Statistical Process Control

MSc: Statistics and Actuarial-Financial Mathematics

Konstantinos Bourazas
kbourazas@aegean.gr

Department of Statistics and Actuarial-Financial Mathematics

Course #8

April 2, 2026

EWMA charts for detecting persistent shifts

- The CUSUM chart accumulates deviations and uses a **restart mechanism** to discard old IC evidence
- An alternative idea, proposed by Roberts[1], is to compute a **weighted average** of all past observations where the weights decay exponentially
- The resulting chart is called the **Exponentially Weighted Moving Average (EWMA)** chart
  - It is easier to understand and implement than the CUSUM
  - Its control limits have a familiar Shewhart-like form
  - Its performance for detecting small, persistent shifts is very similar to that of the CUSUM
- Like the CUSUM, the EWMA is primarily a **Phase II** tool and works best with individual observation data

[1] Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1, 239–250.

The EWMA charting statistic

- Let $\{X_1, X_2, \ldots\}$ be independent observations from a process with IC distribution $N(\mu_0, \sigma^2)$
- The EWMA statistic is defined recursively by

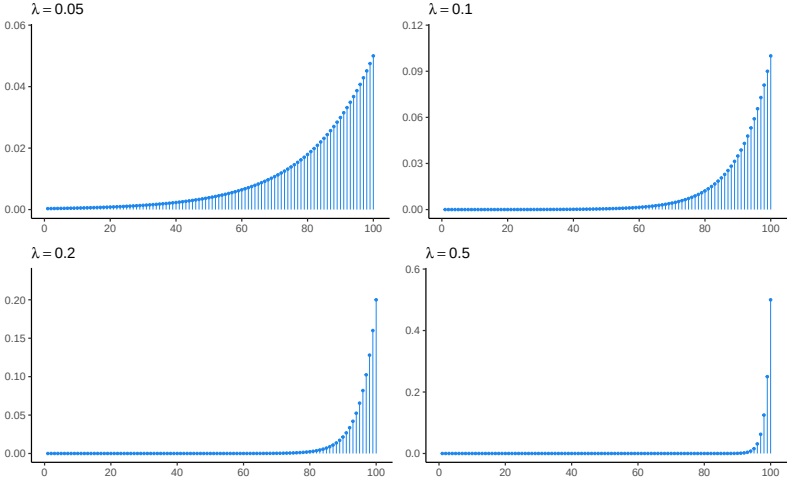
$$E_n = \lambda X_n + (1 - \lambda) E_{n-1}, \qquad E_0 = \mu_0$$

  where $\lambda \in (0, 1]$ is the **weighting parameter**
- Expanding the recursion gives

$$E_n = \lambda \sum_{i=1}^{n} (1 - \lambda)^{n-i} X_i + (1 - \lambda)^n \mu_0$$

  so E_n is a weighted average of μ_0 and all observations up to time n, with weights that decay **exponentially** as observations get older
- When $\lambda = 1$ the statistic reduces to $E_n = X_n$, and the EWMA chart becomes a Shewhart chart

How much history does the EWMA use?



Weights $w_i = \lambda(1-\lambda)^{n-i}$ for $n = 100$ and $\lambda = 0.05, 0.1, 0.2, 0.5$. With $\lambda = 0.05$ about 40 past observations receive meaningful weight; with $\lambda = 0.5$ only 3 or 4 do.

Choosing the weighting parameter λ

- The parameter λ plays a role similar to the allowance k in CUSUM charts: it determines which shift size the chart is best at detecting
- **Small** λ (e.g. 0.05) gives substantial weight to many past observations, making the chart sensitive to **small shifts** but slow for large ones
- **Large** λ (e.g. 0.2 or above) concentrates weight on recent observations, making the chart better for **large shifts** but less effective for small ones
- Unlike the CUSUM, there is no closed-form formula linking λ to a target shift δ. In practice, one can search for the λ that minimises ARL_1 for a given shift, but commonly used defaults are $\lambda \in \{0.05, 0.1, 0.2\}$
- In the extreme case $\lambda = 1$, the statistic becomes $E_n = X_n$ and the EWMA chart reduces to a **Shewhart chart**, confirming that it is designed for large shifts only

