Introduction to (Generalized) Autoregressive Conditional Heteroskedasticity Models in Time Series Econometrics

Bryant Wong

June 2014

Contents

1 Introduction & Motivation

My paper is an examination and application of the ARCH/GARCH models proposed in the 1980's by econometricians such as Robert Engle (who won the Nobel Prize for Economics in 2003 for this work), Tim Bollerslev (one of Engle's PhD students at the time). In particular, we focus on the paper, "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics" by Robert Engle [3], with some supplementation from "Generalized Autoregressive Conditional Heteroskedasticity" by Tim Bollerslev [1].

Since the introduction of ARCH/GARCH models in econometrics, it has widely been used in many applications, especially for volatility modeling. There are many derivatives of ARCH/GARCH used for different applications and different sets of data, etc. Though these days is has largely been superceded by stochastic volatility models in academia, ARCH/GARCH models still have great value and continue to be used heavily in industry and finance.

2 Preliminary Ideas

In order to understand anything about these sort of models, we need to first consider some basic principles from statistics/econometrics.

2.1 Ordinary Least Squares

Acting as the backbone of a large portion of statistics, the sciences, and quantitative analysis in the humanities is the humble Ordinary Least Squares (OLS), which is often called linear regression. Economists seek to derive (pseudo)linear relationships between some number of variables, because this makes the dynamic between them clear - a change in this parameter results in some change in another parameter. As a result of this simplicity in relationship, there is strong predictive and explanatory power in our model.

Of course we know that given any set of data, it would be possible to perfectly fit every single point to some extremely complex function, but this sort of function has no value because it offers little to no explanatory or predictive power. There is simply no way to come up with a coherent relationship between two variables if your equation is some convoluted polynomial with trigonometric functions, etc. Hence, econometricians and other scientists focus on coming up with these simple linear relationships.

The way OLS works is given some set of observations with n parameters: $\{X_{1,i}, X_{2,i}, \cdots, X_{n-1,i}, Y_i\}$ where we designate each X_j to be an "independent" parameter and Y_i is considered to be the "dependent" parameter - that is in the true, but unknown relationship, $Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_{n-1} X_{n-1,i} + u_i$, where β_i is constant and u_i is considered to be an disturbance term. However, since it's impossible to know the true relationship, we try to fit our observations to be close to the real relationship, which is in the form $\hat{Y}_i = b_0 + b_1 X_{1,i} + \cdots + b_{n-1} X_{n-1,i}$. So how do we get b_i close to β_i?

We define the notion of a residual, where $e_i = Y_i - \hat{Y}_i$ and try to minimize the sum of the squares of the residuals (Residual Sum of Squares), i.e. minimize $\sum_{i=1}^{n} e_i^2$ with respect to $b_0, b_1, \cdot, b_n$. Minimizing $\sum_{i=1}^{n} e_i$ is useless since this causes a negative residual to cancel with a positive residual, so we can could just set $b_0 = \overline{Y}$ and $b_1, \cdots, b_n = 0$ and that would be considered a "good fit", since that makes the sum of the residuals 0. Minimizing $\sum_{i=1}^{n} |e_i|$ would be feasible, but when we minimize we want to take derivatives and doing this with an absolute value is difficult (though very possible). Hence, we choose to minimize the sum of the squares.

2.1.1 Sample Case with one Explanatory Variable

In the sample case of one explanatory variable with n observations, our model is in the form $\hat{Y}_i = b_1 + b_2 X_i$. Then, we wish to minimze $\sum_{i=0}^{n} e_i^2$ with respect to b_0 and b_1. So simply apply partial derivatives: $\frac{\partial RSS}{\partial b_0} = 0$ & $\frac{\partial RSS}{\partial b_1} = 0$, where RSS is given as $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$. This yields the two following equations:

$$\sum_{i=1}^{n} Y_i - b_0 - b_1 X_i = 0$$
$$\sum_{i=1} X_i (Y_i - b_0 - b_1 X_i) = 0$$

