On detecting the dependence of time series*

Nikolai Dokuchaev

Department of Mathematics & Statistics, Curtin University,

GPO Box U1987, Perth, 6845 Western Australia

email N.Dokuchaev@curtin.edu.au

October 14, 2010

Abstract

This short note suggests a heuristic method for detecting the dependence of random time series that can be used in the case when this dependence is relatively weak and such that the traditional methods are not effective. The method requires to compare some special functionals on the sample characteristic functions with the same functionals computed for the benchmark time series with a known degree of correlation. Some experiments for financial time series are presented.

Key words: serial dependence, non-parametric methods, technical analysis, econometrics

This short note presents some statistical experiments with the purpose to estimate the dependence for time series. We suggest to compare historical time series with a series with given and known correlation using a functional formed from empirical characteristic functions defined similarly to Hong (1999). It gives a simple empirical method that allows to estimate the dependence by comparing the values of this functional for two time series. The suggested test can be an addition to dependence and correlation tests such as Pearson test, Hoeffding's test, Spearman test, Kendall Tau Rank test, or chi-square test; see, e.g., Conover (1999) and Hollander and Wolfe (1999). In our experiments, we used a simple autoregression as the benchmark process. We present some results of experiments for time

*Accepted to "Communications in Statistics – Theory and Methods"; in press. Submitted: 17 May 2010. Revised: 21 September 2010

series with admittedly weak dependence such as financial time series for returns of stock prices.

Note that the problem of detecting the serial correlations for financial series is very important for applications in finance. In particular, this problem is related to the open problem of validation of "technical analysis" methods that offer trading strategies based on historical observations. The main benefit is that these strategies are model-free: they require only historical data. This is why they are so popular among traders. There are many different strategies suggested in the framework of "technical analysis". Hsu and Kuan (2005) mentioned that there are more than 18,326 different empirical trading rules being used in practice. However, the question remains open if the main hypothesis of technical analysis is correct. This hypothesis suggests that it is possible to make a statistically reliable forecast for future stock price movements using recent prices, and, finally, to find "winning" in statistical sense trading strategies. However, the dependence from the past (if any) if extremely weak for the stock prices, and this dependence is difficult to catch by usual statistical methods. Statistical studies of historical prices made as early as in 1933 didn't support the hypothesis that there is significant dependence from the past and predictability for the stock prices; see the discussion and the bibliography in Chapter 2, pp. 37-38, from Shiryaev (1999). This is the reason why the most common and mainstream model for the stock prices is the random walk or its modifications. Recently, new efforts were devoted to this problem, and some signs of possible presence of statistically significant dependence from the past were found (see, e.g., Lo *et al.* (2000), Hsu and Kuan (2005)), Lorenzoni *et al* (2007)). In particular, Lorenzoni *et al* (2007) found that, for a certain models of stock price evolution, there is a statistically significant informational content in some patterns from technical analysis. The computational experiments with our tests also show that the financial time series have some dependence.

The paper is organized as follows. In Section 1 we describe the method and collect the notation and definitions. Sections 2 contains description of the experiments with financial time series. Sections 3 contains conclusions and some suggestions for future research.

1 The method

Clearly, if random variables ξ and η are independent, then

$$e(q) = |\mathbb{E}e^{i\xi q}\mathbb{E}e^{i\xi q} - \mathbb{E}e^{i(\xi+\eta)q}| \equiv 0, \quad q \in \mathbf{R}. \tag{1}$$

In (1), $\mathbb{E}$ denote the expectation, $i = \sqrt{-1}$ is the imaginary unit. Condition (1) is a necessary but not a sufficient condition of independence; see example in Hamedani and Volkmer (2009). If (1) holds then ξ and η are said to be subindependent.

We suggest to measure the sample analog of the function (1) for the time series and their past history and match it with the similar function for a benchmark AR(1) series with given correlation coefficient.

Let R_t be a time series (not necessary a stationary time series). Let $G_t = \{R_k\}_{k=-\infty}^{k=t} = (R_t, R_{t-1}, R_{t-2}, ...)$ represent the history of the series. We are interested in detecting the dependence of the current value of the series from the history.

We denote by $\mathbf{E}$ the sample mean over available historical data. Let ℓ_∞ denote the set of all bounded sequences $\{x_k\}_{k=0}^{+\infty} \subset \mathbf{R}$.

