MESSY AND MISTAKEN SEASONAL TIME SERIES

Jeremy Penzer
Department of Statistics, London School of Economics
Houghton Street, London, WC2A 2AE, UK.

August 11, 2000

Abstract

Measurement outliers and level shifts can be identified from the output of the Kalman filter smoother. Departures from an established seasonal pattern are harder to detect. We describe several state space representations for seasonal processes. These representations each display different aberrant behaviour when shocked. Simple, computationally efficient, statistics are proposed as a means of detecting change points in seasonal series. These statistics are also effective in finding anomalies caused by mistakes in data entry. Examples are given including a discussion of the UK new car registration series.

KEYWORDS: Kalman filter; Data entry; Outliers; Smoother; State space form; Structural breaks; Structural time series models.

1 Introduction

Many time series are subject to periodic variation. Series recorded on a monthly or quarterly basis often display annual seasonal patterns. Higher frequency periodic behaviour is also common, for example the daily pattern of electricity consumption. It is reasonable to assume that seasonal patterns evolve slowly over time (Harvey 1989, p. 40). However, a shock may alter the seasonal behaviour of a series suddenly. It is these sudden changes that we are interested in modelling and detecting. A shift in seasonal pattern may be caused by a real event such as a change in legislation, however, unusual behaviour also arises when there are mistakes in the record of the data. The process of scanning or typing data into a machine is not flawless and published series contain errors. Some mistakes are obvious from a plot of the series, but others require more sophisticated detection techniques. In seasonal time series, a missed or repeated data entry leads to a shift in seasonal pattern similar to that observed when the series is shocked.

Early methods for detecting outliers (Fox 1972) are based on estimating parameters associated with interventions in autoregressive models. This approach is adopted for autoregressive-moving average (ARMA) models by many authors; for example Tsay (1986, 1988) and Chang, Tiao and Chen (1988). Willsky and Jones (1976) consider shocks to state space models. This approach is put in a statistical context by Atkinson, Koopman and Shephard (1997) whose key contribution is a score statistic for detecting outliers and structural breaks. Harvey and Koopman (1992) suggest diagnostics based on smoothed error components for structural time series models. General smoother based diagnostics for state space models are given by de Jong and Penzer (1998). Recently a number of authors (McCulloch and Tsay 1993; Carter and Kohn 1994) have proposed schemes in which the both location and size of the any shocks are treated as parameters and estimated simultaneously using Markov chain Monte Carlo. A comprehensive review of messy behaviour in time series is given by Harvey, Koopman and Penzer (1998).

The methodological contribution of this paper is a fast and simple diagnostic procedure for changes in seasonal pattern. We concentrate on the seasonal models proposed by Harvey (1989) and West and Harrison (1997). Applying the results of de Jong and Penzer (1998) to these models suggests a number of plausible statistics, all of which can be generated

from a single run of the Kalman filter smoother. The paper is organized as follows. Section 2 describes the modelling framework based on structural models and state space form. Interventions which arise from a single shock to a seasonal model are described. Simple and computationally efficient statistics for detecting aberrant behaviour based on the output of the Kalman filter smoother are given in section 3 while section 4 establishes the effectiveness of the method proposed by application to several real data sets including the UK car registration figures.

2 Seasonal Models and the State Space Form

Structural time series models (Harvey 1989) provide a convenient framework for modeling seasonality. We concentrate on these models although our methods apply to any process that can be given a Markovian form. Seasonality is treated as one of several components which may include a trend or cycle. Exogenous variables can be incorporated readily.

2.1 Seasonal components

In the basic structural model the seasonal component γ_t is combined with a stochastic trend μ_t to give, for $t = 1, \ldots, n$,

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2),$$

where y_t is the observed series and ε_t is the irregular component. NID is used to denote normally and independently distributed. Normality is usually assumed for the basic structural model although it is not essential for the diagnostics given in this paper. The component μ_t is modeled by a local linear trend,

$$\begin{array}{rcccccll} \mu_{t+1} & = & \mu_t & + & \beta_t & + & \eta_t, & \eta_t \sim \text{NID}(0, \sigma_\eta^2), \\ \beta_{t+1} & = & & & \beta_t & + & \zeta_t, & \zeta_t \sim \text{NID}(0, \sigma_\zeta^2), \end{array}$$

where ε_t, η_t and ζ_t are mutually uncorrelated. We consider three representations of the seasonal component γ_t.

A deterministic seasonal component should sum to zero over the period s,

$$\sum_{j=0}^{s-1} \gamma_{t-j} = 0.$$

This ensures that seasonality is not confounded with the trend. We allow for gradual change in seasonal pattern by adding a disturbance term to the sum of seasonal effects yielding

$$\gamma_t = -\gamma_{t-1} - \ldots - \gamma_{t-s+1} + \omega_t, \qquad \omega_t \sim \text{NID}(0, \sigma_\omega^2),$$

where ω_t is not correlated with ε_t, η_t or ζ_t. This is known as *dummy variable* stochastic seasonality.

