MATHEMATICAL MODELING AND COMPUTING, Vol. 11, No. 1, pp. 134–138 (2024)

Machine learning for forecasting some stock market index

Benmoumen M., Salhi I.

LaMSD, Department of Mathematics, Faculty of Science,
Mohammed Premier University, Oujda, Morocco

(Received 28 March 2023; Revised 13 February 2024; Accepted 16 February 2024)

In this paper, we evaluate the QMLKF algorithm, designed in the previous paper [Benmoumen M. Numerical optimization of the likelihood function based on Kalman Filter in the GARCH models. Mathematical Modeling and Computing. **9** (3), 599–606 (2022)] for parameter estimation of GARCH models, by transposing it to real data and then present our machine learning for forecasting the returns of some stock indices.

Keywords: *Machine learning; statistical learning; GARCH model; Kalman filter; stock market index.*

2010 MSC: 62Fxx, 62M10, 68U20, 62P20, 62-04 **DOI:** 10.23939/mmc2024.01.134

1. Introduction

Advanced analysis used by the financial industry has given increasing importance to time series modeling. In particular, nonlinear time series have attracted the widest attention [1]. Numerous financial series, such as stock index, stock and exchange rate returns, exhibit leptokurtosis and time-varying volatility. Both of these features have been studied extensively since Nicholls and Quinn and Engle reported on them. The autoregressive conditional heteroscedastic (ARCH) models [2], and their generalization, the GARCH model [3], provide in fact a convenient framework for studying time varying volatility in financial markets. For instance, financial time series models for stock market index are typical examples of GARCH models.

In this paper, we were concerned with modeling the main index of the Casablanca Stock Exchange, created in 2002, MASI (Moroccan All Shares Index). The main idea was to adopt the machine learning principle into the statistical modeling process, i.e. learn and improve from previous exposures. The statistical learning phase is performed by the QMLKF algorithm [4] then by an existing algorithm in the R software for comparison purpose [5]. According to the study results, the new approach was found to be effective in modeling MASI. This work is part of a series of papers that focus on investigating QMLKF algorithm in the case of GARCH, ARCH and RCA models [4, 6, 7].

The remaining parts of the paper are organized as follows, in section 2, we set out the basic elements used in MASI returns modeling. Finally, the findings from our study are presented and analyzed in section 3.

2. Framework

The singularity of GARCH model lies in its capacity to restore the stylized facts characterizing financial series, i.e. the statistical and empirical properties frequently observed in the usual conditions of a financial market, thanks to the concept of the conditional variance which is expressed as a linear function of the square of the past values of the series. Formally the strong class of the GARCH(p, q) model for a time series y_t, is given by

$$\begin{cases} \varepsilon_t = \sigma_t \eta_t, \\ \sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \end{cases} \tag{1}$$

where (η_t) is a sequence of independent and identically distributed (i.i.d.) random variables such that $E(\eta_t) = 0$, $E(\eta_t^2) = 1$ and $\omega > 0$, $\alpha_i \geqslant 0$ $(i = 1, \ldots, q)$, $\beta_j \geqslant 0$ $(j = 1, \ldots, p)$.

The modeling by GARCH is performed by minimizing the following cost function

$$L_n(\theta) = n^{-1} \sum_{t=1}^{n} \left\{ \frac{\varepsilon_t^2}{\sigma_t^2(\theta)} + \log \sigma_t^2(\theta) \right\}. \tag{2}$$

In the practical case, the process $\sigma_t(\theta)$ represents the assets volatility. Since this process is not observable, the idea was to generate it via the Kalman filter. The use of this filter is motivated by its principle, which is to sequentially estimate the states of a dynamic system from a series of incomplete or perturbed measurements, while minimizing the mean square error. This procedure gives the best linear estimates.

The estimation phase is carried out by the QMLKF algorithm. This algorithm is characterized by three main phases. The first phase consists of checking data stationarity. The cost function is then deduced using the Kalman filter in the second phase. Finally, the algorithm is completed by applying an optimization method, we recommend to use a method capable of finding the global minimum in the presence of a very large number of local minima. In this case-study simulated annealing (SA) [8] method performed well.

This approach is compared to the existing algorithm in rugarch package of R software. The rugarch is a powerful package that provides a set of methods for modeling univariate GARCH processes, including fitting, filtering, forecasting, simulation, as well as diagnostic tools including graphics and various tests.