We suggest to calculate the values

$$e(h, F, q) = \left|\mathbf{E}e^{iqh(G_{t-1})}\mathbf{E}e^{iqF(R_t)} - \mathbf{E}e^{iqh(G_{t-1})+iqF(R_t)}\right|, \tag{2}$$

where $h : \ell_\infty \to \mathbf{R}$ and $F : \mathbf{R} \to \mathbf{R}$ are some real valued functions, $q \in \mathbf{R}$.

If $e(q)$ is sufficiently different from zero under some statistical degree, then one is provided with empirical evidence in favor of dependence.

To measure the degree of the dependence, we suggest to compare a norm of function (2) with the same norm of a similar function calculated for some benchmark the time series $\{\widetilde{R}_t\}$ with certain given level of correlations. We suggest to use as the benchmark series the time series generates by autoregression AR(1)

$$\widetilde{R}_t = a\widetilde{R}_{t-1} + \varepsilon_t, \tag{3}$$

where $a \in (-1, 1)$, ε_t are samples from independent identically distributed random variables such that $\mathbb{E}\varepsilon_t = 0$. These series can be created using Monte-Carlo simulation.

Let $\widetilde{G}_t = \{\widetilde{R}_k\}_{k=-\infty}^{k=t} = (\widetilde{R}_t, \widetilde{R}_{t-1}, \widetilde{R}_{t-2}, ...)$. Let $\widetilde{e}(h, F, q, a)$ be the corresponding value (2) calculated with (R_t, G_t) replaced by $(\widetilde{R}_t, \widetilde{G}_t)$.

The following definition is rather heuristic but still gives an idea how to measure the dependence from the history.

Definition 1.1 *(i) Let functions h and F be given. Let a sample $\{R_t\}$ be given. and let AR(1) autoregression $\{\widetilde{R}_t\}$ be defined by (3) with some given coefficient $a \in \mathbf{R}$. We say that the sets of characteristics $(h(G_{t-1}, F(R_t)))$ and $(h(\widetilde{G}_{t-1}, F(\widetilde{R}_t)))$ have the same level of dependence from the history given (h, F) if $e(h, F, q)$ is similar in some sense to $\widehat{e}(h, F, q, a)$.*

(ii) Let a set $\mathcal{P}$ of the pairs of functions (h, F) be given. We say that the sample $\{R_t\}$ have the same level of dependence from the history as autoregression $\{\widetilde{R}_t\}$ defined by (3) with the coefficient $\widehat{a} = \widehat{a}(\mathcal{P})$ if this $|\widehat{a}|$ is the supremum over all $|a|$ such that there exists $(h, F) \in \mathcal{P}$ such that $(h(G_{t-1}, F(R_t)))$ and $(h(\widetilde{G}_{t-1}, F(\widetilde{R}_t)))$ have the same level of dependence from the history.

Remark 1.1 In Definition 1.1(i), the nature of the required similarity is not specified. In the experiments described below, we have assumed that the similarity is achieved when the $\sup_{q\in[0,\bar{q}]} e(h, F, q) = \sup_{q\in[0,\bar{q}]} \widehat{e}(h, F, a, q)$ for $\bar{q} > 0$ defined by computational abilities and decay of the functions; the interval $[0, \bar{q}]$ should be large enough. In other words, we accepted that the similarity is achieved when these functions have the same norm in $L_\infty(0, \bar{q})$. So far, we have not compare this choice with other possible choices such as comparison of integrals $\int_0^{\bar{q}} e(h, F, q))dq$ and $\int_0^{\bar{q}} e(h, F, q))dq$.

Note that it follows from the definitions that $e(h, F, q) \equiv e(h, F, -q)$ and $\widehat{e}(h, F, a, q) \equiv \widehat{e}(h, F, a, q)$; therefore, it suffices to consider $q \geq 0$.

