Earning yields prediction

Neural networks and econometrics models in forecasting stock returns

Andrey Grishin

Faculty of Economics Moscow State University

January 20, 2023

Agenda

Problem Formalization

Given:

1. Company name.
2. $y_t : t = \overline{1, n}$ - its earning yields series.

Aim:

1. To predict the yield in the *next* period of time.
2. Mathematical view: $\mathbb{E}\left(y_{t+1} | y_t, y_{t-1}, ..., y_1\right)$.

History of Methods:

1. Statistics.
2. Machine Learning.
3. Deep Learning.

Main Techniques

Models:

1. **Decomposition**: deconstruction of time series.
2. **Smooth-based**: removal of anomalies.
3. **Moving-Average**: tracking a single type of data.
4. **Exponential Smoothing**: (2) + exponential window function.

Already existing solutions

Existing solutions for TS forecasting problem:

1. Time-series decomposition.
2. Time-series regression models.
3. Exponential smoothing (**EWMA**).
4. **ARIMA**, SARIMA, SARIMAX.
5. **ARFIMA**, VAR, SVAR.

1. (Recurrent) **Neural Networks**.
2. **GARCH**, **FIGARCH**.
3. **SETARMA**, ADL.
4. **SVM**, **SSA**, TBATS.

etc.

Stationary methods (1)

1. AR(F)IMA(p, d, q) (Autoregressive Fractionally Integrated Moving Average)

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)\nabla^d(y_t-\mu)=\left(1+\sum_{j=1}^{q}\theta_j L^j\right)\varepsilon_t \tag{1}$$

$$ARIMA \Rightarrow d \in \mathbb{Z}^+$$

$$\nabla^d = (1-L)^d$$

$$ARFIMA \Rightarrow d \in (-0.5, 0.5)$$

$$\nabla^d = \sum_{k=0}^{\infty}\frac{\Gamma(k-d)L^k}{\Gamma(-d)\Gamma(k+1)}$$

Stationary methods (2)

1. GARCH(p, q)[1] Conditional Heteroscedasticity:

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 + \sum_{i=1}^{q} \phi_i \sigma_{t-i}^2 = \alpha_0 + \alpha\left(L\right)\varepsilon_t^2 + \beta\left(L\right)\sigma_t^2 \tag{2}$$

2. FIGARCH(p, d, q)

$$\sigma_t^2 = \alpha_0 \left[1 - \beta\left(L\right)\right]^{-1} + \varepsilon_t^2 \left[1 - \left[1 - \beta\left(L\right)\right]^{-1} \cdot \left(1 - L\right)^d \cdot \overbrace{\phi\left(L\right)}^{1-\alpha(L)-\beta(L)}\right] \tag{3}$$

$$d \in (0, 1)$$

[1] Empirically proved that GARCH(1, 1) is better to use in practice.

Stationary methods (3)

1. SETARMA$(k; p_1, \dots, p_k; q_1, \dots, q_k)$ (Self Exciting Threshold ARMA)

$$y_t = \sum_{i=1}^{k} \left[\phi_0^{(i)} + \overbrace{\Phi^{(i)}(L)\, y_t}^{\sum_{j=1}^{p_i} \phi_j^{(i)} y_{t-j}} + \varepsilon_t - \overbrace{\Theta^{(i)}(L)\, \varepsilon_t}^{\sum_{w=1}^{q_i} \theta_w^{(i)} \varepsilon_{t-w}} \right] \cdot I\left(y_{t-d} \in R_i\right) \tag{4}$$

$$\varepsilon \sim N\left(0, \sigma^2\right), d - \text{ threshold delay}, I(\cdot) - \text{ Bernoulli random process}$$

$$R_i = [r_{i-1}, r_i) : -\infty = r_0 < \ldots < r_{i-1} < r_i < \cdots < r_k = \infty : r_i - \text{ threshold}$$

$$\bigcup_{i=1}^{k} R_i = \mathbb{R}$$

Non Stationary Methods (1)

1. EWMA (Exponential Weighted Moving Average[2])

$$EWMA_t = \frac{v_t}{1-\beta^t} = \beta \cdot y_t + (1-\beta) \cdot EWMA_{t-1} \tag{5}$$

2. Neural Networks (sequential adjustment of feature space):

$$\varphi_k(W_{k-1} \cdot \varphi_{k-1}(W_{k-2} \cdot (\ldots(W_1 \cdot X)\ldots))) \tag{6}$$

3. SVM - Support Vector Machine Regressor.

$$\begin{cases} \langle w, w \rangle + C \sum_{j=1}^{n} \xi_j \to \min\limits_{w,b} : C \in \mathbb{R} \\ y_j(\langle w, x_j \rangle + b) \geq 1 - \xi_j : \forall j \\ \xi_j \geq 0 : \forall j \end{cases} \tag{7}$$

[2] $\beta = \frac{2}{n+1}$: n - number of days for averaging, 2 - smoothing factor.

