How Does the Sample Size Affect GARCH Model?

H.S. Ng and K.P. Lam[1]

[1]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Abstract

GARCH model has a long history and permeates the modern financial theory. Most researchers use several thousands of financial data and maximum likelihood to estimate the coefficients of model. Statistically, more samples imply better estimation but are hard to obtain. How many samples are sufficient for estimation? What is the impact of the limited samples on the estimation? In this paper, we examined these questions using GARCH, MEM-GARCH models and NASDAQ composite index. The problems, raised from the limited samples, were discussed. Correlation of the conditional variances of the estimated models between the limited samples and the large samples were calculated. The effectiveness of model estimation for the limited samples was evaluated by the correlation.

Keywords: GARCH model, Multiplicative Error Model, Volatility.

1. Introduction

Volatility permeates the modern financial theory. GARCH model is widely acknowledged to estimate time varying and predictable volatility. Engle[1] originally proposed autoregressive conditional heteroskedasty (ARCH) model to provide a tool for measuring the dynamics of inflation uncertainty. The coefficient of ARCH model is estimated by maximum likelihood. Bollerslev[2] extended his work to Generalized ARCH (GARCH) model by developing a technique that allowed the conditional variance to be an Autoregressive Moving Average (ARMA) process. The coefficients of GARCH model are also estimated by maximum likelihood. A collection of survey articles[3][4][5][6][7][8] appreciates the aspect of this research. In 2002, Engle proposed a multiplicative error model (MEM) by considering a model of time series with non-negative elements [9]. The model specifies an error times the mean. He also estimated the coefficients of MEM-GARCH model by maximum likelihood.

In the past decades, the researchers focused on the explanation of economic phenomenon using GARCH model. Different formulations of GARCH model such as TGARCH [10], AGARCH [11], EGARCH [12] and FIGARCH [13] were proposed. Since the researchers normally used thousands of financial data for model estimation, the sample size normally was not their concerns. The rapid development of high frequency finance recently may face the problem of the insufficient financial data for model analysis. Therefore, it is interesting to study how the sample size affects this classical model. In this paper, we explore the impact of the sample size on GARCH model estimation through the empirical studies. How many samples are sufficient for model estimation? How does it affect the estimated variances?

A review on GARCH model, MEM-GARCH model and maximum likelihood is briefed in section 2. Estimation of both models using various sample sizes is given in section 3. Conclusion is drawn in section 4.

2. GARCH model

Conventional GARCH(1,1) was introduced by Bollerslev in 1986. It is the most famous model in modeling the conditional variance $\sigma^2_{t|t-1}$ of the daily return r_t among other GARCH model. The daily return refers to the difference of logarithmic close prices. The mean equation and the conditional variance equation are defined by

$$r_t = \mu + e_t \qquad (2.1)$$
$$\sigma^2_{t|t-1} = \omega + \alpha e^2_{t-1} + \beta\sigma^2_{t-1|t-2} \qquad (2.2)$$

The innovations e_t is interpreted as the multiplication of conditional variance and the Gaussian noise. The Gaussian noise follows identical independent distribution *(i.i.d)* with zero mean and unit variance. Maximum log-likelihood function in Eq. *2.3* is widely adopted for parameter set $\theta = \{\omega, \alpha, \beta\}$ estimation. The necessary constraints, which ensure that the conditional variances are positive, are $\omega \succ 0, \alpha \succ 0, \beta \succ 0$.

$$L(\theta) = -\sum[\log(\sigma_{t|t-1}^2) + \frac{e_t}{\sigma_{t|t-1}^2}] \qquad (2.3)$$

Owing to the limitation of the conventional GARCH model for handling non-negative time series, parameter estimation using maximum likelihood is difficult and inefficient[9]. Engle proposed a multiplicative error model by considering the non-negative series x_t. The mean equation and conditional variance equation are defined by

$$x_t = \sigma_{t|t-1}^2 \varepsilon_t \qquad (2.4)$$

$$\sigma_{t|t-1}^2 = \omega + \alpha x_{t-1}^2 + \beta \sigma_{t-1|t-2}^2 \qquad (2.5)$$

ε_t is a unit-mean *i.i.d* process. Eq. 2.4 thus specifies an MEM with "an error that is multiplied times the mean" [9]. The parameter set $\theta_m = \{\omega, \alpha, \beta\}$ can be obtained using the conventional GARCH software and maximum likelihood approach by simply taking the positive square root $\sqrt{x_t}$ as r_t and setting the dependent variable in the mean equation to zero.