2 Statistical experiments

We have carried out the following experiment for the time series representing the returns for the historical stock prices, i.e., when $R_t \triangleq S_t/S_{t-1} - 1$ where S_t are the stock prices. Using daily price data from 1984 to 2009 for 19 American and Australian stocks (Citibank, Coca Cola, IBM, AMC, ANZ, LEI, LLC, LLN, MAY, MLG, MMF, MWB, MIM, NAB, NBH, NCM, NCP, NFM and NPC), we generated samples of price data for one synthetic return as $\{R_t\} \triangleq \{S_t/S_{t-1} - 1\}$, where S_t is the price at day t. In fact, the full 47 years of

data was not available for all the stocks; we have the size of sample equal to 69,948. Since a conclusion about a technical analysis strategy can only be made after one collects the results of using it as many times as possible (i.e., either for different stocks or for different time intervals), we claim that our model and our experiment are not unreasonable.

Let $\mathbb{I}$ denote the indicator function.

Let us describe the the analysis done for three possible choices of functions h and F.

Choice 1:

$$h(G_{t-1}) = R_{t-1}, \quad F(R_t) = R_t. \tag{4}$$

We found that $a(h, F) = 0.1$ for this case (see Fig.1).

Choice 2:

$$h(G_{t-1}) = \mathbb{I}_{\{R_{t-1}>0\}}, \quad F(R_t) = \mathbb{I}_{\{R_{t-1}>0\}}. \tag{5}$$

We found that $a(h, F) = 0.1$ for this case (see Fig.2).

Choice 3.

$$\begin{aligned} &h(G_{t-1}\}) = \frac{1}{2}\mathbb{I}_{\{R_{t-1}>0\}} + \frac{1}{2^2}\mathbb{I}_{\{R_{t-2}>0\}} + ... + \frac{1}{2^d}\mathbb{I}_{\{R_{t-d}>0\}}, \\ &F(R_t) = \mathbb{I}_{\{R_t>0\}}, \end{aligned} \tag{6}$$

where d is given. We found that $a(h, F) = 0.15$ for this case with $d = +\infty$ (see Fig.3).

Figures 1-3 show the samples of $e(h, F, q)$ and $\widehat{e}(h, F, a, q)$ for (h, F) defined by (4)-(6) respectively, and for $a = a(h, H)$ that $max_q e(h, F, q) = \max_q \widehat{e}(h, F)$.

On the choice of (h, F)

The functions (h, F) in Choices 1-3 were not selected by some optimal way; we leave it for future research. However, there are certain reasons for the particular Choices 2-3 for the functions h and F instead of more straightforward Choice 3. First, the binary characterization of increments helps to reduce calculations. Second, the Choices 2-3 reduces the impact of volatility and give more emphasize on price movement in the spirit of technical analysis for stock trading, where the sign of the changes is crucial. Special selection of h in the Choice 2 allows to take in account all the history of the signs of the price movements (i.e.,

the history reduced to the binary characteristics); the impact of older movements decays exponentially. In our experiments, we found that Choices 2 and 3 ensures the most robust results; the corresponding curves on the graphs generated by the benchmark AR(1) series behave very regularly with respect to small changes of $|a|$. It can be illustrated by Fig 3 and Fig 4, where the results for Choice 3 with $a = 0.15$ and $a = -0.12$ respectively are presented. The conditions of the experiment were the same except the choice of a.

In some other experiments that we leaved outside of this paper, we found that different combinations of h and F from Choices 1-3 also give robust results that are close to the results presented here. (In particular, we used the pair consisting of function F from Choice 1 and function h from Choice 3).

The experiments show that the maximum matching value of $a(h, F, q)$ can be achieved for the choice of (h, F) defined by (6). Moreover, this result is quite robust with respect to variations of the parameters and data sets. The graphs are practically not changing if we remove any subset from the set of 19 stocks.

For these experiments, we developed a simple MATLAB programm. This programm cannot run over a set of a, so the graphs for every particular a were analyzed one by one. For every particular a, this programm gives the answer for the question: *Is dependence of underlying time series is stronger or weaker that the dependence of AR(1) with the coefficient a?* Obviously, a better programm could make automatic calculation of the best matching a.

On the direct computing the correlation of coefficient

In addition, we tested the hypothesis that the series for R_t is described as linear AR(1) autoregression

$$R_t = \beta R_{t-1} + \varepsilon_t.$$

The standard least square estimator gives the value $\widehat{\beta} = -0.005$ that is too small to indicate the presence of correlation. The same value $\widehat{\beta} = -0.005$ was obtained from Pearson Product-Moment Correlation test for R_t and R_{t-1}. Moreover, this value $\widehat{\beta}$ coefficient appears to be non-robust with changes of the data set; it varies significantly if we add or delete a particular stock.