Non Stationary Methods (SSA) (2)

Singular Spectrum Analysis (SSA)

1. Turn $Y = (y_0, ..., y_{N-1})$ into Hankel's (trajectory) matrix.

$$Y \to X \in \mathbb{R}^{L,K} : L \in [2, \lfloor N/2 \rfloor], K = N - L + 1 \tag{8}$$

2. Singular Vector Decomposition (SVD) of Hankelian.

$$X = \underbrace{U}_{L \times L} \underbrace{\Sigma}_{L \times K} \underbrace{V^*}_{K \times K} \tag{9}$$

$d - \mathbf{rank}(X)$.
For orthonormal $V \in \mathbb{R}^{K,K} \Rightarrow V^{-1} = V^T$ and $V \in \mathbb{C}^{K,K} \Rightarrow V^* = \left(\overline{V}\right)^T$
U - **columns** are performing **orthonormal basis set** in space of **columns** in X.
V - **columns** are performing **orthonormal basis set** in space of **rows** in X.
Σ - "diagonal" matrix composed of singular values of XX^T, which are equal to $X^T X$.

Non Stationary Methods (SSA) (3)

1. X spectral decomposition can be represented:

$$X = \sum_{j=0}^{d-1} \sigma_j U_j V_j^* = \sum_{j=0}^{d-1} \sigma_j \overbrace{U_j}^{L\times 1} \overbrace{V_j^*}^{1\times K} = \sum_{j=0}^{d-1} X_j = \sum_{s\in S} X_s + \sum_{t\in T} X_t + \dots \tag{10}$$

   $s \in S$ - seasonality. $t \in T$ - trend.
   $(\sigma_j, U_j, V_j^*) -$ singular triple : $\sigma_j = \sqrt{\lambda_j}$ - contribution of j elementary matrix to X.

2. Reconstruction of Hankel's matrix from X_j elementary matrix by Hankelisation.

$$\begin{aligned} &\hat{H} : L \times K \to L \times K \\ &\tilde{X}_j = \hat{H} X_j - \text{ anti-diagonal elements are } \approx \text{ equal.} \\ &\tilde{X}_j \to \tilde{Y}_j - \text{ reconstructed } Y \text{ by anti-diagonal averaging.} \end{aligned} \tag{11}$$

Non Stationary Methods (SSA) (4)

$$\hat{H}(A+B) = \hat{H}A + \hat{H}B. \quad \hat{H}(\alpha A) = \alpha \hat{H}A \Rightarrow \text{Linear Operator.} \tag{12}$$

$$X = \hat{H}X = \hat{H}\left(\sum_{j=0}^{d-1} \sigma_j U_j V_j^*\right) = \sum_{j=0}^{d-1} \hat{H}X_j = \sum_{j=0}^{d-1} \tilde{X}_j \tag{13}$$

$$\tilde{x}_{m,n} = \begin{cases} \frac{1}{s+1}\sum_{l=0}^{s} x_{l,s-l} & ,\ 0 \le s \le L-1 \\ \frac{1}{L-1}\sum_{l=0}^{L-1} x_{l,s-l} & ,\ L \le s \le K-1 \\ \frac{1}{K+L-S-1}\sum_{l=s-K+1}^{L} x_{l,s-l} & ,\ K \le s \le K+L-2 \end{cases} \tag{14}$$

$s = m + n$. We achieve the reconstructed times series from its decomposed parts.

Non Stationary Methods (SSA) (5)

Q: How to determine, which components should be grouped together?

A: Introduce the inner weights.

1. Firstly: inner weighted product (frequency of each element in $\tilde{X}_j$).

$$w_k = \begin{cases} k+1 & , 0 \leq k \leq L-1 \\ L & , L \leq k \leq K-1 \\ N-k & , K \leq k \leq N-1 \end{cases} \tag{15}$$

2. Correlation matrix (there is no noiseless dependences in real world data).

$$W_{corr} = \left\{ w_{i,j} = \frac{\left(\tilde{Y}_j, \tilde{Y}_i\right)_w}{\|\tilde{Y}_i\|_w \|\tilde{Y}_j\|_w} \right\}_{i,j=0}^{N-1} : \left(\tilde{Y}_j, \tilde{Y}_i\right)_w = \sum_{k=0}^{N-1} w_k \cdot \tilde{y}_{i,k} \cdot \tilde{y}_{j,k} \tag{16}$$

$w_{i,j} \to 1$ if components $\tilde{Y}_i$ and $\tilde{Y}_j$ are close, else $w_{i,j} \to 0$.
In practice is $w_{i,j} \geq 0.3 \Rightarrow \tilde{Y}_i$ and $\tilde{Y}_j$ should be grouped together.

