J. R. Statist. Soc. B (1999)
61, *Part* 1, *pp.* 39–50

Understanding exponential smoothing via kernel regression

I. Gijbels,

Université Catholique de Louvain, Louvain-la-Neuve, Belgium

A. Pope

University of Newcastle, Callaghan, Australia

and M. P. Wand

Harvard University, Boston, USA

[Received April 1997. Final revision June 1998]

Summary. Exponential smoothing is the most common model-free means of forecasting a future realization of a time series. It requires the specification of a smoothing factor which is usually chosen from the data to minimize the average squared residual of previous one-step-ahead forecasts. In this paper we show that exponential smoothing can be put into a nonparametric regression framework and gain some interesting insights into its performance through this interpretation. We also use theoretical developments from the kernel regression field to derive, for the first time, asymptotic properties of exponential smoothing forecasters.

Keywords: Bandwidth selection; Cross-validation; Dependent errors regression; Kernel smoothing; Limiting distribution; Local polynomial

1. Introduction

Exponential smoothing is an elementary means of forecasting a future realization from a time series (see for example Harvey (1989)). The most basic exponential smoother is the exponentially weighted moving average (EWMA). It has the attraction of being model free and based on a simple algebraic formula, in terms of a smoothing factor, and therefore enjoys common usage as an *ad hoc* forecasting procedure. Thus it is similar in spirit to nonparametric regression where, in its simplest form, the goal is to estimate the underlying trend in a scatterplot without the use of restrictive models. What is generally not recognized is that exponential smoothing can be viewed as a special type of nonparametric regression procedure where the fitting at a particular location uses only data to the left of that location. In fact the EWMA is virtually identical with the Nadaraya–Watson kernel estimator with a kernel function that is 0 in its positive arguments, something which we refer to as a 'half-kernel'. Moreover, the most common data-based choice of the smoothing factor in exponential smoothing is identical with the use of the cross-validation technique from nonparametric regression.

Address for correspondence: M. P. Wand, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA.
E-mail: mwand@hsph.harvard.edu

1369–7412/99/61039

In this paper we take advantage of these links and recent theoretical developments in nonparametric regression to provide some useful insights into the behaviour of exponential smoothing. For example, unlike the nonparametric regression problem, it will be seen that the common smoothing parameter choice is not adversely affected by correlated errors and behaves relatively sensibly for the forecasting problem. This is reflected in the asymptotic behaviour of exponential smoothing, which we derive in a very general time series framework.

This study has benefited greatly from recent contributions to the theory of cross-validatory smoothing parameter selection in nonparametric regression: Härdle *et al.* (1988), Altman (1990), Hart (1991), Chu and Marron (1991) and Hart (1994). Li and Heckman (1996) recently developed a nonparametric regression-type forecaster and studied the performance of cross-validation in that context.

In Section 2 we describe the commonality between exponential smoothing and local polynomial regression. Section 3 contains theory on exponential smoothing and the minimum error sum-of-squares smoothing factor. We discuss some alternative smoothing factor choices in Section 4. All the proofs are given in Appendix A.

2. Exponential smoothing and local polynomial regression

Let $Y=(Y_1, \ldots, Y_T)^{\mathrm{T}}$ be a time series observed at equally spaced time points $x_1, \ldots, x_T$. We consider the problem of using these data to forecast Y_{T+1}. One of the simplest model-free approaches to this problem is the EWMA. It forecasts Y_{T+1} by using a weighted average of past observations with geometrically declining weights:

$$\hat{Y}_{T+1}=(1-\omega)\sum_{j=0}^{T-1}\omega^j Y_{T-j} \tag{1}$$

where $0<\omega<1$ is commonly referred to as the *smoothing constant* (e.g. Harvey (1989)). An advantage of using geometric weights is that equation (1) admits the recurrence formula

$$\hat{Y}_{T+1}=\omega\hat{Y}_T+(1-\omega)Y_T \tag{2}$$

with $\hat{Y}_1\equiv Y_1$. Notice that the sum of the weights in equation (1) is $1-\omega^T$ which is close to 1 for large T. A simple adjustment that has weights summing exactly to 1 is

$$\hat{Y}_{T+1}=\sum_{j=0}^{T-1}\omega^j Y_{T-j}\bigg/\sum_{j=0}^{T-1}\omega^j. \tag{3}$$

If we define $h=-(x_T-x_1)/(T-1)\log_{\mathrm{e}}(\omega)$ and $K_{\mathrm{e}}(u)=\exp(u)\mathbf{1}_{\{u\leqslant 0\}}$ then it is easily shown that equation (3) can be rewritten as

$$\hat{Y}_{T+1}=\sum_{t=1}^{T}K_{\mathrm{e}}\left(\frac{x_t-x_{T+1}}{h}\right)Y_t\bigg/\sum_{t=1}^{T}K_{\mathrm{e}}\left(\frac{x_t-x_{T+1}}{h}\right). \tag{4}$$

This shows that the normalized exponential smooth forecast is equivalent to a Nadaraya–Watson, or zero-degree local polynomial, kernel estimate at x_{T+1} with an exponential kernel and *bandwidth h* (see for example Wand and Jones (1995)).

