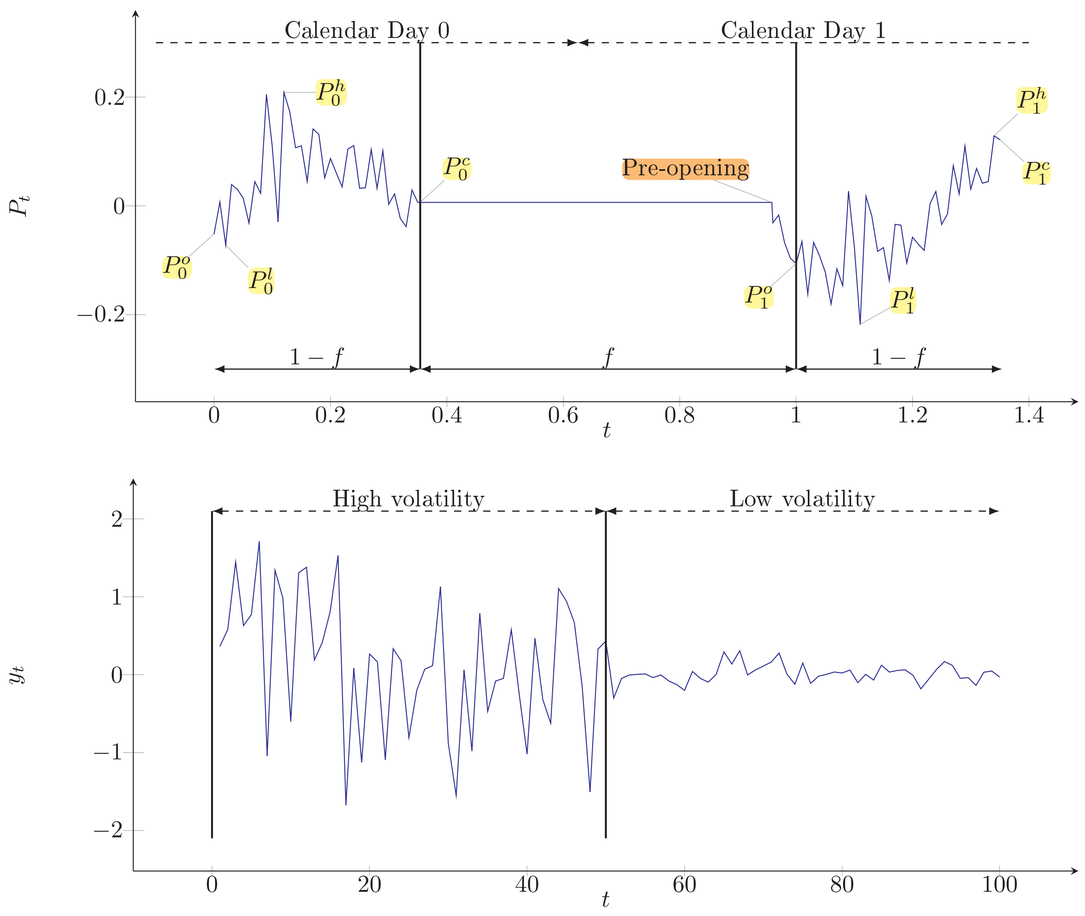
Multi-step-ahead prediction of volatility proxies

Jacopo De Stefani[1], Gianluca Bontempi[1], Olivier Caelen[2], Dalila Hattab[3]
[1] *MLG, Departement d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium*
[2] *Worldline SA/NV, R&D, Bruxelles, Belgium*
[3] *Equens Worldline, R&D, Lille (Seclin), France*

System overview

The goal of this research is employ machine learning techniques to improve the quality of market fluctuation (i.e. volalitity) forecasts. In order to achieve such result we developed the system depicted here. The purpose of our research is twofold:

- First, we aim to perform a statistical assessment of the relationships among the most used proxies in the volatility literature.
- Second, we explore a NARX (Nonlinear Autoregressive with eXogenous input) approach to estimate multiple steps of the output given the past output and input measurements, where the output and the input are two different proxies.



