Modeling Time Series Using Exponentially Weighted Moving Averages

Georg Martin Reinhart*

April 21, 2022

Abstract

This paper presents an algorithm for fitting curves to time series (TS) using their iterated exponentially weighted moving averages (EWMA). The method is well suited for online learning since EWMAs are efficiently updated in real-time. The fits represent the most recent history of a TS more accurately than the more distant past which is often the desirable way to avoid overfitting in applications. If the TS is in the form of a generalized Brownian motion process, optimal choices of the parameters of the fits are presented, maximizing certain performance metrics. In particular, the question of how to choose best time horizons for EWMAs is solved. Practical applications are given in the fields of data reduction, noise filtering and modeling financial data. Software, written in Python, accompanies this paper and can be found at www... On a fundamental level, we develop a complete theory on the expansion of a TS in terms of its iterated EWMAs and give estimates of the expected square error.

***K*eywords** time series, EWMA, curve fitting, machine learning, Brownian motion, data reduction, technical analysis, financial data

1 Introduction

Moving averages have been used extensively in technical analysis of financial markets with the purpose of finding actionable trading strategies (for an introduction see [Murphy (1999)]). One popular trading signal occurs when the 50-day moving average of a market crosses above the 200-day average (Golden Cross) or below (Death Cross). If one believes that past price patterns have predictive power for the future, the question becomes how much information do these moving averages contain about the price action of an asset and whether the values of 50 and 200 are optimal in some (mathematical) sense. Other popular market indicators that make use of moving averages include the Moving Average Convergence Divergence indicator (MACD) and the Relative Strength Index (RSI).

These moving averages are simple moving averages, i.e., rolling windows of arithmetic averages. Another type of moving average is the exponentially weighted moving average (EWMA) that gives the most recent value of a process the most weight and discounts previous values by a factor (precise definitions will be given in the next section). For a mathematical treatment of moving averages it is simpler, or even necessary, to work with EWMAs. The importance of EWMAs extends further

*The author wishes to thank David Madigan for his time and valuable suggestions

with many applications such as smoothing operators on TS or filters and also in statistical process control (e.g., see [Qiu (2014)]).

Some statistical techniques use higher order EWMAs, also called iterated EWMAs, i.e., moving averages of moving averages. The TRIX indicator in technical analysis of market data uses such iterated EWMAs. In [Cacorogna et al. (2001)] and [Müller and Zumbach (2001)] a toolbox of operators on iterated EWMAs is developed giving a rigorous mathematical treatment. One method presented is the delta-operator, giving an estimate between the current value of the TS and the value p time units in the past. Let $e^{(1)}(h)$ be the EWMA with time horizon h of the TS, $e^{(2)}(h)$ the EWMA of $e^{(1)}(h)$, etc. then, using four iterated EWMAs, the delta-operator (equation (3.61) in [Cacorogna et al. (2001)]) is given by

$$\Delta(p) = 1.22208 \left(e^{(1)}(0.3719p) + e^{(2)}(0.3719p) - 2e^{(4)}(0.2418p) \right). \tag{1}$$

The authors of the book give a heuristic argument why this operator gives reasonable estimates. The question of whether this formula is optimal or what error can be expected, however, is not addressed.

1.1 Methods

This paper applies linear regression using the current values of the TS and its iterated EWMAs as predictors of the value of the TS p steps in the past. Minimizing the square error gives optimal time horizons of the EWMAs and the corresponding optimal regression coefficients. If applied to a range, rather than a single value of p, the resulting formulae can be used to fit optimal curves to TS. The fitted curves can be considered as a model of the TS that represents the essential features of the structure of the TS, giving more emphasis to the recent past. The error estimates give a framework to analyze how much information is lost in replacing the TS by its EWMA approximations.

We are interested in deterministic formulae in case the TS is some form of Brownian motion (precise definitions will be given later). The resulting equations can be directly used to model TS rather than performing individual linear regressions to find numerical answers. On a theoretical level, the fitted curves can be considered as *expanding a TS in terms of its iterated EWMAs* in a remotely similar way an analytic function can be expanded in its Taylor series. However, convergence appears to be rather slow which could limit potential applications.

Other applications where EWMAs proved useful were in the field of high frequency data streams. Using EWMAs condenses such a stream into a fixed number of computed values. In [Chen et al. (2001)] and [Chen et al. (2000)], EWMAs were used to efficiently summarize databases of customer transactions in the wireless telecommunications industry. This data reduction technique enabled analysts to profile millions of customers' calling patterns in real time. It is also possible to track virtually all financial transactions worldwide in near-real time using these techniques. While modern computers can handle much larger volumes of data, data reduction techniques still play a major role in data mining. Capturing the essence of a pattern in a TS via data reduction techniques avoids overfitting of statistical models.

The outline of the paper is as follows:

- Section 2 gives the basic definitions of TS and their iterated EWMAs.
- Section 3 develops the algorithm on how to regress iterated EWMAs onto their underlying TS and gives the computational background.

- Section 4 deals with the expected square error.
- Section 5 on optimal time horizons handles the question of how to choose best parameters for the EWMAs, minimizing the square of a point estimate or minimizing the L^1 or L^2 norms over a range. We further conjecture that by adding more iterated EWMAs the expected error converges to zero.
- Section 6 illustrates practical applications, including simulations and their visualizations. Applications considered: curve fitting, forecasting, financial data, data reduction, noise filtering, statistical process control and feature engineering.
- Section 7 compares the method of using iterated EWMAs to other data reduction techniques, such as using EWMAs with different time horizons (rather then iterated EWMAs) or simply storing periodic historical values of the TS. We show that using iterated EWMAs is always superior to using two separate EWMAs with different time horizons.
- Section 8 on further research restates the approximation problem in terms of linear algebra. Solving the regression task in this setting would encompass a much larger class of moving averages.

2 Notation

Definition 2.1 (Random Walk)**.** *Let $\{\xi_i\}_{i\geq 1}$ be a sequence of independent and identically distributed random variables in $\mathbb{R}$ with finite mean μ and variance σ^2. A random walk started at s_0 is the discrete stochastic process given by $s_n = s_0 + \sum_{i=1}^{n} \xi_i$. The ξ_i are called the increments or steps of the RW.*

WLOG, $s_0 = 0$ throughout. μ is also called the *drift* of the RW. If the ξ_i take values in $\{-1, +1\}$, each with probability $\frac{1}{2}$, the RW is called a *simple random walk*.

Remark 2.2. *Expectation and Variance of s_n:*

(a) $E(s_n) = n\mu$,

(b) $\sigma_n^2 = \mathrm{Var}(s_n) = n\sigma^2$.

Definition 2.3 (EWMA)**.**

(a) (Fixed) Let $\{s_n\}$ be a RW. The fixed EWMA of $\{s_n\}$ with weight w, $0 \leq w \leq 1$, is the stochastic process $\{e_n\}$ defined recursively by

$$e_0 = s_0 \ (= 0),$$

$$e_n = we_{n-1} + (1-w)s_n, \ n \geq 1.$$

(b) (Time dependent) Let $\{s_n\}$ be a RW, interpreted as a TS with time increments Δt_n. The EWMA with time horizon h is defined as above replacing w with $w_n = \exp(-\frac{\Delta t_n}{h})$.

(c) (Continuous processes) Let $\{s_t\}$, $t \geq 0$ be a continuous stochastic process. Let $s(\nu) = s_\nu$. The EWMA $\{e_t\}$ of $\{s_t\}$ with time horizon h is defined by

$$e_t = \int_0^t \frac{1}{h} \exp(-\frac{t-\nu}{h}) s(\nu) d\nu.$$

(d) (Higher order EWMAs) Recursively, define the i'th iterated EWMA (fixed or time dependent) by $\{e_n^{(1)}\} = \{e_n\}$ *and* $\{e_n^{(i)}\}$, $i \geq 2$ *to be the EWMA of* $\{e_n^{(i-1)}\}$.

If Δt_n is constant in item (2.), the TS is called *homogeneous* and w is fixed.

This paper is concerned with Brownian Motion type processes, i.e., what happens asymptotically to a RW as $n \to \infty$. By the Central Limit Theorem it is well understood that the standardized process

$$\frac{s_n - n\mu}{\sqrt{n}} \tag{2}$$

has an asymptotically normal distribution with mean zero and variance σ^2. In applications, however, it is necessary to approximate such continuous time series with discrete random walks. For example, stock market prices can be considered as having a continuous underlying process while prices are sampled at fixed (e.g., end of day) or random (e.g., each individual trade) times.

The simplest way to approximate standard Brownian Motion, the ξ_i of Definition 2.1 can be chosen from a simple RW or from a standard normal distribution. To move between continuous and discrete TS, assume that $\bar{n}$ corresponds to one time unit (where $\bar{n}$ is usually large). All quantities measured in time will be indicated by a hat, e.g., if $\bar{n} = 10000$, then $\Delta t = \frac{1}{\bar{n}} = 0.0001$ and the time horizon of an EWMA $h = 1500$ will be given by $\hat{h} = 0.15$ in time units. By (2) ($\mu = 0$), we divide the corresponding s_n by $\sqrt{n}$ to move from simulation to the approximation of the Brownian Motion. E.g., if $s_{20000} = 150$ then $\hat{s}(2) = 150/\sqrt{10000} = 1.5$.

While iterated fixed EWMAs are defined and usually computed recursively, they can also be computed via the increments ξ_i directly. Here we give an explicit formula for this computation. See [Brown et al. (1961)].

Proposition 2.4. *The k'th fixed EWMA of a RW or TS* $\{s_n\}$ *is given by*

$$e_n^{(k)} = (1-w)^k \sum_{i=0}^{n-1} \binom{k-1+i}{k-1} w^i s_{n-i}.$$

Corollary 2.5. *Differences of fixed EWMAs are given by*

$$e_n^{(k)} - e_n^{(k+1)} = (1-w)^k \sum_{i=0}^{n-1} \binom{k+i}{k} w^{i+1} \xi_{n-i}, \text{ for } k \geq 0.$$

(The equation is also true for $k = 0$*, i.e., for* $s_n - e_n^{(1)}$*.)*

3 Regression Coefficients

3.1 Basic Form of the Regression Coefficients

Let t be the current time of a TS and $t_p < t$. Our goal is to find the best possible estimate of the value $\hat{s}_{t_p}$ of TS, knowing only the current value $\hat{s}_t$ and the first m iterated EWMAs at time t.. Best possible is in the sense of a regression minimizing a given loss function. As outlined in section 2, to obtain results for the continuous case we derive results for RWs and then take appropriate limits. Throughout this section, let s_n be the current value of the approximating RW and s_{n-p} the value in the past we want to estimate. To make the proofs and notation simpler,

we will regress $Y = s_n - s_{n-p}$ on $X = (s_n - e_n^{(1)}, e_n^{(1)} - e_n^{(2)}, \ldots, e_n^{(m-1)} - e_n^{(m)})$. We will use $\beta = \mathrm{Cov}(Y,X)\mathrm{Var}^{-1}(X)$ as the method for computing the regression coefficients. Let $\mathrm{Cov}(Y,X)$ be the row-vector of covariances of Y and X and $\mathrm{Var}(X)$ the variance-covariance matrix of X. The case $\mu = 0$ will be addressed first, i.e., processes with no drift. The case of TS with drift will be discussed in section 3.3.

We assume $\mu = 0$ throughout this section; m is the number of iterated EWMAs.