3. MASI stock index modeling

3.1. Pre-processing

In this study, we expect to achieve a double objective. At first, we aim to evaluate the QMLKF estimation algorithm by applying it to real data and the second, we plan to model the MASI stock index data by GARCH(1,1). This series, like all other financial series, shows statistical regularities. These properties are difficult to reproduce artificially using linear stochastic models such as ARMA models. Indeed, the price of this asset is non-stationary in the sense of the second-order stationarity, whereas after logarithmic differentiation, it seems to become stationary, this new process is referred to as return, (see Figures 1 and 2). In other words, if (p_t) is the stochastic process associated with the price of a given financial asset, the return (r_t) between the dates $t-1$ and t would be $\log(p_t) - \log(p_{t-1})$.

Fig. 1. MASI Daily Adjusted Close Price Index.

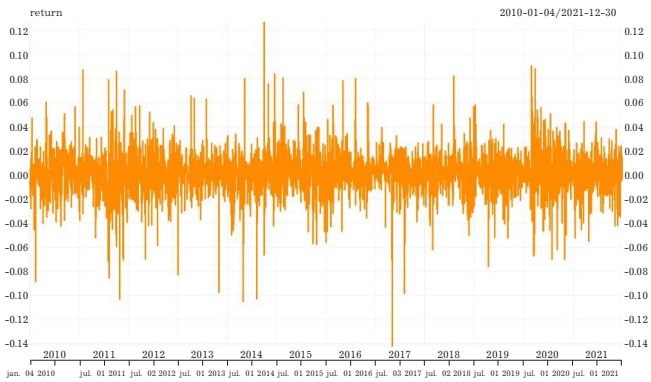

Fig. 2. Daily adjusted MASI closing price returns.

Moreover, the evolution of returns over time brings out another property, which is the accumulation of strong variations in packages, which refutes the hypothesis of constant conditional variance, thus making GARCH and ARCH models suitable candidates for modeling MASI stock index returns. In terms of autocorrelations, returns are uncorrelated, making the hypothesis of weak white noise plausible. In contrast, the autocorrelations of the square returns are strong: we are in the presence of the long memory phenomenon, (see Figures 3 and 4).

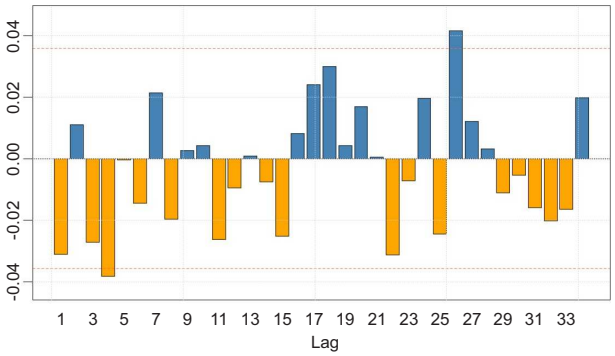

Fig. 3. Autocorrelograms of returns-MASI.

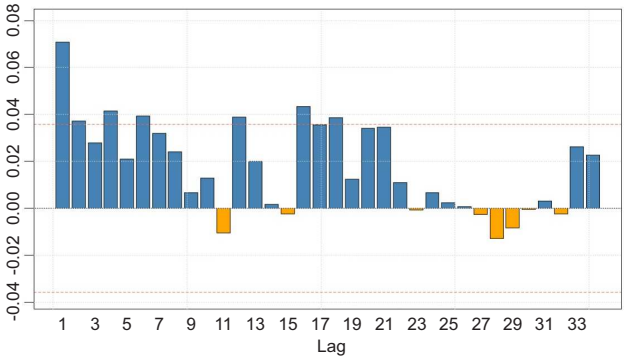

Fig. 4. Autocorrelograms of square returns-MASI.

Another important stylized fact to note concerns the distribution of returns, which is generally asymmetric and leptokurtic, i.e. pointed at its mean and with thick tails, (see Figure 5). All these characteristics support the choice of GARCH models.