Harrison and Stevens (1976) and West and Harrison (1997) put forward a form of stochastic seasonality in which each seasonal component is modelled as a random walk. They refer to this as the *form-free* approach. The seasonal component at time t is given by

$$\gamma_t = \gamma_{1,t},$$

where

$$\begin{aligned} \gamma_{j,t} &= \gamma_{j+1,t-1} + \omega_{j,t}, \quad \text{for } j = 1, \ldots, s-1, \\ \gamma_{s,t} &= \gamma_{1,t-1} + \omega_{s,t}. \end{aligned}$$

The error components $\omega_{1,t}, \ldots, \omega_{s,t}$ are correlated, in fact we impose the zero sum constraint

$$\sum_{j=1}^{s} \omega_{j,t} = 0. \tag{1}$$

for all t. As with the dummy approach, this constraint prevents confusion between seasonal effects and effects which should be attributed to other components.

Sine and cosine functions can be used to capture a deterministic seasonal pattern. Adding a random disturbance to these components leads to the trigonometric form of stochastic seasonality. This is expressed as

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{j,t},$$

where each $\gamma_{j,t}$ is generated by

$$\begin{pmatrix} \gamma_{j,t} \\ \gamma^*_{j,t} \end{pmatrix} = \begin{pmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{pmatrix} \begin{pmatrix} \gamma_{j,t-1} \\ \gamma^*_{j,t-1} \end{pmatrix} + \begin{pmatrix} \omega_{j,t} \\ \omega^*_{j,t} \end{pmatrix}, \qquad j = 1, \ldots, [s/2],$$

$\lambda_j = 2\pi j/s$ is frequency in radians, ω_t and ω^*_t are two mutually uncorrelated white noise disturbances with zero means and common variance σ^2_ω and $\gamma^*_{j,t}$ is a dummy element. The notation $[s/2]$ refers to the integer part of $s/2$. For s even, $[s/2] = s/2$ and the component at $j = s/2$ collapses to

$$\gamma_{j,t} = \gamma_{j,t-1}\cos\lambda_j + \omega_{j,t}. \tag{2}$$

Without the disturbance terms this model will give the same deterministic pattern as the dummy variable model. However, for stochastic seasonality, the trigonometric representation allows the seasonal pattern to evolve more smoothly.

2.2 Seasonal models in state space form

All linear time series models have a state space representation. This representation relates the disturbance vector $\{\epsilon_t\}$ to the observation vector $\{y_t\}$ via a Markov process $\{\alpha_t\}$. A convenient expression of the state space form is, for $t = 1, \ldots, n$,

$$\begin{aligned} y_t &= Z_t\alpha_t + G_t\epsilon_t, \\ \alpha_{t+1} &= T_t\alpha_t + H_t\epsilon_t, \end{aligned} \tag{3}$$

where $\epsilon_t \sim (0, I)$ are the disturbances and $\alpha_1 \sim (a_1, P_1)$ is the initial state. The disturbances are mutually uncorrelated white noise variables which are also uncorrelated with the initial state. The system matrices Z_t, T_t, G_t and H_t are deterministic and any unknown elements which they contain are estimated by maximum likelihood. For a univariate series with $m \times 1$ state vector, α_t, and $k \times 1$ vector of disturbances, ϵ_t, the system matrices Z_t, T_t, G_t and H_t have dimensions $1 \times m$, $m \times m$, $1 \times k$ and $m \times k$ respectively.

The basic structural model has a time invariant state space representation whose exact form depends on the choice of seasonal component. Some illustrative examples are given below. Using a dummy seasonal model for quarterly data, the third element of the state vector represents the current seasonal effect. The matrices in the state space representation are

$$Z = \begin{pmatrix} 1 & 0 & \vdots & 1 & 0 & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} \sigma_\varepsilon & 0 & 0 & \vdots & 0 \end{pmatrix},$$

$$T = \begin{pmatrix} 1 & 1 & \vdots & 0 & \cdots & 0 \\ 0 & 1 & \vdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \vdots & -1 & -1 & -1 \\ \vdots & \vdots & \vdots & 1 & 0 & 0 \\ 0 & 0 & \vdots & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} 0 & \sigma_\eta & 0 & \vdots & 0 \\ 0 & 0 & \sigma_\zeta & \vdots & 0 \\ \cdots & \cdots & \cdots & \vdots & \cdots \\ 0 & \cdots & 0 & \vdots & \sigma_\omega \\ \vdots & & \vdots & \vdots & 0 \\ 0 & \cdots & 0 & \vdots & 0 \end{pmatrix}. \qquad (4)$$

Note that the seasonal component does not interact with with any of the other components in the model. This is reflected by the block diagonal structure of the matrix T. An immediate benefit of this observation is that it allows us to concentrate solely on the seasonal parts of the state space representation. For efficiency, the sub-matrices of Z, T and H, which are involved respectively in selecting, propagating and disturbing the seasonal effects, are denoted Z_s, T_s and H_s. In the quarterly dummy seasonal case $Z_s = (\,1 \quad 0 \quad 0\,)$, $H_s = (\,\sigma_\omega \quad 0 \quad 0\,)'$ and T_s is the 3×3 lower diagonal block of T given in (4).