GARCH model considers the additive error model while MEM-GARCH model considers the multiplicative error model. It is interesting to study the impact of the sample size on two different models.

3. Impact of sample size

NASDAQ composite index from *5* February *1971* to *20* Nov *1990* (in Fig. *1*) are used in this paper. There are totally *5000* data. The index price started at *100* on *5* February and ended at *348.3* on *20* November. The extreme values of index are *54.87* and *487*. The volatility of NASDAQ composite index was plotted in Fig. *2*. The peak volatility appears at the *4219*th day while the volatilities on the other days have a great fluctuation.

In this section, we studied the conventional GARCH(1,1) model and MEM-GARCH model using various sample sizes. Selection of appropriate model depends on the initial values of parameter set as they drive the estimation to different local optimal solutions. Carefully selection should be considered to avoid choose the wrong model. If the smallest sample size is *200*, we look into the relationship of the first *200* conditional variances of the estimated model among different sample sizes. For example, suppose we consider *200* and *3000* samples to estimate their models. The first *200* conditional variances and *3000* conditional variances for the first and second model respectively could be estimated. Since we have only 200 variances for the first model, only the first *200* conditional variances of both models are used to calculate their correlation. Intuitively, if the sample size is too small, the estimated conditional variances may be a noise only. Would *200* samples are sufficient for the estimation? If yes, what is the major difference of the estimated model between it and 3*000* samples? If no, how much number of samples should we use? Since we also consider MEM-GARCH model, would this model reduce the sample requirement?

3.1. Conventional GARCH model

The initial values of parameter set, which drove the optimal solution, were set in advance. Empirical result showed that if the sample size was fewer than *1000*, two optimal solutions might be found. For example, *600* samples of starting day of *1300* were used to verify this argument. The first and second initial values were set to {*0.500, 0.100, 0.500*} and {*0.045, 0.100, 0.100*} respectively. By taking the maximum log-likelihood, the optimal solutions of the parameter set θ were {*2.3219e-5, 0.2149, -0.3224*} and {*1.5755e-5, 3.2792e-3, 9.9825e-2*} respectively. These conditional variances of both models were used to calculate the correlation with those of model using *3000* samples and the same starting day. The correlation was *0.4378* and *0.5499* respectively. The second solution had higher correlation. Empirical result also showed that most of the initial values gave the first solution by maximum likelihood. To determine whether it is the correct solution, we check the condition of $\omega \succ 0, \alpha \succ 0, \beta \succ 0$. The first solution was rejected as β=*-0.3224<0*. Furthermore, we found that only a small change of the second initial values would make the solution dedicated to the wrong solution.

#	**200**	**300**	**400**	**500**	**600**
Correl	0.5478	0.9391	*0.9849*	0.9866	0.9805
#	**700**	**800**	**900**	**1k**	**1.1k**
Correl.	0.9810	0.9872	0.9830	0.9845	0.9815
#	**1.3k**	**1.5k**	**2k**	**3k**	
Correl	0.9813	0.9859	0.9987	1	

Table 1: Correlation of the conditional variances of GARCH model using the sample size between *x* and *3000*

Based on the selection criterion, we extended our study to the different sample sizes. They included *200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1500, 2000* and *3000*. The correlation of the conditional variances of estimated model using the sample size between *x* and *3000* were computed and tabulated in Table *1*. "#" represents the sample sizes while "Correl" represents the correlation. The correlation increased with the sample size up to *0.98*. When the sample size was *700*, the correlation reached *0.98*.

Instead of looking at the same starting day, we conducted further experiments with different starting day. The starting days were *{3, 300, 500, 800, 1k, 1.3k, 1.5k, 1.8k, 2k}*. For each starting day, we used the same number of sample sizes as those of the starting day *300* in the former experiment. The correlation of the conditional variances of the estimated model with the same starting day between different sample sizes and 3000 sample size were calculated and were plotted in Fig. 3. For each curve, it represented the same starting day and different sample sizes. There were totally *9* curves. The lowest curve was the data sets with the starting day *1300*. Citing this curve as an example, when the sample size was fewer than *700*, the correlation varied greatly. The correlation with sample size greater than *700* was higher than *0.86*. For the other curves, the correlation with sample size greater than *700* was higher than *0.86*. Such high correlation implied that even though the estimated model using *700* samples was as good as that using *3000* samples.

