BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
RESEARCH REPORT SERIES
No. RR-93/03

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

by

Professor James Durbin
Statistics Department
London School of Economics
and Political Science
Houghton Street
London WC2A 2AE
United Kingdom

Magdalena Cordero
Institute Nacional de Estadistica
po Castellana, 183
28046 Madrid
Spain

Report issued: April 7, 1993

22/2/93

HANDLING STRUCTURAL SHIFTS, OUTLIERS AND HEAVY-TAILED DISTRIBUTIONS IN STATE SPACE TIME SERIES MODELS[†]

J.Durbin and Magdalena Cordero,

London School of Economics and Political Science, U.K. and Instituto Nacional de Estadistica, Spain.

Summary

Time series containing abrupt structural shifts or outliers or both are considered. Techniques are developed for handling these using mixtures of densities, one component of which is a Gaussian density with a large variance. State space models are fitted to the series. The state vectors are estimated by the mode of their posterior density given the observations. The mode is found by Gauss-Newton iteration using Kalman filtering and smoothing. Three approximations to the likelihood function for estimating the hyperparameters are given. The techniques are illustrated by applying them to simulated and real series. The treatment is extended to deal with heavy-tailed densities.

Key words: Kalman filter; level shifts; mixtures of densities; outliers; robustification; seasonal; smoothing; state space models; time series; trend.

[†] Work done in the Statistics Research Division, US Bureau of the Census

1. Introduction

The object of this paper is to develop methods for dealing with outliers, structural shifts and heavy-tailed distributions in time series analysis. The state space model we use is very general and covers a wide range of applications including ARIMA models and spline smoothing. However, our main focus is on structural time series models in which the observed series is made up of trend, seasonal and irregular. Although we set out to devise tools that have general application, we were particularly anxious that our techniques should be applicable to monthly time series, partly because we hope that our methods will prove useful in the development of model-based techniques of seasonal adjustment for which automatic or nearly automatic means of handling outliers and structural shifts are essential.

Our basic tool is a mixture of densities in which structural shifts and outliers are allowed for by including in the model Gaussian components with large variances. The remaining densities can be Gaussian or they can be non-Gaussian such as Student's t. Given a sample of observations our object is to estimate the state vector. Since we wish the methods to apply to monthly time series the state transition matrices will be quite large, at least 13x13, so techniques based on numerical integration, such as that of Kitagawa (1987), are impractical. Our approach is to consider the posterior density of the series of state vectors given the observations and to estimate the state by the mode of this density. Our estimates can therefore be regarded as Bayes estimates. We use the mode rather than the mean because it is not feasible to use the mean. The mode is found by modified Gauss-Newton iteration using a Kalman filter and smoother at each step. Hyperparameters are estimated by approximate maximum likelihood.

There is a huge literature on state-space modelling with non-Gaussian data, going back more than twenty years. Some of the early work is reviewed in Chapter 8 of Anderson and Moore's (1979) text book. Key references are Alspach and Sorensen (1972) who introduced approximating by Gaussian mixture densities for filtering and Masreliez (1975) who gave filtering formulae when either the observation noise or the state noise is non-Gaussian. Good reviews of earlier work are given in Kitagawa (1987) and in the discussion of this paper, particularly in the extensive comments of Martin and Raftery (1987). Gaussian mixtures were used by Harrison and Stevens (1971,1976) under the name multi-process models for the treatment of a variety of problems including non-Gaussian data; see also Harrison and West (1989), Chapter 12. Peña and Guttman (1989) robustify the Kalman filter by Gaussian mixtures and give useful references to previous work.

Much of the work referred to above deals only with filtering. The most comprehensive treatment of both filtering and smoothing by Gaussian mixtures is by Kitagawa (1989,1991). At each updating step he takes a Gaussian mixture as the prior state density and then updates exactly to obtain a Gaussian mixture for the posterior. However, if this process were to be continued unmodified the number of components in the mixture would increase exponentially and so would rapidly become unmanageable. Consequently he "collapses" the posterior into a mixture of a smaller number of components at each update using the Kullback-Leibler distance for each pair of components as his criterion for collapsing. This is computationally time-consuming since many comparisons must be made at each update.

A different approach to handling outliers and level shifts has been developed by the time series section of the Statistics Research Division of the US Bureau of the Census as part of the X-12 seasonal adjustment procedure (see, for example, Bell (1983) for an early version of the method and Bruce and Jurke (1992) for discussion of later developments). They define regressors x_s for handling outliers and w_s for handling level shifts, where

$$x_s = 1, \quad s = t \quad \text{and} = 0, \quad s \geq t,$$

$$w_s = 0, \quad s < t \quad \text{and} = 1, \quad s \geq t,$$

and then fit ARIMA models with these considered as explanatory variables as t varies over 1,...,n. They have devised an automatic procedure for deciding when an outlier or level shift has occurred. While the method seems to be effective and robust we believe it lacks the direct elegance of our approach.

The feature which differentiates our approach from all these contributions is that, as far as we are aware, none of the other work uses posterior mode estimation. The most time-consuming part of our procedure is hyperparameter estimation, which has to be done by any method aiming at a complete solution; apart from this our method is very fast and is normally implemented in about five seconds on our unix machine (21 mips) for a seasonal series of around 250 observations..

In the next section we present the basic theory for the mixture method for both the case where the main density is Gaussian and the case where it is non-Gaussian. Section 3 gives three different approximations to the likelihood function for use in hyperparameter estimation. In section 4 we give three alternatives for modelling heavy-tailed densities using posterior mode estimation. Section 5 discusses some further aspects of the implementation of the theory. In section 6 the theory is illustrated by applying it to simulated and real time series.

2. Dealing with structural shifts and outliers by means of mixtures

The state space model we consider has the form

$$y_t = Z_t\alpha_t + \varepsilon_t, \quad t=1,\dots,n \tag{1a}$$

$$\alpha_{t+1} = T_t\alpha_t + G\eta_t, \quad t=0,\dots,n \tag{1b}$$

where y_t is a p×1 observational vector, α_t is an m×1 unobserved state vector and Z_t, T_t and G are non-stochastic matrices. We assume that ε_t and η_t are white noise series independent of each other with non-singular densities, and that $T_o\alpha_o = a_1$ where, in theoretical work, a_1 will normally be treated as fixed and known, whereas in applications it will normally be treated as diffuse or as an unknown vector to be estimated.

Our objective is to estimate $\alpha_1,\dots\alpha_n$ when the observations y_t contain outliers and the state α_t has structural shifts such as abrupt changes of level or slope of trend. We allow for structural shifts and outliers by using as our models for state and observation errors mixtures of densities which include Gaussian components with large variances. Denoting the set of observations $y_1,\dots,y_t$ by Y_t, we adopt as our state estimates the posterior mode estimates (PME's) of $\alpha_1,\dots,\alpha_n$ given Y_n, that is, the estimates obtained by maximising the posterior density of $\alpha_1,\dots,\alpha_n$ given Y_n. Since this density is just the joint density of $\alpha_1,\dots,\alpha_n$ and Y_n divided by the marginal density of Y_n, and since this latter density does not depend on the α's, we can obtain the PME's by maximising the joint density. These estimates are intuitively appealing. They can be thought of as analogues for random parameters of maximum likelihood estimates for fixed parameters. They can also be regarded as Bayes estimates that are obtained by taking the mode of the posterior distribution instead of the mean when, as for the problems considered in this paper, calculating the mean is impractical. PME's have been considered for related problems by Whittle (1991), Fahmeir and Kaufmann (1991), Fahmeir (1992) and Durbin and Koopman (1993).

We assume that $p(y_t|\alpha_1,\ldots,\alpha_t,Y_{t-1}) = p(y_t|\alpha_t)$ and that $p(\alpha_{t+1}|\alpha_t,\ldots,\alpha_1,Y_t) = p(\alpha_{t+1}|\alpha_t)$. In general the densities of ε_t and η_t, and possibly the matrices Z_t and T_t also, will depend on an unknown hyperparameter vector ψ, but for the development of the theory in this section we shall assume that ψ is known, deferring the estimation of ψ until section 3.

The main reason for the inclusion of the matrix G in (1b) is that some of the constituent relations in (1b) may be identities. The function of G is then to select those relations that have non-degenerate error terms. We therefore confine ourselves to the case where G is the identity matrix I_m or the columns of G are a subset of the columns of I_m. Thus $G'G = I_r$ where r is the number of non-degenerate error terms in (1b) and $\eta_t = G'(\alpha_{t+1} - T_t\alpha_t)$.

Returning to the problem of dealing with outliers and structural shifts, we shall show that the following simple device is remarkably effective in handling both problems. Let x be a component of ε_t or η_t and suppose that the density of x in the absence of outliers or structural shifts is $f(x,\sigma^2)$ where σ^2 is the variance of x. Then take the density of x with allowance for outliers and shifts to be the mixture

$$(1-\beta)f(x,\sigma^2) + \beta\ N(0,\lambda^2\sigma^2) \qquad (2)$$

where β is a pre-assigned small number and λ^2 is a pre-assigned large number. The results are relatively insensitive to the values β and λ^2; we have found values $\beta = 0.01$ and $\lambda^2 = 100$ to be effective. For simplicity we have considered here a mixture of two components only; later we present a general theory for an arbitrary number of components.

Before treating the general mixture model let us consider from the PME point of view the classical Gaussian model where $\varepsilon_t \sim N(0,H_t)$ and $\eta_t \sim N(0,Q_t)$, with H_t and Q_t positive-definite, since this provides the basis for our treatment of the general case. The density of $\alpha_1,\ldots,\alpha_{t+1},y_1,\ldots,y_t$ is $p_t(\alpha_1,\ldots,\alpha_{t+1},y_1,\ldots,y_t)$ where, apart from irrelevant constants,

$$\log p_t = -\frac{1}{2}\sum_{s=0}^{t}(\alpha_{s+1}-T_s\alpha_s)'GQ_s^{-1}G'(\alpha_{s+1}-T_s\alpha_s)$$
$$-\frac{1}{2}\sum_{s=1}^{t}(y_s-Z_s\alpha_s)'H_s^{-1}(y_s-Z_s\alpha_s). \qquad (3)$$

Note that $\mathrm{Var}(\alpha_{s+1}-T_s\alpha_s) = \mathrm{Var}(G\eta_s) = GQ_sG'$ and that this and $GQ_s^{-1}G'$ are Moore-Penrose generalised inverses of each other (see, for example, Rao (1973) section 1b.5).

The PME's of $\alpha_1,\ldots,\alpha_{t+1}$ given Y_t are obtained by differentiating (3) with respect to $\alpha_1,\ldots,\alpha_{t+1}$ and equating to zero. This gives

$$-GQ_{s-1}^{-1}G'(\alpha_s-T_{s-1}\alpha_{s-1})+T_s'GQ_s^{-1}G'(\alpha_{s+1}-T_s\alpha_s)+Z_s'H_s^{-1}(y_s-Z_s\alpha_s)=0 \quad (4a)$$

for s = 1,...,t together with

$$GQ_t^{-1}G'(\alpha_{t+1} - T_t\alpha_t) = 0. \qquad (4b)$$

Since the conditional distribution of $\alpha_1,\ldots,\alpha_{t+1}$ given Y_t is Gaussian, its conditional mode is equal to its conditional mean. Thus the PME of α_s given Y_t is equal to $E(\alpha_s|Y_t)$, $s = 1,\ldots,t+1$.

