The exponentially weighted moving average applied to the control and monitoring of varying sample sizes

J. E. Everett
Centre for Exploration Targeting,
The University of Western Australia, Australia

Abstract

The exponentially weighted moving average (EWMA) can be used to report the smoothed history of a production process, and has some considerable advantages over a simple moving average (MA). Discussion of these advantages includes comparison of the filter characteristics of the EWMA and MA in the frequency domain. It is shown that the EWMA provides a much smoother filter than does the MA, and the corresponding implications of this difference are examined in the time domain. In smoothing a production process, the successive entities being smoothed commonly have varying “weights”, where the weights may be such quantities as tonnage, value or time interval. Standard textbook treatments of moving averages and exponential smoothing are generally confined to equal spaced data of equal weight. Adapting the average to cope with items of varying weight is shown to be trivial for the case of MA, but is not so obvious for the EWMA. This paper shows how the exponential smoothing constant has to be adapted to provide a consistent EWMA. Applications of the EWMA in process control are discussed, with particular reference to quality control in the mining industry.

Keywords: quality control, forecasting, exponential smoothing, sample size.

1 Introduction

It is common to consider a series of observations, x_n, where each observation is equivalently spaced in time or distance or some other relevant dimension.

For forecasting and for system control purposes, it is useful to have some summary of the performance up to the n^{th} observation.

WIT Transactions on Modelling and Simulation, Vol 51, © 2011 WIT Press
www.witpress.com, ISSN 1743-355X (on-line)
doi:10.2495/CMEM110011

The summary could be calculated as the mean (M) of all the observations since the first one:

$$M_n= {}_1\Sigma^n x_m/n \tag{1}$$

Usually we are mainly interested in recent history, so a straight average over the entire history of observations. Two approaches are to consider either a Moving Average (MA), applying equal weight to the past k observations, or an Exponentially Weighted Moving Average (EWMA), where successively declining weights are applied as we go further back in history.

1.1 Moving average (MA)

Usually we are mainly interested in recent history, perhaps over the past k observations, so a moving average (MA) over those k observations would be more appropriate:

$$MA_n = {}_{m=0}\Sigma^{k-1} x_{n-m}/k \tag{2}$$

Figure 1 shows the uniform weights of 1/k that are applied to the past k observations.

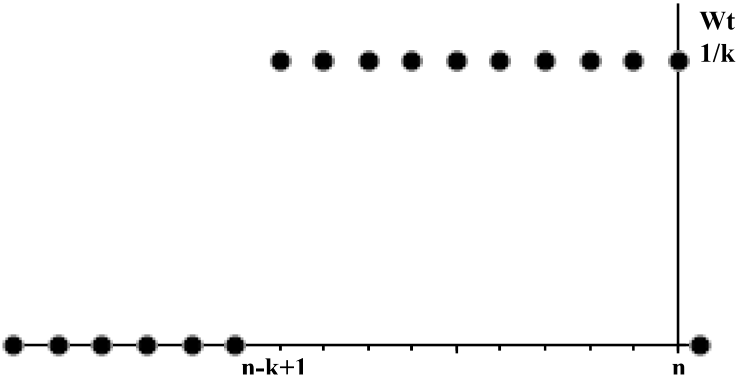

Figure 1: Moving average (MA) weights applied to recent data.

The moving average has the disadvantage that, for the first k intervals, each of the observations is treated as being of equal importance, but then is suddenly disregarded, as soon as it falls off the end of the data being averaged. This discontinuity has several disadvantages that will be discussed more fully in a later section.

1.2 Exponential smoothing (EWMA)

Exponential smoothing, across an “exponentially weighted moving average” (EWMA), provides a smoother means of averaging, where data becomes gradually less influential as it ages.

$$\text{EWMA}_n = S_n \quad = (1\text{-}\alpha)S_{n\text{-}1} + \alpha x_n$$
$$= (1\text{-}\alpha)((1\text{-}\alpha)S_{n\text{-}2} + \alpha x_{n\text{-}1}) + \alpha x_n$$
$$= {}_{m=0}\Sigma^{\text{Infinity}}\alpha(1\text{-}\alpha)^m x_{n\text{-}m} \qquad (3)$$

Figure 2 shows how the weights applied to earlier data die off exponentially as we go back through the data history.