Lemma 3.1. *Let $\{s_n\}$ be a random walk with no drift ($\mu = 0$), X, Y, m, n and p ($p < n$) as above, w the weight of the EWMAs, and let $\mathrm{Cov}(Y,X) = (c_0, \ldots, c_{m-1})$. Then*

$$c_k = \sigma^2(1-w)^k \sum_{i=0}^{p-1} \binom{k+i}{k} w^{i+1}.$$

Proof. The proof is given in the case where the steps of the RW are drawn from a discrete distribution. By Corollary 2.5, $E(e_n^{(k)} - e_n^{(k+1)}) = 0$ and therefore

$$\begin{aligned}
&\mathrm{Cov}(s_n - s_{n-p}, e_n^{(k)} - e_n^{(k+1)}) \\
&= \sum_{\xi_1,\ldots,\xi_n} \left[p(\xi_1,\ldots,\xi_n) \left(\sum_{i=0}^{p-1} \xi_{n-i} \right) \left((1-w)^k \sum_{i=0}^{n-1} \binom{k+i}{k} w^{i+1} \xi_{n-i} \right) \right] \\
&= (1-w)^k \sum_{\xi_1,\ldots,\xi_n} \left[p(\xi_1,\ldots,\xi_n)(w\xi_n^2 + w\xi_{n-1}\xi_n + (k+1)w^2\xi_{n-1}\xi_n + (k+1)w^2\xi_{n-1}^2 + \cdots) \right],
\end{aligned} \tag{3}$$

where the outer sum is over all paths $\xi_i, \ldots, \xi_n$ with probability $p(\xi_1, \ldots, \xi_n)$. By the independence of steps note that if $i \neq j$

$$\sum_{\xi_1,\ldots,\xi_n} p(\xi_1,\ldots,\xi_n)\xi_i\xi_j = \sum_{\xi_i,\xi_j} p(\xi_i,\xi_j)\xi_i\xi_j = \sum_{\xi_i,\xi_j} p(\xi_i)p(\xi_j)\xi_i\xi_j$$

$$= \sum_{\xi_i} p(\xi_i)\xi_i \sum_{\xi_j} p(\xi_j)\xi_j = \mu^2 = 0.$$

Furthermore, we have

$$\sum_{\xi_1,\ldots,\xi_n} p(\xi_1,\ldots,\xi_n)\xi_i^2 = \sum_{\xi_i} p(\xi_i)\xi_i^2 = \sigma^2.$$

Hence all mixed terms in ξ_i in (3) sum to zero and considering the square terms, the expression (3) simplifies to

$$\sigma^2(1-w)^k \sum_{i=0}^{p-1} \binom{k+i}{k} w^{i+1}.$$

□

The covariance does not depend on n, i.e., it is independent of how long the process has evolved in time (as long as $n > p$). Note that $w = \exp(-1/h)$ depends on h.

Next, we shift our focus from fixed to time dependent EWMAs. As outlined in section 2, we will interpret a RW as a (homogeneous) sampling of a continuous TS. We want to estimate the change in value $s_n - s_{n-p}$ of the RW, given iterated EWMAs at step n with time horizon h. We will see

that the regression estimate, taking limits as $n \to \infty$, is a function of the ratio of p and h, denoted by $r = \frac{p}{h}$. Thus estimating a past value of a TS will depend on how far back we want to estimate relative to h, independent on the time unit one step of the RW represents. As $\Delta t \to 0$, $n \to \infty$, and p and h will approach ∞ at the same rate with fixed ratio r.

Lemma 3.2. *Using the notation of Lemma 3.1, let h be the time horizon of a RW, $w = \exp(-1/h)$, r a constant, and $p = rh$. Then for $0 \le k \le m-1$*

$$c_k \sim h\sigma^2 \left(1 - e^{-r} \sum_{\nu=0}^{k} \frac{r^\nu}{\nu!}\right) \quad \text{as } h \to \infty.^1$$

Proof. By Lemma 3.1 and Proposition A.2 in the Appendix

$$\begin{aligned} c_k &= \sigma^2(1-w)^k \sum_{i=0}^{p-1} \binom{k+i}{k} w^{i+1} \\ &= \frac{\sigma^2}{1-w} \left(w + \frac{w^{p+1}}{k!} \sum_{i=0}^{k} (-1)^{i+1} \binom{k}{i} \left[\prod_{j=0, j\neq i}^{k} (p+j) \right] w^i \right). \end{aligned} \tag{4}$$

We will now examine the sum in the expression above (call it A). Expand

$$w^i = e^{\frac{-i}{h}} = \sum_{\nu=0}^{k} (-1)^\nu \frac{i^\nu}{\nu! h^\nu} + O(\frac{1}{h^{k+1}}). \tag{5}$$

A, as a polynomial in p, is of degree at most k. Since $p = rh$ it follows that

$$\sum_{i=0}^{k} (-1)^{i+1} \binom{k}{i} \left[\prod_{j=0, j\neq i}^{k} (p+j) \right] O(\frac{1}{h^{k+1}}) = o(h) \text{ as } h \to \infty.$$

By Proposition A.4 in the Appendix

$$\frac{1}{k!} \sum_{i=0}^{k} (-1)^{i+1} \binom{k}{i} \left[\prod_{j=0, j\neq i}^{k} (p+j) \right] \frac{i^\nu}{\nu! h^\nu} = \frac{-(-p)^\nu}{\nu! h^\nu}.$$

Plugging (5) into (4) and using the last two observations yields:

$$\begin{aligned} c_k &= \frac{\sigma^2}{1-w} \left(w + w^{p+1} \sum_{\nu=0}^{k} -\frac{(-1)^\nu (-p)^\nu}{\nu! h^\nu} + o(h) \right) \\ &= \frac{\sigma^2}{1 - e^{\frac{-1}{h}}} \left(e^{\frac{-1}{h}} + e^{-r-\frac{1}{h}} \sum_{\nu=0}^{k} \frac{-(-1)^\nu (-rh)^\nu}{\nu! h^\nu} + o(h) \right) \\ &= \sigma^2 h (1 + o(h)) \left(1 - e^{-r} \sum_{\nu=0}^{k} \frac{r^\nu}{\nu!} + o(h) \right). \end{aligned}$$

□

[1] "e^*" denotes both exp and EWMAs depending on the context, the iteration index of an EWMA is always given in parentheses in the exponent

We now turn our attention to $\mathrm{Var}(X)$, where X is the random vector of differences of EWMAs, i.e., $x_k = e_n^{(k)} - e_n^{(k+1)},\ k = 0, \ldots, m-1$.

Lemma 3.3. *Let $c_{kl}(n,h)$ be the entry in row k and column l, $0 \le k, l \le m-1$, of $\mathrm{Var}(X)$. Then $c_{kl}(h) = \lim_{n\to\infty} c_{kl}(n,h)$ exists $\forall h > 0$ and*

$$c_{kl}(h) \sim h\sigma^2 \binom{k+l}{k} \frac{1}{2^{k+l+1}} \text{ as } h \to \infty.$$

Proof. By Corollary 2.5

$$\begin{aligned} c_{kl}(n,h) &= \mathrm{Cov}(e_n^{(k)} - e_n^{(k+1)}, e_n^{(l)} - e_n^{(l+1)}) \\ &= \sum_{\xi_1,\ldots,\xi_n} \left[p(\xi_1,\ldots,\xi_n)(1-w)^{k+l} \left(\sum_{i=0}^{n-1} \binom{k+i}{k} w^{i+1} \xi_{n-i} \sum_{j=0}^{n-1} \binom{l+j}{l} w^{j+1} \xi_{n-j} \right) \right], \end{aligned}$$

where the outer sum is over all paths $\xi_1, \ldots, \xi_n$. As in the proof of Lemma 3.1, the mixed terms in ξ_i and ξ_j sum to zero and the square terms sum to σ^2. Hence,

$$c_{kl}(n,h) = \sigma^2 (1-w)^{k+l} \sum_{i=0}^{n-1} \binom{k+i}{k} \binom{l+i}{l} w^{2i+2}.$$

By Proposition A.5 in the Appendix

$$c_{kl}(n,h) = \sigma^2 \frac{(1-w)^{k+l} w^2}{(1-w^2)^{k+l+1}} \left[\sum_{i=0}^{k} \binom{k}{i} \binom{l}{i} w^{2i} + P_{kl}(w,n) w^{2n} \right],$$

where $P_{kl}(w,n)w^{2n}$ is a polynomial in w with coefficients depending on n and each term having degree at least $2n$. Since $w = \exp(-1/h)$ we have $lim_{n\to\infty} P_{kl}(n,w)w^{2n} = 0$. Therefore

$$\begin{aligned} c_{kl}(h) &= \sigma^2 \frac{(1-w)^{k+l} w^2}{(1-w)^{k+l+1}(1+w)^{k+l+1}} \sum_{i=0}^{k} \binom{k}{i} \binom{l}{i} w^{2i} \qquad (6) \\ &= \sigma^2 \frac{(1+o(h))^2}{(1-w)(2+o(h))^{k+l+1}} \sum_{i=0}^{k} \binom{k}{i} \binom{l}{i} (1+o(h))^{2i} \\ &= \sigma^2 \left(\frac{h}{2^{k+l+1}} + o(h) \right) \binom{k+l}{k} + o(h). \end{aligned}$$

The last simplification follows since $1/(1-w) = 1/(1-e^{-1/h}) \sim h$ and by Proposition A.1. □

Whereas $\mathrm{Cov}(Y,X)$ (see Lemma 3.1) is not dependent on n, $\mathrm{Var}(X)$ is, i.e., depends on how long the process has evolved. However, as n gets large relative to h, the effect of n on $\mathrm{Var}(X)$ becomes arbitrarily small. In applications, if the built-up time of the EWMAs is large relative to the time horizon of the EWMAs, the effect becomes negligible.

The regression coefficients $(\beta_0, \ldots, \beta_{m-1}) = COV(Y,X)$ can now be computed. In the case $m = 1$:

$$\beta_0 = \frac{cov(s_n - s_{n-p}, s_n - e_n^{(1)})}{var(s_n - e_n^{(1)})} \sim \frac{h\sigma^2(1-e^{-r})}{h\sigma^2 1/2} = 2(1-e^{-r}).$$

Therefore the best estimate is given by

$$s_{n-p} = s_n - \beta_0(s_n - e_n^{(1)}) = s_n - 2(1 - e^{-r})(s_n - e_n^{(1)}).$$

In general, since each entry in $\mathrm{Cov}(Y, X)$ and $\mathrm{Var}(X)$ has the factor $h\sigma^2$, the β's will not depend on σ. The regression is given by

$$s_n - s_{n-p} = \sum_{i=0}^{m-1} \beta_i^{(m)}(r) \left(e_n^{(i)} - e_n^{(i+1)} \right),$$

where each $\beta_i^{(m)}$ has the form

$$\beta_i^{(m)}(r) = c + \left(\sum_{j=0}^{m-1} a_j r^j \right) e^{-r}.$$

Here are the first few:

$$\begin{aligned}
m = 1 &: \\
&\beta_0 = 2 - 2e^{-r} \\
m = 2 &: \\
&\beta_0 = 4re^{-r} \\
&\beta_1 = 4 + (-4 - 8r)e^{-r} \\
m = 3 &: \\
&\beta_0 = 2 + (-2 + 4r - 4r^2)e^{-r} \\
&\beta_1 = -4 + (4 - 8r + 16r^2)e^{-r} \\
&\beta_2 = 8 + (-8 - 16r^2)e^{-r} \\
m = 4 &: \\
&\beta_0 = (8r - 8r^2 + \tfrac{8}{3}r^3)e^{-r} \\
&\beta_1 = 8 + (-8 - 32r + 40r^2 - 16r^3)e^{-r} \\
&\beta_2 = -16 + (16 + 48r - 64r^2 + 32r^3)e^{-r} \\
&\beta_3 = 16 + (-16 - 32r + 32r^2 - \tfrac{64}{3}r^3)e^{-r}
\end{aligned}$$

3.2 Computation of the Regression Coefficients

This section investigates how the β's can be effectively computed without inverting the matrix $\mathrm{Var}(X)$. The recursion relations given in either Theorem 3.4 or Theorem 3.7 will make a computer implementation easy and efficient. It is unknown whether there is a *simple* closed-form solution for the β's. These recursions will also be needed in subsequent theorems.

Theorem 3.4.

(a) In the case of m EWMAs, all $\beta_k^{(m)}$, $0 \le k \le m-1$ have the form

$$\beta_k^{(m)} = -\alpha_0 + (\alpha_0 + \alpha_1 r + \cdots + \alpha_{m-1} r^{m-1})e^{-r}.$$

(b) Let $\beta(m, k, q) = \alpha_q$ for $0 \le k \le m-1$, $0 \le q \le m-1$ and $\beta(m, k, q) = 0$ for all other indices, then the following recursion holds:

(i)

$$\beta(m,0,0) = (-1)^m - 1 \text{ for } m \in \mathbb{N},$$

(ii)

$$\beta(m,k,0) = 2\beta(m-1,k-1,0) - \beta(m-1,k,0) \text{ for } m \geq 2,\ 1 \leq k \leq m-1,$$

(iii)

$$\beta(m,k,q) = \frac{4}{q}\beta(m-1,k-1,q-1) - \frac{2}{q}\beta(m-1,k,q-1) - 2\beta(m-1,k-1,q) + \beta(m-1,k,q)$$

$$\text{for } m \in \mathbb{N},\ 0 \leq k \leq m-1,\ 1 \leq q \leq m-1$$

Definition 3.5. *Let the $(m+1 \times m+1)$ matrix $\hat{C}$ be defined as follows (rows and columns are indexed from -1 to $m-1$ to maintain consistency in notation):*

(a) All entries in the top row are 1's, i.e., $\hat{c}_{-1,l} = 1$ for $-1 \leq l \leq m-1$,

(b) Entries in the leftmost column are given by:

$$\hat{c}_{k,-1} = 2^k \left(1 - \sum_{i=0}^{k} \frac{r^i}{i!} e^{-r}\right) \text{ for } 0 \leq k \leq m-1,$$

(c) The remaining entries are the Pascal matrix

$$\hat{c}_{kl} = \binom{k+l}{k} \text{ for } 0 \leq k,l \leq m-1.$$

E.g., the matrix for $m = 3$ is given by

$$\hat{C} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 - e^{-r} & 1 & 1 & 1 \\ 2 - 2(1+r)e^{-r} & 1 & 2 & 3 \\ 4 - 4(1 + r + \frac{1}{2}r^2)e^{-r} & 1 & 3 & 6 \end{pmatrix}.$$

Further let $M(A,k,l)$ denote the k,l minor of a matrix A.