Since the mean of X, $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and similarly, the mean of Y $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$, we can rewrite the first equation in the form:

$$\begin{aligned}
\sum Y_i - \sum b_0 - b_1 \sum X_i &= 0 \\
\frac{1}{n} \sum Y_i - \frac{1}{n} \sum b_0 - \frac{b_1}{n} \sum X_i &= 0 \\
\bar{Y} - b_0 - b_1 \bar{X} &= 0 \\
b_0 &= \bar{Y} - b_1 \bar{X}
\end{aligned}$$

Now solving for b_1, substituting our result for b_0:

$$
\begin{aligned}
\sum X_iY_i - b_0\sum X_i - b_1\sum X_i^2 &= 0\\
\sum X_iY_i - (\bar{Y} - b_1\bar{X})\sum X_i - b_1\sum X_i^2 &= 0\\
\sum X_iY_i - \bar{Y}\sum X_i + b_1\bar{X}\sum X_i - b_1\sum X_i^2 &= 0\\
\sum X_iY_i - n\bar{Y}\bar{X} + nb_1\bar{X}^2 - b_1\sum X_i^2 &= 0\\
b_1\left(n\bar{X}^2 - \sum X_i^2\right) &= n\bar{Y}\bar{X} - \sum X_iY_i\\
b_1 &= \frac{\sum X_iY_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2}
\end{aligned}
$$

This method generalizes to however many explanatory variables there are in our equation. However, in cases with large numbers of explanatory variables, it is much quicker to solve via linear algebra methods rather than with algebraic methods [2].

2.2 Heteroskedasticity

Ordinary Least Squares works great (assuming we meet some preliminary conditions), but one assumption that must be made for OLS to work is that the disturbance terms, u_i are homoskedastic - that is, the variance of each disturbance term is the same ($\sigma^2_{u_i} = \sigma^2_u \quad \forall i$) However, this is not always a very realistic assumption in real life, since variance is not necessarily always constant. For example, consider the case where a researcher examines the relationship between income and consumption in households. They would likely find that consumption is more closely tied to income in low-income households rather than higher ones, since savings/deficit is likely to be much smaller in absolute value for those households. Then, the variance of those households with higher incomes appears to be much higher, so variance is not constant across the sample. [3].

The problem with this is then when running statistical tests and such, we weight each data point equally, despite the fact that some will vary from the true model more than others. This makes any statistical analysis inaccurate since our confidence intervals and standard errors will end up being too small, so we end up thinking we have more precision than we actually have.

There exist methods of dealing with heteroskedasticity such as White's test, etc. in order to produce better statistical measure and estimates, but instead what ARCH/GARCH models do is take this notion of heteroskedastic disturbance terms and treat them as their own data set to be modeled. Much like how we tried to fit a a straight line to our intial set of straight lines, ARCH/GARCH models aim to find a model for the disturbance terms. ARCH/GARCH utilizes heteroskedasticity as a parameter to be modeled, where the variance of these disturbance terms is often considered to be the volatility or risk of some asset.

2.3 Time Series Modeling

One small thing to consider is that ARCH/GARCH models are only used in time-series econometrics, so we can only apply this with time series data. When working with cross-sectional data (eg. comparing inflation to unmployment), often times there isn't an implicit ordering, but in time series data there is only one order the data can be worked with. Another thing is that time-series phenomena is generally considered to be continuous, which can make things tricky since we can only really work with discrete data sets. Finally, unlike in cross-sectional data where we can sometimes make the naive assumption that regressors are drawn randomly from populations, in time-series models, it is almost always the case that there is some level of correlation between a regressor at time t and $t+1$, simply due to the nature of time series [2].

3 Building Blocks

ARCH/GARCH models can basically be considered to be a composition of several simpler models. We first need to gain an understanding of these models in order to begin comprehending ARCH/GARCH. Here we can understand why ARCH/GARCH only applies to time series data - these models utilize lagged terms, which only makes sense in the context of time-ordered data.

3.1 AutoRegressive models

The generalized AR(p) model uses p lag variables which can be written in the form:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t$$

The basis behind the AR model comes from the idea that the output/dependent variable is a linear function of its previous values as lag variables. The easiest way to understand this is via example: the simplest case of an AR model is AR(1): $Y_t = c + \phi_1 Y_{t-1} + \epsilon_t$, where ϕ_1 is constant and ϵ_t is the error term at time t, which we consider to be white noise. White noise can be considered to be some independent identically distributed random distribution centered around 0. Commonly used is a Gaussian white noise distribution, which is basically a normal distribution with mean 0.