Let $a_t = E(\alpha_t|Y_{t-1})$, $P_t = \mathrm{Var}(\alpha_t|Y_{t-1})$, $v_t = y_t-Z_ta_t$, $F_t = E(v_tv_t') = Z_tP_tZ_t' + H_t$ and $K_t = T_tP_tZ_t'F_t^{-1}$. It is well known that a_{t+1} can be calculated recursively by the Kalman filter, which can be written in the form

$$a_{t+1} = T_ta_t + K_tv_t \qquad (5a)$$

$$P_{t+1} = T_tP_t(T_t-K_tZ_t)' + GQ_tG' \qquad (5b)$$

for t = 1,...,n. (See, for example, Harvey (1989)(3.2.4)).

Let $\hat{\alpha}_t = E(\alpha_t | Y_n)$. Then $\hat{\alpha}_t$ is called the smoothed value of α_t and can be calculated by a variety of smoothers including those of Anderson and Moore (1979), Chapter 7, de Jong (1989) and Koopman (1992). The most convenient for our purpose is Koopman's since it is the fastest. It is calculated by the forwards recursion

$$\hat{\alpha}_{t+1} = T_t\hat{\alpha}_t + G\hat{\eta}_t, \quad t = 0,\ldots,n-1 \tag{6}$$

where $\hat{\eta}_t = E(\eta_t | Y_n)$ and is given by

$$\hat{\eta}_t = Q_t G' r_t, \quad t = n,\ldots,0 \tag{7}$$

where r_t can be calculated by the backwards recursion

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, \quad t = n,\ldots,1 \tag{8}$$

with $r_n = 0$ and $L_t = T_t - K_t Z_t$. Relations (7) and (8) can be obtained by a slight modification to allow for the inclusion of G of the derivation given in section 3 of Durbin and Koopman (1993).

We now investigate methods of filtering and smoothing for model (1) when ε_t has the mixture density

$$h_t(\varepsilon_t) = \sum_{i=1}^{k} \beta_i h_{it}(\varepsilon_t), \quad \beta_i \geq 0, \quad \sum_{i=1}^{k} \beta_i = 1 \tag{9}$$

and η_t has the mixture density

$$q_t(\eta_t) = \sum_{i=1}^{l} \delta_i q_{it}(\eta_t), \quad \delta_i \geq 0, \quad \sum_{i=1}^{l} \delta_i = 1, \tag{10}$$

where h_{it} and q_{it} are suitable densities. We shall begin by supposing that all the component densities are Gaussian with $h_{it} \sim N(0,H_{it})$, $q_{it} \sim N(0,Q_{it})$. The density of $\alpha_1,\ldots,\alpha_{t+1},Y_t$ is $p_t(\alpha_1,\ldots,y_t)$ where, apart from constants,

$$\log p_t = \sum_{s=0}^{t} \log q_s(\ \ _s) + \sum_{s=1}^{t} \log h_s(\varepsilon_s)$$

with $\eta_s = G'(\alpha_{s+1} - T_s\alpha_s)$, $\varepsilon_s = y_s - Z_s\alpha_s$. Differentiating with respect to $\alpha_1,\ldots,\alpha_{t+1}$, equating to zero and using (9) and (10) gives

$$-G\tilde{Q}_{s-1}^{-1}G'(\alpha_s-T_{s-1}\alpha_{s-1})+T_s'G\tilde{Q}_s^{-1}G'(\alpha_{s+1}-T_s\alpha_s)+Z_s'\tilde{H}_s^{-1}(y_s-Z_s\alpha_s)=0 \quad (11a)$$

for s = 1,...,t together with

$$G\tilde{Q}_t^{-1}G'(\alpha_{t+1} - T_t\alpha_t) = 0, \quad (11b)$$

where

$$\tilde{Q}_s^{-1} = \frac{1}{q_s(\eta_s)} \sum_{i=1}^{1} \delta_i q_{is}(\eta_s) Q_{is}^{-1}, \quad (12a)$$

$$\tilde{H}_s^{-1} = \frac{1}{h_s(\varepsilon_s)} \sum_{i=1}^{k} \beta_i h_{is}(\varepsilon_s) H_{is}^{-1} \quad (12b)$$

with $\eta_s = G'(\alpha_{s+1} - T_s\alpha_s)$ and $\varepsilon_s = y_s - Z_s\alpha_s$. The PME 's of $\alpha_1,\ldots,\alpha_n$ given Y_n are obtained by solving equations (11) with t=n.

We solve equations (11) iteratively by a modified Gauss-Newton technique similar to the one developed by Durbin and Koopman (1993) for exponential family observations. Suppose that $\alpha_s^{(j)}$ is the jth approximation to the smoothed value $\hat{\alpha}_s$ of α_s where $\hat{\alpha}_1,\ldots,\hat{\alpha}_{n+1}$ are the solution of equations (11) when t=n. Let

$$Q_{s(j)}^{-1} = \frac{1}{q_s(\eta_s^{(j)})} \sum_{i=1}^{1} \delta_i q_{is}(\eta_s^{(j)}) Q_{is}^{-1} \quad (13a)$$

and

$$H_{s(j)}^{-1} = \frac{1}{h_s(\varepsilon_s^{(j)})} \sum_{i=1}^{k} \beta_i h_{is}(\varepsilon_s^{(j)}) H_{is}^{-1} \quad (13b)$$

where $\eta_s^{(j)} = G'(\alpha_{s+1}^{(j)} - T_s\alpha_s^{(j)})$ and $\varepsilon_s^{(j)} = y_s - Z_s\alpha_s^{(j)}$. Now substitute $Q_{s(j)}^{-1}$ for $\tilde{Q}_s^{-1}$ and $H_{s(j)}^{-1}$ for $\tilde{H}_s^{-1}$ in (11). The resulting equations are

$$-GQ_{s-1(j)}^{-1}G'(\alpha_s - T_{s-1}\alpha_{s-1}) + T_s'GQ_{s(j)}^{-1}G'(\alpha_{s+1} - T_s\alpha_s)$$
$$+Z_s'H_{s(j)}^{-1}(y_s - Z_s\alpha_s) = 0 \quad (14a)$$

for s = 1,...,t together with

$$GQ_{t(j)}^{-1}G'(\alpha_{t+1} - T_t\alpha_t) = 0 \quad (14b)$$

These equations are linear with exactly the same structure as equations

(4) for the classical Gaussian model. Let $a_{t+1}^{(j+1)}$ be the value of α_{t+1} obtained in the solution of these equations. It follows from (4) and (5) that $a_{t+1}^{(j+1)}$ can be calculated recursively by the Kalman filter relations

$$a_{t+1}^{(j+1)} = T_t a_t^{(j+1)} + K_t^{(j)} v_t^{(j)} \qquad (15a)$$

$$P_{t+1}^{(j)} = T_t P_t^{(j)} (T_t - K_t^{(j)} Z_t)' + G Q_{t(j)} G' \qquad (15b)$$

where $v_t^{(j)} = y_t - Z_t a_t^{(j+1)}$ and $K_t^{(j)} = T_t P_t^{(j)} Z_t' F_{t(j)}^{-1}$ with $F_{t(j)} = Z_t P_t^{(j)} Z_t' + H_{t(j)}$ for $t = 1,\dots,n$.

Similarly, let $\alpha_s^{(j+1)}$ be the value of α_s in the solution of the equations when $t=n$. Then $\alpha_t^{(j+1)}$ can be computed by the smoother (5), (6) and (7) with Q_t replaced by $Q_{t(j)}$, F_t replaced by $F_{t(j)}$ and L_t replaced by $L_t^{(j)} = T_t - K_t^{(j)} Z_t$ for $t=1,\dots,n$. It follows that $\alpha_t^{(j+1)}$ is the (j+1)th approximation to the PME $\hat{\alpha}_t$ of α_t for $t = 1,\dots,n$. Iteration is continued until suitable convergence has been achieved. The PME's of $\alpha_1,\dots,\alpha_n$ can therefore be computed iteratively by the standard Kalman filter and smoother, provided, of course, that the iterative process converges. Since there is no guarantee that the function log p_t is always unimodal care must be taken in the provision of starting values for the iteration. A suitable technique for the purpose will be given in section 5.

We now demonstrate how the theory is modified when one or more of the component densities of the mixture are non-Gaussian. For simplicity let us take the case where y_t is univariate, the elements of η_t are independent, each mixture has only two component densities, the distributions of ε_t and η_t are the same for all t and the second component of each mixture is Gaussian. To begin with, let us assume that whereas the density of η_t is a Gaussian mixture , the density of ε_t is

$$h(\varepsilon_t) = \beta_1 h_1(\varepsilon_t) + \beta_2 h_2(\varepsilon_t), \quad \beta_1 + \beta_2 = 1,$$

where h_1 is a non-Gaussian density with variance σ_ε^2 and $h_2 \sim N(0,\lambda^2\sigma_\varepsilon^2)$. Let $g(\varepsilon) = -\log h_1(\varepsilon)$ and let $\dot{g}(\varepsilon)$ be the derivative of $g(\varepsilon)$. Then

$$\frac{d\log h(\varepsilon_s)}{d\alpha_s} = \frac{Z_s'}{h(\varepsilon_s)}\left[\frac{\beta_1 h_1(\varepsilon_s)\dot{g}(\varepsilon_s)}{\varepsilon_s} + \frac{\beta_2 h_2(\varepsilon_s)}{\lambda^2\sigma_\varepsilon^2}\right]\varepsilon_s$$

where $\varepsilon_s = y_s - Z_s\alpha_s$. With $\alpha_s^{(j)}$ as the jth approximation to $\hat{\alpha}_s$ we approximate the right-hand side of this by

$$\frac{Z_s'}{h(\varepsilon_s^{(j)})}\left[\frac{\beta_1 h_1(\varepsilon_s^{(j)})\dot{g}(\varepsilon_s^{(j)})}{\varepsilon_s^{(j)}} + \frac{\beta_2 h_2(\varepsilon_s^{(j)})}{\lambda^2\sigma_\varepsilon^2}\right](y_s - Z_s\alpha_s)$$

where $\varepsilon_s^{(j)} = y_s - Z_s\alpha_s^{(j)}$. We therefore obtain the same equations (14) as in the Gaussian mixture case except that, instead of (13b), $H_{s(j)}^{-1}$ is given by the scalar

$$H_{s(j)}^{-1} = \frac{1}{h(\varepsilon_s^{(j)})}\left[\frac{\beta_1 h_1(\varepsilon_s^{(j)})\dot{g}(\varepsilon_s^{(j)})}{\varepsilon_s^{(j)}} + \frac{\beta_2 h_2(\varepsilon_s^{(j)})}{\lambda^2\sigma_\varepsilon^2}\right]. \qquad (16)$$

Similarly, if the density of the ith element η_{it} of the state error η_t has density

$$q_i(\eta_{it}) = \delta_1 q_{1i}(\eta_{it}) + \delta_2 q_{2i}(\eta_{it}), \quad \delta_1 + \delta_2 = 1,$$

where q_{1i} is a non-Gaussian density with variance $\sigma_{\eta i}^2$ and $q_{i2} \sim N(0,\lambda_i^2\sigma_{\eta i}^2)$, the same equations (14) are obtained except that $Q_{s(j)}^{-1}$ is given, not by (13a) but by

$$Q_{s(j)}^{-1} = \mathrm{diag}\left\{\frac{1}{q_i\eta_{is}^{(j)}}\left[\frac{\delta_1 q_{1i}(\eta_{is}^{(j)})\dot{f}(\eta_{is}^{(j)})}{\eta_{is}^{(j)}} + \frac{\delta_2 q_{2i}(\eta_{is}^{(j)})}{\lambda_i^2\sigma_{\eta i}^2}\right]\right\} \qquad (17)$$

where $f(\eta) = -\log q_{1i}(\eta)$ and $\dot{f}(\eta)$ is its derivative.