Volatility proxies

$\sigma_{\mathbf{i}}$ family [1]

$$\sigma_t^0 = \left[\ln\left(\frac{P_{t+1}^{(c)}}{P_t^{(c)}}\right)\right]^2 = r_t^2$$

$$\sigma_t^1 = \underbrace{\frac{1}{2f}\cdot\left[\ln\left(\frac{P_{t+1}^{(o)}}{P_t^{(c)}}\right)\right]^2}_{\text{Nightly volatility}} + \underbrace{\frac{1}{2(1-f)}\cdot\left[\ln\left(\frac{P_t^{(c)}}{P_t^{(o)}}\right)\right]^2}_{\text{Intraday volatility}}$$

$$\sigma_t^2 = \frac{1}{2\ln 4}\cdot\left[\ln\left(\frac{P_t^{(h)}}{P_t^{(l)}}\right)\right]^2$$

$$\sigma_t^3 = \underbrace{\frac{a}{f}\cdot\left[\ln\left(\frac{P_{t+1}^{(o)}}{P_t^{(c)}}\right)\right]^2}_{\text{Nightly volatility}} + \underbrace{\frac{1-a}{1-f}\cdot\sigma_2(t)}_{\text{Intraday volatility}}$$

$$\sigma_t^4 = 0.511(u-d)^2 - 0.019[c(u+d) - 2ud] - 0.383c^2$$

$$\sigma_t^5 = 0.511(u-d)^2 - (2\ln 2 - 1)c^2$$

$$\sigma_t^6 = \underbrace{\frac{a}{f}\cdot\log\left(\frac{P_{t+1}^{(o)}}{P_t^{(c)}}\right)^2}_{\text{Nightly volatility}} + \underbrace{\frac{1-a}{1-f}\cdot\sigma_t^4}_{\text{Intraday volatility}}$$

where:

$$f \in [0,1]$$
$$a = 0.17$$

$$u = \ln\left(\frac{P_t^{(h)}}{P_t^{(o)}}\right) \quad d = \ln\left(\frac{P_t^{(l)}}{P_t^{(o)}}\right) \quad c = \ln\left(\frac{P_t^{(c)}}{P_t^{(o)}}\right)$$

GARCH (1,1) model [2]

$$\sigma_t^G = \sqrt{\omega + \sum_{j=1}^{p}\beta_j(\sigma_{t-j}^G)^2 + \sum_{i=1}^{q}\alpha_i\varepsilon_{t-i}^2}$$

where $\varepsilon_{t-i} \sim \mathcal{N}(0,1)$, with the coefficients $\omega, \alpha_i, \beta_j$ fitted according to [3].

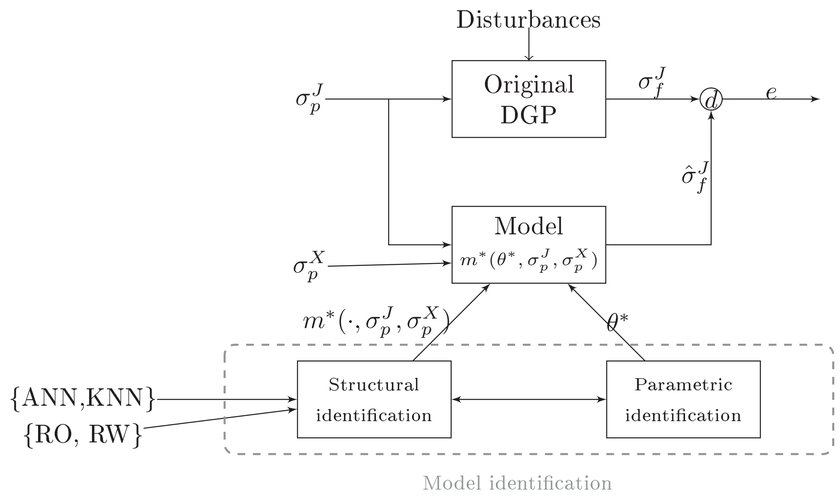
Sample standard deviation

$$\sigma_t^{SD,n} = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}(r_{t-i} - \bar{r})^2}$$

where

$$r_t = \ln\left(\frac{P_t^{(c)}}{P_{t-1}^{(c)}}\right) \qquad \bar{r_n} = \frac{1}{n}\sum_{j=t-n}^{t} r_j$$

