ISSN 1440-771X

MONASH University

Australia

Department of Econometrics
and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

# A Pedants Approach to Exponential Smoothing

Ralph D Snyder

March 2005

Working Paper 5/05

# A Pedants Approach to Exponential Smoothing

Ralph D. Snyder
Department of Econometrics and Business Statistics,
Monash University
Clayton, Victoria 3800, Australia

March 22, 2005

**Abstract**

An approach to exponential smoothing that relies on a linear single source of error state space model is outlined. A maximum likelihood method for the estimation of associated smoothing parameters is developed. Commonly used restrictions on the smoothing parameters are rationalised. Issues surrounding model identification and selection are also considered.

It is argued that the proposed revised version of exponential smoothing provides a better framework for forecasting than either the Box-Jenkins or the traditional multi-disturbance state space approaches.

Keywords: time series analysis, prediction, exponential smoothing, ARIMA models, Kalman filter, state space models

JEL CLASSIFICATION: C22

## 1 Introduction

Given that exponential smoothing is one of the most widely used methods of forecasting for inventory control and operations management (Gardner, 1985), and given that suitable adaptations of it are increasingly being used in finance applications to measure volatility, one could be forgiven for thinking that all the details for its proper use would have been resolved since its inception in the 1950s. Traditional implementations of it, however, were based on heuristics instead of a proper statistical framework. Highly questionable practices arose as a consequence, particularly in relation to the estimation of prediction error variances (Johnston and Harrison, 1986; Snyder, Koehler and Ord, 1999).

The Bayesian forecasting framework (Harrison and Stevens, 1971) emerged as an attempt to avoid these pitfalls. Being built on traditional multi-disturbance state space models (Kalman, 1960; Kalman and Bucy 1961), it proved to be necessary to use the Kalman filter in place of exponential smoothing. Statistical rigor, it seemed, could only be achieved by discarding exponential smoothing.

A later development (Ord, Koehler and Snyder, 1997), however, revealed that exponential smoothing could still retain a central place in forecasting. The multi-disturbance state space model of Bayesian forecasting was replaced by an innovations state space model (Anderson and Moore, 1979; Snyder, 1985). It was then possible to propose a maximum likelihood approach to the estimation of seed states and smoothing parameters in place of the old heuristics. It was also possible to replace the ad hoc approaches for measuring prediction error variances with a logically sound model-based approach. This development, therefore, provided the missing statistical framework for exponential smoothing.

Most of the issues surrounding exponential smoothing have since been resolved (Ord et. al., 1997; Hyndman, Koehler, Snyder and Grose, 2002) but, some matters of detail still remain to be addressed. First, the likelihood was seen as a function of the seed state variables, smoothing parameters and the variance and it was optimised with respect to all these quantities. Unlike the parameters, however, the seed state variables are random. Moreover, they are not observable so they cannot be fixed at observed values like the series values. Their randomness must be made to disappear in some way. The strategy (Ord et. al., 1997) to resolve this issue was to condition on *fixed* but unknown values of the seed variables, resulting in a *conditional likelihood* function. Nevertheless, the seed variables are really random and they strictly induce randomness in the value of conditional likelihood. A more satisfactory approach, from a theoretical perspective at least, is to average the conditional likelihood with respect to a distribution of seed state variables to give the exact likelihood function, something that is deterministic and hence suitable for optimisation purposes. In other words, there is a need to explore the possibility of replacing the conditional likelihood with the exact likelihood in the theory of exponential smoothing. This is one of the main issues addressed in this paper.

Second, under certain conditions (Ord et. al., 1997; Hyndman, Akram and Archibald, 2003) for the smoothing parameters, exponential smoothing discounts the importance of older sample values in associated calculations. However, these conditions differ markedly from much tighter restrictions (Gardner, 1985) commonly used in practice. It has been found (Hyndman et. al., 2002) that tighter restrictions can translate into better forecasts, a point that supports current practice. The practical restrictions, however, are somewhat arbitrary. They have never been justified with respect to an underlying principle. A purpose of this paper is to show that a set of narrower restrictions similar to those used in practice can be derived from first principles.

To set the scene, multiple error structural time series models are introduced in section 2. Single source of error state space models are derived from them. In the process, the tighter restrictions on the smoothing parameters are derived. The most general linear form of exponential smoothing is introduced in section 3. Its links with the single source of error models is also outlined. The exact likelihood function is derived in section 4 and its use in estimation is outlined. Model selection is considered in section 5.

# 2 State Space Models

## 2.1 Multiple Source of Error State Space Model (MSOE)

State space models and exponential smoothing are known to be closely linked (Harrison and Stevens, 1971; Harvey, 1991). A new unorthodox form for the state space framework that serves the purpose of this paper best is:

$$\begin{aligned} Y_t &= \overline{h}' X_t + U_t & \text{(1a)} \\ X_t &= T\left(X_{t-1} + V_t\right). & \text{(1b)} \end{aligned}$$

Equation (1a) is the *measurement equation.* It shows how the observable series value $Y_t$ is related to a random $k-$vector $X_t$ called the *state vector*, and a random variable $U_t$ called the *measurement disturbance.* The $U_t$ are normally and independently distributed with mean 0 and a common variance $\sigma^2$. Each $U_t$ measures *temporary unanticipated* change, that is stochastic change that impacts on only the period in which it occurs. The $k$-vector $\overline{h}$ is fixed.

The state vector $X_t$ summarises the history of the process. Its evolution through time is governed by the first-order recurrence relationship (1b) where $T$ is a fixed $k \times k$ matrix called the *transition matrix* and $V_t$ is a random $k$-vector of what are called the *system disturbances.* The $V_t$ are normally and independently distributed with mean 0 and variance matrix $\sigma^2 Q$ where $Q$ is a symmetric, positive semi-definite matrix. The purpose of $V_t$ is to model the effect of *structural change*, that is unanticipated change that persists through time.

The covariance between $V_t$ and $U_t$ is given by $\sigma^2 q$ where $q$ is a fixed $k$-vector. $U_t$ and $V_s$ are independent for all distinct periods $s$ and $t$. The unorthodox feature of this model is that the prior state vector is amended by the system disturbance before it is transformed by the transition matrix.

The model (1) is *invariant* because the vectors $\overline{h}, q$ and matrices $T$, $Q$ are independent of time. In most applications the elements of $h$, $q$, $T$ and $Q$ are a mix of known and unknown quantities. The unknown quantities are represented by the vector $\theta$. A problem is to estimate $\theta$ from a sample $y_1$, $y_2$,...,$y_n$ where $n$ is the sample size.

The elements of $q$ and $Q$ are usually unknown. In the quest for parsimony the following additional assumptions are often made:

1. The elements of $V_t$ are mutually independent; hence the off-diagonal elements of $Q$ are zero.

2. $U_t$ and $V_t$ are independent; hence $q = 0$.