A natural extension of equation (4) is to general degree local polynomial kernel estimates:

$$\hat{Y}_{T+1}=e_1^{\mathrm{T}}\{X_p(x_{T+1})^{\mathrm{T}}W_{\mathrm{e},h}(x_{T+1})X_p(x_{T+1})\}^{-1}X_p(x_{T+1})^{\mathrm{T}}W_{\mathrm{e},h}(x_{T+1})Y, \qquad T\geqslant p+1, \tag{5}$$

where

$$X_p(x) = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_T - x & \cdots & (x_T - x)^p \end{pmatrix}, \qquad W_{e,h}(x) = \operatorname*{diag}_{1 \leqslant u \leqslant T} \left\{ K_e\left(\frac{x_u - x}{h} \right) \right\} \tag{6}$$

and e_t is the column vector with 1 in the tth position and 0s elsewhere. The case $p = 1$ is often used in the forecasting literature for handling a local linear trend, where it appears as the technique of double-exponential smoothing (e.g. Harvey (1989), pages 27–28). Fig. 1 illustrates the local linear exponential smoothing forecast.

The equivalent bandwidth h is a monotone transformation of the smoothing factor ω. Their relationship is depicted in Fig. 2 for the case $x_t = t$. We can see from this that $\omega = 0$ corresponds to $h = 0$ (no smoothing) and $\omega = 1$ corresponds to $h = \infty$ (fitting a global pth-degree polynomial).

The most popular approach to automatic smoothing parameter choice in exponential smoothing is to take h, or ω, to minimize the average of the squared residuals (ASR) of the previous one-step-ahead forecasts. Thus we define the error sum-of-squares choice of h, $\hat{h}_{\mathrm{ASR}}$, as

$$\hat{h}_{\mathrm{ASR}} = \operatorname*{arg\,min}_{h \geqslant 0} \{\mathrm{ASR}(h)\}$$

where

$$\mathrm{ASR}(h) = T^{-1} \sum_{t=1}^{T} (\hat{Y}_t - Y_t)^2.$$

In the next section we show that this is equivalent to choosing h by cross-validation.

It is known that EWMA forecasting is optimal (i.e. is the minimum mean-square error predictor) for the state space (or structural) model:

$$\begin{aligned} Y_t &= \mu_t + \epsilon_t, \\ \mu_t &= \mu_{t-1} + \eta_t, \end{aligned} \tag{7}$$

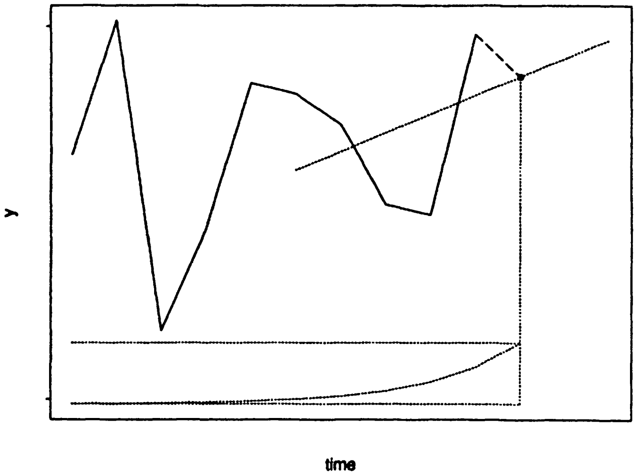

Fig. 1. Pictorial representation of $\hat{Y}_{T+1}$ of degree $p = 1$ (··········, fit by weighted least squares, with weights proportional to the height of the exponential function shown at the base of the plot)

