On the Volatility Prediction of the HAR-RV Model

Wat Street

November 7, 2024

1 Problem Formulation

The following equation follows from the derivation of the HAR-RV model

$$RV_{t+1} = C + \beta_d RV_t + \beta_w RVW_t + \beta_m RVM_t + \overline{\omega}_t,$$

where $\beta_d, \beta_w, \beta_m$ are the learned parameters, $\overline{\omega}$ represents the errors, C is a constant, and we have

$$RVW_t = \frac{1}{5}\left(\sum_{i=1}^{5} RV_{t-i+1}\right)$$

and

$$RVM_t = \frac{1}{22}\left(\sum_{i=1}^{22} RV_{t-i+1}\right).$$

which represent the weekly and monthly volatilities in terms of the daily volatility.

An interesting thing to note here is that the traditional HAR-RV model has a "symmetry problem" - if you swap two volatility values some RV_i and RV_j, $i \neq j$, the model's predictions don't change, even though in reality, the ordering of volatility events should matter in financial markets.

To address this symmetry issue, we propose introducing additional terms that ensure any pair-wise swap of volatility values affects the model's predictions. Our goal is to find a minimal set of additional terms, constructed from averages or sums, that captures these temporal dependencies while keeping the model parsimonious. The solution should generalize beyond the standard weekly and monthly intervals to arbitrary time periods, without compromising the model's predictive accuracy. Mathematically, we can formulate this as an optimization problem: finding the smallest collection of sets whose elements correspond to time indices, such that these sets can distinguish between any

two different time points through their averages.

Since we are limiting our scope to averages, we may formulate the problem as follows: We would like to find a set of sets $\{a_j\}_{j=1}^{c}$ and for all $j \in \{1, 2, \cdots, c\}$ we denote the set a_j by $\{a_{j_i}\}_{i=1}^{k}$, we have the extra term

$$\frac{1}{k}\sum_{i=1}^{k} RV_{t-a_{j_i}+1}.$$

We would like to minimize c and find $\{a_j\}$ such that for all pairs $(x, y) \in \{1, 2, \cdots, n\}, x \neq y$, where n is the limit of the terms, there exists some set a_j such that x is in a_j but y is not in a_j. We propose the following three solutions:

1. $a_j = \{1, 2, \cdots, j\}$ for all j.
2. Using the binary representation theorem for Hamming codes to formulate a solution to the problem.
3. Find the least positive integer greater than n, denoting it by N, such that N is a product of distinct primes. Then, let

$$\{a_j\} = \bigcup_{p|N,\ p \text{ prime}} \{[a]_p : 0 \leq a \leq p-1\}.$$

2 Exhaustive Search

The first solution is to let $a_j = \{1, 2, \cdots, j\}$ for all j. In this case, we stay true to the HAR-RV model by considering consecutive windows starting from the current time. Hence, this is likely to stay the most accurate to the HAR-RV model while also solving the problem of the volatility model staying the same when two terms are swapped. Note that $c = n$ in this case.

To prove that this indeed satisfies our desired conditions, consider any arbitrary pair $(x, y) \in \{1, 2, \cdots, n\}, x \neq y$. Without loss of generality, let $x < y$. Then, a_x contains x but not y, as desired.

Let's understand what this means in practice. Let's say we're considering $n = 5$ time periods. We would create sets like this:

$$\begin{aligned} a_1 &= \{1\} \\ a_2 &= \{1, 2\} \\ a_3 &= \{1, 2, 3\} \\ a_4 &= \{1, 2, 3, 4\} \\ a_5 &= \{1, 2, 3, 4, 5\} \end{aligned}$$

For each of these sets, we would create a new term in our model that averages the realized volatilities corresponding to the indices in that set. So our modified HAR-RV model would look like:

$$RV_{t+1} = C + \beta_d RV_t + \beta_w RV_t^W + \beta_m RV_t^M + \sum \left(\gamma_j \cdot \frac{1}{|a_j|} \sum_{i \in a_j} RV_{t-i+1} \right) + \omega_t \tag{1}$$