The effect of these assumptions is to reduce the number of unknown parameters in $Q$ and $q$ from $k^2 + k$ to $k$.

Time series methods account for the intertemporal dependencies that may exist between the values of a time series. These independence assumptions can be imposed on the disturbances without destroying the possibility of dependencies between series values. In fact, it will be seen in Sections 2.3-2.5 that the independence assumptions can often be imposed on special cases of the model without loss of generality because the multi-disturbance state space model, in its most general form, contains many redundant parameters.

## 2.2 Single Source of Error State Space Models (SSOE)

If Equation (1b) is substituted into Equation (1a) the equation $y_t = h'b_{t-1} + h'V_t + U_t$ is obtained where $h' = \overline{h}'T$. The term $h'b_{t-1}$ is the one-step ahead prediction of $y_t$. The remainder $E_t = h'V_t + U_t$ is the one-step ahead prediction error. Its composition reflects that fact that prediction errors can possess two sources of error: the error $h'V_t$ induced by structural change and the temporary error $U_t$.

An alternative to the independence assumptions, to achieve a more parsimonious representation, is to assume that $U_t$ and $V_t$ are perfectly correlated with $E_t$. Then $V_t = \overline{\alpha}E_t$ and $U_t = \beta E_t$ where $\overline{\alpha}$ is a non-negative fixed $k$-vector and $\beta$ is a non-negative scalar. The state space model can then be rewritten as

$$\begin{aligned} Y_t &= h'X_{t-1} + E_t && \text{(2a)} \\ X_t &= T\left(X_{t-1} + \overline{\alpha}E_t\right). && \text{(2b)} \end{aligned}$$

In effect, the number of parameters is again reduced from $k^2 + k$ to $k$. The scalar $\beta$ is ignored because it does not directly appear in this single source of error specification. At first sight it might be thought that this perfect correlation assumption is likely to be very restrictive. However, the examples considered in Sections 2.3-2.5 indicate that this need not be the case.

An interesting byproduct of this specification is that $E_t = h'\overline{\alpha}E_t + \beta E_t$, something that must be true for all non-zero values of $E_t$. It follows that the parameter vector $\overline{\alpha}$, as well as being non-negative, must satisfy the linear restriction

$$h'\overline{\alpha} \leq 1. \tag{3}$$

It suggests that the elements of $\overline{\alpha}$ effectively allocate the prediction error amongst the unobserved components of the model. Because it relates to a model formulated in terms of the one-step ahead predictions, the restriction (3) will be referred to as the *prediction condition*.

An equivalent variation of the specification of the the single source of error state space model is

$$\begin{aligned} Y_t &= h'X_{t-1} + E_t && \text{(4a)} \\ X_t &= TX_{t-1} + \alpha E_t. && \text{(4b)} \end{aligned}$$

where $\alpha = T\overline{\alpha}$. It is the more traditional form of the single source of error state space model (Ord et. al., 1997). Equivalent restrictions on the $k$-vector $\alpha$ can be derived from the prediction condition on $\overline{\alpha}$. If $T$ is non-singular, the restrictions take the form

$$\begin{aligned} T^{-1}\alpha &\geq 0 && \text{(5a)} \\ \overline{h}\alpha &\leq 1. && \text{(5b)} \end{aligned}$$

In some applications $T$ may be singular, in which case it is simplest to elucidate the restrictions on a case by case basis.

The recurrence relationship

$$X_t = DX_{t-1} + \alpha Y_t, \tag{6}$$

where $D = T - \alpha h'$, may be derived by eliminating the error from (2). The solution to this relationship is

$$X_t = D^t X_0 + \sum_{j=0}^{t-1} D^j \alpha Y_{t-j} \tag{7}$$

It shows that the state vector depends on past values of a series. In the presence of structural change, it would be expected that the state vector is influenced less by older series values than more recent ones. Structural change implies that $\alpha$ should take values that ensure that $\alpha D^j \to 0$ as $j \to \infty$. Unless $\alpha = 0$, the case of no structural change, this condition holds when the eigenvalues of $D$ lie within the unit circle. This leads to additional restrictions on the vector $\alpha$, herein referred to as the *structural* change conditions.

The perfect correlation assumption is not necessary to derive the single source of error model (4) from a multiple source of error model. The Kalman filter, for *any* invariant multiple source of error model, has a steady state that is suggestive of a single source of error model with the same ouput covariance structure (Anderson and Moore, 1979). In other words, it is always possible to find an SSOE that is equivalent to a given MSOE. In this more general context, it is normal to impose the structural change condition instead of the prediction condition.

The framework (4) is particularly important because it underpins the most general linear form of exponential smoothing (Ord et. al., 1997), something that is explored in Section 3 using the new but equivalent model formulation (2). Within this context, it is normally applied with what effectively amounts to the prediction error condition imposed on $\alpha$. However, its form appears to have first emerged in Box and Jenkins (1976) where it was proven to be the first-order recurrence relationship representation (eventual forecast functions) of the ARIMA family of models. When the structural change condition is imposed instead of the prediction condition, it is actually more general than the ARIMA class because invertibility excludes the possibility that $\alpha = 0$. It encompasses, for example, the classical linear trend line that is precluded by the invertibility condition. It will be seen in Sections 2.3-2.5 that the prediction condition normally imposes much tighter restrictions on the vector $\alpha$ than the structural change conditions.

It is interesting to speculate as to why the SSOE model has not played a more central role in time series analysis. Because they were wedded to the use of autocorrelation functions and partial autocorrelation functions for the important issue of model identification, Box and Jenkins saw considerable value in the ARIMA form for identification and estimation purposes. The first-order form of their framework was relegated to the limited role of generating the final forecasts. In taking this stance they overlooked another possible approach to identification: the use of unobserved components in conjunction with the 'stylised facts' of time series analysis to model the intertemporal dependencies in a time series (Harvey, 1991).

The state space model (4) has been also referred to as an innovations model because of its close link with the steady state of a Kalman filter applied to time invariant multi-disturbance state space models (Anderson and Moore, 1979). Of the infinite number of possible state space models with a common output autocovariance function, it is the only one with an input noise process that corresponds to the innovations from the Kalman filter in the steady state. When it is used directly for representing time series without reference to an equivalent multi-disturbance model (Snyder, 1985), it is referred to as a single source of error state space model (SSOE).

## 2.3 Local Level Model

One of the simplest state space models involves a local level $A_t$ that follows a random walk over time. The series values are randomly scattered about the local levels. More specifically

$$\begin{aligned} Y_t &= A_t + U_t & \text{(8a)} \\ A_t &= A_{t-1} + V_t . & \text{(8b)} \end{aligned}$$

The correlation between $U_t$ and $V_t$ is designated by $\rho$.

