A semi-parametric GARCH (1, 1) estimator under serially dependent innovations

Cassim, Lucius

Economics Department, University of Malawi

5 May 2018

A semi-parametric GARCH (1, 1) estimator under serially dependent innovations

Lucius Cassim[1]

May 5 2018

Abstract

The main objective of this study is to derive semi parametric GARCH (1, 1) estimator under serially dependent innovations. The specific objectives are to show that the derived estimator is not only consistent but also asymptotically normal. Normally, the GARCH (1, 1) estimator is derived through quasi-maximum likelihood estimation technique and then consistency and asymptotic normality are proved using the weak law of large numbers and Linde-berg central limit theorem respectively. In this study, we apply the quasi-maximum likelihood estimation technique to derive the GARCH (1, 1) estimator under the assumption that the innovations are serially dependent. Allowing serial dependence of the innovations has however brought problems in terms of methodology. Firstly, we cannot split the joint probability distribution into a product of marginal distributions as is normally done. Rather, the study splits the joint distribution into a product of conditional densities to get around this problem. Secondly, we cannot use the weak laws of large numbers or/and the Linde-berg central limit theorem. We therefore employ the martingale techniques to achieve the specific objectives. Having derived the semi parametric GARCH (1, 1) estimator, we have therefore shown that the derived estimator not only converges almost surely to the true population parameter but also converges in distribution to the normal distribution with the highest possible convergence rate similar to that of parametric estimators

Key Words: GARCH(1,1), semi parametric , Quasi Maximum Likelihood Estimation, Martingale

[1] Economics Department, Chancellor College, University of Malawi, P.O.Box 280, Zomba, Malawi : Email: luciuscassim@gmail.com or lcassim@cc.ac.mw

1. INTRODUCTION

1.1 Background of the study

Recently, volatility modelling has been a very active and extensive research area in empirical finance and time series econometrics for both academics and practitioners (Chung, 2012). Chung (2012) argued that volatility is important in investment, security valuation, pricing derivatives, calculating measures of risk, and hedging against portfolio risk. The simplest way to estimate volatility would be to use standard deviation, which is time-invariant (Chung, 2012). However, Tsay (2010) showed that this approach is contradicted by empirical evidence. According to Tsay (2010),there are some financial data empirical regularities that violate the constant volatility assumption which include: (i) existence of volatility clusters; (ii) evolution of volatility over time in a continuous manner; (iii) variability in volatility within some fixed range; and (iv) difference in volatility reaction to big increases and decreases, which is known as the leverage effect. As a consequence, volatility has been modelled as a time-dependent variable and not as a time-invariant standard deviation.

There are several approaches to modelling time-dependent volatility in financial literature. The first is deterministic approach. This approach models volatility as

conditional variance expressed as a function of lagged conditional variance and lagged squared innovations. Deterministic models come under parametric, semi-parametric or non-parametric sub- approaches depending on the assumptions about the structure of the volatility. At one extreme, parametric models make explicit both the functional form of the volatility model, while assuming a specific probability distribution of the innovations. The most popular parametric structure is the Generalized Autoregressive Conditional Heteroskedaticity (GARCH here-in-after) (see Hansen & Lunde, 2001) and many variants of GARCH type models that have been proposed in the literature (see Taylor, 1986; Engle & Ng., 1993; Zakoian, 1994; Glosten & Runkle, 1993; Geweke, 1986; Pantula, 1986; Higging & Bera, 1992; Sentana, 1995; Hentshel, 1995; Duan, 1997) where volatility is a deterministic function of its own lagged values and lagged squared innovations. At the other extreme, nonparametric approach makes no specification of the volatility model and no explicit assumption of the probability distribution of the innovations. It lets the data guide the process (Buhlman & McNeil, 2000). Semi parametric estimation approaches are a hybrid of the two extremes. Under this, the volatility model is explicitly specified but the distribution of the innovations is left unspecified.

The second approach is stochastic volatility (SV) approach. This is where an innovation (a stochastic component) is added to the deterministic component of volatility. The most popular stochastic volatility model is the one proposed by Shephard (2008). According to Shephard (2008), adding an innovation substantially increases the model flexibility in describing the evolution of volatility but it increases the difficulty in parameter estimation. Shephard (2008) showed that SV models do not perform any better than deterministic models. Shephard (2008) claimed that deterministic modelling is the most popularly used approach in literature, even though recently stochastic volatility approach has been gaining ground. It has been shown that stochastic volatility models produce no statistically better results over the deterministic approach (see Shephard, 2008).

Among the sub-approaches of deterministic modelling, the parametric approach is used the most in financial literature (Chung, 2012). Just like any estimation technique, it has its pros and cons. The main advantage of using parametric approach is that the estimators converge at higher rate than those of non-parametric approach, though almost at the same rate as semi-parametric approach (Yang & Song, 2012). However, the parametric approach suffers from a high risk of getting inconsistent estimators if the model and/or the probability distribution of the innovations are not correctly specified. Using semi-parametric/non-parametric approaches increases the flexibility of the model as they do not impose strict assumptions on the model specification/probability distribution. This flexibility comes at the expense of low rate of convergence of the estimators asymptotically in the case of non-parametric approach. Other than that, non-parametric models do not perform any better than the parametric approach unless there is leverage effect (Buhlman & McNeil, 2000). In fact, Buhlman and McNeil (2000) stressed that non-parametric models should only be used when there is evidence of leverage effect because that is the only time that they perform better than the parametric models. This is

why non-parametric approaches are not highly used to estimate volatility in financial literature.

As we have already stated earlier, semi parametric approach is a hybrid of parametric and non-parametric approaches. Just like the non-parametric approach it introduces flexibility in the model while producing estimators that converge almost at the same rate (very high speed of convergence) as the parametric approach (Yang & Song, 2012). One would therefore expect semi parametric approach to be extensively used and developed in the literature. Surprisingly, very few theoretical papers (e.g. Linton & Mammen, 2003; Drost & Klassenn, 1996; Yang & Song, 2012; Engle & Gonzale-Rivera,1991) have employed semi-parametric approach though it produces estimators that converge at the same rate as the parametric estimators while being more flexible (Buhlman & McNeil, 2000).

To the contrary, parametric deterministic models have been extensively developed (e.g. the ARCH model of Engle(1982); the GARCH model of Bollerslev(1986); the I-GARCH model of Taylor(1986); the T-GARCH model of Zakoian(1994); the H-GARCH model of Hentshel(1995); the TS-GARCH, the NA-GARCH and the V-GARCH models of Engle & Ng(1993); and the Aug-GARCH model of Duan(1997)). These are theoretical papers that have been trying to extend and develop the parametric deterministic modelling further. There are also a lot more empirical papers that have applied these models but we will not mention them here since our interest is theoretical not empirical in this study.

GARCH models define the time-varying variance as a deterministic function of past squared innovations and lagged conditional variances (Bollerslev, 1986). That is to say, in general terms, a GARCH model with order $p(\geq 1)$ and $q(\geq 1)$ is defined as;

$$y_t = f(x_t, \beta) + \varepsilon_t \tag{1.01}$$

$$\varepsilon_t = z_t \sigma_t$$

$$\sigma_t^2 = \vartheta + \sum_{i=1}^{p} \alpha_i L^i \varepsilon_t^2 + \sum_{i=1}^{q} \beta_i L^i \sigma_t^2 \tag{1.02}$$

Here, ϑ, α, β are non-negative parameters, L is a lag operator, x_t are factors affecting y_t , ε_t are innovations, σ_t is the conditional standard deviation of the innovations and z_t are independent standardized innovations. It should be mentioned here that equation (1.01) is called conditional mean equation while equation (1.02) is called conditional variance equation. This technically means that GARCH (1, 1) is defined as;

$$y_t = f(x_t, \beta) + \varepsilon \tag{1.03}$$

$$\varepsilon_t = z_t \sigma_t$$

$$\sigma_t^2 = \vartheta + \alpha_1 L \varepsilon_t^2 + \beta_1 L \sigma_t^2 \quad (1.04)$$

GARCH (1, 1) model is popularly used in parameterization of volatility in financial literature (Hansen & Lunde, 2001). This is partly because of their simple specification and interpretability (Chung, 2012; Hansen & Lunde, 2001). Additionally, although the model doesn't take into account the leverage effect, it outperforms other volatility models, in terms of volatility predictive power (see Hansen & Lunde, 2001; and White, 2001). In fact, Hansen and Lunde (2001) compared a total of about 330 volatility models and GARCH (1, 1) outperformed all of them in terms of ability to predict volatility. Parametric estimation of the GARCH (1, 1) is mostly done by using nonlinear maximum (or quasi-maximum) likelihood estimation based on two key assumptions (Bollerslev, 1986; Chung, 2012; Buhlman & McNeil, 2000; Drost & Klassenn., 1996); (1) that the innovations have a specific known density law, mostly the normal distribution and (2) that the innovations are independently and identically distributed (here-in after referred to as i.i.d).

1.2 Problem statement

As we have seen in section 1.1, GARCH (1, 1) is the most used volatility model in financial literature and parametric approach is the most used approach in estimating GARCH (1, 1) model in literature. The parametric estimation of GARCH (1, 1) model is based on the following two key assumptions. Firstly, the innovations have a known distribution e.g. mostly, normal distribution or, recently, the student t-distribution (Rossi., 2004; Bollerslev, 1987; Gallant & Hsieh, 1989; Baillie & Bollerslev, 1987). Secondly, the innovations are i.i.d (Rossi., 2004; Bollerslev., 1987; Gallant & Hsieh, 1989; Baillie & Bollerslev, 1987). On the one hand, the i.i.d assumption technically means that the innovations are treated as having the same probability distribution (i.e. identically distributed). For example, if one assumes that the innovations have a student t distribution then each and every realisation of the innovations' stochastic process is assumed to have a student t distribution, without exceptions. On the other hand, the i.i.d assumption also means that the innovations are taken to be statistically independent(i.e. independent).In other words, the value of an innovation today does not, in any way, influence the value of an innovation tomorrow or any other future values of innovations. This study then observes that the two key assumptions(i.e. the assumption that the innovations have a specific known distribution and that the innovations are statistically independent(within the i.i.d assumption)) that the parametric approach relies on expose the derived estimators to very high risk of inconsistency as explained in detail in the following paragraphs.

Assuming a specific distribution exposes the model to high risk of producing inconsistent estimators in the event that the assumed distribution is not correct (Chung, 2012; Dahl & Levine, 2010). Further, financial time series often exhibit leptokurtosis; meaning that their distribution is symmetrical in shape, similar to a normal distribution, but the centre peak is much higher with fatter tails (Holly & Montifort, 2010; Chung, 2012; Drost &

Klassenn, 1996; Bollerslev, 1986; Gallant & Hsieh, 1989; Baillie & Bollerslev, 1987). Therefore, assuming normality of the innovations may technically lead to wrong likelihood functions and hence inconsistent results. How then can one address the issue of non-normality in literature? The issue of non-normality has been generally addressed in general econometric literature. It is possible to obtain consistent results even if the innovations are not normally distributed. One of the ways is to use non-parametric estimation approach where you do not make any assumption about the distribution. This, as we have already discussed above, provide flexibility in estimation such that there is no chance of making a wrong specification. But, as we have stated above, literature shows that non-parametric approaches are only good for volatility models with confirmed leverage effect. GARCH (1, 1), unfortunately, does not take leverage effect into account.
Another way is to use parametric approach but under quasi maximum likelihood estimation principle. Hood and Koopman (1953) demonstrated that the conditionally Gaussian Maximum Likelihood estimator is consistent and asymptotically Gaussian, even if the true distribution is not conditionally Gaussian, as long as the first two conditional moments are well specified. They coined the label quasi maximum likelihood estimator"(QMLE here-in after) for this kind of estimator. So it means QMLE would still be consistent in the face of non-normal innovations in this case.

As regards to the statistical independence assumption (within the i.i.d assumption), empirical regularities of time series financial returns show the innovations are dependent and not independent. The following characteristics are frequently observed in financial data (see Holly, 2010; Chung, 2012; Drost & Klassenn, 1996; Engle & Gonzale-Rivera, 1991). The first is volatility clustering. This is where large changes tend to be followed by large changes and small changes tend to be followed by small changes. Second is that squared returns exhibit serial correlation whereas little or no serial dependence can be detected in the return series itself. In addition, financial returns exhibit fading memory i.e., distant innovations have little effect on financial returns compared to recent innovations. This means that the i.i.d assumption is not correct when it comes to financial time series. That being said, one may wonder as to what exactly is the problem with continuing with the statistical independence assumption when in fact the innovations are statistically dependent. The statistical independence assumption is very critical when the volatility function is allowed to be time dependent. This is because it ensures that the parameters entering the conditional mean function are time-independent (Dahl & Levine, 2010). Dahl and Levine (2010) argued that if the conditional mean function is estimated assuming time invariant parameters, when they are time variant, its estimators will be inconsistent and the effect of this misspecification will carry over into the volatility estimation. Technically, what we are saying here is that the statistical independence assumption (within the i.i.d assumption) is not correct in time series financial data. If we continue making it, when in fact the innovations are statistically dependent, we are bound to get inconsistent estimators.

This means that parametric estimation of GARCH (1, 1) model, the most popular estimation technique for GARCH (1, 1), is not appropriate since it is based on assumptions that are more ad hoc than based on economic reasoning in finance. This paper is coming in to offer an estimation approach that is based on economic reasoning

that is in line with empirical financial data regularities by relaxing these two key assumptions that the parametric approach is based on.

This study proposes use of semi-parametric estimation approach while relaxing the statistical independence assumption. In this way we will solve both the normality issue and the statistical independence issue. After all, we have explained in section 1.1 above that semi-parametric approach is better than both parametric and non-parametric approaches. It must be said however that this paper is not the first to propose semi-parametric estimation approach to estimating GARCH (1, 1) volatility model. However, the papers that have proposed semi-parametric approach in literature (e.g. Linton & Mammen, 2003; Drost & Klassenn, 1996; Yang & Song, 2012; Engle & Gonzale-Rivera, 1991) continue making the statistical independence assumption. Much as this may be better than the parametric approach due to increased flexibility in the model, making the statistical independence assumption is not really ideal when it comes to volatility modelling as we have explained above. This is because there is high risk of getting inconsistent estimators in the event that the innovations are statistically dependent. This means that the semi-parametric approaches proposed so far are no better than the parametric approach.