Technically the simplest AR model is AR(0) which is just $Y_t = c + \epsilon_t$, which is basically just white noise centered around c, but this doesn't give us any intuition about how the model should actually look.

This model can be seen as a case of OLS - we can determine/solve for our constants c and ϕ_i by using the OLS procedure detailed above. We just treat ϵ as white noise, so we don't have to do anything with that. This sort of model is valuable in financial applications, where the information used to predict the value of some asset is heavily based on the prior values of the asset in earlier time periods, from the information set. For example, in economics we make the assumption that the stock price on one day is extremely correlated to the price of the stock from the day before.

3.2 Moving Average models

The generalized MA(q) model uses q lag error terms which can be written in the form:

$$Y_t = d + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

Unlike AR models, moving average models utilize "past" error terms in order to forecast future terms. Of course, this is not a true regression model, because each ϵ_t is not actually known (again, we consider these ϵ terms to be white noise or random shocks). However, we can write a mathematical representation of the MA model, which will be useful later on in ARMA/ARIMA/ARCH/GARCH models. The simplest (nontrivial) case of MA is MA(1), which can be written in the form $Y_t = d + \epsilon_t + \theta_1 \epsilon_{t-1}$, where θ_1 is constant and each ϵ is a white noise term.

Again, this model is extremely valuable in financial applications, where we can think of the price of some asset being affected by a sum of stochastic shocks over time, again from the information set. However, unlike in the AR model, we cannot simply just apply OLS to solve for the θ coefficients, since each ϵ term is completely unknown. The method of solving MA coefficients involves solving a system of nonlinear equations, so most of the time we just use software to compute this since it is entirely numerical and has no bearing on the theory.

3.3 ARMA/ARIMA models

The generalized ARMA(p, q) with p autoregressive and q moving average parameters can be written in the form:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_i \epsilon_{t-j}$$

The ARMA model is derived by combining autoregressive terms and moving average terms to create a more complete model (AutoRegressive + Moving Average = ARMA). A very simple case of ARMA is ARMA(1, 1): $Y_t = c + \epsilon_t + \phi_1 Y_{t-1} + \theta_1 \epsilon_{t-1}$, with one autoregressive term and one moving average term. A possible intepretation of this model could be Y_t being the price of some asset at time t, which is a function of the price of the asset at time $t-1$ (Y_{t-1}), a random shock at time t, (ϵ_t), random shock at time $t-1$, (ϵ_{t-1}) along with a constant c.

In general, p and q are not large because 1) the coefficients are likely to get small and not statistically significant with too many lag terms, 2) the interpretations can get difficult with such large models, and 3) With too many terms, we lose our predictive power due to overfitting. Overfitting is the case where we have too many parameters and end up modeling the random noise rather than the actual underlying relationships.

We can consider ARIMA to be a further generalization of ARMA. However, to understand this, we need to address the topics of stationarity, differencing, etc.

3.3.1 Stationarity

ARMA are time series models for stationary data - that is despite our data being stochastic, the probability distribution of our data remains constant. Any time series data with trends or seasonality (regular cycles) cannot be considered to be stationary. In very simple terms, we want our data to look roughly similar at any point in time. Time series with cyclic behavior can actually be stationary, as long as it is non regular (cycles w/o fixed length), so at some point in time it's impossible to know if you are at some sort of peak or trough. Most of the time stationary data stays relatively flat, with constant variance due to the fact that the probability distribution is constant.