A model with only one primary source of randomness may be derived from the multi-disturbance model (8) employing a suitable adaptation of an argument from Harvey and Koopman (2000). First, the reduced form is obtained by eliminating the unobservable $A_t$. The result is the ARIMA(0,1,1) model

$$\Delta Y_t = U_t - U_{t-1} + V_t \tag{9}$$

with autocovariance function is given by

$$\gamma_j = \begin{cases} 2\sigma_u^2 + \sigma_v^2 + 2\rho\sigma_u\sigma_v & \text{for } j = 0 \\ -\sigma_u^2 - \rho\sigma_u\sigma_v & \text{for } j = 1 \\ 0 & j > 1 \end{cases} . \tag{10}$$

where $\gamma_j$ is the autocovariance of lag $j$. This autocovariance function depends on the three parameters $\sigma_u$, $\sigma_v$ and $\rho$, but has only two non-zero values. Ostensibly, the three parameters cannot be uniquely determined. However, $\gamma_0 + 2\gamma_1 = \sigma_v^2$, so that $\sigma_v$ is uniquely determined. Only $\sigma_u$ and $\rho$ cannot be uniquely determined. It seems sensible to choose a value for $\rho$; then a unique value of $\sigma_u$ can be obtained. The most common strategy is to assume that $\rho = 0$ (Harrison and Stevens, 1971; Harvey, 1991). Since any value of $\rho$ may be used, however, there is no loss of generality in assuming that $\rho = 1$.

Second, Equation (8b) may be substituted into Equation (8a) to give

$$Y_t = A_{t-1} + V_t + U_t$$

The term $A_{t-1}$ is the one-step ahead prediction, while $E_t = V_t + U_t$ is the one-step ahead prediction error. The prediction error has two components: one

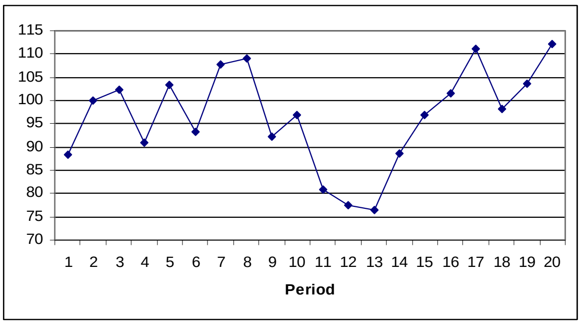

Figure 1: Simulated time series from a local level model with $a_0 = 100$, $\sigma = 10$ and $\alpha_1 = 0.5$.

*permanent* ($V_t$) and the other *temporary* ($U_t$). The permanent component might reflect the effect of new customers or the impact of new suppliers (competitors) in a market.

Third, perfectly correlated permanent and temporary disturbances also correlate perfectly with the one-step ahead prediction error $E_t$; in other words $V_{1t} = \alpha E_t$ and $U_t = \beta E_t$ where $\alpha$ and $\beta$ are non-negative parameters. The local level model (8) can be rewritten as

$$\begin{aligned} Y_t &= A_{t-1} + E_t && \text{(11a)} \\ A_t &= A_{t-1} + \alpha E_t && \text{(11b)} \end{aligned}$$

It is the single source of error version of the local level model (Ord, 1997). The associated prediction condition is

$$0 \leq \alpha \leq 1. \tag{12}$$

The size of the parameter $\alpha$ is a measure of the impact of structural change in a time series. When $\alpha = 0$, successive levels are equal: the case of no structural change. When $\alpha = 1$, the model reduces to a random walk, a case at the other extreme where a time series has no parametric structure (except the variance parameter).

A time series simulated from a local level model is shown in Figure 1. Successive values of the series have a tendency to be close to each other, a phenomena that may be attributed to structural change. This closeness property arises because the local level equation (11b) transmits the history of the process through time.

A recurrence relationship corresponding to to the general relationship (6) is

$$A_t = \delta A_{t-1} + \alpha Y_t. \tag{13}$$

It describes the evolution of the level over time. Note that $\delta = 1 - \alpha$. Under the condition (12), the level can be viewed as a weighted average. In traditional expositions of exponential smoothing (Winters, 1960), the condition (12) is imposed to permit this interpretation. As has been seen here, there is a more fundamental reason for this condition. It was derived from structural considerations, not imposed as an assumption.

The structural change condition requires that $\alpha\delta^j \to 0$ as $j \to \infty$. This occurs if $-1 < \delta \leq 1$. The equivalent condition, in terms of $\alpha$, is

$$0 \leq \alpha < 2. \tag{14}$$

Advocates of the broader condition (14) argue that it provides greater flexibility. Indeed maximum likelihood estimates of $\alpha$ obtained under this restriction often exceed one on typical economic time series. Proponents of the narrower condition (12), however, argue that the added flexibility is counterproductive. An $\alpha$ in excess of one is seen as evidence of the existence of patterns in a time series such as a trend that are not covered by a local level model. It is seen as a signal that the local level model is not appropriate for the data and will yield inferior forecasts.

There are close links between the local level model and the ARIMA(0,1,1) model in the Box-Jenkins framework. Another ARIMA(0,1,1) model is obtained by differencing Equation (11a) and eliminating the level variables with Equation (11b). Condition (14) corresponds to the invertibility condition for an ARIMA(0,1,1) model.

Given that the tighter condition (12) seems to be more appropriate for a local level model, it might also be argued that the ARIMA(0,1,1) model provides more flexibility and is therefore more likely to work better. However, the same basic criticism applies. The ARIMA(0,1,1) model, and indeed the entire ARIMA family of models, largely ignore structural considerations. As all time series emerge from processes or systems, the additional information conveyed by their structure should not be ignored. The perceived additional generality of the Box-Jenkins approach is really illusory. The now growing view (Durbin, 2000) that the Box-Jenkins approach is an inadequate framework for forecasting is reinforced by this argument.

## 2.4 Local Trend Model

A local level may be supplemented by a time dependent growth rate $B_t$ which follows a random walk $B_t = B_{t-1} + V_{2t}$ where $V_{2t}$ is another disturbance. The resulting local trend model is

$$\begin{aligned} Y_t &= A_t + U_t && (15a) \\ A_t &= A_{t-1} + B_t + V_{1t} && (15b) \\ B_t &= B_{t-1} + V_{2t} && (15c) \end{aligned}$$

Unlike the usual local trend model, the current level in (15b) is updated with the *current* growth rate. Equation (15c) may be used to eliminate $B_t$ from Equation (15b) to yield the relationship

$$A_t = A_{t-1} + B_{t-1} + V_{1t} + V_{2t}. \quad (16)$$

This model then becomes a special case of the general framework (1).

