of model \( M_j \) given observations \( y(1), y(2), \ldots, y(t) \). I is an elementary matrix.
2.3 Parameter Estimation

Using the training data \( y(1), y(2), \ldots, y(n) \), we estimate noise parameters \( \Theta \) and initial state \( x_0 \). To use the observation data efficiently in parameter estimation, we propose the use of the backward Kalman filtering technique. The procedure of this technique is as follows.

1. We assume \( \Theta = \Theta^{(0)}, \hat{R}(n,n) = kI \), where \( k \) is large enough to weaken the effect of initial state \( \hat{R}(n,n) \) which is assumed to be an arbitrary value.

2. Using Kalman filter for the data \( y(n), y(n-1), \ldots, y(1) \), we estimate \( \hat{R}(1,1) \).

3. Considering \( \hat{R}(1,1) \) as the initial state, we calculate the likelihood \( L^{(0)} \) by Eq. (4) using a Kalman filter for the data \( y(t), y(t+1), \ldots, y(n) \) and \( \Theta^{(0)} \).

Given \( \Theta^{(0)} \), we can calculate the likelihood \( L^{(0)} \). We determine the optimal noise parameter \( \Theta \) by numerical maximization of the likelihood of the noise parameter. We use the Davidson method [6] for non-linear maximization. Because the stationary Kalman gain \( K \) depends only on the ratio of system noise variance and observation noise variance \([7],[8]\), we normalize the observation noise variance \( R \) into 1, regard the system noise variance \( Q \) as the noise ratio \( Q/R \), and maximize the likelihood of the system noise variance \( Q \) only.

2.4 Bayesian Forecasting

After Kalman filtering for the i-th segment, the conditional likelihoods \( L_{i,j}^{(0)} \) of sub-model \( M_j \) \((j=1,2,\ldots,n)\) is calculated as shown in the previous sections. On the basis of these likelihoods, each sub-model's weight \( w_{i+1,j} \) \((j=1,2,\ldots,n)\) for the \((i+1)\)-th segment is calculated.

In Bayesian statistics, posterior probability is proportional to the product of the prior probability and the likelihood [5].

\[
\text{posterior probability} \propto (\text{prior probability}) \times (\text{likelihood}) \quad (13)
\]

In this Bayesian forecasting \( w_{i+1,j} \), is considered the prior probability of the \((i+1)\)-th segment and the posterior probability of the i-th segment in sub-model \( M_j \). The proportional of the posterior probability can be obtained from Eq. (13). As the summation of all the posterior probabilities is one, the sub-model's weight is calculated by normalizing the product of prior probability and the likelihood of all sub-models in the multiple state space model, as follows.

\[
w_{i+1,j} = \frac{w_{i,j} L_{i,j}}{\sum_{j=1}^{n} w_{i,j} L_{i,j}} \quad (j=1,\ldots,n) \quad (14)
\]

The prior probability for the first segment is defined as

\[
w_{1,j} = \frac{1}{n} \quad (j=1,\ldots,n) \quad (15)
\]

This prior probability is called ignorant prior in Bayesian statistics. In a multiple state space model, each sub-model's weight is equal when there are no observations. With these weights we calculate the Bayesian forecasting value \( \hat{y}(t+1) \) \((t \in (i+1)-\text{th segment})\).

\[
\hat{y}(t+1) = \sum_{j=1}^{n} w_{i+1,j} \hat{y}_{j}(t+1) \quad (16)
\]

The procedures of calculating posterior probability and Bayesian forecasting are shown Fig. 1 and 2.

3. NUMERICAL EXAMPLES

Let us apply the method described above to the forecasting of monthly telephone revenues and trunk group loads. In each example, we consider two different multiple state space models.
3.1 Monthly Telephone Revenue Forecasting

Monthly telephone revenue data is shown in Fig. 3. The total number of months is 60 and a tariff change is imposed at the 21st month. The effect of a tariff change continues from the 21st to the 23rd month. In an earlier work [8] we analyzed the effect of a tariff change using a state space model and estimated it under the condition that it was exactly known when the effect appeared. In Sect. 3.1.1 we show that the multiple state space model is able to adapt to sudden changes e.g., a tariff change, even if we do not know the time of change a priori. In Sect. 3.1.2 parameter estimation of trend model is shown using a multiple state space model.

3.1.1 Multiple Model with Trend Model and Random Walk Model

Let us consider a multiple model composed of a trend model and a random walk model. In a trend model the mean increment is assumed to be constant, that is, \( E[X(t+1) - X(t)] = E[X(t) - X(t-1)] \). \( E \) is the operator to take the expectation. While in a random walk model the mean at time \( n+1 \) is assumed to be the value at time \( n \), that is, \( E[X(t+1)] = E[X(t)] \).

System matrices, driving matrices and observation matrices of each sub-model are as follows.

\[ M_1 : \text{trend model} \]
\[
\begin{align*}
F_1 &= \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}, & G_1 &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & H_1 &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
Q_1 &= 0.05, & R_1 &= 1.0, & X_1(t) &= (X_1(t), X_1(t-1))', \\
\end{align*}
\]

\[ M_2 : \text{random walk model} \]
\[
\begin{align*}
F_2 &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, & G_2 &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & H_2 &= \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
Q_2 &= 0.25, & R_2 &= 1.0, & X_2(t) &= (X_2(t), X_2(t-1))'. \\
\end{align*}
\]

\[ M : \text{Multiple state model (} M_1, M_2 \) \]