3.3.2 Differencing

However, a lot of useful time series data is non-stationary - eg. stock indices like the Dow Jones experiences obvious trends over periods of time, some phenomena experiences regular literal seasonal changes, such as the cost of heating oil. So one way we can still work with non-stationary time series data is with differencing. Since we are working with discrete data, there isn't exactly a notion of derivative, so we just define $Y_t' = Y_t - Y_{t-1}$. (Minor detail, given n observations, our differenced data will have $n-1$ observations). Differencing allows us to potentially stabilize the mean of our time series so that we can remove trends and seasoning. For example, our Dow Jones data might have trends over some period of time, but on a day to day basis, the change in the Dow Jones is very likely to be centered at 0. Along with methods like taking logarithms to normalize variances, we can turn non-stationary time series data into stationary time series data that we can work with

Of course, differencing will not always work, and so we may have to reiterate the process. For example, the second differencing is given by $Y_t'' = Y_t' - Y_{t-1}'$, which can be generalized to anything we want. However, in basically every real world case, the maximum order we would ever go to is 2, because we lose explanatory power when going to the 3rd derivative.

3.3.3 ARIMA

The generalized ARIMA model is hard to write because differencing is hard to capture explicitly, but using our previous notation, we can write the general ARIMA(p, d, q) model has p autoregressive terms and q moving average terms, with d degree of differencing in the form:

$$Y_t^{(d)} = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_i \epsilon_{t-j}$$

The purpose of ARIMA (AutoRegressive Integrated Moving Average) models is to generalize ARMA models to analysis non-stationary time series data, by combining differencing, moving average terms and autoregresive terms. We can also think of AR, MA, and ARMA all as special cases of the more general ARIMA model. ARIMA models are extremely useful in time series econometrics and statistics and have a variety of applications [2].

4 ARCH/GARCH models

However, the question is not always whether a model is a good fit for our data, but sometimes we wish to consider the accuracy of the model itself and whether its predictions are valid or not. One method of testing the accuracy is to look at the variance of our error terms.

Before the ARCH model was introduced by Engle in 1982, the way that econometricians described variances of models was to use a rolling standard deviation, where one could equally weigh all the observations the standard deviation over some number of previous observations, ie:

$$\sigma_{u_{t+1}} = \frac{1}{n}\sum_{i=0}^{n} \sigma_{u_{t-i}}$$

A good way to think about ARCH is to think of it as a generalization of this formulation - instead of weighting each value equally, ARCH treats the weights as parameters to be estimated. This is more realistic because 1) Assuming equal weights seems inaccurate since we presume that more recent observations are more likely to be more relevant, and 2) Restricting our weights only to some finite number of observations is not ideal.

We can write the general ARCH(m) model of order m as:

$$u_t^2 = c + \left(\sum_{i=1}^{m} \alpha_i u_{t-i}^2\right) + w_t$$

where w_t is some white noise process [4]. So we see that ARCH is essentially the combination of an AR and MA model but applied to disturbance terms. Unfortunately I don't have enough background to go much further with the math behind ARCH, but it gives us the capability to measure a lot of interesting data.

The generalization of ARCH to GARCH is analogous to the generalization of ARMA to ARIMA - GARCH basically says that the best predictor of the variance in the next time period is given by a weighted average of the long-run average variance, the variance predicted in this period by (G)ARCH, and new information given in this period [3]. Writing the generalized GARCH model down is difficult without introducing new notation, and it will not add any value to the paper, so we skip it here. For further details, please look into the paper by Tim Bollerslev [1].