The equation $Y_t = A_{t-1} + B_{t-1} + V_{1t} + V_{2t} + U_t$ is obtained when $A_t$ is eliminated from Equation (15a). Given that $A_{t-1} + B_{t-1}$ is now the one-step ahead prediction, the prediction error is given by $E_t = V_{1t} + V_{2t} + U_t$. The prediction error has three components, two of them permanent. As before, one of the permanent disturbances is associated with the change in the underlying level. The other is the permanent change in the rate of growth. It is assumed that the three disturbances are potentially correlated.

The reduced form of this local trend model is the ARIMA(0,2,2) process $\Delta Y_t^2 = V_{2t} + V_{1t} - V_{1,t-1} + U_t - 2U_{t-2} + U_{t-2}$. It is readily seen that all autocovariances of $\Delta Y_t$ satisfy the condition $\gamma_j = 0$ for $j > 2$. The first three covariances, which are potentially non-zero, depend on the the three disturbance variances and the three correlation coefficients between the disturbances. Again there is an identification problem. A common resolution is to assume that the disturbances are contemporaneously uncorrelated. Then the variances can be uniquely determined. A second, but observationally equivalent possibility, is to assume that the disturbances are all perfectly correlated.

Under the perfect correlation assumption, the three disturbances are also perfectly correlated with the one-step ahead prediction error, so that $V_{1t} = \overline{\alpha}_1 E_t$, $V_{2t} = \overline{\alpha}_2 E_t$ and $U_t = \beta E_t$ where $\overline{\alpha}_2$ is a parameter. The resulting single source of error model is

$$\begin{aligned} Y_t &= A_{t-1} + B_{t-1} + E_t && (17a) \\ A_t &= A_{t-1} + B_{t-1} + (\overline{\alpha}_1 + \overline{\alpha}_2) E_t && (17b) \\ B_t &= B_{t-1} + \overline{\alpha}_2 E_t && (17c) \end{aligned}$$

It can be rewritten as

$$\begin{aligned} Y_t &= A_{t-1} + B_{t-1} + E_t && \text{(18a)} \\ A_t &= A_{t-1} + B_{t-1} + \alpha_1 E_t && \text{(18b)} \\ B_t &= B_{t-1} + \alpha_2 E_t && \text{(18c)} \end{aligned}$$

where $\alpha_1 = \overline{\alpha}_1 + \overline{\alpha}$ and $\alpha_2 = \overline{\alpha}_2$ This is the more traditional form of the local linear trend model found in Hyndman et. al. (2002). It may be established that the region for the parameters then becomes $\alpha_1 \geq 0$, $\alpha_2 \geq 0, \alpha_1 < 1$,and $\alpha_2 \leq \alpha_1$.

Yet another way of writing the model is

$$\begin{aligned} Y_t &= A_{t-1} + B_{t-1} + E_t && \text{(19)} \\ A_t &= A_{t-1} + B_{t-1} + \alpha_1 E_t && \text{(20)} \\ B_t &= B_{t-1} + \alpha_2^* \left( A_t - A_{t-1} - B_{t-1} \right) && \text{(21)} \end{aligned}$$

where $\alpha_2^* = \alpha_2 / \alpha_1$. It is obtained by solving (18b) for $E_t$ and substituting the result into Equation (18c). It is the model underlying the original form of trend corrected exponential smoothing (Holt, 2002). The above feasible region for the parameters can be re-expressed as $0 \leq \alpha_1 \leq 1$ and $0 \leq \alpha_2^* \leq 1$, conditions that have been traditionally advocated (Makridakis, Wheelwright and Hyndman, 1998) for trend corrected exponential smoothing. A contribution of this paper has been to show that these conditions can be derived from structural considerations, instead of being imposed by assumption as has been the tradition.

The invertibility conditions for an ARIMA(0,2,2) process are $\alpha \geq 0$, $\alpha_2 \geq 0$ and $2\alpha_1 + \alpha_2 \leq 4$. This region is larger than the one derived from structural considerations. It again highlights a problem with the Box-Jenkins approach.

## 2.5 Local Seasonal Model

An extension involving a seasonal factor $C_t$ is

$$\begin{aligned} Y_t &= A_t + C_t + U_t && \text{(22a)} \\ A_t &= A_{t-1} + B_t + V_{1t} && \text{(22b)} \\ B_t &= B_{t-1} + V_{2t} && \text{(22c)} \\ C_t &= C_{t-m} + V_{3t}. && \text{(22d)} \end{aligned}$$

where $m$ is the number of seasons per year. Substituting Equations (22b) and (22d) into Equation (22a) yields $Y_t = A_{t-1} + B_{t-1} + C_{t-m} + E_t$ where

$E_t = V_{1t} + V_{2t} + V_{3t} + U_t$. Adapting the perfect correlation argument above, the equivalent single source of error model is

$$\begin{aligned}
Y_t &= A_{t-1} + B_{t-1} + C_{t-m} + E_t && \text{(23a)}\\
A_t &= A_{t-1} + B_{t-1} + (\overline{\alpha}_1 + \overline{\alpha}_2)\, E_t && \text{(23b)}\\
B_t &= B_{t-1} + \overline{\alpha}_2 E_t && \text{(23c)}\\
C_t &= C_{t-m} + \overline{\alpha}_3 E_t && \text{(23d)}
\end{aligned}$$

where $\overline{\alpha}_1 \geq 0$, $\overline{\alpha}_2 \geq 0$, $\overline{\alpha}_3 \geq 0$ and $\overline{\alpha}_1 + \overline{\alpha}_2 + \overline{\alpha}_3 \leq 1$. An equivalent representation is

$$\begin{aligned}
Y_t &= A_{t-1} + B_{t-1} + C_{t-m} + E_t && \text{(24a)}\\
A_t &= A_{t-1} + B_{t-1} + \alpha_1 E_t && \text{(24b)}\\
B_t &= B_{t-1} + \alpha_2 E_t && \text{(24c)}\\
C_t &= C_{t-m} + \alpha_3 E_t && \text{(24d)}
\end{aligned}$$

where $0 \leq \alpha_2 \leq \alpha_1$, $\alpha_3 \geq 0$ and $\alpha_1 + \alpha_3 \leq 1$. The latter conditions define a region for the smoothing parameters that is smaller than the region associated with the invertibility conditions [1](Hyndman et. al., 2003).