As an example, when h_1 is the scaled t-density with variance σ_ε^2,

$$h_1(\varepsilon) = \frac{c(\nu)}{\left[1 + \frac{\varepsilon^2}{(\nu-2)\sigma_\varepsilon^2}\right]^{(\nu+1)/2}} \qquad (\nu>2) \qquad (18)$$

then $\dot{g}(\varepsilon_s^{(j)})/\varepsilon_s^{(j)}$ in (16) is equal to $(\nu+1)/[(\nu-2)\sigma_\varepsilon^2 + \varepsilon_s^{(j)2}]$; this obviously converges to $1/\sigma_\varepsilon{}^2$ as ν tends to infinity. This density is one of those that we shall consider in section 6 for fitting heavy-tailed distributions.

3. Approximate maximum likelihood estimation of hyperparameters

We now consider the estimation of the unknown hyperparameter vector ψ. Our approach is to construct approximations to the likelihood which are then maximised by numerical optimisation algorithms. We begin by considering the Gaussian form of model (1) in which $\varepsilon_t \sim N(0,H_t)$ and $\eta_t \sim N(0,Q_t)$. Let $p_G(Y_n|\psi)$ be the density of Y_n given ψ or, equivalently, the likelihood of ψ given Y_n.

It is well known that p_G is given by the prediction error decomposition form

$$p_G(Y_n|\psi) = (2\pi)^{-np/2}\prod_{t=1}^{n}|F_t|^{-1/2}\exp\Big(-\frac{1}{2}\sum_{t=1}^{n}v_t'F_t^{-1}v_t\Big)$$

(see for example Harvey (1989) equation (3.4.5)) where F_t and v_t are defined in the paragraph containing (2). Let $p(Y_n|\psi)$ be the density of Y_n given ψ in the general case.By analogy with this expression our first approximate form for the likelihood in the mixture case is

$$p_1(Y_n|\psi) = (2\pi)^{-np/2}\prod_{t=1}^{n}|F_{t(c)}|^{-1/2}\exp\Big(-\frac{1}{2}\sum_{t=1}^{n}v_t^{(c)\prime}F_{t(c)}^{-1}v_t^{(c)}\Big) \tag{19}$$

where $F_{t(c)}$ and $v_t^{(c)}$ are the final values of $F_{t(j)}$ and $v_t^{(j)}$ at convergence of the iterative estimation of the $\hat{\alpha}_t$'s, and where $F_{t(c)}$ and $v_t^{(c)}$ are defined in the paragraph containing (15) with $j = c$.

To derive our second approximation form, let α, $\hat{\alpha}$ be the stacked vectors $[\alpha_1',\dots,\alpha_n']'$, $[\hat{\alpha}_1',\dots,\hat{\alpha}_n']'$. For the Gaussian and mixture models denote $E[(\alpha-\hat{\alpha})(\alpha-\hat{\alpha})']$ by V_G and V respectively and let $p_G(\alpha,Y_n|\psi)$, $p(\alpha,Y_n|\psi)$ be the joint densities of α,Y_n under the two models. Then

$$p_G(\alpha, Y_n|\psi) = p_G(Y_n|\psi)(2\pi)^{-nm/2}|V_G|^{-1/2}\exp\left[-\frac{1}{2}(\alpha-\hat{\alpha})' V_G^{-1}(\alpha-\hat{\alpha})\right],$$

so on putting $\alpha = \hat{\alpha}$ we obtain

$$p_G(Y_n|\psi) = (2\pi)^{nm/2}|V_G|^{1/2}p_G(\hat{\alpha}, Y_n|\psi).$$

Durbin and Koopman (1993), equation (20), have shown that

$$|V_G| = \prod_{t=1}^{n} |Q_{t-1}||H_t||F_t|^{-1}.$$

Hence,

$$p_G(Y_n|\psi) = (2\pi)^{nm/2} \prod_{t=1}^{n} |Q_{t-1}|^{1/2}|H_t|^{1/2}|F_t|^{-1/2}p_G(\hat{\alpha}, Y_n|\psi).$$

By analogy with this our second approximate form for the likelihood is

$$p_2(Y_n|\psi) = (2\pi)^{nm/2} \prod_{t=1}^{n} |Q_{t-1(c)}|^{1/2}|H_{t(c)}|^{1/2}|F_{t(c)}|^{-1/2}p(\hat{\alpha}, Y_n|\psi) \tag{20}$$

where $F_{t(c)}$ is as in (20) and $Q_{t-1(c)}^{-1}$, $H_{t(c)}^{-1}$ are the final forms in the iteration of $Q_{t-1(j)}^{-1}$ and $H_{t\ (j)}^{-1}$ given by (13) or (16) and (17).

Our third approximation is obtained by a special case of Kitagawa's (1989) "collapsing" method. This applies only to Gaussian mixtures. The basic idea is to collapse mixtures after updating at each time point into a single Gaussian density and then use this to obtain a manageable expression for $p(y_{t+1}|Y_t,\psi)$ for $t=1,\ldots,n-1$, from which the likelihood is obtained. In fact, the idea of collapsing to a single Gaussian goes back to Alpbach and Sorenson (1972), Masreliez (1975) and Harrison and Stevens (1976). To simplify the notation we suppress the dependence on ψ in the derivation. We first show how to get from a single Gaussian for $p(\alpha_{t-1}|Y_{t-1})$ to a mixture for $p(\alpha_t|Y_t)$ and how to collapse this to a single Gaussian. We then show how to obtain $p(y_{t+1}|Y_t)$ from this as a mixture.

Write the observational and state mixtures as

$$p(y_t|\alpha_t) = \sum_i \beta_i p_i(y_t|\alpha_t), \quad \sum_i \beta_i = 1 \tag{21a}$$

$$p(\alpha_{t+1}|\alpha_t) = \sum_j \delta_j p_j(\alpha_{t+1}|\alpha_t), \quad \sum_j \delta_j = 1. \tag{21b}$$

Denoting Kitagawa (1989) by K we have from K(5)

$$p(\alpha_t|Y_t) = \frac{p(y_t|\alpha_t)p(\alpha_t|Y_{t-1})}{p(y_t|Y_{t-1})} \tag{22}$$

$$= \frac{\sum_i \beta_i p_i(y_t|\alpha_t)p(\alpha_t|Y_{t-1})}{p(y_t|Y_{t-1})} \tag{23}$$

and from K(4)

$$p(\alpha_t|Y_{t-1}) = \int_{-\infty}^{\infty} p(\alpha_t|\alpha_{t-1})p(\alpha_{t-1}|Y_{t-1})d\alpha_{t-1}$$

$$= \sum_j \delta_j \int_{-\infty}^{\infty} p_j(\alpha_t|\alpha_{t-1})p(\alpha_{t-1}|Y_{t-1})d\alpha_{t-1}$$

$$= \sum_j \delta_j p_j(\alpha_t|Y_{t-1}).$$

Substituting in (23) gives

$$p(\alpha_t|Y_t) = \frac{\sum_{i,j} \beta_i \delta_j p_i(y_t|\alpha_t)p_j(\alpha_t|Y_{t-1})}{p(y_t|Y_{t-1})}$$

$$= \sum_{i,j} \beta_i \delta_j \rho_{ijt} p_{ij}(\alpha_t|Y_t) \tag{24}$$

on using (22), where

$$\rho_{ijt} = \frac{p_{ij}(y_t|Y_{t-1})}{p(y_t|Y_{t-1})} = \frac{p_{ij}(y_t|Y_{t-1})}{\sum_{i,j} \beta_i \delta_j p_{ij}(y_t|Y_{t-1})} \tag{25}$$

and where $p_{ij}(\alpha_t|Y_t)$ and $p_{ij}(y_t|Y_{t-1})$ are obtained by standard Kalman filtering assuming $p(y_t|\alpha_t) = p_i(y_t|\alpha_t)$ and $p(\alpha_t|\alpha_{t-1}) = p_j(\alpha_t|\alpha_{t-1})$ and also assuming that $p(\alpha_{t-1}|Y_{t-1})$ is a single Gaussian.

To collapse the right-hand side of (24) into a single Gaussian, assume that $p_{ij}(\alpha_t|Y_t) \sim N(\mu_{ijt}, V_{ijt})$ where μ_{ijt}, V_{ijt} are the values given by the Kalman filter, and let $\mu_t = E(\alpha_t|Y_t)$ and $V_t = Var(\alpha_t|Y_t)$. Then

$$\mu_t = \sum_{i,j} \beta_i \delta_j \rho_{ijt} \mu_{ijt} \tag{26a}$$

$$V_t = \sum_{i,j} \beta_i \delta_j \rho_{ijt} [V_{ijt} + (\mu_{ijt} - \mu_t)(\mu_{ijt} - \mu_t)'] . \tag{26b}$$

Now assume that $p(\alpha_t | Y_t) \sim N(\mu_t, V_t)$. This completes the operation of updating $p(\alpha_{t-1} | Y_{t-1})$.

We now calculate $p(y_{t+1} | Y_t)$ from the single Gaussian density $p(\alpha_t | Y_t)$. From K p.505

$$\begin{aligned} p(y_{t+1} | Y_t) &= \int_{-\infty}^{\infty} p(y_{t+1} | \alpha_{t+1}) p(\alpha_{t+1} | Y_t) d\alpha_{t+1} \\ &= \sum_{i,j} \beta_i \delta_j \int_{-\infty}^{\infty} p_i(y_{t+1} | \alpha_{t+1}) p_j(\alpha_{t+1} | Y_y) d\alpha_{t+1} \\ &= \sum_{i,j} \beta_i \delta_j p_{ij}(y_{t+1} | Y_t) \end{aligned} \tag{27}$$

where $p_{ij}(y_{t+1} | Y_t)$ is obtained by standard Kalman filtering steps assuming $p(y_{t+1} | \alpha_{t+1}) = p_i(y_{t+1} | \alpha_{t+1})$ and $p(\alpha_{t+1} | \alpha_t) = p_j(\alpha_{t+1} | \alpha_t)$. Our third approximation to the likelihood is then

$$p_3(Y_n | \psi) = \prod_{t=0}^{n-1} p(y_{t+1} | Y_t, \psi) \tag{28}$$

where $p(y_{t+1} | Y_t, \psi)$ is calculated using (27) for $t \geq 1$.

These approximate likelihoods are to be maximised numerically. For the first two approximations, values of the likelihood are calculated from the results of the iterated estimation of $\hat{\alpha}$ and are therefore subject to small errors since the iteration cannot be continued until these errors are infinitesimal. Consequently,they are not suitable for maximising routines that calculate gradients from adjacent values of the hyperparameters and proceed from the last approximation along directions determined solely by these gradients, so routines should be used such as the downhill simplex method, given in section 10.4 of Press _et_. _al_. (1988), Powell's method, given in section 10.5 of the same book or the Gill-Murray-Pitfield algorithm, which is E04JBF in the NAG library. The third approximation is not subject to errors of the same type so a wider range of algorithms can be employed. Since this approximation is computationally economical we suggest that in the Gaussian-mixture case it should be used either as the sole method or to provide starting values for one of the other two.

4. Heavy-tailed distributions

In the last section we assumed that the main densities for both observations and state errors are Gaussian. However, it is well known that in many areas of application, particularly with economic data, actual distributions tend to have heavier tails than the Gaussian distribution. In this section we therefore consider how to modify the theory of the previous sections in order to accommodate heavy-tailed densities.