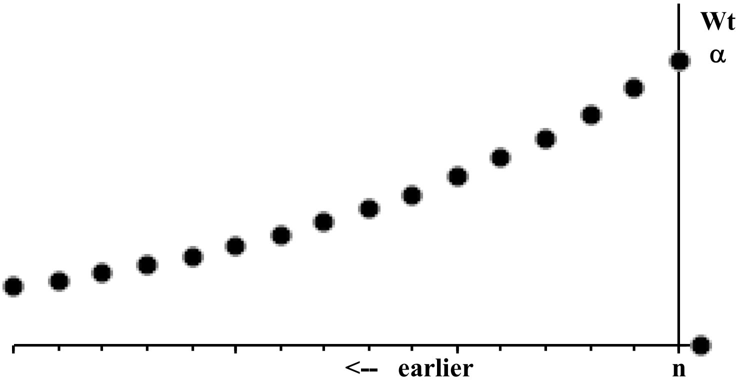

Figure 2: Exponential smoothing (EWMA) weights applied to recent data.

Exponential smoothing is commonly used in forecasting, and is generally described in forecasting textbooks. Diebold [1]) provides a good description.

It can be shown that the EWMA is a minimum mean squared error predictor when the true data generating process is ARIMA(0,1,1).ARIMA processes cover the very wide field of “Autoregressive Integrated Moving Average” processes, identified by Box and Jenkins [2]. As its parameters imply, an ARIMA(0,1,1) process is not autoregressive, but is first-order integrated moving average. Ramjee et al. [3] show that the EWMA method can also provide simple, yet useful, forecasts for other types of ARIMA processes.

Treatments in the literature are generally confined to equally spaced observations of equal weight, so that each new observation is of equal importance. However, it is commonly the case that the desired quality control or forecasting relates to observations that are of varying weight.

An example of this situation, with observations of varying weight, would be a mine’s production, shift by shift, of ore of varying tonnage and grade. In this particular example, the objective may be to forecast the grade of the next shift's production. Alternatively, the purpose may be to summarise the grade of the recent production so that the next shift's ore can be selected so as to restore the smoothed grade back to its target value. Everett [4] provides an example of this type of EWMA application in the iron ore mining industry.

Whether the averaging is being used to generate a forecast or to control a production system, it is being used to summarise recent behaviour. In doing so, it needs to respond to sustained changes in the data, but not be over sensitive to short-term variations. The averaging process is therefore being required to act as a low-pass filter.

Sections 2 and 3 will discuss more fully the advantages of an exponentially weighted EWMA over an MA. Comparing the Fourier transforms of the filters enables their performance as low-pass filters to be evaluated, and clearly demonstrates the advantages of the EWMA over the MA.

Adjustment for varying sample size is comparatively straightforward for the MA. For the EWMA, the adjustment for varying sample size is not so obvious, and appears to have been neglected in the literature.

Section 4 will consider the appropriate treatment of data where sample sizes vary.

Both for MA and EWMA, the choice of weighting constant, and the consequent length over which the data is averaged, depends upon the purpose for which the average is being used.

Section 5 considers the choice of the alpha constant for an EWMA, and its relation to the length of a comparable MA.

2 MA and EWMA compared

Figure 3 shows a signal x with a wavelet disturbance, first up and then down.

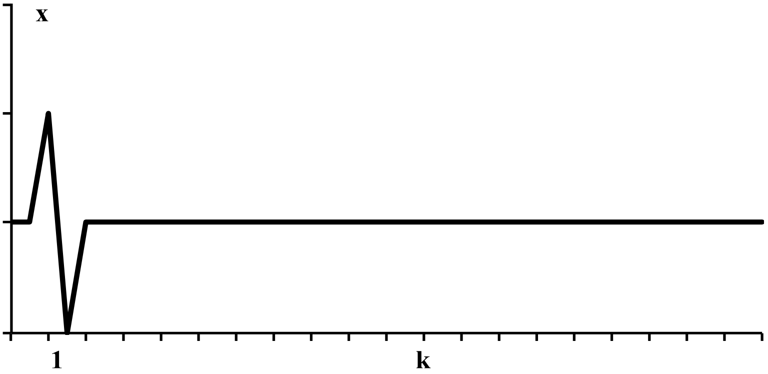

Figure 3: Signal "x" with a wavelet disturbance.