Lemma 3.6.

$$\beta_k = (-1)^k 2^{k+1} M(\hat{C}, -1, k).$$

Proof. We will drop the $h\sigma^2$ factors in all expressions, since they will cancel in β_i. Let $c_{kl} = \binom{k+l}{k}/2^{k+l+1}$ be the entries of $\mathrm{Var}(X)$ (indexed from 0 to $m-1$) as in Lemma 3.3. Multiplying row k of $\mathrm{Var}(X)$ by 2^k and column l by 2^{l+1} gives the Pascal matrix whose determinant is 1. A total of $m(m-1)/2 + (m+1)m/2 = m^2$ factors of 2 have been applied. Hence $\det(\mathrm{Var}(X)) = 2^{-m^2}$.

Let c_k be as in Lemma 3.2, b_{kl} the entries of $\mathrm{Var}^{-1}(X)$ and P the Pascal matrix. Then

$$\begin{aligned} \beta_k &= \sum_{i=0}^{m-1} c_i b_{ki} = \sum_{i=0}^{m-1} c_i 2^{m^2} (-1)^{i+k} M(\mathrm{Var}^{-1}(X), i, k) \\ &= \sum_{i=0}^{m-1} c_i 2^{m^2} (-1)^{i+k} M(P,i,k) 2^{-m^2+i+(k+1)} = \sum_{i=0}^{m-1} c_i 2^{i+k+1} (-1)^{i+k} M(P,i,k). \end{aligned}$$

The Lemma follows since expanding $M(\hat{C}, -1, k)$ across the leftmost column yields

$$M(\hat{C}, -1, k) = \sum_{i=0}^{m-1} (-1)^i 2^i c_i M(P, i, k).$$

□

Proof of Theorem. Part (a) follows directly from the lemma. For part (b) note that $\beta(m, k, q)$ can be computed as the determinant of the matrix $\hat{D}_{kq}$, defined as the matrix $\hat{C}$ with the top row and k'th column removed and the entries in the leftmost column replaced by the coefficient of $r^q e^{-r}$ multiplied by $(-1)^k 2^{k+1}$. The rows and column entries $\hat{d}_{ij}$ will be indexed from 0 to $m-1$ (dropping k and q from the notation).

We will show (iii) first. Let $q \geq 1$. We have

$$\hat{d}_{i0} = \begin{cases} 0 & \text{if} \quad i < q \\ \frac{(-1)^{k+1}}{q!} 2^{i+k+1} & \text{if} \quad i \geq q \end{cases}, \quad \hat{d}_{ij} = \begin{cases} \binom{i+j-1}{j-1} & \text{if} \quad 1 \leq j \leq k \\ \binom{i+j}{j} & \text{if} \quad j \geq k+1 \end{cases}.$$

Subtracting row i from row $i+1$ in $\hat{D}_{kq}$ gives (again we use $\hat{d}_{ij}$):

$$\hat{d}_{i0} = \begin{cases} 0 & \text{if} \quad i < q \\ \frac{(-1)^{k+1}}{q!} 2^{q+k+1} & \text{if} \quad i = q \\ \frac{(-1)^{k+1}}{q!} 2^{i+k} & \text{if} \quad i > q \end{cases},$$

$$\hat{d}_{ij} = \begin{cases} \binom{i+j-1}{j-1} - \binom{i+j-2}{j-1} = \binom{i+j-2}{j-2} & \text{if} \quad 1 \leq j \leq k \\ \binom{i+j}{j} - \binom{i+j-1}{j} = \binom{i+j-1}{j-1} & \text{if} \quad j \geq k+1 \end{cases}.$$

Subtracting column j from column $j+1$, $j \geq 1$ gives (columns with index 0 and 1 remain the same):

$$\hat{d}_{ij} = \begin{cases} \binom{i+j-2}{j-2} - \binom{i+j-3}{j-3} = \binom{i+j-3}{j-2} & \text{if} \quad 2 \leq j \leq k \\ \binom{i+j-1}{j-1} - \binom{i+j-3}{j-3} & \text{if} \quad j = k+1 \\ \binom{i+j-1}{j-1} - \binom{i+j-2}{j-2} = \binom{i+j-2}{j-1} & \text{if} \quad j > k+1 \end{cases}.$$

Note that for $k = 0, 1, m-2, m-1$ not all cases above occur, but the argument that follows still holds. Since $q \geq 1$, $\hat{d}_{00} = 0$, $\hat{d}_{01} = 1$ and all other entries in the top row are zero. Hence we can compute $\det(\hat{D})$ by computing the negative of the $0, 1$ minor after the row and column operations. Let A be the corresponding matrix of this minor, folding the negative sign into the first column. Indexed from 0 to $m-2$ the entries of A are given by:

$$a_{i0} = \begin{cases} 0 & \text{if} \quad i < q-1 \\ \frac{(-1)^k}{q!} 2^{q+k+1} & \text{if} \quad i = q-1 \\ \frac{(-1)^k}{q!} 2^{i+k+1} & \text{if} \quad i \geq q \end{cases}, \quad a_{ij} = \begin{cases} \binom{i+j-1}{j-1} & \text{if} \quad 1 \leq j \leq k-1 \\ \binom{i+j+1}{j} - \binom{i+j-1}{j-2} & \text{if} \quad j = k \\ \binom{i+j}{j} & \text{if} \quad j > k \end{cases}. \quad (7)$$

Assume first that $k \geq 1$. Then $\frac{4}{q}\beta(m-1, k-1, q-1)$ can be computed as the determinant of the following matrix (indexed from 0 to $m-2$):

$$(\cdot)_{i0} = \begin{cases} 0 & \text{if} \quad i < q-1 \\ \frac{(-1)^k}{q!} 2^{i+k+2} & \text{if} \quad i \geq q-1 \end{cases}, \quad (\cdot)_{ij} = \begin{cases} \binom{i+j-1}{j-1} & \text{if} \quad 1 \leq j \leq k-1 \\ \binom{i+j}{j} & \text{if} \quad j \geq k \end{cases}.$$

Similarly, compute $-\frac{2}{q}\beta(m-1,k,q-1)$ via

$$(\cdot)_{i0} = \begin{cases} 0 & \text{if} \quad i < q-1 \\ \frac{(-1)^k}{q!}2^{i+k+2} & \text{if} \quad i \geq q-1 \end{cases}, \quad (\cdot)_{ij} = \begin{cases} \binom{i+j-1}{j-1} & \text{if} \quad 1 \leq j \leq k \\ \binom{i+j}{j} & \text{if} \quad j \geq k+1 \end{cases}.$$

The previous two matrices agree except for column k. Hence, the sum of the determinants can be computed by evaluating the determinant of the matrix identical to the two matrices, expect that the columns k have been added. Call this matrix A_1. To summarize, A_1 consists of the columns:

$$(\cdot)_{i0} = \begin{cases} 0 & \text{if} \quad i < q-1 \\ \frac{(-1)^k}{q!}2^{i+k+2} & \text{if} \quad i \geq q-1 \end{cases}, \quad (\cdot)_{ij} = \begin{cases} \binom{i+j-1}{j-1} & \text{if} \quad 1 \leq j \leq k-1 \\ \binom{i+j}{j} + \binom{i+j-1}{j-1} & \text{if} \quad j = k \\ \binom{i+j}{j} & \text{if} \quad j \geq k+1 \end{cases}.$$

A similar argument shows that $-2\beta(m-1,k-1,q) + \beta(m-1,k,q)$ can be computed via the determinant of the matrix A_2 with columns:

$$(\cdot)_{i0} = \begin{cases} 0 & \text{if} \quad i < q \\ \frac{(-1)^{k+1}}{q!}2^{i+k+1} & \text{if} \quad i \geq q \end{cases}, \quad (\cdot)_{ij} = \begin{cases} \binom{i+j-1}{j-1} & \text{if} \quad 1 \leq j \leq k-1 \\ \binom{i+j}{j} + \binom{i+j-1}{j-1} & \text{if} \quad j = k \\ \binom{i+j}{j} & \text{if} \quad j \geq k+1 \end{cases}. \quad (8)$$

The matrices A_1 and A_2 agree except for the first column. Hence $\det(A_1) + \det(A_2)$ can be computed by the determinant of A_3 which is the same as A_1, except the first column is the sum of the first columns of A_1 and A_2. The first column o fA_3 is given by:

$$\hat{a}_{i0} = \begin{cases} 0 & \text{if} \quad i < q-1 \\ \frac{(-1)^k}{q!}2^{q+k+1} & \text{if} \quad i = q-1 \\ \frac{(-1)^k}{q!}2^{i+k+1} & \text{if} \quad i \geq q \end{cases}.$$

Finally, note that

$$\binom{i+j}{j} + \binom{i+j-1}{j-1} = \binom{i+j+1}{j} - \binom{i+j-1}{j-2}.$$

Hence the matrices A and A_3 are identical. This proves (iii) in the case $k \geq 1$.

If $k = 0$ then $\beta(m-1,k-1,q-1) = \beta(m-1,k-1,q) = 0$. This means that A_1 and A_2 are not sums of two matrices, but defined by $-2/q\beta(m-1,k,q-1)$ and $\beta(m-1,k,q)$, respectively. The argument above still holds.

To prove (ii), note that in this case the first two terms of the RHS of (iii) are zero and the remaining terms are the negative of the RHS of (ii). Showing that $\det(A) = -\det(A_3)$ will prove (ii). If $q = 0$ and $k \geq 1$, the matrix $\hat{D}_{k0}$, after the row and column transformations, has the form

$$\begin{pmatrix} (-1)^{k+1}2^{k+1} & 1 & 0 & 0 & \cdots \\ (-1)^{k+1}2^{k+1} & 0 & \cdot & \cdot & \cdots \\ (-1)^{k+1}2^{k+2} & 0 & \cdot & \cdot & \cdots \\ (-1)^{k+1}2^{k+3} & 0 & \cdot & \cdot & \cdots \\ \cdots & & & & \end{pmatrix}.$$

Again, let A be the matrix associated with the $0,1$ minor of this matrix. The first column of A is given by $a_{i0} = (-1)^k 2^{k+1+i}$ (agrees with (7)). This case differs since $A_1 = 0$ and hence $A_3 = A_2$. By (8) the entries of the first column of A_3 are given by $(-1)^{k+1}2^{i+k+1}$. This is the negative of the first column of A. All other columns of A and A_3 are equal.

Finally, the matrix $\hat{D}_{00}$ associated with $\beta(m,0,0)$ before and after row and column operations is given by

$$\begin{pmatrix} -2 & 1 & 1 & 1 & \cdots & 1 \\ -4 & 2 & 3 & & & \\ -8 & 3 & 6 & & \ddots & \\ -16 & 4 & 10 & & & \\ \cdots & & & & & \end{pmatrix}, \quad \begin{pmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ -2 & & & & & \\ -4 & & & P & & \\ -8 & & & & & \\ \cdots & & & & & \end{pmatrix},$$

where P is the Pascal matrix. The case (i) differs from the other cases since the determinant of $\hat{D}_{00}$ cannot be computed via the $0,1$ minor. However, expanding across the first row gives

$$\beta(m,0,0) = -2\det(P) - \beta(m-1,0,0) = -2 - \beta(m-1,0,0).$$

Since $\beta(1,0,0) = -2$, it follows that $\beta(m,0,0) = (-1)^m - 1$. □

The βs can also be computed via another recursion relation. Recall that

$$\beta_k^{(m)} = -\beta(m,k,0) + p_k^{(m)}(r)e^{-r},$$

where $p_k^{(m)}$ is a polynomial of degree $m-1$ in r with coefficients $\beta(m,k,q)$ of r^q. Let $\bar{p}_k^{(m)} = p_k^{(m)}/\beta(m,k,0)$ if $\beta(m,k,0) \neq 0$ (note that $\beta(m,m-1,0) \neq 0$). The following theorem gives a recursive way to define the β's with respect to β_{m-1}, i.e., w.r.t. the regression coefficient of the highest order EWMA. These formulae allow us to express other quantities (such as the expected error in the next section) in terms of the highest β. Part (b) of the theorem shows that adding another iteration of an EWMA changes the regression coefficients by multiples of the newly added highest order β.