To begin with, let us leave aside questions of handling structural shifts and outliers. We shall consider three different forms of heavy-tailed density. Since our prime concern is with univariate series and with the observational error ε_t we start with this case. The first form of heavy-tailed density we shall take for ε_t is the Gaussian mixture

$$h(\varepsilon) = (1-\beta)h_1(\varepsilon) + \beta h_2(\varepsilon), \quad 0\leq\beta\leq1$$

where $h_1 \sim N(0,\tau^2)$ and $h_2 \sim N(0,\lambda^2\tau^2)$, $\lambda>1$. The variance of ε is $\sigma^2 = (1-\beta)\tau^2 + \beta^2\lambda^2\tau^2 = [1 + \beta(\lambda^2-1)]\tau^2$. In the type of application we are concerned with in this paper it is important to keep the number of parameters to be estimated as low as possible. We therefore recommend that the value of λ should normally be pre-assigned by the investigator, say in the range 2 to 4 and possibly after experimentation. The values of β would normally be estimated as part of the likelihood maximisation procedures though if the estimation turns out to be difficult this could also be pre-assigned or estimated roughly by trial and error. We suggest that the pair σ^2, β be estimated rather than τ^2, β since σ^2 is relatively independent of β. This discussion has been based on the assumption that the density is homogeneous over time, which is the normal situation. There is no difficulty in principle in extending the treatment to densities which change over time in a predetermined way. Similar considerations apply to state error densities. Since these models are Gaussian mixtures, the theory required for handling them carries over from sections 2 and 3 unchanged, subject to the inclusion of additional unknown parameters in the hyperparameter vector.

The second heavy-tailed distribution we consider is Student's t-distribution, which we write in the form

$$h(\varepsilon) = \frac{\Gamma(\nu/2 + 1/2)}{\sigma[(\nu-2)\pi]^{1/2}\Gamma(\nu/2)} \frac{1}{\left[1 + \frac{\varepsilon^2}{(\nu-2)\sigma^2}\right]^{(\nu+1)/2}}, \quad \nu > 2$$

so that $\mathrm{Var}(\varepsilon) = \sigma^2$ for all ν thus keeping estimation of σ^2 and ν relatively independent in the estimation process. Since

$$-\frac{d\log h(\varepsilon)}{d\varepsilon} = \frac{\nu+1}{(\nu-2)\sigma^2} \frac{\varepsilon}{\left[1 + \frac{\varepsilon^2}{(\nu-2)\sigma^2}\right]} = \frac{(\nu+1)\varepsilon}{[(\nu-2)\sigma^2+\varepsilon^2]}$$

the contribution of $\partial\log h(y_s - Z_s\alpha_s)/\partial\alpha_s$ to $\partial\log p_t(\alpha_1,\ldots,\alpha_{t+1},y_1,\ldots,y_t)/\partial\alpha_s$ is

$$\frac{(\nu+1)Z_s'(y_s-Z_s\alpha_s)}{[(\nu-2)\sigma^2 + (y_s-Z_s\alpha_s)^2]}.$$

Suppose that in the iterative estimation of the PME's of $\alpha_1,\ldots,\alpha_{n+1}$ the jth approximation to $\hat{\alpha}_s$ is $\alpha_s^{(j)}$. Let

$$H_{s(j)}^{-1} = \frac{\nu+1}{[(\nu-2)\sigma^2 + (y_s-Z_s\alpha_s^{(j)})^2]} \qquad (29)$$

Substituting this in (14a), where $Q_{r(j)}^{-1}$ is suitably defined for $r = s-1,s$, we can then obtain the (j+1)th approximation $\alpha_s^{(j+1)}$ by Kalman filtering and smoothing as before.

Since the t-density is not a Gaussian mixture, except as an integral form, the third approximate likelihood (28) cannot be used for estimation and one of the other two approximate forms (19) or (20) should be used instead. The Gaussian density, corresponding to $\nu = \infty$ in (29), could be employed to start the iterations. Mixtures in which the principal density is a t-density with variance σ^2 and the subsidiary density is Gaussian with variance $\lambda^2\sigma^2$ may be used as in (16) with h_1 given by (18). Similar considerations apply when the t-distribution is used to model the principal component of state error densities, with (17) being used to bring in Gaussian subsidiaries.

Our third heavy-tailed distribution is the general error distribution with density

$$h(\varepsilon) = \frac{w(\kappa)}{\sigma} \exp\left[-c(\kappa) \left|\frac{\varepsilon}{\sigma}\right|^{\kappa}\right], \quad 1<\kappa<2 \qquad (30)$$

where

$$w(\kappa) = \frac{2[\Gamma(3\kappa/4)]^{1/2}}{\kappa[\Gamma(\kappa/4)]^{3/2}}, \qquad c(\kappa) = \left[\frac{\Gamma(3\kappa/4)}{\Gamma(\kappa/4)}\right]^{\kappa/2}.$$

Some details about this distribution are given by Box and Tiao (1973), section 3.2.1, from which it follows that $\mathrm{Var}(\varepsilon) = \sigma^2$ for all κ. We include this density for completeness since it is often advocated for the representation of heavy-tailed data but we have not investigated its performance and it may turn out that it is not suitable in the present context because $H_{s(j)}^{-1}$ in (31) below takes large values when $|y_s - Z_s\alpha_s^{(j)}|$ is small.

Since

$$\frac{d\log h(\varepsilon)}{d\varepsilon} = -\frac{c(\kappa)\kappa}{\sigma^{\kappa}} |\varepsilon|^{\kappa-2}\varepsilon$$

the contribution of $\partial\log h(y_s-Z_s\alpha_s)/\partial\alpha_s$ to $\partial\log p_t(\alpha_1,\dots,\alpha_{t+1},y_1,\dots,y_t)/\partial\alpha_s$

is

$$\frac{c(\kappa)\kappa}{\sigma^{\kappa}} |y_s - Z_s\alpha_s|^{\kappa-2}(y_s - Z_s\alpha_s).$$

Thus if $\alpha_s^{(j)}$ is the jth approximation to $\hat{\alpha_s}$ we can take

$$H_{s(j)}^{-1} = \frac{c(\kappa)\kappa}{\sigma^{\kappa}} |y_s - Z_s\alpha_s^{(j)}|^{\kappa-2} \qquad (31)$$

and proceed as with the t-distribution. Since $\kappa-2<0$, it is desirable to put an arbitrary ceiling on the value of $H_{s(j)}^{-1}$ to guard against the possibility that y_s is equal to or very nearly equal to $Z_s\alpha_s^{(j)}$. Similar considerations apply to the state error densities.

In principle κ can be estimated by approximate maximum likelihood by incorporating it into the hyperparameter vector and using one of the first two approximations of section 3. However, it is important to bear in mind with regard to all three densities and particularly the t-density and this density that the objective is to achieve a better performance than with the Gaussian density rather than to achieve excellence in the estimation of parameters of the densities other than their variances. Thus rough and ready preassignment based on residual analysis assuming Gaussianity, or on experimentation, may be justifiable on occasion.

5. Further aspects of the mixture-PME technique

The key to our method of handling outliers and structural shifts is the use of mixtures of the form

$$h(\varepsilon) = (1-\beta)N(0,\sigma^2) + \beta N(0,\lambda^2\sigma^2) \qquad (32)$$

Consider the use of this mixture for the observational density in the univariate case; similar considerations apply in the multivariate case and to state errors. The first question we discuss in this section is the extent of the downweighting of the contribution of observation y_t to the estimation of state when y_t is an outlier so $|y_t - Z_t\alpha_t|$ is large. The contribution of y_t to $\partial \log p(\alpha_1,\ldots,\alpha_{n+1},y_1,\ldots,y_n)/\partial\alpha_t$ is, using (11a),

$$-Z_t' \frac{d\log h(\varepsilon_t)}{d\varepsilon_t} = \frac{Z_t'}{\sigma^2}(y_t - Z_t\alpha_t)f(x_t)$$

where

$$f(x_t) = -\frac{\sigma^2}{\varepsilon_t}\frac{d\log h(\varepsilon_t)}{d\varepsilon_t}$$

with $\varepsilon_t = y_t - Z_t\alpha_t$ and $x_t = \varepsilon_t/\sigma$. When $\beta = 0$, $f(x) = 1$ for all x, emphasising the linearity of (11a) in the Gaussian case. When $\beta > 0$ and $\lambda^2 > 1$, the weight function $f(x_t)$ indicates the downweighting of the contribution from y_t as $|y_t - Z_t\alpha_t|$ becomes large. For the mixture (32),

$$f(x) = \frac{(1-\beta)e^{-x^2/2} + \beta\lambda^{-3}e^{-x^2/2\lambda^2}}{(1-\beta)e^{-x^2/2} + \beta\lambda^{-1}e^{-x^2/2\lambda^2}} \qquad (33)$$

which tends to $[(1-\beta) + \beta\lambda^{-3}]/[(1-\beta) + \beta\lambda^{-1}]$ as $x \to 0$ and to λ^{-2} as $|x| \to \infty$ for $\beta > 0$. Of course it is only the relative magnitudes which matter, so the fact that f(x) is always < 1 when $\beta > 0$ whereas it always equals 1 when the errors are Gaussian is of no significance.

Figure 1a gives the weight function (33) for the case $\beta = 0.01$, $\lambda^2 = 100$ which are the values we have used for all the work on outliers and structural shifts in this paper. We see that the weight drops from over 0.9 when the standardised irregular x is 2 to near to 0.01 when it is 4 or more; this is consistent with the kind of treatment that is often advocated for outliers in applied work. Figure 1b gives the function for $\beta = 0.25$, $\lambda^2 = 9$, values that might be considered appropriate for a heavy-tailed density in the absence of large outliers. Here the weight drops smoothly from x = 0 to x = 4 by a factor of around 1/7. Figure 1c has $\beta = 0.35$, $\lambda^2 = 4$ and is intended for moderately heavy tails. Figure 1d adds provision for large outliers to allowance for heavy tails by taking a mixture of 0.65 times $N(0,\sigma^2)$, 0.34 times $N(0,4\sigma^2)$ and 0.01 times $N(0,100\sigma^2)$. This gives a relatively slow decline from 0.74 to 0.08 as x moves from 0 to 6. This diagram suggests the possibility of the investigator designing his own function f(x) a priori either with an appropriate Gaussian mixture or by taking f(x) as a piecewise linear function and integrating $d\log h(\varepsilon)/d\varepsilon$ to give a density with a piecewise cubic function of ε in the exponent; however, these possibilities will not be pursued further in this paper.

For Student's t-distribution considered in section 4 we have

$$f(x) = \frac{\nu+1}{\nu-2+x^2}, \quad \nu > 2.$$

This is plotted for $\nu = 8$ in Figure 1e. The function drops fairly rapidly from 1.5 at $x = 0$ to 0.6 at $x = 3$. For the general error distribution of section 4,

$$f(x) = c(\kappa)\kappa|x|^{\kappa-2}, \quad 1 < \kappa < 2$$

which is plotted for $\kappa = 1.5$ and $f(x) \leq 2$ in Figure 1f. The steep drop for $x < 1$ seems to confirm the impression that this density might be unsuitable for the approach of this paper; however, we have not investigated the matter in detail.

The manner in which the mixture model (32) handles outliers by downweighting the contribution to trend estimation of large deviations $y_t-Z_t\alpha_t$ is evident from Figure 1a. However, it is not immediately obvious why the same device should be equally helpful in handling abrupt structural shifts. In fact, similar considerations apply. If a component of $\alpha_{t+1}-T_t\alpha_t$ is large, it is given less weight by the mixture model than the Gaussian model and is correspondingly more acceptable in the state estimation process.

