AN INTRODUCTION TO MULTIPROCESS CLASS II MIXTURE MODELS

I.R. Korsgaard and P. Løvendahl

Department of Animal Breeding and Genetics, Danish Institute of Agricultural Sciences, P.O. Box 50, DK 8830 Tjele, Denmark

INTRODUCTION

Somatic cell count is an accepted indicator of mastitis. However, measurements of cell count are subject to noise and outliers, which decrease their potential use in decision support. Statistical tools to separate noise from biologically relevant changes can help improving the interpretation of somatic cell count (SCC) data. The extension (Smith and West, 1983) of the multiprocess Kalmanfilter (Harrison and Stevens, 1976) to provide probabilities of different kinds of changes may be used in decision support - for example an action of treatment should be taken if the probability of an increase in SCC is above a critical level. The purpose of this presentation is to introduce dynamic linear model and in particular multiprocess class II mixture models with the recursive updating procedure for providing probabilities of different kinds of changes.

MODEL

Dynamic linear model. For the time series $\{y_t\}_{t=1,\ldots,n}$ consisting of n observations (of e.g. ln(somatic cell count)) a dynamic linear model (DLM) is described by an observation equation:

$$Y_t = F_t\theta_t + v_t$$

a system equation:

$$\theta_t = G_t\theta_{t-1} + w_t$$

and initial information:

$$\theta_0 \sim N(m_0, C_0)$$

where F_t is the observation matrix, θ_t is a latent vector (or scalar) and v_t, with $v_t \sim N(0, V_t)$, is the observation noise. The latent process $\{\theta_t\}_{t=1,\ldots,n}$ is given by the system equation (and the initial information) with evolution matrix (system matrix) G_t and evolution error w_t. It is assumed that $w_t \sim N(0, W_t)$, with $v_1,\ldots,v_n, w_1,\ldots,w_n$ mutually independent and independent of the initial information. The model specified by $\{F_t, G_t, V_t, W_t\}$ will be denoted M_t.

Example: The sire model given by $Y_t = s + e_t$, *for* $t = 1,..,n$; *with* $s \sim N(0, \sigma_s^2)$ *independent of* $e = (e_1,\ldots,e_n)' \sim N_n(0, I_n\sigma_e^2)$ *is equivalent to the DLM given by observation equation* $Y_t = s_t + e_t$, *system equation* $s_t = s_{t-1}$ *and initial information* $s = s_0 \sim N(0, \sigma_s^2)$. *Note that*

$F_t = G_t = 1$ *and* $V_t = \sigma_e^2$, *for* $t = 1,...,n$; $m_0 = 0$ *and* $C_0 = \sigma_s^2$, *and the model is without evolution error.*

Multiprocess class II mixture model. If the observations do not follow the same DLM for all values of t, it is useful to introduce mixture models, where, at each time t, we may choose between J different models. The Multiprocess class II mixture model is defined as follows: Let, for some integer $J > 1$, $\mathrm{A} = \{\alpha_1,...,\alpha_J\}$ denote the parameter space for α, and suppose, that at each time t, there exist an $\alpha \in \mathrm{A}$ so that $M_t(\alpha)$ holds. If the value, α_j, of α defining the model at time t, $M_t(\alpha_j)$, is selected with known probability, $\pi_t(j) = P\left(M_t(\alpha_j) \mid D_{t-1}\right)$, then the series $\{Y_t\}_{t=1,...,n}$ is said to follow a multiprocess class II mixture model. We will use $M_t(j)$ as short notation for $M_t(\alpha_j)$. Furthermore we let D_t denote the information available at time t, $t \geq 0$. Here we will assume that $D_t = D_{t-1} \cup \{Y_t\}$ for $t > 0$.