Variances of system noise \( Q_1 \) and \( Q_2 \) are estimated by numerical maximization. The training data to estimate them is the initial 12 data of the observation series. The transitions of the posterior probability of \( M_1 \) and \( M_2 \) are shown in Fig. 4. We find the posterior probability of the random walk model is relatively large near the tariff change. In general the forecasting by trend model tends to overshoot after a sudden change. In this case the forecasting accuracy of the random walk model is relatively high. Figure 4 shows that Bayesian multiple model \( M \) approaches trend model \( M_1 \), as the number of observations increases.
<table>
<thead>
<tr>
<th>Table 1 Forecasting accuracy.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
</tr>
<tr>
<td>MS</td>
</tr>
<tr>
<td>MA</td>
</tr>
</tbody>
</table>

3.1.2 Multiple Model with Trend Models of Different Noise Variances

Let us consider a multiple state space model composed of 5 trend models of different system noises. We adopt several arbitrary values as the system noise candidates and apply a multiple state space model to parameter estimation. For the case when the unknown parameter is varies with time, this estimation is useful. System matrices, driving matrices, observation matrices and noise variances of sub-models are as follows.

$$ M_j : 2 \times 2 \text{ trend model} \ (j=1,2,\ldots,5) $$

$$ F_j = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \quad G_j = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad H_j = \begin{bmatrix} 1 \\ 0 \end{bmatrix} $$

$$ R_j = 1.0 $$

$$ Q_1 = 0, \ Q_2 = 0.1, \ Q_3 = 0.5, \ Q_4 = 1.0, \ Q_5 = 10 $$

$$ X_j(t) = (X_j(t), X_j(t-1))' $$

Using this multiple state space model, we can obtain the sub-optimal system noise variance of trend model. The selected value by this method is the best as long as considered. Figure 5 shows the transition of each sub-model's posterior probability. The posterior probability of sub-model $M_2$ gradually increases. We know from this figure that the optimal system noise variance comes close to $Q_2$.

3.2 Monthly Trunk Group Load Forecasting

Monthly trunk group load data is shown in Fig.6. The total number of months is 48. In Sect.3.2.1 the multiple state space model is applied to this data and the forecasting ability is shown. In Sect.3.2.2 parameter estimation of a random walk model is shown using a multiple state space model.

<table>
<thead>
<tr>
<th>Month</th>
<th>1</th>
<th>9</th>
<th>17</th>
<th>25</th>
<th>33</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$</td>
<td>0.0</td>
<td>0.1</td>
<td>0.5</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Fig.6 Trunk group load data

3.2.1 Multiple Model with Trend Model and Random Walk Model

Let us forecast trunk group load. We apply the same multiple state space model as in Sect.3.1.1. System matrices, driving matrices and observation matrices of each sub-model are the same as in sect 3.1.1 except a system noise. In this forecasting, we use $Q_1 = 0.1, Q_2 = 0.1$ as the noise variances. The transition of the posterior probability of $M_1$ and $M_2$ is shown in Fig.7. It shows that Bayesian multiple model $M$ approaches random walk model $M_2$, as the number of observations increases.

Fig.7 Posterior probability of multiple model
Table 2 shows the one-step ahead forecasting accuracy.

<table>
<thead>
<tr>
<th></th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>0.068</td>
<td>0.059</td>
<td>0.057</td>
</tr>
<tr>
<td>MA</td>
<td>0.214</td>
<td>0.185</td>
<td>0.185</td>
</tr>
</tbody>
</table>

From these results we know that a random walk model has a good forecasting ability for changeable data compared with a trend model.

3.2.2 Multiple Model with Random Walk Models of Different Noise Variances

Let us consider a multiple state space model composed of 5 random walk models of different system noises. System matrices, driving matrices, observation matrices and noise variances of sub-models are as follows.

$$M_j : \text{the } j-\text{th random walk model } (j=1,2,\ldots,5)$$

$$F_j = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad G_j = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad H_j = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$R_j = 1.0, \quad \sigma^2_1 = 0, \quad \sigma^2_2 = 0.1, \quad \sigma^2_3 = 0.5, \quad \sigma^2_4 = 1.0, \quad \sigma^2_5 = 10$$

$$x_j(t) = (x_j(t), x_j(t-1))^T$$

Figure 8 shows the transition of each sub-model's posterior probability. The most suitable sub-model changed from $M_1$ to $M_2$ at the 31th month.

This change can be explained as follows. As you can see in Fig.6, a level is about 8.0 for the first half (before the 31th month), and about 6.5 for the latter half. Though the levels are different for both periods, we can regard that it remains constant within each period. This means that the system noise is very small for all period. On the other hand, it is known that the observation noise variance is proportional to trunk group load [10]. Thus, the observation noise variance is greater for the first half than for the latter half. This turns out that the system noise variance becomes greater relatively for the latter half, that is, the most suitable sub-model changes from $M_1$ to $M_2$. The level change might occur as a result of trunk group replacement.

4. CONCLUSIONS

Bayesian forecasting with multiple state space models has been proposed. Using the likelihoods of state space model, Bayesian posterior probabilities were obtained. The structure of the social phenomenon like a traffic trend is unclear and likely to change. To conquer them in model selecting, several state space models were prepared as candidates to represent the true behavior. Using this Bayesian forecasting method, the posterior probability of the best fitting model increases automatically, as the number of observations increases. In the numerical example, it is demonstrated that forecasting with multiple state space models is robust and more accurate than that with single state space models. This method is applicable to situations which include exceptional data when an insensitive model is added to the multiple model. A problem remaining to be studied is how to develop more versatile sub-models.

5. ACKNOWLEDGEMENT

The authors would like to express their appreciation to Mr. Kunio Kodaira, Chief of Teletraffic Section in the Musashino Electrical Communication Laboratory, NTT, and to Dr Jun Matsuda, Staff Engineer of the same section, for their many helpful comments and encouragement.

REFERENCES