As with the dummy variable representation, the form-free approach takes the current seasonal component to be the first seasonal element of the state vector, however, there are now s seasonal elements. Thus, for a quarterly model, $Z_s = (\,1 \quad 0 \quad 0 \quad 0\,)$. The seasonal elements are permuted at each time step to ensure that the first is the correct effect for the current time point. For quarterly data this is achieved by setting

$$T_s = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

An additional feature which distinguishes the form-free representation from the dummy seasonal approach is the presence of s seasonal disturbance terms, $\omega_{1,t}, \ldots, \omega_{s,t}$, at each time step. Taking $\omega_t = (\,\omega_{1,t} \quad \ldots \quad \omega_{s,t}\,)'$, the zero sum constraint (1) implies that every row and column of the variance matrix $\mathrm{var}(\omega_t)$ must sum to zero. This restriction is satisfied by taking $\mathrm{var}(\omega_t) = \sigma_\omega^2(sI - ii')$ where i is an $s \times 1$ vector of ones.

Using a trigonometric representation of seasonality, the matrix T_s has a block diagonal

structure. For example, with monthly data,

$$T_s = \begin{pmatrix} C_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & C_5 & 0 \\ 0 & \cdots & 0 & -1 \end{pmatrix},$$

where

$$C_j = \begin{pmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{pmatrix},$$

and $\lambda_j = 2\pi j/12$. The seasonal part of the state consists of the elements

$$\begin{pmatrix} \gamma_{1,t} & \gamma^*_{1,t} & \gamma_{2,t} & \cdots & \gamma_{4,t} & \gamma^*_{4,t} & \gamma_{5,t} \end{pmatrix}',$$

and the constraint, $\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{j,t}$, is satisifed by taking

$$Z_s = \begin{pmatrix} 1 & 0 & 1 & \ldots & 1 & 0 & 1 \end{pmatrix}.$$

The matrices G_s and H_s have the obvious forms.

2.3 Interventions generated by shocks to seasonal models

De Jong and Penzer (1998) establish that many useful intervention structures correspond to a single shock to the transition equation of a state space model. Suppose that we have fitted a model of the form given by (3). To test whether this model captures the behaviour of the data adequately we propose the alternative, for $t = 1, \ldots, n$,

$$\begin{aligned} y_t &= X_t\delta + Z_t\alpha_t + G_t\epsilon_t, \\ \alpha_{t+1} &= W_t\delta + T_t\alpha_t + H_t\epsilon_t, \end{aligned} \tag{5}$$

where X_t and W_t are regression variables which determine how shocks enter the system and δ is a parameter measuring shock magnitude. We confine our interest to simple interventions, that is interventions which can be represented by the introduction of a single shock. For a simple intervention with origin i, $X_t = 0$ and $W_t = 0$ for $t \neq i$.

The shape or *signature* of an intervention depends on the model and on the way in which the shock enters the system. For a simple intervention, de Jong and Penzer (1998) show

that the additive effect of a shock at time i on point t of the series is $D_t(i)\delta$ where

$$D_t(i) = \begin{cases} 0, & t = 1, \ldots, i-1, \\ X_i, & t = i, \\ Z_t T_{t-1,i+1} W_i, & t = i+1, \ldots, n, \end{cases} \tag{6}$$

$T_{j,i+1} = T_j \ldots T_{i+1}$ for $j > i$, $T_{i,i+1} = I$ and $T_{j,i+1} = 0$ for $j < i$. For a local linear trend, $X_i = 1$, $W_i = 0$ models a single outlying value while $X_i = 0$, $W_i = (\,1 \quad 0\,)'$ corresponds to a level shift. The interventions arising from shocking seasonal components are somewhat more complicated.

We return to the models described in section 2.2 for examples. These models are all time invariant and have the property that $T_s^s = I$. Our only concern is changes in seasonal pattern so we can, without loss of generality, consider models which do not contain any other components, that is $T = T_s$. For a quarterly dummy seasonal model

$$ZT = (\,-1 \quad -1 \quad -1\,), \qquad ZT^2 = (\,0 \quad 0 \quad 1\,),$$
$$ZT^3 = (\,0 \quad 1 \quad 0\,), \qquad ZT^4 = (\,1 \quad 0 \quad 0\,).$$

The shape of the intervention is determined by the choice of W_i, that is the way in which the shock enters the system. For example, if we shock the first seasonal element of the state vector then $W_i = (\,1 \quad 0 \quad 0\,)'$ and

$$D_t(i) = \begin{cases} 0 & \text{for } t = 1, \ldots, i, \\ 1 & \text{for } t = i+1,\ i+5, \ldots, \\ -1 & \text{for } t = i+2,\ i+6, \ldots, \\ 0 & \text{for } t = i+3,\ i+7, \ldots, \\ 0 & \text{for } t = 1+4,\ i+8, \ldots. \end{cases}$$

This corresponds to a shift in one season, compensated for immediately by a shift in the opposite direction in the following season. Figure 1 shows the signature along with those for a shock to the second seasonal element, $W_i = (\,0 \quad 1 \quad 0\,)'$, third seasonal element $W_i = (\,0 \quad 0 \quad 1\,)'$ and all seasonal elements simultaneously $W_i = (\,1 \quad 1 \quad 1\,)'$.