On the other hand, we found, using our test, that the series R_t has the same degree of dependence as AR(1) regression (3) with coefficient $a = 0.15$. Therefore, we can conclude

that the dependence cannot be expressed via straightforward calculation of the correlation as the coefficient for the linear autoregression model.

Selection of $\operatorname{sign} a$ and $\operatorname{Var} \varepsilon_t$ for the benchmark series

Remark 2.1 In our criterion, we use only the value $|a|$ and ignore the sign for the coefficient a that defines the correlation of the benchmark series. The reason is that, as we observed in the experiments, the shape of $\widehat{e}(h, F, q, a)$ is very close to the shape of $\widehat{e}(h, F, q, -a)$. It can be illustrated by Fig 3 and Fig 4, where the results for Choice 3 with $a = 0.15$ and $a = -0.12$ respectively are presented. The conditions of the experiment were the same except the selection of a. Other experiments showed the same independence from the sign of a for other choices of (h, F).

In the experiments, we considered benchmark series with $\widehat{R}_0 = 0$. For the Choices 2 and 3, the results are not affected by the selection of the value for $\operatorname{Var} \varepsilon_t$. We used $\operatorname{Var} \varepsilon_t = 1$ for the Choices 2 and 3. For the Choice 1, the selection of $\operatorname{Var} \varepsilon_t$ defines the scaling of the function $\widehat{e}(h, F, a, q)$: for instance, let the series ε_t generates the function $\widehat{e}(h, F, a, q)$. If we replace ε_t by $k\varepsilon_t$ for some $k > 0$, then the function $\widehat{e}(h, F, a, q)$ will be replaced by the function $\widehat{e}(h, F, a, kq)$, i.e., the value of $\sup_q \varepsilon(h, F, q)$ will not be affected but the visual image of the graph of the function will be changed.

We found that a convenient scaling can be achieved with $\operatorname{Var} \varepsilon_t = (1 - a^2)V$, where V is the sample second moment for R_t. In this case, $\operatorname{Var} \widetilde{R}_t$ is asymptotically close to V, and it ensures a satisfactory scaling.

3 Other choices of benchmark processes

We have suggested to use the simplest AR(1) series as the benchmark series $\{R_t\}$. Alternatively, other models with certain predetermined level of serial dependence can be used, such as Markov chains with a given size of non-diagonal elements in the matrix transitional probabilities, or with an ARCH or GARCH process, or with autoregression of a higher order. These model with multidimensional parameters have more flexibility. However, it is more difficult to use them to order the series R_t with respect to the degree of dependence.

Consider, for example, ARCH series for the purpose to generate a benchmark dependence. Let us consider the following model for the benchmark process:

$$\widetilde{R}_{t+1} = a\widetilde{R}_t + \sigma_t \varepsilon_t, \qquad \sigma_t = b + c\varepsilon_t^2.$$

For this model, the degree of the dependence is defined by (a, b, c). The case of $c = 0$ corresponds to AR(1) model that was used above. Matching the degree of dependence for the ARCH model and for the observed series $\{R_k\}$ leads for situation when the same degree of dependence can be achieved with selection of parameters (a_1, b_2, c_1) and (a_2, b_2, c_2), where $(a_1, b_2, c_1) \neq (a_2, b_2, c_2)$. For example, we obtained that $(a, b, c) = (0.02, 1, 0.08\mathbf{E}\varepsilon_t^2)$ gives the same maximum of $\widehat{e}$ as $(a, b, c) = (0.1, 0, 0)$ (i.e., without ARCH), with (h, F) defined by Choice 1. The corresponding plot is shown on Fig. 5 below. Therefore, the presence of vector parameters lead to analysis of one-dimensional surfaces $\{(a, b, c)\} \in \mathbf{R}^3$ that correspond to different level of dependence of stock returns. We leave it for future research.

4 Conclusion

We suggested a method that allows to make a fast detecting of dependence from the past and some estimate of the degree of dependence via comparison with a benchmark AR(1) series. This method requires to compare visually the graphs for functions e and $\widehat{e}$. This estimate is not very precise; however, it is quite robust with respect to variations of the parameters and data sets. We used this method in statistical experiments with stock prices. We found that the result of the experiments support the hypothesis that there is certain dependence for financial time series. It gives a reason in favor of an existence of a statistically winning strategy based on observations of recent prices (i.e., a winning "technical analysis" trading strategy).

