EWMA Weighted Linear Ridge Regression

By David Romoff

Time series forecasting requires balancing adaptability to new data with robustness against overfitting. Traditional methods such as ARIMAX / GARCH or penalized regression address this challenge. This paper presents a different quick first-principles approach that integrates Exponentially Weighted Moving Average (EWMA), Weighted Linear Regression, and Ridge Regression (L2 regularization) into a single closed-form solution.

By applying exponential decay to older observations and shrinking coefficients to manage multicollinearity, this method provides a fast and intuitive tool for short-term forecasting. While it does not capture every structural aspect (autocorrelation or volatility) like ARIMAX / GARCH does, it provides a quick, efficient way to address nonstationarity and overfitting.

Below, each technique is introduced along with matrix-algebra derivations of the resulting regression coefficients. The combined closed-form formula for *EWMA-weighted ridge regression* is then presented, followed by a discussion of fitting approaches and time-series cross-validation.

Weighted Linear Regression

Weighted regression is a variant of ordinary linear regression in which each observation has a user-specified weight. This approach is beneficial in contexts such as time series or heteroskedastic data, where some observations may be more important or more reliable than others.

Definitions of the Variables

- β: The regression coefficients.
- n: The number of observations (rows).
- p: The number of predictors (columns) in the design matrix.
- $X \in \mathbb{R}^{n \times p}$: The design matrix of predictor variables, where each row $x_i^\top$ represents observation i and each column represents a particular predictor.
- $y \in \mathbb{R}^n$: The response vector, where y_i is the response for observation i.
- $w_i \geq 0$: The weight assigned to observation i, reflecting its relative importance.

- $W = \mathrm{diag}(w_1, w_2, \ldots, w_n)$: A diagonal matrix whose main diagonal entries are the weights. The notation $\mathrm{diag}(\cdot)$ refers to the operator that places the given vector elements on the main diagonal of a square matrix, and zeroes elsewhere.

Mathematical Intuition and Derivation

Weighted linear regression (WLR) modifies the usual sum of squared residuals by incorporating weights. Specifically, it minimizes

$$S(\beta) \;=\; (y - X\beta)^{\mathsf{T}}\, W\, (y - X\beta).$$

Expanding this expression gives

$$S(\beta) \;=\; y^{\mathsf{T}} W\, y \;-\; 2\,\beta^{\mathsf{T}}\,(X^{\mathsf{T}} W\, y) \;+\; \beta^{\mathsf{T}}\,(X^{\mathsf{T}} W\, X)\,\beta.$$

Taking the gradient with respect to β and setting it to zero leads to

$$\nabla_{\beta} S(\beta) \;=\; -2\, X^{\mathsf{T}} W\, y \;+\; 2\, X^{\mathsf{T}} W\, X\, \beta \;=\; 0,$$

which implies

$$X^{\mathsf{T}} W\, X\, \beta \;=\; X^{\mathsf{T}} W\, y \quad\Rightarrow\quad \hat{\beta}_{\mathrm{WLR}} \;=\; (X^{\mathsf{T}} W\, X)^{-1}\,(X^{\mathsf{T}} W\, y).$$

Larger weights w_i cause the corresponding observations (x_i, y_i) to have a stronger influence on the estimated parameters.

EWMA

EWMA is an approach for weighting observations in a time series so that more recent data carry greater influence than older data. This is especially useful in scenarios where the underlying process evolves over time.

EWMA Formulation and Interpretation

Consider a time series $\{x_t\}$. The EWMA s_t is defined by the recursion:

$$s_t \;=\; \alpha\, x_t \;+\; (1 - \alpha)\, s_{t-1}, \quad 0 < \alpha < 1.$$

Unrolling this recursion reveals a geometric decay in the influence of past values:

$$s_t \;=\; \alpha\, x_t \;+\; \alpha(1 - \alpha)\, x_{t-1} \;+\; \alpha(1 - \alpha)^2\, x_{t-2} \;+\; \cdots.$$

In many financial and statistical contexts, the parameter α is expressed as $1 - \lambda$, with λ denoting the decay factor. Since there is no requirement to normalize these weights for the purposes of linear regression, one can simply express them as a geometric series in terms of λ. The sequence can therefore be understood as a geometric series (with common ratio λ) that is not normalized by its sum yet still provides a valid weighting scheme for regression.

Ridge Regression

Ridge regression is a penalized form of linear regression that shrinks coefficients to mitigate collinearity and overfitting. It is especially useful when the design matrix X has closely correlated predictors or when the number of predictors (p) exceeds the number of observations (n).

Mathematical Intuition and Derivation (L2 Penalty)

Ordinary least squares determine β such that

$$\min_{\beta} \quad (y - X\beta)^{\mathsf{T}}(y - X\beta).$$

Ridge regression adds an $L2$ penalty $\lambda \parallel \beta \parallel_2^2$, where $\lambda \geq 0$, leading to

$$\min_{\beta} \quad (y - X\beta)^{\mathsf{T}}(y - X\beta) \; + \; \lambda \parallel \beta \parallel_2^2.$$

Writing

$$S_{\text{ridge}}(\beta) \; = \; (y - X\beta)^{\mathsf{T}}(y - X\beta) \; + \; \lambda\, \beta^{\mathsf{T}}\beta,$$

taking derivatives, and setting them to zero gives

$$(X^{\mathsf{T}}X + \lambda\, I)\, \beta \; = \; X^{\mathsf{T}}y \quad \Rightarrow \quad \hat{\beta}_{\text{ridge}} \; = \; (X^{\mathsf{T}}X + \lambda I)^{-1}\, X^{\mathsf{T}}y.$$

The scalar λ scales the strength of the penalty relative to the residual sum of squares. The optimal choice of λ is typically determined via an out-of-sample or cross-validation procedure.