Theorem 3.7.

(a) The β_k with highest index $k = m-1$ are given by

$$\begin{aligned} \beta(m,m-1,0) &= -2^m, \\ \bar{p}_0^{(1)} &= 1, \\ \bar{p}_{m-1}^{(m)} &= 2 - \bar{p}_{m-2}^{(m-1)} + 2\int_0^r \bar{p}_{m-2}^{(m-1)}(s)ds, \quad \textit{for } m \geq 2. \end{aligned} \tag{9}$$

(b)

$$\beta_k^{(m)} = \beta_k^{(m-1)} + (-1)^{m+k+1}\frac{\binom{m-1}{k}}{2^{m-k-1}}\beta_{m-1}^{(m)}, \quad \textit{for } 0 \leq k \leq m-2,\ m \geq 2.$$

Proof. By Theorem 3.4 (b) part (iii)

$$-\frac{1}{2^m}\beta(m,m-1,q) = -\frac{1}{2^{m-1}}\left(\frac{2}{q}\beta(m-1,m-2,q-1) - \beta(m-1,m-2,q)\right),$$

which proves part (a) by comparing coefficients of r^q.

To prove (b), let $P_{m,k}$ be the $m \times m-1$ matrix obtained by dropping column k from the Pascal matrix, $k = 0, \ldots, m-1$. Rewrite the equation in (b) as

$$(-1)^{-k}2^{-(k+1)}\beta_k^{(m)} = (-1)^{-k}2^{-(k+1)}\beta_k^{(m-1)} + (-1)^{m+1}\frac{\binom{m-1}{k}}{2^m}\beta_{m-1}^{(m)}.$$

By Lemma 3.6 the left hand side is the determinant of a matrix whose first column contain the $\hat{c}$ of Definition 3.5 (b) and the remaining columns form $P_{m,k}$. The first term on the RHS is the determinant of the similar $m-1 \times m-1$ matrix and the second term is $\binom{m-1}{k}$ times the determinant of the matrix formed using $P_{m,m-1}$. Call these matrices A, B and C. Add a row and column to B to obtain $\hat{B}$:

$$B = \begin{pmatrix} \hat{c}_0 & \\ \hat{c}_1 & \\ \cdots & P_{m-1,k} \\ \hat{c}_{m-2} & \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} \hat{c}_0 & & & & 0 \\ \hat{c}_1 & & & & 0 \\ \cdots & & P_{m-1,k} & & \cdots \\ \hat{c}_{m-2} & & & & 0 \\ \hat{c}_{m-1} & p_1 & \cdots & p_{m-2} & 1 \end{pmatrix},$$

where the p_i are extending the columns of the Pascal matrix $P_{m-1,k}$ in the natural way. Note that $\det(B) = \det(\hat{B})$.

Let $\hat{C}$ be the matrix obtained by exchanging column k and the last column of C (exchange column k with $k+1$, $k+1$ with $k+2$, etc., performing $m-k-2$ operations) and then multiplying the last column by $(-1)^{m-k-2}\binom{m-1}{k}$. Then $\det(\hat{C}) = \binom{m-1}{k}\det(C)$. Now $\hat{B}$ and $\hat{C}$ have identical columns except for the last column. Hence $\det(B) + \binom{m-1}{k}\det(C)$ can be computed as the determinant of the matrix D obtained from $\hat{B}$, adding the last columns of $\hat{C}$ and $\hat{B}$.

We need to show that $\det(D) = \det(A)$. A and D are identical except for the last column. The last column of A has entries in row i: $\binom{i+m-1}{m-1}$, $0 \le i \le m-1$ (note that $k \le m-2$) whereas D has last column entries: $(-1)^{m-k-2}\binom{m-1}{k}\binom{i+k}{k}$ if $0 \le i < m-1$ and $(-1)^{m-k-2}\binom{m-1}{k}\binom{m-1+k}{k} + 1$ if $i = m-1$. We will show that adding multiples of columns of D to the last column of D will give the last column of A and hence $\det(A) = \det(D)$. The result follows if we can show:

$$\begin{aligned} \binom{i+m-1}{m-1} &= (-1)^{m-k-2}\binom{m-1}{k}\binom{i+k}{k} + \sum_{\mu=0,\mu\neq k}^{m-2}(-1)^{m+\mu}\binom{i+\mu}{\mu}\binom{m-1}{\mu} \\ &= \sum_{\mu=0}^{m-2}(-1)^{m+\mu}\binom{i+\mu}{\mu}\binom{m-1}{\mu}, \quad \text{for } i < m-1. \end{aligned}$$

If $i = m-1$ add 1 to the RHS.

By Proposition A.6 (using $a = m-1$, $b = 0$, $c = i$, $d = i$ and dividing by $(-1)^m$)

$$\sum_{\mu=0}^{m-1}(-1)^{m+\mu}\binom{i+\mu}{\mu}\binom{m-1}{\mu} = \begin{cases} 0 & \text{if} \quad i < m-1 \\ -1 & \text{if} \quad i = m-1 \end{cases}.$$

Therefore,

$$\sum_{\mu=0}^{m-2}(-1)^{m+\mu}\binom{i+\mu}{\mu}\binom{m-1}{\mu} - \binom{m-1+i}{m-1} = \begin{cases} 0 & \text{if} \quad i < m-1 \\ -1 & \text{if} \quad i = m-1 \end{cases}.$$

□

Remark 3.8. *A few facts (without proof):*

(a) Using the initial polynomial $\hat{p}_1 = 2$ and defining higher order polynomials $\hat{p}_m$ by the recurrence relation (9) (replacing bars with hats) then $\hat{p}_m = (-1)^m p_0^m$. In other words, using the initial condition $p_1 = 1$ yields the polynomial part divided by 2^m of the highest order β while using $p_1 = 2$ yields $(-1)^m$ times the polynomial part of the lowest order β.

(b)

$$\frac{d}{dr}\bar{p}_{m-1}^{(m)}(r) = (-1)^{m-1} p_0^{(m-1)}(r), \tag{10}$$

i.e., the derivative of the normalized polynomial part of the highest order β is a multiple of the non-normalized polynomial part of the lowest order β using one EWMA iteration less.

(c) While closed form expression for $\beta(m, k, q)$ are currently unknown, such expressions can be found for subsets of β's, e.g.:

$$\begin{aligned} \beta(m, 0, 3) &= \frac{1}{9}m(m-2)(2m-5) + \frac{1}{6}((-1)^m - 1) \\ \beta(m, 1, 1) &= (-1)^{m-1}(2m-1) - 2m(m-1) - 1 \end{aligned}$$

3.3 Time Series with drift

If $\mu \neq 0$ by Remark 2.2, Corollary 2.5 and Proposition A.3

$$E(s_n - s_{n-p}) = p\mu \tag{11}$$

$$E(e_n^{(k)} - e_n^{(k+1)}) = (1-w)^k \sum_{i=0}^{n-1} \binom{k+i}{k} w^{i+1}\mu$$

$$\lim_{n\to\infty} E(e_n^{(k)} - e_n^{(k+1)}) = \frac{w\mu}{1-w} \sim h\mu \text{ as } h \to \infty. \tag{12}$$

Therefore the least squares regression can be computed by

$$\begin{aligned} s_n - s_{n-p} - p\mu &= \beta_0(e^{(0)} - e^{(1)} - h\mu) + \cdots + \beta_{m-1}(e^{(m-1)} - e^{(m)} - h\mu) \\ s_n - s_{n-p} &= \beta_0(e^{(0)} - e^{(1)}) + \cdots + \beta_{m-1}(e^{(m-1)} - e^{(m)}) + h\mu(r - \sum \beta_i) \end{aligned}$$

Note on computer simulations/applications: a more precise asymptotic approximation in (12) is $(h - 0.5)\mu$. This 0.5 adjustment is negligible for large h. In practical applications, however, this adjustment will often give better results. The associated software package uses this 0.5 adjustment.

4 Expected Square Error

In this section, except otherwise noted, all processes are assumed to have zero drift. Without using EWMAs (i.e., the estimate is the current value s_n), the expected square error of estimating s_{n-p} equals the variance: $\mathrm{Var}(s_n - s_{n-p}) = p\sigma^2$. To normalize the square error and enable comparisons, all square errors are divided by $p\sigma^2$ throughout this section. In ordinary least squares regression, dividing the square error of the residuals by the variance yields $1 - R^2$, where R^2 is the coefficient of determination. In this sense, R^2 will be defined as $1 - SSE$ given in the following theorem:

Theorem 4.1. *Using m EWMAs, the normalized expected square error is given by*

$$SSE = 1 - \frac{2}{r}\sum_{k=1}^{m}\left(\frac{\beta_{k-1}^{(k)}}{2^k}\right)^2.$$

Lemma 4.2. *Using the notation of section 3 and setting* $\bar{c}_i = 1-(1+r+\cdots+r^i/i!)e^{-r}$,

$$\beta_k^{(k+1)} = \sum_{i=0}^{k}(-1)^{k+i}2^{i+k+1}\bar{c}_i\binom{k}{i}.$$

Proof. We will use that fact that if $P(m,i,j)$ is the i,j minor of the $m\times m$ Pascal matrix, indexed from $0,\ldots,m-1$, then $P(m,m-1,i)=\binom{m-1}{i}$. By Lemma 3.6, $(-1)^k2^{-(k+1)}\beta_k^{(k+1)}$ is the determinant of the following matrix:

$$\begin{pmatrix} \bar{c}_0 & 1 & 1 & 1 & \cdots & 1 \\ 2\bar{c}_1 & 1 & 2 & 3 & \cdots & k \\ \vdots & & & & & \vdots \\ 2^k\bar{c}_k & 1 & k+1 & \binom{k+2}{2} & \cdots & \binom{2k-1}{k-1} \end{pmatrix}.$$

Expanding across the first column,

$$(-1)^k2^{-(k+1)}\beta_k^{(k+1)} = \sum_{i=0}^{k}(-1)^i2^i\bar{c}_iP(k+1,k,i),$$

and hence,

$$\beta_k^{(k+1)} = \sum_{i=0}^{k}(-1)^{k+i}2^{i+k+1}\bar{c}_i\binom{k}{i}.$$

□

Proof of Theorem. The square error SSE is given by

$$\begin{aligned} SSE &= V(s_n - s_{n-p}) - \mathrm{Cov}(Y,X)\mathrm{Var}^{-1}(X)\mathrm{Cov}(Y,X)^T = p\sigma^2 - \beta\mathrm{Cov}(Y,X)^T \\ &= p\sigma^2 - h\sigma^2\sum_{i=0}^{m-1}\beta_i^{(m)}\bar{c}_i, \end{aligned}$$

where $\bar{c}_i = 1-(1+r+\cdots+r^i/i!)e^{-r}$ (see Lemma 3.2). Dividing by $p\sigma^2$ and using Theorem 3.7 (b) gives

$$SSE = 1 - \frac{1}{r}\sum_{i=0}^{m-1}\beta_i^{(m)}\bar{c}_i = 1 - \frac{1}{r}\sum_{i=0}^{m-1}\left[\sum_{k=i}^{m-1}\left(\frac{(-1)^{k+i}}{2^{k-i}}\binom{k}{i}\beta_k^{(k+1)}\right)\bar{c}_i\right].$$

Interchanging the order of summation and using the Lemma gives

$$SSE = 1 - \frac{1}{r}\sum_{k=0}^{m-1}\sum_{i=0}^{k}(-1)^{k+i}2^{i-k}\binom{k}{i}\beta_k^{(k+1)}\bar{c}_i = 1 - \binom{1}{r}\sum_{k=0}^{m-1}2^{-1-2k}\left(\beta_k^{(k+1)}\right)^2.$$

The result follows by re-indexing the sum from $k=1$ to $k=m$. □

The theorem says that the m^{th} order EWMA reduces the square error by $\frac{2}{r}(\beta_{m-1}^m/2^m)^2$, an expression involving only the highest order β.