Acknowledgment

This work was supported by NSERC grant of Canada 341796-2008 to the author.

References

Conover, W. J. (1999). Practical Nonparametric Statistics. 3rd edition. Wiley.

G. G. Hamedani, H. W. Volkmer. (2009). Letter to the Editor. *The American Statistician* **63** (3), 295-295

Hsu, P.-H., Kuan, C.-M. (2005). Reexaming the profitability of technical analysis with data snooping checks. *Journal of Financial Econometrics* **3**, iss. 4, 606-628.

Hollander and Wolfe (1999). Non-parametric statistical method. Wiley.

Hong, Y. (1999) Hypothesis Testing in Time Series via the Empirical Characteristic Function: A Generalized Spectral Density Approach *Journal of the American Statistical Association* **94**, No. 448, 1201–1220.

Lo, A.W., Mamaysky, H., and Wang, Jiang. (2000). Foundation of technical analysis: computational algorithms, statistical inference, and empirical implementation. *Journal of Finance* **55** (4), 1705-1765.

Lorenzoni, G., Pizzinga, A., Atherino, R., Fernandes, C,, Freire, R.R.. (2007). On the Statistical Validation of Technical Analysis. *Revista Brasileira de Finanças*. Vol. 5, No. 1, pp. 328.

Shiryaev, A.N. (1999) Essentials of Stochastic Finance. Facts, Models, Theory. World Scientific Publishing Co., NJ, 1999.

Figure 1: Shapes for $e(h, F, q)$ for (h, F) defined by (4) and for $\widehat{e}(h, F, a, q)$ with $a = 0.1$, $q \in [0, 180)$, $\operatorname{Var} \varepsilon_t = (1 - a^2)V$; ——: values of $e(h, F, q)$; $- - -$: values of $\widehat{e}(h, F, a, q)$.

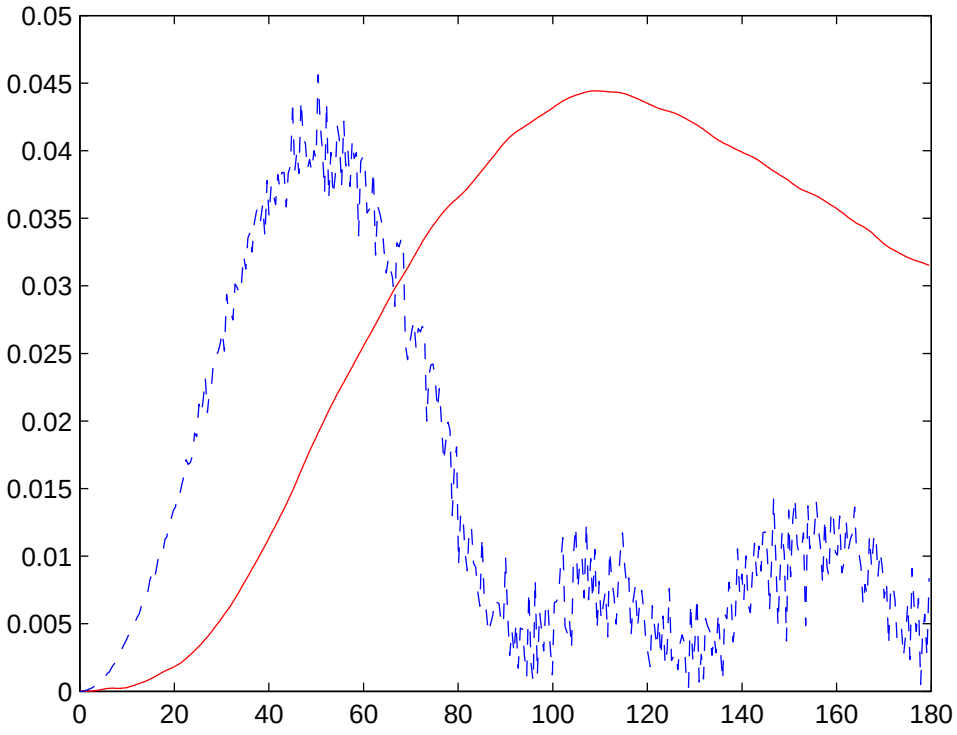

Figure 2: Shapes for $e(h, F, q)$ for (h, F) defined by (5) and for $\widehat{e}(h, F, a, q)$ with $a = 0.10$, $q \in [0, 50)$; ——: values of $e(h, F, q)$; $- - -$: values of $\widehat{e}(h, F, a, q)$.