Figure 4 shows the effects of applying a Moving Average (MA) and Exponential Smoothing (EWMA) to this signal x.

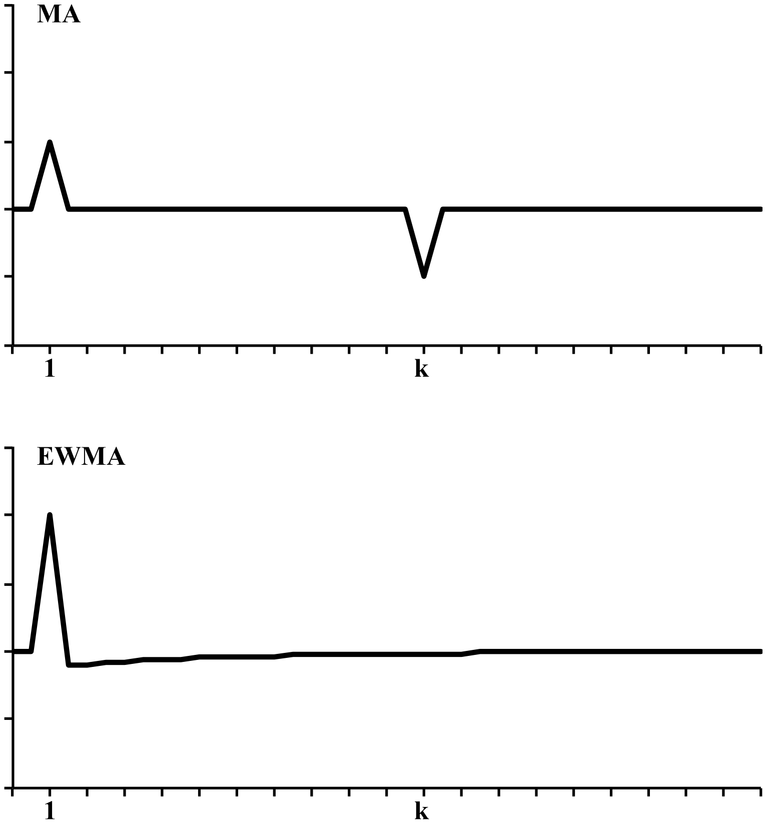

Figure 4: Signal “x” with a wavelet disturbance.

In both cases the disturbance appropriately appears in the smoothed trace at the time it occurs in the signal x.

With the MA an equal and opposite disturbance appears at a delay equal to the length of the Moving Average. This delayed rebound effect is spurious, since its occurrence is dependent solely on the length of the MA and has no relation to the wavelet disturbance.

The EWMA, by contrast, is well behaved, with a gradual return to normal after the disturbance.

3 MA and EWMA considered as low-pass filters

3.1 The Fourier transform

Fourier analysis provides a standard procedure for converting data from the time or distance domain to the equivalent frequency domain [5].

Consider a set of N data values, x_n, equally spaced in time or distance. Their Fourier transform generates N points in the frequency spectrum. These N points in the frequency spectrum carry exactly the same information as the N points in the time or distance domain. The lowest frequency has a period equal to the data length.

Fitting cosine and sine waves of this wavelength to the data generates the real and imaginary components of this fundamental frequency.

Further, fitting cosine and sine waves of each multiple of the fundamental frequency generates its real and imaginary components, up to the “Nyquist” frequency. The Nyquist frequency is N times the fundamental frequency and has a wavelength equal to twice the data interval. Any signal frequency higher than the Nyquist frequency cannot be detected, but will “fold back” to add to the amplitude of a corresponding lower frequency.

Each frequency value can be expressed either as real and imaginary components (the cosine and sine fits), or as an amplitude and phase

The Fourier transform converts the N values in the time (or distance) domain to the equivalent N values in the frequency domain.

Applying the Fourier transform in turn to the frequency domain data converts them back to the time (or distance) domain.