So, this study proposes a semi-parametric approach while relaxing the statistical independence assumption. However, unlike other semi-parametric approaches, instead of completely leaving the distribution of the innovations unspecified this study proposes use of a family of distributions. The assumption is that the true distribution of the innovations is not known but it is assumed to belong to a known family of probability densities. In this way, we will be introducing some reasonable flexibility in the model unlike in the parametric case.

So the study assumes that the innovations belong to the Generalised Error Distribution (GED hereinafter) while allowing them to be serially dependent. Then quasi-maximum likelihood estimation is applied to ensure that the estimator is still consistent even when the true distribution is not in the GED family. The estimator derived in this manner is definitely semi-parametric. This is because, even though the conditional variance equation is explicitly specified, the distribution of the innovations is not explicitly specified as in parametric case. The distribution is partly specified by assuming that it belongs to the GED. In this way flexibility is introduced in the model while relaxing the statistical independence assumption. As such, the estimator derived is based on reliable assumptions that are in line with financial data empirical regularities explained above.

1.3 Objectives of the study

The main objective of this paper is to derive the semi-parametric GARCH (1, 1) estimator under serially dependent GED innovations.

To achieve the main objective, the following specific objectives shall be pursued:

- To show that the semi-parametric GARCH (1, 1) estimator under serially dependent GED innovations is consistent.

- To prove that the semi-parametric GARCH (1, 1) estimator under serially dependent GED innovations is asymptotically normal.

1.4 Significance of the study

This study contributes to the financial econometrics literature by providing an estimator of the GARCH (1, 1) volatility model that allows the relaxation of both the normality distribution and the statistical independence assumptions of innovations. In that way, the study offers an estimator that is based on assumptions that are in line with empirical financial data regularities.

1.5 Organization of the study

The first Chapter focused on introduction. The rest of the paper proceeds as follows: Chapter two reviews theoretical literature regarding estimation techniques of GARCH (1,1) that have been proposed already in literature; Chapter three outlines the methodology employed to achieve the objectives, Chapter four derives and explains the main theoretical results of the study and then Chapter five offers conclusion and theoretical implications.

2. LITERATURE REVIEW

2.1 Introduction

It should be mentioned here that ordinarily, this chapter was supposed to review both empirical and theoretical literature. However, this is a theoretical paper. As such this chapter reviews only the theoretical literature. In this chapter, therefore, we explain more on different approaches of estimating deterministic GARCH (1, 1) that have been proposed in literature. We expound in turn the three broad approaches of estimating volatility models.

2.2 Parametric Approach

According to Chung (2012), a parametric approach imposes a specific linear structure on volatility and a specific probability distribution of the innovations. For GARCH (1, 1) model, conditional variance is expressed as a deterministic function of lagged conditional variance and the lagged squared innovations as shown in equation (2.01) through equation (2.02) below.

$$y_t = f(x_t, \beta) + \varepsilon_t \tag{2.01}$$

$$\varepsilon_t = z_t(\lambda_t)^{\frac{1}{2}}, z_t \sim NIID(0,1)$$

$$\lambda_t = \zeta + \alpha(L)\varepsilon_t^2 + \pi(L)\lambda_t = \frac{\zeta}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2 = \zeta\left[1 + \sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2 + \pi\right)\right] \tag{2.02}$$

Where $y_t, x_t, \varepsilon_t, \beta, \zeta, \alpha, \pi, \lambda_t \in \mathrm{R}$, ; ε_t is the innovation of the model and is generated by GARCH (1, 1) process, L is the lag operator, f is a function, x_t are factors affecting the dependent variable y_t and β are parameters of the mean equation showing how x_t impact on y_t , on average and λ_t represents conditional variance of ε_t . As can be seen in the model above, mostly the innovations are assumed to be Gaussian (see Bollerslev, 1986; Choi, 2004; Andersen, 1996; Hansen & Lunde, 2001). Recently, practitioners in the trade have been assuming the student-t distribution and the gamma distribution (see Bollerslev, 1986; Bollerslev, 1987; Holly & Pentsak, 2004; Holly, 2009; Gallant & Hsieh, 1989; Rossi, 2004; Chung, 2012). From the model, it can be seen that,

$$\varepsilon_t = y_t - f(x_t, \beta) \Rightarrow \varepsilon_t \sim N(0, \lambda_t) = N\left(0, \zeta\left[1 + \sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2 + \pi\right)\right]\right) \tag{2.03}$$

Assuming that the innovations are really normally distributed, the estimation of the model is done through Maximum Likelihood Estimation (MLE) procedure. This approach involves maximization of a log likelihood function constructed under the auxiliary assumption of an i.i.d. distribution for the standardized innovation $z_t(\theta) = \frac{\varepsilon_t(\theta)}{\lambda_t(\theta)}$; (where $\theta = (\beta \quad \zeta \quad \alpha \quad \pi)'$). Let $\phi(z_t(\theta)) = (2\pi)^{-0.5}\exp\left(-\frac{z_t(\theta)}{2}\right)$

$$\Rightarrow l_t(\theta) = -0.5\ln 2\pi - 0.5 z_t(\theta) \tag{2.04}$$

$$\Rightarrow \hat{\theta}_{MLE} = \arg\max \sum l_t(\theta) \tag{2.05}$$

$$S_t(\theta) = \frac{\partial f(x_t, \beta)}{\partial \theta}\frac{\varepsilon_t}{\lambda_t(\theta)} + 0.5\lambda_t(\theta)^{-1}\frac{\partial \lambda_t(\theta)}{\partial \theta}\left[\frac{\varepsilon_t^2(\theta)}{\lambda_t(\theta)} - 1\right] \tag{2.06}$$

Where $l_t(\theta)$ and $S_t(\theta)$ are log likelihood and score functions respectively and $\phi(z_t(\theta))$ is the standard normal function. The score function, $S_t(\theta)$, is then solved using a gradient numerical method called BHHH algorithm. This algorithm uses equation (2.07) below;

$$\hat{\theta}_{s+1} = \hat{\theta}_s + A_s g_s, s = 1, 2, 3..., S \tag{2.07}$$

Where; $A_s = -B_{HHH,S}^{-1} = -\left[-\sum_{t=1}^{T} \frac{\partial l_t(\theta)}{\partial \theta} \frac{\partial l_t(\theta)}{\partial \theta'} \mid \hat{\theta}_s \right]^{-1}$, $g_s = \frac{\partial \sum_{t=1}^{T} l_t(\theta)}{\partial \theta} \mid \hat{\theta}_s$, the gradient vector.

Under certain regularity conditions, the ML estimator converges at the rate of $\sqrt{T}$ and is asymptotically normally distributed (Davidson., 2000; Davidson & Mackinnon., 1993).That is,

$$\sqrt{T}\left(\hat{\theta}_{MLE} - \theta\right) \xrightarrow{d} N\left(0, \left(\sum_{t=1}^{T} E\left[\frac{\partial l_t(\theta)}{\partial \theta} \frac{\partial l_t(\theta)}{\partial \theta'}\right]\right)^{-1}\right) \tag{2.08}$$

The log-likelihood function in equation (2.04) is determined under the assumption of conditional normality, which is more ad hoc than based on statistical or economic reasoning. In the empirical literature on GARCH processes, it turns out that conditional normality of speculative returns is more an exception than the rule (Hewartz, 2004). As a result, assuming normality when in fact the innovations are not normal would result in wrong functional form hence inconsistent estimators. As such some researchers (e.g. Rossi, 2004; Bollerslev, 1987; Gallant & Hsieh, 1989; Baillie & Bollerslev, 1987) have proposed usage of other distributions other than the Gaussian density, for instance student t-distribution, the gamma distribution and the generalized error distribution.

If one assumes that the innovations have the student t-distribution with v (>2) degrees of freedom then its probability density ($f(\varepsilon_t \mid \theta, v)$) will be as specified in equation (2.09) below. It must be mentioned here that most derivations and formulae in this subsection are taken from Hewartz (2004).

$$f(\varepsilon_t \mid \theta, v) = \left(\left(\sqrt{\pi}\Gamma\left(\frac{v}{2}\right)\sqrt{\frac{(v-2)\sigma_t^2}{v}}\right)^{-1}\left(v^{\frac{v}{2}}\Gamma\left(\frac{v+1}{2}\right)\right)\right)\left(v + \frac{v.\varepsilon_t^2}{(v-2)\sigma_t^2}\right)^{-\left(\frac{v+1}{2}\right)} \tag{2.09}$$

Where $\Gamma(.)$ denotes a gamma function specified in equation (2.10) below.

$$\Gamma(h) = \int_0^{\infty} x^{h-1} \exp(-x) dx, h > 0 \tag{2.10}$$

$$\Rightarrow l_t(\theta, v) = \ln\left(v^{\frac{v}{2}}\Gamma\left(\frac{v+1}{2}\right)\right) - \ln\left(\sqrt{\pi}\Gamma\left(\frac{v}{2}\right)\sqrt{\frac{(v-2)\sigma_t^2}{v}}\right) - \frac{v+1}{2}\ln\left(v + \frac{v.\varepsilon_t^2}{(v-2)\sigma_t^2}\right) \tag{2.11}$$

Letting $\sigma_t^2 = z_t'\theta$, the score functions then can be derived as in equations (2.12) through (2.13) below.

$$\Rightarrow \frac{\partial l_t(\theta, v)}{\partial \theta} = -0.5\sigma_t^{-2} z_t + \frac{v+1}{2}\left(v + \frac{v.\varepsilon_t^2}{(v-2)\sigma_t^2}\right)^{-1}\left(\frac{v.\varepsilon_t^2}{(v-2)\sigma_t^2} z_t\right) \tag{2.12}$$

$$\Rightarrow \frac{\partial l_t(\theta,v)}{\partial v} = 0.5\left[\ln v + 1 + \Gamma^{-1}\left(\frac{v+1}{2}\right)\Gamma'\left(\frac{v+1}{2}\right) - \Gamma^{-1}\left(\frac{v}{2}\right)\Gamma'\left(\frac{v}{2}\right) - \frac{1}{v-2} + \frac{1}{v}\right]$$
$$-0.5\left[\ln\left(v + \frac{v.\varepsilon_t^2}{(v-2)\sigma_t^2}\right) - (v+1)\left(v + \frac{v.\varepsilon_t^2}{(v-2)\sigma_t^2}\right)^{-1}\left(1 - \frac{2.\varepsilon_t^2}{(v-2)\sigma_t^2}\right)\right] \quad (2.13)$$

In the same vein, if the innovations have a generalized error distribution ($f(\varepsilon_t \mid \theta, v)$);

$$f(\varepsilon_t \mid \theta, v) = \left[2^{\frac{v+1}{2}}\left(\int_0^\infty x^{\frac{1}{v}-1}\exp(-x)dx\right)\lambda.\sigma_t\right]^{-1} v\exp\left(-\frac{1}{2}\left|\frac{\varepsilon_t}{\lambda\sigma_t}\right|^v\right)$$

$$\text{Where } \lambda = \left[\left(2^{\frac{2}{v}}\left(\int_0^\infty x^{\frac{3}{v}-1}\exp(-x)dx\right)\right)^{-1}\left(\int_0^\infty x^{\frac{1}{v}-1}\exp(-x)dx\right)\right]^{0.5}$$

This means that the log likelihood function is given by equation (2.14) below;

$$\Rightarrow l_t(\theta,v) = \ln v - 0.5\left|\frac{\varepsilon_t}{\lambda\sigma_t}\right|^v - \ln\left(2^{\frac{v+1}{2}}\left(\int_0^\infty x^{\frac{1}{v}-1}\exp(-x)dx\right)\lambda\right) - 0.5\ln\sigma_t^2 \quad (2.15)$$

The score functions then can be derived as in equations (2.16) through (2.17) below

$$\Rightarrow \frac{\partial l_t(\theta,v)}{\partial\theta} = 0.25v\left|\frac{\varepsilon_t}{\lambda\sigma_t}\right|^v \sigma_t^{-2}z_t - 0.5\sigma_t^{-2}z_t \quad (2.16)$$

$$\Rightarrow \frac{\partial l_t(\theta,v)}{\partial v} = \frac{1}{v} - 0.5\left|\frac{\varepsilon_t}{\lambda\sigma_t}\right|^v\left(\ln\left(\left|\frac{\varepsilon_t}{\lambda\sigma_t}\right|\right) - v^{\frac{\lambda'}{\lambda}}\right) + v^{-2}\left(\begin{array}{c}\ln 2 + \left(\int_0^\infty x^{\frac{1}{v}-1}\exp(-x)dx\right)^{-1} \\ \times\left(\int_0^\infty x^{\frac{1}{v}-1}\exp(-x)dx\right)'\end{array}\right) - \frac{\lambda'}{\lambda} \quad (2.17)$$

$$\text{Where } \lambda' = \lambda^{-1}2^{\frac{v-2}{2}}v^{-2}\Gamma\left(\frac{3}{v}\right)\left[2\ln 2\Gamma\left(\frac{1}{v}\right) - \Gamma'\left(\frac{1}{v}\right) + 3\Gamma^{-1}\left(\frac{3}{v}\right)\Gamma'\left(\frac{3}{v}\right)\Gamma\left(\frac{1}{v}\right)\right]$$

Just like under normality assumption; the t, gamma and the generalized error score functions are also solved using the BHHH algorithm. Similarly, if the true distribution of

the innovations is not the specified one, then we are more likely to derive wrong likelihood function and hence inconsistent estimators. As such the Maximum likelihood procedure (where a specific distribution of the innovations is specified) is not commonly used in estimating GARCH (1, 1) volatility model. Rather estimation of GARCH (1, 1) model is mostly done through the use of QML.