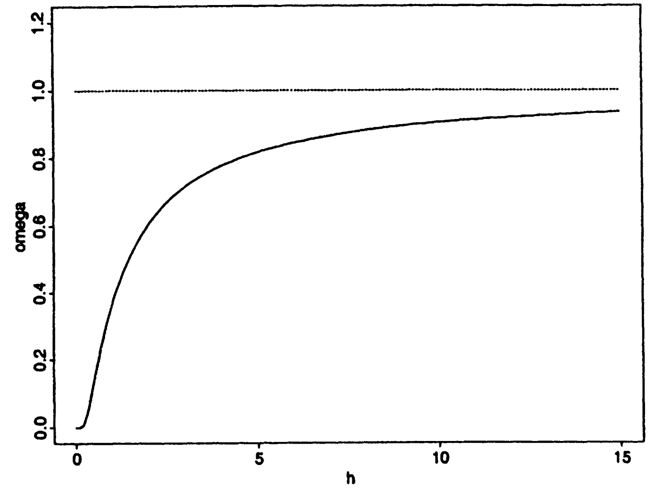

Fig. 2. Relationship between ω and h

where ϵ_t and η_t are independent white noise processes. This model is sometimes referred to as a random walk plus noise. The optimality may be seen by recognizing that the EWMA procedure reduces to the Kalman filter in this case (Harvey (1989), p. 175), and it is well known that the Kalman filter enjoys many optimality properties for state space models. By the same token, since it is clearly not the same as the Kalman filter in other cases, the EWMA is not in general an optimal procedure. This has been the source of considerable discussion in the literature about the appropriateness of using EWMA procedures when the model has not been correctly identified. (See Gardner (1985) and Newbold and Bos (1989) for example.) The optimal weights in these cases come from the Kalman filter, therefore, and reflect the error structure. However, a clear message from the theory of kernel smoothing is that within broad limits the choice of the kernel shape is not important, and what matters much more is choosing the correct bandwidth. This reinforces the thesis of this paper that the asymptotic theory of smoothing factor selection is important.

3. Bandwidth selection, including asymptotic theory

We now suppose that the data are generated according to the model

$$Y_t = m(x_t) + \epsilon_t, \qquad t = 1, \ldots, T,$$

where $x_t = t/T$ and the ϵ_t are realizations of a zero-mean causal autoregressive moving average (ARMA) process (see definitions 3.1.2 and 3.1.3 of Brockwell and Davis (1991)) with covariance function $\gamma(k) = E(\epsilon_t \epsilon_{t-k})$. The signal m is deterministic here. In structural time series models such as model (7) the trend μ_t is random. As we do not intend to take averages over the space of possible signals m, we do not consider the random signal model. Although the case that we consider and the random signal model are subtly different, they have in common the ability to incorporate our uncertainty about the form of the model. In the random walk plus noise model (7) this is achieved through the white noise process η_t, which makes μ_t random; in the situation that we investigate, the uncertainty is handled through a nonparametric regression approach. In both situations the local level or local trend is a nuisance parameter for the forecasting problem, but they can both be estimated as well if we desire. Note that $T \to \infty$ corresponds to the number of observations becoming denser in the interval [0, 1] with the dependence structure remaining the same.

Let

$$\hat{m}(x) = e_1^{\mathrm{T}} \{X_p(x)^{\mathrm{T}} W_h(x) X_p(x)\}^{-1} X_p(x)^{\mathrm{T}} W_h(x) Y$$

where $X_p(x)$ is the same as in equations (6) with x_{T+1} replaced by x and

$$W_h(x) = \operatorname*{diag}_{1 \leqslant j \leqslant T} \left\{ K\left(\frac{x_j - x}{h}\right) \right\}$$

for a general kernel K with the properties $K(x) = 0$, $x > 0$, and $\int K = 1$. We shall refer to such a kernel as a half-kernel. With this notation we can rewrite equation (5) as $\hat{Y}_{T+1} = \hat{m}(x_{T+1})$.

Although our main interest lies in forecasting Y_{T+1}, we can think of $\hat{m}(x)$ as being a nonparametric estimate of $m(x)$ for any $x \in [0, x_{T+1}]$. But, since K is a half-kernel, the estimate is always based on data to the left of x or at x itself. Our first results concern the bias and variance of $\hat{m}(x)$.

Let N_p be the $(p+1) \times (p+1)$ matrix having (i, j) entry equal to $\int u^{i+j-2} K(u) \, \mathrm{d}u$ and $M_p(u)$ be the same as N_p, but with the first column replaced by $(1, u, \ldots, u^p)^{\mathrm{T}}$. Then, as in Ruppert and Wand (1994), define the half-kernel

$$K_p(u) = \{|M_p(u)| / |N_p|\} K(u).$$

Theorem 1. Under assumptions given in Appendix A, and assuming that $h = h_T \to 0$ and $Th \to \infty$ as $T \to \infty$, for all $x \in [0, 1]$,

$$E\{\hat{m}(x) - m(x)\} = \frac{\int u^{p+1} K_p(u) \, \mathrm{d}u}{(p+1)!} m^{(p+1)}(x) h^{p+1} + o(h^{p+1})$$

and

$$\operatorname{var}\{\hat{m}(x)\} = \int K_p(u)^2 \, \mathrm{d}u \left\{ \sum_{k=-\infty}^{\infty} \gamma(k) \right\} (Th)^{-1} + o\{(Th)^{-1}\}.$$

Unlike ordinary local polynomial kernel smoothing, the bias of $\hat{m}(x)$ has leading term proportional to $h^{p+1} m^{(p+1)}(x)$ for both odd and even p. This is because estimation with a half-kernel corresponds to estimation at a boundary for all x (see p. 128 of Wand and Jones (1995)). For full kernels and p even this term vanishes when x is away from the boundary and a more complicated $O(h^{p+2})$ term becomes the leading term.