3.2. MEM-GARCH model

Similarly, we evaluated the MEM-GARCH model using the same starting values and the same sample sizes.

Regarding the multiple optimal solutions, empirical results showed that it happened comparatively less frequent for >700 sample sizes. Similarly, most initial values of parameter set directed to the wrong solution by maximum likelihood.

Regarding different sample sizes with the starting day *300*, the correlation of the conditional variances (in Table 2) increased with the sample size. However, the correlation for *200* samples size was lower than that of the conventional GARCH model. Unlike the conventional GARCH model, the correlation could reach 0.99 when *900* samples were used.

#	**200**	**300**	**400**	**500**	**600**
Correl	0.4983	0.9203	0.9805	0.9878	0.9956
#	**700**	**800**	**900**	**1k**	**1.1k**
Correl	0.9991	1.0000	0.9996	0.9980	0.9967
#	**1.3k**	**1.5k**	**2k**	**3k**	
Correl	0.9946	0.9953	0.9993	1.0000	

Table 2: Correlation between *x* samples and *3000* samples

Similarly, we conducted experiments using the different starting day and the same sample sizes as those in the conventional GARCH model. Their correlations were plotted with sample sizes in Fig. *4*. Unlike the conventional GARCH model, the correlation could reach *0.88* for *600* samples only. It further reached *0.95* for *1k* samples while that of the conventional GARCH model reached *0.90*. In general, the correlation of conditional variance for MEM-GARCH model is higher than that for the conventional GARCH model.

4. Conclusion

In conclusion, if the number of samples is less than *700*, two or more optimal solutions may be found by maximum likelihood. Also, most initial values direct to the wrong optimal solution. By examining the parameters' condition, the correct optimal solution can be identified. If the number of samples is more than *1000*, the correlation of conditional variances of estimated model between the limited samples and the large samples reaches the high values of 0.90. Thus, we recommend using *1000* samples for the model estimation for the conventional GARCH model. On the other hand, the sample size requirement for MEM-GARCH model is lower than the conventional GARCH model. Only 800 samples can give the comparatively high correlation of *0.90*.

Recently, the study of high frequency financial data plays important role in volatility study. The study of this paper supports the reduction of data requirement. However, the study on the intraday effects on volatility index is pending for the sufficient data.

5. References

[1] R. F. Engle, ``Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, Vol. 50, pp. 987-1007, 1982.

[2] T. Bollerslev, "Generalized Autoregressive Conditional Heteroscedasticity," *J. of Econometrics*, Vol. 32, pp 307-27, 1986.

[3] A. Bera and M.L. Higgins, "ARCH Models: Properties, Estimation and Testing," *J. of Economic Surveys*, Vol. 7, pp305-66, 1993.

[4] T. Bollerslev, R.Y. Chou, and K.F. Kroner, "ARCH modeling in Finance: A Review of the Theory and Empirical Evidence," *J. of Econometrics*, Vol. 52, pp 5-59, 1992.

[5] T. Bollerslev, R.F. Engle and D.B. Nelson, "ARCH Model," *Chapter 49 of Handbook of Econometrics*, Vol. IV, Ed. Engle and McFadden, Elsevier Science, pp. 2291-3031, 1994.

[6] R. Engle, "GARCH 101: An Introduction to the Use of ARCH/GARCH models in Applied Econometrics," *J. of Economic Perspectives*, 2001.

[7] R. Engle and A. Patton, “What Good is a Volatility Model?” *Quantitative Finance*, pp 237-45, 2001.

[8] S. Ling and M. McAleer, “Stationarity and the existence of moments of a family of GARCH processes,” *J. of Econometrics*, Vol. 106, pp. 109-27, 2002.

[9] R.F. Engle, “New Frontiers for ARCH Models,” *J. of Applied Econometrics*, Vol. 17, pp 425-46, 2002.

[10] L. R. Glosten, R. Jagannathan, and D.E. Runkle, “On the relation between the expected value and the volatility of the nominal excess return on stocks,” *J. of Finance*, Vol. 48, pp 1779-801, 1993.

[11] R.F. Engle and V. K. Ng, “Measuring and Testing the Impact of News on Volatility,” J. of Finance, Vol. 48, pp 1749-78, 1993.