A key difference between these two methods is that the former allows for possible misspecification of the likelihood function (Chung, 2012). By contrast, the conventional ML method assumes that the postulated likelihood function is specified correctly, so that specification errors are assumed away. As such, the results in the ML method are just special cases of the QML method (Chung, 2012). Technically, QML uses the Kullback-Leiber Information Criterion (KLIC) (Cameron & Trivedi, 2005; Chung, 2012). In the spirit of KLIC, let $\zeta(\varepsilon)$ be the true distribution of the innovations (which is unobserved), and $\zeta(\varepsilon \mid \theta)$ be the `assumed' distribution, then KLIC can be given as in equation (2.18) below.

$$KLIC = E_{\zeta(\varepsilon)}\left[\ln\left(\frac{\zeta(\varepsilon)}{\zeta(\varepsilon \mid \theta)}\right)\right] = \sum_{t=1}^{T}\ln\left(\frac{\zeta(\varepsilon)}{\zeta(\varepsilon \mid \theta)}\right)\zeta(\varepsilon)\partial\varepsilon \tag{2.18}$$

KLIC takes a minimum value of 0 when there is a θ_0 such that $\zeta(\varepsilon) = \zeta(\varepsilon \mid \theta)$ i.e. the density is correctly specified. Larger values of $KLIC$ indicate greater ignorance about the true density (Chung, 2012).Then equation (2.19) gives the QMLE,

$$\hat{\theta}_{QMLE} = \arg\min KLIC = \arg\min \sum_{t=1}^{T}\ln\left(\frac{\zeta(\varepsilon)}{\zeta(\varepsilon \mid \theta)}\right)\zeta(\varepsilon)\partial\varepsilon = E\left[\ln \zeta(\varepsilon \mid \theta)\right] \tag{2.19}$$

Following Hewartz (2004), if the normality assumption is violated, the covariance matrix of the QMLE[2] is:

$$\sqrt{T}\left(\hat{\theta}_{QMLE} - \theta\right) \xrightarrow{d} N\left(0, \left(-E\left[\frac{\partial l_t(\theta)}{\partial\theta}\frac{\partial l_t(\theta)}{\partial\theta'}\right]\right)^{-1}\left(\sum_{t=1}^{T}\left[\frac{\partial l_t(\theta)}{\partial\theta}\frac{\partial l_t(\theta)}{\partial\theta'}\right]\right)\left(-E\left[\frac{\partial l_t(\theta)}{\partial\theta}\frac{\partial l_t(\theta)}{\partial\theta'}\right]\right)^{-1}\right)$$

The parametric approach has an advantage that the estimators converge at a high rate, $\sqrt{T}$ (which is the highest rate an estimator can achieve). In addition, QMLE is consistent regardless of whether the functional forms are correct/ incorrect as long as the first two moments are correctly specified (Hood and Koopman, 1953). However, volatility is a latent variable hence explicit functional form assumptions and/or the explicit probability distribution assumption on the models might be too strong. This exposes the models to high risk of providing inconsistent and/or inefficient estimators assuming the assumed functional form and/or the assumed probability density laws are wrong (Chung, 2012; Yang & Song, 2012; Dahl & Levine, 2010; Drost & Klassenn, 1996) .Lastly, it assumes

[2]Check appendix D below and Appendix A in Gonzalez-Rivera (1991) to see regularity conditions for QMLE

that the innovations are i.i.d. And again if the innovations are not i.i.d, we are bound to make wrong likelihood function and then inconsistent estimators (Dahl & Levine, 2010).

2.3 Non parametric Approach

Technically there are two definitions of the term non-parametric statistics. The first meaning incorporates methods that do not rely on data belonging to any particular distribution (i.e. distribution free methods).The other meaning incorporate techniques that do not assume that the structure of a model is fixed a priori. In such methods, variables are assumed to belong to parametric distributions (Chung, 2012).It should be noted here that in this definition, even though structural assumptions about the model are not made a priori, statistical assumptions about the variables are made. In either definition, it is clear that, generally, non-parametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from data. In other words, the term non-parametric is not meant to insinuate that such models completely lack parameters but that the number and the nature of the parameters are flexible and not fixed a priori (Chung, 2012).

To ease the structural assumptions in parametric models, nonparametric models make no structural assumptions. According to Chung (2012), under this approach, the conditional variance is not explicitly specified and the distribution of the innovations is kept unknown. Compared to parametric models, there is limited literature on nonparametric volatility models. Linton and Mammen (2003) and Yang and Song (2012) examined some recent advances in nonparametric volatility modelling. Before going any further, let's look at some basics of non-parametric estimation techniques. The following derivations and formulae are due to Tschernig (2004).

Assume equations (2.20) and (2.21) hold;

$$y_t = \mu(x_t) + \sigma(x_t)\varepsilon_t \tag{2.20}$$

$$x_t = \left(y_{t-i_1}, y_{t-i_2}, \ldots\ldots, y_{t-i_n}\right) \tag{2.21}$$

Where x_t is the $(m\times 1)$ vector of all m current lagged values; $i_1 < i_2 < \ldots < i_m$; $\varepsilon_t, t = i_m + 1, i_m + 2, \ldots.$, denotes a sequence of i.i.d random variables with zero mean and variance unity; $\mu(x_t)$ and $\sigma(x_t)$ denote the conditional mean and volatility function, respectively. Estimation of $\mu(x_t)$ and $\sigma(x_t)$ in equation (2.20) is mostly done locally, meaning it is estimated separately for each $(m\times 1)$ vector $x = \left(x_1, x_2, \ldots x_m\right)'$ of interest. Under this approach, although $\mu(x_t)$ is not observable it appears in a first order Taylor expansion of $\mu(x_t)$ taken at x as can be seen in equation (2.22) below·

$$\mu(x) = \mu(x_t) + \frac{\partial \mu(x_t)}{\partial x_t}(x_t - x) + R(x_t, x) \tag{2.22}$$

Where $R(x_t, x)$ denotes the remainder term. But equation (2.20) can now be written as equation (2.23) below.

$$y_t = \mu(x_t) + \sigma(x_t)\varepsilon_t \Rightarrow y_t = \mu(x_t)1 + \frac{\partial \mu(x_t)}{\partial x_t}(x_t - x) + R(x_t, x) + \varepsilon_t \quad (2.23)$$

From this, we observe that only 1 and $(x_t - x)$ are observable. This means that if $R(x_t, x) = 0$, one would estimate $\mu(x_t)$ by OLS, where $\mu(x_t)$ and $\frac{\partial \mu(x_t)}{\partial x_t}$ are parameters to be estimated. But, whenever the conditional mean is non-linear, the remainder term may not be zero and in such a case using standard OLS would give biased results for which the size of biasness depends on the sizes of all the remainder terms, $R(x_t, x), t = 1, 2, ..T$. One possibility to reduce the biasness is to use only those observations x_t that are in some sense close to x. That is to say, down weighing those observations that are not in local neighbourhood of x. If more data become available, it is possible to decrease the size of the local neighbourhood where the estimation variance and bias can reduce. The approximation error of the model can decline with sample size. This is the main idea behind non-parametric estimation approach.

There are so many streams of non-parametric estimation techniques in the literature depending on the weighing scheme used. Technically, the weighing is controlled by the so-called kernel function $K(.)$. A kernel function is a continuous function symmetric around zero that integrates to unity and satisfies the following additional boundness conditions (Cameron & Trivedi, 2005).

1) $K(z)$ is symmetric around 0 and is continuous
2) $\int K(z)dz = 1, \int zK(z)dz = 0, \&, \int |K(z)|dz < \infty$
3) $K(z) \to 0,$ as $|z| \to \infty$
4) $\int z^2 K(z)dz = \delta$ where δ is a constant

To adjust the size of the neighbourhood one introduces a band-width h, such that for a scalar x, the kernel function becomes $\frac{1}{h}K\left(\frac{x_t - x}{h}\right)$. If $m > 1$ and $x = (x_1, x_2, ..x_m)'$ is a vector, one uses a product kernel in equation (2.24) below;

$$K_h(x_t - x) = \prod_{i=1}^{m} \frac{1}{h^2} K\left(\frac{x_{ti} - x}{h}\right) \quad (2.24)$$

Here x_{ti} denotes the i-th component of x_t. The larger the h, the larger the neighbourhood and the larger the estimation bias. The band-width is also called a smoothing parameter.

Since the observations in the local neighbourhood of x are the most important, this estimation is also called local estimation.

Owing to the introduction of a kernel function, one now has to solve a weighted least-squares problem as shown in equation (2.25);

$$(c_1, c_2, \ldots c_m) = \arg\min \sum_{t=i_m+1}^{T} \left(y_t - c - \sum_{i=1}^{m} c_i (x_{ti} - x) \right)^2 K_h (x_t - x) \tag{2.25}$$

This delivers the local linear function estimate $\hat{\mu}(x_t, h)$ at the point x .Technically, with matrices;

$$e = \begin{pmatrix} 1 & 0_{1\times m} \end{pmatrix}'; \; z(x) = \begin{pmatrix} 1 & \ldots. & 1 \\ x_{i_m+1} & \ldots. & x_T - x \end{pmatrix}'; \; W(x,h) = diag\left(\frac{K_h(x_t - x)}{T} \right)_{i=i_m+1}^{T}$$

$$y = \left(y_{i_m+1,\ldots.}, y_T \right)' \Rightarrow \hat{\mu}(x,h) = e'\left(z'(x) W(x,h) z(x) \right)^{-1} z'(x) W(x,h) y \tag{2.26}$$

The most popular weighting scheme is that of Watson (1964) and Nadaraya(1964) given in equation (2.27) below;

$$\hat{\mu}_{NW}(x,h) = e'\left(z'_{WN}(x) W(x,h) z(x) \right)^{-1} z'_{WN}(x) W(x,h) y \tag{2.27}$$

$$= \left(\sum_{t=i_m+1}^{T} K_h (x_t - x) \right)^{-1} \left(\sum_{t=i_m+1}^{T} K_h (x_t - x) \right) y_t ; z_{NW} = \begin{pmatrix} 1 & \ldots\ldots. & 1 \end{pmatrix}'_{1\times(T-i_m)}$$

A theoretical formulation of Nadaraya -Watson conditional variance estimator using squared residuals obtained from the conditional mean function is proposed by Fan and Yao (1998) and Fan and Gijbels (1995) and we sketch their formalization below. The explanation below is due to Chung (2012). Assume that a strictly stationary process $\{x_t; t = 1,2,\ldots.,T\}$ is generated by equation (2.28) below.

$$x_t = m(x_{t-1}) + \sqrt{\lambda}(x_{t-1}) \varepsilon_t \tag{2.28}$$

Where ε_t is i.i.d, $E(\varepsilon_t \mid \Xi_{t-1}) = 0$, $Var(\varepsilon_t \mid \Xi_{t-1}) = 1$, and Ξ_{t-1} is a sigma algebra generated by x_{t-1} (or some past information). Fan and Yao (1998) proposed a two stage method to obtain a local linear estimator for conditional variance (Chung, 2012; Dahl & Levine, 2010).

1) Obtain local linear estimator $\hat{m}(x_t) = \hat{a}$ which is the minimization intercept in the following weighted least-squares problem;

$$\left(\hat{a},\hat{b}\right)=\arg\min\sum_{t=i_m+1}^{T}\left(x_t-a-b\left(x-x\right)\right)^2K_h\left(\frac{x_{t-1}-x}{h_1}\right) \tag{2.29}$$

2) Obtain the squared residuals $\hat{r}_t=\left[x_t-\hat{m}\left(x_{t-1}\right)\right]^2$ to use in equation (2.30) below.

$$\left(\hat{\alpha},\hat{\beta}\right)=\arg\min\sum_{t=i_m+1}^{T}\left(\hat{r}_t-\alpha-\beta\left(x-x\right)\right)^2K_h\left(\frac{x_{t-1}-x}{h_2}\right) \tag{2.30}$$

Where the bandwidth $h_2>0$ is different from h_1 (Chung, 2012; Dahl & Levine, 2010)

As we have seen in equation (2.27) above, the non-parametric mean estimator is;

$$\hat{\mu}_{NW}\left(x,h\right)=e'\ \left(z'_{WN}\left(x\right)W\left(x,h\right)z\left(x\right)\right)^{-1}z'_{WN}\left(x\right)W\left(x,h\right)y$$

$$\left(\sum_{t=i_m+1}^{T}K_h\left(x_t-x\right)\right)^{-1}\left(\sum_{t=i_m+1}^{T}K_h\left(x_t-x\right)\right)y_t$$

According to Tschernig (2004);

$$\sqrt{Th^m}\left(\hat{\mu}\left(x,h\right)-\mu\left(x\right)-b\left(x\right)h^2\right)\rightarrow N\left(0,v\left(x\right)\right)$$

Where the asymptotic bias $b\left(x\right)$ and asymptotic variance $v\left(x\right)$ are as given respectively in equations (2.31) and (2.32) below

$$b\left(x\right)=\frac{\sigma_k^2}{2}tr\left(\frac{\partial\mu\left(x\right)}{\partial x\partial x'}\right) \tag{2.31}$$

$$v\left(x\right)=\frac{\sigma^2\left\|K\right\|_2^{2m}}{f\left(x\right)} \tag{2.32}$$

$$\Rightarrow\hat{\mu}\left(x,h\right)\rightarrow N\left(\mu\left(x\right)+\frac{\sigma_k^2}{2}tr\left(\frac{\partial\mu\left(x\right)}{\partial x\partial x'}\right)h^2,\frac{1}{Th^m}\frac{\sigma^2\left\|K\right\|_2^{2m}}{f\left(x\right)}\right)$$

Thus, if we denote any positive constant β, any optimal band-width for which $h=\beta T^{\left(-\frac{1}{(m+4)}\right)}$ holds has an optimal rate of decline of bias (Tschernig, 2004).

$$\Rightarrow\sqrt{Th^m}\left(\hat{\mu}\left(x,h\right)-\mu\left(x\right)-b\left(x\right)h^2\right)\rightarrow N\left(0,v\left(x\right)\right)$$

$$\Rightarrow T^{\frac{2}{(M+4)}}\left(\hat{\mu}(x,h)-\mu(x)\right)\rightarrow N\left(\frac{\sigma_k^2}{2}tr\left(\frac{\partial\mu(x)}{\partial x\partial x'}\right)h^2,\frac{1}{h^m}\frac{\sigma^2\|K\|_2^{2m}}{f(x)}\right)$$

From this, the rate of convergence is $T^{\frac{2}{(M+4)}}$ which is less than the parametric rate $T^{\frac{1}{2}}$ for any $m(>0)$ (Tschernig, 2004).This implies that the rate of convergence of the local linear estimator depends on the number of lags, m, and becomes slow even when $m=1$ (the case of GARCH (1, 1)).In fact for GARCH(1,1), $m=1$ implying that the rate of convergence is $T^{\frac{2}{5}}$ which is far less than the convergence rate of the parametric case; $T^{\frac{1}{2}}$ .The fact that non-parametric approach results in low speed of convergence makes it undesirable in volatility estimations. In addition, we note that this approach, just like the parametric approach, assumes that the innovations are i.i.d not to mention the complexity it brings in estimation even though its results are no better than the parametric ones in case i.i.d holds. This means that even though this approach offers greater flexibility to the model, we may still get inconsistent estimates, if the innovations are non-i.i.d.

