OUTLIERS, STRUCTURAL SHIFTS AND HEAVY-TAILED DISTRIBUTIONS IN STATE SPACE TIME SERIES MODELS

JUAN CARLOS ABRIL
Faculty of Economics,
National University of Tucumán and CONICET,
Casilla de Correo 209, 4000 Tucumán, Argentina

ABSTRACT

In this work a general method is developed for handling outliers, structural shifts and heavy-tailed distributions in linear state space time series models. The basic tool we use for dealing with outliers and structural shifts is to model observation or state error densities by a mixture of densities, one component of which is a Gaussian density with a large variance. The other component can be a Gaussian density, a non-Gaussian density such as Student's t or it can itself be a Gaussian mixture. The underlying idea is to estimate the state vector by its posterior mode using linearisation, iteration and the Kalman filter and smoother.

KEYWORDS: Approximate Kalman filtering and smoothing; Heavy-tailed distributions; Kalman filtering and smoothing; Outliers; Posterior mode estimate; Structural shifts.

MATHEMATICS SUBJECT CLASIFICATION (1991): Primary 62M10, Secondary 62E17 and 62M20.

1. INTRODUCTION

We begin with the linear Gaussian state space model. Although our main concern is with non-Gaussian models, the linear Gaussian model provides the basis from which all our methods will be developed. The model can be formulated in a variety of ways; we shall take the form

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim N(0, H_t), \\ \alpha_t &= T_t\alpha_{t-1} + R_t\eta_t, & \eta_t &\sim N(0, Q_t), \end{aligned} \tag{1}$$

for $t = 1,...,n,$ where y_t is a $p \times 1$ vector of observations, α_t is an unobserved $m \times 1$ state vector, R_t is a selection matrix composed of g columns of the identity matrix I_m, which need not be adjacent, and the variance matrices H_t and Q_t are nonsingular. The disturbance vectors ε_t and η_t are serially independent and independent of each other. Matrices H_t, Q_t, Z_t and T_t are assumed known apart for possible dependence on a parameter vector ψ which in classical inference is assumed fixed and unknown and in Bayesian inference is assumed to be random. The first line of (1) is called the observation equation and the second line, the state equation of the state space model. Matrices Z_t and T_t are permitted to depend on $y_1,...,y_{t-1}$. The initial state α_0 is assumed to be $N(a_0, P_0)$ independently of $\varepsilon_1,...,\varepsilon_n$ and $\eta_1,...,\eta_n$, where a_0 and P_0 are first assumed known; later, we consider how to proceed in the absence of knowledge of a_0 and P_0 and particularly in the diffuse case when $P_0^{-1} = 0$.

Let Y_{t-1} denote the set $y_1,...,y_{t-1}$ together with any information prior to time $t = 1$. Starting at $t = 1$ and building up the distributions of α_t and y_t recursively, it can be shown that the conditional densities $p(y_t \mid \alpha_1,...,\alpha_t, Y_{t-1}) = p(y_t \mid \alpha_t)$ and $p(\alpha_t \mid \alpha_1,...,\alpha_{t-1}, Y_{t-1}) = p(\alpha_t \mid \alpha_{t-1})$, thus establishing the truly Markovian nature of the model. Since in model (1) all distributions are Gaussian, conditional distributions are also Gaussian. Assume that α_t given Y_{t-1} is $N(a_t, P_t)$ and that α_t given Y_t is $N(a_{t|t}, P_{t|t})$. The object of the Kalman filter is to calculate $a_{t|t}, P_{t|t}, a_{t+1}$ and P_{t+1} given

a_t, P_t recursively. On the other hand, the Kalman smoother has the purpose of calculating recursively $\hat{\alpha}_t = E(\alpha_t | Y_n)$, $Var(\hat{\alpha}_t - \alpha_t)$ and $Cov(\hat{\alpha}_s - \alpha_s, \hat{\alpha}_t - \alpha_t)$ for $s < t$. Both procedures jointly are known as the Kalman filter and smoother (KFS).

In this work a general method is developed for handling outliers, structural shifts and heavy-tailed distributions in linear state space time series models. The basic tool we use for dealing with outliers and structural shifts is to model observation or state error densities by a mixture of densities, one component of which is a Gaussian density with a large variance. The other component can be a Gaussian density, a non-Gaussian density such as Student's *t* or it can itself be a Gaussian mixture. The underlying idea is to estimate the state vector by its posterior mode using linearisation, iteration and the Kalman filter and smoother.

Gaussian mixtures have been employed by many authors for filtering aspects of state space treatment of non-Gaussian data, notably Harrison and Stevens (1971, 1976), Sorenson and Alspach (1971), Alspach and Sorenson (1972), Guttman and Peña (1988, 1989) and Durbin and Cordero (1994). A comprehensive treatment of both filtering and smoothing using Gaussian mixtures has been given by Kitagawa (1994) but his techniques are onerous computationally. However, none of these authors but Durbin and Cordero estimated the posterior mode. In this work we present techniques for posterior mode estimation (PME) for these problems.

In section 2 we present a general method for handling outliers, structural shifts, heavy-tailed distributions and non-Gaussian observations in linear state space time series models. The method is based on the idea of estimating the state vector by its posterior mode. More details about it can be seen in Abril (2001).