Properties of E_n under IC

- When the process is IC up to time n, the statistic has

$$\mu_{E_n} = \mu_0, \qquad \sigma^2_{E_n} = \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right]\sigma^2$$

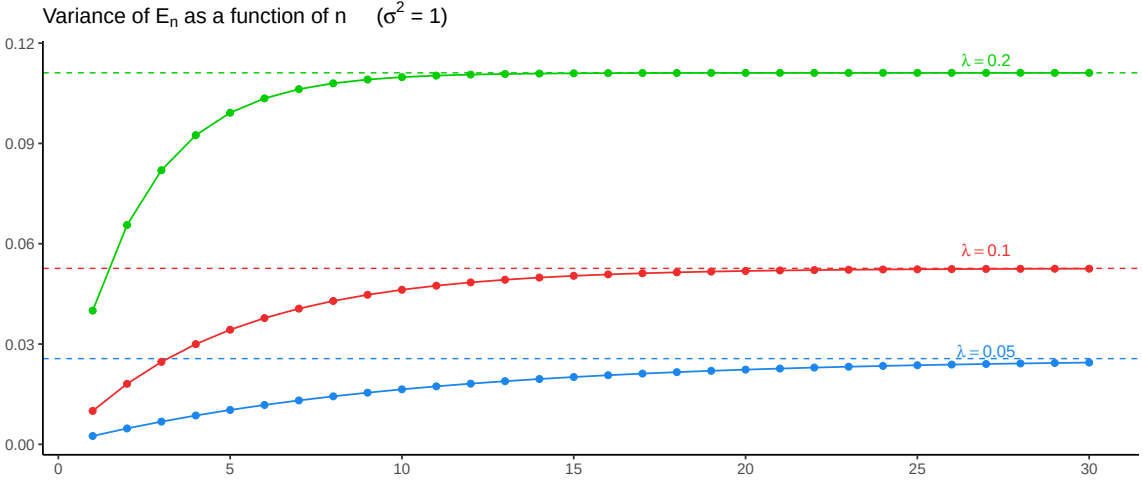
- As $n \to \infty$, the variance converges to the **asymptotic** value

$$\tilde{\sigma}^2_{0,\lambda} = \frac{\lambda}{2-\lambda}\,\sigma^2$$

  The convergence is fast, especially for larger λ. For $\lambda = 0.2$ the variance is essentially at its limit by $n = 10$; for $\lambda = 0.05$ by about $n = 20$

- This is the key advantage over the plain cumulative sum C_n, whose variance grows without bound. The EWMA has a **stable variance**, so fixed control limits make sense

Convergence of $\sigma^2_{E_n}$ for different λ



Behaviour of E_n after a mean shift

- Suppose the process mean shifts from μ_0 to μ_1 at time τ. Then for $n \geq \tau$ the mean of E_n becomes

$$\mu_{E_n,\tau} = \mu_0 + \left[1 - (1-\lambda)^{n-\tau+1}\right](\mu_1 - \mu_0)$$

- This is a weighted average of μ_0 and μ_1, with the weight on μ_1 growing towards 1 as n increases past τ
- Crucially, the **variance of E_n is unchanged** by the mean shift, so the shift shows up as a drift in the mean of the statistic while its spread stays the same
- Both facts together confirm that E_n carries useful information about the shift and can be compared to **fixed control limits**

Control limits of the EWMA chart

- The **exact** (time-varying) control limits are

$$U = \mu_0 + \rho \sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right]}\,\sigma, \qquad L = \mu_0 - \rho \sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right]}\,\sigma$$

  where $\rho > 0$ is a parameter that controls the ARL_0

- The **asymptotic** (constant) control limits replace the time-varying factor by its limit

$$U = \mu_0 + \rho \sqrt{\frac{\lambda}{2-\lambda}}\,\sigma, \qquad L = \mu_0 - \rho \sqrt{\frac{\lambda}{2-\lambda}}\,\sigma$$