Multiprocess Kalman filter (extended). In the following we outline the recursive updating procedure for providing posterior probabilities $P(M_t(j) \mid D_t)$, of model j at time t, as well as one and two step back smoothed probabilities, $P(M_{t-1}(j) \mid D_t)$ and $P(M_{t-2}(j) \mid D_t)$, for the different models at different time points. The procedure is outlined for a model with $J = 4$, $G_t = G$ and $F_t = F$ for all t. The observation error as well as system error are assumed to depend on the model at time t, but are otherwise independent of time. Model j is assumed to be selected with probability $P\left(M_t(\alpha_j) \mid D_{t-1}\right) = \pi_0(j)$ independently of the past, D_{t-1}, $j = 1,...,4$ (fixed model selection probabilities). A priori it is assumed that $\theta_0 \sim N(m_0, C_0)$ and that all of the parameters are known. *For t=1:* From the system equation and the prior distribution of θ_0 we obtain $\theta_1 \mid M_1(j), D_0 \sim N\left(Gm_0, GC_0G' + W(j)\right)$ for $j = 1,...,4$. This, together with the observation equation, gives, conditional on $M_1(j)$, the forecast distribution of Y_1:

$$Y_1 \mid M_1(j) \sim N\left(FGm_0, F\left(GC_0G' + W(j)\right)F' + V(j)\right)$$

Next, the posterior probability of the different models at time 1 are calculated from

$$P\left(M_1(j) \mid D_1\right) \propto p\left(y_1 \mid M_1(j)\right) P\left(M_1(j) \mid D_0\right)$$

where $P\left(M_1(j) \mid D_0\right)$ by assumption is equal to $\pi_0(j)$. The posterior distribution of θ_1 is then given by a mixture of $\theta_1 \mid M_1(j), D_1 \sim N\left(m_1(j), C_1(j)\right)$, $j = 1,...,4$ with mixture probabilities $P\left(M_1(j) \mid D_1\right)$; *For time* $t > 1$*:* The steps in obtaining the (an approximate) posterior distribution of θ_t, as well as one (and two) step back smoothed probabilities of the different states/models at time $t-1$ ($t-2$ for $t > 2$) become more involved. Here we refer to Smith and West (1983) or West and Harrison (1997) for further details.

EXAMPLE

In order to illustrate the methodology, data $\{y_t\}_{t=1,...,50}$ were generated according to the linear growth model with exceptions, conditional on $M_t(j)$, given by observation equation: $Y_t = F_t\theta_t + v_t$, with $F_t = (1 \quad 0)$ and $\theta_t = (\mu_t, \beta_t)'$; system equation(s) $\mu_t = \mu_{t-1} + \beta_t + \varepsilon_{\mu t}$ and $\beta_t = \beta_{t-1} + \varepsilon_{\beta t}$ with $v_t \mid M_t(j) \sim N(0, V(j))$, $\varepsilon_{\mu t} \mid M_t(j) \sim N(0, E_\mu(j))$ and $\varepsilon_{\beta t} \mid M_t(j) \sim N(0, E_\beta(j))$, $j = 1,...,4$ and $t = 1,...,50$ assumed to be mutually independent.

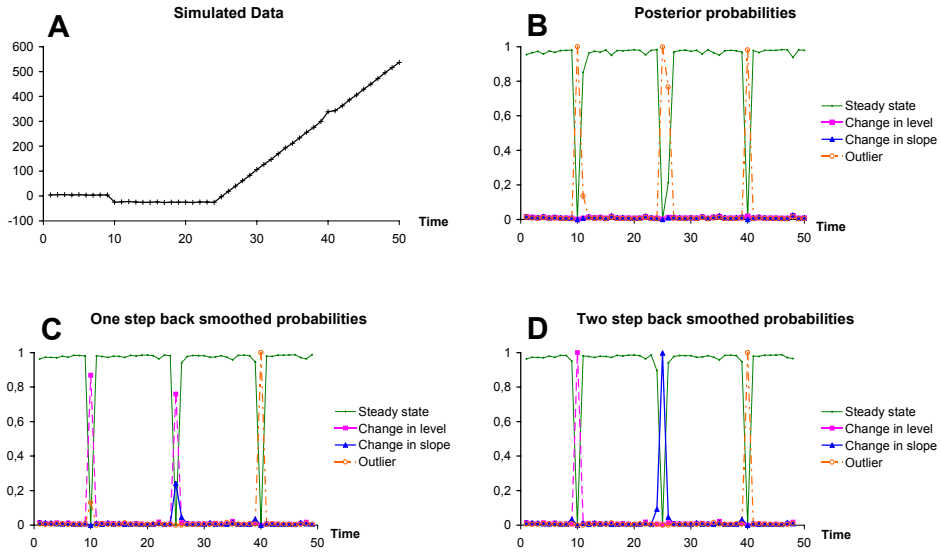

Figure 1. A Simulated data, B Posterior probabilities of the 4 different models at time $t = 1,...,50$, C and D One and two step back smoothed probabilities of the 4 different models at time $t = 1,...,49$ and time $t = 1,...,48$, respectively

$$(\theta_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\theta_{t-1} + w_t \text{ with } w_t \mid M_t(j) \sim N_2(0, W_t(j)); \; W_t(j) = \begin{pmatrix} E_\mu(j) + E_\beta(j) & E_\mu(j) \\ E_\mu(j) & E_\mu(j) \end{pmatrix})$$

Values of the different parameters are given in Table 1. θ_0 was arbitrary set to $(4 \quad 0)'$ and all of the data were simulated from Model 1 (steady state) except for a change in level at time 10, a change in slope at time 25 and an outlier at time 40. The parameters from Table 1 were used

in the analysis and the initial information was (arbitrary) assumed to be
$\theta_0 \sim N_2\left(\begin{pmatrix}4\\0\end{pmatrix},\begin{pmatrix}20 & 0\\0 & 10\end{pmatrix}\right)$.

Table 1. Parameters used for analysing simulated data

Model j	Name	$\pi_0(j)$	$V(j)$	$E_\mu(j)$	$E_\beta(j)$
1	steady state	0.94	1.0	0.0	0.0
2	chance in level	0.02	1.0	20.0	0.0
3	change in slope	0.02	1.0	0.0	10.0
4	outlier	0.02	50.0	0.0	0.0

The simulated data are shown in Figure 1.A. Posterior probabilities (Figure 1.B) of the outlier model are very high at times 10, 25 and 40. I.e. abrupt changes are detected - but not the true nature of the changes. The outlier is pointed out (wihtout false positive detections of outliers) from one step back smoothed probabilities (Figure 1.C). Finally from two step back smoothed probabilites (Figure 1.D) we obtain high probabilities of the models used in the simulation - at all timepoints.

CONCLUSION

We have briefly summarised (with minor modifications) part of the methodology for modelling and monitoring biological time series subject to outliers and changes in the underlying latent variables presented in Smith and West (1983). They used the method successfully to provide on-line probabilities of serious changes in kidney function in individual patients who had recently received transplants. The method is relevant in agriculture. We may for example, based on regular measurement of ln(somatic cell count), or other indicators of mastitis, provide probabilities of mastitis and hopefully be able to detect mastitis earlier and more reliable compared to other methods - because of the flexibility in these models. In applications of multiprocess II mixture models parameters have been found from empirical trials with the system (see e.g. Smith and West (1983) and Thysen (1992)). Methods for assessing the adequacy of mixture models are lacking. Increasing the dimension of the observations as well as the dimension of the state vector off course increases the complexity in finding suitable parameters. Another challenge is to incorporate information from relatives in mixture models and/or to integrate with breeding value estimation.

REFERENCES

Harrison P.J. and Stevens C.F. (1976) *J. Stat. Soc. Ser. B* **38** : 205-247.
Smith A.F.M. and West M. (1983) *Biometrics* **39** : 867-878.
Thysen I. (1993) *Acta Agric. Scand. Sect. A. Anim. Sci*. **43** : 58-64.
West M. and Harrison J. (1997) "Bayesian Forecasting and Dynamic Models". Springer-Verlag.