2.1 Concrete Example

Say we have historical RV values:

$$\begin{aligned} RV_t &= 0.2 \\ RV_{t-1} &= 0.3 \\ RV_{t-2} &= 0.4 \\ RV_{t-3} &= 0.5 \\ RV_{t-4} &= 0.6 \end{aligned}$$

Our new terms would calculate:

$$\begin{aligned} \text{For } a_1 &: \frac{0.3}{1} = 0.3 \\ \text{For } a_2 &: \frac{0.3 + 0.4}{2} = 0.35 \\ \text{For } a_3 &: \frac{0.3 + 0.4 + 0.5}{3} = 0.4 \\ \text{For } a_4 &: \frac{0.3 + 0.4 + 0.5 + 0.6}{4} = 0.45 \end{aligned}$$

Remember, the original issue was that swapping two RV values in the traditional model wouldn't change the predictions. Let's see how this solution fixes that:

If we swap RV_{t-1} and RV_{t-2} in our example:

$$\begin{aligned} \text{Original values} &: 0.3, 0.4 \\ \text{After swap} &: 0.4, 0.3 \end{aligned}$$

In the traditional HAR-RV model, the weekly and monthly averages would remain unchanged. However, in our new terms:

- a_1 would change from 0.3 to 0.4
- a_2 would maintain the same average but with different temporal information
- a_3, a_4, etc. would maintain their averages but with different temporal patterns

2.2 The Mathematical Structure

This solution creates a triangular structure of averages, where each new term adds one more historical value to the average. This creates a hierarchical system of temporal dependencies that can capture both short-term and long-term effects while maintaining sensitivity to the ordering of volatility events.

2.3 Prefacing Improvements

Some issues we've come across during implementation and testing are computational efficiency and parameter proliferation, multicollinearity, temporal weighting, scaling, sand noise. We've addressed some of the following with methods proposed below.

2.3.1 Decay Pattern

Replace the linear averaging with exponential weights. Instead of giving equal weight $(1/|a_j|)$ to each volatility in our sets, we could use exponentially decreasing weights. Our modified term would look like:

$$RV_{t+1} = C + \beta_d RV_t + \beta_w RV_t^W + \beta_m RV_t^M + \sum \left(\gamma_j \cdot \frac{\sum_{i \in a_j} \alpha^i RV_{t-i+1}}{\sum_{i \in a_j} \alpha^i} \right) + \omega_t \tag{2}$$

where α is a decay parameter between 0 and 1. This better aligns with how markets actually process and "forget" information.

2.3.2 Reducing Parameter Space

Implement a grouping strategy. Instead of having a separate γ_j for each set a_j, we could group our sets into meaningful time horizons. For example:

- Very short-term (1-3 days)
- Short-term (4-7 days)
- Medium-term (8-15 days)
- Long-term (16+ days)

Each group would share a single γ parameter, dramatically reducing the number of parameters while maintaining the ability to capture temporal ordering effects.

2.3.3 Incorporating Market Microstructure

We can add a filtering component to handle market microstructure noise. Before computing our averages, we could apply a noise-reduction technique like:

$$RV_{filtered} = RV_{raw} - k\sqrt{IV} \cdot \eta \tag{3}$$

where IV is the integrated variance and η is a noise term. This helps ensure our model captures true volatility patterns rather than market microstructure effects.

2.3.4 Adding Regime Switching

To capture changing market conditions, we could introduce a regime-switching component. Our modified model would look like:

$$RV_{t+1} = C(s_t) + \beta_d(s_t)RV_t + \beta_w(s_t)RV_t^W + \beta_m(s_t)RV_t^M + \sum(\gamma_j(s_t) \cdot Term_j) + \omega_t \tag{4}$$

where s_t represents the market regime (could be high/low volatility, or more complex states). This allows the model to adapt its behavior based on market conditions.