- Because $\sigma^2_{E_n}$ converges quickly, the two versions give very similar results unless a shift occurs in the first few observations. Most software uses the asymptotic form
- A signal is given whenever $E_n > U$ or $E_n < L$

Example 5.1 The EWMA chart in action

- IC distribution $N(10, 2^2)$.
- EWMA chart with $\lambda = 0.1$ and $\rho = 2.703$, giving $\text{ARL}_0 = 200$ under the exact limits
- The chart signals at the **21st time point**, detecting an upward mean shift that a Shewhart chart would likely miss

Example 5.1 EWMA chart with $\lambda = 0.1$

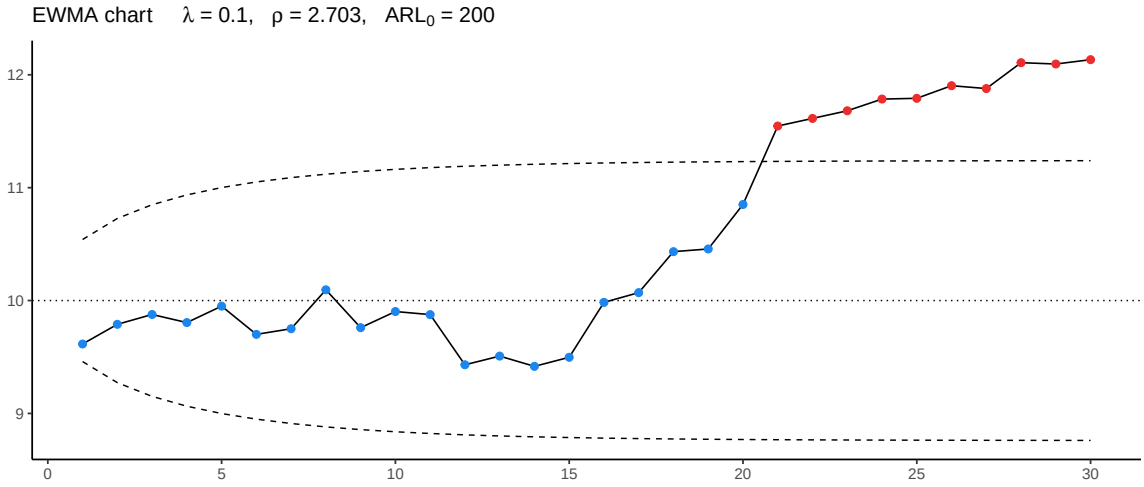
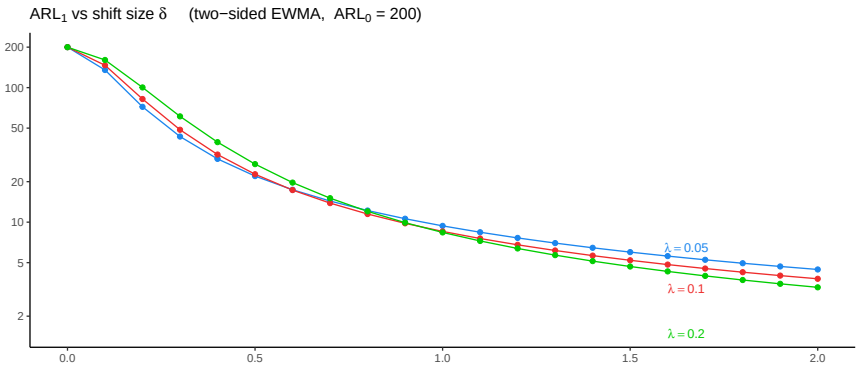


Table of ρ values for common ARL_0 and λ (Qiu, Table 5.1)