Recent research in local polynomial smoothing (e.g. Fan and Gijbels (1992), Hastie and Loader (1993) and Cheng *et al.* (1997)) has established that local lines handle boundary effects considerably better than local constants do. Since forecasting always involves smoothing near a boundary we would expect the local linear exponential smoother to outperform the ordinary EWMA. We may also note that (two-sided) smoothing (as opposed to filtering) estimators of the signal based on structural time series models such as model (7) may also be regarded as nonparametric regression estimators and the theory of bandwidth choice taken over to them as well. Harvey (1989), p. 230, discusses this smoothing approach. Because of the two-sided nature of the problem, the advantage of the local linear estimator over the local constant estimator will only be seen at the ends of the data, but for the forecasting problem it is present at each time point.

A common measure of the global error in $\hat{m}$ as an estimate of m is the average squared error (ASE)

$$\mathrm{ASE}(h) = T^{-1} \sum_{t=1}^{T} \{\hat{m}(x_t) - m(x_t)\}^2.$$

Note that ASE(h) differs from

$$\mathrm{ASR}(h) = T^{-1} \sum_{t=1}^{T} \{\hat{m}(x_t) - Y_t\}^2$$

in that the former is concerned with the distance between $\hat{m}$ and m, whereas the latter is concerned with the distance between $\hat{m}$ and the time series. In nonparametric regression it is convenient to work with

$$\mathrm{MASE}(h) = E\{\mathrm{ASE}(h)\}.$$

The bandwidth that minimizes this quantity for a particular m is denoted by h_{MASE}. Let

$$V_p = \int K_p(u)^2 \,\mathrm{d}u \sum_{k=-\infty}^{\infty} \gamma(k)$$

and

$$B_p^2 = \left\{ \int u^{p+1} K_p(u) \,\mathrm{d}u \Big/ (p+1)! \right\}^2 \int m^{(p+1)}(x)^2 \,\mathrm{d}x.$$

Then an easily derived corollary of theorem 1 is as follows.

Corollary 1. Under the assumptions of theorem 1,

$$\mathrm{MASE}(h) = V_p(Th)^{-1} + B_p^2 h^{2p+2} + o\{(Th)^{-1} + h^{2p+2}\}$$

and

$$h_{\mathrm{MASE}} = C_{\mathrm{MASE}} T^{-1/(2p+3)} \{1 + o(1)\}$$

where $C_{\mathrm{MASE}} = \{V_p/(2p+2)B_p^2\}^{1/(2p+3)}$.

We now return to the automatic forecasting problem with smoothing factor choice based on $\hat{h}_{\mathrm{ASR}}$. For $p = 0$ note that, because K is a half-kernel,

$$\hat{Y}_t = \sum_{j=1}^{t-1} K\left(\frac{x_j - x_t}{h}\right) Y_j \Big/ \sum_{j=1}^{t-1} K\left(\frac{x_j - x_t}{h}\right) = \sum_{j \neq t} K\left(\frac{x_j - x_t}{h}\right) Y_j \Big/ \sum_{j \neq t} K\left(\frac{x_j - x_t}{h}\right)$$

so $\hat{h}_{\mathrm{ASR}}$ minimizes $\Sigma_{t=1}^{T} \{Y_t - \hat{m}_{-t}(x_t)\}^2$ where $\hat{m}_{-t}(x)$ is the same as $\hat{m}(x)$, but based on the data with (x_t, Y_t) omitted. The same result can be easily shown to hold for general p. Therefore, $\hat{h}_{\mathrm{ASR}}$ is the same as cross-validation with a half-kernel and so its asymptotic distribution can be obtained by extending the results of Chu and Marron (1991) to half-kernels and higher degree fits. Let

$$(f * g)(x) = \int f(u)\, g(x-u) \,\mathrm{d}u$$

denote the convolution of two functions f and g.