2.3.5 Regularization Framework

To improve statistical efficiency, we could add a regularization term to our objective function:

$$L = \|RV_{actual} - RV_{predicted}\|^2 + \lambda_1 \sum |\gamma_j| + \lambda_2 \sum(\gamma_j - \gamma_{j-1})^2 \tag{5}$$

The first penalty term (λ_1) encourages sparsity, while the second (λ_2) encourages smooth transitions between adjacent coefficients.

3 Hamming Codes

The second solution is related to the binary representation theorem for Hamming codes. In summary, the theorem says that the minimum c is $\lceil \log_2(n) \rceil$ and provides a construction that works: For all $\{1, 2, \cdots, n\}$, we let

$$a_j = \{x \in \{1, 2, \cdots, n\} : \text{the } j-\text{th digit of } x \text{ in binary is } 1\}.$$

To prove this construction works, if we let $(x, y) \in \{1, 2, \cdots, n\}, x \neq y$ be arbitrary, then since $x \neq y$, then x and y must differ by at least one digit in binary. Suppose that x and y differ by the $i-$th digit, and without loss of generality, let x have its $i-$th digit in binary be 0 and let y have its $i-$digit in binary be 1. Then, a_i contains y but not x, as desired. Furthermore, there are

$\lceil \log_2(n) \rceil$ such sets since there are $\lceil \log_2(n) \rceil$ digits in the binary representation of n.

This solution minimizes c but might not stay true to the intended behavior of the HAR-RV model because of the randomness in the construction of the sets a_j; they may not make sense in a financial sense.

4 Prime Modulo Classes

For this solution we give a description of the solution and omit a formal proof for minimality and simplicity. Given large $n \in \mathbb{Z}^+$, we wish to split $\{1, 2, \cdots, n\}$ into modulo classes

$$\bigcup_i \{[a]_{k_i} : 0 \leq a \leq k_i - 1\}$$

for possibly multiple values of k_i such that all k_i are pairwise coprime and $\prod_i k_i \geq n$. To prove that this works, let $(x, y) \in \{1, 2, \cdots, n\}$ be arbitrary with $x \neq y$. Assume for the sake of contradiction that for all sets that contain x, they also contain y. Then, $x \equiv y \pmod{k_i}$ for all k_i. Since all k_i are pairwise coprime, then by the Chinese Remainder Theorem, we get that $x \equiv y \pmod{\prod_i k_i}$, and letting $k := \prod_i k_i$, we get that $k|(x - y)$. Since $k \geq n$, and $|x - y| \leq n - 1 < k$, this implies that $|x - y| = 0$, so $x = y$, which is a contradiction. Therefore, this construction works. It remains to find $\{k_i\}$ that minimizes c, the number of extra terms we add.

We claim that for large $n \in \mathbb{Z}^+$, the following construction works: find the minimal $N \in \mathbb{Z}^+$ with $N \geq n$ such that $N = \prod_i p_i$ for distinct primes p_i. The intuitive idea behind this construction is we can find that powers of primes are extremely bad for the summation. To find a general minimum, we can use the AM-GM inequality: we have that

$$\sum_i p_i \geq k \sqrt[k]{\prod_i p_i} = k \sqrt[k]{N} \geq k \sqrt[k]{n},$$

where k is the number of primes that multiple to N. By the Prime Number Theorem (maybe?), it can be shown that

$$O\left(\sum_{i=1}^{k} p_k\right) \in O\left(\frac{(\ln(N))^2}{(\ln(\ln(N)))^2}\right).$$

In more intuitive terms, find the smallest number N that's greater than our time horizon n, where N is a product of distinct primes. Then, for each prime p that divides N, we create sets based on modular arithmetic classes.

4.1 Concrete Example

Say we're looking at a 5-day horizon ($n = 5$). The smallest number N that's greater than 5 and is a product of distinct primes would be $6 = 2 \times 3$.