2.4 Semi Parametric Approach

This approach basically combines elements of both parametric and non-parametric approaches. Under this approach, the conditional variance is explicitly specified, just like in parametric case but the distribution of the innovations are assumed un-known just like in non-parametric case (Drost & Klassenn, 1996; Yang & Song, 2012; Engle &Gonzale-Rivera, 1991). Just like the non-parametric case, there is limited literature on semi-parametric volatility models. Tschernig (2004), Engle and Ng (1993), Hafner and Rombonts (2002), Haafner (2003), Yang and Song (2012), Buhlman and McNeil (2000), Linton and Mammen (2003), and Yang and Song (2012) are the most popular. As we can notice, semi parametric approach is a hybrid of parametric and non-parametric approaches as it combines elements of these two approaches.

As such estimation of GARCH-type volatility models semi-parametrically uses a combination of both the parametric approach and the non-parametric approach. The first semi-parametric paper in GARCH context which was done by Engle and Gonzale-Rivera (1991) was partially successful. Using Discrete Maximization Penalized Likelihood Estimation (DMPLE) of Tapia & Thompson (1991), Engle and Gonzale-Rivera (1991) showed that the efficiency of estimates was improved only up to 50 per cent over QMLE but did not hit the Cramer-Rao lower bound. Under this approach, the following model given in equation (2.33) is considered;

$$y_t=E(y_t\mid\Omega_{t-1})+\varepsilon_t$$

$$\varepsilon_t=z_t(\lambda_t)^{\frac{1}{2}}$$

$$\lambda_t=\zeta+\alpha(L)\varepsilon_t^2+\pi(L)\lambda_t=\frac{\zeta}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^kL^k\varepsilon_{t-1}^2=\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^iz_t^2+\pi\right)\right] \quad (2.33)$$

$$\varepsilon_t(\lambda_t)^{-\frac{1}{2}} = z_t \sim i.i.dg((0,1))$$

Where f is the unknown density function of ε_t conditional on the set of past information. Estimation of λ_t and f, involves maximizing the following likelihood function;

$$L_T(\theta) = -0.5\sum_{t=1}^{T} \ln \lambda_t + \sum_{t=1}^{T} \ln g\left(\frac{\varepsilon_t^2}{(\lambda_t)^{\frac{1}{2}}}\right) \tag{2.34}$$

According to Engle and Gonzalez-Rivera (1991), this is done in the following steps;

Firstly, choose initial consistent estimates of θ. This is done by applying Ordinary Least Squares (OLS) on the model (Engle, 1986) or applying QML (Bollerslev& Woodridge, 1992; Weiss, 1986). Secondly, save the residuals ε_t and the conditional variance λ_t from step 1 and construct $\frac{\varepsilon_t^2}{(\lambda_t)^{\frac{1}{2}}}$.

Then use DMPLE technique to estimate the non-parametric density. DMPLE involves maximizing the actual likelihood of the sample in which the arguments of the function are the heights $(p_1, p_{2,\ldots,} p_{m-1})$ of the generalized histogram at some given knots $(n_1, n_{2,\ldots,} n_{m-1})$. For a sample $x_1, x_{2,\ldots,} x_n \left(\text{in our case} \frac{\varepsilon_1^2}{(\lambda_1)^{\frac{1}{2}}}, \frac{\varepsilon_2^2}{(\lambda_2)^{\frac{1}{2}}}, \ldots, \frac{\varepsilon_n^2}{(\lambda_n)^{\frac{1}{2}}}\right)$ and interval (a,b), divide into m sub-intervals of length q. Then solve the following;

$$MaxL(p_1, p_2, \ldots, p_{m-1}) = \sum_{i=1}^{m} \ln g(x_i) - \frac{\lambda}{q}\sum_{k=1}^{m-1}(p_{k+1} - 2p_k + p_{k-1})^2$$

$$Subj.q\sum_{k=1}^{m-1} p_k = 1, p_{k\geq 0, k=1,2,\ldots,m-1}$$

$$\text{where } g(n) = p_k + \frac{p_{k+1} - p_k}{q}(n - n_k), n \in [n_k, n_{k+1}]; g(n) = 0, n \notin (n_0, n_m)$$

Where λ is the penalty term (chosen by the researcher to ensure smoothness of the estimate of p)Lastly, perform the maximization of the log-likelihood function from stem three, keep $\hat{g}$ fixed and iterate to convergence.

Following Klassenn et al. (1996), this procedure produces estimators that are not adaptive in the class of densities with mean 0 and variance 1. Semi-parametric estimators converge almost at the same rate as parametric case while at the same time offering flexibility to the model by not/partially specifying the distribution of the innovations (see Linton,

2005; Yang, 1998; Levine et al., 2012; Engle et al., 1993). This technically means that semi parametric approach is the best among the three approaches since it has good aspects of both the parametric and non-parametric approaches; that is, high convergence rate and flexibility. Nevertheless, just like in the parametric and non-parametric methods reviewed above; this approach still assumes that the innovations are i.i.d.

It must be said that under almost all these approaches to estimating GARCH (1, 1) model the innovations are assumed i.i.d. As explained before, assuming independence when in fact the innovations are dependent for GARCH (1, 1) could lead to inconsistent estimators.

2.5 Critique of existing literature

In this chapter we have noted that there are three broad approaches to estimating GARCH (1, 1) model in literature. These are parametric, nonparametric and semi-parametric approaches. The following criticisms can be directed at these approaches.

The existing parametric approaches in the literature, still assume independence of innovations and/or specific distribution (i.e. normal distribution, student t distribution etc). These assumptions are not only ad hoc than based on economic or statistical reasoning but also expose the estimators to high risk of inconsistency. These assumptions therefore lender the parametric approaches in the existing literature very risky hence undesirable.

Non-parametric approaches reduce the risk of getting inconsistent estimators by not assuming any distribution about the innovations. However most of them assume that the innovations are independent hence still exposing the estimators to some risk of inconsistency as well. This however is not the main critique that this study directs towards non-parametric approaches since there have been non-parametric estimators recently that have been derived while relaxing the i.i.d assumption (see Dahl & Levine, 2010). The main issue with non-parametric estimation of GARCH (1, 1) model is that the estimators have very low convergence rate. Non parametric approaches proposed in the literature so far have very low asymptotic convergence rate of about $T^{\frac{2}{5}}$ (the parametric convergence rate is $T^{\frac{1}{2}}$).This means that non-parametric estimators produce good asymptotic properties when the sample size is very big. However, most of the times we do not have the luxury of having very large sample sizes in time series analysis.This lender the non-parametric approaches in the existing literature undesirable.

With regard to semi-parametric GARCH (1, 1) estimators in the current literature we note that they seem to be the best since they have both high convergence rate and offers great flexibility. However, the assumption of statistical independence of innovations is still maintained. We have noted in sections above that this assumption is more ad hoc than based on economic reasoning. This means that the existing semi-parametric GARCH (1,1) estimators are no better than the parametric and non-parametric ones since they are still exposed to high risk of inconsistency in the event that the statistical independence assumption is not holding (a more likely phenomenon in time series financial data).

It is apparent therefore that the existing GARCH (1, 1) estimators are very risky to use. The existing literature lacks a GARCH (1, 1) estimator that is flexible while having high convergence rate (i.e $T^{\frac{1}{2}}$). This is the research gap that this study is geared to fill. The study aims at filling that research gap by deriving a GARCH (1, 1) estimator using semi-parametric approach while allowing the innovations to be serially dependent. In this way such an estimator will not only have nice asymptotic properties but it will also be based on assumptions that are not mere ad hoc but based on economic reasoning in line with time series financial data empirical regularities.

3. METHODOLOGY

3.1 Introduction

This chapter presents the methodology used to achieve the objectives. As we have seen from the literature review above, one way to obtain better estimators of GARCH(1,1) model is to employ semi parametric approach(after all it is the best approach) but then relax the i.i.d assumption(that is, allowing the innovations to be serially dependent). So, in this paper we apply semi parametric approach and relax the i.i.d assumption to derive a GARCH (1, 1) estimator. However, relaxing the i.i.d assumption can make establishment of consistency and asymptotic normality of the estimator derived very difficult. This is because the commonly used techniques (i.e. weak law of large numbers and Lindeberg central limit theorem) to showing consistency and normality use the i.i.d assumption. Nevertheless, it is possible to prove consistency and asymptotic normality of an estimator by using martingale technique (i.e. using the almost sure convergence instead of convergence in probability and the martingale central limit theorem) under non i.i.d assumption (see Hansen &Lunde, 2001; Kouassi, 2015). As such we will establish consistency and asymptotic normality in this study using martingale techniques.

This section is outlined as follows; we first introduce the GARCH (1, 1) model. After that we present the likelihood function. Then we present assumptions adopted in our analysis. We will then present and prove lemmas that will set the ground ideal for us to prove consistency and asymptotic normality.

3.2 The Model and Assumptions

Before proceeding any further it is imperative that we present the model we will be working with and the underlying assumptions that we will make in order to achieve our objectives. The observed volatility from a sample may not necessarily be the true volatility. As such it is standard in QML estimation literature to specify the observed and the unobserved volatility models (Rossi, 2004; Kouassi, 2015; Posedel, 2005; Chung, 2012). We will therefore specify the observed and the unobserved volatility models respectively, from which we will derive the observed and unobserved likelihood functions. It should be mentioned here that the model, and the lemmas are standard. They were not developed in this paper. They are/may be found in many theoretical econometrics papers(e.g. Rossi, 2004; Posedel, 2005; Chung, 2012; Kouassi, 2015).As such, necessary references have been made in that regard. This paper has, however, proved lemmas, given the key assumptions that this paper is making. In addition to that, in the course of the proofs, whenever some idea has been taken from someone/somewhere necessary referencing has accordingly been made.

3.2.1 The Model

3.2.1.1 Unobserved GARCH (1, 1) model

Following Rossi (2004) and Engle & Gonzale-Rivera (1991) among others, the unobserved model with unknown parameters $\theta = (\alpha, \beta, \zeta, \pi)'$ will be given as shown in equation (3.01) through equation (3.01) below. It must be mentioned here that equation (3.01) is technically called the mean equation while equation (3.02) is technically called conditional variance equation. It is equation (3.0) that is normally referred to as the GARCH (1, 1) model.

$$y_t = f(x_t, \beta) + \varepsilon_t \tag{3.01}$$

$$\varepsilon_t = z_t(\lambda_t)^{\frac{1}{2}}, t = 1,2,\ldots,..$$

$$\lambda_t = \zeta + \alpha(L)\varepsilon_t^2 + \pi(L)\lambda_t \tag{3.02}$$

The conditional variance equation can be simplified as follows[3];

$$\lambda_t = \zeta + \alpha(L)\varepsilon_t^2 + \pi(L)\lambda_t$$

[3]This simplification was completely done by the Author although it may also be found in other papers since the mathematics involved is not uncommon.

$$= \zeta + \alpha\varepsilon_{t-1}^2 + \pi\lambda_{t-1}$$

$$\Rightarrow \lambda_t - \pi\lambda_{t-1} = \zeta + \alpha\varepsilon_{t-1}^2$$

$$\Rightarrow \lambda_t(1-\pi L) = \zeta + \alpha\varepsilon_{t-1}^2 \Rightarrow \lambda_t = \frac{\zeta}{(1-\pi L)} + \frac{\alpha\varepsilon_{t-1}^2}{(1-\pi L)}$$

$$= \zeta(1 + \pi L + \pi^2 L^2 + \pi^3 L^3 + ..) + \alpha\varepsilon_{t-1}^2(1 + \pi L + \pi^2 L^2 + \pi^3 L^3 + ...)$$

$$= \zeta(1 + \pi + \pi^2 + \pi^3 + ...) + \alpha(\varepsilon_{t-1-0}^2 + \pi\varepsilon_{t-1-1}^2 + \pi^2\varepsilon_{t-1-2}^2 + \pi^3\varepsilon_{t-1-3}^2 + ...)$$

$$= \frac{\zeta}{(1-\pi)} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2$$

Here, L is a lag operator. Following Rossi (2004), the conditional variance equation can also be expressed as[4];

$$\lambda_t = \zeta + \alpha(L)\varepsilon_t^2 + \pi(L)\lambda_t = \zeta + \lambda_{t-1}(\alpha z_{t-1}^2 + \pi)$$

$$\therefore \lambda_t = \zeta + \lambda_{t-1}(\alpha z_{t-1}^2 + \pi)$$

$$\text{But } \lambda_{t-1} = \zeta + \lambda_{t-2}(\alpha z_{t-2}^2 + \pi)$$

$$\Rightarrow \lambda_t = \zeta + [\zeta + \lambda_{t-2}(\alpha z_{t-2}^2 + \pi)](\alpha z_{t-1}^2 + \pi)$$

$$= \zeta + \zeta(\alpha z_{t-1}^2 + \pi) + \lambda_{t-2}(\alpha z_{t-2}^2 + \pi)(\alpha z_{t-1}^2 + \pi)$$

$$= \zeta + \zeta(\alpha z_{t-1}^2 + \pi) + \zeta(\alpha z_{t-1}^2 + \pi)(\alpha z_{t-2}^2 + \pi) + \lambda_{t-3}(\alpha z_{t-3}^2 + \pi)(\alpha z_{t-2}^2 + \pi)(\alpha z_{t-1}^2 + \pi)$$

.

.