We give in section 3 an introduction to the basic technique by considering outliers and level shifts for a simple special case, the local level (LL) model. In section 4 the treatment is extended to cover general Gaussian mixtures for the general linear state space model.

In many areas of applications, observed distributions tend to have heavier tails than the normal distribution. In section 5, posterior mode estimation of the state vector is considered for two different models for heavy-tailed distributions, the Gaussian mixture and Student's *t*. These distributions can be used without or with the addition of an extra Gaussian component to handle large structural shifts or large outliers. Theory for this option is given in section 6.

Approximate maximum likelihood estimation of hyperparameters is considered in section 7.

2. POSTERIOR MODE ESTIMATE

Let α be the stacked vector $(\alpha_1',\ldots,\alpha_n')'$ and let $p(\alpha \mid Y_n)$ be the conditional density of α given Y_n. The *posterior mode estimate* (PME) of α is defined to be the value $\hat{\alpha}$ of α that maximises $p(\alpha \mid Y_n)$. When the model is linear and Gaussian, $\hat{\alpha} = E(\alpha \mid Y_n)$. When the observations are non-Gaussian, however, $E(\alpha \mid Y_n)$ is generally difficult or impossible to compute and to use the mode instead is a natural alternative. More than that, it can be argued that in the non-Gaussian case the PME is preferable to $E(\alpha \mid Y_n)$ since it is the value of α which is the most probable given the data. In this respect it can be thought of analogous to the maximum likelihood (ML) estimate of a fixed parameter vector. The *t*th subvector of $\hat{\alpha}$ is called the smoothed value of α_t and is denoted by $\hat{\alpha}_t$.

Let A_t be the stacked vector $(\alpha_1',\ldots,\alpha_t')'$, $t = 1,\ldots,n$. Then $\hat{\alpha}$ is the PME of A_n and for filtering, the PME of A_t given Y_{t-1} is the value of A_t that maximises the conditional density of $p(A_t \mid Y_{t-1})$, $t = 1,\ldots,n$. The *t*th subvector of this is denoted by a_t.

In all cases we shall consider, the PME $\hat{\alpha}$ is the solution of the equations

$$\frac{\partial \log p(\alpha \mid Y_n)}{\partial\, \alpha_t} = 0, \qquad t = 1,\ldots,n.$$

Since, however, $\log p(\alpha \mid Y_n) = \log p(\alpha, Y_n) - \log p(Y_n)$, it is more easily obtained from the joint density as the solution of the equations

$$\frac{\partial \log p(\alpha, Y_n)}{\partial\, \alpha_t} = 0, \qquad t = 1,\ldots,n. \tag{2}$$

Similarly, at the filtering stage the PME of A_t given Y_{t-1} is the solution of the equations

$$\frac{\partial \log p(A_t \mid Y_{t-1})}{\partial \alpha_s} = 0, \quad s = 1, \ldots, t. \tag{3}$$

3. OUTLIERS AND LEVEL SHIFTS FOR THE LL MODEL

In this section we consider filtering and smoothing for the LL model in the presence of outliers and level shifts. We begin by considering outliers, that is, observations which differ from their immediately preceding and immediately succeeding observations by amounts which are so large that the observations are regarded as anomalous. Let us write the LL model in the form

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t , \\ \alpha_t &= \alpha_{t-1} + \eta_t , \quad t = 1, \ldots, n. \end{aligned} \tag{4}$$

Our first objective is to construct a model for the ε_t 's which allows for the presence of outliers in the data, so that when PME of the state vector α_t are obtained, the effect of the outliers are effectively smoothed out. To achieve this objective we assume that the ε_t has the Gaussian mixture density

$$h(\varepsilon_t) = (1 - \beta)\, h_1(\varepsilon_t) + \beta\, h_2(\varepsilon_t), \tag{5}$$

where $h_1(\varepsilon_t) = N(0, \sigma^2)$ and $h_2(\varepsilon_t) = N(0, \lambda^2 \sigma^2)$, and where β is small, say 0.01 and λ is large, say 10. Here, β can be thought of as the prior probability of an outlier and λ can be regarded as indicating the magnitude of the deviation of an outlier from its mean relative to that of an ordinary observation. Thus the density (5) can be regarded as a genuine model for observations containing outliers and not just a device for detecting them and eliminating its effects. However, experience shows that the results obtained are relatively insensitive to the values of β and λ within reasonable limits. In fact it is not normally worth while treating β and λ as parameters to be estimated from the data since the number of anomalous observations available for their estimation is, by definition, small. It is usually preferable to assign prior values such as 0.01 and 10, or if there is doubt, to proceed by trial and error. In the work of Durbin and Cordero (1994) the values 0.01 and 10 worked very well.