3.2. Modeling and validation phase

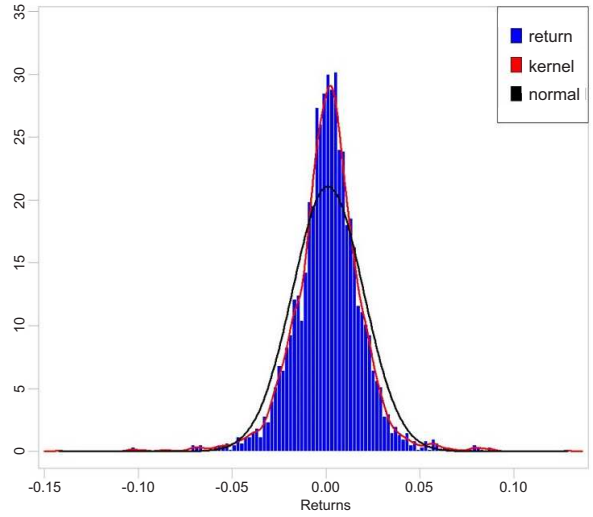

Fig. 5. Distribution of MASI returns.

In the following, we present an attempt to model the returns of the MASI index over the period from 01/01/2010 to 31/12/2021. For this purpose, we operate in two phases, one of construction and the other of validation, by creating two samples: a training sample where model parameters are estimated and a test sample on which the model is assessed.

Parameter estimation is performed using two tools: at first by the R package "rugarch" and then by the QMLKF algorithm. The resulting models are called Model 1 and Model 2, respectively. Models identification and validation are based on the analysis of the residuals.

At the end of the training phase, analysis of the autocorrelogram of the residuals and the squared residuals in two models revealed that the autocorrelations are generally insignificant, that means that both residuals are white noise and besides the models best explain the correlations between the observations.

Table 1. Summary of the two models GARCH(1,1).

	Model 1 (rugarch)	Model 2 (QMLKF)
$\hat{\omega}$	0.00009	0.000622
$\hat{\alpha}$	0.039221	0.003012
$\hat{\beta}$	0.940192	0.477476
Kurtosis	3.772746	5.46538
Skewness	0.975702	-0.2168179
Akaike	-5.28560	-5.755973
MSE	0.116857	0.001119

A summary of the resulting models is given in the table below (see Table 1). Note that the distribution of the $(\eta_t)_t$ process used in the definition of the GARCH model in this treatment is an asymmetric Student (Skew Student). This choice takes into account the characteristics of the distribution of the observed data (see Figure 5). The parameters Kurtosis and Skewness indicate in the table respectively the skewness and kurtosis coefficients of (η_t) in each model.

Having applied both models to the test sample, we extracted the residuals from the predictions. In Figures 10–13, we notice that the residuals take the structure of white noise. A comparative analysis based on the Akaike criterion reveals no great difference between two models, but in terms of MSE Model 2 is significantly more accurate than Model 1, (see Table 1).

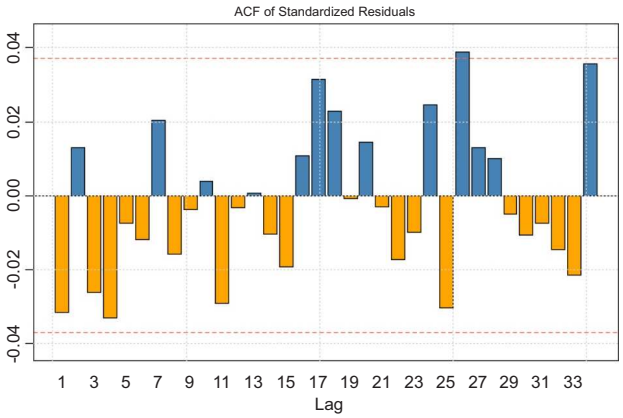

Fig. 6. Autocorrelogram of residuals Model 1.

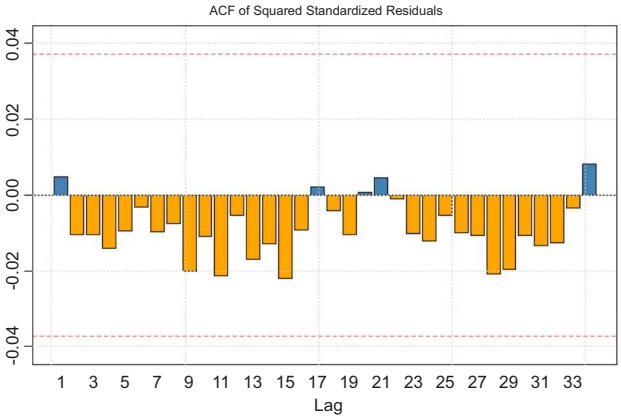

Fig. 7. Autocorrelogram of squared residuals Model 1.