.

$$\lambda_t = \zeta\left[1 + \sum_{k=1}^{\infty}\prod_{i=1}^{k}(\alpha L^i z_t^2 + \pi)\right] \tag{3.03}$$

$$\therefore \lambda_t = \zeta + \alpha(L)\varepsilon_t^2 + \pi(L)\lambda_t = \frac{\zeta}{(1-\pi)} + \alpha\sum_{k=-0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2$$

$$= \zeta\left[1 + \sum_{k=1}^{\infty}\prod_{i=1}^{k}(\alpha L^i z_t^2 + \pi)\right], \lambda_1 = \zeta$$

[4]For more on this simplification, check Rossi(2004;page 15)

This implies that the conditional volatility equation can also be expressed as:

$$\lambda_t = \zeta + \alpha(L)\varepsilon_t^2 + \pi(L)\lambda_t = \frac{\zeta}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2$$

$$= \zeta\left[1 + \sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2 + \pi\right)\right], \lambda_1 = \zeta$$

Technically, this means that our model can be presented as[5]:

$$y_t = f(x_t, \beta) + \varepsilon_t \tag{3.04}$$

$$\varepsilon_t = z_t(\lambda_t)^{\frac{1}{2}}, t = 1,2,3,.....$$

$$\lambda_t = \zeta + \alpha(L)\varepsilon_t^2 + \pi(L)\lambda_t = \frac{\zeta}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2$$

$$= \zeta\left[1 + \sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2 + \pi\right)\right], \lambda_1 = \zeta \tag{3.05}$$

We define $\Omega_t = \left\{x_1, x_2, ..., x_t, (\lambda_t)^{\frac{1}{2}}(\varepsilon_t, \varepsilon_{t-1}, ...)\right\}$ as a set of information at time t

It should be mentioned here once again that the unobserved model given in equations (3.01) through equation (3.01) and later simplified to equations (3.04) and (3.05), just like the observed GARCH (1, 1) model presented below, are already there in literature (see Rossi,2004;Chung,2012 & Engle and Gonzale-Rivera,1991).In the same vein the simplification of equation (3.02) to yield equation (3.03) has been taken from Rossi(2004).It is reproduced here for simplification and expository purposes.

[5]See Rossi(2004),Engel and Gonzale-Rivera(1991) and Chung(2012)

3.2.1.2 Observed GARCH (1, 1) model

Again Following Rossi (2004) and Engle & Gonzale-Rivera (1991) among others, the observed GARCH (1, 1) model will be given as shown in equation (3.06) through equation (3.07) below.

$$y_t = f(x_t, \beta_0) + \varepsilon_{0t} \tag{3.06}$$

$$\varepsilon_{0t} = z_t(\lambda_{0t})^{\frac{1}{2}}, t = 1,2,....,T$$

$$\lambda_{0t} = \zeta_0 + \alpha_0(L)\varepsilon_{0t}^2 + \pi_0(L)\lambda_{0t} \tag{3.07}$$

Where y_t , x_t , ε_{0t} , β_0 , ζ_0 , α_0 , γ , $\lambda_{0t} \in \mathrm{R}$; ε_{0t} is the error term of the model and is generated by GARCH (1, 1) process, L is the lag operator, f is a function, x_t are factors affecting the dependent variable y_t and β_0 are parameters of the mean equation showing how x_t impacts on y_t , λ_{0t} represents conditional variance of ε_{0t} ,and the subscript 0 connotes "observed". Normally z_t is taken to be standard Gaussian random variable, but as we will see under assumptions below, z_t is no longer going to be assumed to be normal but rather generalised error.

It should be noted here that the key difference between the un-observed and the observed model presented above is that the unobserved model is spanning from period 1 to infinity while the observed model is finite, spanning from period 1 to period T. That implies that we are treating it as a stochastic process, a realisation of which is the one we are calling the observed model. This means that the observed model may be stationary while the unobserved model is stationary given the assumptions that this paper is making in sections below. This is standard notation[6]. Following Rossi (2004) and Posedel (2005), the conditional volatility equation for the observed model can be derived analogously as:

$$\lambda_{0t} = \frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=-0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \tag{3.08}$$

&

$$\lambda_{0t} = \zeta_0\left[1 + \sum_{k=1}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2 + \pi_0\right)\right] \tag{3.09}$$

$$\therefore \lambda_{0t} = \zeta_0 + \alpha_0(L)\varepsilon_t^2 + \pi_0(L)\lambda_{0t} = \frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=-0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2$$

[6] See Posedel (2005)

$$= \zeta_0 \left[1 + \sum_{k=1}^{t-2} \prod_{i=1}^{k} \left(\alpha_0 L^i z_t^2 + \pi_0 \right) \right]$$

Technically, this means that the observed model can be presented as:

$$y_t = f(x_t, \beta_0) + \varepsilon_{0t} \tag{3.10}$$

$$\varepsilon_{0t} = z_t (\lambda_{0t})^{\frac{1}{2}}, t = 1,2,.....,T$$

$$\lambda_{0t} = \zeta_0 + \alpha_0(L)\varepsilon_{0t}^2 + \pi_0(L)\lambda_{0t} = \frac{\zeta_0 (1 - \pi_0^{t-1})}{(1 - \pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2$$

$$= \zeta_0 \left[1 + \sum_{k=0}^{t-2} \prod_{i=1}^{k} \left(\alpha_0 L^i z_t^2 + \pi_0 \right) \right], \lambda_{01} = \zeta_0 \tag{3.11}$$

We define $\Omega_t = \left\{ x_1, x_2, ..., x_t, (\lambda_t)^{\frac{1}{2}} (\varepsilon_t, \varepsilon_{t-1}, ...) \right\}$ as a set of information at time t

3.2.2Assumptions of the Model

This section outlines some of the assumptions made in this analysis. They include both the regularity assumptions (i.e. standard assumptions)[7] that are necessary for identifiability, stationarity etc and some additional assumptions necessary for our analysis. To achieve the objectives, we make (following Kouassi, 2015; Choi, 2004; Chung, 2012; Posedel, 2005 and Hansen, 2006) the following assumptions:

3.2.2.1 Assumption one (A1): The innovations are martingale differences

This is one of very important assumptions in this analysis. Before we explain more on what this means, let us look at the technical explanation. To do this let us first look at some basic concepts. It must be mentioned here that these definitions are due to Ibragimov and Philips (2010), Williams (1991), Avran (1988), Hansen and Heyde (1980), Amemiya (1985), Sousi (2013) and Hansen (2006).

In martingale mathematics, a filtration on a probability space is a sequence $\{\Xi_t; t = 1,2,...\}$ of sub-sigma fields of Ξ_t such that for all t, $\Xi_t \in \Xi_{t+1}$ .Basically, in the theory of martingales, filtration represents our knowledge at successive betting times. This increases with time so that the sigma fields increase. Secondly, a stochastic process $X = \{x_t; t = 0,1,..\}$ is adapted to the filtration Ξ_t if for all t, X_t is Ξ_t -measurable. If a process is Ξ_t -measurable, it depends only on the past before t . If a process (X_t) is

[7] Check appendix D of this study and Appendix A in Engle and Gonzalez-Rivera (1991)to see some regularity conditions for QMLE to appreciate where some of these assumptions come from

adapted to Ξ_t , then each X_t depends only on what has already happened before time t . Having laid the ground like that, we can now define a martingale.

A process $X = \{x_t; t = 0,1,..\}$ is a martingale if for each $t = 0,1,2,..$

1) $\{\Xi_t; t = 1,2,...\}$ is a filtration and X is adapted to Ξ_t .

2) E $(X_t \mid \Xi_t) = X_{t-1}, a.s$

In other words, a process is a martingale if its current value depends only on what has happened before the current period. One sometimes calls such processes "non-anticipating" because, quite simply, they cannot look into the future. Secondly, since we cannot look into the future the expected value of a process today is the same as yesterday's value. It is only safe to expect that what happens tomorrow is the same as what happened today.

Therefore, by assuming that the innovations are martingales we mean that their current values depend only on what happened before the current period and that the expected value today is the same as their yesterday's value. It should be mentioned here that we are technically assuming that the innovations are martingale differences not just martingales. In this way, the value of an innovation today is taken to be the difference between today's value and yesterday's value. In other words, A process $X = \{x_t; t = 0,1,..\}$ is a martingale difference if for each $t = 0,1,2,..$ $X_t^* = X_t - LX_t$ (i.e. the value today is the same as the difference between today's and yesterday's values). It should be noted here that in this way; $E[X_t^* \mid \Xi_t] = E[X_t \mid \Xi_t] - E[LX_t \mid \Xi_t] = LX_t - LX_t = 0$ .That is, by assuming that the innovations (ε_t) are martingale differences, then;

$$E[\varepsilon_t \mid \Xi_t] = 0$$

It is appropriate to treat innovations volatility models as martingale differences since they do satisfy all the properties of martingales (Chung, 2012; Linton & Mammen, 2003; Bollerslev, 1987; Dahl & Levine, 2010). For more information on this argument see Yao (1998), and Lu (1999). This assumption will help us prove consistence by using almost sure convergence. If we assume that our innovations are dependent, we cannot be able to use convergence in probability since this needs the i.i.d assumption. However, if we assume that the innovations are martingale differences, then we will be able to show almost sure convergence under dependent innovations which will ultimately help us prove convergence in probability (Rao, 1973; Stout, 1974).

Another importance of this assumption is that it will help us show asymptotic normality in the face of dependent innovations. Normally, practitioners use the Linder-berg central limit theorem to show convergence in distribution. However, this theorem requires that

the innovations be i.i.d. In our study here, the i.i.d assumption has been relaxed which means we cannot be able to use the cerebrated Linde-berg central limit theorem. Luckily though, if we assume that our innovations are martingales we can be able to use the martingale central limit theorem[8] to show normality in the face of dependent innovations (Buhlman & McNeil, 2000).

3.2.2.2 Assumption two (A2): Differentiability and continuity

$$\forall_{x_t,\theta} f\left(x_t,\beta\right) \text{ and the likelihood functions are both continous and differentiable}$$

3.2.2.3 Assumption three (A3): Conditional Variance positivity

$$\zeta > 0; 0 \le \alpha < 1; 0 \le \pi < 1$$

This assumption ensures the positivity of both conditional and unconditional variances.

3.2.2.4 Assumption four (A4): Parameter identification

$$\lambda_{(0)t} = \zeta_{(0)}\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha_{(0)}L^i z_t^2 + \pi_{(0)}\right)\right], \&, \lambda_{(0)t} = \zeta^*_{(0)}\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha^*_{(0)}L^i z_t^2 + \pi^*_{(0)}\right)\right]$$

$$\&, f\left(x_t,\beta^*_{(0)}\right) = f\left(x_t,\beta_{(0)}\right) \Leftrightarrow c_{(0)} = c^*_{(0)}, c \in \{\zeta,\alpha,\beta,\pi\}$$

This assumption is simply saying that all parameters in GARCH (1, 1) model are identified (i.e. they can be uniquely estimated).

3.2.2.5 Assumption five (A5): Existence of the moments of innovations

$$\text{For some } \delta > 0, \exists\, S_\delta < \infty, s.t, E\left(z_t^{2+\delta}\right) \le S_\delta < \infty$$

This assumption is simply saying that all moments (i.e. the mean, the variance etc.) of the innovations do exist.

3.2.2.6 Assumption six (A6): ergodicity and distribution of innovations

z_t are ergodic process that belong to a probability law that belongs to the generalized error distribution

This is a relaxation of the normality assumption that is generally used in parametric approaches to volatility modelling.

3.2.2.7 Assumption seven (A7): Uniform boundness of the likelihood function

$$f\left(x_t,\beta\right) = O(1)$$

[8]Check Appendix C for more details on martingale central limit theorem.

This means that $f(x_t,\beta)$ is bounded uniformly. This ensures that our likelihood function is bounded as well. The boundness of the likelihood function is important in establishing estimator consistency.

3.3 The Quasi Maximum Likelihood Function

For us to be able to derive the QMLE we need a likelihood function, the Quasi Maximum Likelihood Function (QMLF, here-in-after). However, since we have both the observed and unobserved models, we will also have observed and unobserved likelihood functions. We will firstly provide the unobserved likelihood function after which we will provide the observed likelihood. These likelihood functions have been derived in this paper using the standard steps[9] of deriving likelihood functions. As such the spirit (i.e the steps) in the derivation process may be similar the spirit of derivations in other similar theoretical papers(e.g. Ross,2004; Possedel,2005; Chung,2012).Under parametric approach we would use the normal distribution function to derive the likelihood function. However, since in this paper we are assuming a generalised error distribution, we will use the generalised error distribution instead.

Assume that $\Phi=\left\{f_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right),t_t\in T,\theta\in\Theta\right\}$ is our probability model that postulate a plausible form of the joint distribution, $f_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)$of random variables associated with T observations, θ the vector of unknown parameters to be estimated, Θ the parametric space and Φ the parametric density functions. If we were maintaining the i.i.d assumption the joint density function would be presented as,

$$f_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)=\prod_{t=1}^{T}f_{Y_t}\left(y_t,\theta\right)$$

Where $f_{Y_i}\left(y_i,\theta\right)$ represent the marginal densities. In this study we are assuming that the probability distribution of innovations (and hence Y) falls in a generalised error distribution. Given this assumption, the marginal density function is presented as (see Holly &Montifort, 2010; Nielsen, 1978; McCullagh, 1994):

$$f\left(y_t,\Theta\right)=\left(\lambda 2^{\left(1+v^{-1}\right)}\Gamma\left(v^{-1}\right)\right)^{-1}v\exp\left(-0.5\left|\frac{y_t-f\left(x_t,\beta\right)}{\sigma_t\lambda}\right|^{v}\right)\quad -\infty<y_t<\infty\quad -\infty<v\leq\infty$$

This means that the joint density function could be presented as:

$$f_{Y_{t_1},Y_{t_2},....,Y_{tT}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)=\prod_{t=1}^{T}f_{Y_t}\left(y_t,\theta\right)$$

[9] See (Cameron & Trivedi, 2005) for the steps.

$$= \prod_{t=1}^{T} \left(\lambda 2^{\left(1+v^{-1}\right)} \Gamma\left(v^{-1}\right) \right)^{-1} v \exp\left(-0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} \right)$$

However, the i.i.d assumption has been relaxed in this analysis. It is therefore not possible to break the joint density function into the same simple form as above. Nevertheless, due to the intrinsic order of temporal data and the fact that the innovations have been assumed to be martingale differences, the joint density function can be written as the product of conditional densities, that is;

$$f_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta \right) = \prod_{t=1}^{T} f_{Y_{t_1} | Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_t} \mid y_{t_2}, \ldots, y_{t_{t-1}}; \theta \right)$$