The contribution of y_t to $\partial\log p(\alpha_1,\ldots,y_n)/\partial\alpha_t$ can also be written as

$Z_t'(y_t-Z_t\alpha_t)/\tilde{\sigma}^2$ where

$$\tilde{\sigma}^2 = -\frac{\varepsilon_t}{d\log h(\varepsilon_t)/d\varepsilon_t} = \frac{(1-\beta)e^{-\varepsilon^2/2\sigma^2} + \beta\lambda^{-1}e^{-\varepsilon^2/2\lambda^2\sigma^2}}{(1-\beta)e^{-\varepsilon^2/2\sigma^2} + \beta\lambda^{-3}e^{-\varepsilon^2/2\lambda^2\sigma^2}}\sigma^2. \quad (34)$$

We call $\tilde{\sigma}^2$ the quasi-variance and have found it a useful concept for indicating outliers. Similar considerations apply to state errors and structural shifts. It is interesting to relate (34) to the posterior probability of an outlier. Suppose we postulate a model in which for most of the time the error has density $N(0,\sigma^2)$ while if it is an outlier its error has density $N(0,\lambda^2\sigma^2)$, the overall density being given by (32). Then the prior probability of an outlier is β and the posterior probability is

$$\tilde{\beta} = \frac{\beta\lambda^{-1}e^{-\varepsilon^2/2\lambda^2\sigma^2}}{(1-\beta)e^{-\varepsilon^2/2\sigma^2} + \beta\lambda^{-1}e^{-\varepsilon^2/2\lambda^2\sigma^2}} \quad (35)$$

We see that as $|\varepsilon|$ increases, $\tilde{\sigma}^2$ increases to $\lambda^2\sigma^2$ and $\tilde{\beta}$ increases to one. In effect, both quantities are indicating the same behaviour pattern. Some examples are shown in the next section.

Of these two measures we prefer $\tilde{\sigma}^2$, for two reasons. First, we see no need to introduce a probabilistic mechanism for the occurrence of outliers. To us, (32) is a statistical model for a particular type of data set, one thought to contain a small proportion of large deviations, and its justification is to be found in the extent to which it facilitates an acceptable analysis of the data, rather than whether it correctly reflects the mechanism by which these large deviations occurred. Secondly, we find the scale of measurement in terms of variance more directly intelligible than a scale of measurement in terms of probabilities.

The second aspect we consider in this section is the estimation of trend. Our basic model for the trend μ_t is the local linear trend (LLT) model

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \xi_\tau$$

$$\beta_t = \beta_{t-1} + \zeta_t$$

where ξ_t and ζ_t are white noise. In the absence of slope in the trend this collapses to the random walk (RW) model $\Delta\mu_t = \xi_t$. A competitor to the LLT is the integrated random walk (IRW) model $\Delta^2\mu_t$ = white noise, favoured by P. Young and his collaborators; see for example Young _et_ _al_ (1991). We have found all three models capable of coping with level shifts in non-seasonal series when used with our mixture model, but in the presence of seasonality we have found that the IRW sometimes has a tendency to follow the seasonal pattern. At the same time some workers feel that the output of the RW or the LLT model is somewhat lacking in smoothness for an estimate of trend. As will be illustrated in the next section, our general recommendation is therefore that when allowing for level shifts the RW or the LLT should be used, but that if a smoother final trend estimate is desired the IRW or equivalently the LLT with level variance set equal to zero, together with the mixture model for the errors, should be applied subsequently to the deseasonalised series or the initial trend estimate.

We mentioned in section 2 the importance of good starting values for the state iteration. We now discuss a two-filter method for obtaining first approximations to the state vectors which have been very good in the examples we have considered. Let $Y^n_{t+1} = (y_{t+1},\dots,y_n)$. On using the Markovian assumptions made for model (1) we have

$$p(\alpha_t|Y_n) = \frac{p(Y_n|\alpha_t)p(\alpha_t)}{p(Y_n)}$$

$$= \frac{p(Y_t|\alpha_t)p(Y^n_{t+1}|\alpha_t)p(\alpha_t)}{p(Y_n)}$$

$$= \frac{p(\alpha_t | Y_t) p(\alpha_t | Y_{t+1}^n) c(Y_n)}{p(\alpha_t)}$$

where $c(Y_n)$ depends only on Y_n and $p(\alpha_t)$ is the marginal density of α_t. This formula has been derived in a wider context by Solo (1989). If, as in the cases considered in this paper, we initialise our Kalman filters by a diffuse density for a_1 then $p(\alpha_t)$ is also diffuse so the estimate of α_t obtained from Y_t using a Kalman filter going forwards in time and the estimate obtained from Y_{t+1}^n using a Kalman filter going backwards in time are effectively independent. Of course, the method only applies when, under the model, backwards travel is as valid as forwards travel, as is the case in the examples considered here.

Our suggestion is that Kalman filters are run through the data in both directions for our Gaussian mixture model, collapsing to a single Gaussian density after each update in the way described for the third approximation to the likelihood in section 3. Each filter automatically produces an estimate of the variance matrix of the estimated state vector at each step. Suppose that the forwards and backwards estimates of a particular element of α_t are f and b and that their estimated variances are v_f and v_b. We then take as our starting value for this element the ordinary weighted average $(f/v_f + b/v_b)/(1/v_f + 1/v_b)$. Although we have been greatly surprised by how good these approximations are, they cannot be recommended as final estimates since they are based on two approximations, first the collapse to a single Gaussian after each update and secondly the weighting of the two estimates by reciprocal variances instead of weighting the two estimates of the state vectors by reciprocals of variance matrices. Since only filtering is involved the process is very fast and in our examples took only a few seconds on our

Sun network. It should be noted that the state transition matrix and the state error variance matrix usually need adjustment for travel in the reverse direction. For densities which are not Gaussian mixtures the same two-filter technique could be employed using an extended Kalman filter in each direction. This idea could be used, for example, to obtain improved starting values for hyperparameter estimation in the exponential-family models considered by Durbin and Koopman (1993).

It is worth observing that the two-filter technique offers a solution to the "k adjacent outlier" problem for state space models considered by Martin and Bruce (1989) for ARIMA models. Using a mixture model, the investigator could track the quasi-variances considered above for the observation error in each direction. When these indicated a group of adjacent possible outliers he could then decide whether to treat them as outliers or as representing a genuine level shift; if the former, it would be straightforward to treat them as missing values in the standard way for state space models. See for example Harvey (1989) sections 3.4.7 and 6.4.1. While this possibility will not be pursued in general in this paper, as an experiment we considered the special case of two adjacent outliers from a slightly different standpoint in one of the examples given in section 6.

A final observation is that although we have concentrated in this paper on the special case of symmetric error densities, it is obvious that the methods could be adapted to deal with non-symmetric densities.

6. Applications to real and simulated series

We now consider the application of the techniques to a number of simulated and real series. Our simulated series are all of 120 observations, intended to represent ten years of monthly data. Our first simulated series is generated by the RW plus noise model $y_t = \mu_t + \varepsilon_t$, $\mu_t = \mu_{t-1} + \xi_t$, where ε_t and ξ_t are Gaussian white noise. Superimposed on this are an outlier at $t = 40$ and a level shift at $t = 80$. Figure 2a shows the original series and the trend as estimated by the standard filter and smoother for the state space model (10) assuming that ε_t and ξ_t are Gaussian. Figure 2b shows the original series and the posterior mode estimate of trend, assuming mixture model (32) for both ε_t and ξ_t with pre-assigned values $\beta = 0.01$ and $\lambda^2 = 100$. In this, as in all the examples considered in this section, hyperparameters have been estimated by approximate maximum likelihood using the third approximation of section 3.

It is evident that a dramatic improvement has been achieved in the handling of both the outlier and the level shift by using the mixture-PME technique. This improvement is reflected by the large difference in the values of the loglikelihood which is -152.0 for the Gaussian model and -100.4 for the mixture model. The number of iterations needed to obtain the PME of the state vector for the mixture model was 6. The convergence criterion was $\max_{t,i} |\alpha_{ti}^{(j)} - \alpha_{ti}^{(j)}| < 10^{-7}$, that is, the maximum change in any component of the state vector over the last iteration is 10^{-7}. We appreciate that this criterion is severe but we found it worked well with our examples; indeed the number of iterations was always between 5 and 9, and we used it for all our illustrations.

Our second simulated series is intended to represent a seasonal series in which there is a shift in trend and in seasonal amplitude as well as an outlier and a level shift. For the seasonal we took a sine wave of period a year whose amplitude was a random walk. For the trend we took the LLT. The underlying series is generated by the model

$$y_t = \mu_t + \gamma_t + \varepsilon_t \qquad \gamma_t = \phi_t \sin \frac{\pi t}{6}$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \xi_t \qquad \phi_t = \phi_{t-1} + \omega_t$$

$$\beta_t = \beta_{t-1} + \zeta_t$$

This was modified by including a trend slope shift at t = 24, a fifty percent amplitude reduction of the seasonal at t = 48, an outlier at t = 72 and a level shift at t = 96. Figure 3a shows the original series and estimate of trend while figure 3b shows the true seasonal and its estimate as obtained by assuming the Gaussian model. We see that the Gaussian method has failed to deal with either the outlier or the level shift satisfactorily and that it produces a slow change in the seasonal instead of indicating an abrupt change at t = 48. Figures 3c and 3d present the corresponding comparisons for the mixture-PME technique which show that it has coped satisfactorily with slope shift , the outlier and the level shift and in addition has responded well to what was an abrupt change in seasonal. The difference between the values of the loglikelihoods was enormous; for the Gaussian model it was -142.3 while for the mixture model it was -9.2. The number of iterations needed for the calculation of the PME of the state vector was 5.

The first real series we consider is monthly retail sales of automobiles in the US from 1977 to 1985. We chose this series because it seemed free of outliers and structural shifts and we wished to illustrate aspects of the behaviour of the mixture model for this type of series. Figure 4a shows the original series and the trend given by the mixture model. (Note that for all the real series in Figures 4,5 and 6 the vertical scale for observations y_t is 1,000 log y_t). Figures 4b and 4c compare the Gaussian and mixtures estimates of the trends and seasonals. As can be seen, the two estimates are close, illustrating that when a series has no outliers or structural shifts, the introduction of the mixture components does not distort the analysis. This is also bourne out by the two values of the loglikelihood which for the Gaussian model was -527.1 while for the mixture model it was very little different at -529.4. In this comparison no adjustment needs to be made for number of parameters fitted since it is the same in both cases. Thus our loglikelihood comparisons are essentially equivalent to AIC comparisons. Figures 4d and 4e compare the quasi-variances of the observation errors with the posterior probabilities of outliers. The two graphs are virtually indistinguishable, illustrating the claim in section 5 that the quasi-variances and posterior probability are more or less equally informative about the presence, or in this case the absence, of outliers.

For figure 5 we took the same automobile sales series as in Figure 4 and added to it outliers of 200 at Sept 78 and July 81 and a level shift of 300 at Aug 83. Figure 5a shows the original series and the mixture model trend estimate. Figure 5b graphs the observation and the level quasi-variances. These graphs demonstrate the power of the quasi-variances as diagnostic pointers to outliers and structural shifts. Figures 5c and 5d compare the seasonally adjusted series and the estimated trend for the Gaussian model and the mixture model respectively for the case of adjacent outliers of 200 and 325 at Sept and Oct 78. For handling adjacent outliers we found it advisable to experiment with proportions in the mixtures. For the present example we found that the mixture $0.98\ N(0,\sigma^2) + 0.02\ N(0,100\sigma^2)$ worked best. It is clear that the mixture model has coped well with the outliers and level shift in all cases whereas the Gaussian model has not. The number of iterations needed for convergence of the PME with the mixture model was 5 for all cases with this series.