Theorem 2. Under assumptions given in Appendix A,

$$T^{1/(4p+6)}\left(\frac{\hat{h}_{\mathrm{ASR}}}{h_{\mathrm{MASE}}} - \frac{C_{\mathrm{ASR}}}{C_{\mathrm{MASE}}}\right) \xrightarrow{\mathrm{D}} N(0, \sigma^2_{\mathrm{ASR}})$$

where

$$C_{\mathrm{ASR}} = \left[\left\{V_p - 2\,K_p(0)\sum_{k=1}^{\infty}\gamma(k)\right\}\Big/(2p+2)B_p^2\right]^{1/(2p+3)},$$

$$\sigma^2_{\mathrm{ASR}} = \frac{8(C_{\mathrm{MASE}}/C_{\mathrm{ASR}})^{4p+3}\left\{\sum_{k=-\infty}^{\infty}\gamma(k)\right\}^{1/(2p+3)}\int Q_p^2}{(2p+3)^2\left\{(2p+2)B_p^2\left(\int K_p^2\right)^{4p+5}\right\}^{1/(2p+3)}},$$

$$Q_p = K_p * K_p^- - \tfrac{1}{2}K_p * L_p^- - \tfrac{1}{2}L_p * K_p^- - (K_p - L_p)$$

and

$$K_p^-(u) = K_p(-u),$$

$$L_p(u) = -u\,K_p'(u).$$

It follows from theorem 2 that, for large T,

$$\hat{h}_{\mathrm{ASR}} \simeq C_{\mathrm{ASR}}T^{-1/(2p+3)}.$$

Useful insight into the behaviour of $\hat{h}_{\mathrm{ASR}}$ can be obtained by considering the cases

(a) independent errors,
(b) positively correlated errors and
(c) negatively correlated errors.

If the errors are independent then $C_{\mathrm{ASR}} = C_{\mathrm{MASE}}$ so $\hat{h}_{\mathrm{ASR}}$ has asymptotic behaviour similar to that of h_{MASE}, the globally optimal bandwidth for estimating the mean function m across the time series. In the case of independence, the best mean-squared error predictor of Y_{T+1} is $m(x_{T+1})$ so the quality of $\hat{h}_{\mathrm{ASR}}$ depends on how appropriate the *globally* optimal bandwidth h_{MASE} is for *local* estimation of m at x_{T+1} (see for example Fan and Gijbels (1995) or Wand and Jones (1995), section 2.9).

If m has relatively uniform curvature then the difference between the global and local optimal bandwidths is not very large and therefore $\hat{h}_{\mathrm{ASR}}$ should perform reasonably well. But if the curvature in m varies significantly across the series then the quality of $\hat{h}_{\mathrm{ASR}}$ worsens. A localized bandwidth choice, based on the later observations in the series, would be expected to perform better.

Suppose instead that the errors exhibit positive correlation. Cross-validation is known to perform poorly for mean estimation in nonparametric regression (e.g. Hart (1991)). This is essentially because cross-validation mistakes the smoothness of the series caused by the positive correlations for low variability. This shows up in theorem 2 with $\Sigma_k\,\gamma(k)$ being positive and therefore $\hat{h}_{\text{ASR}}$ usually being smaller than h_{MASE}. This behaviour is disastrous for mean estimation, but it is the correct type of behaviour in the forecasting context since averaging over a small number of past observations is more likely to be close to the next value in the series when there are positive correlations. A concrete example of this is the case of positive serial correlation where the errors satisfy $\epsilon_t = \rho\epsilon_{t-1} + u_t$ for some $0 < \rho < 1$, with u_t a white noise process. In this case the minimum mean-squared error predictor of Y_{T+1} is

$$m(x_{T+1}) + \rho\{Y_T - m(x_T)\} = (1-\rho)\,m(x_T) + \rho Y_T + m(x_{T+1}) - m(x_T).$$

If m is relatively flat near x_T then the main component of this quantity is $(1-\rho)\,m(x_T) + \rho Y_T$. When ρ is close to 1 then the best predictor is close to Y_T. This corresponds to interpolation so a choice of a small bandwidth is desirable. From theorem 2 we have, for $K = K_{\text{e}}$,

$$\frac{\hat{h}_{\text{ASR}}}{h_{\text{MASE}}} \simeq \left(\frac{1-3\rho}{1+\rho}\right)^{1/(2p+3)}.$$

For positive ρ this ratio is close to 0 or even negative. This means that $\hat{h}_{\text{ASR}}$ is near the left-hand end of H_T (defined in Appendix A) asymptotically, which is an indication of $\hat{h}_{\text{ASR}}$ choosing small bandwidths in the case of positive serially correlated data. (Theorem 2 of Hart (1991) gives a more concise quantification of this behaviour.) Therefore, for the forecasting problem, cross-validation *does* provide about the right amount of smoothing in the case of positive serially correlated data.

If there are negative correlations then the opposite tends to happen. Cross-validation chooses a larger bandwidth to smooth out the extra variability caused by the negative correlations. For the nonparametric regression problem this is not good if the mean has varied considerably over the same time interval. But for forecasting it is advantageous to average over a larger span when there is so much variability.