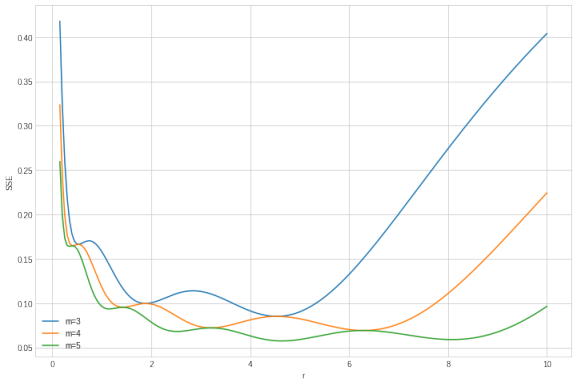

Figure 1: SSE for $m = 3, 4, 5$. Note that $SSE^{(m)}$ is minimal where $SSE^{(m+1)}$ is maximal.

5 Optimal Time Horizons

5.1 Point Estimates

For $m = 1$, Theorem 4.1 gives $SSE = 1 - \frac{2}{r}(1 - 2e^{-r} + e^{-2r})$. This is minimal for $r \approx 1.2564$ and $SSE \approx 0.1855$ for this r. This means the optimal time horizon is $1/r = 0.7959$ times the approximation period ($p = rh$). The standard deviation is $\sqrt{0.1855} = 0.431$. In other words, adding one moving average reduces the standard deviation from 1 to 0.431.

If $m = 2$,

$$SSE = 1 - \frac{2}{r}\left(2 - (4r + 4)e^{-r} + (4r^2 + 4r + 2)e^{-2r}\right) \tag{13}$$

is minimal for $r = 2.8427$ giving $SSE = 0.1139$ or a standard deviation of 0.3374, an additional reduction of 0.094. The optimal time horizon is 0.3518 times the approximation period. Figure 1 shows SSE for selected m. The reason that the curves are *wavy* results from the fact that each of the m EWMAs is best for estimating a previous value at a different time. Hence s_{n-p} is approximated better for certain values of p. Also, $SSE \to 1$ as $r \to 0$ or $r \to \infty$ (note that $SSE = 1 - R^2$).

5.2 Estimates over a Time Period

So far the question has been how to optimally estimate a fixed point in the past. Now our attention will focus on finding the optimal time horizon if the TS has to be approximated over a time interval, optimal meaning minimizing the SSE in some L^p norm over the period in question. We'll give answers for the L^1 and L^2 norms.

If the ξ_i in Definition 2.1 are chosen from a normal distribution, then $s_n - s_{n-p}$ as well as est_{n-p} are normally distributed, where est_{n-p} is the regression estimate of the difference. This follows from Corollary 2.5 and Theorem 3.4, showing that the EWMAs and the estimate are sums of iid normal variables. If the ξ_i are drawn from another distribution (e.g., the simple RW), then the estimates will be approximately normal by a Central Limit Theorem argument (proof omitted). Hence facts known about normal distributions can be applied to SSE.

L^1 **norm:** By the normality argument above and noting that for a $N(0, \sigma^2)$ distribution the mean absolute error is given by $MAE = \sqrt{2/\pi}\sigma$. To find the optimal time horizon h over the period p_0 to n, note that the *unadjusted* L^1 norm is given by

$$\frac{1}{p_0}\int_0^{p_0} MAE(p)\ dp = \frac{1}{p_0}\int_0^{p_0} \sqrt{\frac{2}{\pi}p\sigma^2 SSE(\frac{p}{h})}\ dp = \sqrt{\frac{2}{\pi}}\sqrt{p_0}\sigma\int_0^1 \sqrt{uSSE(\frac{u}{\hat{h}})}du,$$

where $\hat{h} = h/p_0$ is measured as a percentage of the period in question. Dividing by $\sqrt{p_0}\sigma$ yields the expected adjusted L^1 norm. Using numerical integration, for $m = 1$ the optimal time horizon in the L^1 norm is $h = 0.496$ with a normalized $L^1 = 0.2683$, and for $m = 2$ it is $h = 0.266$ with $L^1 = 0.2062$. These values are smaller than those for point estimates (note that approximations are better for small p).

L^2 **norm:** The above argument needs to be adjusted because SSE is the expected square error, i.e., a probabilistic expectation. The square root prohibits switching the order of taking the expectation and integral. However, we have the following inequality (let sse be the square error of individual instances and SSE its expectation given by Theorem 4.1):

$$\begin{aligned} E\left(\left[\frac{1}{p_0}\int_0^{p_0} sse(p)dp\right]^{\frac{1}{2}}\right) \leq &\quad \left[\tfrac{1}{p_0}\textstyle\int_0^{p_0} E\left(sse(p)\right)dp\right]^{\frac{1}{2}} &(14)\\ = &\quad \left[\tfrac{1}{p_0}\textstyle\int_0^{p_0} p\sigma^2 SSE(\frac{p}{h})dp\right]^{\frac{1}{2}} \\ = &\quad \sqrt{p_0}\sigma\left(\textstyle\int_0^1 uSSE(\frac{u}{\hat{h}})du\right)^{\frac{1}{2}} &(15)\end{aligned}$$

Last last expression in (15) can again be evaluated numerically and dividing by $\sqrt{p_0}\sigma$ yields a normalized error measure, denoted by $\hat{L}^2$. Using numerical integration, for $m = 1$ the optimal time horizon is given by $h = 0.528$ with a normalized $\hat{L}^2 = 0.3448$, and for $m = 2$ it is $h = 0.275$ with $\hat{L}^2 = 0.2665$. The simulations in Figure 2 illustrate the situation. The red curve labeled 'expected2' of the plot in the lower right hand corner shows the expected $\hat{L}^2$ values given in (15), the blue markers labeled "2MA11" the values obtained by simulations and the green markers labeled 'MA11' give the simulated results for the left hand side of (14), i.e., the true expected L^2 error. It appears L^2 and $\hat{L}^2$ differ by a fixed factor such that the optimal h for either L^2 or $\hat{L}^2$ is (at least approximately) the same.

5.3 Convergence

In this section we shed insight into the question of whether the estimates converge to the underlying TS as $m \to \infty$. One way to think about this is whether adding more iterated EWMAs expand the TS in a similar way as a Taylor series approximates an analytic function by adding more terms. Another explanation is statistical. In section 4 it was shown that $1 - SSE$ can be considered as an R^2 measure. If $R^2 = 1 - SSE \to 1$ then the moving averages would explain all the variance of the TS. In other words, the current values of the EWMAs, given a large enough number m, would explain the TS's behavior arbitrarily precise at any time in the past. At this point we only have the conjecture given below with some idea on how to go about proving it. The result is mostly of theoretical interest. Simulations using $m \geq 20$ results in jaggy (non-precise) estimates. If at all, the convergences is extremely slow.

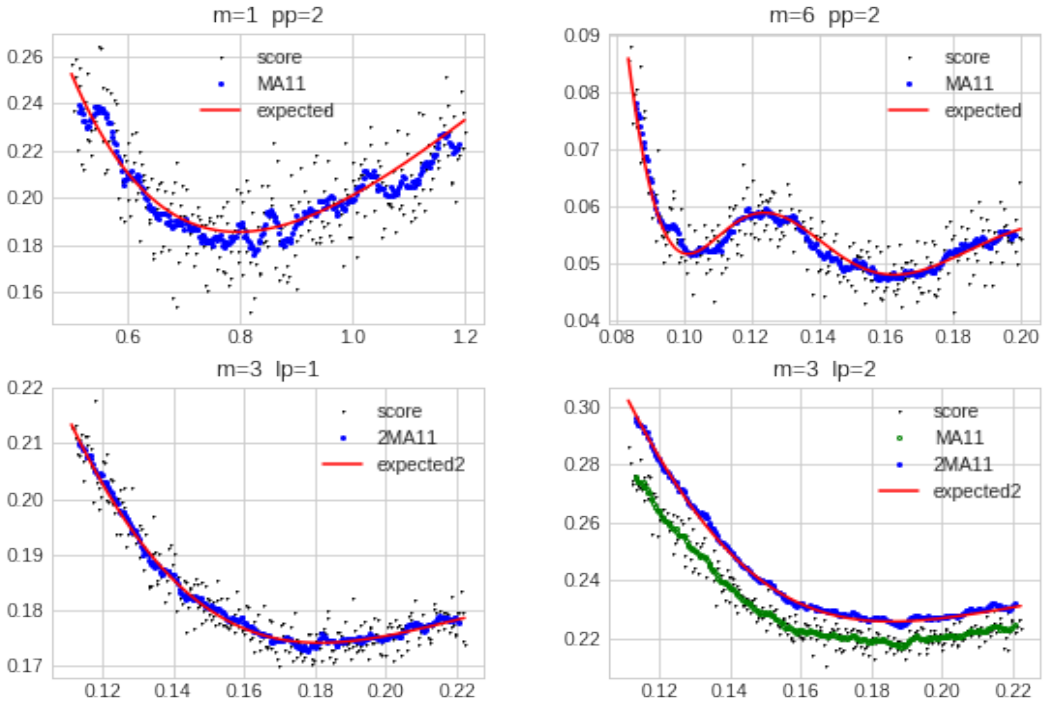

Figure 2: Simulation of errors. The top row shows simulations for point estimates, the lower row for L^1 and L^2 norms. The units of the horizontal axis are values of h and the vertical axis gives errors in the appropriate norm. The small black markers are the average error of 400 runs for each value of h. The blue markers smooth the individual errors by using a rolling window average of 11 (5 values to the left and 5 to the right of a given h). The red line shows the expected error given by the formulas in this section.

Conjecture 5.1. *For any compact interval $I = [a, b]$, $a > 0$, SSE converges to zero uniformly on I as $m \to \infty$.*

An idea of a proof is given below and will give additional insight into the structure of the regression coefficients.

Lemma 5.2. *Let $\alpha(m, q)$ be the coefficient of r^q in $\bar{p}_{m-1}^{(m)}$, where $\bar{p}$ is defined as in Theorem 3.7. Then*

$$\alpha(m, q) = \frac{2^q}{q!} \sum_{i=1}^{m-q-1} (-1)^i \binom{q-1+i}{q-1}.$$

Proof. Note that $\alpha(m, q) = -1/2^m \beta(m, m-1, q)$ and therefore by Lemma 4.2 ($\bar{c}_i$ contains $r^q/q!$ for $i \geq q$):

$$\alpha(m, q) = \sum_{i=q}^{m-1} (-1)^{m-1+i} \frac{2^i}{q!} \binom{m-1}{i} = 1 - \sum_{i=0}^{q-1} (-1)^{m-1+i} \frac{2^i}{q!} \binom{m-1}{i}.$$

The result follows from the second equality in Proposition A.2. □

Lemma 5.3.

$$m\bar{p}_{m-1}^{(m)} + (1+2r)\bar{p}_m^{(m+1)} - 2r\bar{p}_m'^{(m+1)} = (m+1)\bar{p}_{m+1}^{(m+2)},$$

where the prime indicates differentiation wrt r.