Figure 6 refers to US retail sales of liquor from Jan 67 to Feb 89. We chose this series as an example of a series that is generally well-behaved but has a moderate outlier and moderate shift. Figure 6a displays the original series and the Gaussian trend, Figure 6b shows the seasonally adjusted series and the mixture trend and Figure 6c graphs the seasonal. Inspection reveals a level shift at around Jan 84 and an outlier at Sept 85. To highlight these, Figures 6d and 6e graph the seasonally adjusted series and the Gaussian and mixture trends for the period Jan 83 to Dec 86. They show a worthwhile improvement by the mixture model, particularly in dealing with the outlier. The number of iterations for the PME was 9. The loglikelihoods were -1,104.9 for the Gaussian model and -1,130.7 for the mixture model, indicating a substantially better fit for the mixture model.

Summing up our experience based on the analysis of these series, we believe that the mixture model provides an effective automatic technique for handling certain types of outliers and structural shifts. Apart from estimation of the hyperparameters, computation was very fast. We think that further work is needed on choice and development of maximisation routines suited to hyperparameter estimation for models of the kind we have employed. We used mainly the standard maximisation optmisers in the S-PLUS system, which are basically quasi-Newton algorithms, applied to the third approximation to the likelihood function in Section 3. We also experimented with the downhill simplex method and Powell's method mentioned in Section 3, together with the first approximation to the likelihood given in that section; these are capable of working adequately in suitable cases. Finally, we concluded at the end of this work that the two-filter estimate of state was so good as a starting value for the PME iteration that any questions we had about possible complications due to multiple modes in the posterior state density had been effectively resolved.

Acknowledgements. This research was carried out in the time series group of the Statistics Research Division of the US Bureau of the Census. We are grateful to the members of the group, David Findley, William Bell, Matthew Kramer, Brian Monsell, Mark Otto and Bor-Chung Chen for their generous help and hospitality. Durbin's work was supported by an ASA/NSF/Census Research Fellowship. Cordero's work was supported by an MAP/Fullbright Fellowship from the Spanish Government and an ASA/NSF/Census Research Associateship.

References

Alspach,D.L. and Sorenson,H.W. (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. IEEE Trans. Autom. Contr., AC-17, 439-448.

Anderson,B.D.O. and Moore,J.B. (1979). Optimal Filtering. Englewood Cliffs NJ: Prentice-Hall.

Bell,W. (1983). A computer program (TEST) for detecting outliers in time series. 1983 Proc. Bus. and Econ. Section, Amer. Statist. Assoc., 634-639.

Box,G.E.P. and Tiao,G.C. (1973). Bayesian Inference in Statistical Analysis. Reading, Massachussetts:Addison-Wesley.

Bruce,A.G. and Jurke,S.R. (1992). Non-Gaussian seasonal adjustment: X-12 ARIMA versus robust structural models. (To appear).

de Jong,P. (1989). Smoothing and interpolation with the state space model. J. Amer. Statist. Assoc., 84,1085-1088.

Durbin,J. and Koopman,S.J. (1993). Filtering, smoothing and estimation for time series models when the observations come from exponential family distributions. (To appear).

Fahmeir,L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalised linear models. J. Amer. Statist. Assoc., 87, 501-509.

Fahmeir,L. and Kaufmann,H. (1991). On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. Metrika, 38, 37-60.

Harrison,J. and Stevens,C.F. (1971). A Bayesian approach to short-term forecasting. Oper. Res. Quart., 22, 341-362.

Harrison,J. and Stevens,C.F. (1976). Bayesian forecasting (with discussion). J. Roy. Statist. Soc. B, 38 205-247.

Harvey,A.C. (1989). Forecasting, Structural Time series Models and the Kalman Filter. Cambridge:Cambridge University Press.

Koopman,S.J. (1992). Disturbance smoother for state space models. (To appear in Biometrika).

Kitagawa,G. (1989). Non-Gaussian seasonal adjustment. Computers Math. Applic. 18, 503-514.

Kitagawa,G. (1987). Non-Gaussian state-space modelling of nonstationary time series (with discussion). J. Amer. Statist. Assoc. 82, 1032-1063.

Kitagawa,G. (1990). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. Technical report, Institute of Statistical Mathematics, Tokyo.

Martin,R.D. and Bruce,A.G. (1989). Leave-k-out diagnostics for time series (with discussion). J. R. Statist. Soc. B, 51, 363-424.

Martin,R.D. and Raftery,A.E. (1987). Robustness, computation and non-Euclidean models. (Contribution to discussion of Kitagawa (1987)). J. Amer. Statist. Assoc., 82, 1044-1050.

Masreliez,C.J. (1975). Approximate non-Gaussian filtering with linear state and observation relations. IEEE Transactions on Automatic Control. 20, 107-110.

Otto,M.C. and Bell,W.R. (1990). Two issues in time series outlier detection using indicator variables. _1990 Proc. Bus. and Econ. Section, Amer. Statist. Assoc._, 182-187

Peña,D. and Guttman,I. (1998) Baysian approach to robustifying the Kalman filter. _Baysian Analysis of Time Series and Dynamic Models._ Ed. J.C.Spall. New York: Marcel Dekker

Press,W.H., Flannery,B.P., Teukolsky,S.A. and Vetterling,W.T. (1988). _Numerical Recipes in C._ Cambridge: Cambridge University Press.

Rao,C.R. (1973). _Linear Statistical Inference and its Applications._ New York: Wiley.

Solo,V. (1989). A simple derivation of two filter smoothing formula. Technical Report, Department of Electrical and Computer Engineering, Johns Hopkins University.

Whittle,P. (1991). Likelihood and cost as path integrals (with discussion). _J. Roy. Statist. Soc. B_, 53, 505-538.

Young,P., NG,C.N., Lane,K. and Parker,D. (1991). Recursive forecasting, smoothing and seasonal adjustment of non-stationary environmental data. _J. Forecasting._, 10, 57-89.

Fig.1 Weight function for different error models

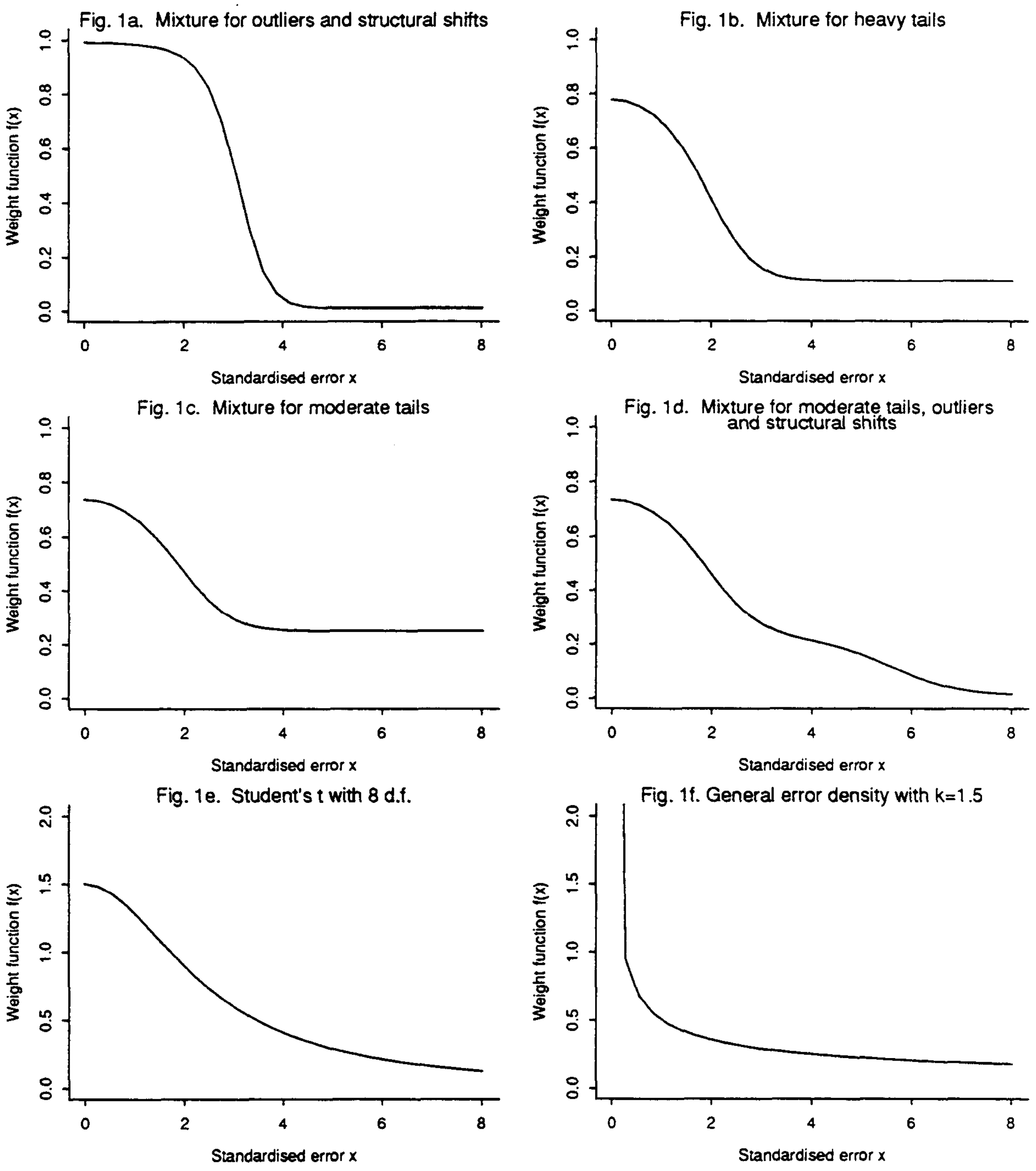


Fig. 1a. Mixture for outliers and structural shifts

Fig. 1b. Mixture for heavy tails

Fig. 1c. Mixture for moderate tails

Fig. 1d. Mixture for moderate tails, outliers and structural shifts

Fig. 1e. Student's t with 8 d.f.