# 3 Exponential Smoothing

## 3.1 Simple Exponential Smoothing

The single source of error local level model underpins what has traditionally been called the simple exponential smoothing algorithm. The model treats $Y_t$ as a random variable. It describes the situation before $Y_t$ is observed. After it is observed, $Y_t$ becomes a fixed value designated by $y_t$, and certain calculations become possible. If $A_{t-1}$ is known to be equal to a fixed value $a_{t-1}$ from preceding calculations, the measurement equation may be used to calculate a fixed value $e_t = y_t - a_{t-1}$ for the error $E_t$. The level equation can then be used to obtain the fixed value $a_t = a_{t-1} + \alpha_1 e_t$ for $A_t$. If the process is started with the seed $A_0$ equal to a fixed trial value $a_0$, these steps can be repeated for successive values of a time series. The resulting algorithm corresponds to classical simple exponential smoothing (Brown, 1959). The $a_t$ form what have traditionally been called the *smoothed* series, but this terminology is inconsistent with modern usage of the term smoothed. The

[1] My thanks to Muhammad Akram for producing plots that confirm this relationship.

typical $a_t$ depends on a sub-sample $y_1$, $y_2$,..., $y_t$ rather than the entire sample $y_1$, $y_2$,..., $y_n$ through the relationship

$$a_t = \delta^t a_0 + \alpha_1 \sum_{j=0}^{t-1} \delta^j y_{t-j}. \tag{25}$$

where $\delta = 1 - \alpha_1$ is the so-called discount factor. Thus, the $a_t$ are more akin to a filtered series. For future reference, it should be noted that $a_t$ is a linear function of the seed $a_0$.

## 3.2 General Exponential Smoothing

Similar arguments can be applied to the local trend model to give trend corrected exponential smoothing (Holt, 2004). The Winters additive method (Winters, 1960) can also be obtained from the local seasonal model. The details of these approaches is not covered here because they are special cases of the general form of exponential smoothing.

The general exponential smoothing algorithm is based on the the general linear single source of error model (4). It begins in typical period $t$ with a fixed value $x_{t-1}$ for the random state vector $X_{t-1}$ obtained from earlier calculations. The one-step ahead prediction is obtained with $\widehat{y}_t = h'x_{t-1}$. On observing the fixed value $y_t$ for $Y_t$ the fixed value $e_t = y_t - \widehat{y}_t$ for the error $E_t$ is computed. The fixed value $x_t = Tx_{t-1} + \alpha e_t$ is then calculated for the state vector $X_t$. This process, which is seeded with a fixed trial value $x_0$ for the seed state vector $X_0$, is repeated for each successive observation in the sample. The resulting sequence of $\widehat{y}_t$ values is the *smoothed* series.

# 4 Estimation

A challenge is to find appropriate values for the seed vector $X_0$ and the parameters $\alpha$ and $\sigma^2$. Then estimates of subsequent state vectors may be generated recursively with the transition equation. Once the final state vector is obtained it may be used to generate predictions.

## 4.1 Estimation of the Seeds

### 4.1.1 Heuristic Approaches

Traditionally, a variety of heuristics (Gardner, 1985) have been used to estimate the seed state vector. Examples include:

- Local level model: the seed level is approximated by a simple average of the first few series values.

- Local trend model: a trend line in fitted using the principle of least-squares to the first five observations in a time series; the seed level is set to the intercept and the seed growth rate is set to the trend rate of growth.

- Seasonal model: a linear trend with seasonal dummy variables is fitted to the few years of observations from a time series; the seed level and seed rate are set as for the local trend; the seed seasonal effects are set to the seasonal averages from the approximating model.

The heuristic methods implicitly assume that structural change has been fairly limited over the short stretch of data to which they are applied. As such they usually provide plausible estimates of the local structure in this short time span. An approach that does not need this approximation is possible and will now be considered. It will be based on the assumption that $\alpha$ has a known value, possibly an assigned trial value.

### 4.1.2 Simple Exponential Smoothing

The seed value $a_0$ in Equation (25) is unknown. A seemingly futile tactic is to let $a_0 = 0$. The typical pattern that emerges for the errors is shown in Figure 2. It is obtained by applying simple exponential smoothing to the time series in Figure 1. The errors are quite large initially but quickly settle to a stable state with a zero mean. The initial positive bias in the errors reflects the effect of the poor trial value of 0 for the seed level. However, the bias disappears quickly.

Suppose the errors, based on a zero seed value, are designated by $e_t^*$. From the theory in the Appendix for the general linear case of exponential smoothing, it can be shown that

$$e_t^* = \delta^{t-1} a_0 + e_t \tag{26}$$

The first term on the right hand side of Equation (26) is the bias term. It is this that leads to the initial distortion depicted in the errors in Figure 2. It depends on the seed state but its size decreases with increases in $t$ when $|\delta| < 1$. By assumption the $e_t$ are drawn from identical and independent normal distributions. Equation (26) can therefore be viewed as a simple homogeneous regression. The formula for the least-squares estimate of the seed level is

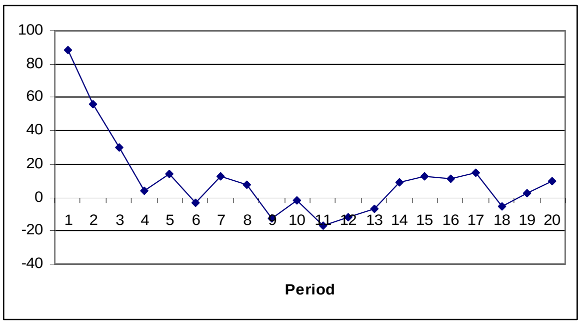

Figure 2: Plot of errors from simple exponential smoothing with $a_0 = 0$ and $\alpha_1 = 0.5$.

$$\widehat{a}_0 = \sum_{t=1}^{n} \delta^{t-1} e_t^* \Big/ \sum_{t=1}^{n} \delta^{2(t-1)}.$$

Although seeding simple exponential smoothing with a zero level seemed initially to be counterproductive, this tactic can now be seen as a convenient steppingstone to getting a statistically sound estimate of the seed. Once obtained, it is then possible to calculate the unbiased one-step ahead prediction errors with the formula $e_t = e_t^* - \delta^{t-1}\widehat{a}_0$. An equivalent tactic is to undertake a second pass of the data with simple exponential smoothing seeded with $\widehat{a}_0$ rather than 0 for the unbiased one-step ahead prediction errors.

It may be thought that the least-squares estimate of the seed level is more accurate than its heuristic counterpart. The latter, however, also gives quite plausible results in a wide range of circumstances. The reason for preferring the least-squares approach is that it provides a more general framework. It works for the special case $\alpha_1 = 0$. Then $e_t^* = y_t$ and the least-squares estimate reduces to the classical simple average $\widehat{a}_0 = \sum_{t=1}^{n} y_t \big/ n$. In other words, exponentially weighted averages and simple averages are properly reconciled under the least squares approach. The heuristic approaches are based on the assumption that $0 < \alpha_1 \leq 1$ and that any adverse effects are washed out as the method convergence to a stable state. However, when $\alpha_1$ is small, convergence is slow. And when $\alpha_1 = 0$, there is no convergence, in which case any adverse effects from a heuristic approach persist. The advantage of the proposed approach is that it works reliably over the entire interval $0 \leq \alpha_1 \leq 1$.