	λ						
ARL_0	0.01	0.05	0.1	0.2	0.3	0.5	0.75
50	0.845	1.520	1.811	2.054	2.166	2.268	2.315
100	1.152	1.879	2.148	2.360	2.453	2.534	2.568
200	1.500	2.216	2.454	2.635	2.713	2.777	2.802
370	1.819	2.490	2.701	2.859	2.925	2.978	2.996
500	1.973	2.615	2.814	2.962	3.023	3.071	3.087
1000	2.308	2.884	3.059	3.187	3.238	3.277	3.289

- As expected, ρ increases with both ARL_0 and λ
- The values can also be obtained using the function `xewma.crit()` from the R package `spc`, or by Monte Carlo simulation

ARL_1 for different values of λ and shift size δ

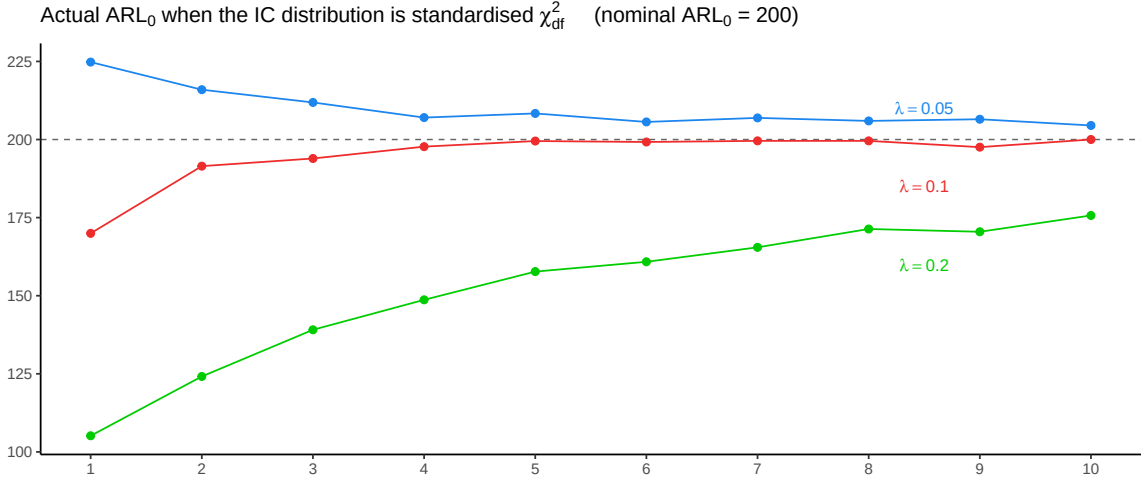


IC distribution $N(0,1)$, $\mathrm{ARL}_0 = 200$. Smaller λ wins for small shifts; larger λ wins for large shifts.

Robustness to non-normality (Example 5.2)

- Suppose the IC distribution is not $N(0,1)$ but the standardised χ^2_{df} distribution (mean 0, variance 1)
- Because E_n averages many past observations, the CLT suggests its distribution will be closer to normal when λ is small
- Numerical results confirm this: the actual ARL_0 stays near the nominal value of 200 for small λ, but can deviate substantially when λ is large and the underlying distribution is highly skewed
- However, the robustness comes at a cost. Choosing a very small λ means the chart is ineffective for detecting shifts larger than about 1σ
- When the IC distribution is far from normal, Qiu recommends using nonparametric charts (Chapter 8) rather than relying on the EWMA's approximate robustness

Example 5.2 Actual ARL_0 under non-normality



Effect of autocorrelation (Example 5.3)

- As with CUSUM charts, autocorrelation can severely distort the EWMA's performance
- With $(\lambda, \rho) = (0.1, 2.454)$ (nominal $\text{ARL}_0 = 200$) and an AR(1) model $X_n = \mu_0 + \phi(X_{n-1} - \mu_0) + e_n$

ϕ	-0.5	-0.25	0	0.25	0.5
Actual ARL_0	876	582	202	69	28

- Positive autocorrelation **inflates** the false alarm rate; negative autocorrelation makes the chart too conservative
- The standard remedy is the same **residual approach** used for CUSUM charts: fit a time series model to the IC data, compute residuals, and apply the EWMA to those
- Zhang (1998) also proposed adjusting the EWMA control limits directly to account for the correlation structure