In the dummy variable representation the zero sum constraint is incorporated into the transition matrix. By design, the seasonal intervention effects sum to zero over the period. For form-free seasonality this is no longer true, in fact, in the quarterly case,

$$ZT = (\,0 \quad 1 \quad 0 \quad 0\,), \qquad ZT^2 = (\,0 \quad 0 \quad 1 \quad 0\,),$$

$$ZT^3 = (\,0 \quad 0 \quad 0 \quad 1\,), \qquad ZT^4 = (\,1 \quad 0 \quad 0 \quad 0\,).$$

Thus, if a shock is applied to the first seasonal element, $W_i = (\,1 \quad 0 \quad 0 \quad 0\,)'$, the signature is

$$D_t(i) = \begin{cases} 1 & \text{for } i+1,\ i+5, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Shocks to each of the other seasonal elements will have the same signature but shifted in time. For example, shocking the fourth will produce an intervention which is non-zero for the first time at $t = i + 4$.

A trigonometric seasonal component is a sum of harmonics with frequencies $2\pi j/s$ for $j = 1, \ldots, [s/2]$. The harmonics are represented in the transition matrix T via the sub-matrices C_j each of which has the property

$$C_j^k = \begin{pmatrix} \cos k\lambda_j & \sin k\lambda_j \\ -\sin k\lambda_j & \cos k\lambda_j \end{pmatrix},$$

where $\lambda_j = 2\pi j/s$ and k is a positive integer. Thus, by appropriate choice of W_i, we can determine the combination of frequencies in the intervention. For example, the trigonometric model for quarterly data has

$$Z = (\,1 \quad 0 \quad 1\,) \text{ and } T = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

and so

$$ZT = (\,0 \quad 1 \quad -1\,), \qquad ZT^2 = (\,-1 \quad 0 \quad 1\,),$$
$$ZT^3 = (\,0 \quad -1 \quad -1\,), \qquad ZT^4 = (\,1 \quad 0 \quad 1\,).$$

A shock to either or both of the first two seasonal elements will set up an intervention which has the fundamental frequency $2\pi/s$ whereas shocking the final component results in a signature with the Nyquist frequency π.

2.4 Shock designs for given intervention structures

Equation (6) indicates how the signature, $D(i) = (\,D_1(i) \quad \cdots \quad D_n(i)\,)'$, can be generated from a given shock design, $(X_i,\ W_i)$. In this section we demonstrate how, given the intervention signature, we can calculate the corresponding shock design.

Assume that the intervention is seasonal with arbitrary origin i so $D_t(i) = D_{t+s}(i)$ for $t > i$. As in the previous section, seasonal components do not interact with other elements of the state so, without loss of generality, we take $T = T_s$. The design vector W_i determines how the shocks enter the seasonal part of the model. If there are m seasonal elements, we need to evaluate an $m \times 1$ design vector W_i. From equation (6)

$$ZT^{t-i-1}W_i = D_t(i), \qquad t > i. \tag{7}$$

We need m such equations to calculate the m unknowns in W_i. Stacking (7) for $t = i+1, \ldots, i+m$ yields

$$\Lambda_i W_i = D_*(i),$$

where Λ_i is the known $m \times m$ matrix $(\, Z' \quad (ZT)' \quad \cdots \quad (ZT^{m-1})' \,)'$ and $D_*(i)$ is the known $m \times 1$ vector $(\, D_{i+1}(i)' \quad \cdots \quad D_{i+m}(i)' \,)'$ defined by the signature. Provided that the model is observable, Λ_i is non-singular and

$$W_i = \Lambda_i^{-1} D_*(i). \tag{8}$$

Thus, given a model and the required intervention shape, the appropriate shock design can be constructed.

Suppose, for example, we suspect that there has been a switch between two adjacent quarters. Using the trigonometric model $m = 3$ and we could take

$$D_*(i) = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix},$$

to represent a switch. We construct Λ_i directly from the transition matrix, in fact

$$\Lambda_i = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \Lambda_i^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

Now (8) implies that using the design

$$W_i = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix},$$

will produce an intervention with the desired properties.

3 Smoother Based Intervention Statistics

The statistical treatment of state space models is based on the Kalman filter and the associated smoother. We refer to this combination as the Kalman filter smoother (KFS). The KFS can be used to construct the likelihood (Harvey and Phillips 1979), evaluate the score vector of the parameters (Koopman and Shephard 1992), estimate unobserved components (Koopman 1993) and compute diagnostic statistics for detecting structural changes (Harvey and Koopman 1992; De Jong and Penzer 1998). All models which have a Markovian representation can be handled using the KFS, however for ARIMA models the smoother output is hard to interpret.