Let us leave aside for the moment the initialisation question by assuming that α_0 is fixed and known. The initialisation problem considers how to start the filter at the beginning of the series when nothing is known about the distribution of α_0, and the object of the filtering is to update our knowledge of the system each time a new observation y_t is brought in. Taking η_t to be $N(0,\ \sigma_\eta^2)$ and $\alpha = (\alpha_1, \ldots, \alpha_n)'$ as before, the log joint density of α and Y_n is, ignoring constants,

$$\log p(\alpha, Y_n) = -\frac{1}{2\sigma_\eta^2}\sum_{t=1}^{n}(\alpha_t - \alpha_{t-1})^2 + \sum_{t=1}^{n}\log h(y_t - \alpha_t).$$

Thus the PME of α is the solution of the equations

$$\begin{aligned}\frac{\partial \log p}{\partial \alpha_t} &= \frac{1}{\sigma_\eta^2}\left(\alpha_{t-1} - 2\alpha_t + \alpha_{t+1}\right)\\ &\quad + \frac{1}{h(\varepsilon_t)}\left[\frac{1-\beta}{\sigma^2}h_1(\varepsilon_t) + \frac{\beta}{\lambda^2\sigma^2}h_2(\varepsilon_t)\right](y_t - \alpha_t) = 0,\end{aligned} \tag{6}$$

for $t = 1, \ldots, n-1$ with $\varepsilon_t = y_t - \alpha_t$ and with the first term replaced by $\sigma_\eta^{-2}(\alpha_{n-1} - \alpha_n)$ for $t = n$.

We solve these equations by linearising them, putting the linearised equations in the same form as the equations for the analogous Gaussian model and then using the KFS. We use a general linearisation technique from the Taylor expansion method which is not a specific one to the particular model (5).

We note first that the analogous Gaussian model is (4) with $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ so the equations analogous to (6) for the Gaussian case are

$$\frac{\partial \log p}{\partial \alpha_t} = \frac{1}{\sigma_\eta^2}\left(\alpha_{t-1} - 2\alpha_t + \alpha_{t+1}\right) + \frac{1}{\sigma_\varepsilon^2}\left(y_t - \alpha_t\right) = 0, \tag{7}$$

with an analogous modification when $t = n$. Suppose that $\tilde{\alpha}_t$ is a trial value of α_t. Putting $\tilde{\varepsilon}_t = y_t - \tilde{\alpha}_t$ and comparing (6) and (7) suggests taking

$$\frac{1}{\tilde{\sigma}_{\varepsilon}^{2}}=\frac{1}{h(\tilde{\varepsilon}_{t})}\left[\frac{1-\beta}{\sigma^{2}}h_{1}(\tilde{\varepsilon}_{t})+\frac{\beta}{\lambda^{2}\,\sigma^{2}}h_{2}(\tilde{\varepsilon}_{t})\right]. \tag{8}$$

Replacing the corresponding term in ε_t in (6) by this gives a linear set of equations with the same form as (7) and which can therefore be solved by the KFS. Taking the solution as a new trial value of α, the process is repeated until suitable convergence is achieved.

At this point a word of caution is appropriate. There is evidence that with models of this kind it is possible for densities to be multimodal. In consequence, care is needed in choosing the initial trial value $\tilde{\alpha}$ of α. It is inappropriate to start with completely arbitrary values such as $\tilde{\alpha}_t = 0$ for all *t* as is sometimes possible when maximising a well-behaved function with a single maximum. The two-filter smoother considered by Abril (2001) for exponential family observations seems to be very suitable for the purpose.

Let us now consider why this technique is effective in neutralising the effect on the solution of equations (6) of a large outlier at time *t*. If $\beta = 0$, so the model is Gaussian, then the contribution of the second term of (6) is $\sigma^{-2}(y_t - \alpha_t)$ which, since $|y_t - \alpha_t|$ is large, will have a distorting influence on the estimation of α_τ for τ near *t*. If, however, we are using the mixture model (5), then the contribution of $h_1(\varepsilon_t)$ is approximately zero relative to that of $h_2(\varepsilon_t)$ so that the second term of (6) reduces approximately to $\lambda^{-2}\,\sigma^{-2}(y_t - \alpha_t)$ which is much smaller; indeed, with $\lambda = 10$ the contribution of the outlier is reduced to about 1% of its Gaussian value.

Now suppose that there are no outliers but that there are abrupt changes of α_t; we call these *level shifts*. These occur when $|\eta_t|$ is large. In order to deal with them we use a similar mixture model to the one used for outliers, taking for the density of η_t,

$$q(\eta_t) = (1-\delta)\,q_1(\eta_t) + \delta\,q_2(\eta_t),$$

where $q_1(\eta_t) = N(0, \sigma_\eta^2)$, $q_2(\varepsilon_t) = N(0, \kappa^2\,\sigma_\eta^2)$, δ is small and κ is large. Assume that $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. The log density of α and Y_n is now

$$\log p(\alpha, Y_n) = \sum_{t=1}^{n} \log q(\alpha_t - \alpha_{t-1}) - \frac{1}{2\sigma_\varepsilon^2} \sum_{t=1}^{n} (y_t - \alpha_t)^2 .$$

On differentiating and equating to zero we obtain the PME $\hat{\alpha}$ as the solution of the equations

$$\begin{aligned} \frac{\partial \log p}{\partial \alpha_t} = & - \frac{1}{q(\eta_t)} \left[\frac{1-\delta}{\sigma_\eta^2} q_1(\eta_t) + \frac{\delta}{\kappa^2 \sigma_\eta^2} q_2(\eta_t) \right] (\alpha_t - \alpha_{t-1}) \\ & + \frac{1}{q(\eta_{t+1})} \left[\frac{1-\delta}{\sigma_\eta^2} q_1(\eta_{t+1}) + \frac{\delta}{\kappa^2 \sigma_\eta^2} q_2(\eta_{t+1}) \right] (\alpha_{t+1} - \alpha_t) \\ & + \frac{1}{\sigma_\varepsilon^2} (y_t - \alpha_t) = 0, \qquad t = 1, \ldots, n-1, \end{aligned} \tag{9}$$

with $\eta_t = \alpha_t - \alpha_{t-1}$. There is a slightly different form for $t = n$ which we need not specify.