Comparison with CUSUM charts

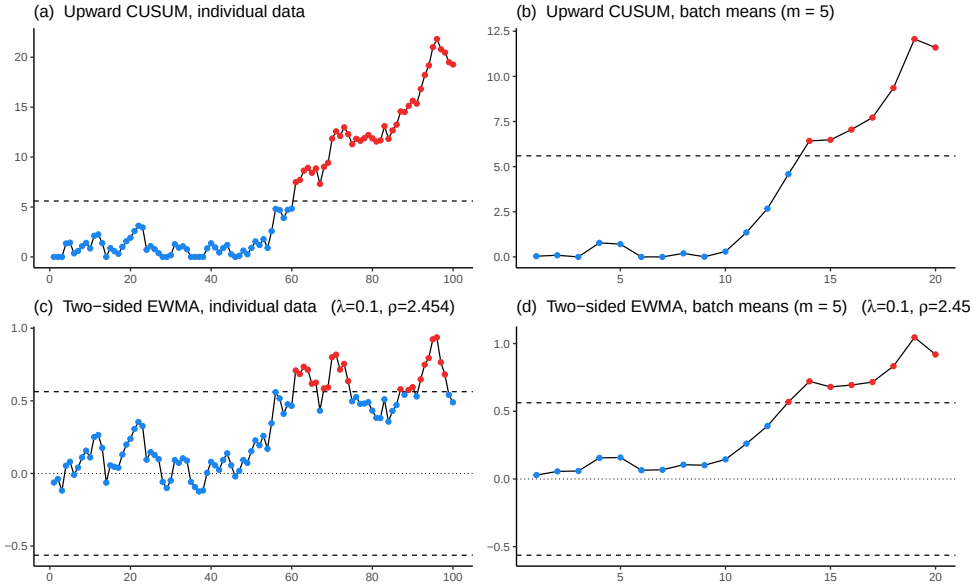
- Both CUSUM and EWMA are designed for small, persistent shifts and both are primarily Phase II tools
- Both prefer **individual observation data**, because grouping observations into batches delays detection
- In terms of $\mathbf{ARL_1}$, the two charts have very similar optimal performance across the full range of shift sizes (Qiu, Table 5.3). Neither consistently dominates the other
- The CUSUM has the advantage of **proven optimality** (Moustakides, 1986). No corresponding optimality result exists for the EWMA
- The EWMA has the advantage of **simplicity**. Its control limits look like those of a Shewhart chart, and the exponentially decaying weights are intuitive
- In practice, the choice between them is often a matter of convenience. Many practitioners run **both** charts together

Optimal ARL_1 of upward EWMA and CUSUM (Qiu, Table 5.3)

	EWMA		CUSUM	
δ	λ^*	ARL_1	k^*	ARL_1
0.2	0.020	54.3	0.10	54.1
0.5	0.069	19.7	0.25	19.3
0.8	0.135	10.5	0.40	10.2
1.0	0.185	7.7	0.50	7.4
1.5	0.327	4.3	0.75	4.0
2.0	0.496	2.8	1.00	2.6

- IC distribution $N(0, 1)$, $\text{ARL}_0 = 200$ in all cases
- The CUSUM is slightly faster, but the difference is small

Example 5.4 Individual vs. batch data



EWMA charts for monitoring the process variance

- The mean-monitoring EWMA has **some** ability to detect upward variance shifts, but cannot detect downward ones and is not efficient for that purpose
- For individual observations, a dedicated variance EWMA is based on the squared standardised observations $Y_n = [(X_n - \mu_0)/\sigma_0]^2$

$$E_n = \lambda\, Y_n + (1-\lambda)\, E_{n-1}, \qquad E_0 = 1$$

- When the process is IC, $Y_n \sim \chi^2_1$ so $\mu_{E_n} = 1$ and $\sigma^2_{E_n} = \frac{2\lambda}{2-\lambda}[1-(1-\lambda)^{2n}]$
- A signal of **upward** variance shift is given when

$$E_n > U = 1 + \rho_U \sqrt{\frac{2\lambda}{2-\lambda}\left[1-(1-\lambda)^{2n}\right]}$$