Proof. The equality, collecting of coefficients r^q, becomes

$$m\alpha(m, q) + \alpha(m+1, q) + 2\alpha(m+1, q-1) - 2q\alpha(m+1, q) = (m+1)\alpha(m+2, q).$$

Since $q\alpha(m+2, q) = 2\alpha(m+1, q-1) - q\alpha(m+1, q)$ by Theorem 3.4 (iii) one needs to show

$$m\alpha(m, q) - (q-1)\alpha(m+1, q) = (m+1-q)\alpha(m+2, q). \tag{16}$$

By the previous lemma, the LHS equals

$$(m-q+1)\frac{2^q}{q!} \sum_{i=1}^{m-q-1} (-1)^i \binom{q-1+i}{q-1} - \frac{2^q}{q!}(q-1)(-1)^{m-q} \binom{m-1}{q-1}.$$

The result follows since

$$(q-1)\binom{m-1}{q-1} = (m-q+1)\left(\binom{m}{q-1} - \binom{m-1}{q-1}\right),$$

giving the $i = m-q$ and $i = m-q+1$ terms in the sum above, which then equals the RHS of (16) by Lemma 5.2. □

Theorem 5.4.

$$\frac{d}{dr} SSE^{(m)}(r) = \frac{m}{r^2 2^{2m}} \beta_{m-1}^{(m)}(r) \beta_m^{(m+1)}(r).$$

Proof. All functions depend on r and we drop the notation. By Theorem 4.1

$$SSE^{(m)} = 1 - \frac{2}{r}\sum_{k=1}^{m}(1-\bar{p}_{k-1}^{(k)})^2 = 1 - \frac{2}{r}\left(m - 2e^{-r}\sum_{k=1}^{m}\bar{p}_{k-1}^{(k)} + e^{-2r}\sum_{k=1}^{m}\bar{p}_{k-1}^{(k)^2}\right).$$

After differentiating, the result of collecting terms of $\frac{2}{r^2}e^{-r}$ is given by

$$-2\sum_{k=1}^{m}\bar{p}_{k-1}^{(k)} - 2r\sum_{k=1}^{m}\left(\bar{p}_{k-1}^{(k)} - \bar{p}_{k-1}^{\prime(k)}\right).$$

We will show by induction on m that this equals $-m(\bar{p}_{m-1}^{(m)} + \bar{p}_{m}^{(m+1)})$. The case $m = 1$ yields $-2 - 2r$ ($\bar{p}_0^{(1)} = 1$ and $\bar{p}_1^{(2)} = 1 + 2r$). For the induction step, adding the term $k = m + 1$ to the sums and using the induction hypothesis,

$$-m(\bar{p}_{m-1}^{(m)} + \bar{p}_{m}^{(m+1)}) - 2(1+r)\bar{p}_{m}^{(m+1)} + 2r\bar{p}_{m}^{\prime(m+1)} = -(m+1)(\bar{p}_{m}^{(m+1)} + \bar{p}_{m+1}^{(m+2)})$$

by the previous lemma.

After differentiating SSE, the result of collecting terms of $\frac{2}{r^2}e^{-2r}$ is given by

$$\sum_{k=1}^{m}(\bar{p}_{k-1}^{(k)})^2 - 2r\sum_{k=1}^{m}\left(\bar{p}_{k-1}^{(k)}\bar{p}_{k-1}^{\prime(k)} - \bar{p}_{k-1}^{(k)^2}\right).$$

Again, using induction on m and the previous lemma, this equals $m\bar{p}_{m-1}^{(m)}\bar{p}_{m}^{(m+1)}$. Therefore

$$\begin{aligned}
\frac{d}{dr}SSE^{(m)} &= \frac{2}{r^2}\left(m - m(\bar{p}_{m-1}^{(m)} + \bar{p}_{m}^{(m+1)})e^{-r} + m\bar{p}_{m-1}^{(m)}\bar{p}_{m}^{(m+1)}e^{-2r}\right) \\
&= \frac{2m}{r^2}(1 - \bar{p}_{m-1}^{(m)}e^{-r})(1 - \bar{p}_{m}^{(m+1)}e^{-r}) \\
&= \frac{m}{r^2 2^{2m}}\beta_{m-1}^{(m)}\beta_{m}^{(m+1)}.
\end{aligned}$$

□

The theorem indicates that $SSE^{(m)}$ has local extrema only if the highest order $\beta_{m-1}^{(m)}$ or $\beta_{m}^{(m+1)}$ (next higher order) are zero. The interpretation is given by Theorem 4.1. Adding an extra EWMA of order $m + 1$ reduces $SSE^{(m)}$ by $1/(r2^{2m+1})(\beta_{m}^{(m+1)})^2$. Hence at points where $\beta_{m}^{(m+1)} = 0$, $SSE^{(m)}$ is optimal and forms a local minimum. However $SSE^{(m+1)}$ is not reduced at these points and so forms local maxima. In other words, wherever $SSE^{(m)}$ has a local min, $SSE^{(m+1)}$ forms a local max; the points where $SSE^{(m)}$ is minimal correspond to points where $\beta_{m}^{(m+1)} = 0$. Figure 3 illustrates the situation.

Lemma 5.5.

$$\bar{p}_{m-1}^{(m)}(r) \to e^r \text{ as } m \to \infty \text{ for } r \geq 0.$$

Proof.

$$\begin{aligned}
SSE^{(m)}(r) &= 1 - \frac{2}{r}\sum_{k=1}^{m}\left(\frac{\beta_{k-1}^{(k)}(r)}{2^k}\right)^2 = 1 - \frac{2}{r}\sum_{k=1}^{m}\left(1 - \bar{p}_{k-1}^{(k)}(r)e^{-r}\right)^2 \\
&= 1 - \frac{2}{re^{2r}}\sum_{k=1}^{m}\left(e^r - \bar{p}_{k-1}^{(k)}\right)^2
\end{aligned}$$

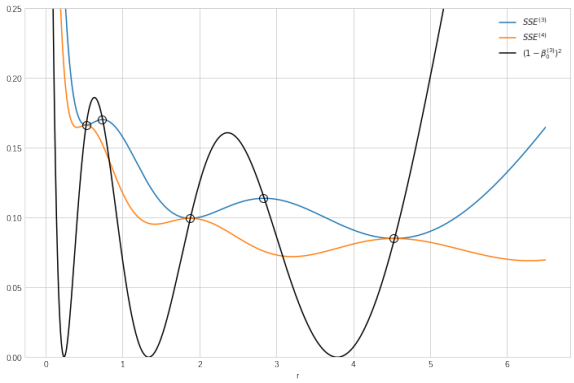

Figure 3: $SSE^{(3)}$ is minimal where $SSE^{(4)}$ is maximal. $(1-\beta_0^{(3)})^2$ passes through the local extrema of $SSE^{(3)}$.

Since $SSE > 0\ \forall r \geq 0$, the sum must converge and hence the lemma follows. □

The lemma implies that the $\beta_{m-1}^{(m)}/2^m$ converge to zero. By the integral equation (9) it is clear that if $\bar{p}_{m-1}^{(m)}$ converges, it must converge to e^r. This is true if we replace the initial condition $\bar{p}_0^{(1)} = 1$ by any other constant. Let $p_0^{(m)}$ be the polynomials resulting from the integral equation replacing the initial condition by $\bar{p}_0^{(1)} = 2$. Then by Remark 3.8 (a) we have $\beta_0(m)(r) = p_0^{(m)}(r)e^{-r}$ if m is even and $2 - p_0^{(m)}(r)e^{-r}$ if m is odd. Again, if the $p_0^{(m)}$ converge they must converge to e^r. Hence $\beta_0^{(m)} \to 1$.

Another indication that $\beta_0^{(m)} \to e^r$ for $r > 0$ is the following: Lemma 5.5 implies that $\bar{p}_{m-1}^{(m)}(r)e^{-r} \to 1$ for $r > 0$. It is then plausible that the derivative converges to zero. We have

$$\begin{aligned} \frac{d}{dr}\bar{p}_{m-1}^{(m)}(r)e^{-r} &= \bar{p}\prime_{m-1}^{(m)}(r)e^{-r} - \bar{p}_{m-1}^{(m)}(r)e^{-r} = (-1)^{m-1}p_0^{m-1}(r)e^{-r} - \bar{p}_{m-1}^{(m)}(r)e^{-r} \\ &\approx (-1)^{m-1}p_0^{m-1}(r)e^{-r} - 1 = \pm(\beta_0^{(m-1}(r) - 1), \end{aligned}$$

the $\pm$ depending on whether m is even or not (see Remark 3.8 (b)). The convergence of the derivative, however, is not obvious. Note that $1 - \beta_0^{(m)}$ appears similar to a sine curve with increasing frequency as m increases. The derivatives could theoretically become unbounded. Also convergence of the $p_0^{(m)}$ for $r > 0$ needs more work, e.g., the polynomials will diverge for $r \leq 0$.

Outline of a possible proof of the Conjecture: It is very likely true that if $\hat{r}$ is a local extremal point of $SSE^{(m)}$ then $SSE^{(m)}(\hat{r}) = (1 - \beta_0^{(m)}(\hat{r}))^2$. This can possibly be shown by algebraic simplifications using binomial identities (along the lines of the proof of Theorem 5.4). Figure 3 illustrates the situation. By the arguments above, $1 - \beta_0^{(m)}(r)$ probably converges to zero for $r > 0$ and the result follows.

Another approach to prove convergence is to address the rate at which SSE approaches zero. We conjecture that SSE is asymptotic to $1/(\pi\sqrt{2mr - r^2})$. More precisely:

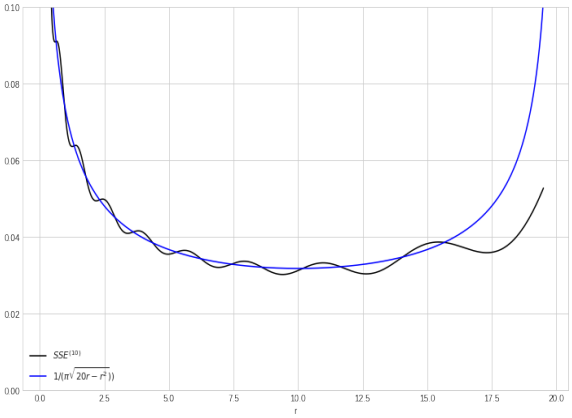

Figure 4: The square error for $m = 10$ and $1/(\pi\sqrt{2mr - r^2})$.

Conjecture 5.6. *For any closed interval* $[a, b]$, $a > 0$ *and* $\epsilon > 0$, *there exists* M *such that* $\forall m > M$

$$|SSE^{(m)}(r) - \frac{1}{\pi\sqrt{2mr - r^2}}| < \epsilon \ \forall r \in [a, b].$$

If this were true then the rate of convergence would be $O(\sqrt{m})$. (See Figure 4).

6 Applications and Simulations

6.1 Curve Fitting

Figure 5 shows some examples of curves fitted to simulated random walks. Unless otherwise noted, steps of the random walks were sampled from a standard normal distribution. The initial value of the RW was set to 100 and the approximation period was chosen as the last 1000 steps, i.e., from index 3000 to 4000.

Upper Left: displays iterated EWMAs for a RW with time horizon 100. **Upper Right:** the same RW with fitted curves for various choices of m. The time horizon for each fit was chosen to be optimal under the L^1 norm. **Middle Left:** the same random walk with fixed $m = 4$ and varying time horizons. Note that, on average, $h = 500$ is approximately optimal under the L^1 norm and $h = 800$ is approximately optimal for point estimates at index 3000. In this particular example, however, the point estimate using $h = 500$ is actually better than the one using $h = 800$. **Middle Right:** Another RW where a drift of 0.1 was added. The plot shows fitted curves with and without the drift adjustment. One must know the drift in advance, or estimated by some other method. For fits over short ranges the difference is negligible (and using no drift adjustment might be the preferred fit if the drift is unknown). Longer term, the drift adjustment becomes essential. **Lower Left:** Another example, this time showing that the fitted curves lose information further in the past. Indeed, the fits become asymptotic to horizontal lines. **Lower Right:** The RW was drawn from a simply RW with a small probability of a larger jump. In this case the RW drops 50 points at

index 3866. The fitted curves appear to handle the jump well; the theory appears to be applicable to a wide range of distributions of the steps of the RW.

6.2 Forecasting

All formulae developed in this make paper produce valid results for negative r. If one believes that past values of a TS have predictive power for the future, negative values of r may be interpreted as forecasts for the TS. Note that the theory breaks down for $r < 0$, e.g., the sum in Theorem 4.1 does not converge if $r < 0$. The regression coefficients have a factor of e^{-r} meaning that future estimates become unbounded rather quickly. In addition, there is a well defined slope of the approximations at the right edge of the TS which can easily be computed (there is a method for the slope in the accompanying software). This slope may be interpreted as the *momentum* of the TS at a given time.

Figure 6 exhibits an example. It is interesting to see that for $m = 6$ the forecast shows a small increase in value before dropping rapidly.

6.3 Modeling Financial Data

In the financial literature, stock prices are often assumed to be log-normally distributed, e.g., see [Ladde and Wu (2009)] or [Hilpisch (2020)] page 79. This means the prices, denoted by S_t, of assets follow a Geometric Brownian Motion (GBM). Let $\log S_t$ have mean μ and variance σ^2. To simulate S_t, the following algorithm can be used: choose and initial price S_0, then

$$S_n = S_{n-1} \exp\left((\mu - 0.5\sigma^2)\Delta t + \sigma\sqrt{\Delta t} x_n\right),$$

where the x_n are sampled from a standard normal distribution. The $-0.5\sigma^2$ adjustment is necessary since if $\log(S_t) \sim \mathcal{N}(\mu, \sigma^2)$, then S_t has expectation $\mu + 0.5\sigma^2$. However, this adjustment is often negligible for short-term simulations, e.g., for daily stock data, assuming an annualized volatility of 20%, then $0.5\sigma^2 \approx 0.000063$.

