Principled Derivation of Some Moving Average (Sliding Window) Statistics

William H. Press

February 26, 2019

Introduction

Moving (or sliding window) averages are widely used to estimate the present parameters of noisy nonstationary time series. This is a difficult undertaking almost by definition unless the nonstationarity is in some sense "slow", allowing the accumulation of a sufficient number of samples for a meaningful measurement. That difficulty leads us to here consider moving averages in general and to ask what are the underlying principles that should govern their construction. In particular, should we use an equal-weights simple moving average (SMA), or something more complicated like an exponentially weighted moving average (EWMA)? Or should we use something entirely different?

Sliding window averages are also used in a completely different application, that of data smoothing. For that case, there are many better techniques than those described here, for example Fourier filtering in the frequency domain, or Savitzky-Golay filtering. Here, the applications of interest are distinguished by our intense interest in the present moment, and our access to data only from the present and past, not the future. Financial time series furnish the most relevant examples.

1 Simple Moving Average

The simple moving average (SMA) estimates the expectation present value $\langle x \rangle$ as a uniformly weighted mean of the most recent N data points,

$$\widehat{x}_{\text{sma}} \equiv \sum_{i=0}^{N-1} w_i x_i, \quad w_i \equiv \frac{1}{N}, \quad \sum_{i=0}^{N-1} w_i = 1 \tag{1}$$

written in this way to call out its structure as a weighted mean with uniform weights. Here x_0 is the most recent data point, with $x_1, x_2, \ldots$ extending by uniform time steps to the past.

The estimator $\widehat{x}_{\text{sma}}$ is unbiased, as can be seen by

$$\langle \widehat{x}_{\text{sma}} \rangle = \left\langle \sum_{i=0}^{N-1} w_i x_i \right\rangle = \left(\sum_{i=0}^{N-1} w_i \right) \langle x_i \rangle = \langle x \rangle \tag{2}$$

Writing $\langle x_i \rangle = \langle x \rangle$ is the necessary approximation when we know nothing about the non-stationary nature of the series except that its changes are slow. The variance of $\widehat{x}$ is

$$\text{Var}(\widehat{x}_{\text{sma}}) = \sum_{i=0}^{N-1} w_i^2 \, \text{Var}(x_i) = \left(\sum_{i=0}^{N-1} w_i^2 \right) \text{Var}(x) = \frac{1}{N} \text{Var}(x) \tag{3}$$

In other words, the standard deviation of the estimator, $\sigma(\widehat{x}_{\text{sma}})$, decreases as the square root of the sample size, $N^{-1/2}$.

We might want to estimate not just the present value of a nonstationary series, but also the present variance of its values, $\text{Var}(x)$. A sample estimate is

$$\widehat{\text{Var}(x)} \equiv \left(\sum_{i=0}^{N-1} w_i x_i^2 \right) - \left(\sum_{i=0}^{N-1} w_i x_i \right)^2 \tag{4}$$

("the mean of the square minus the square of the mean"). As is well known, this estimator is only asymptotically unbiased, because one can calculate that, for the SMA,

$$\left\langle \widehat{\text{Var}(x)}_{\text{sma}} \right\rangle = \left(\frac{N-1}{N} \right) \text{Var}(x) \tag{5}$$

Below, we will give a derivation of equation (5) in a more general setting.

1.1 SMA is Not Ideal

There are two reasons to be dissatisfied with the simple moving average. First, since we are estimating the *present* value of the mean or variance, it seems perverse to count equally data as old as sample number $N-1$. Surely the more recent data should be considered as having greater relevance.

Second, if we are interested in *changes in* the the estimated present parameters—that is, in a time series of moving averages going forward—then the sharp distinction between sample $N-1$ (counted in the average) and sample N (not counted) is an undesirable feature.

To understand this latter point, imagine that one data point x_j is a large positive fluctuation. It first affects the SMA as an uptick when it is first observed (that is, $j=0$). Much later, when it goes from $j=N-1$ to $j=N$ it produces an equal downtick in the SMA. That downtick is essentially spurious: It conveys no new information about the *present* parameters; it is an entirely predictable consequence of old data falling off the end of the measurement window.