### 4.1.3 General Exponential Smoothing

Mimicking the logic used for simple exponential smoothing, the algorithm begins with $x_0 = 0$. The resulting errors, designated by $e_t^*$ are again biased. It is shown in the Appendix that the bias in these errors is linearly dependent on the true seed state $x_0$. Thus, the biased errors can be written as a linear function of the true seed state vector and the unbiased errors

$$e_t^* = Zx_0 + e_t. \tag{27}$$

The matrix $Z$, which depends on the smoothing parameter vector $\alpha$, is derived in the Appendix.

Again the principle of least-squares may be applied to give the estimate of the seed vector

$$\widehat{x}_0 = \left(Z'Z\right)^{-1} Z'e_t^*. \tag{28}$$

Then the unbiased errors may be calculated with $e_t = e_t^* - Zx_0$ or by applying the general form of exponential smoothing for a second pass of the data with $X_0 = \widehat{x}_0$. In the first approach the 'smoothed' series values may be recovered with $\widehat{y}_t = y_t - e_t$. In the second approach these values are generated as part of the second pass of the exponential smoothing algorithm.

## 4.2 Estimation of Parameters

Estimation of the smoothing parameter vector $\alpha$ and the variance $\sigma^2$ provide a further challenge. Simple heuristics (Gardner, 1985) were often used in early implementations of exponential smoothing to avoid the computational overheads of nonlinear optimisers. As computers became more powerful, however, it became feasible to adopt Winters' earlier suggestion of selecting those values that minimise the sum of squared errors. The evidence (Fildes, Hibon, Makridakis and Meade, 1998) suggests that optimisation leads to better forecasts. The standard deviation is then typically estimated with

$$\widehat{\sigma} = \sqrt{\sum_{t=1}^{n} \frac{e_t^2}{n}}. \tag{29}$$

or a variation $\widehat{\sigma} = 1.25\Delta$ where $\Delta$ is the mean deviation (Brown, 1959).

The *conditional* likelihood function (Ord et. al., 1997) can be used in place of the sum of squared errors criterion to yield the same estimates for both the seed vector $X_0$ and the smoothing parameter vector $\alpha$. It is based on the distribution of $Y|\alpha, x_0$ where $Y$ is a random vector formed from $n$ series values $Y_1, Y_2, ..., Y_n$. The random seed vector $X_0$ is set to a fixed but unknown value $x_0$; hence the use of the term *conditional.* Given that exponential smoothing transforms the original series autocorrelated series $Y$ to the uncorrelated error series $E$ and that this transformation has a unit Jacobian, it follows that $Y|\alpha, x_0$ has the same distribution as $E$, namely a multivariate normal distribution with mean of zero and variance matrix $\sigma^2 I$. In other words, the conditional likelihood is given by $\mathcal{L}(\alpha, x_0) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{t=1}^{n} e_t^2}{2\sigma^2}\right)$.

Likelihood, strictly speaking, should be based on the distribution of $Y|\alpha$, not the distribution of $Y|\alpha, x_0$ because $X_0$ is not observable. Some approach must be adopted to eliminate the dependence of the joint distribution of $Y, X_0|\alpha$ on $X_0$. One possibility is to eliminate the state variables $X_t$ from the SSOE. A lag operator $L$ may be introduced and used to write Equation (4b) as $X_t = TLX_t + \alpha E_t$. This solution $X_t = (1 - TL)^{-1} E_t$ can be used to eliminate $X_{t-1}$ from the measurement equation (4a) to give

$$Y_t = h(1 - TL)^{-1} \alpha E_{t-1} + E_t \tag{30}$$

This is a *reduced form* of the SSOE because it does not reference the state variables. It is the *integrated* reduced form because it represents the series in its original form.

It would now appear to be a simple matter to derive the exact likelihood. The reduced form is expanded to give the moving average representation

$$Y_t = \sum_{j=1}^{\infty} h(TL)^j E_{t-j} + E_t \tag{31}$$

$Y_t$ is clearly normally distributed with a zero mean. However, in most most business and economic applications, time series are non-stationary, in which case the transition matrix $T$ has unit roots and the variance of $Y_t$ is arbitrarily large. This is certainly true for time series represented by the local level, local trend or local seasonal models. The distribution of $Y_t$ is then not properly defined for the purpose of forming the likelihood function.

The essential problem is that that the series is not stationary. However, there usually exists a linear transformation that does not depend on the unknown parameters $\theta$ and which may be applied to the data to derive an equivalent non-stationary series. The exact likelihood of the original non-stationary series is then defined as the likelihood of the transformed series. More specifically, any matrix inverse can be rewritten in terms of its adjoint and determinant. Hence

$$(1 - TL)^{-1} = (1 - TL)^{\dagger} / |1 - TL| \tag{32}$$

where $(1-TL)^{\dagger}$ designates the adjoint and $|1-TL|$ designates the determinant of the matrix $1-TL$. The determinant is a polynomial function of the lag operator $L$ of degree $k$. It can be written as the product $|1-TL| = \Psi(L)\Phi(L,\alpha)$ where $\Psi(L)$ and $\Phi(L,\alpha)$ are polynomial functions of the lag operator $L$, the latter depending on the parameter vector $\alpha$. When some or all of the state variables are non-stationary the polynomial formed from the determinant $|1-TL|$ has unit roots. Some of the unit roots can be seasonal unit roots. As unit roots are independent of $\alpha$, $\Psi(L)$ is formed from the unit root components of $|1-TL|$.

The reduced form (30) can be rewritten as:

$$\Psi(L)\Phi(L,\alpha)Y_t = \Theta(L,\alpha)e_t \tag{33}$$

where $\Theta(L,\alpha) = h'(1-TL)^{\dagger} L\alpha + \Psi(L)\Phi(L,\alpha)$ is a polynomial function of degree $k$. Equation (33) is also a reduced form because it contains no state variables. The right hand side of Equation (33) is stationary, so that its left hand side is also stationary. It follows that $Z_t = \Psi(L)Y_t$ is a stationary series. $\Psi(L)$ is the means by which the original series $Y$ is transformed to the 'equivalent' stationary series $Z$. As it only has unit roots, this transformation process is undertaken with a succession of differencing operations, some of which may be seasonal differencing operations.