For real-world data, the time (or distance) values are strictly real, while the frequency values will have real (sine wave) and imaginary (cosine wave) components corresponding to their amplitude and phase.

If the data length N is a power of 2 (i.e. N = 2r, where r is an integer), the very efficient Fast Fourier transform algorithm can be used. Cooley and Tukey [6] first publicised this algorithm in 1965 (although it was discovered by Gauss in 1805).

Sequentially averaging a set of data is equivalent to applying a low-pass filter to the frequency data.

Applying averaging weights as in equations (2) or (3) to the time (or distance) data is a “convolution” operation. Multiplying the frequency spectrum of the filter weights by the frequency spectrum of the data set is exactly equivalent to convolving the time (or distance) domain data. The Fourier transform of the resulting product of the two frequency spectrums gives the same result as is obtained by convolving the corresponding MA or EWMA with the time (or distance) data.

MA and EWMA each act as low-pass filters, so it is instructive to compare the frequency spectrums.

3.2 Frequency spectrum for the moving average (MA)

The amplitude of the frequency spectrum for the Moving Average filter of Figure 1 is shown in Figure 5.

The amplitude is the square root of the summed squares of the cosine and sine Fourier components. (The phase would be the arctangent of the ratio of the sine and cosine Fourier components, but is not being considered here).

The amplitude spectrum of the MA filter is seen to have side lobes. Instead of the low-pass filter steadily reducing the amplitude of higher frequencies, it

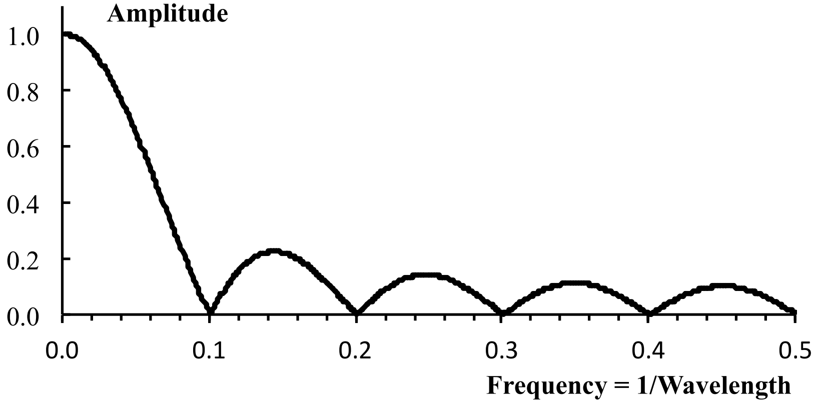

Figure 5: The amplitude spectrum for an MA filter of length = 10.

completely cuts out frequencies of 0.1, which corresponds to a wavelength of 10, the length of the Moving Average filter in Figure 1.

As we increase the frequency, the amplitude rises again, before again falling to zero at a frequency of 0.2 (5 units wavelength). This behaviour is repeated, allowing through ever diminishing side lobes, with complete cut-off at each harmonic of the filter length.

So, as we consider frequencies increasing from the fundamental lowest frequency, they will alternately be filtered out, allowed through, filtered out, and so on repeatedly, with the proportion of signal amplitude allowed through steadily diminishing for each side lobe.

The non-monotonic behaviour of the MA amplitude spectrum is a direct consequence of the MA filter's discontinuity in the time (or distance) domain that we saw in Figure 1.

The operational implication is that some high-frequency disturbances will pass through the filter, while lower-frequency disturbances will be completely blocked if they happen to be close to one of the harmonic frequencies.

For this reason, we must conclude that the Moving Average (MA) filter is unsatisfactory.

3.3 Exponential smoothing (EWMA)

The amplitude of the frequency spectrum for an Exponentially Smoothed filter of Figure 2 is shown in Figure 6.

The amplitude spectrum now has no side lobes, but declines steadily and exponentially. So the EWMA filter is much better behaved than the MA filter. The EWMA filter monotonically decreases the amplitude passed as the frequency increases.