2 Weighted Averages in General

One can mitigate both of SMA's undesirable features by choosing not uniform weights w_i, but weights which taper (generally smoothly and monotonically) towards zero to the past, but that still sum to unity. So,

$$\widehat{x} \equiv \sum_{i=0}^{\infty} w_i x_i, \quad \sum_{i=0}^{\infty} w_i = 1, \quad w_0 > w_1 > w_2 > \cdots \tag{6}$$

If we are unhappy with sums that go to infinity, we can further impose a condition of *compact support* (finite sliding window width),

$$w_i = 0 \text{ for } i \geq M \tag{7}$$

for some value M.

For any choice of weights, equation (6) remains an unbiased estimator, because equation (2) (with now $N=\infty$) still goes through. Similarly, equation (3) for the variance of the estimator $\widehat{x}$ goes through (except for the last equality specializing to SMA). Since this variance, $\mathrm{Var}(\widehat{x})$, is a quantity to be made small, equation (3) shows that we should make $\sum_i w_i^2$ as small

as possible. In this connection it is useful to define a "variance reduction factor",

$$N_{\text{eff}} \equiv \left(\sum_i w_i^2 \right)^{-1}, \tag{8}$$

which is the factor by which the estimate $\text{Var}(\widehat{x})$ is smaller than the single point variance $\text{Var}(x)$. In the case of SMA, $N_{\text{eff}} = N$, as we have seen.

2.1 Calculation of the Bias of the Variance Estimator

Here is the promised derivation of equation (5) in a general setting: Define $\overline{x}$ and x' as the mean and fluctuating parts of x, so that

$$x_i \equiv \overline{x} + {x'}_i, \quad \langle {x'}_i \rangle = 0, \quad \langle {x'}_i {x'}_j \rangle = \delta_{ij} \text{Var}(x), \tag{9}$$

(the last equality making the assumption that the fluctuation values ${x'}_i$ are i.i.d.) Then,

$$\begin{aligned}
\left\langle \widehat{\text{Var}(x)} \right\rangle &= \left\langle \sum_{i=0}^{\infty} w_i x_i^2 - \left(\sum_{i=0}^{\infty} w_i x_i \right)^2 \right\rangle \\
&= \sum_{i=0}^{\infty} w_i \left\langle (\overline{x} + {x'}_i)^2 \right\rangle - \left\langle \left(\sum_{i=0}^{\infty} w_i (\overline{x} + {x'}_i) \right) \left(\sum_{j=0}^{\infty} w_j (\overline{x} + {x'}_j) \right) \right\rangle \\
&= \left(\overline{x}^2 + \sum_{i=0}^{\infty} w_i \left\langle {x'}_i^2 \right\rangle \right) - \left(\overline{x}^2 + \sum_{i,j=0}^{\infty} w_i w_j \left\langle {x'}_i {x'}_j \right\rangle \right) \\
&= \left(1 - \sum_{i=0}^{\infty} w_i^2 \right) \text{Var}(x) = \left(\frac{N_{\text{eff}} - 1}{N_{\text{eff}}} \right) \text{Var}(x)
\end{aligned} \tag{10}$$

We see that making N_{eff} large not only minimizes the variance of $\widehat{x}$ but also minimizes the bias of its estimator $\widehat{\text{Var}(x)}$.

2.2 A List of Desiderable Properties

Thus far, we know that we must have $\sum_i w_i = 1$, and we have seen that maximizing N_{eff} minimizes the variance and bias of the estimator $\widehat{x}$. For practical reasons, we may want to specify or the window size M in equation

(7). Monotonicity is good thing, because it supports the principle that older data is less relevant. What else should be on our list of desirable properties?

We can formalize the idea that data more recent on average is more relevant than data older on average, and effect this by seeking to minimize a moment of the weights, for example,

$$L_n \equiv \langle i^n \rangle = \sum_i i^n w_i \tag{11}$$

We will refer to $L_1 = \langle i \rangle$ as the *estimator lag* of the moving average. The estimator lag of the SMA is easily calculated to be $(M-1)/2$, that is, the halfway point (or center of mass) in its interval of compact support.