But $f_{Y_{t_1} | Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_t} \mid y_{t_2}, \ldots, y_{t_{t-1}}; \theta \right) = f\left(y_t \mid \Omega_{t-1,}; \theta \right)$

$$\left(\lambda 2^{\left(1+v^{-1}\right)} \Gamma\left(v^{-1}\right) \right)^{-1} v \exp\left(-0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} \right)$$

This implies that the joint density function, $f_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta \right) = F\left(y_t; \theta, \Omega_{t-1} \right)$, will be :

$$f_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta \right) = \prod_{t=1}^{T} f_{Y_{t_1} | Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_t} \mid y_{t_2}, \ldots, y_{t_{t-1}}; \theta \right)$$

$$= \prod_{t=1}^{T} \left(\lambda 2^{\left(1+v^{-1}\right)} \Gamma\left(v^{-1}\right) \right)^{-1} v \exp\left(-0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} \right)$$

It should be noted that even though this joint density seems the same as the one under i.i.d,they are technically different.This is because this joint density function is a product of conditional densities and not just unconditional densities as it would have been under i.i.dassumption.But the joint density function is always the same as the likelihood function, denoted $L_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(\theta; y_{t_1}, y_{t_2}, \ldots, y_{t_T} \right) = L\left(\theta \mid y_t, \Omega_{t-1} \right)$ in this study. That is;

$$f_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta \right) = L_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta \right) = L\left(\Theta \mid y_t, \Omega_{t-1} \right)$$

$$\therefore L_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(\theta; y_{t_1}, y_{t_2}, \ldots, y_{t_T} \right)$$

$$= \prod_{t=1}^{T} \left(\lambda 2^{\left(1+v^{-1}\right)} \Gamma\left(v^{-1}\right)\right)^{-1} v \exp\left(-0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} \right)$$

$$\Rightarrow \ln L_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right)$$

$$= \ln \left(\prod_{t=1}^{T} \left(\lambda 2^{\left(1+v^{-1}\right)} \Gamma\left(v^{-1}\right)\right)^{-1} v \exp\left(-0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} \right) \right)$$

$$= \sum_{t=1}^{T} \ln \left(\left(\lambda 2^{\left(1+v^{-1}\right)} \Gamma\left(v^{-1}\right)\right)^{-1} v \exp\left(-0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} \right) \right)$$

$$\sum_{t=1}^{T} \left[\ln\left(\frac{v}{\lambda}\right) - 0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} - \left(1 + v^{-1}\right) \ln(2) - \ln \Gamma\left(v^{-1}\right) - 0.5 \ln\left(\sigma_t^2\right) \right]$$

$$\sigma_t^2 = \frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2$$

$$\sum_{t=1}^{T} \left[\ln\left(\frac{v}{\lambda}\right) - 0.5 \left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} - \left(1 + v^{-1}\right) \ln(2) - \ln \Gamma\left(v^{-1}\right) - 0.5 \ln\left(\sigma_t^2\right) \right]$$

$$\sum_{t=1}^{T} \left[0.5 \left(\left| \frac{y_t - f\left(x_t, \beta\right)}{\sigma_t \lambda} \right|^{v} - \ln\left(\frac{v}{\lambda}\right) \right); \ \ln\left(\frac{v}{\lambda}\right) - \left(1 + v^{-1}\right) \ln(2) - \ln \Gamma\left(v^{-1}\right) - 0.5 \ln\left(\sigma_t^2\right) \right]$$

$$\sum_{t=1}^{T} \left[0.5 \left(\left| \left(\frac{\Gamma(3/v)}{(1-\gamma-\alpha)\Gamma(1/v)} + \gamma L^k \right)^{\frac{1}{2}} \varepsilon_t - f\left(x_t, \beta\right) \right|^{v} - \ln\left(\frac{\Gamma(3/v)}{(1-\gamma-\alpha)\Gamma(1/v)} + \gamma L^k \right) \right) \right]$$

$$\text{Let } l_t(\theta) = l_t = \Omega - 0.5 \left(\left| \lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(y_t - f\left(x_t, \beta\right)\right) \right|^{v} - \ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) \right)$$

$$\Rightarrow L_T(\theta) = \ln L_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right) (\text{The log - likelihood function}) = \sum_{t=1}^{T} l_t(\theta)$$

Where $l_t(\theta) = l_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right)$ denotes log likelihood function for a single observation. We will be using $l_t(\theta)$ and $l_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right)$ interchangeably in this study.

$$\therefore \text{Unobserved Likelihood} : L_T(\theta) = \ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right) = \sum_{t=1}^{T} l_t(\theta)$$

$$= \sum_{t=1}^{T}\left[\Omega - 0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t - f(x_t,\beta)\right)\right|^{v} - \ln\left(\frac{\xi}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2\right)\right)\right]$$

The observed log-likelihood, $L_{0T}(\theta)$, is derived analogously;

$$L_{0T}(\theta) = \ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right)$$

$$= \sum_{t=1}^{T}\left[\Omega - 0.5\left(\left|\lambda^{-1}\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k \varepsilon_{0t-1}^2\right)^{-\frac{1}{2}}\left(y_t - f(x_t,\beta_0)\right)\right|^{v} - \ln\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k \varepsilon_{0t-1}^2\right)\right)\right]$$

Where, just like before, $l_{0t}(\theta) = l_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right)$ denotes the likelihood function for a single observation.

3.4 Lemmas

In order to achieve our objectives, we will provide and prove lemmas that will enable us achieve the objectives. In this section we will mainly provide these lemmas and prove them.It must be emphasized here that most of these lemmas are standard[10] lemmas that practitioners in econometrics use to establish consistency and normality. As such most if not all these lemmas can be found in many theoretical econometrics papers in the literature (e.g. (Buhlman & McNeil, 2000);(Engle & Gonzale-Rivera, October,1991);(Holly & Montifort, 2010);(Rossi, 2004);(Posedel, 2005)).The spirit of proving these lemmas is almost the same in these paper with the only difference being that different papers use different probability laws and assumptions. In this paper, a generalised error distribution has been used. As such the paper has entirely proved these lemmas using the generalised error probability law in the same spirit. That is to say, the only difference here is that we will be proving these lemmas under the assumptions of non-i.i.d and the generalised error distribution of the innovations. To achieve our objectives, we prove the following standard lemmas (see Kouassi, 2015; Choi, 2004; Chung, 2012; Hansen & Lunde, 2001; Holly & Montifort, 2010; Engle R, 1982; Bollerslev T, 1986; Rossi, 2004; Buhlman & McNeil, 2000).

3.4.1 Lemma one

$$Sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right) - \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right)\right| \xrightarrow{p} 0$$

[10]Check appendix D and Appendix A in Gonzalez-Rivera (1991) to see regularity conditions of QMLE

Remark: The following lemma technically implies that the observed likelihood function which is not stationary is asymptotically approximated by the unobserved likelihood function such that we can ignore the differences that may exist between the observed and the unobserved likelihood functions. This lemma, just like all the other lemmas below have been proved in most parametric papers (for example; Choi, 2004; Chung, 2012; Holly & Montifort, 2010) using normal distribution assumption. In this paper, we are proving it in the context of generalised error probability law. As mentioned in Kouassi (2015), this lemma may not necessarily be directly involved in proving neither consistency nor asymptotic normality, but it justifies the use of unobserved likelihood function over observed likelihood function in the other lemmas.

Proof: Technically for us to prove this lemma we just have to show that

$$\lim_{T\to\infty} prob\left(\left|\frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)-\frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)\right|>\varepsilon\right)=0$$

Using the notation used in this study, the above expression can also be written as;

$$\lim_{T\to\infty} prob\left(\left|\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)-\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)\right|>\varepsilon\right)=0$$

The unobserved single observation likelihood function is given as;

$$l_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)=\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]$$

Likewise, observed single observation likelihood function is given as;

$$l_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)=\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{\left(1-\pi_0\right)}+\alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k\varepsilon_{0t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta_0\right)\right)\right|^{v}-\ln\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{\left(1-\pi_0\right)}+\alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k\varepsilon_{0t-1}^2\right)\right)\right]$$

$$and, \text{using triangle inequality};\left|\sum_{t=1}^{T}X_t\right|\le\sum_{t=1}^{T}\left|X_t\right|$$

$$\Rightarrow\left|\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)-\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)\right|\le$$

$$\frac{1}{T}\sum_{\forall t}\left|\ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)-\ln L_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},....,y_{t_T};\theta\right)\right|$$

$$=\frac{1}{T}\sum_{\forall t}\left|\begin{array}{l}\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]\\-\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{\left(1-\pi_0\right)}+\alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k\varepsilon_{0t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta_0\right)\right)\right|^{v}-\ln\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{\left(1-\pi_0\right)}+\alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k\varepsilon_{0t-1}^2\right)\right)\right]\end{array}\right|$$

$$= \frac{1}{T}\sum_{\forall t}\left| \sum_{t=1}^{T}\left[\Omega - 0.5\left(\left| \lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} (y_t - f(x_t,\beta)) \right|^{v} - \ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) \right) \right] - \right.$$

$$\left. - \sum_{t=1}^{T}\left[\Omega - 0.5\left(\left| \lambda^{-1}\left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-\frac{1}{2}} (y_t - f(x_t,\beta_0)) \right|^{v} - \ln\left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right) \right) \right] y_t^i - \chi_3(y_t) \right|$$

$$\leq \frac{1}{T}\sum_{\forall t} |\Omega - \Omega| + \frac{1}{T}\sum_{\forall t} \left| 0.5 \left| \lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} (y_t - f(x_t,\beta)) \right|^{v} - 0.5 \left| \lambda^{-1}\left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-\frac{1}{2}} (y_t - f(x_t,\beta_0)) \right|^{v} \right|$$

$$\frac{1}{T}\sum_{\forall t} \left| -\ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) - \left(-\ln\left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right) \right) \right|$$

$$= \frac{1}{T}\sum_{\forall t} \left[0.5 \left| \lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} y_t - \left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-\frac{1}{2}} y_t \right| \right]$$

$$-\frac{1}{T}\sum_{\forall t} \left[\left| \lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} (-f(x_t,\beta)) - \left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-\frac{1}{2}} (-f(x_t,\beta_0)) \right| \right]$$

$$+ \frac{1}{T}\sum_{\forall t} \left| \ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-1} - \left(\ln\left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-1} \right) \right|$$

Since $E\left(\varepsilon_{t-k-1}^2\right), E\lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} y_t - \left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)$,

$\lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} (-f(x_t,\beta)) - \left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-\frac{1}{2}} (-f(x_t,\beta_0))$

and $\ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-1} - \left(\ln\left(\frac{\zeta_0(1-\pi_0^{t-1})}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-1} \right)$ are finite, we can apply markov inequality. According to markov inequality (see Hansen B. C, 2006): let z be a random variable with "finite" pth moment. Then,

$$\text{Prob}(|z| \geq c) \leq \frac{E|z|}{c}, c \in \mathrm{R}^{+}$$

$$\therefore \forall \varepsilon > 0, prob\left(\left| \frac{1}{T} \ln L_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}}\left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right) - \frac{1}{T} \ln L_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}(0)}\left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right) \right| > \varepsilon \right)$$

$$= \text{Prob}\left((T)^{-1} \sum_{\forall t} \left| \lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} y_t - \left(\frac{\zeta_0 \left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right) \right| > \varepsilon \right)$$

$$+ \text{Prob}\left((T)^{-1} \sum_{\forall t} \left| \lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(-f(x_t, \beta)\right) - \left(\frac{\zeta_0 \left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-\frac{1}{2}} \left(-f(x_t, \beta_0)\right) \right| > \varepsilon \right)$$

$$+ \text{Prob}\left((T)^{-1} \sum_{\forall t} \left| \ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-1} - \left(\ln\left(\frac{\zeta_0 \left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-1} \right) \right| > \varepsilon \right)$$

$$\leq \lim_{T\to\infty} E\left((T\varepsilon)^{-1} \sum_{\forall t} \left(\lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} y_t - \left(\frac{\zeta_0 \left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right) \right) \right)$$

$$+ \lim_{T\to\infty} E\left((T\varepsilon)^{-1} \sum_{\forall t} \left(\lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(-f(x_t, \beta)\right) - \left(\frac{\zeta_0 \left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-\frac{1}{2}} \left(-f(x_t, \beta_0)\right) \right) \right)$$

$$+ \lim_{T\to\infty} E\left((T\varepsilon)^{-1} \alpha \sum_{\forall t} \ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-1} - \left(\ln\left(\frac{\zeta_0 \left(1-\pi_0^{t-1}\right)}{(1-\pi_0)} + \alpha_0 \sum_{k=0}^{t-2} \pi_0^k L^k \varepsilon_{0t-1}^2 \right)^{-1} \right) \right)$$

$$= 0 + 0 + 0 - 0 + 0 + 0 - 0 = 0$$

$$\therefore \forall \varepsilon > 0, \lim_{T\to\infty} prob\left(\left| \frac{1}{T} \sum_{t=1}^{T} l_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right) - \frac{1}{T} \sum_{t=1}^{T} l_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}(0)} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right) \right| > \varepsilon \right) = 0$$

This means that;

$$Sup_{\theta\in\Theta} \left| \frac{1}{T} \sum_{t=1}^{T} l_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right) - \frac{1}{T} \sum_{t=1}^{T} l_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}(0)} \left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right) \right| \xrightarrow{p} 0$$

Alternatively, we can use the other definitions of conditional variances outlined above, to prove the same lemma. According to equations (3.03);

$$\sigma^2{}_t = \zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]\&,\sigma^2{}_{0t} = \zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]$$

This means that;

$$\left|\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)-\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right|\le$$

$$\frac{1}{T}\sum_{\forall t}\left|\ln L_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)-\ln L_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right| =$$

$$= \frac{1}{T}\sum_{\forall t}\left|\begin{array}{l}\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]\right)\right] \\ -\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta_0\right)\right)\right|^{v}-\ln\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]\right)\right]\end{array}\right|^{v}$$

$$= \frac{1}{T}\sum_{\forall t}\left|\begin{array}{l}\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]\right)\right]- \\ -\sum_{t=1}^{T}\left[\Omega-0.5\left(\left|\lambda^{-1}\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta_0\right)\right)\right|^{v}-\ln\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]\right)\right]y_t^i\end{array}\right|$$

$$\le \frac{1}{T}\sum_{\forall t}|\Omega-\Omega|+\frac{1}{T}\sum_{\forall t}\left|0.5\left|\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-0.5\left|\lambda^{-1}\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta_0\right)\right)\right|^{v}\right|$$

$$\frac{1}{T}\sum_{\forall t}\left|-\ln\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]-\left(-\ln\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]\right)\right|$$

$$= \frac{1}{T}\sum_{\forall t}\left[0.5\left|\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}}y_t-\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-\frac{1}{2}}y_t\right|\right]$$

$$-\frac{1}{T}\sum_{\forall t}\left[\left|\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}}\left(-f\left(x_t,\beta\right)\right)-\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-\frac{1}{2}}\left(-f\left(x_t,\beta_0\right)\right)\right|\right]$$

$$+\frac{1}{T}\sum_{\forall t}\left|\ln\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-1}-\left(\ln\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-1}\right)\right|$$