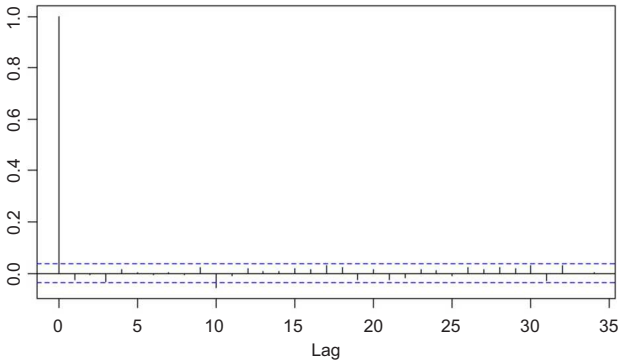

Fig. 8. Autocorrelogram of residuals Model 2.

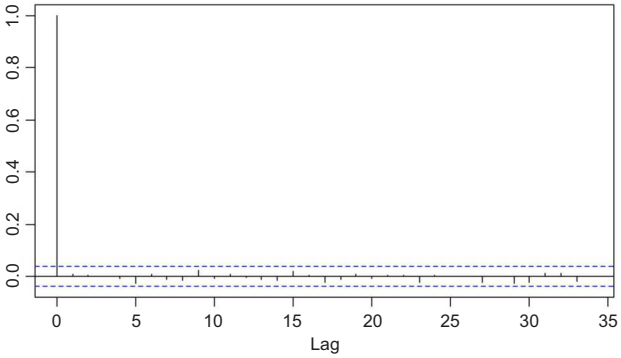

Fig. 9. Autocorrelogram of squared residuals Model 2

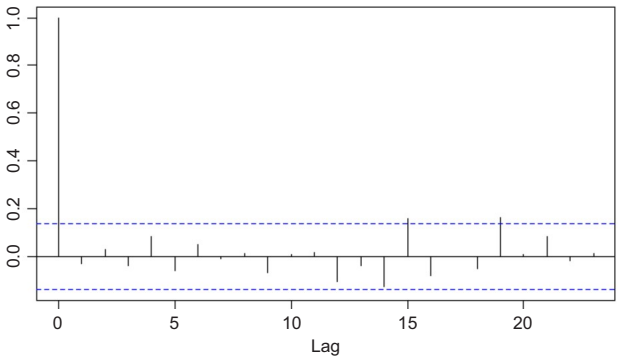

Fig. 10. Autocorrelogram of residuals of forecasts by Model 1.

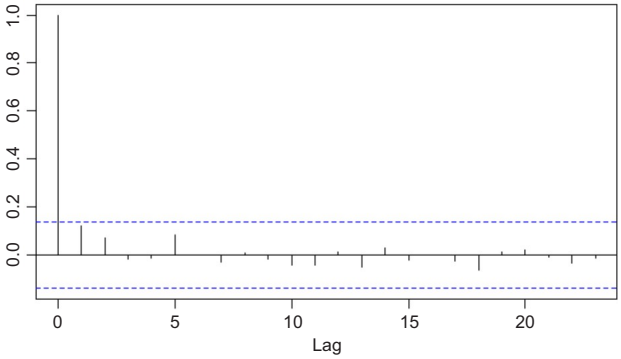

Fig. 11. Autocorrelogram of squared residuals of forecasts by Model 1.

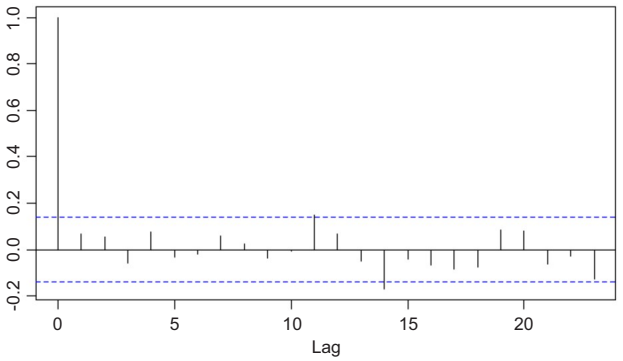

Fig. 12. Autocorrelogram of the residuals of the forecasts by Model 2.