We can also formalize the previous discussion of the time-smoothness of $\widehat{x}$ as old data moves through the sliding window (and possibly drops discontinuously off the end). Write as $\widehat{x}_t$ the value of the moving average at timestep t, that is,

$$\widehat{x}_t = \sum_{i=0}^{\infty} x_{t-i} w_i, \tag{12}$$

Then, for the difference of two consecutive moving averages, we have

$$\Delta\widehat{x}_t \equiv \widehat{x}_t - \widehat{x}_{t-1} = x_t w_0 + \sum_{i=1}^{\infty} x_{t-i}(w_i - w_{i-1}) \tag{13}$$

Now taking the variance of equation (13),

$$\mathrm{Var}(\Delta\widehat{x}) = \left[w_0^2 + \sum_{i=0}^{\infty} (w_i - w_{i+1})^2 \right] \mathrm{Var}(x) \equiv \left[w_0^2 + S^2 \right] \mathrm{Var}(x) \tag{14}$$

The first term in square brackets, w_0^2, is the necessary result of adding new present information to the estimate. The summation term, which we denote S^2 and refer to as the "timestep variance", represents the undesirable time fluctuation in the estimate $\widehat{x}_t$ that does not reflect any new information; we may wish to minimize it.

Typically, both w_0^2 and S^2 turn out to be much smaller than $1/N_{\mathrm{eff}}$, so, in magnitude, they contribute only negligibly to $\mathrm{Var}(\widehat{x})$. Nevertheless, they become important when we are interested in $\widehat{x}_t$ (equation (12)) as a time series, and in particular interested in its point-to-point change. Equation

(14) shows that these "upticks" or "downticks" will depend exclusively on new information only if $S^2 \ll w_0^2$.

To summarize, here is our list of desirable attributes for moving averages. We will see in subsequent sections that it is not possible to have all of these properties at the same time—we will have to make tradeoffs.

1. Normalization: $\sum_i w_i = 1$ must always be true.
2. Monotonicity: require $w_i \geq w_{i+1}$
3. Compact support: minimize M such that $w_i = 0,\ i \geq M$
4. Variance reduction factor: maximize $N_{\text{eff}} = 1/\sum_i w_i^2$
5. Timestep variance: minimize $S^2 = \sum_i (w_{i+1} - w_i)^2$; or at least require $S^2 \ll w_0^2$
6. Estimator lag: minimize $L_1 \equiv \langle i \rangle = \sum_i i w_i$

3 SMA and EWMA Derived from Principles

3.1 Simple Moving Average (SMA)

We here show that the simple moving average is the weighted average that maximizes N_{eff}, the variance reduction factor, for a window limited to M samples (i.e., for fixed compact support); and that it achieves $N_{\text{eff}} = M$. The proof is via a Lagrange multiplier to impose the normalization contraint:

$$\begin{aligned} \mathcal{L} &= \sum_{i=0}^{M-1} w_i^2 - 2\lambda \left(\sum_{i=0}^{M-1} w_i - 1 \right) \\ \frac{\partial \mathcal{L}}{\partial w_i} = 0 &\implies w_i = \lambda = \frac{1}{M} \\ &\implies N_{\text{eff}} = M \end{aligned} \tag{15}$$

The estimator lag of the SMA is readily calculated to be $L_1 = \langle i \rangle = \frac{1}{2}(N_{\text{eff}} - 1)$. To see that the SMA has a poor (i.e., large) timestep variance, note that $S^2 = (w_M - w_{M-1})^2 = w_{M-1}^2 = w_0^2$, which is not $\ll w_0^2$.

3.2 Exponentially Weighted Moving Average (EWMA)

We here show that the exponentially weighted moving average is the weighted average that minimizes the timestep variance S^2 for a specified variance reduction factor N_{eff}. Proof:

$$\begin{aligned} \mathcal{L} &= \sum_{i=0}^{\infty}(w_{i+1}-w_i)^2 - 2\lambda_1\left(\sum_{i=0}^{\infty} w_i - 1\right) + \lambda_2\left(\sum_{i=0}^{\infty} w_i^2 - \frac{1}{N_{\text{eff}}}\right) \\ \frac{\partial \mathcal{L}}{\partial w_i} &= 0 \implies w_{i+1} = (2+\lambda_2)w_i - w_{i-1} - \lambda_1, \quad i = 1, \ldots, \infty \end{aligned} \tag{16}$$

This is a second order inhomogeneous linear recurrence. Its homogeneous solutions are

$$w_i \propto r_\pm^i, \text{ where } r_\pm = \frac{1}{2}\left[(2+\lambda_2) \pm \sqrt{\lambda_2(\lambda_2+4)}\right] \tag{17}$$