If there are $d$ non-stationary states, $\Psi(L)$ is a polynomial of order $d$, and so $Z$ is smaller than $Y$, its length being $n-d$; the initial $d$ observations are lost in the transformation process. It is not possible to reconstruct $Y$ from $Z$, so some information is lost by the transformation. Nevertheless, the exact likelihood of the original non-stationary state space model is defined as the likelihood of the reduced form model governing the stationary series $Z$.

This transformation process is fine provided that all the values of a time series have been observed. When there are missing values the reduction to stationary reduced form is not possible. Another, equivalent approach is needed to define the exact likelihood.

Conditional probability theory implies that associated density functions are related by

$$p\left(y|\alpha,\sigma^2\right) = p\left(y|x_0,\alpha,\sigma^2\right)p\left(x_0|\alpha,\sigma^2\right)/p\left(x_0|y,\alpha,\sigma^2\right). \tag{34}$$

When *all* the states are non-stationary, $X_0|\alpha$ has a non-informative distribution and so $p(y|\alpha,\sigma^2) \propto p(y|x_0,\alpha,\sigma^2)/p(x_0|y,\alpha,\sigma^2)$. Furthermore, from the theory of least-squares, $x_0|y,\alpha,\sigma^2 \sim N\left(0,\sigma^2(Z'Z)^{-1}\right)$. Thus, the exact likelihood function is given by

$$\mathcal{L}(\alpha) = \frac{|Z'Z|^{-1/2}}{(2\pi\sigma^2)^{(n-k)/2}}\exp\left(-\frac{\sum_{t=1}^{n}e_t^2}{2\sigma^2}\right). \tag{35}$$

The errors for this likelihood are calculated using the usual exponential smoothing recursions. When an observation is missing in period $t$, the usual error is replaced by $e_t = 0$. Relevant information about the past is carried forward through the period with the missing value by the state variables. Thus calculations in the presence of missing values is possible with exponential smoothing.

The determinant $|Z'Z\ |$depends on the smoothing parameter vector $\alpha$ so that estimates based on this exact likelihood differ from those obtained by minimising the traditional sum of squared errors. The findings of Kang (1975) and Davidson (1981) relating to an MA(1) process carry over to this context for the case of simple exponential smoothing. They indicate that in small samples, the differences between both estimates can be quite marked when the true value of $a$ is small. Moreover, exact maximum likelihood estimates display less bias than least-squares estimates.

A feature of the exact likelihood is that it involves the degrees of freedom $n-k$ instead of the sample size $n$. This gives a hint as to why exact likelihood estimates are less biased. The maximum exact likelihood estimator of the variance is

$$\widehat{\sigma}^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-k}. \tag{36}$$

Division by the degrees of freedom $n-k$ rather than the sample size $n$ means that this estimate is less biased than the estimate 29. This point is reinforced by examining the exact likelihood in the special case of a local level model. It is easily seen for this case that $Z'Z = \sum_{t=1}^{n} \delta^{2(j-1)}$. For a random walk with $\delta = 1$, the least-squares approach yields $\widehat{a_0} = y_1$ so that $e_1 = 0$. Furthermore, $Z'Z = n$ so that

$$\mathcal{L}\left(1, \sigma^2\right) = \frac{n^{-1/2}}{\left(2\pi\sigma^2\right)^{(n-1)/2}} \exp\left(-\frac{\sum_{t=2}^{n} e_t^2}{2\sigma^2}\right).$$

Ignoring the factor of proportionality $n^{-1/2}$, this is the usual likelihood for a random walk. The exact maximum likelihood estimate of $\sigma^2$ becomes $\widehat{\sigma}^2 = \frac{\sum_{t=2}^{n} e_t^2}{n-1}$. In contrast, the conditional likelihood yields the biased estimator $\widehat{\sigma}^2 = \frac{\sum_{t=2}^{n} e_t^2}{n}$. A divisor of $n$ makes little sense when there are only $n-1$ terms in the sum.

# 5 Model Selection

The choice between the various forms of exponential smoothing for forecasting from a particular set of data has been undertaken traditionally with approaches such as prediction validation that make no direct recourse to a statistical framework. Now that exponential smoothing has been provided with such a framework, the choice between method can be recast as a model selection problem. It opens up many traditional possibilities from time series analysis, that are new in the context of exponential smoothing, for the problem of choice.

One approach to model selection is to seek the model with the smallest estimated standard deviation $\widehat{\sigma}$. However, it is now widely recognised that good fit does not necessarily translate into good forecasts. An approach like this has a tendency to favour model complexity and projections form such models can be rather strange.

Forecast validation, where the end of a sample is reserved to evaluate the forecasting capacity of a model, is a way of circumventing the over-fitting problem. It has worked well in practice but whether it is the best way of model selection is open to question. Not using the final part of the sample for fitting means that the estimation error, by necessity, is larger than if the whole sample had have been used.

Likelihood might seem to be another possibility for choosing between models. The conditional likelihood is equivalent to the use of the estimates of $\sigma$ and so suffers from the same problem of overfitting. The exact likelihood cannot be used for a more subtle reason. More specifically, the factor of proportionality that was side-stepped in the above derivation of the exact likelihood has a term $\tau^{-d/2}$ where $\tau$ is an arbitrarily large number and $d$ is the number of *non-stationary* state variables (Ansley and Kohn, 1985). The exact likelihood of models with different values of $d$ are non-comparable.

The Akaike information criterion (Akaike, 1973) has become a common way of adjusting the likelihood to avoid over-fitting. It is tempting to calculate it with the exact likelihood but this does not work because of the comparability problem. It does, however, appear to work with the conditional likelihood - see Hyndman et. al. (2002) for details. The AIC has the advantage over prediction validation that estimation is undertaken with the entire sample. A recent comparative study (Billah, King, Snyder and Koehler) suggests that it is the better model selection criterion .

It would be wrong to conclude from this that the conditional likelihood should be used in preference to exact likelihood with exponential smoothing. The estimators obtained with the exact likelihood are less biased. So it seems that exponential smoothing should utilise both types of likelihoods: the exact likelihood for estimation, and the conditional likelihood for model selection in conjunction with the AIC.

# 6 Conclusions

Two things were done in this paper. First, the prediction restrictions on the smoothing parameters were derived from first principles; restrictions that are tighter than those associated with the traditional invertibility principle. In the process, some restrictions commonly used in practice were properly rationalised for the first time. A new restriction was derived for seasonal exponential smoothing. Second, the exact likelihood for the exponential smoothing models was derived for the first time. It was argued that it and its conditional counterpart can both play a useful role in exponential smoothing, one for estimation and the other for model selection based on the Akaike information criterion.