If a level shift occurs at time *t*, then as in the outlier case, $q_1(\eta_t)$ is nearly zero so $q(\eta_t) \approx q_2(\eta_t)$; if no level shift occurs then $q(\eta_t) \approx q_1(\eta_t)$. In order to understand how the model handles large level shifts, suppose that these approximate equalities are equalities and suppose also that a large level shift occurs at time *t* and nowhere else. The three equations involving α_t are then, after multiplying by σ_η^2 and eliminating small quantities,

$$\begin{aligned} -(1-\delta)(\alpha_{t-1} - \alpha_{t-2}) + \frac{\delta}{\kappa^2}(\alpha_t - \alpha_{t-1}) + \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}(y_{t-1} - \alpha_{t-1}) = 0, \\ \frac{\delta}{\kappa^2}(\alpha_t - \alpha_{t-1}) + (1-\delta)(\alpha_{t+1} - \alpha_t) + \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}(y_t - \alpha_t) = 0, \\ -(1-\delta)(\alpha_{t+1} - \alpha_t) + (1-\delta)(\alpha_{t+2} - \alpha_{t+1}) + \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}(y_{t+1} - \alpha_{t+1}) = 0. \end{aligned} \tag{10}$$

We see that when $|\alpha_t - \alpha_{t-1}|$ is O(κ) then $\kappa^{-2} |\alpha_t - \alpha_{t-1}|$ is O(κ^{-1}) which is small, so an anomalously large value of $|\alpha_t - \alpha_{t-1}|$ has relatively little distorting effect on the solution of the equations. In fact, equation (9) effectively break up into two sets, one involving $\alpha_1, \ldots, \alpha_{t-1}$ only and the other involving $\alpha_t, \ldots, \alpha_n$ only, and the solution is relatively unaffected by the large value of $|\alpha_t - \alpha_{t-1}|$. On the other hand, the equations we would get if we used the Gaussian model would have in place of (10) the same set but with $\delta = 0$ and κ equal to unity. It can be seen that the effect of a large value of $|\alpha_t - \alpha_{t-1}|$ will spill over into neighbouring equations and thus distort the estimates of α_τ when τ is near to *t*.

Another way of understanding why the mixture model is better than the Gaussian model in dealing with structural shifts is to recognise that it is a more realistic model when shifts are present and can therefore be expected to fit the data better. When both outliers and structural shifts are present, mixtures models can be used simultaneously for both observation and state errors and they can be dealt with straightforwardly by the linearisation and KFS techniques used above.

4. GAUSSIAN MIXTURES FOR MODELLING THE ERRORS IN THE GENERAL LINEAR STATE SPACE MODEL

In this section we extend the results of the previous section by developing methods for dealing with outliers and structural shifts for the general state space model

$$\begin{aligned} y_t &= Z_t \alpha_t + \varepsilon_t, \\ \alpha_t &= T_t \alpha_{t-1} + R_t \eta_t, \quad t = 1, \ldots, n, \end{aligned} \tag{11}$$

by means of Gaussian mixtures. However, the modelling of the distributions of ε_t and η_t by Gaussian mixtures has much wider application than to outliers and structural shifts alone. These mixtures can in fact be used to model arbitrary types of departure from normality. For this reason we shall present a general theory here, giving the application to outliers and structural shifts as a special case. We assume that R_t is an $m \times g$ selection matrix with $g \leq m$ so $(R_t\, Q_t\, R_t')^{-1} = R_t\, Q_t^{-1}\, R_t'$ where Q_t is positive definite. For simplicity we assume to begin with that α_0 is fixed and known, relaxing this assumption later.

We consider the general case where ε_t has the mixture density

$$h_t(\varepsilon_t) = \sum_{i=1}^{k} \beta_i \, h_{it}(\varepsilon_t), \quad \beta_i \geq 0, \quad \sum_{i=1}^{k} \beta_i = 1 \tag{12}$$

and η_t has the mixture density

$$q_t(\eta_t) = \sum_{i=1}^{l} \kappa_i \, q_{it}(\eta_t), \quad \kappa_i \geq 0, \quad \sum_{i=1}^{l} \kappa_i = 1, \tag{13}$$

where $h_{it}(\varepsilon_t) = N(0, H_{it})$ and $q_{it}(\eta_t) = N(0, Q_{it})$. The log joint density of α and Y_n is

$$\log p(\alpha, Y_n) = \sum_{t=1}^{n} \log q_t(\eta_t) + \sum_{t=1}^{n} \log h_t(\varepsilon_t),$$

with $\eta_t = R_t'(\alpha_t - T_t\alpha_{t-1})$ and $\varepsilon_t = y_t - Z_t\alpha_t$. To obtain the PME $\hat{\alpha}$ of α we differentiate with respect to α_t and equate to zero giving

$$\begin{aligned} -R_t\overline{Q}_t^{-1}R_t'(\alpha_t - T_t\alpha_{t-1}) + T_{t+1}'R_{t+1}\overline{Q}_{t+1}^{-1}R_{t+1}'(\alpha_{t+1} - T_{t+1}\alpha_t) \\ + Z_t'\overline{H}_t^{-1}(y_t - Z_t\alpha_t) = 0, \end{aligned} \tag{14}$$