3.1 Kalman filter smoother

In a Gaussian state space model, the Kalman filter evaluates the minimum mean squared estimator (MMSE) of the state vector α_{t+1} given observations $Y_t = \{y_1, \ldots, y_t\}$, denoted $a_{t+1} = E(\alpha_{t+1}|Y_t)$. The variance matrix $P_{t+1} = \text{var}(\alpha_{t+1}|Y_t)$ is also computed. The Kalman filter is, for $t = 1, \ldots, n$,

$$\begin{aligned} v_t &= y_t - Z_t a_t, & F_t &= Z_t P_t Z_t' + G_t G_t', \\ & & K_t &= (T_t P_t Z_t' + H_t G_t')\, F_t^{-1}, \\ a_{t+1} &= T_t a_t + K_t v_t, & P_{t+1} &= T_t P_t L_t' + H_t J_t', \end{aligned} \tag{9}$$

where $L_t = T_t - K_t Z_t$ and $J_t = H_t - K_t G_t$. The derivation of the Kalman recursions can be found in Anderson and Moore (1979) and Harvey (1989). There is an extensive literature on filter initialisation for non-stationary models; for example de Jong (1988, 1991), Ansley and Kohn (1985, 1990), Snyder and Saligari (1996) and Koopman (1997). The one-step ahead prediction error of the observation vector is $v_t = y_t - E(y_t|Y_{t-1})$ with covariance matrix $F_t = \text{var}(y_t|Y_{t-1}) = \text{var}(v_t)$. The quantities v_t are referred to as *innovations*.

Papers by de Jong (1988, 1989), Kohn and Ansley (1989) and Koopman (1993) lead to a smoothing algorithm from which estimators based on the full sample Y_n can be computed. Smoothing takes the form of a backwards recursion, $t = n, \ldots, 1$,

$$u_t \;=\; F_t^{-1} v_t - K_t' r_t, \qquad M_t = F_t^{-1} + K_t' N_t K_t,$$

$$r_{t-1} = Z_t' u_t + T_t' r_t, \qquad N_{t-1} = Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \tag{10}$$

where $r_n = 0$ and $N_n = 0$. These recursions require the filter output v_t, F_t and K_t to be stored, for $t = 1, \ldots, n$. The u_t are referred to as *smoothations*. As we will see in subsection 3.2, u_t and r_t play a pivotal role in the construction of diagnostic tests for changes in seasonal pattern.

3.2 Statistics to detect change in seasonal patterns

The approach we adopt to detecting departures from a null time series model is based on fitting an intervention at each point of the series in turn. Functions of the estimated magnitude of the intervention are used to suggest the location of any unusual behaviour. It is hard to construct formal tests on the basis of these statistics. In general, the statistics are serially correlated and a simultaneous testing problem arises. However, in the spirit of Brown, Durbin and Evans (1975), a plot of the statistics against time provides useful diagnostic information.

De Jong and Penzer (1998) establish that the generalised least squares (GLS) esimate of the magnitude, δ_i, of a simple intervention at point i is given by

$$\hat{\delta}_i = S_i^{-1} s_i,$$

where

$$s_i = X_i' u_i + W_i' r_i, \qquad S_i = X_i' F_i^{-1} X_i + Q_i' N_i^{-1} Q_i, \tag{11}$$

and $Q_i = W_i - K_i X_i$. The quantities u_i, r_i, F_i and N_i are all generated by applying the KFS to the null model (3). This allows us to compute diagnostic statistics based on the estimated magnitude of an intervention at each point of the series from null model KFS output. The analogue of the t-statistic used to test the significance of explanatory variables in regression is

$$\tau_i = S_i^{-\frac{1}{2}} s_i.$$

Another useful diagnostic is the quadratic form

$$\tau_i^2 = s_i' S_i^{-1} s_i,$$

which, given the model parameters, is approximately χ^2_p where p is the rank of S_i.

For a local linear trend, outliers and structural breaks can be represented by single shocks in the measurement and first component of the transition equations respectively. Thus, result (11) indicates that the smoothations, u_i, and the first component of the vector r_i, will be key quantities in detecting these types of departure. Practical examples can be found in Harvey, Koopman and Penzer (1998) and de Jong and Penzer (1998).

For seasonal models, many different signatures can be generated by shocking different combinations of seasonal elements. In fact, under the assumptions given in subsection 2.4, we can generate any seasonal intervention by choice of W_i and hence compute statistics to detect any seasonal intervention using result (11). However, in practice we may not have a clear idea of the shape of intervention we should impose. A sensible first approach can be constructed by analogy to the level shift case. The elements of the vector r_i corresponding to the seasonal part of the model may provide useful diagnostic information for shifts in seasonal pattern. For a univariate model this suggests the statistics $N_{i,kk}^{-1/2} r_{i,k}$ or $N_{i,kk}^{-1} r_{i,k}^2$, where $r_{i,k}$ is the kth element of the vector r_i, $N_{i,kk}$ is the kth diagonal entry of the matrix N_i and k indexes the seasonal elements of the state.