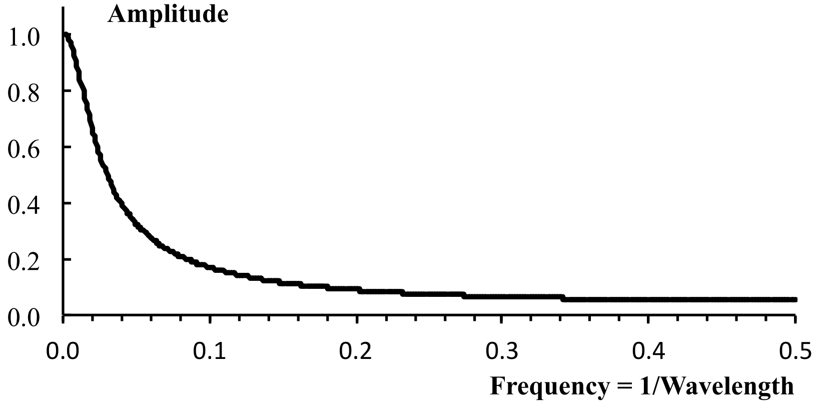

Figure 6: The amplitude spectrum for an EWMA filter with alpha = 0.1.

4 Adjustment for varying sample size

The treatment so far has assumed that the data are of equal importance. However, in many real situations, successive observations may need to exert varying influence. For example, if we are forecasting the grade of ore from previous shifts of varying tonnage, the higher tonnage shifts should have more influence than those of lower tonnage.

We will now consider such a situation of varying tonnage, so that observations x_n relate to tonnages w_n.

4.1 Moving average (MA)

If the MA is to be taken as the moving average over a total tonnage T, then equation (2) becomes:

$$MA_n = {}_{m=0}\Sigma^{k[n]} w_{n-m} x_{n-m}/T, \text{ where } {}_{m=0}\Sigma^{k[n]} w_{n-m} = T \qquad (4)$$

For a Moving Average, the length k[n] over which the average is taken will therefore have to be varied so that it encompasses the same tonnage (or as nearly as possible, the same tonnage).

4.2 Exponential smoothing (EWMA)

The treatment for exponentially smoothing over observations with varying tonnages is not so immediately obvious.

It is clear that the appropriate alpha value is a function of the tonnage: if the tonnage w increases we should use a larger $\alpha[w]$, so that a larger tonnage has more influence on the smoothed grade.

Consider two scenarios. Under the first scenario, two successive shifts have identical grade x and equal tonnage w.

Under the second scenario a single shift delivers ore of twice the tonnage, 2w but again with the same grade x.

If we start with a smoothed grade S_O, it is clear that under either scenario we should end up with the same grade, which we shall call S_F.

Under the first scenario, where each of the two shifts has grade x_n and tonnage w_n:

$$\begin{aligned} S_F &= (1-\alpha[w])((1-\alpha[w])S_O + \alpha[w]x) + \alpha[w]x \\ &= (1-\alpha[w])^2 S_O + \alpha[w](2-\alpha[w])x \end{aligned} \tag{5}$$

Under the second scenario, the single shift has grade x and tonnage 2w:

$$S_F = (1-\alpha[2w])S_O + \alpha[2w]x \tag{6}$$

Equating the coefficients of S_O and of x in equations (5) and (6) appears to give rise to two conditions that have to be satisfied.

For the coefficients of S_O in equations (5) and (6) to be the same:

$$(1-\alpha[2w]) = (1-\alpha[w])^2 \tag{7}$$

For the coefficients of x in equations (5) and (6) to be the same:

$$\alpha[2w] = \alpha[w](2-\alpha[w]) \tag{8}$$

We see that these two conditions are in fact identical, both being equivalent to:

$$\alpha[2w] = 1 - (1-\alpha[w])^2 \tag{9}$$

By induction, the condition can be extended to:

$$\alpha[nw] = 1 - (1-\alpha[w])^n \tag{10}$$

If w = 1, unit tonnage, then:

$$\alpha[W] = 1 - (1-\alpha[1])^W \tag{11}$$

Equation (11) has the satisfactory properties that $\alpha[0]$ is zero, and also that $\alpha[W]$ tends to 1 as W becomes very large.