L1 vs. L2, and Standardization

Penalty functions can be of two primary types:

- **L1 (Lasso):** $\parallel \beta \parallel_1 = \sum_j \left|\beta_j\right|$. This penalty may drive some coefficients to zero (sparsity) but does not have a closed-form solution.
- **L2 (Ridge):** $\parallel \beta \parallel_2^2 = \beta^{\mathsf{T}}\beta$. This penalty shrinks all coefficients continuously toward zero (though not exactly to zero) and does have a closed-form solution.

Standardizing each predictor is generally recommended, so that the penalty applies uniformly across coefficients. This ensures that predictors on vastly different scales are not penalized disproportionately.

EWMA-Weighted Ridge Regression

This section combines weighted linear regression and ridge regression, letting the weights themselves follow an EWMA decay pattern over time. Suppose there are n observations

indexed by time, and each observation t has an EWMA-based weight $w_t(\lambda_{\text{EWMA}})$. Collect these into

$$W_{\text{EWMA}} \;=\; \text{diag}(w_1, \ldots, w_n).$$

The objective for EWMA-weighted ridge regression is

$$\min_{\beta} \quad (y - X\beta)^{\top} W_{\text{EWMA}} (y - X\beta) \;+\; \lambda_{\text{ridge}} \parallel \beta \parallel_2^2.$$

Taking the gradient with respect to β and setting it to zero yields

$$\left(X^{\top} W_{\text{EWMA}} X + \lambda_{\text{ridge}}\, I\right) \beta \;=\; X^{\top} W_{\text{EWMA}}\, y,$$

so the closed-form estimator becomes

$$\hat{\beta} \;=\; (X^{\top} W_{\text{EWMA}} X + \lambda_{\text{ridge}}\, I)^{-1} (X^{\top} W_{\text{EWMA}}\, y).$$

Hence, each data point is downweighted exponentially by age, while the ridge penalty shrinks coefficients to reduce variance and collinearity issues.

Fitting Approach and Cross-Validation in Time Series

Two hyperparameters must be tuned:

- λ_{EWMA} (the decay factor for exponential weighting),
- λ_{ridge} (the L2 penalty coefficient).

These are chosen via an out-of-sample approach. In time series, a rolling or forward-chaining cross-validation strategy is more appropriate than random splitting, since it respects the chronological order of observations:

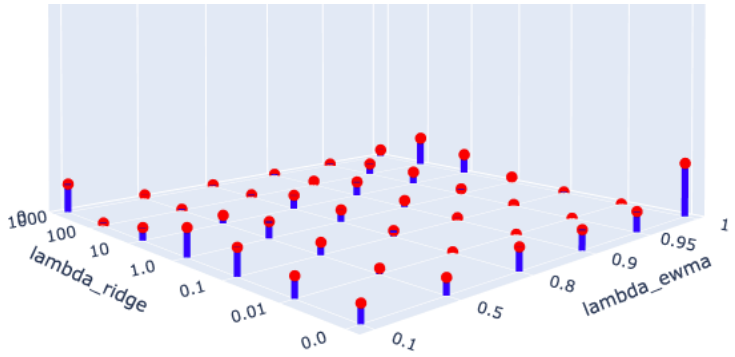
- Train on an initial time block,
- Validate on the subsequent time block,
- Test on a further block,
- Move (roll) the window forward and repeat.

Testing

The proposed approach is tested by randomly drawing four stocks from the S&P 500 and regressing the returns of three of them onto the returns of the fourth. This is repeated 1000 times. On each repetition, the rolling cross validation strategy steps through time blocks. For each time block, the combinations of λ_{EWMA} and λ_{ridge} are looped through and the mean residual sum of squares (MSE) are recorded in the validation set which is five days in

the future relative to the training data. The $(\lambda_{\text{EWMA}},\ \lambda_{\text{ridge}})$ combination with the lowest MSE is recorded.

The image below displays the fraction of the time that the $(\lambda_{\text{EWMA}},\ \lambda_{\text{ridge}})$ combination performed best out of sample. The right corner combination (1,0) represents runs with effectively no regularization and no decay. That combination performs best about 1/10th of the time.



This sample distribution suggests that 9/10ths of the time, some combination of decay weighting and regularization outperform standard OLS.

Another prevalent combination is the left corner combination (1000, 0.1) where high regularization enables high decay (small λ_{EWMA}) forecasting. That combination performs best about 1/20th of the time.

Lastly, a diagonal ripple in the implied surface of the image suggests that a higher regularization penalty facilitates forecasting with higher decay.

Conclusion and Further Study

The outperformance of non (1,0) combinations of λ_{EWMA} and λ_{ridge} is observable. The benefit of regularization may be more pronounced when more securities are used to forecast. Readers can perform their own investigation of this approach. Python code is available at this link .

David Romoff, MBA, MS, is a Lecturer in the Enterprise Risk Management program at Columbia University. He can be reached at djr2132@columbia.edu