More generally, it has been shown that the framework espoused in this paper and its antecedents (Ord et. al., 1997; Hyndman et. al., 2002) has important implications for the future direction of time series analysis and forecasting. Time series analysis has been dominated by the Box-Jenkins approach but the findings of this paper confirms that the latter has inherent weaknesses that can only be avoided by a structural approach. Moreover, it has been shown that a structural approach need not be cast in terms of the common multi-disturbance state space framework that depends on the Kalman filter for the evaluation of the associated likelihood function. The equally general single source of error state space approach can be used instead, something that allows the relatively complex Kalman filter to be replaced with exponential smoothing. This paper therefore provides further tantalising support for the growing view that the central roles of Box-Jenkins analysis and the Kalman filter in time series analysis are questionable and that they should be replaced by the the enhanced version of exponential smoothing outlined in this paper.

# A Seed State Vector Estimates

The purpose of this section is to outline the theory for obtaining least squares estimates of the seed state vector. The basic strategy is to convert the general single source of error state space model (4) into an equivalent regression.

Equation (7) has the general form

$$X_t = P_t X_0 + Q_t \tag{37}$$

where $P_t$ is a matrix and $Q_t$ is a vector. Equations for recursively computing $P_t$ and $Q_t$ are obtained by substituting Equation (37) into Equation (6) to give

$$\begin{aligned} P_t &= DP_{t-1} & \text{(38a)} \\ Q_t &= DQ_{t-1} + \alpha Y_t & \text{(38b)} \end{aligned}$$

Substituting $t = 0$ into Equation 37 suggests that $P_0 = I$ and $Q_0 = 0$. It follows that

$$\begin{aligned} P_t &= D^t & \text{(39a)} \\ Q_t &= TQ_{t-1} + \alpha \left(Y_t - Q_{t-1}\right). & \text{(39b)} \end{aligned}$$

The Equation (39b) corresponds to the rule used in the general linear form of exponential smoothing. It is seeded with $Q_0 = 0$, so justifying the step in the body of the paper where exponential smoothing is applied with a zero seed vector.

Equation (37) may be substituted into the measurement Equation (4) to give $Y_t = h'P^{t-1}X_0 + h'Q_{t-1} + E_t$. A rearrangement of the terms results in the regression

$$Y_t^* = z_t'X_0 + E_t. \tag{40}$$

where $Y_t^* = Y_t - h'Q_{t-1}$ and $z_t = h'P^{t-1}$. This justifies the regression (27) where the the $Y_t^*$ correspond to the biased one-step ahead prediction errors.

The Equation (40) has stochastic regressors. The theory in Duncan and Horn (1972) applies so that the least squares estimates are best, unbiased linear predictors using these terms in the sense that they define them. Because of the linear relationships involved, the filtered values $\widehat{y}_t$ are best, linear unbiased predictors of the series values $y_t$.

When $D$ has eigenvalues all lying within the unit circle, it acts as a discount matrix in the sense that $D^t \to 0$. Then $z_t \to 0$, the implication being that the bias term $z_t'X_0$ disappears from (40), thereby ensuring that the biased error series converges to the unbiased errors. Whether or not this condition is satisfied depends on the values adopted by $\alpha$.

# B References

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in Second International Symposium on Information Theory, Akademiai Kiado: Budapest, pp. 267-281.

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

Ansley, C., and Kohn, R. (1985), "Estimation, Filtering, and Smoothing in State Space Models with Incompletely Specified Initial Conditions," The Annals of Statistics, 13, 1286-1316.

Billah, B., King, M. L., Snyder, R. D., Koehler, A. B. (2005), "Exponential Smoothing Model Selection for Forecasting", (unpublished paper).

Box, G. E. P., and Jenkins, G. M. (1976), Time Series Analysis: Forecasting and Control (Revised ed.), San Francisco: Holden Day.

Brown, R. G. (1959), Statistical Forecasting for Inventory Control, New York: McGraw-Hill.

Davidson, J. E. H. (1981), "Problems with the Estimation of Moving Average Processes," Journal of Econometrics, 16, 295-310.

Duncan, D. B., and Horn, S. D. (1972), "Linear Dynamic Regression from the Viewpoint of Regression Analysis," Journal of the American Statistical Association, 67, 815-821.

Durbin, J. (2000), "The Foreman Lecture: The State Space Approach to Time Series Analysis and Its Potential for Official Statistics," Australian & New Zealand Journal of Statistics, 42, 1-23.

Fildes, R., Hibon, M., Makridakis, S., and Meade, N. (1998), "Generalising About Univariate Forecast Methods: Further Empirical Evidence," International Journal of Forecasting, 14, 339-258.

Gardner, E. S. (1985), "Exponential Smoothing: The State of the Art," Journal of Forecasting, 4, 1-28.

Harrison, P. J., and Stevens, C. F. (1971), "A Bayesian Approach to Short-Term Forecasting," Operational Research Quarterly, 22, 341-362.

Harvey, A. C. (1991), Forecasting, Structural Time Series Models and the Kalman Filter, Cambridge: Cambridge University Press.

Harvey, A. C., and Koopman, S. (2000), "Signal Extraction and the Formulation of Unobserved Components Models," Econometrics Journal, 3, 84-107.

Holt, C. (2004), "Forecasting Seasonals and Trends by Exponentially Weighted Averages," International Journal of Forecasting, 20.

Hyndman, R., Koehler, A. B., Snyder, R. D., and Grose, S. (2002), "A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods," International Journal of Forecasting, 18, 439-454.

Hyndman, R. J., Akram, M., and Archibald, B. (2003), "Invertibility Conditions for Exponential Smoothing Models," Department of Econometrics and Business Statistics, Monash University, Working Paper Series.

Johnston, F. R., and Harrision, P. J. (1986), "The Variance of Lead-Time Demand," Journal of the Operational Research Society, 37, 303-398.

Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," Journal of Basic Engineering, Transactions of the ASME, Series D, 82, 35-45.

Kalman, R. E., and Bucy, R. S. (1961), "New Results in Linear Filtering and Prediction Theory," Journal of Basic Engineering, Transactions of the ASME, Series D, 83, 95-108.

Kang, K. M. (1975), "A Comparison of Estimators for Moving Average Processes," Technical, Australian Bureau of Statistics.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (1998), Forecasting: Methods and Applications, New York: John Wiley & Sons.

Ord, J. K., Koehler, A. B., and Snyder, R. D. (1987), "Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models," Journal of the American Statistical Association, 92, 1621-1629.

Snyder, R. D. (1985), "Recursive Estimation of Dynamic Linear Models," Journal of the Royal Statistical Society: Series B, 47, 272 276.

Snyder, R. D., Koehler, A. B., and Ord, J. K. (1999), "Lead Time Demand for Simple Exponential Smoothing: An Adjustment Factor for the Standard Deviation," Journal of the Operational Research Society, 50, 1079-1082.

Winters, P. R. (1960), "Forecasting Sales by Exponentially Weighted Moving Averages," Management Science, 1960, 324-342.