for $t = 1, \ldots, n-1$, together with

$$-R_n\overline{Q}_n^{-1}R_n'(\alpha_n - T_n\alpha_{n-1}) + Z_n'\overline{H}_n^{-1}(y_n - Z_n\alpha_n) = 0,$$

where

$$\overline{Q}_t^{-1} = \frac{1}{q_t(\eta_t)} \sum_{i=1}^{l} \kappa_i q_{it}(\eta_t) Q_{it}^{-1} \tag{15}$$

and

$$\overline{H}_t^{-1} = \frac{1}{h_t(\varepsilon_t)} \sum_{i=1}^{k} \beta_i h_{it}(\varepsilon_t) H_{it}^{-1}, \tag{16}$$

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

We solve equations (14) by linearising and iterating as before, starting with a trial value $\tilde{\alpha}$ of α. Let

$$\widetilde{Q}_t^{-1} = \frac{1}{q_t(\tilde{\eta}_t)} \sum_{i=1}^{l} \kappa_i q_{it}(\tilde{\eta}_t) Q_{it}^{-1}$$

and

$$\widetilde{H}_t^{-1} = \frac{1}{h_t(\tilde{\varepsilon}_t)} \sum_{i=1}^{k} \beta_i h_{it}(\tilde{\varepsilon}_t) H_{it}^{-1},$$

where $\tilde{\eta}_t = R_t'(\tilde{\alpha}_t - T_t\tilde{\alpha}_{t-1})$ and $\tilde{\varepsilon}_t = y_t - Z_t\tilde{\alpha}_t$ with $\left(\tilde{\alpha}_1', \dots, \tilde{\alpha}_n'\right)' = \tilde{\alpha}$. Substituting $\tilde{Q}_t^{-1}$ for $\overline{Q}_t^{-1}$ and $\tilde{H}_t^{-1}$ for $\overline{H}_t^{-1}$ in (14) gives the linear equations

$$\begin{aligned} -R_t\tilde{Q}_t^{-1}R_t'(\alpha_t - T_t\alpha_{t-1}) + T_{t+1}'R_{t+1}\tilde{Q}_{t+1}^{-1}R_{t+1}'(\alpha_{t+1} - T_{t+1}\alpha_t) \\ + Z_t'\tilde{H}_t^{-1}(y_t - Z_t\alpha_t) = 0, \end{aligned} \tag{17}$$

for $t = 1, \dots, n-1$, together with

$$-R_n\tilde{Q}_n^{-1}R_n'(\alpha_n - T_n\alpha_{n-1}) + Z_n'\tilde{H}_n^{-1}(y_n - Z_n\alpha_n) = 0.$$

Comparing these equations with the corresponding equations for the analogous Gaussian version of (11) for when $\varepsilon_t \sim N(0, H_t)$ and $\eta_t \sim N(0, Q_t)$ we see that the two sets have the same form. Equation (17) can therefore be solved by the KFS to obtain a new trial value and the process is repeated until convergence. The assumption that α_0 is fixed and known is dropped at the filtering stage, using the initialisation procedures considered by Abril (1999).

5. HEAVY-TAILED DISTRIBUTIONS

In many areas of applications of time series analysis, particularly with economic data, observed distributions tend to have heavier tails than those of the normal distribution, even when no outliers or structural shifts are present. In this section we shall consider the fitting of linear state space models to data of this kind, initially in the absence of outliers and structural shifts. We shall employ two models, the Gaussian mixture and Student's *t*.

We begin by considering univariate series where the observational error ε_t has a heavy-tailed distribution and the state error η_t is Gaussian. The first form of heavy-tailed distribution we take has density

$$h(\varepsilon_t) = (1-\beta)\,h_1(\varepsilon_t) + \beta\,h_2(\varepsilon_t), \quad 0 < \beta < 1,$$

where $h_1(\varepsilon_t) = N(0, \sigma^2)$ and $h_2(\varepsilon_t) = N(0, \lambda^2 \sigma^2)$, $\lambda > 1$. The variance of ε_t is $\sigma_\varepsilon^2 = \left[1 + \beta(\lambda^2 - 1)\right]\sigma^2$. This is a three-parameter model, as compare with the Gaussian model which is a one-parameter model. If sufficient data are available to estimate all three parameters, it is the set $\sigma_\varepsilon^2, \beta, \lambda^2$ that should be estimated rather than $\sigma^2, \beta, \lambda^2$ since σ_ε^2 is relatively independent of β and λ^2. If not enough data are available to estimate both β and λ^2 as well as σ_ε^2, then one of them could be fixed *a priori* and the other estimated; for example, we could pre-assign λ arbitrarily, say in the range 2 to 4 and then estimate σ_ε^2 and β by the approximate maximum likelihood (ML) methods to be discussed later. Several values of λ could be tried and the value chosen which gives the highest likelihood. Similar considerations apply to state densities. Since the models are Gaussian mixtures, they can be handled by the theory given in section 4 above.