The statistics described above may be useful when the signature generated by the shock to an individual element of the seasonal component is a good representation of departures in the observed series. However, the approach provides an abundance of diagnostic information which is not always readily interpretable. For example, in an annual dummy seasonal model with monthly data, 11 diagnostic series would be generated. A single diagnostic test for shifts in seasonal pattern is desirable. If every element of the seasonal component is shocked independently, that is $W_i = I$, the resulting τ^2 statistic is

$$\tau_i^{2*} = r_i' N_i^{-1} r_i. \tag{12}$$

De Jong and Penzer (1998) show that statistics of this form are maximal. In the seasonal context this implies that (12) provides an upper bound on all statistics generated by shocks to seasonal elements of the state space form. Another candidate for a single statistic arises by applying a single shock to all seasonal elements simultaneously, that is $W_i = (\,1 \quad \ldots \quad 1\,)'$, a vector of 1s of the same length as the number of seasonal elements

in the state. Result (11) now yields the statistic

$$\tau_i^{2\dagger} = \left(\sum_k \sum_j N_{i,kj} \right)^{-1} \left(\sum_k r_{i,k} \right)^2, \tag{13}$$

where $N_{i,kj}$ is the element on row k and column j of the matrix N_i.

Calculating the statistics given by (12) and (13) exactly would involve storage and, for τ^{2*}, inversion of the matrix N_i. For models with a large number of seasonal elements, although not infeasible, this becomes computationally cumbersome. In our experience, the matrix N_i is close to being diagonal which suggests the approximations

$$\tau_i^{2*} \approx \sum_k N_{i,kk}^{-1} r_{i,k}^2, \quad \text{and} \quad \tau_i^{2\dagger} \approx \left(\sum_k N_{i,kk} \right)^{-1} \left(\sum_k r_{i,k} \right)^2.$$

These approximations can be computed readily and in most instances will be near numerical equivalence to the exact statistics.

4 Practical Examples

Diagnostic statistics are best assessed on the basis of their performance when applied to real data sets. By demonstrating that our statistics detect known events, we hope to establish their usefulness as diagnostics. All computations are carried out using STAMP (Koopman, Harvey, Doornik, and Shephard 1995), Ox (Doornik 1996) and Ssfpack (Koopman, Shephard, and Doornik 1998).

4.1 Marriages data

Figure 2 shows the number of marriages (thousands of registrations) in the United Kingdom from quarter 1 of 1965 to quarter 4 of 1970 taken from the UK Monthly Digest of Statistics (West and Harrison 1997). Fitting a basic structural model with quarterly dummy seasonality yields a model in which all of the hyperparameters are zero except that for seasonality, $\sigma_\omega = 8.363$. The τ statistics associated with shocks to the first and third seasonal elements are given in figures 3(a) and 3(b) respectively. Note that 3(b) is just a

shifted reflection of 3(a), behaviour which could be anticipated by inspection of plots 1(c) and 1(a). Figure 3(a) shows a marked low at the fourth quarter of 1968. Recalling the delay between introduction of a shock and its impact on the series (subsection 2.3), this suggest a change in seasonal pattern starting in quarter 1 of 1969. The signature in figure 1(a) indicates that this change is a drop in the first quarter of each year matched by a rise in the second quarter. This result is confirmed by figure 4 which shows τ_i^{2*} and $\tau_i^{2\dagger}$, although the exact location and nature of the intervention are harder to identify using these statistics.

The alteration in seasonal pattern has a historic explanation. Up to the end of 1968 the law allowed couples to claim the married persons tax allowance retrospectively for the entire year in which they were married. As the tax year runs from April to March this provides an incentive to marry in the first quarter of the calendar year, a period which would otherwise be fairly unappealing given the climate in the UK. In 1969 this rule was abolished causing a marked drop in quarter 1 registrations and a compensatory rise in the second quarter as the financial disincentive for a spring wedding had been removed. Fitting an explanatory variable to take into account the change in seasonal pattern yields a marked drop in the variability of the seasonal component, $\sigma_\omega = 3.3161$. This explanatory variable is highly significant with a τ (approximate t) value of 8.417 once hyperparameters are re-estimated.

The marriages data illustrate another prediction of our analysis of signatures in subsection 2.3. In free-form seasonality shocks to each of the seasonal elements produce the same signatures just shifted in time. This is reflected in the corresponding τ statistics as illustrated in figure 5 for shocks to the first (a) and fourth (b) elements of a form-free model for the marriages data.

Although this example illustrates the effectiveness of the proposed statistics, it could be argued that the marriages data is very simple and the shift is seasonal pattern is apparent by inspection of the data plot, figure 2. With this in mind, we turn to examples in which the aberrant seasonal behaviour is less obvious.

4.2 Car registration data

Figure 6 shows the monthly average (thousands) of new cars registered in Great Britain each quarter from the first quarter of 1955 to the first quarter of 1997. The figures are those provided by the Office for National Statistics (ONS) in table 4.9 of Lewin (1997). A basic structural model with trigonometric seasonality fits the data well with hyperparameter estimates given in table 1. No sudden changes in seasonal pattern are immediately apparent from the data plot. However, the τ^{2*} statistics in figure 7 give a very strong indication of unusual behaviour at the beginning of 1996 with the peak in the second quarter of 1996. Closer inspection reveals that the figures for quarters 1 and 2 of 1996 are identical, both being 175.4. This seems implausible and is in fact a typographic error in the data record. The Department of the Environment, Transport and the Regions provided the true figures. Hyperparameter estimates for these corrected data are given in table 1 and the τ^{2*} statistics in figure 8. The peak towards the end of 1979 is due to the fact that sales in quarter 2 were higher than in quarter 3 in that year which may have been caused by an increase in car registration in the immediate aftermath of the 3rd May 1979 general election. This represents a temporary change in seasonal pattern; see section 5.