To compute the estimates using EWMAs, the algorithm is applied to the log of the prices, adjusting the drift with suitable values of μ and σ (which may have to be estimated), resulting in estimates of $\log S$. Exponentiation then reverts the estimates to the scale of actual prices.

In real applications it may be of advantage not to perform this conversion to log prices and treat actual stock prices as BM rather than GBM. The question is how to model variance, which does not appear in the regression coefficients if the variance is constant. If the variance is assumed constant in percentage terms (GBM) rather than absolute terms (BM), the theory of fitting curves no longer applies. However, if the price changes from one level to another, the EWMAs *learn* this new level of variance. Hence, the fitted curves are more accurate in the most recent history and lose their power further in the past (which may be desirable in applications). For long term modeling, using the geometric transformation become essential.

Figure 7 shows the fitted curves for Tesla corporation with and without the geometric adjustment. Data Source: Yahoo, retrieved via the pandas_datareader module.

6.4 Data Reduction

Computing EWMAs is extremely efficient and can be done online, requiring minimal computer memory requirements. in [Chen et al. (2001)] and [Chen et al. (2000)] EWMAs were used to track

Figure 5: Curve fitting in action using simulated data.

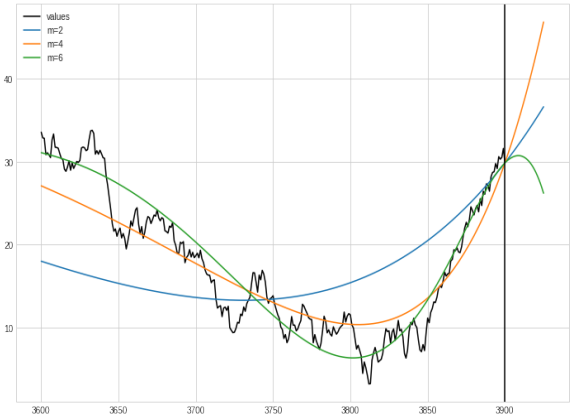

Figure 6: Forecasting a TS

Figure 7: Chart of Tesla stock prices

calling behavior of millions of wireless customers in real-time. It is possible to monitor financial instruments traded on exchanges in near real-time.

Any statistical algorithm of time series can be performed on the fitted curves instead. The brute force method would be to compute the fitted curves first and then perform the analysis on these rather than the original TS. Future work may include deriving formulae using only the available EWMA values rather than estimating the entire history. E.g., knowing only the current values of the EWMAs of TS A and B, what is the best estimate for the correlation of A and B?

6.5 Noise Filtering

The fitted curves, depending on the choice of m and the time horizon, exhibit the larger structure of the underlying TS. Using these estimates may thus avoid overfitting of models and be considered as a sort of noise reduction. The fits are more accurate in the recent history and lose information in the further past which may be an advantage. What looks like a significant pattern during the past week may look like noise a year from now.

6.6 Statistical Process Control

We propose that the EWMA estimates of a TS with drift μ and standard deviation σ may be used to detect changes in the drift μ. This is analogous to statistical process control using CUSUM or EWMA charts (for an overview see [Qiu (2014)]). WLOG assume the TS has been standardized so that $\mu = 0$ and we want to detect a change in μ away from zero. Let $est^{(m)}$ be the fitted curve using m EWMAs and time horizon h. One standard deviation of $s_{n-p} - est^{(m)}_{n-p}$ is given by $\sigma\sqrt{pSSE^{(m)}(p/h)}$. Hence the difference of the TS and the estimate can be plotted in terms of standard deviations. Preliminary simulations have shown that changes in μ can be detected quickly. However, no formal analysis has been performed.

Figure 8 gives an example of detecting 3σ (σ of $s_{n-p} - est^{(m)}_{n-p}$, not of the TS). The figure on the left shows a random walk and its fitted curve, using $m = 4$ and the time horizon chosen optimal under the L^1 norm for the past 1000 steps. The figure on the right shows the difference $s_{n-p} - est^{(m)}_{n-p}$ and the expected error expressed in standard deviations ($\sigma = 1, 2$ and 3). The dip around index 3250 forms a 3σ event, whereas the max around index 3480 just falls below that threshold. Note that the boundaries for σ-events are relatively narrow during the approximation period and widen considerably further in the past.

6.7 Feature Engineering

The various quantities explored in this paper can be used as inputs to other machine learning algorithms. Just to give a few examples, the EWMAs themselves or the slopes of the fitted curves could be features of other algorithms. σ events of the previous section can be used to label data (retrospectively) and then fed to other data mining techniques to see if these events can be predicted in real-time. Other possibilities could be points where various fitted curves cross (for different m and/or different time horizons). It has been a standard technical analysis technique of market data to find moving average crosses as trading signals. Future work may investigate whether using iterated EWMAs with optimal time horizons yield better signals.

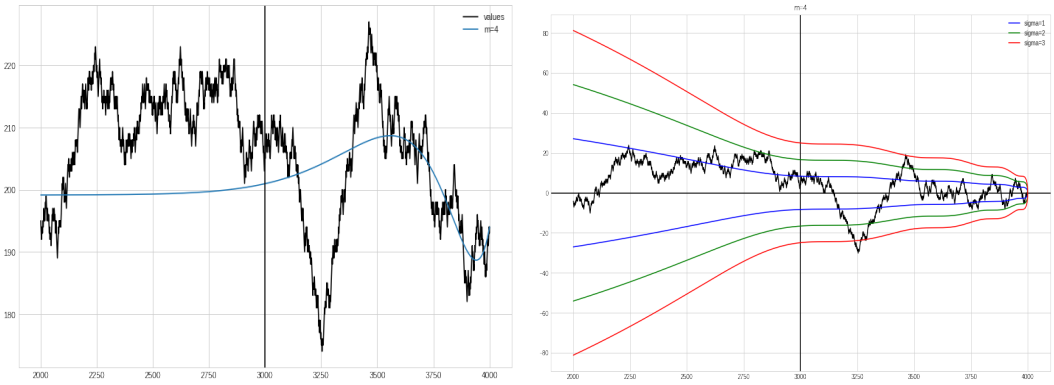

Figure 8: Detecting σ-events

7 Comparisons to other Data Reduction Techniques

7.1 EWMAs with Different Time Horizons

It is interesting to compare the case of using m iterated EWMAs with the same time horizon to using m EWMAs with different time horizons. We will show, that if $m = 2$ it is always best to use iterated EWMAs.

The setup and notation is identical to the case of iterated EWMAs, using a standardized random walk. The two EWMAs with time horizons h_1 and h_2 are denoted by $e^{(1)}$ and $e^{(2)}$. Set $Y = s_n - s_{n-p}$, $X_1 = s_n - e_n^{(1)}$, $X_2 = s_n - e_n^{(2)}$, $\lambda = h_2/h_1$ and $r = p/h_1$. Then $\mathrm{Cov}(Y, X_1) = h_1(1-e^{-r})$ and $\mathrm{Cov}(Y, X_2) = \lambda h_1(1 - e^{-r/\lambda})$. By arguments similar to the proof of Lemma 3.3 it follows that

$$\mathrm{Var}(X_1, X_2) = \begin{pmatrix} \frac{h_1}{2} & \frac{\lambda}{1+\lambda} h_1 \\ \frac{\lambda}{1+\lambda} h_1 & \frac{\lambda h_1}{2} \end{pmatrix}.$$

The regression coefficients are given by

$$(\beta_1, \beta_2) = \left(\frac{1-e^{-r}}{\frac{1}{2} - \frac{2\lambda}{(1+\lambda)^2}} - \frac{1-e^{-\frac{r}{\lambda}}}{\frac{(1+\lambda)}{4\lambda} - \frac{1}{1+\lambda}}, -\frac{1-e^{-r}}{\frac{1}{4}(1+\lambda) - \frac{\lambda}{1+\lambda}} + \frac{1-e^{-\frac{r}{\lambda}}}{\frac{1}{2} - \frac{2\lambda}{(1+\lambda)^2}} \right).$$

The expected square error equates to

$$SSE = 1 - \frac{1}{r}(\beta_1, \beta_2)(1-e^{-r}, \lambda(1-e^{-\frac{r}{\lambda}}))^T,$$

which, after some algebraic simplifications, becomes

$$SSE = 1 - \frac{1}{r}\frac{(1+\lambda)^2}{(1-\lambda)^2}\left(2(1-e^{-r})^2 + 2\lambda(1-e^{-\frac{r}{\lambda}})^2 - \frac{8\lambda}{(1+\lambda)}(1-e^{-r})(1-e^{-\frac{r}{\lambda}}) \right).$$

The limit as $\lambda \to 1$ is given by (using l'Hôpital on a computer algebra system)

$$1 - \frac{2}{r}\left(2 - (4r+4)e^{-r} + (4r^2+4r+2)e^{-2r}\right),$$

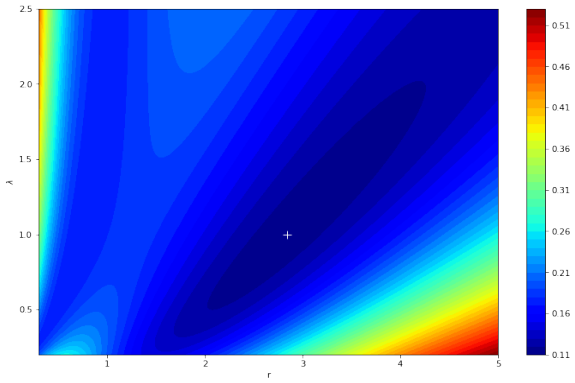

Figure 9: Heat map of SSE using two time horizons. The minimum SSE is at $\lambda = 1$ and $r = 2.8427$.

which is the same as (13), the normalized $SSE^{(2)}$ using two iterated EWMAs. The heatmap in Figure 9 illustrates the situation. The result means that for $m = 2$, iterated EWMAs with optimal time horizon always outperform two separate simple EWMAs with different time horizons. More generally, we have the following

Conjecture 7.1. *Given m EWMAs, either iterated, or with different time horizons, or a mixture of simple and iterated EWMAs, using m iterated EWMAs with a fixed time horizon that minimizes SSE in Theorem 4.1, yields the minimum SSE.*

For practical applications, there is no optimal setting using two single EWMAs, but if r is close to 2.8427 and $\lambda \approx 1$, i.e., both h_1 and $h_2 \approx 0.3518p$, but $h_1 \neq h_2$, the square error for estimating s_{n-p} is close to the optimal level using two iterated EWMAs ($SSE \approx 0.1139$). If $\lambda \approx 1$, however, both β_1 and β_2 are large in magnitude and the approximations become unstable. In other words, using two iterated EWMAs does not only give the best approximations, but the estimates are also stable under small fluctuation of the random walk.

7.2 Comparison to Storing Historical Values

Another way to approximate the past of a TS is to store historical values and then regress the TS on the stored values which results in linear interpolation.

The problem will be set analogous to the case of EWMAs. Let m be the number of values stored with values $x_0, \ldots, x_{m-1}$, with x_0 being the most recent value. We assume the values are spaced equally with time difference h. Once a time period of h has elapsed, the latest value of the TS is stored and the most distant value is dropped. Given a RW $\{s_n\}$, the problem is then to regress $Y = s_n - s_{n-p}$ on $X = (s_n - x_0,\ x_0 - x_1,\ \ldots,\ x_{m-2} - x_{m-1})$ (there are m stored values plus the latest value s_n, the time difference between s_n and x_0 varies, whereas the time differences between x_i and x_{i+1} are fixed at h). The RW is assumed to have no drift and be standardized such that $\sigma^2 = 1$. If $x_0 = s_{n-\nu}$, i.e., ν (≥ 1) is the number of steps elapsed since the last value was stored

and $\min_+(a,b) = \max(0, \min(a,b))$, then

$$\begin{aligned}\mathrm{Cov}(Y,X) = (&\min(p,\nu), \min_+(p-\nu,h), \min_+(p-h-\nu,h), \ldots,\\ &\min_+(p-(m-2)h-\nu,h)).\end{aligned}$$

This follows by arguments similar to the proof of Lemma 3.1. $\mathrm{Var}(X)$ is given by

$$\mathrm{Var}(X) = \begin{pmatrix} \nu & & & \\ & h & 0 & \\ & & & \\ & 0 & \ddots & \\ & & & h \end{pmatrix},$$

hence $\mathrm{Var}^{-1}(X)$ has entries in the main diagonal $\frac{1}{\nu}, \frac{1}{h}, \ldots, \frac{1}{h}$ and zeros otherwise. Therefore the regression coefficients are given by

$$\begin{aligned}(\beta_1, \ldots, \beta_m) = (&\frac{\min(p,\nu)}{\nu}, \frac{\min_+(p-\nu,h)}{h}, \frac{\min_+(p-h-\nu,h)}{h}, \ldots,\\ &\frac{\min_+(p-(m-2)h-\nu,h)}{h}).\end{aligned}$$

Note that this is regression is equivalent to linear interpolation between the stored points.