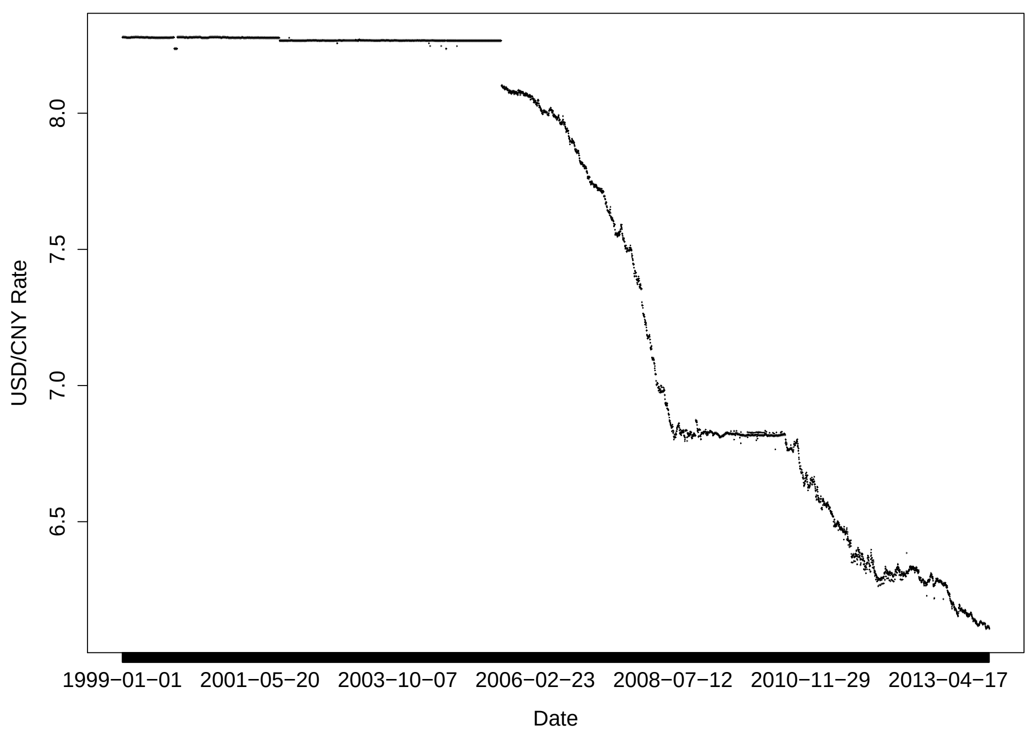
4.1 Sample Application

Most of the rest of Engle's paper focuses on various examples, so I will be presenting my own. In this example, I am working with exchange rate data between the USD and CNY from 1984 - 2014, and trying to find a model for the volatility of this data. I will be using the GARCH(1,1) model which is one of the most commonly used models. The reason for this choice is mostly because this is simpler and easier to work with than a GARCH model with many lagged and autoregressive terms, but also has some modeling power.

The GARCH(1,1) model can be expresed as $h_t = \omega + \alpha(r_{t-1} - \mu)^2 + \beta h_{t-1}$, where h_t is the variance of r_t, the return at time t conditional on all prior returns.

Here is what the data looks like - there are approximately 5480 observations so it's an extremely large set:

Plot 1: USD/CNY Exchange Rate from 1999−2014



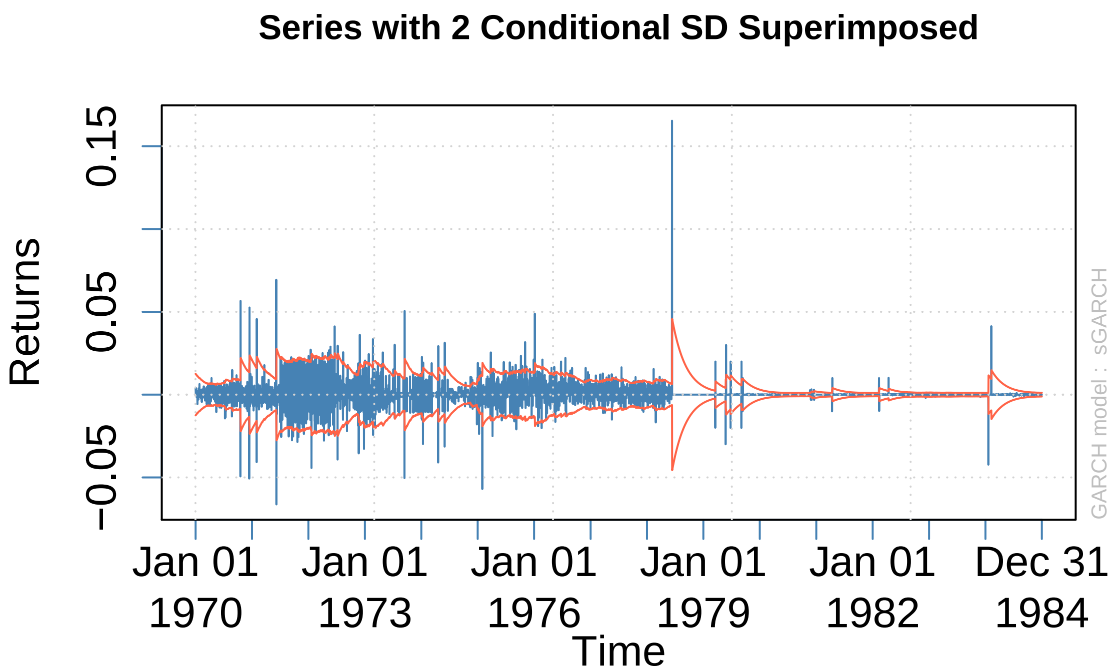
For this data analysis, I am using a package in the statistical software R, which is called rugarch. Here is the code I used to get my volatility model:

```
> spec <- ugarchspec(variance.model = list(model = "sGARCH",
garchOrder = c(1, 1)), mean.model = list(armaOrder = c(0, 0),
include.mean = FALSE), distribution.model = "norm")
> modelfit = ugarchfit(spec, exRatesDiff, solver = 'hybrid')
> modelfit
```

```
*---------------------------------*
*          GARCH Model Fit        *
*---------------------------------*

Conditional Variance Dynamics
-----------------------------------
GARCH Model : sGARCH(1,1)
Mean Model : ARFIMA(0,0,0)
Distribution : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error    t value Pr(>|t|)
omega   0.000000    0.000000 4.2788e-02  0.96587
alpha1  0.018674    0.000238 7.8447e+01  0.00000
beta1   0.978236    0.000351 2.7909e+03  0.00000

Robust Standard Errors:
        Estimate  Std. Error  t value Pr(>|t|)
omega   0.000000    0.000653 0.000008 0.999994
alpha1  0.018674    0.061960 0.301385 0.763121
beta1   0.978236    0.329942 2.964876 0.003028

LogLikelihood : 23621.16
....
> mean(exRatesdiff)
[1] 0.0003964775
```

Some of the output is deleted since it's not particularly relevant. What is good is that we get that omega is 0, which means that our data is probably differenced correctly since we're getting the average variance (returns) to be 0, which is exactly what we want. So our model looks something like $h_t = 0.018674(r_{t-1} - 0.0003964775)^2 + 0.978236h_{t-1}$. Some small interpretations we can do is that there is a stronger effect from the prior day's variance than from prior returns.

Attached is a plot of what the model looks like. The dates are all wrong, but the shape is what we want. However, there are some problems with this model because despite mathematically working out, politically we know that in the "flat" periods of our plot, China fixed its currency to the USD, which throws off our model completely. In addition, both currencies are constantly manipulated by each countries government, so it's very difficult to actually apply the model since we don't know whether the exchange rates are truly random or not. However, in some other applications, we might get some better results.



5 Conclusions and Analysis

In this paper, I have barely touched upon the surface of volatility modeling. ARCH and GARCH are the most simple of a whole family of models based off of AR/MA models, which are all optimized to perform best under many different conditions. So ARCH/GARCH are definitely not the only or even best models for this sort of modeling. However, what they are useful for are getting a quick estimate on the data we are modeling. Since a lot of industry needs data to be processed extremely quickly and computationally easily, ARCH/GARCH still see common use in those sort of applications. However, with today's computational power, most academic econometricians have moved on to stochastic volatility models, which produce "better" results based on a variety of metrics [6].

References

[1] Tim Bollerslev. "Generalized Autoregressive Conditional Heteroskedasticity" *Journal of Econometrics* 31 (1986): 307-327

[2] Christopher Dougherty. *Introduction to Econometrics* Oxford University Press, New York, 4th Edition, 2011

[3] Robert Engle. "GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics." *Journal of Economic Perspectives* 15.4 (Fall 2011): 157-168

[4] James D. Hamilton. *Time Series Analysis* Princeton University Press, Princeton, New Jersey, 1st Edition, 1994

[5] Roberto Perrelli. *Introduction to ARCH and GARCH models*. University of Illinois, 2001

[6] Stephen J. Taylor. *Asset Price, Dynamics, Volatility, and Prediction*. Princeton University Press, Princeton, New Jersey, 1st Edition, 2005.