The basic difference is that, in forecasting, we do not care about what is attributable to the mean and what is attributable to the correlation, whereas in nonparametric regression the mean is of central interest with the correlations being a nuisance. Therefore, forecasting is somewhat easier than nonparametric regression in the presence of correlated errors.

Since the most common form of exponential smoothing involves $p = 0$ and $K = K_{\text{e}}$ it is also of interest to see what the asymptotic distribution of $\hat{h}_{\text{ASR}}$ is in this case.

Corollary 2. For the EWMA forecast procedure (1), under assumptions given in Appendix A,

$$T^{1/6}\left[\frac{\hat{h}_{\text{ASR}}}{h_{\text{MASE}}} - \left\{\frac{\sum\limits_{k=-\infty}^{\infty}\gamma(k) - 4\sum\limits_{k=1}^{\infty}\gamma(k)}{\sum\limits_{k=-\infty}^{\infty}\gamma(k)}\right\}^{1/3}\right] \xrightarrow{\text{D}} N(0, \sigma^2_{\text{ASR}})$$

where

$$\sigma^2_{\text{ASR}} = \frac{5\left\{\sum\limits_{k=-\infty}^{\infty}\gamma(k)\right\}^{4/3}}{9(2^{2/3})\left\{\sum\limits_{k=-\infty}^{\infty}\gamma(k) - 4\sum\limits_{k=1}^{\infty}\gamma(k)\right\}\left\{\int_0^1 m'(x)^2\,\mathrm{d}x\right\}^{1/3}}.$$

4. Local smoothing factor choices

There are alternatives to $\hat{h}_{\mathrm{ASR}}$ which, depending on one's belief about the series, may lead to better forecasts. For example if the errors appear to be heteroscedastic or if the trend varies in curvature then it may be beneficial to localize the ASR criterion towards the end of the time series by inserting a weight function into the error sum of squares:

$$T^{-1}\sum_{t=1}^{T}(\hat{Y}_t - Y_t)^2\, w\left(\frac{x_t - x_T}{\delta}\right)$$

where w is non-increasing in its negative argument and $\delta > 0$. This adaptation is a version of local cross-validation studied by Mielniczuk *et al.* (1989) and Hall and Schucany (1989). Other local smoothing parameter choices with good boundary properties, such as those proposed by Fan and Gijbels (1995) and Ruppert (1997), could also be adapted to the correlated errors forecast setting. There is clearly plenty of computing and simulating that could be done along these lines to see whether modern bandwidth selection algorithms offer a significant improvement over the cross-validatory $\hat{h}_{\mathrm{ASR}}$. Preliminary simulations involving an appropriate modification of Ruppert's (1997) algorithm have not yet yielded any significant improvements.

5. Seasonality

Since seasonality is important in applications, the simple EWMA approach to forecasting described above is commonly extended to accommodate seasonal effects by adding recursions similar to equation (2) operating at a fixed lag (the seasonal period). This may be done in several different ways, depending on whether the seasonal effect is additive or multiplicative for instance, and the details may be messy, so we do not discuss these methods here. There are descriptions in Harvey (1989), section 2.2, Bowerman and O'Connell (1993), chapter 8, and Newbold and Bos (1994), chapter 6, for example of how this is done. (These references also contain descriptions of the EWMA in practice.) Our methods and general conclusions carry over to the seasonal case. The extra features are the existence of several parameters to be determined and the possibly complex interaction between the recursions.

6. Conclusions

Our main insights have been that

(a) simple exponential and double-exponential smoothing are types of kernel regression,
(b) the usual data-driven choice of smoothing parameter for exponential smoothing is equivalent to cross-validation,
(c) as a global bandwidth choice this procedure is difficult to beat, at least asymptotically, and
(d) forecasting is easier than regression when the errors are correlated.

These insights tend to support the usual exponential smoothing practices, including the cross-validated choice of smoothing parameter, and they provide another explanation for the effectiveness of double-exponential smoothing, through its inheritance of the superior boundary behaviour of the local linear regression estimator. Thus our message can be interpreted as saying that exponential smoothing is, as everybody suspected, a sensible procedure. However, for the first time we have now been able to provide theory in support of the 'commonsense' view that minimizing the historical average squared prediction error is a good way to choose the smoothing parameter. In addition, we expect that other insights from kernel regression theory can be incorporated: for example, local bandwidth choices may provide an improved forecasting procedure.

Acknowledgements

We are grateful for the excellent comments provided by two referees which led to improved insights and conclusions from our results. We also thank Chih Kang Chu for providing us with a copy of his doctoral dissertation. Research of the first author was supported by Projet d'Actions de Recherche Concertées (93/98-164), Fonds National de la Recherche Scientifique grant 1.5.001.95F and the European Human Capital and Mobility Programme (grant CHRX-CT94-0693). Research of the third author was supported by a grant from l'Institut de Mathématique Pure et Appliquée of Université Catholique de Louvain, Belgium, and by Sonderforschungsbereich 373 at Humboldt University, Germany.