Fig. 1f. General error density with k=1.5

Fig. 2. Performance of Gaussian and mixture models on a simulated series with one outlier and one level shift

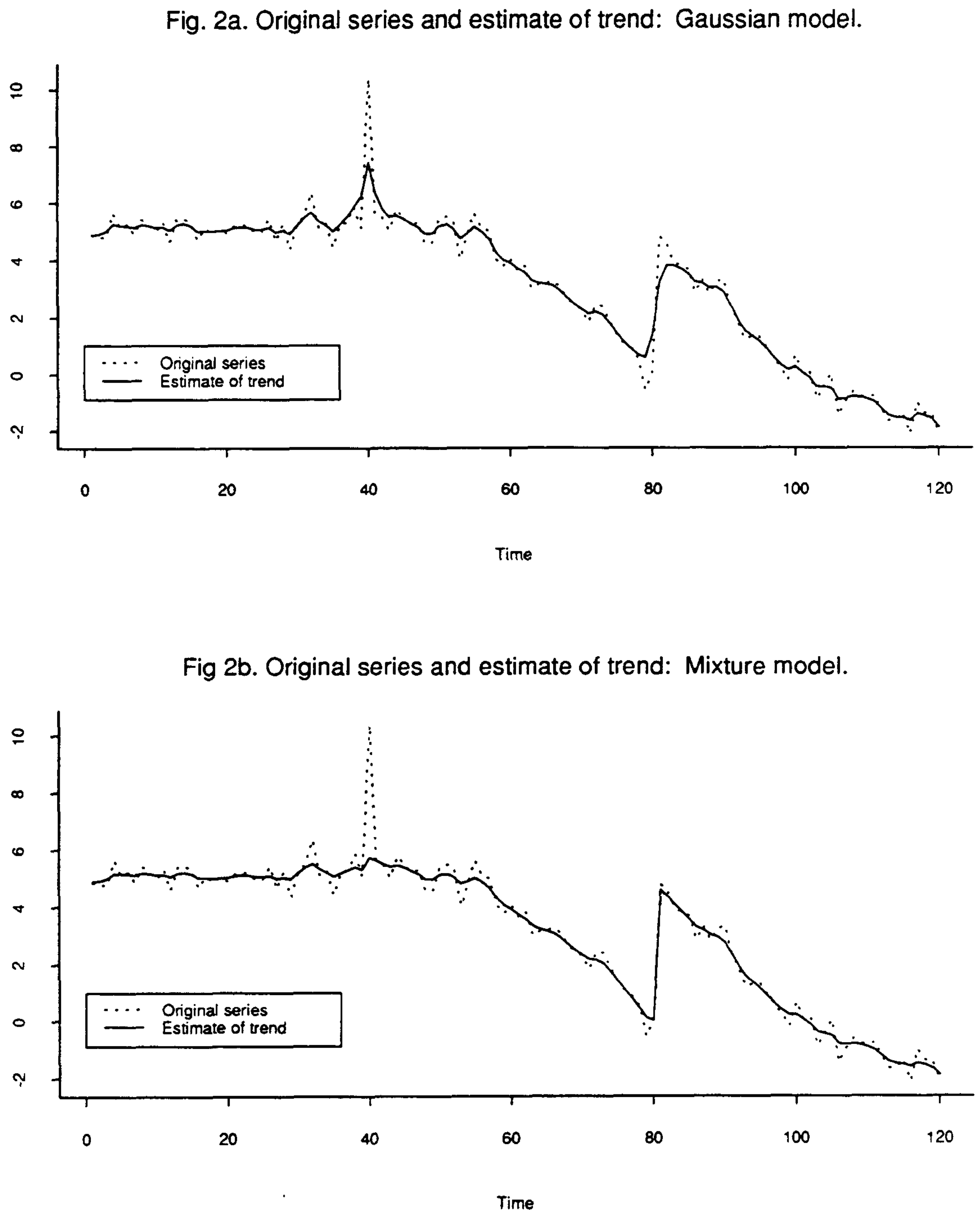


Fig. 2a. Original series and estimate of trend: Gaussian model.

Fig 2b. Original series and estimate of trend: Mixture model.

Fig. 3. Performance of Gaussian and Mixture model on a simulated seasonal time series with structural changes

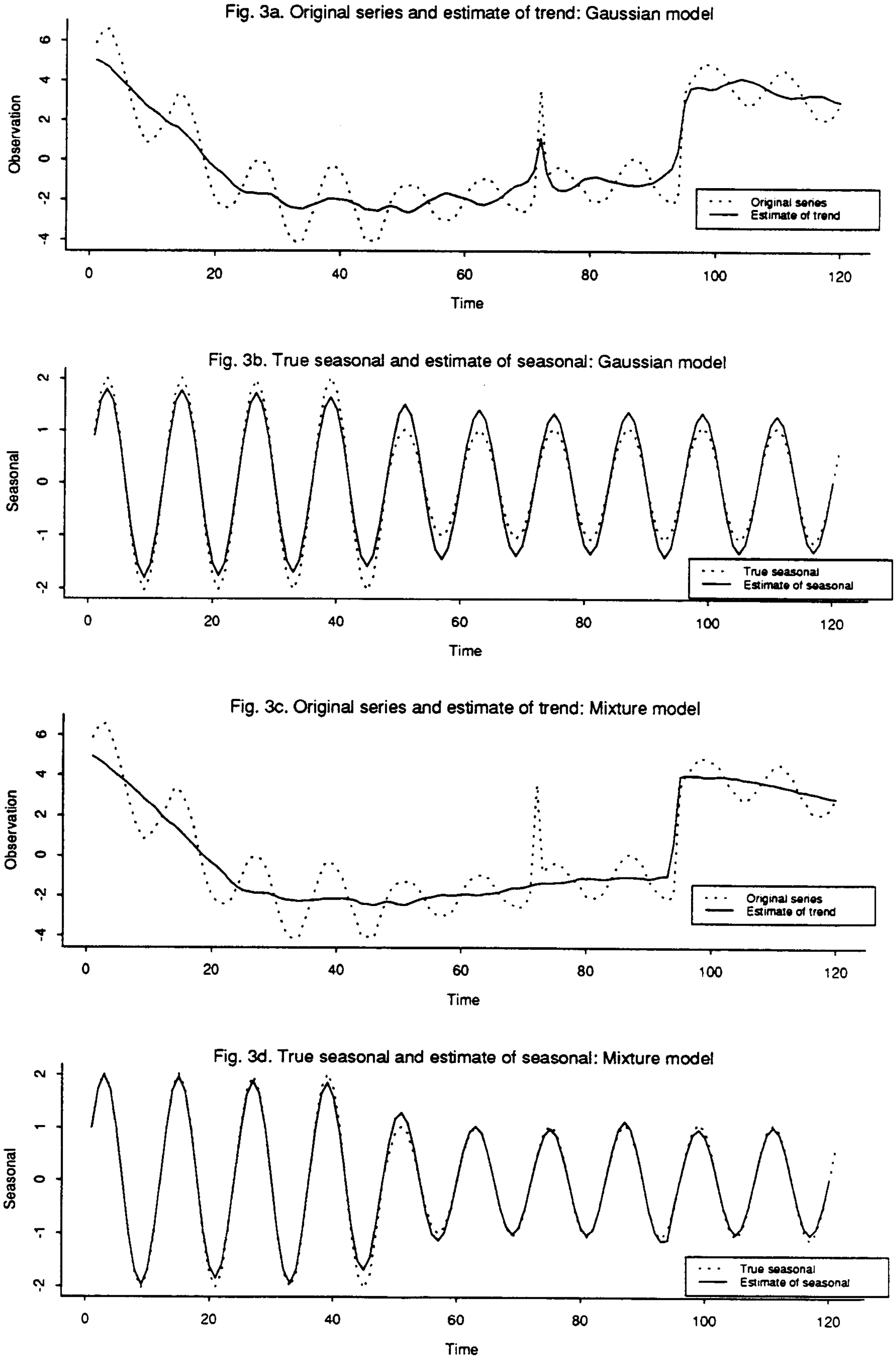


Fig. 3a. Original series and estimate of trend: Gaussian model

Fig. 3b. True seasonal and estimate of seasonal: Gaussian model

Fig. 3c. Original series and estimate of trend: Mixture model

Fig. 3d. True seasonal and estimate of seasonal: Mixture model

Fig. 4. Performance of Gaussian and Mixture models on U. S. retail sales of automobiles 1977-85

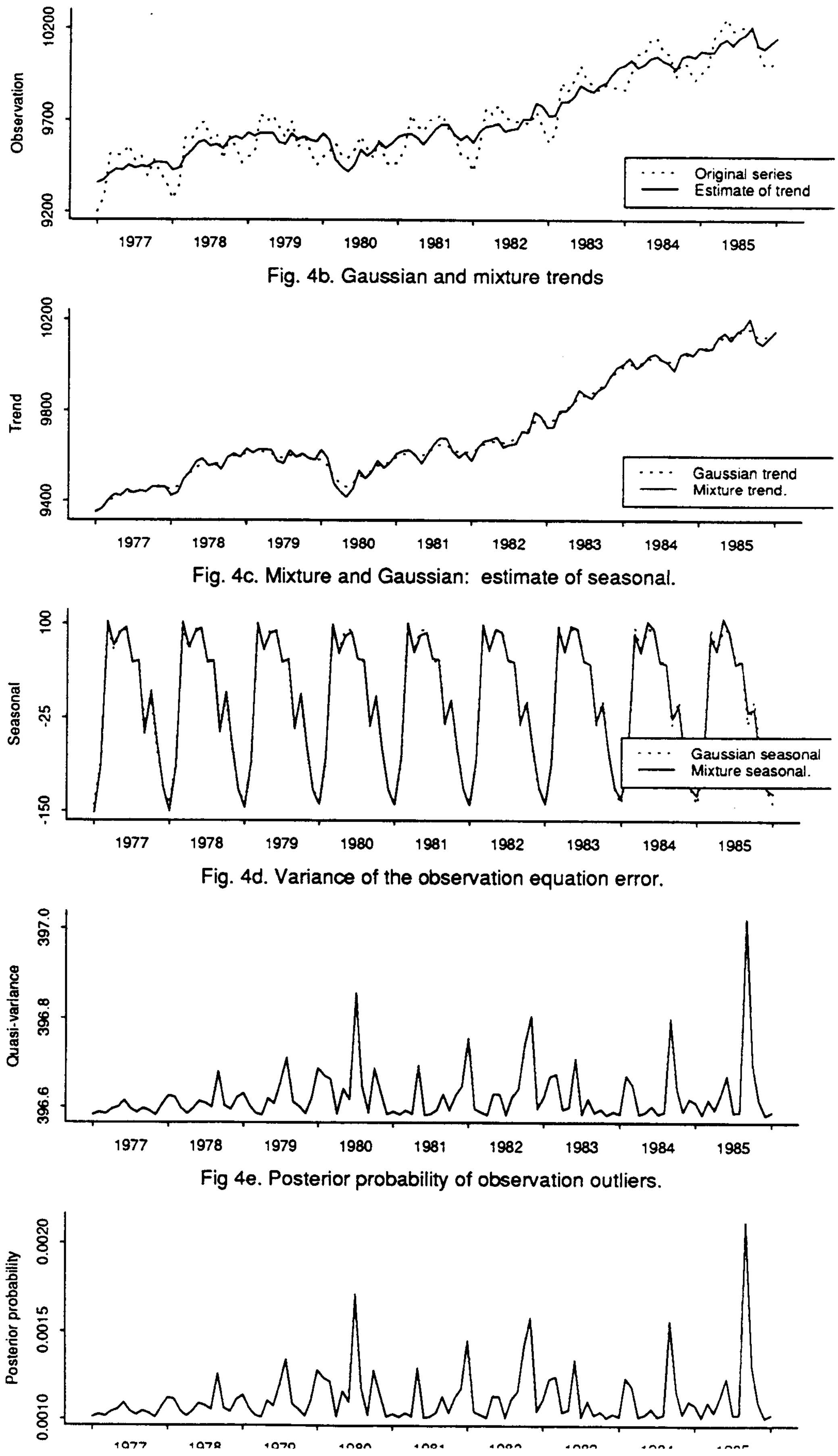


Fig. 4a. Original series and estimate of trend: mixture model

Fig. 4b. Gaussian and mixture trends

Fig. 4c. Mixture and Gaussian: estimate of seasonal.

Fig. 4d. Variance of the observation equation error.

Fig 4e. Posterior probability of observation outliers.