We note that $r_+ r_- = 1$, so only r_- takes on values in the range $(0, 1)$ that give normalizable solutions. A particular solution (to be added to the homogeneous solution) is $w_i = \text{constant} = \lambda_1/\lambda_2$, so normalizability implies $\lambda_1 = 0$. With some algebra, we can eliminate λ_2 in favor of N_{eff} giving the solution,

$$w_i = \left(\frac{2}{1+N_{\text{eff}}}\right)\left(\frac{N_{\text{eff}}-1}{N_{\text{eff}}+1}\right)^i, \quad i \geq 0 \tag{18}$$

More algebra gives the achieved minimum value $S^2 = w_0^2/N_{\text{eff}}$. As desired, we have $S^2 \ll w_0^2$ when $N_{\text{eff}} \gg 1$. In that same limit equation (18) can be written

$$w_i \approx \left(\frac{2}{1+N_{\text{eff}}}\right)\exp\left[-\frac{i}{(N_{\text{eff}}/2)}\right] \tag{19}$$

The estimator lag of the EWMA, equation (18), is $L_1 = \langle i \rangle = \frac{1}{2}(N_{\text{eff}} - 1)$, which is the same as the SMA with the same N_{eff}.

4 Better Moving Averages

4.1 Linearly Weighted Moving Average (LWMA)

What weighted average minimizes the estimator lag L_1 for fixed variance reduction factor N_{eff}? We have:

$$\mathcal{L} = \sum_{i=0}^{\infty} i w_i - \lambda_1 \left(\sum_{i=0}^{\infty} w_i - 1 \right) + \tfrac{1}{2}\lambda_2 \left(\sum_{i=0}^{\infty} w_i^2 - \frac{1}{N_{\text{eff}}} \right) \tag{20}$$

which implies for all $i \geq 0$,

$$w_i = \frac{\lambda_1 - i}{\lambda_2} \tag{21}$$

Since $w_0 > 0$, w_i must become negative for some positive i. However, since the equations for the w_i's are independent of one another, we can simply allow negative w_i's to "pin" with the constraint $w_i = 0$ for $i \geq M$, for some M. A normalized solution is then

$$w_i = \frac{2}{M(M+1)}(M - i), \quad 0 \leq i \leq M, \tag{22}$$

a linear ramp from $2/(M+1)$ to zero. This linearly weighted moving average or LWMA (as it is known to stock traders, also called just WMA) has the variance reduction factor

$$N_{\text{eff}} = \frac{3M(M+1)}{2(2M+1)} \;\Leftrightarrow\; M = \frac{1}{6}\left[(4N_{\text{eff}} - 3) + \sqrt{16{N_{\text{eff}}}^2 + 9}\right] \tag{23}$$

Thus we can specify N_{eff} and choose M as the next larger integer to equation (23). Or, since we are getting compact support automatically, we can instead choose a window size M and achieve (for large M), $N_{\text{eff}} \approx \frac{3}{4}M$. Note that, for $M \gg 1$, the LWMA has a standard deviation $\sigma(\widehat{x}) = \sqrt{\text{Var}(x)/N_{\text{eff}}}$ that is only about 15% larger than the simple moving average with the same M.

The estimator lag of equation (22) is

$$L_1 = \sum_{i=0}^{M} i w_i = \frac{M-1}{3}, \tag{24}$$

somewhat better (i.e., more weighted to the present) than simple moving average's $(M-1)/2$.

The timestep variance of equation (22) is

$$S^2 = \sum_{i=0}^{M-1} (w_i - w_{i+1})^2 = \frac{4}{M(M+1)^2} = \frac{w_0^2}{M} \ll w_0^2 \quad \text{for} \quad M \gg 1 \tag{25}$$

This satisfies the desired criterion that the timestep variance be negligible, even if it is not strictly minimized. Overall the LWMA seems like an excellent general-purpose choice, superior in most ways to both the SMA (e.g., smaller estimator lag and timestep variance) and the EWMA (e.g., compact support).