The second heavy-tailed distribution we consider is Student's *t* with ν degrees of freedom, the density of which we write as

$$h(\varepsilon_t) = \frac{c(\nu)}{\sigma_\varepsilon^2} \frac{1}{\left[1 + \dfrac{\varepsilon_t^2}{(\nu - 2)\sigma_\varepsilon^2}\right]^{\frac{\nu+1}{2}}}, \quad \nu > 2, \tag{18}$$

where $c(\nu)$ is a constant. We write this in this form so that $Var(\varepsilon_t) = \sigma_\varepsilon^2$ for all ν, thus keeping the estimation of σ_ε^2 and ν relatively independent in the estimation process. The contribution of $\partial \log h(y_t - Z_t\alpha_t)/\partial\alpha_t$ to $\partial \log p(\alpha, Y_n)/\partial\alpha_t$ is

$$\frac{\nu+1}{(\nu-2)\sigma_\varepsilon^2} \frac{Z_t'(y_t - Z_t\alpha_t)}{\left[1 + \dfrac{(y_t - Z_t\alpha_t)^2}{(\nu-2)\sigma_\varepsilon^2}\right]} = \frac{(\nu+1) Z_t'(y_t - Z_t\alpha_t)}{\left[(\nu-2)\sigma_\varepsilon^2 + (y_t - Z_t\alpha_t)^2\right]}. \tag{19}$$

Suppose that in the iterative estimation of the PME $\hat{\alpha}$ of α, we have a trial value $\tilde{\alpha}_t$ of α_t. We linearise (19) by putting

$$\frac{1}{\tilde{\sigma}_\varepsilon^2} = \frac{(\nu+1)}{(\nu-2)\sigma_\varepsilon^2 + (y_t - Z_t\alpha_t)^2}. \tag{20}$$

Substituting in (19) we have

$$\frac{\partial \log h(y_t - Z_t\alpha_t)}{\partial \alpha_t} = \frac{1}{\tilde{\sigma}_\varepsilon^2} Z_t'(y_t - Z_t\alpha_t),$$

which has the same form as if $h(\varepsilon_t)$ were $N(0, \tilde{\sigma}_\varepsilon^2)$. We can therefore use the KFS as in the last section to solve the linearised equation $\partial \log p(\alpha, Y_n)/\partial \alpha_t = 0$ and hence obtained a better trail value of α. Analogous considerations apply when the *t* distribution is used to model independent components of the state error η_t. A similar approach could be employed based on the multivariate *t* density when the components of η_t cannot be treated as independent. The value of ν in (18) can be regarded as an unknown parameter to be estimated in the ML process or it can be assigned arbitrarily on the basis of experience or trial and error.

Throughout this section we have been concerned with distributions which are non-Gaussian but are symmetric and homogeneous through time. We did so because, apart from exponential family distributions, these are the most important for time series analysis. However, the techniques we have used do not in fact require symmetry. Consider for simplicity the case where y_t is univariate and where the distribution of the observation error ε_t has density $h(\varepsilon_t)$ which possesses derivative $\dot{h}_t(\varepsilon_t) = d\, h_t(\varepsilon_t)/d\varepsilon_t$. Then the contribution of $\partial \log h(y_t - Z_t\alpha_t)/\partial \alpha_t$ to the left hand side of the equation $\partial \log p(\alpha, Y_n)/\partial \alpha_t = 0$ is

$$\dot{h}_t(y_t - Z_t\alpha_t)Z_t' = \frac{\dot{h}(y_t - Z_t\alpha_t)}{y_t - Z_t\alpha_t} Z_t'(y_t - Z_t\alpha_t).$$

Given a trial value $\tilde{\alpha}_t$ of α_t, this can be linearised and put in Gaussian form by taking

$$\tilde{\sigma}_\varepsilon^2 = \frac{y_t - Z_t\tilde{\alpha}_t}{\dot{h}_t(y_t - Z_t\tilde{\alpha}_t)}$$

and hence treated by the KFS, exactly as for Student's *t*, irrespective of whether the distribution is symmetric. However, in order to avoid singularity it may be necessary to replace $y_t - Z_t\tilde{\alpha}_t$ by an arbitrary small positive number ε where $|y_t - Z_t\tilde{\alpha}_t| < \varepsilon$. Similar considerations apply to state error densities.

6. OUTLIERS AND STRUCTURAL SHIFTS WHEN THE STATE OR OBSERVATION ERRORS ARE NON-GAUSSIAN

Suppose that we have decided to treat state or observation errors as heavy-tailed or otherwise non-Gaussian but that we also wish to allow for exceptionally large outliers or structural shifts. The question is how this cases can be treated. If the heavy-tailed distributions have been modelled by Gaussian mixtures there is no difficulty in principle. Taking the case of univariate observations for which the basic model chosen for the observation errors is $(1-\beta_1)N(0,\sigma^2)+\beta_1 N(0,\lambda_1^2\,\sigma^2)$, this is augmented by adding a $N(0,\lambda_2^2\,\sigma^2)$ component giving the overall density

$$(1-\beta_2)\left[(1-\beta_1)N(0,\sigma^2)+\beta_1 N(0,\lambda_1^2\,\sigma^2)\right]+\beta_2 N(0,\lambda_2^2\,\sigma^2),$$

where β_2 would normally be pre-assigned at say 0.01 and λ_2 would normally be pre-assigned at say $\lambda_2=10$. The parameters β_1 and λ_1^2 may be pre-assigned or estimated from data. The theory of section 4 may then be applied in routine fashion.