Fig. 5. Performance of Gaussian and Mixture models on U. S. retail sales of automobiles with outliers and level shift superimposed

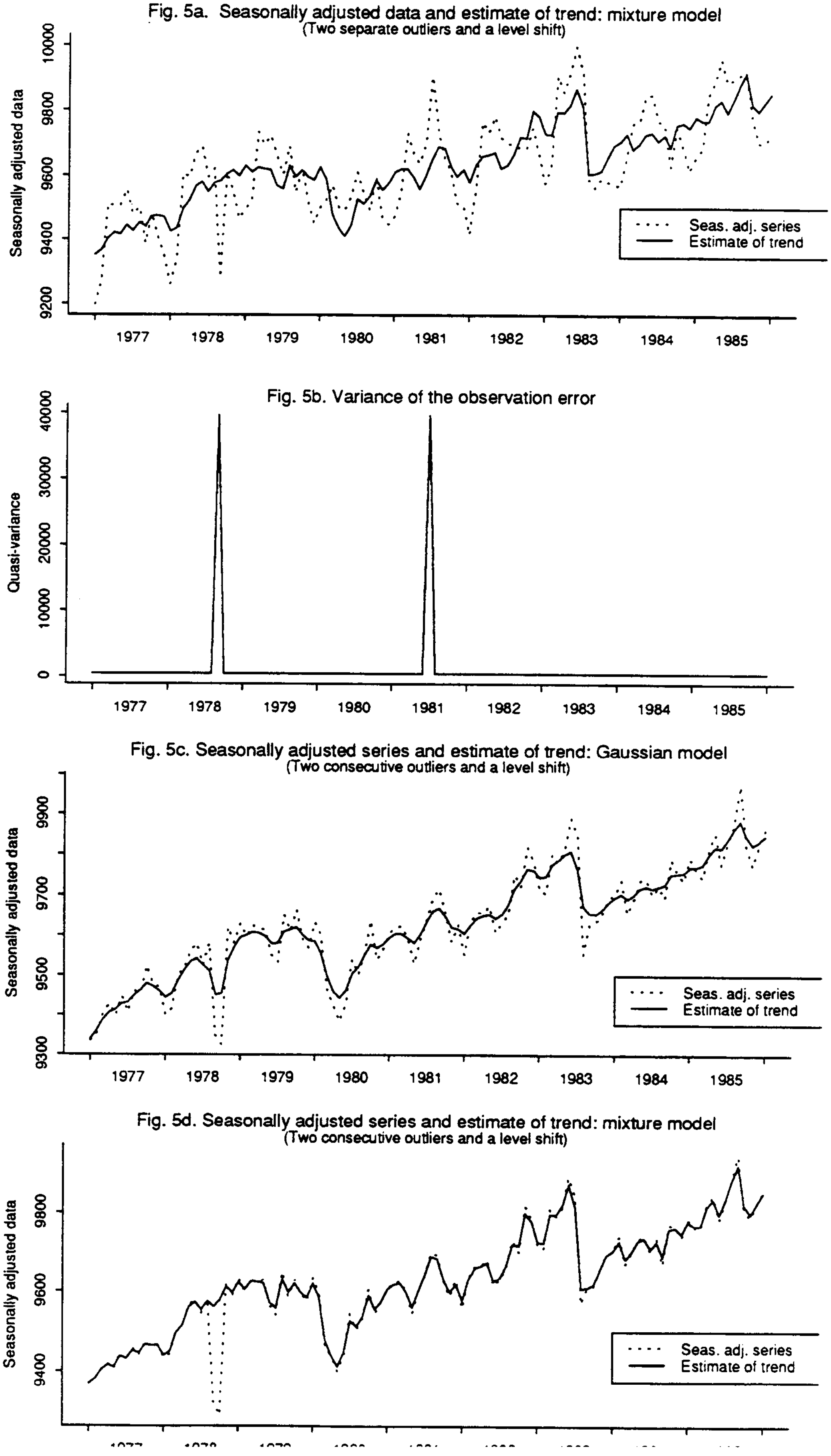


Fig. 5a. Seasonally adjusted data and estimate of trend: mixture model
(Two separate outliers and a level shift)

Fig. 5b. Variance of the observation error

Fig. 5c. Seasonally adjusted series and estimate of trend: Gaussian model
(Two consecutive outliers and a level shift)

Fig. 5d. Seasonally adjusted series and estimate of trend: mixture model
(Two consecutive outliers and a level shift)

Fig. 6. U.S. retail sales of liquor Jan.67-Feb.89

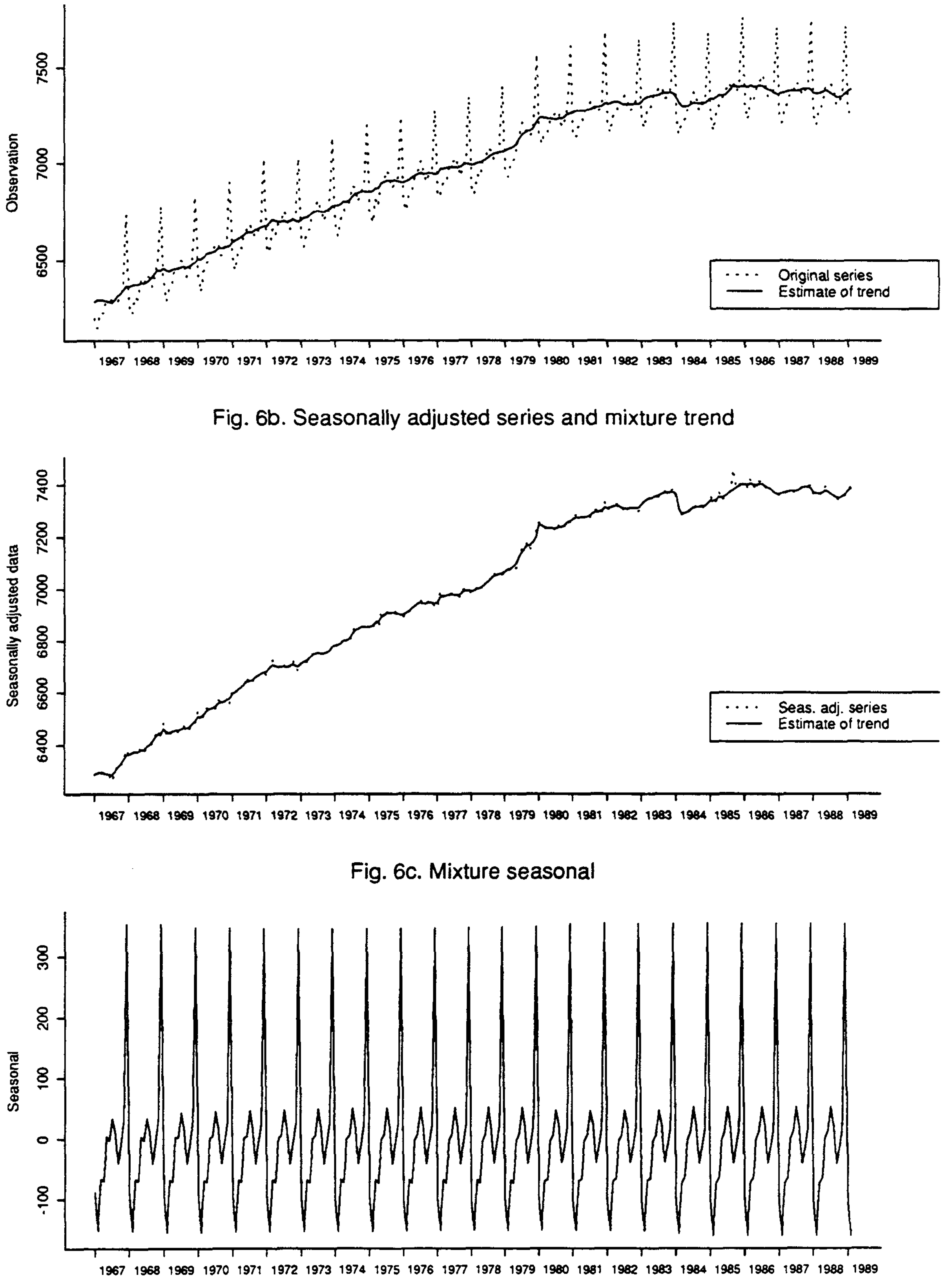


Fig 6a. Original series and Gaussian trend

Fig. 6b. Seasonally adjusted series and mixture trend

Fig. 6c. Mixture seasonal

Fig. 6. U.S. retail sales of liquor Jan.83-Dec.86

Fig. 6d. Seasonally adjusted series and Gaussian trend.

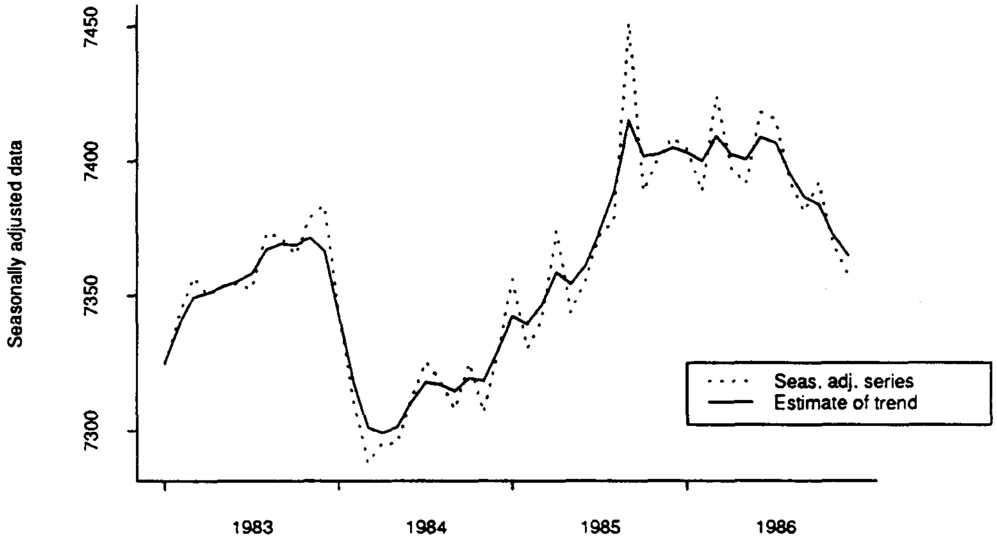
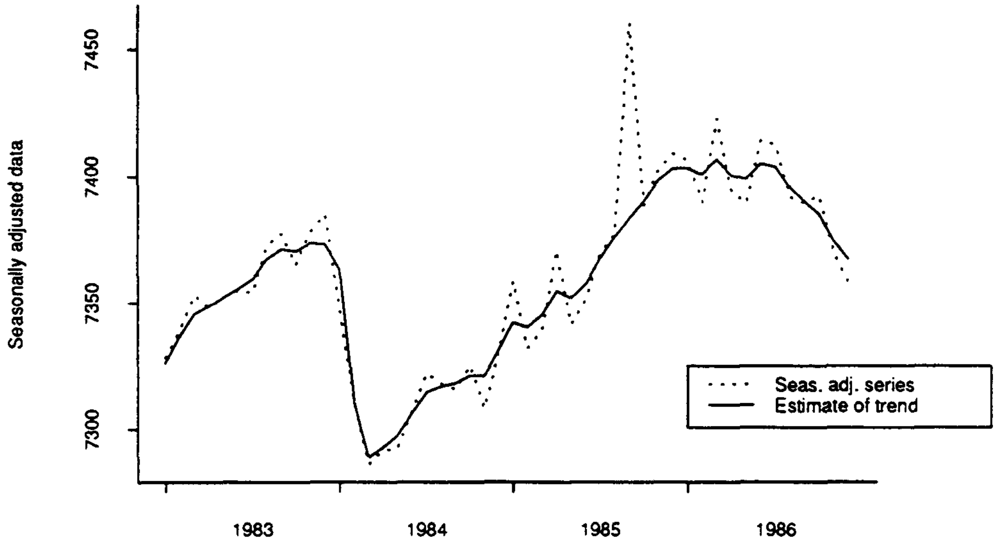


Fig. 6e. Seasonally adjusted series and mixture trend.


Seasonally adjusted data
Seas. adj. series
Estimate of trend