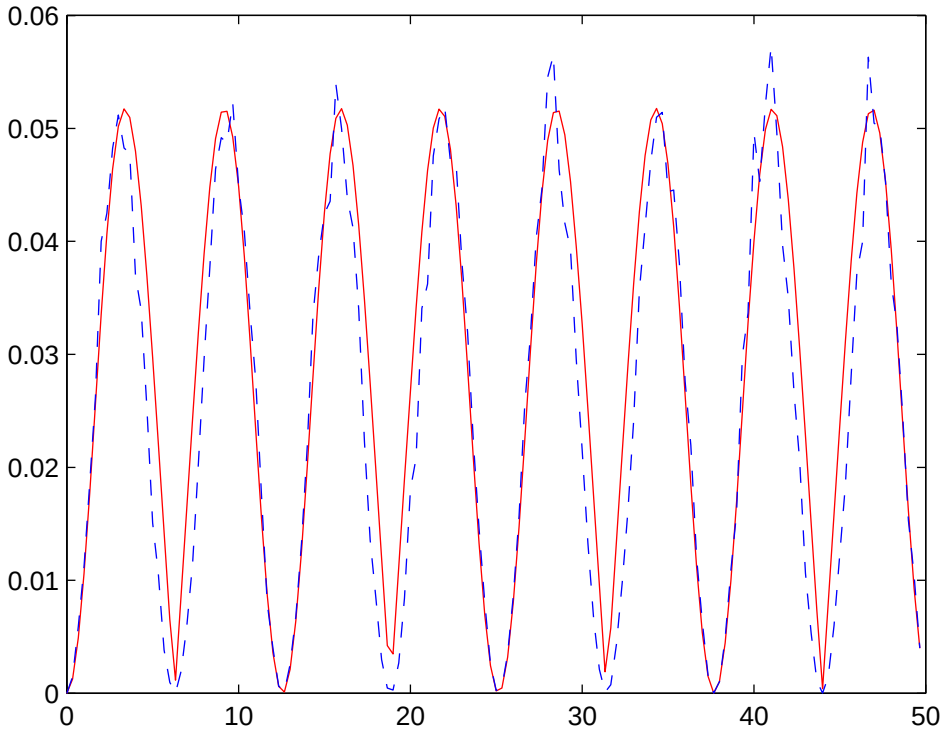

Figure 3: Shapes for $e(h, F, q)$ for (h, F) defined by (6) with $d = +\infty$ and for $\widehat{e}(h, F, a, q)$ with $a = 0.15$, $q \in [0, 50)$; ——: values of $e(h, F, q)$; $-\ -\ -$: values of $\widehat{e}(h, F, a, q)$.