5 How large should alpha be?

We have seen that alpha for an observation of tonnage W should be a monotonically increasing function of the tonnage W, and of $\alpha[1]$, the alpha for unit tonnage.

The question remains as to the appropriate choice for $\alpha[1]$. Clearly, this must depend upon the purpose for which the exponentially smoothed grade or other variable is being monitored.

In the control system discussed by Everett [4], ore was selected for each shift so that the expected grade of the selected ore, exponentially smoothed into the shift history, gave a grade on target. The ore was being blended onto stockpiles of 200 kilotonnes. So if a Moving Average (MA) were being used, it would be appropriate average over a tonnage T = 200 kt, as in equation (4), so the averaging weight applied to each kilotonne is 1/T.

For Exponential Smoothing, the choice of $\alpha[1]$ is not so clear cut. One criterion is to consider the average "age" of the sample. For a moving average, or for a completed stockpile of tonnage T, the average age is T/2. For an exponentially smoothed average to have the same average age of sample:

$$T/2 = {}_{m=0}\Sigma^{\text{Infinity}} m\alpha[1](1-\alpha[1])^m \tag{12}$$

$$= (1-\alpha[1])/\alpha[1] \tag{13}$$

$$\alpha[1] = 2/(2+T) \approx 2/T \tag{14}$$

So the starting weight for an EWMA should be about twice that of an equivalent MA, as shown in Figure 7:

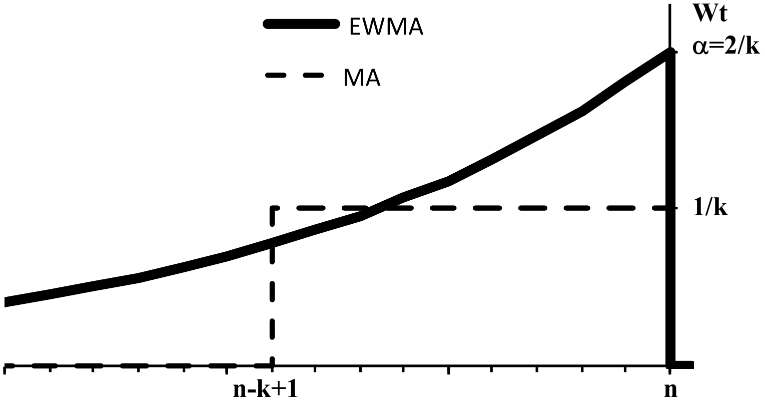

Figure 7: Equivalent EWMA and MA weights applied to recent data.

In a production process, such as a mining operation, T would be the subsequent blending tonnage, achieved either by blending directly onto stockpiles or inherent in the processing and transportation system.

6 Conclusions

By considering both the time (or distance) domain and the frequency domain, this paper has shown that Exponential Smoothing (EWMA) has considerable advantages over Moving Averages (MA).

The problem of varying sample sizes has been considered, and we have shown that the appropriate exponential smoothing factor for a sample of size w is given by equation [11], $\alpha[W] = 1 - (1-\alpha[1])^{W}$, where $\alpha[1]$ is the exponential smoothing factor to be applied to samples of unit weight.

We have further shown, in equation (14), that $\alpha[1]$ should be approximately 2/T, where T is the comparable MA tonnage, or the blending tonnage in a production process.

References

[1] Diebold, F.X. *Elements of Forecasting.* Fourth ed. Mason, OH: South-Western, 2008.

[2] Box, G. & Jenkins, G. *Times Series Analysis: Forecasting and Control.* San Francisco, CA: Holden-Day, 1970.

[3] Ramjee, R., Crato, N. & Ray, B.K. A note on moving average forecasts of long memory processes with an application to quality control. *International Journal of Forecasting,* **18,** pp. 291–297, 2002.

[4] Everett, J.E. Computer aids for production systems management in iron ore Mining. *International Journal of Production Economics,* **110/1**, pp. 213-223, 2007.

[5] Marks R.J. Handbook of Fourier Analysis and Its Applications. Oxford University Press, 2009.

[6] Cooley, J.W. & Tukey, J.W. An algorithm for the machine calculation of complex Fourier series, *Mathematics of Computation,* **19,** pp. 297–301, 1965.