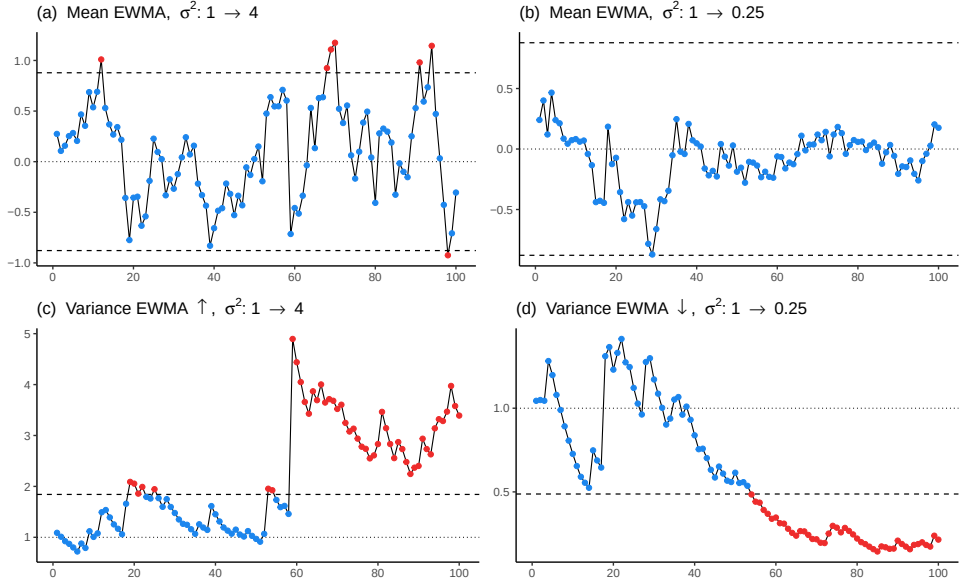
- The downward version signals when $E_n < L$ with an analogous lower limit. The parameters ρ_U and ρ_L differ because the IC distribution of E_n is skewed

Table of ρ_U and ρ_L for variance charts (Qiu, Table 5.4)

	ρ_U				ρ_L			
ARL_0	λ=0.05	0.1	0.2	0.3	λ=0.05	0.1	0.2	0.3
50	0.901	1.380	1.916	2.259	0.865	1.100	1.195	1.166
100	1.455	1.988	2.606	2.996	1.201	1.366	1.360	1.273
200	2.017	2.595	3.258	3.702	1.510	1.580	1.480	1.349
370	2.518	3.133	3.854	4.354	1.746	1.731	1.563	1.401
500	2.796	3.419	4.184	4.722	1.862	1.808	1.605	1.426

- The values of ρ_U and ρ_L are quite different because the IC distribution of E_n is right-skewed
- ρ_U increases with both ARL_0 and λ; ρ_L shows a non-monotone pattern in λ

Example 5.5 Detecting variance shifts



Variance EWMA with batch data

- When a sample of size m is collected at each time point, the natural charting statistic uses the sample variance

$$E_n = \lambda \frac{s_n^2}{\sigma_0^2} + (1-\lambda)\, E_{n-1}, \qquad E_0 = 1$$

- Under IC, $\mu_{E_n} = 1$ and $\sigma^2_{E_n} = \frac{2\lambda}{(2-\lambda)(m-1)}[1-(1-\lambda)^{2n}]$
- The control limits keep the same form, with an extra factor of $1/(m-1)$ inside the square root
- An important advantage of batch-based variance charts is that they are **not affected by a simultaneous mean shift**, because $\bar{X}_n$ and s_n^2 are independent for normal data