[12] D.B. Nelson, “Conditional Heteroskedasticity in Asset Returns: A New Approach,” Econometrica, Vol 59, pp. 347-70, 1991.

[13] R. T. Baillie, T. Bollerslev and H.O. Mikkelsen, ”Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity,” *J. of Econometrics*, Vol. 74, pp. 3-30, 1996.

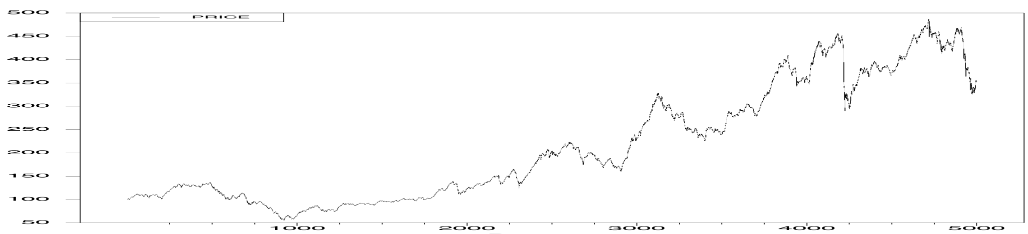

Fig. 1: NASDAQ Composite index during 5 February 1971 to 20 November 1990

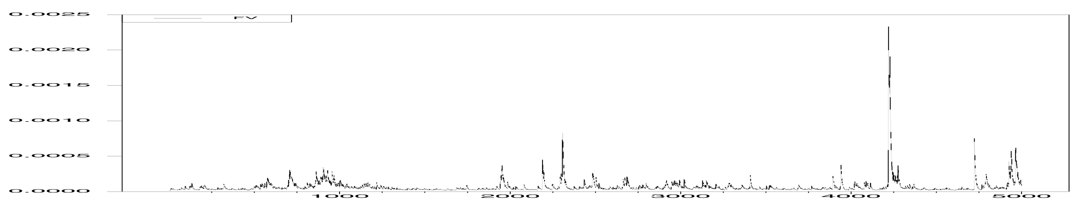

Fig. 2: Conditional variance of the conventional GARCH model during 5 February to 20 November 1990

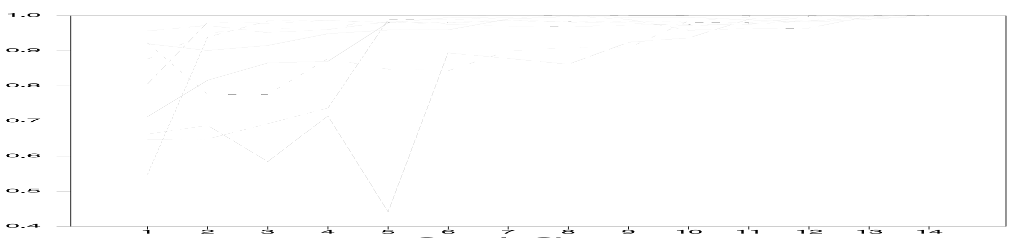

Fig. 3: Correlation of conditional variances for the GARCH mode using various sample sizes and starting days

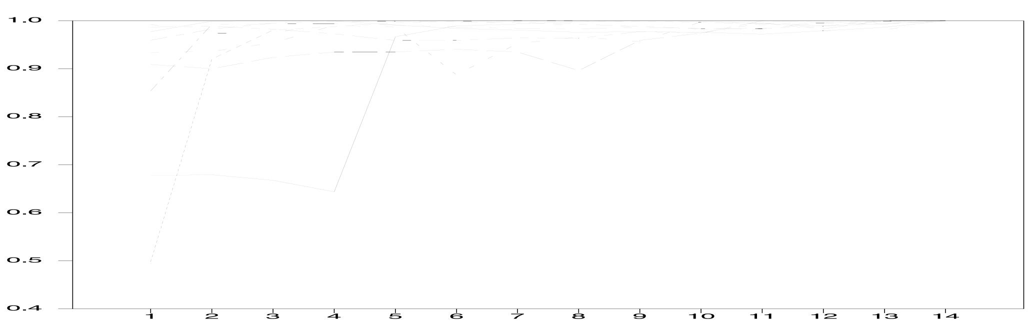

Fig. 4: Correlation of conditional variances for MEM-GARCH model using various sample sizes and starting days