Theorem 7.2. *If $r = \frac{p}{h}$, the normalized expected square error is given by*

(a) $m = 1$:

$$SSE = \begin{cases} 1 - \frac{1}{2}r + r\log(r) & \text{if} \quad r \le 1, \\ 1 - \frac{1}{2r} & \text{if} \quad r > 1 \end{cases}$$

(b) $m \ge 2$:

$$SSE = \begin{cases} 1 - \frac{1}{2}r - \frac{1}{3}r^2 + r\log(r) & \text{if} \quad r \le 1, \\ \frac{1}{6r} & \text{if} \quad 1 < r \le m-1, \\ \frac{1+2(r-m+1)^3}{6r} & \text{if} \quad m-1 < r \le m, \\ 1 - \frac{2m-1}{2r} & \text{if} \quad r > m \end{cases}.$$

Proof. One difference to the analog proof for EWMAs is that the square error depends on ν, i.e., on the time elapsed since the last value was stored. To compute the expected square error, we average the square error for $\nu = 1, \ldots, h$. Using $SSE = \mathrm{Var}(s_n - s_{n-p}) - \mathrm{Cov}(Y,X)\mathrm{Var}^{-1}(X)\mathrm{Cov}(Y,X)^T$, for $r < 1$ the expected square error is given by

$$SSE = p + \frac{1}{h}\sum_{\nu=1}^{p}\left(-\nu - \frac{(p-\nu)^2}{h}\right) + \frac{1}{h}\sum_{\nu=p+1}^{h} -\frac{p^2}{\nu}.$$

Since $\sum_{\nu=p+1}^{h} 1/\nu$ is asymptotic with $\log(\frac{h}{p})$,

$$SSE \sim p - \frac{p(p+1)}{2h} - \frac{(p-1)p(2p-1)}{6h^2} - \frac{p^2}{h}\log(\frac{h}{p}).$$

Dividing by p and letting $h \to \infty$ and hence $p \to \infty$ gives the stated result.

If $1 < r \leq m-1$ let k and l be such that $p = kh + l$. Then

$$SSE = p - \frac{1}{h}\sum_{\nu=1}^{h}(\nu + (k-1)h) - \frac{1}{h}\sum_{\nu=1}^{l}(h + \frac{(p-kh-\nu)^2}{h}) - \frac{1}{h}\sum_{\nu=l+1}^{h}\frac{(p-(k-1)h-\nu)^2}{h}.$$

If $m-1 < r \leq m$ let $l = p-(m-1)h$ and then

$$SSE = p - \frac{1}{h}\sum_{\nu=1}^{h}(\nu + (m-2)h) - \frac{1}{h}\sum_{\nu=1}^{l}h - \frac{1}{h}\sum_{\nu=l+1}^{h}\frac{(p-(m-2)h-\nu)^2}{h}.$$

If $r > m$

$$SSE = p - \frac{1}{h}\sum_{\nu=1}^{h}(-\nu - (m-1)h).$$

These equations simplify to the formulae given in the theorem. The case for $r < 1$ and $m = 1$ needs to be handled separately, since for large values of ν, $s_{n-\nu}$ is further in the past than the most distant (and only) stored value x_0. Note that the r^2 term is missing for $m = 1$ so that SS is continuous at $r = 1$. □

For $m = 1$, SSE is minimal for $e^{-0.5} = 0.6064$ and the minimal $SSE = 1 - exp^{-0.5} = 0.3925$. In comparison, the minimal SSE using one EWMA is $SSE = 0.1855$. Performing the computations of the EWMA (two multiplications and one addition) adds extra information about the TS compared to just storing a value. For larger m, SSE is minimal if r is slightly bigger than $m-1$ and the optimal SSE is asymptotic to $1/6(m-1)$. Intuitively, this appears reasonable. Assume the objective is to approximate a value 80 steps in the past and one is allowed to store 5 values at a time. It is optimal to store a new value every 19 or 20 steps since this produces the densest packing of stored values such that s_{n-80} lies either between two stored values or just beyond the oldest stored value. Denser spacing of values results in better linear interpolations. Storing historical values becomes more accurate for larger m and outperforms using EWMAs. However, there is a trade-off between computational complexity and having information going further back in time. This warrants further investigation.

8 Further Research

This paper gives a framework for modeling time series using EWMAs. The problem can be expressed in terms of linear algebra by realizing that each EWMA update is a linear combination of the current value of the TS and m stored numbers. This can be written as

$$e_{n+1} = \alpha s_{n+1} + Ae_n. \tag{17}$$

Now $e_n = (e_n^{(1)}, \ldots, e_n^{(m)})^T$ and $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ are m column vectors and $A = (a_{ij})$ is an $m \times m$ matrix. In the case of one EWMA, $\alpha_1 = 1 - w$ and $a_{11} = w$; the case of two iterated EWMAs can be written as

$$\begin{pmatrix} e_{n+1}^{(1)} \\ e_{n+1}^{(2)} \end{pmatrix} = \begin{pmatrix} 1-w \\ (1-w)^2 \end{pmatrix} s_{n+1} + \begin{pmatrix} w & 0 \\ w(1-w) & w \end{pmatrix} \begin{pmatrix} e_n^{(1)} \\ e_n^{(2)} \end{pmatrix}.$$

Note that $\alpha_i + \sum_j a_{ij} = 1$ for $i = 1, 2$. This motivates the definition:

The system (17) is called a moving average if all rows of coefficients add up to 1, i.e., $\alpha_i + \sum_j a_{ij} = 1$ for $i = 1, \ldots, m$.

The equivalent formulation of Proposition 2.4 becomes (again $s_0 = e_0 = 0$):

$$e_n = \sum_{i=1}^{n} A^{n-i} \alpha s_i,$$

or

$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} s_n - e_n = \sum_{i=0}^{n-1} \left[\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} - \sum_{j=0}^{i} A^j \alpha \right] \xi_{n-i}.$$

The inner sum $\sum A^j$ converges iff all eigenvalues of A are strictly less than 1 in magnitude. In this case, $I - A$ is invertible and the expression above equals

$$\begin{aligned} &\sum_{i=0}^{n-1} \left[\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} - (I - A^{i+1})(I - A)^{-1} \alpha \right] \xi_{n-i} \\ &= \left[\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} - (I - A)^{-1} \alpha \right] s_n + (I - A)^{-1} \sum_{i=0}^{n-1} A^{i+1} \alpha \xi_{n-i}. \end{aligned}$$

Hence, $E(s_n - e_n^{(k)})$ converges for all k iff the expression in square brackets equals zero. This is equivalent to the system forming a moving average. More importantly, regressing $s_n - s_{n-p}$ on $s_n - e_n^{(k)}$ yields zero regression coefficients unless the system forms a moving average. In other words, only moving average systems contain information about the past. The problem has been reduced to investigating the sum of the second term using linear algebra techniques. It is not known whether this approach produces simpler proofs of the theorems of this paper. Any results using this setting would solve a much larger class of problems.

A Binomial Identities

A few binomial identities used in this paper are given in this appendix. Most of these can be found in books on the subject such as [Graham et al. (1994)]. Proves are given for the less well known (or new) identities.

The following is a special case of Vandermonde's convolution (5.22) in [Graham et al. (1994)].

Proposition A.1. *Let $k, l \in \mathbb{N}_0$. Then*

$$\sum_{i=0}^{k} \binom{k}{i} \binom{l}{i} = \binom{k+l}{k}.$$

Proposition A.2. *Let $k \in \mathbb{N}_0$, $p \in \mathbb{N}$ and $w \in \mathbb{C}$. Then*

$$\begin{aligned}\sum_{i=0}^{p-1}\binom{k+i}{k}w^i &= (1-w)^{-(k+1)}\left(1+\frac{1}{k!}\sum_{i=0}^{k}(-1)^{i+1}\binom{k}{i}\left[\prod_{j=0,j\neq i}^{k}(p+j)\right]w^{p+i}\right),\\ &= (1-w)^{-(k+1)}\left(1-\sum_{i=0}^{k}\binom{p+k}{i}(1-w)^i w^{p+k-i}\right).\end{aligned}$$

Proof. The first equality can be shown by induction and the second equality follows from (2) and (4) in [Wilf (2004)]. □

Proposition A.3.

$$\sum_{i=0}^{\infty}\binom{k+i}{k}w^i = \frac{1}{(1-w)^{k+1}}$$

Proposition A.4. *For $0 \le \nu \le k$*

$$\frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}\left[\prod_{j=0,j\neq i}^{k}(p+j)\right]i^\nu = (-p)^\nu.$$

Proof. Consider both sides as polynomials in p, of degree at most k (hence the proposition cannot be true for $\nu > k$). Substitute $p = 0, -1, -2, \ldots, -k$. We will show that the LHS and RHS agree on these $k+1$ values and hence the polynomials are identical. Note that for these values of p, only the $i = -p$ term in the sum is non-zero. Hence the LHS equals

$$\begin{aligned}&\frac{1}{k!}(-1)^{-p}\frac{k(k-1)\cdots(k+p+1)}{(-p)!}p(p+1)\cdots(-2)(-1)\cdot 1\cdot 2\cdots(p+k)(-p)^\nu\\ &= \frac{1}{k!}(-1)^{-p}\frac{k!}{(-p)!}(-1)^{-p}(-p)!(-p)^\nu = (-p)^\nu.\end{aligned}$$

□

Proposition A.5.

$$\sum_{i=0}^{n-1}\binom{k+i}{k}\binom{l+i}{l}w^i = \frac{1}{(1-w)^{k+l+1}}\left[\sum_{i=0}^{k}\binom{k}{i}\binom{l}{i}w^i + P_{kl}(w,n)w^n\right],$$

where $P_{kl}(w,n)$ is a polynomial of degree $k+l$ with coefficients that depend on n.

Proof. By induction on l. Note that the proposition is true $\forall n, k$ if $l = 0$ by Proposition A.2. □

Proposition A.6. *$a, b, c, d \in \mathbb{Z}$, $a \ge 0$*

$$\sum_{\mu\in\mathbb{Z}}\binom{a}{b+\mu}\binom{c+\mu}{d}(-1)^\mu = (-1)^{a+b}\binom{c-b}{d-a}.$$

Proof. (5.24) in [Graham et al. (1994)]. □

References

Brown, R. G., Meyer, R. F., and D'Esopo, D. A. (1961). The fundamental theorem of exponential smoothing. *Operations Research*, 9(5):673–687.

Cacorogna, M., Gençay, R., Muller, U. A., and Pictet, O. (2001). *An Introduction to High Frequency Finance*. Academic Press.

Chen, F., Lambert, D., and Pinheiro, J. C. (2001). Updating timing profiles for millions of customers in real-time. *Journal of the American Statistical Association*.

Chen, F., Lambert, D., Pinheiro, J. C., and Sun, D. X. (2000). Reducing transaction databases, without lagging behind the data or losing information. *Proceedings of KDD*.

Graham, R. L., Knuth, D. E., and Patashnik, O. (1994). *Concrete Mathematics, A Foundation for Computer Science*. Addison-Wesley Professional, second edition.

Hilpisch, Y. (2020). *Python for Algorithmic Trading*. O'Reilly.

Ladde, G. and Wu, L. (2009). Development of modified geometric brownian motion models by using stock price data and basic statistics. *Nonlinear Analysis*, 71(12):e1203–e1208.

Murphy, J. J. (1999). *Technical Analysis of the Financial Markets*. New York Institute of Finance.

Müller, U. A. and Zumbach, G. (2001). Operators on inhomogeneous time series. *International Journal of Theoretical and Applied Finance*, 4(01):147–177.

Qiu, P. (2014). *Introduction to Statistical Process Control*. Chapman and Hall/CRC.

Wilf, H. S. (2004). The method of characteristics, and 'problem 89' of graham, knuth and patashnik. *arXiv:math/0406620*.