Other non-Gaussian densities can be handled in the following way. Suppose for example that the density of ε_t is chosen to be

$$h(\varepsilon_t)=(1-\beta)\,h_1(\varepsilon_t)+\beta\,h_2(\varepsilon_t),$$

where $h_1(\varepsilon_t)$ is the Student density (18), $h_2(\varepsilon_t)$ is $N(0,\lambda^2\,\sigma_\varepsilon^2)$ and β and λ^2 are small and large pre-assigned values respectively. Here, $h_1(\varepsilon_t)$ is intended to represent the bulk of the distribution of the ε_t's and $h_2(\varepsilon_t)$ is intended solely to deal with a small number of very large outliers. The contribution of $\partial\log h(y_t-Z_t\alpha_t)/\partial\alpha_t$ to $\partial\log p(\alpha\,,Y_n)/\partial\alpha_t$ is

$$\frac{Z_t'}{h(\varepsilon_t)}\left\{\frac{(1-\beta)(\nu+1)\,c(\nu)}{(\nu-2)\,\sigma_\varepsilon^2\left[1+\dfrac{\varepsilon_t^2}{(\nu-2)\,\sigma_\varepsilon^2}\right]^{\frac{\nu+3}{2}}}+\frac{\beta\,h_2(\varepsilon_t)}{\lambda^2\,\sigma_\varepsilon^2}\right\}(y_t-Z_t\alpha_t)$$

$$= \frac{Z_t'}{h(\varepsilon_t)} \left[\frac{(1-\beta) h_1(\varepsilon_t)}{\tilde{\sigma}_\varepsilon^2(\varepsilon_t)} + \frac{\beta h_2(\varepsilon_t)}{\lambda^2 \sigma_\varepsilon^2} \right] (y_t - Z_t \alpha_t), \tag{21}$$

where $\varepsilon_t = y_t - Z_t \alpha_t$ and, analogously to (20)

$$\frac{1}{\tilde{\sigma}_\varepsilon^2(\varepsilon_t)} = \frac{\nu + 1}{(\nu - 2)\sigma_\varepsilon^2 + \varepsilon_t^2}.$$

Now put $\tilde{\varepsilon}_t = y_t - Z_t \tilde{\alpha}_t$ where $\tilde{\alpha}_t$ is a trial value of α_t and let

$$\frac{1}{\bar{\sigma}_\varepsilon^2} = \frac{1}{h(\tilde{\varepsilon}_t)} \left[\frac{(1-\beta) h_1(\tilde{\varepsilon}_t)}{\tilde{\sigma}_\varepsilon^2(\tilde{\varepsilon}_t)} + \frac{\beta h_2(\tilde{\varepsilon}_t)}{\lambda^2 \sigma_\varepsilon^2} \right] \tag{22}$$

as in (8). Then (21) becomes

$$\frac{1}{\bar{\sigma}_\varepsilon^2} Z_t'(y_t - Z_t \alpha_t),$$

which is the standard Gaussian form so the KFS can be used to obtain a better trial value of α as before. We note the similarity between (22) and (8).

Although this section has been included for the sake of completeness, it should be recognised that if heavy-tailed densities have been adopted for the state and observation errors, these may in many cases provide adequate handling of structural shifts and outliers where these are not too extreme; certainly their performance in this respect can be expected to be better than that of the Gaussian density.

7. APPROXIMATE MAXIMUM LIKELIHOOD ESTIMATION OF HYPERPARAMETERS

As in Abril (2001), exact ML estimation of the hyperparameter vector ψ is not feasible so we employ approximate methods. Our first approximate form of the likelihood for this case as for the exponential family case is given by formula (10) of Abril (2001) where $v_t = y_t - Z_t \hat{\alpha}_t$ and $F_t = Var(v_t)$ are the values produced in the final pass of the Kalman filter after the final smoothed value $\hat{\alpha}$ of α has been computed. This is maximised numerically as proposed by Abril (1999, 2001).

Analogous to (16) of Abril (2001), our second approximate form is

$$L = (2\pi)^{nm/2} |V|^{1/2} p(\hat{\alpha}, Y_n | \psi), \tag{23}$$

where, using (14) of Abril (2001),

$$|V|^{1/2} = \prod_{t=1}^{n} |\overline{Q}_t|^{1/2} |\overline{H}_t|^{1/2} |F_t|^{-1/2}$$

and where $p(\alpha, Y_n | \psi)$ is the joint density of α and Y_n given ψ with $\overline{Q}_t$ and $\overline{H}_t$ given by (15) and (16) above.

Our third approximate form applies only to Gaussian mixtures and arises from the fact that in principle it is possible to update a Gaussian mixture exactly as each new observation comes in and in this way construct an exact likelihood. The problem is that the number of components increases exponentially with time and therefore rapidly becomes unmanageable. For example, consider the simple case in which the state error density is a single Gaussian and the observation error density is a mixture of two Gaussian components. Then $p(y_1)$ has two components, $p(y_2 | y_1)$ has four components, $p(y_3 | Y_2)$ has eight components, and so on until $p(y_n | Y_{n-1})$ has 2^n components. The situation would be worse with a mixture for the state error density and with more components in the observation error density. It follows that calculation of the exact likelihood by exact updating is not feasible. However, an approximation to this approach can be obtained in the following way. Suppressing dependence on ψ and assuming that $p(\alpha_{t-1} | Y_{t-1})$ is a single Gaussian, obtain $p(\alpha_t | Y_t)$ as a mixture using standard Kalman filtering formulae and collapse it into a single Gaussian by replacing it by the single Gaussian with the same mean vector and variance matrix. From this, obtain $p(y_{t+1} | Y_t)$ as a mixture and take the product for t = 0 to $n-1$ as the approximate likelihood. This approach has the advantage that only filtering is required, so state iteration to obtain a smoothed value of α is not needed to compute the approximate likelihood for a particular value of ψ. Numerical maximisation with respect to ψ can then be carried out in a routine way. Because this method is so economical computationally, it should either be used as the sole method or to provide a starting value for one of the other two methods. Details of the collapsing process are as follows.