Since $E\left(\varepsilon_{t-k-1}^2\right), E\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}} y_t-\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]$,

$\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(-f\left(x_t,\beta\right)\right)-\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-\frac{1}{2}}\left(-f\left(x_t,\beta_0\right)\right)$

and $\ln\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-1}-\left(\ln\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{\left(1-\pi_0\right)}+\alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k\varepsilon_{0t-1}^2\right)^{-1}\right)$ are finite, we can apply markov inequality. According to markov inequality (see Hansen B. C, 2006): let z be a random variable with "finite" p-th moment. Then,

$$\text{Prob}\left(|z|\geq c\right)\leq\frac{E|z|}{c}, c\in \mathrm{R}^{+}$$

$$\therefore\forall\varepsilon>0, prob\left(\left|\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)-\frac{1}{T}\ln L_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}(0)}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right|>\varepsilon\right)$$

$$=\text{Prob}\left((T)^{-1}\sum_{\forall t}\left|\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}} y_t-\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]\right|>\varepsilon\right)$$

$$+\text{Prob}\left((T)^{-1}\sum_{\forall t}\left|\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}}\left(-f\left(x_t,\beta\right)\right)-\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-\frac{1}{2}}\left(-f\left(x_t,\beta_0\right)\right)\right|>\varepsilon\right)$$

$$+\text{Prob}\left((T)^{-1}\sum_{\forall t}\left|\ln\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-1}-\left(\ln\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-1}\right)\right|>\varepsilon\right)$$

$$\leq\lim_{T\to\infty}E\left((T\varepsilon)^{-1}\sum_{\forall t}\left(\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}} y_t-\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]\right)\right)$$

$$+\lim_{T\to\infty}E\left((T\varepsilon)^{-1}\sum_{\forall t}\left(\lambda^{-1}\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-\frac{1}{2}}\left(-f\left(x_t,\beta\right)\right)-\left(\frac{\zeta_0\left(1-\pi_0^{t-1}\right)}{\left(1-\pi_0\right)}+\alpha_0\sum_{k=0}^{t-2}\pi_0^k L^k\varepsilon_{0t-1}^2\right)^{-\frac{1}{2}}\left(-f\left(x_t,\beta_0\right)\right)\right)\right)$$

$$+\lim_{T\to\infty}E\left((T\varepsilon)^{-1}\alpha\sum_{\forall t}\ln\zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]^{-1}-\left(\ln\zeta_0\left[1+\sum_{k=0}^{t-2}\prod_{i=1}^{k}\left(\alpha_0 L^i z_t^2+\pi_0\right)\right]^{-1}\right)\right)$$

$$= 0 + 0 + 0 - 0 + 0 + 0 - 0 = 0$$

$$\therefore \forall \varepsilon > 0, \lim_{T\to\infty} prob\left(\left| \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right) - \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right) \right| > \varepsilon \right) = 0$$

This means that;

$$Sup_{\theta\in\Theta}\left| \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right) - \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}(0)}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right) \right| \xrightarrow{p} 0$$

As explained in section 4.2 in chapter four below, the implication of this lemma is that;

$$\tilde{\theta}_{SEM} = \arg\max \sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},....,Y_{t_T}}\left(y_{t_1}, y_{t_2},....,y_{t_T};\theta\right)$$

$$= \arg\max \sum_{t=1}^{T}\left[\sum_{i=0}^{2} \chi_i\left(f\left(x_t,\beta\right), \frac{\zeta}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2 \right) y_t^i + \chi_3\left(y_t\right) \right]$$

Where $\tilde{\theta}_{SEM}$ is the semi-parametric GARCH (1, 1) estimator. That is, the estimator is a vector of estimators $\tilde{\theta}_{SEM} = \left(\tilde{\beta}_{SEM} \quad \tilde{\zeta}_{SEM} \quad \tilde{\alpha}_{SEM} \quad \tilde{\pi}_{SEM}\right)'$ that maximise the unobserved likelihood function.

3.4.2 Lemma two

The Processes $\lambda_t, l_t(\theta)$ and their derivatives are strictly stationary and ergodic.

Remark: Before we present the proof let's review the concepts of strict stationarity and ergodicity. A process $\{Yt\}$, is covariance stationary if $E(Y_t)$ and $\text{cov}(Y_t, Y_{t-k}) = \gamma(k)$ are time-invariant (see Hansen B. C, 2006). A process $\{Yt\}$ is strictly stationary if the joint distribution of $(Y_t, Y_t, ..., Y_t)$ is independent of time. A stationary time series is said to be ergodic if $\gamma(k) \to 0, a.s$,as $k \to \infty$ .This loosely means that, ergodicity imply that statistical properties of a series can be deduced from a single, sufficiently long, random sample of a process. As shown in Hansen B. C (2006), if $\{Yt\}$ is strictly stationary and ergodic then $\{Xt\} = f(Yt)$ is also strictly ergodic and stationary and if $E|Xt| < \infty$ then as $T \to \infty$, $\frac{1}{T}\sum_{t=1}^{T} X_t \to E(Xt)$,a.s. The necessity of this lemma is therefore that it will help us apply the strict law of large numbers on expressions that are functions of λ_t ,for

example the likelihood function. If we can be able to apply the strong law of large numbers then we can be able to show almost sure convergence[11] which ultimately implies convergence in probability (Chung, 2012; Rao, 1973). Therefore, this lemma is very important.

Proof: To prove this lemma we just have to show that σ_t^2 and its derivatives are functions of z or ε.From assumption six z is strictly stationary and ergodic. Since ε is a function of z, it therefore means that ε is also strictly stationary and ergodic. Technically any function of ε or z will also be strictly stationary and ergodic. That is why all we need to show is that σ and its derivatives are functions of ε or z to prove that they are strictly stationary and ergodic.

$$\sigma_t^2 = \zeta\left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right] = \frac{\zeta}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2 = f\left(z_t\right)$$

This shows that σ_t^2, by being a function of z_t, a strictly stationary and ergodic variable, is ergodic and strictly stationary. For the derivatives of σ_t^2 with respect to the respective parameters, $\theta=\begin{pmatrix}\beta & \zeta & \alpha & \pi\end{pmatrix}'$;

$$\frac{\partial\sigma_t^2}{\partial\zeta} = \left[1+\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha L^i z_t^2+\pi\right)\right]+\zeta\sum_{k=0}^{\infty}\prod_{i=1}^{k}\left(\alpha\frac{\partial L^i z_t^2}{\partial\zeta}\right) = f\left(z_t\right)$$

$$Simillarly, \sigma_t^2 = \frac{\zeta}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2$$

$$\Rightarrow \frac{\partial\sigma_t^2}{\partial\alpha} = \sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2+\alpha\sum_{k=0}^{\infty}\pi^k L^k\frac{\partial\varepsilon_{t-1}^2}{\partial\alpha} = f\left(\varepsilon_t\right)$$

$$\frac{\partial\sigma_t^2}{\partial\pi} = -\frac{\zeta}{\left(1-\pi\right)^2}+\alpha\sum_{k=0}^{\infty}\left(k\pi^{k-1}\varepsilon_{t-k-1}^2+\pi^k\frac{\partial\varepsilon_{t-k-1}^2}{\partial\pi}\right) = f\left(\varepsilon_t\right)$$

$$\sigma_t^2 = \frac{\zeta}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k\varepsilon_{t-k-1}^2 = \lambda_t = \frac{\zeta}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k\left(y_t-f\left(x_t,\beta\right)\right)_{t-k-1}^2$$

$$\Rightarrow \frac{\partial\sigma_t^2}{\partial\beta} = -2\alpha\sum_{k=0}^{\infty}\left(\pi^k\left(y_t-f\left(x_t,\beta\right)\right)_{t-k-1}^2\frac{\partial f\left(x_t,\beta\right)}{\partial\beta}\right) = f\left(\varepsilon_t\right)$$

[11]Check Appendix B for more information on basics of statistical convergence.

Without loss of generality, it can be seen that even the second derivatives will be functions of ε_t process. Therefore, process σ_t^2 and its derivatives are measurable functions of an ergodic process (ε_t), and so they are also ergodic. Similarly for $l_t(\theta)$,

$$l_t(\theta) = \Omega - 0.5\left(\left| \lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(y_t - f(x_t, \beta) \right) \right|^v - \ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) \right) = f(\varepsilon_t)$$

This implies that the log-likelihood function is a function of anergodic process.This means that the likelihood function itself is ergodic.

$$\Rightarrow \frac{\partial l_t(\theta)}{\partial \beta} = \frac{\partial l_t(\theta)}{\partial f(x_t, \beta)} \frac{\partial f(x_t, \beta)}{\partial \beta}$$

$$= \frac{\partial}{\partial f(x_t, \beta)} \left[\Omega - 0.5\left(\left| \lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(y_t - f(x_t, \beta) \right) \right|^v - \ln\left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) \right) \right]$$

$$\times \left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)$$

$$= f(\varepsilon_t)$$

Just like the log-likelihood function, its first derivative here is also a function of an ergodic process. This means that it itself is ergodic as well.

$$\Rightarrow \frac{\partial^2 l_t(\theta)}{\partial \beta \partial \beta'} = \frac{\partial}{\partial \beta'} \left[\frac{\partial l_t(\theta)}{\partial \beta} \right]$$

$$= \frac{\partial}{\partial \beta'} \left[\frac{\partial \left[\Omega - 0.5 \left(\left| \lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(y_t - f(x_t, \beta) \right) \right|^{v} - \ln \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) \right) \right]}{\partial f(x_t, \beta)} \times \frac{\partial f(x_t, \beta)}{\partial \beta} \right]$$

$$= \frac{\partial}{\partial f(x_t, \beta)} \left[\frac{\Omega - 0.5 \left(\left| \lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(y_t - f(x_t, \beta) \right) \right|^{v} - \ln \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) \right)}{\partial f(x_t, \beta)} \times \frac{\partial f(x_t, \beta)}{\partial \beta} \right] \times \frac{\partial f(x_t, \beta)}{\partial \beta'}$$

$$= f(\varepsilon_t)$$

Without loss of generality, this also implies that the second derivative of the likelihood function is also ergodic by virtue of being a function of an ergodic process.

$$\frac{\partial l_t(\theta)}{\partial \zeta} = \frac{\partial l_t(\theta)}{\partial \sigma_t^2} \frac{\partial \sigma_t^2}{\partial \zeta}$$

$$= \frac{\partial}{\partial \sigma_t^2} \left[\Omega - 0.5 \left(\left| \lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(y_t - f(x_t, \beta) \right) \right|^{v} - \ln \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right) \right) \right]$$

$$\times \left(\frac{\partial \sigma_t^2}{\partial \zeta} \right)$$

$$= \left[\left[\frac{5}{32} v \sigma_t^2 \left(\left| \lambda^{-1} \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-\frac{1}{2}} \left(y_t - f(x_t, \beta) \right) \right|^{v-1} - \left(\frac{\xi}{1-\pi} + \alpha \sum_{k=0}^{\infty} \pi^k L^k \varepsilon_{t-1}^2 \right)^{-1} \right) \right] \right]$$

$$\times\left(\left(\frac{1}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}\,\frac{\partial\varepsilon_{t-k-1}^{2}}{\partial\zeta}\right)\right)$$

$$=f\left(\varepsilon_{t}\right)$$

$$\frac{\partial^{2}l_{t}\left(\theta\right)}{\partial\zeta\partial\zeta'}=\frac{\partial}{\partial\zeta'}\left[\frac{\partial l_{t}\left(\theta\right)}{\partial\zeta}\right]$$

$$=\frac{\partial}{\partial\zeta'}\left[\left[\left[\frac{5}{32}v\sigma_{t}^{2}\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-\frac{1}{2}}\left(y_{t}-f\left(x_{t},\beta\right)\right)\right|^{v-1}-\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-1}\right)\right]\right]\left(\left(\frac{1}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}\,\frac{\partial\varepsilon_{t-k-1}^{2}}{\partial\zeta}\right)\right)\right]$$

$$\times\left(\frac{\partial\left[\dfrac{\zeta}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right]}{\partial\zeta'}\right)=f\left(\varepsilon_{t}\right)$$

$$\frac{\partial l_{t}\left(\theta\right)}{\partial\alpha}=\frac{\partial l_{t}\left(\theta\right)}{\partial\lambda_{t}}\frac{\partial\sigma_{t}^{2}}{\partial\alpha}$$

$$=\frac{\partial}{\partial\sigma_{t}^{2}}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-\frac{1}{2}}\left(y_{t}-f\left(x_{t},\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)\right)\right]$$

$$\times\left(\frac{\partial\sigma_{t}^{2}}{\partial\alpha}\right)$$

$$=\left[\left[\frac{5}{32}v\sigma_{t}^{2}\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-\frac{1}{2}}\left(y_{t}-f\left(x_{t},\beta\right)\right)\right|^{v-1}-\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-1}\right)\right]\right]$$

$$\times\left(\sum_{k=0}^{\infty}\pi^{k}\varepsilon_{t-k-1}^{2}+\alpha\sum_{k=0}^{\infty}\pi^{k}\frac{\partial\varepsilon_{t-k-1}^{2}}{\partial\alpha}\right)$$

$$\frac{\partial^{2}l_{t}(\theta)}{\partial\alpha'\partial\alpha}=\frac{\partial}{\alpha'}\left[\frac{\partial l_{t}(\theta)}{\partial\alpha}\right]=$$