	Original	Corrected
Irregular σ_ε	8.785	5.934
Level σ_η	7.262	7.325
Slope σ_ζ	0.0	0.0
Seasonal σ_ω	2.374	1.483

Table 1: Hyperparameter estimates for basic structural model fitted to quarterly car registration data.

The monthly figures for new car registrations are provided in electronic form by the ONS via Manchester Information Datasets and Associated Services (MIDAS). Figure 9 shows the logged series from January 1955 to September 1991 once two extreme outliers have been corrected using the quarterly data. The τ^{2*} statistics for a basic structural model with trigonometric seasonality are given in figure 10. Two features of the plot stand out; a period of large τ^{2*} values between mid-1966 and mid-1968 and a peak at the end of 1974.

A history of vehicle registration is available from the UK Driver and Vehicle Licensing Agency (DVLA). This provides some explanation of the peaks observed in figure 10. In January 1963 a suffix was introduced to indicate the year in which the vehicle was registered with the letter A corresponding to 1963, B for 1964 and so forth. In order to shift the focus of new car sales away from the start of the year, the month for the new registration letter was changed to August in 1967. Thus the letter F was used for all cars registered between August 1967 and July 1968, G for August 1968 to July 1969 and so forth. The series of high τ^{2*} values around 1967 indicate that the change in date for the new suffix letter had an effect on the seasonal pattern of registration although the shift to August was not particularly strong at this point. The sudden change in seasonal pattern detected 1974 is harder to explain. There are a number of possible contributing factors. In 1974 the Driver and Vehicle Licensing Centre (DVLC) took over the registration of all new vehicles which may have altered the way in which new registrations were recorded. Emergence from the oil crisis and the political upheaval of 1974 combined with the introduction of a new style of number plates in 1973 may have focused attention on new registration cars as status symbols and led to a more marked jump in August sales from 1975 onwards. This is confirmed by figure 11 which shows the τ^2 statistics for an intervention in the first component of a dummy seasonal model. The peak in June 1975 indicates a shift in seasonal pattern corresponding to an sudden drop in July registrations compensated by a rise in August.

4.3 Confounding with outliers

In many instances sudden seasonal changes are associated with large values in other diagnostic statistics. In particular seasonal shifts may cause large values in smoothation based statistics which are usually thought of as detecting outliers. If the converse were true, that is outliers caused large values in the statistics designed to detect seasonal shifts, the value of the statistics would be seriously undermined. Fortunately this is not the case as shown below using two simple corruptions of the well known airline data (Box, Jenkins, and Reinsel 1994) shown in figure 12.

The first corrupted data set, figure 12(b), is the log airline data with a typographic error similar to that observed in the quarterly car registration series. This is introduced

in January 1955 by repeating the December 1954 value thus shifting the whole series forward. Figure 13(a) shows the τ^2 statistic corresponding to a shock to the measurement equation for this corrupted series. There are a number of peaks between January 1954 and December 1956 suggesting the presence of measurement outliers. Figure 13(b) shows the seasonal τ^{2*} statistic with the large peak clearly locating the corruption in the series. Once this seasonal shift is accounted for both statistics for detecting outliers and aberrant seasonal behaviour indicate that there is nothing unusual in the series. In the second corrupted series, figure 12(c), an outlier is introduced in January 1955 by replacing the observed value of 5.489 with the value 5.6. This is clearly indentified by the spike in figure 14(a). However, there is no peak in figure 14(b) indicating that outliers are not confounded with seasonal shifts.

5 Conclusion

Sudden changes in seasonal pattern, caused by real events or mistakes in the data record, arise in series which are of practical importance. In order to get good estimates of model parameters, these changes must be detected and accounted for. The statistics described in this paper are generated as a by product of the estimation process – they are free. Despite this, as the examples in section 4 indicate, plots of the statistics are effective in detecting aberrant seasonal behaviour.

The examples also suggest a seasonal shock diagnostic methodology. The τ^{2*} statistics detect the presence and general location of seasonal shifts. Statistics associated with shocks to individual seasonal elements then determine the nature of the change. Another area which could be developed is the detection of temporary changes, for example one unusual year. The composite interventions (De Jong and Penzer 1998) required to model this sort of behaviour can readily be constructed using two shocks.

6 Acknowledgements

I am indebted to Andrew Harvey and Siem Jan Koopman whose original ideas were the catalyst for this work. Derek Jones at the Department of the Environment, Transport and the Regions provided the true quarterly car registration figures for which I am very grateful. Thanks also to Anthony Atkinson and Jane Galbraith for their helpful suggestions.