4.2 Lowered Exponential Moving Average (LEMA)

The EWMA has the disadvantage of formally requiring an infinite number of data points to the past. In reality, one has only finite data. A common practice is to use the first M (say) EWMA weights and renormalize,

$$\widehat{x} = \sum_{i=0}^{M-1} w_i x_i \Bigg/ \sum_{i=0}^{M-1} w_i \tag{26}$$

where the weights are given by equation (18). This practice does not have much to recommend it. In particular, it negates the smoothness of the EWMA and adds an unnecessary amount w_{M-1}^2 to the timestep variance S^2.

A better idea, we now show, is to obtain compact support by using "lowered exponential" weights,

$$w_i \propto (r^i - r^M) = \frac{1-r}{1-(M+1)r^M + Mr^{M+1}}(r^i - r^M), \quad 0 \le i \le M \tag{27}$$

Here the last equality introduces the normalization that makes the weights sum to 1, while r in the range $(0, 1)$ is the parameter analogous to the ratio of successive terms in the EWMA. We can refer to equation (27) as a lowered exponential moving average (LEMA).

Exact formulas for the properties of LEMA are complicated, but they simplify in two relevant limits: If r is sufficiently less than one so as to make r^M neglible, then weights will have decayed effectively to zero before the window size M is reached. In that case,

$$w_i = (1-r)r^i, \quad 0 \le i \le M \tag{28}$$

which is exactly equation (18) with

$$N_{\text{eff}} = \frac{1+r}{1-r} \;\Leftrightarrow\; r = \frac{N_{\text{eff}}-1}{N_{\text{eff}}+1} \tag{29}$$

So, in this limit, LEMA is equivalent to EWMA, with properties in terms of N_{eff} previously given.

The other limit is when r is sufficiently close to 1 so that $1-1/M < r < 1$. In this case, the weights have decayed by less than a single e-fold in M steps, and the limiting formula for the weights as $r \to 1$ is exactly equation (22), the linearly weighted moving average, LWMA, described (and praised) above.

When neither limit holds, the LEMA interpolates naturally between the two. Its variance reduction factor is

$$N_{\text{eff}} = \frac{1+r}{1-r}\,\frac{[1-r^M(1+M-Mr)]^2}{1-2(1+r)r^M+[1+2r+M(1-r^2)]r^{2M}} \tag{30}$$

in which the limit equation (29) is evident. The lag is given in general by

$$L_1 = \frac{2r-M(M+1)r^M+2(M^2-1)r^{M+1}-M(M-1)r^{M+2}}{2(1-r)[1-(M+1)r^M+Mr^{M+1}]} \tag{31}$$

A possible application of equations (30) and (31) is, for fixed M, to adjust r until a desired small value for the estimator lag L_1 is obtained (smaller than the maximum value given by equation (24)), and then check to see of an acceptably large value of N_{eff} is obtained (in turn bounded by its maximum, equation (23)). L_1 and N_{eff} will always be of the same order.

The timestep variance is given in general by

$$S^2 = \frac{w_0^2}{N_{\text{eff}}}\,\frac{\left(1+r^M\right)\left(1-(M+1)r^M+Mr^{M+1}\right)^2}{\left(1-r^M\right)\left[1-2r^M-2r^{M+1}+(M+1)r^{2M}+2r^{2M+1}-Mr^{2M+2}\right]} \tag{32}$$

For any $M \geq 2$ and any $0 < r < 1$, the last factor in equation (32) lies between 0.73 and 1. Thus, when $N_{\text{eff}} \gg 1$, we automatically get $S^2 \ll w_0^2$, as desired.

5 Recommendations

Other than tradition, it is hard to think of a good justification for using either the simple moving average (SMA) or the exponentially weighted moving

average (EWMA) instead of one of the two alternatives that we have given above, the linearly weighted moving average (LWMA) and lowered exponential moving average (LEMA).

The only disadvantage of our recommended linearly weighted moving average (LWMA) is that, for fixed window size M, it has a standard deviation $\sigma(\widehat{r})$ about 15% larger than the SMA with the same M. Its advantages are its smaller estimator lag (it is more current) and, crucially, a small timestep variance, so that artifacts due to old data moving through the sliding window are insignificant (in contrast to the SMA).

The lowered exponential moving average (LEMA) naturally generalizes EWMA to a finite window size M. It has an adjustable estimator lag L_1. When L_1 is set significantly smaller than M, LEMA reduces to EWMA. In the other limit of maximizing the variance reduction factor N_{eff}, it reduces to LWMA. Between these limits, it is the natural interpolation.