Appendix A

A.1. Assumptions

The assumptions which we make for the theorems given in this paper are as follows.

(a) $m^{(p+1)}$ is continuous and square integrable on (0, 1).

(b) K is square integrable and has compact support on the interval $[-\tau, 0]$ or $\tau > 0$ such that $K(0) > 0$. Also, K is $p+1$ times differentiable on its support and $K^{(p+1)}$ is Lipschitz continuous.

(c) Data are available in the interval $[-h\tau, 0]$ and are used in the construction of $\hat{m}(x)$. This condition ensures that there are no left-hand boundary effects.

(d) The errors ϵ_t are obtained by application of a causal linear filter (see definition 3.1.3 of Brockwell and Davis (1991)) to independent and identically distributed random variables with mean 0 and all moments finite.

(e) The autocovariance function γ of ϵ_t satisfies $0 < \Sigma_{k=-\infty}^{\infty} |\gamma(k)| < \infty$.

(f) The minimizer of $\Sigma_{t=1}^{T} (\hat{Y}_t - Y_t)^2$ is searched on the interval $H_T = [aT^{-1/(2p+3)}, bT^{-1/(2p+3)}]$ for each T, for some $b > a > 0$.

A.2. Proof of theorem 1

Theorem 1 is a relatively straightforward extension of theorem 4.1 of Ruppert and Wand (1994). The main difference is that K is now a half-kernel and so the odd order moments do not vanish. Additionally, the generalization from independence to causal ARMA dependence means that the variance expression has multiplicative factor $\Sigma_{k=-\infty}^{\infty} \gamma(k)$ rather than the variance of the errors. See, for example, Hart (1991) for the details of such an extension.

A.3. Proof of theorem 2

Let $\text{ASR}(h) = T^{-1} \Sigma_{t=1}^{T} (\hat{Y}_t - Y_t)^2$ denote the ASR. The notation $U_T = o_u(V_T)$ is defined to mean that, as $T \to \infty$, $|U_T/V_T| \to 0$ almost surely and uniformly on H_T. Following Chu and Marron (1991) we note that

$$\text{ASR}(h) = \text{ASE}(h) - 2\,\text{cross}(h) + T^{-1} \sum_{t=1}^{T} \epsilon_t^2 + R(h)$$

where $\text{cross}(h) = T^{-1} \Sigma_{t=1}^{T} \epsilon_t \{\hat{Y}_t - m(x_t)\}$ and

$$R(h) = T^{-1} \sum_{t=1}^{T} \{\hat{Y}_t - \hat{m}(x_t)\}\{\hat{Y}_t - m(x_t) + \hat{m}(x_t) - m(x_t)\}.$$

Through straightforward calculations we have, as $T \to \infty$, $R(h) = o_u\{\mathrm{MASE}(h)\}$ and

$$\mathrm{cross}(h) = (Th)^{-1}\, K_p(0) \sum_{k=1}^{\infty} \gamma(k) + o_u\{\mathrm{MASE}(h)\}.$$

Therefore

$$\mathrm{ASR}(h) = (Th)^{-1}\left\{ V_p - 2\,K_p(0) \sum_{k=1}^{\infty} \gamma(k) \right\} + h^{2p+2} B_p^2 + o_u\{(Th)^{-1} + h^{2p+2}\}$$

and so

$$\hat{h}_{\mathrm{ASR}} = C_{\mathrm{ASR}} T^{-1/(2p+3)}\{1 + o_u(1)\}.$$

Similarly, the minimizer of the asymptotic mean ASR,

$$\mathrm{AMASR}(h) = E\left\{ \mathrm{ASE}(h) - 2\,\mathrm{cross}(h) + T^{-1} \sum_{t=1}^{T} \epsilon_t^2 \right\},$$

is

$$h_{\mathrm{AMASR}} = C_{\mathrm{ASR}} T^{-1/(2p+3)}\{1 + o(1)\}.$$

Noting that

$$\mathrm{ASR}(h) = \mathrm{AMASR}(h) + G(h) + R(h) + T^{-1} \sum_{t=1}^{T} \{\epsilon_t^2 - E(\epsilon_t^2)\}$$

where

$$G(h) = \mathrm{ASE}(h) - 2\,\mathrm{cross}(h) - E\{\mathrm{ASE}(h) - 2\,\mathrm{cross}(h)\}$$

we have

$$0 = \mathrm{ASR}'(\hat{h}_{\mathrm{ASR}}) = (\hat{h}_{\mathrm{ASR}} - h_{\mathrm{AMASR}})\, \mathrm{AMASR}''(h^*) + G'(\hat{h}_{\mathrm{ASR}}) + R'(\hat{h}_{\mathrm{ASR}}) \tag{8}$$