List of figures

- **Figure 1**: Signatures for quarterly dummy seasonal models, shocks to (a) first, (b) second, (c) third and (d) all seasonal elements.
- **Figure 2**: Quarterly UK marriages (thousands) quarter 1 of 1965 to quarter 4 of 1970.
- **Figure 3**: τ statistics for shocks to (a) first and (b) third seasonal elements in dummy model of marriages data.
- **Figure 4**: (a) τ^{2*} and (b) $\tau^{2\dagger}$ statistics for dummy seasonal model of marriages data.
- **Figure 5**: τ statistics for shocks to (a) first and (b) fourth seasonal elements in form-free model of marriages data.
- **Figure 6**: Quarterly UK new car registrations (thousands) from quarter 1 of 1955 to quarter 1 of 1997 as published in ETAS 23.
- **Figure 7**: τ^{2*} statistics for a trigonometric model of the original quarterly car registrations series.
- **Figure 8**: τ^{2*} statistics for a trigonometric model of the corrected quarterly car registrations series.
- **Figure 9**: Log monthly UK new car registrations (thousands) from January 1955 to September 1991.

- **Figure 10**: τ^{2*} statistics for a trigonometric model of the log monthly car registrations series.
- **Figure 11**: τ^2 statistics for shock to the first component of a dummy seasonal model of the log monthly car registrations series.
- **Figure 12**: Log monthly number of airline passengers from January 1949 to December 1960 (a) original data (b) with seasonal shift and (c) with an outlier.
- **Figure 13**: τ^2 statistics for (a) outliers and (b) seasonal shifts in log airline series with seasonal shift.
- **Figure 14**: τ^2 statistics for (a) outliers and (b) seasonal shifts in log airline series with an outlier.

References

Anderson, B. D. O. and J. B. Moore (1979). *Optimal filtering*. Englewood Cliffs, New Jersey: Prentice-Hall.

Ansley, C. F. and R. Kohn (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *Annals of Statistics 13*, 1286–1316.

Ansley, C. F. and R. Kohn (1990). Filtering and smoothing in state space models with partially diffuse initial conditions. *Journal of Time Series Analysis 11*, 275–294.

Atkinson, A. C., S. J. Koopman, and N. Shephard (1997). Detecting shocks: outliers and breaks in time series. *Journal of Econometrics 80*, 387–422.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control* (3rd ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Brown, R. L., J. Durbin, and J. M. Evans (1975). Techniques of testing constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B 37*, 141–192.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*, 541–553.

Chang, I., G. C. Tiao, and C. Chen (1988). Estimation of time series parameters in presence of outliers. *Technometrics 30*, 193–204.

De Jong, P. (1988). A cross-validation filter for time series models. *Biometrika 75*, 594–600.

De Jong, P. (1989). Smoothing and interpolation with the state-space model. *Journal of the American Statistical Association 84*, 1085–1088.

De Jong, P. (1991). The diffuse Kalman filter. *Annals of Statistics 19*, 1073–1083.

De Jong, P. and J. R. Penzer (1998). Diagnosing shocks in time series. *Journal of the American Statistical Association 93*, 796–806.

Doornik, J. A. (1996). *Ox: Object Oriented Matrix Programming*. London: Chapman and Hall.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society,* Series B *34*, 350–363.

Harrison, P. J. and C. F. Stevens (1976). Bayesian forecasting. *Journal Royal Statistical Society, Series B 38*, 205–247.

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

Harvey, A. C. and S. J. Koopman (1992). Diagnostic checking of unobserved component time series models. *Journal of Business and Economic Statistics 10*, 377–389.

Harvey, A. C., S. J. Koopman, and J. Penzer (1998). Messy time series: a unified approach. (to appear).

Harvey, A. C. and G. D. A. Phillips (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika 66*, 49–58.

Kohn, R. and C. F. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika 76*, 65–79.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–126.

Koopman, S. J. (1997). Exact initial Kalman filter and smoother for non-stationary time series models. *Journal of the American Statistical Association 92*, 1630–1638.

Koopman, S. J., A. C. Harvey, J. A. Doornik, and N. Shephard (1995). *STAMP 5.0: Structural Time Series Analyser, Modeller and Predictor*. London: Chapman and Hall.

Koopman, S. J. and N. Shephard (1992). Exact score for time series models in state space form. *Biometrika 79*, 823–826.

Koopman, S. J., N. Shephard, and J. A. Doornik (1998). Ssfpack 2.1: Statistical algorithms for models in state space (http://center.kub.nl/stamp/ssfpack.htm). Submitted.

Lewin, P. (1997). *Economic Trends Annual Supplement, No. 23*. London: The Stationary Office.

McCulloch, R. E. and R. S. Tsay (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association 88*, 968–978.

Snyder, R. D. and G. R. Saligari (1996). Initialization of the Kalman filter with partially diffuse initial conditions. *Journal of Time Series Analysis 17*, 409–424.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association 81*, 132–141.

Tsay, R. S. (1988). Outliers, level shifts and variances changes in time series. *Journal of Forecasting 7*, 1–20.

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

Willsky, A. S. and H. L. Jones (1976). A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control 21*, 108–112.