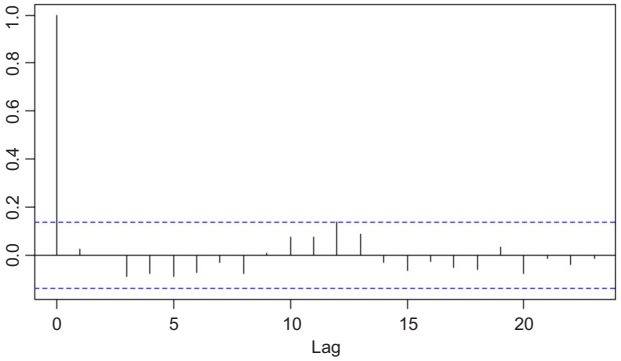

Fig. 13. Autocorrelogram of squared residuals of forecasts by Model 2.

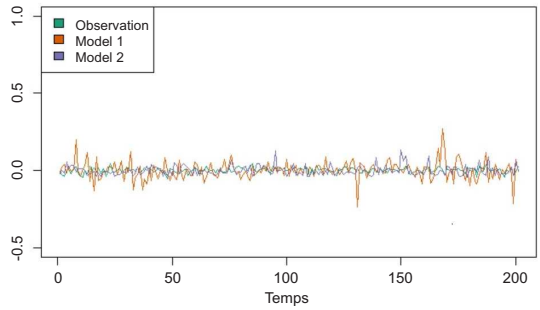

Fig. 14. Returns-MASI observed and predicted.

Graphically, by plotting the predictions and observations (see Figure 14), we can say that the predictions of model 1 show strong and more frequent fluctuations, thereby making the prediction far from the observations, unlike the predictions of model 2. Hence, for all the above reasons, we opt for Model 2 instead of Model 1 to fit the MASI returns.

4. Conclusion

In this paper, we developed a model for MASI stock index returns. The novelty we have proposed consists in using the QMLKF algorithm. This practical application provided positive feedback that, on the one hand, confirmed the simulation results of a previous study. On the other hand, this result revealed the competitiveness of the algorithm compared to the rugarch package algorithm.

[1] Franses P. H., Van Dijk D. Non-linear time series models in empirical finance. Cambridge University Press (2000).

[2] Engle R. E. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica. **50** (4), 987–1007 (1982).

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

[4] Benmoumen M. Numerical optimization of the likelihood function based on Kalman Filter in the GARCH models. Mathematical Modeling and Computing. **9** (3), 599–606 (2022).

[5] Ghalanos A. Introduction to the rugarch package (Version 1.3-1), (2020). http://cran.r-project.org/web/packages/rugarch.

[6] Benmoumen M. Numerical optimization of the likelihood function based on Kalman filter in the ARCH models. AIP Conference Proceedings. **2074**, 020020 (2019).

[7] Benmoumen M., Allal J., Salhi I. Parameter Estimation for p-Order Random Coefficient Autoregressive (RCA) Models Based on Kalman Filter. Journal of Applied Mathematics. **2019**, 8479086 (2019).

[8] Corana A., Marchesi M., Martini C., Ridella S. Minimizing Multimodal functions of continuous variables with "Simulated Annealing" Algorithm. ACM Transactions on Mathematical Software. **13** (3), 262–280 (1987).

Машинне навчання для прогнозування деяких індексів фондового ринку

Бенмумен М., Салхі І.

LaMSD, кафедра математики, природничий факультет, Прем'єрський університет Мохаммеда, Уджда, Марокко

У цій статті оцінюється алгоритм QMLKF, розроблений у попередній статті [Benmoumen M. Numerical optimization of the likelihood function based on Kalman Filter in the GARCH models. Mathematical Modeling and Computing. **9** (3), 599–606 (2022)] для оцінки параметрів моделей GARCH, шляхом перенесення його на реальні дані, а потім представляємо наше машинне навчання для прогнозування прибутковості деяких фондових індексів.

Ключові слова: *машинне навчання; статистичне навчання; модель GARCH; фільтр Калмана; індекс фондового ринку.*