Non Stationary Methods (SSA) (6)

Forecasting algorithm:

Note: U_j - j column of U.

1. Count $r = |\{\sigma_j : \sigma_j > 0\}|$.
2. Take $\left\{\underline{U}_{j,k} : 1 \leq j \leq r, 1 \leq k < L\right\}$: U - matrix of orthonormal column-vectors.
3. Take $\left\{\pi_j : \pi_j = U_{j,L} : j = \overline{1,r}\right\}$ - π_j - the last element of each of r columns in matrix U.
4. Evaluate $\nu = \sum_{j=1}^{r} \pi_j^2$.
5. Compute vector of coefficients $R = (a_{L-1}, \ldots, a_1)^T = \frac{1}{1-\nu} \sum_{j=1}^{r} \pi_j \underline{U}_j : \underline{U}_j \in \mathbb{R}^{L-1,1}$
6. Evaluate the forecast formula:

$$y_t = \begin{cases} \tilde{y}_t & , t = \overline{1, N} \\ \sum\limits_{j=1}^{L-1} a_j y_{t-j} & , t = \overline{N+1, N+h} \end{cases} \tag{17}$$

$\tilde{y}_t$ - reconstructed TS without noise components.

Simple RNN[4] (1)

- Simple RNN block[3] specified for our problem:

$$
\begin{aligned}
&h_t = \tanh\left(W_{hh}h_{t-1} + b_{hh} + W_{hx}x_t + b_{hx}\right)\\
&W_{hx} \in \mathbb{R}^{H_{out}\times H_{in}}, x_t \in \mathbb{R}^{H_{in}\times 1}\\
&W_{hh} \in \mathbb{R}^{H_{out}\times H_{out}}, h_t \in \mathbb{R}^{H_{out}\times 1}\\
&b_{hx}, b_{hh} \in \mathbb{R}^{H_{out}\times 1}
\end{aligned}
$$

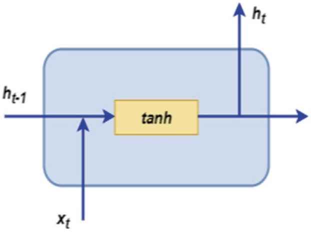

Figure 1: Simple RNN cell

[3] Usually h_0 is initialized with $\mathbf{0}$.
[4] Do not look at bidirectional way.

Gate Recurrent Unit

- GRU:

$$
\begin{aligned}
r_t &= \sigma\left(W_{ir} x_t + b_{ir} + W_{hr} h_{t-1} + b_{hr}\right) \\
z_t &= \sigma\left(W_{iz} x_t + b_{iz} + W_{hz} h_{t-1} + b_{hz}\right) \\
n_t &= \tanh\left(W_{in} x_t + b_{in} + r_t * \left(W_{hn} h_{t-1} + b_{hn}\right)\right) \\
h_t &= (1 - z_t) * n_t + z_t * h_{t-1}
\end{aligned}
$$

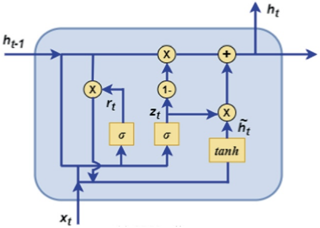

Figure 2: GRU cell

Long-Short Term Memory

- LSTM:

$$
\begin{aligned}
i_t &= \sigma\left(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}\right)\\
f_t &= \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right)\\
g_t &= \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right)\\
o_t &= \sigma\left(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}\right)\\
c_t &= f_t * c_{t-1} + i_t * g_t\\
h_t &= o_t * \tanh\left(c_t\right)
\end{aligned}
$$

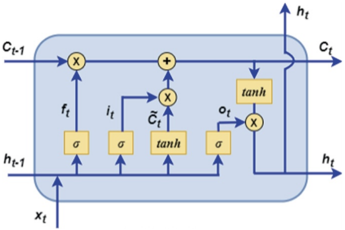

Figure 3: LSTM cell

Comparison of results

content...

Conclusion

content...

Thank you for attention!