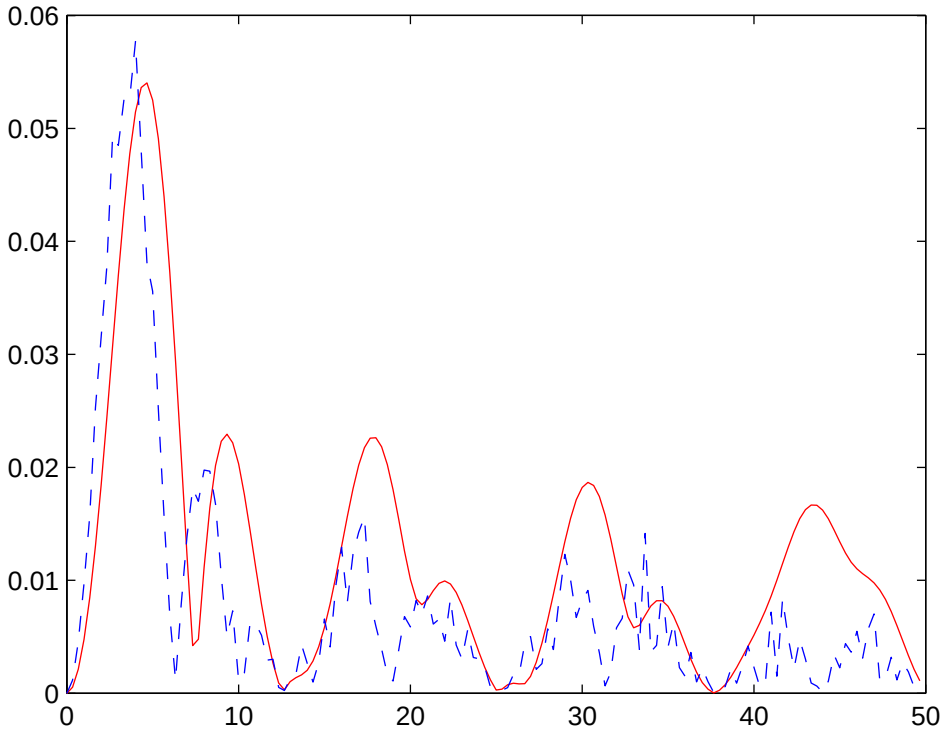

Figure 4: Shapes for $e(h, F, q)$ for (h, F) defined by (6) with $d = +\infty$ and for $\widehat{e}(h, F, a, q)$ with $a = -0.12$, $q \in [0, 50)$; ——: values of $e(h, F, q)$; $-\ -\ -$: values of $\widehat{e}(h, F, a, q)$.

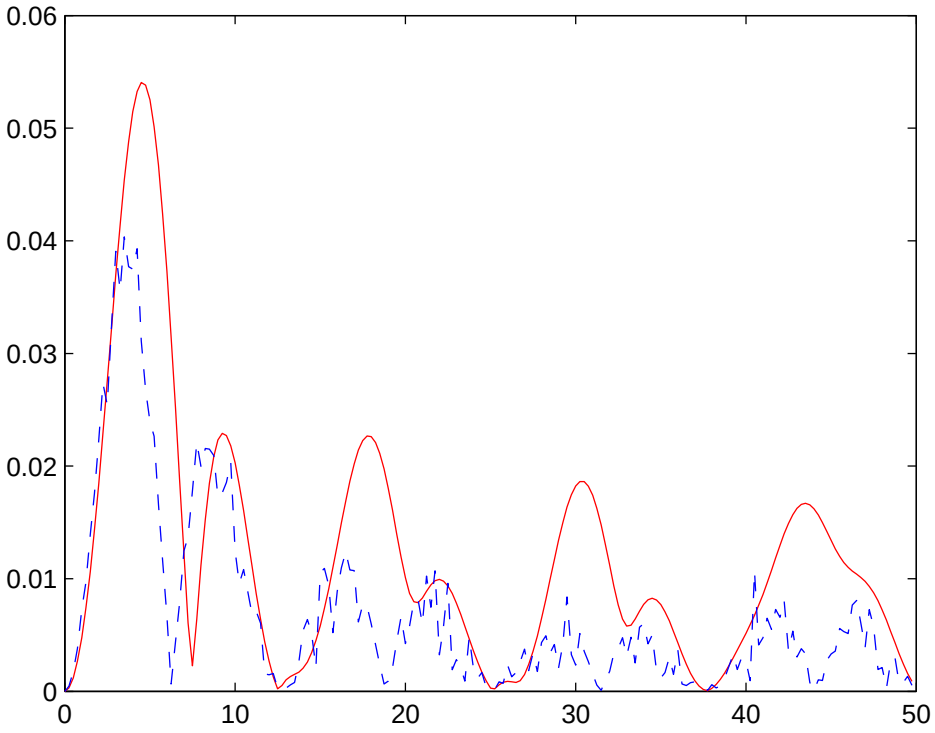

Figure 5: Shapes for $e(h, F, q)$ and for $\widehat{e}(h, F, a, q)$ with (h, F) defined for Choice 1 and for ARCH benchmark model with $a = 0.02$, $b = 1$, $c = 0.08\mathbf{E}\varepsilon_t^2$; ——: values of $e(h, F, q)$; $- - -$: values of $\widehat{e}(h, F, a, q)$.