Write the observational and state mixture as

$$p(y_t|\alpha_t) = \sum_i \beta_i p_i(y_t|\alpha_t),$$
$$p(\alpha_t|\alpha_{t-1}) = \sum_j \delta_j p_j(\alpha_t|\alpha_{t-1}). \tag{24}$$

Let $p_{ij}(\alpha_t|Y_t)$ and $p_{ij}(y_t|Y_{t-1})$ be the densities given by the standard Kalman filtering for $p(\alpha_t|Y_t)$ and $p(y_t|Y_{t-1})$ assuming that $p(y_t|\alpha_t) = p_i(y_t|\alpha_t)$, $p(\alpha_t|\alpha_{t-1}) = p_j(\alpha_t|\alpha_{t-1})$ and that $p(\alpha_{t-1}|Y_{t-1})$ is a single Gaussian. Then Durbin and Cordero (1994) show that

$$p(\alpha_t|Y_t) = \sum_{i,\,j} \beta_i \delta_j \rho_{ijt} p_{ij}(\alpha_t|Y_t), \tag{25}$$

where

$$\rho_{ijt} = \frac{p_{ij}(y_t|Y_{t-1})}{\sum_{i,\,j} \beta_i \delta_j p_{ij}(y_t|Y_{t-1})}.$$

To collapse the mixture (25) into a single Gaussian, suppose that $p_{ij}(\alpha_t|Y_t)$ is $N(\mu_{ijt}, V_{ijt})$ where μ_{ijt} and V_{ijt} are given by the Kalman filter and let $\mu_t = E(\alpha_t|Y_t)$ and $V_t = Var(\alpha_t|Y_t)$. Compute

$$\mu_t = \sum_{i,\,j} \beta_i \delta_j \rho_{ijt} \mu_{ijt},$$

$$V_t = \sum_{i,\,j} \beta_i \delta_j \rho_{ijt} [V_{ijt} + (\mu_{ijt} - \mu_t)(\mu_{ijt} - \mu_t)']$$

and take $p(\alpha_t|Y_t)$ to be $N(\mu_t, V_t)$. This completes the collapsing ready for the next updating step. To compute the approximate likelihood we also need

$$p(y_t|Y_{t-1}) = \sum_{i,\,j} \beta_i \delta_j p_{ij}(y_t|Y_{t-1}).$$

The third approximate form for the likelihood is then

$$L = \prod_{t=1}^{n} p(y_t|Y_{t-1}),$$

with appropriate initialisation as before.

8. CONCLUSIONS

This paper presents a methodology for the treatment outliers, structural shifts, heavy-tailed distributions and non-Gaussian time series observations, particularly when they come from linear state space time series models. The methodology is developed that can be used by applied researchers for dealing with real time series data without them having to be time series specialists. The idea underlying the techniques is to put everything in state space form and then, linearised it to obtain an approximation to the Gaussian case in order to apply the KFS, estimating the state vector by its posterior mode. The PME is clearly a reasonable estimate because it maximises the corresponding density and is analogous to the maximum likelihood estimate.

REFERENCES

ABRIL, JUAN CARLOS. (1999). *Análisis de Series de Tiempo Basado en Modelos de Espacio de Estado* (In Spanish). EUDEBA: Buenos Aires.

ABRIL, JUAN CARLOS. (2001). On time series of observations from exponential family distributions. *Pak. J. Statist.*, **17(3)**, 235-48.

ALSPACH, D. L. AND H. W. SORENSON. (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Trans. Automatic Control*, **AC-17**, 439-48.

DURBIN, J. AND M. CORDERO. (1994). Handling structural shifts, outliers and heavy-tailed distributions in state space time series models. London School of Economics Statistics Research Report.

GUTTMAN, I. AND D. PEÑA. (1988). Bayesian approach to robustifying the Kalman filter. In *Bayesian Analysis of Time Series and Dynamic Models* (J. Spall, Ed.). Marcel Dekker: New York.

GUTTMAN, I. AND D. PEÑA. (1989). Optimal collapsing of mixture distributions in robust estimation. *Communications in Statist. (Theory and Methods)*, **18**, 817-34.

HARRISON, P.J. AND C. F. STEVENS. (1971). A Bayesian approach to short-term forecasting. *Operational Research Quarterly*. **22**, 341-62.

HARRISON, P.J. AND C. F. STEVENS. (1976). Bayesian forecasting. *J. R. Statist. Soc.*, **B, 38**, 205-47.

KITAGAWA, GENSHIRO. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Ann. Inst. Statist. Math.*, **46**, 605-23.

SORENSON, H. W. AND D. L. ALSPACH. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica*, **7**, 465-79.