where h^* is between $\hat{h}_{\mathrm{ASR}}$ and h_{AMASR}. Using

$$\hat{h}_{\mathrm{ASR}} = h_{\mathrm{AMASR}}\{1 + o_u(1)\}$$

and noting that

$$\mathrm{AMASR}''(h) = 2(Th^3)^{-1}\left\{ V_p - 2\,K_p(0) \sum_{k=1}^{\infty} \gamma(k) \right\} + (2p+1)(2p+2) B_p^2 h^{2p} + o\{(Th^3)^{-1} + h^{2p}\}$$

we obtain

$$\mathrm{AMASR}''(h^*) = C_1 T^{-2/(2p+3)}\{1 + o_u(1)\}$$

where

$$C_1 = (2p+3)\left[\left\{ V_p - 2\,K_p(0) \sum_{k=1}^{\infty} \gamma(k) \right\}^{2p} \{(2p+2)B_p^2\}^3 \right]^{1/(2p+3)}.$$

Also,

$$G'(\hat{h}_{\mathrm{ASR}}) = G'(h_{\mathrm{AMASR}}) + o_p(T^{-(4p+3)/(4p+6)})$$

so multiplication of equation (8) by $T^{(4p+3)/(4p+6)}$ gives

$$0 = T^{3/(4p+6)}(\hat{h}_{\mathrm{ASR}} - h_{\mathrm{AMASR}})C_1 + T^{(4p+3)/(4p+6)}\, G'(h_{\mathrm{AMASR}}) + o_p(1).$$

Theorem 2 then follows from this and, for all $\alpha > 0$,

$$T^{(4p+5)/(4p+6)}\,G_1(\alpha T^{-1/(2p+3)}) \xrightarrow{\mathrm{D}} N\left[0,\ \frac{2}{\alpha}\left\{\sum_k \gamma(k)\right\}^2 \int Q_p^2\right]$$

where $G_1(h) = (h/2)\,G'(h)$. This result can be obtained by reworking the arguments of Härdle *et al.* (1988) for half-kernels, dependent errors and higher degree polynomial fits. A more detailed argument of this type (which treats dependent errors) can be found in Chu (1989).

References

Altman, N. S. (1990) Kernel smoothing of data with correlated errors. *J. Am. Statist. Ass.*, **85**, 749–759.

Bowerman, B. L. and O'Connell, R. T. (1993) *Forecasting and Time Series: an Applied Approach.* Belmont: Wadsworth.

Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*, 2nd edn. New York: Springer.

Cheng, M.-Y., Fan, J. and Marron, J. S. (1997) On automatic boundary corrections. *Ann. Statist.*, **25**, 1691–1708.

Chu, C. K. (1989) Some results in nonparametric regression. *PhD Dissertation.* Department of Statistics, University of North Carolina, Chapel Hill.

Chu, C. K. and Marron, J. S. (1991) Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, **19**, 1906–1918.

Fan, J. and Gijbels, I. (1992) Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008–2036.

———(1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Statist. Soc.* B, **57**, 371–394.

Gardner, E. S. (1985) Exponential smoothing: the state of the art. *J. Forecast.*, **41**, 1–28.

Hall, P. and Schucany, W. R. (1989) A local cross-validation algorithm. *Statist. Probab. Lett.*, **8**, 109–117.

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum (with discussion)? *J. Am. Statist. Ass.*, **83**, 86–101.

Hart, J. D. (1991) Kernel regression estimation with time series errors. *J. R. Statist. Soc.* B, **53**, 173–187.

———(1994) Automated kernel smoothing of dependent data by using time series cross-validation. *J. R. Statist. Soc.* B, **56**, 529–542.

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

Hastie, T. J. and Loader, C. (1993) Local regression: automatic kernel carpentry (with discussion). *Statist. Sci.*, **8**, 120–143.

Li, X. and Heckman, N. E. (1996) Local linear forecasting. *Technical Report 167.* Department of Statistics, University of British Columbia, Vancouver.

Mielniczuk, J., Sarda, P. and Vieu, P. (1989) Local data-driven bandwidth choice for density estimation. *J. Statist. Planng Inf.*, **23**, 53–69.

Newbold, P. and Bos, T. (1989) On exponential smoothing and the assumption of deterministic trend plus white noise data-generating models. *Int. J. Forecast.*, **5**, 523–527.

———(1994) *Introductory Business and Economic Forecasting*, 2nd edn. Cincinnati: South-Western.

Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Ass.*, **92**, 1049–1062.

Ruppert, D. and Wand, M. P. (1994) Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**, 1346–1370.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing.* London: Chapman and Hall.