NARX Forecaster

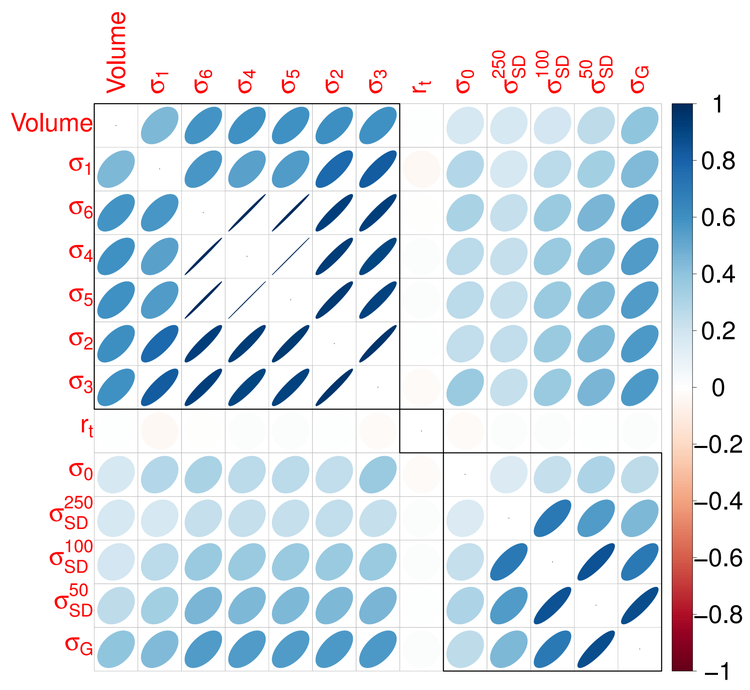


σ^X	ANN	kNN	ANN_X	kNN_X	GARCH(1,1)
σ^6	0.07	0.08	0.06	0.11	1.34
$Volume$	0.07	0.08	0.07	0.14	1.34
$\sigma^{SD,5}$	0.07	0.08	0.07	0.09	1.34
$\sigma^{SD,15}$	0.07	0.08	0.06	0.10	1.34
$\sigma^{SD,21}$	0.07	0.08	0.06	0.10	1.34

σ^X	ANN	kNN	ANN_X	kNN_X	GARCH(1,1)
σ^6	0.58	0.49	0.53	0.56	1.15
$Volume$	0.58	0.49	0.57	0.66	1.15
$\sigma^{SD,5}$	0.58	0.49	0.58	0.58	1.15
$\sigma^{SD,15}$	0.58	0.49	0.65	0.65	1.15
$\sigma^{SD,21}$	0.58	0.49	0.56	0.65	1.15

This setup includes **two volatility proxies simultaneously**. Two ML techniques (**Artificial Neural Networks and K-Nearest Neighbors**) are employed in the forecaster. The purpose is to investigate how the combination of proxies affects the forecasting performance of the algorithms.

Correlation analysis



The figure shows the aggregated correlation (over all the 40 time series) between the proxies, obtained by meta-analysis [4]. The black rectangles indicate the results of an hierarchical clustering using [5] with k=3. We observed that:

- As expected, a correlation clustering phenomenon exists between proxies belonging to the same family, i.e. σ_t^i and $\sigma_t^{SD,n}$.
- The presence of σ_t^0 in the $\sigma_t^{SD,n}$ cluster can be explained by the fact that the former represents a degenerate case of the latter when $n = 1$.
- A significant correlation between the volume and the σ_t^i family.

Conclusions - ⚠ Preliminary results

- Correlation clustering among proxies belonging to the same family, i.e. σ_t^i and $\sigma_t^{SD,n}$.
- Both machine learning methods outperform the benchmark methods (naive and GARCH).
- ANN can take advantage of the additional information provided by the exogenous proxy better than k-NN
- Combination of proxies coming from different families could improve forecast accuracy

For the final version we expect to provide additional comparisons in terms of the number of series, forecasting horizons h and model orders m.

References

[1] Mark B Garman and Michael J Klass. On the estimation of security price volatilities from historical data. *Journal of business*, pages 67–78, 1980.

[2] Peter R Hansen and Asger Lunde. A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics*, 20(7):873–889, 2005.

[3] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

[4] Andy P Field. Meta-analysis of correlation coefficients: a monte carlo comparison of fixed-and random-effects methods. *Psychological methods*, 6(2):161, 2001.

[5] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Find the paper at: