On the Volatility Prediction of the HAR-RV Model

Wat Street

November 7, 2024

1 Problem Formulation

In the HAR-RV model, we have the equation

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

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

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

However, a problem (insert discussion about proof with Google Collab) that arises is when RV_i and RV_j, $i \neq j$ are swapped, since in a financial sense, this should lead to fluctuations in the volatility, but under our HAR-RV model, this does not change any of the terms.

As such, it is sensible to introduce "extra terms" which will mitigate the issue: we would like every arbitrary swap being a pair RV_i and RV_j, $i \neq j$, to affect the model. Limiting our scope to averages (or sums) of terms, we would like to minimize the amount of extra terms we introduce to the model to minimize the number of parameters the model must learn. Furthermore, we would like our solution to apply to arbitrary finite intervals of time, not just weekly and monthly, while also maintaining or improving the accuracy of the HAR-RV model.

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 c to minimized and to 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 of consider, 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 Solution 1: 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 mostly 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.

3 Solution 2: 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 n but might not stay true to the intended behaviour of the HAR-RV model because of the arbitraryness and "randomness" in the construction of the sets a_j; they may not make sense in a financial sense.

4 Solution 3: Prime Modulo Classes

For this solution we give a description of the solution and omit a formal proof of minimality for 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 $\displaystyle\prod_i k_i \geq n$. 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 $x \equiv y \pmod{\displaystyle\prod_i k_i}$, and letting $k := \displaystyle\prod_i k_i$, 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 $N \in^+$ with $N \geq n$ such that $N = \displaystyle\prod_i p_i$ 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

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