For each prime (2 and 3), we create sets based on remainders when divided by that prime:

For prime $p = 2$:

- $[0]_2 = \{2, 4\}$ (numbers that give remainder 0 when divided by 2)
- $[1]_2 = \{1, 3, 5\}$ (numbers that give remainder 1 when divided by 2)

For prime $p = 3$:

- $[0]_3 = \{3, 6\}$ (numbers that give remainder 0 when divided by 3)
- $[1]_3 = \{1, 4\}$ (numbers that give remainder 1 when divided by 3)
- $[2]_3 = \{2, 5\}$ (numbers that give remainder 2 when divided by 3)

These sets have a beautiful property: for any two different numbers x and y in our range, there will always be at least one set that contains x but not y. This is because if two numbers are in exactly the same remainder class for every prime divisor of N, they must be equal (modulo N).

In terms of our volatility model, this means we'd add terms that average the RV values corresponding to each of these sets. Our modified HAR-RV model would look like:

$$RV_{t+1} = C + \beta_d RV_t + \beta_w RV_t^W + \beta_m RV_t^M + \sum (\gamma_{ij} \cdot \text{average of RVs in } [i]_j) + \omega_t \tag{6}$$

where $[i]_j$ represents the set of numbers with remainder i modulo prime j.

4.2 Prefacing Improvements

Similarly to before, issues have arisen revolving efficiency, grouping, weighting, scaling, etc., which has been addressed with the following:

4.2.1 Market-Weighted Modulo Classes

We modify our original modulo classes by introducing temporal decay weights. For a prime $p|N$, instead of using simple modulo classes $[a]_p$, we define weighted modulo classes:

$$[a]_p^w = \left\{ (x, \alpha^{|t-x|}) : x \in [a]_p \right\}$$

where $\alpha \in (0,1)$ is a decay parameter and t represents the current time. Our modified term becomes:

$$\sum_{p|N} \sum_{a=0}^{p-1} \gamma_{p,a} \left(\frac{\sum_{(x,w)\in[a]_p^w} w \cdot RV_{t-x+1}}{\sum_{(x,w)\in[a]_p^w} w} \right)$$

4.2.2 Market-Aligned Prime Selection

Instead of using the minimal product of primes exceeding n, we select primes based on known market cycles. Let $\mathcal{M}$ be the set of market-relevant time periods. We define our set of primes as:

$$\mathcal{P} = \{p : p \text{ is prime and } p \approx m \text{ for some } m \in \mathcal{M}\}$$

Then we let $N = \prod_{p\in\mathcal{P}} p$. Typical choices include:

$$\mathcal{M} = \{2, 5, 23\} \text{ (representing daily, weekly, and monthly patterns)}$$

4.3 Regime-Dependent Modulo Classes

To account for varying market conditions, we introduce regime-dependent parameters. Let $s_t \in \{1, \ldots, K\}$ represent the market regime at time t. Our final enhanced model becomes:

$$RV_{t+1} = C(s_t) + \beta_d(s_t)RV_t + \beta_w(s_t)RVW_t + \beta_m(s_t)RVM_t +$$
$$\sum_{p\in\mathcal{P}} \sum_{a=0}^{p-1} \gamma_{p,a}(s_t) \left(\frac{\sum_{(x,w)\in[a]_p^w} w \cdot RV_{t-x+1}}{\sum_{(x,w)\in[a]_p^w} w} \right) + \overline{\omega}_t$$

where $C(s_t), \beta_d(s_t), \beta_w(s_t), \beta_m(s_t)$, and $\gamma_{p,a}(s_t)$ are regime-dependent parameters.

The regime s_t can be determined by threshold values $\tau_1 < \tau_2$ on the recent volatility:

$$s_t = \begin{cases} 1 & \text{if } \bar{RV}_t < \tau_1 \text{ (low volatility)} \\ 2 & \text{if } \tau_1 \leq \bar{RV}_t \leq \tau_2 \text{ (normal volatility)} \\ 3 & \text{if } \bar{RV}_t > \tau_2 \text{ (high volatility)} \end{cases}$$

where $\bar{RV}_t$ is a local average of realized volatility.