$$\frac{\partial}{\partial\alpha'}\left[\left[\left[\frac{5}{32}v\sigma_{t}^{2}\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-\frac{1}{2}}\left(y_{t}-f\left(x_{t},\beta\right)\right)\right|^{v-1}-\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-1}\right)\right]\right]\left(\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}+\alpha\sum_{k=0}^{\infty}\pi^{k}\frac{\partial\varepsilon_{t-k-1}^{2}}{\partial\alpha}\right)\right]$$

$$=f\left(\varepsilon_{t}\right)$$

$$\frac{\partial l_{t}(\theta)}{\partial\pi}=\frac{\partial l_{t}(\theta)}{\partial\sigma_{t}^{2}}\frac{\partial\sigma_{t}^{2}}{\partial\pi}$$

$$=\frac{\partial}{\partial\sigma_{t}^{2}}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-\frac{1}{2}}\left(y_{t}-f\left(x_{t},\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)\right)\right]\times$$

$$\left(-\frac{\zeta}{(1-\pi)^{2}}+\alpha\sum_{k=0}^{\infty}\left(k\pi^{k-1}\varepsilon_{t-k-1}^{2}+\pi^{k}\frac{\partial\varepsilon_{t-k-1}^{2}}{\partial\pi}\right)\right)$$

$$=f\left(\varepsilon_{t}\right)$$

$$\frac{\partial^{2}l_{t}(\theta)}{\partial\pi'\partial\pi}=\frac{\partial\left[\frac{\partial}{\partial\lambda_{t}}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)^{-\frac{1}{2}}\left(y_{t}-f\left(x_{t},\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^{k}L^{k}\varepsilon_{t-1}^{2}\right)\right)\right]\right]\times}{\partial\pi'}$$

$$\left[\left(-\frac{\zeta}{(1-\pi)^{2}}+\alpha\sum_{k=0}^{\infty}\left(k\pi^{k-1}\varepsilon_{t-k-1}^{2}+\pi^{k}\frac{\partial\varepsilon_{t-k-1}^{2}}{\partial\pi}\right)\right)\right]+$$

$$\frac{\partial}{\partial\pi'}\left[\left(-\frac{\zeta}{(1-\pi)^{2}}+\alpha\sum_{k=0}^{\infty}\left(k\pi^{k-1}\varepsilon_{t-k-1}^{2}+\pi^{k}\frac{\partial\varepsilon_{t-k-1}^{2}}{\partial\pi}\right)\right)\right]\times$$

$$\left[\frac{\partial}{\partial\lambda_t}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]\right]$$

$$=f\left(\varepsilon_t\right)$$

This implies that ψ and its derivatives are functions of k and/or z proving that they are strictly stationary and ergodic. The above lemma implies that,

$$\tfrac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\to E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]$$

$$E\left|\nabla l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right|<\infty,\&E\left|\nabla^2 l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right|<\infty$$

So we will treat these as our next lemmas.

3.4.3 Lemma three

$$\boxed{\frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\xrightarrow{p} E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]}$$

Remark: This is to say that our criterion to maximize the likelihood function converges in probability to a non-stochastic function $E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]$ .This is a necessary condition for the convergence of QMLE.

Proof: From lemma two above, $E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]=f\left(\varepsilon_t\right)$ which is stationary, implying $E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]$ is finite, guarantying the conditions for applying the strong law of large numbers. By the strong law of large numbers,

$$\frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\xrightarrow{a.s} E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]$$

Since almost sure convergence imply convergence in probability (Rao, 1973), then

$$\frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\xrightarrow{p} E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]$$

For formality sake, let's show indeed that $E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]=f\left(\varepsilon_t\right)$

$$l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)=\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)$$

$$E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]=E\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]$$

$$=f\left(\varepsilon_t\right)$$

So $E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]=f\left(\varepsilon_t\right)$, which implies that $E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]<\infty,$ since ε_t is stationary. This technically means that we can apply the strong law of large numbers (SLL). That is;.

$$\frac{1}{T}\sum_{t=1}^{T}l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\xrightarrow{a.s}E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]$$

$$\Rightarrow\frac{1}{T}\sum_{t=1}^{T}l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\xrightarrow{p}E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right]$$

It must be mentioned here that without A1, A6, and lemma 2 above, we could not be able to show this lemma. We have derived convergence in probability of the likelihood function to its expected value through almost sure convergence which mainly depends on the assumptions of strict stationarity, ergodicity and martingales. This lemma will be heavily used in theorem 1 below when we will be showing consistency of the estimator.

3.4.4 Lemma four

$$\boxed{E\left|\nabla l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right|<\infty,\&E\left|\nabla^2 l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right|<\infty}$$

Remark: What this lemma is saying is that the absolute score function together with its derivatives is bounded. This is a sufficient condition for convergence of QMLE (Hood & Koopman, 1953).

Proof:

$$l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)=\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)$$

$$\Rightarrow E\left|\frac{\partial l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)}{\partial\beta}\right| = E\left|\frac{\partial}{\partial\beta}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]\right|$$

$$= E\left|\frac{\partial}{\partial f\left(x_t,\beta\right)}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]\frac{\partial f\left(x_t,\beta\right)}{\partial\beta}\right|$$

$$= E\left[\left|\frac{\partial}{\partial f\left(x_t,\beta\right)}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]\right|\right]$$

$$\times\left[\left|\frac{\partial f\left(x_t,\beta\right)}{\partial\beta}\right|\right]$$

$$\leq\left|\frac{\partial}{\partial f\left(x_t,\beta\right)}\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k E\left(\varepsilon_{t-1}^2\right)\right)^{-\frac{1}{2}}\left(E\left(y_t\right)-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k E\left(\varepsilon_{t-1}^2\right)\right)\right)\right]\right|$$

$$\times\left[\left|\frac{\partial f\left(x_t,\beta\right)}{\partial\beta}\right|\right]$$

$$=\left|\frac{\partial}{\partial f\left(x_t,\beta\right)}\left[\Omega-0.5\left(-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k E\left(\varepsilon_{t-1}^2\right)\right)\right)\right]\right|\times\left[\left|\frac{\partial f\left(x_t,\beta\right)}{\partial\beta}\right|\right]$$

By assumption seven and assumption two it implies that;

$$\frac{\partial f\left(x_t,\beta\right)}{\partial\beta}=O(1)\ \&\ E\left(\varepsilon_t^2\right)=O(1)\Rightarrow E\left|\frac{\partial l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)}{\partial\beta}\right|<\infty$$

Similarly,

$$E\left|\frac{\partial l_t(\theta)}{\partial \pi}\right| = E\left|\frac{\partial}{\partial \pi}\left[\Omega - 0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\left(\varepsilon_{t-1}^2\right)\right)^{-\frac{1}{2}}\left(y_t - f(x_t,\beta)\right)\right|^{v} - \ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\left(\varepsilon_{t-1}^2\right)\right)\right)\right]\right|$$

$$= E\left|\frac{\partial}{\partial \sigma_t^2}\left[\Omega - 0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\left(\varepsilon_{t-1}^2\right)\right)^{-\frac{1}{2}}\left(y_t - f(x_t,\beta)\right)\right|^{v} - \ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\left(\varepsilon_{t-1}^2\right)\right)\right)\right]\frac{\partial \sigma_t^2}{\partial \pi}\right|$$

$$= E\left[\left|\frac{\partial}{\partial \sigma_t^2}\left[\Omega - 0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k E\left(\varepsilon_{t-1}^2\right)\right)^{-\frac{1}{2}}\left(y_t - f(x_t,\beta)\right)\right|^{v} - \ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\left(\varepsilon_{t-1}^2\right)\right)\right)\right]\right|\left|\frac{\partial \sigma_t^2}{\partial \pi}\right|\right]$$

$$\leq \left[\left[\frac{5}{32}v\sigma_t^2\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k E\left(\varepsilon_{t-1}^2\right)\right)^{-\frac{1}{2}}\left(E(y_t) - f(x_t,\beta)\right)\right|^{v-1} - \left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k E\left(\varepsilon_{t-1}^2\right)\right)^{-1}\right)\right]\right]\times\left(-\frac{\zeta}{(1-\pi)^2}+\alpha\sum_{k=0}^{\infty}\left(k\pi^{k-1}E\left(\varepsilon_{t-k-1}^2\right)+\pi^k\frac{\partial E\left(\varepsilon_{t-k-1}^2\right)}{\partial \pi}\right)\right)$$

Again by assumption seven; $\Rightarrow E\left|\frac{\partial l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)}{\partial \pi}\right| < \infty$

Without loss of generality we can do the same for all the other parameters and indeed for the second derivatives.

$$\therefore E\left|\nabla l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right| < \infty, \& E\left|\nabla^2 l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1},y_{t_2},\ldots,y_{t_T};\theta\right)\right| < \infty$$

Lemma 3 and lemma 4 are necessary and sufficient conditions for the convergence of QMLE respectively.

3.4.5 Lemma five

$$\boxed{\nabla\sum_{t=0}^{T} l_t\left(\tilde{\theta}_{SEM}\right) = 0 = \nabla\sum_{t=0}^{T} l_t(\theta_0) + \left(\tilde{\theta}_{SEM} - \theta_0\right)\nabla\sum_{t=0}^{T} l_t(\theta^*), \text{ where } \theta^* \text{is between } \tilde{\theta}_{SEM} \ \& \ \theta_0}$$

Remark: What this lemma is saying is that, we can analyse the asymptotic behaviour of the likelihood function by simply analysing the behaviour of the right hand side of lemma 5 above.

Proof: According to the mean value theorem, Let $f : \mathrm{R}^k$ be defined on an open convex set $\Theta \subset \mathrm{R}^k$ such that f is continuously differentiable on Θ . Then there exists θ^* in the interval, $[\theta - \theta_0]$ such that $f(\theta^*) = \frac{f(\theta_0) - f(\theta)}{\theta_0 - \theta}$) (Avran, 1988).In our case, $\nabla l_t(\tilde{\theta}_{SEM})$ is differentiable in the interval $[\theta_0 - \tilde{\theta}_{SEM}]$ such that its mean value expansion about θ^* is:

$$\nabla^2 \sum_{t=0}^{T} l_t(\theta^*) = (\tilde{\theta}_{SEM} - \theta_0)^{-1} \left(\nabla \sum_{t=0}^{T} l_t(\tilde{\theta}_{SEMT}) - \nabla \sum_{t=0}^{T} l_t(\theta_0) \right) \Rightarrow (\tilde{\theta}_{SEM} - \theta_0) \nabla^2 \sum_{t=0}^{T} l_t(\theta^*)$$

$$= \left(\nabla \sum_{t=0}^{T} l_t(\tilde{\theta}_{SEM}) - \nabla \sum_{t=0}^{T} l_t(\theta_0) \right)$$

$$\Rightarrow \nabla \sum_{t=0}^{T} l_t(\tilde{\theta}_{SEM}) = (\tilde{\theta}_{SEM} - \theta_0) \nabla^2 \sum_{t=0}^{T} l_t(\theta^*) + \nabla \sum_{t=0}^{T} l_t(\theta_0), \text{but} \nabla \sum_{t=0}^{T} l_t(\tilde{\theta}_{SEM}) = 0$$

$$\Rightarrow \nabla \sum_{t=0}^{T} l_t(\tilde{\theta}_{SEM}) = 0 = \nabla \sum_{t=0}^{T} l_t(\theta_0) + (\tilde{\theta}_{SEM} - \theta_0) \nabla^2 \sum_{t=0}^{T} l_t(\theta^*) \Rightarrow \nabla \sum_{t=0}^{T} l_t(\tilde{\theta}_{SEM}) = 0$$

$$= \nabla \sum_{t=0}^{T} l_t(\theta_0) + (\tilde{\theta}_{SEM} - \theta_0) \nabla \sum_{t=0}^{T} l_t(\theta^*)$$

This proves that;

$$\nabla \sum_{t=0}^{T} l_t(\tilde{\theta}_{SEM}) = 0 = \nabla \sum_{t=0}^{T} l_t(\theta_0) + (\tilde{\theta}_{SEM} - \theta_0) \nabla \sum_{t=0}^{T} l_t(\theta^*)$$

This technically implies that the asymptotic normality of $\tilde{\theta}_{SEM}$ is determined by the RHS of lemma 5;

$$\nabla \sum_{t=0}^{T} l_t(\theta_0) + (\tilde{\theta}_{SEM} - \theta_0) \nabla \sum_{t=0}^{T} l_t(\theta^*)$$

3.4.6 Lemma six

$$\boxed{Sup_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \left| \nabla^2 \sum_{t=0}^{T} l_t(\theta^*) - E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta^*) \right) \right| \xrightarrow{p} 0}$$

Remark: This Lemma is saying that $\nabla^2 \sum_{t=0}^{T} l_t(\theta^*)$ obeys the weak uniform law of large numbers

Proof: Before we prove this; let us review some concepts in asymptotic theory that will be necessary in this section. Now, we say that $q_{T_t(Z_t;\theta)}$ obeys strong uniform law of large numbers, SULLN if;

$$Sup_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \left| q_{T_t(Z_t;\theta)} - E\left(q_{T_t(Z_t;\theta)} \right) \right| \to 0, a.s$$

On the other hand, $q_{T_t(Z_t;\theta)}$ is said to obey WULLN if the convergence condition above holds in probability. That is; $Sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|q_{T_t(Z_t;\theta)}-E\left(q_{T_t(Z_t;\theta)}\right)\right|\xrightarrow{p}0$

Here we want to show that; $Sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\nabla^2\sum_{t=0}^{T}l_t\left(\theta^*\right)-E\left(\nabla^2\sum_{t=0}^{T}l_t\left(\theta^*\right)\right)\right|\xrightarrow{p}0$. There are two conditions that a sequence, like $\nabla^2\sum_{t=0}^{T}l_t\left(\theta^*\right)$, must satisfy in order to obey WULLN (Posedel, 2005); first, $\nabla^2\sum_{t=0}^{T}l_t\left(\theta^*\right)$ should follow the weak law of large numbers, WLLN. Second, $\nabla^2\sum_{t=0}^{T}l_t\left(\theta^*\right)$, should be Lipchitz continuous. Let's start with WLLN;

$$l_t(\theta)=\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta\right)\right)\right|^{v}-\ln\left(\frac{\xi}{1-\pi}+\alpha\sum_{k=0}^{\infty}\pi^k L^k\varepsilon_{t-1}^2\right)\right)\right]$$

$$\nabla^2\sum_{t=0}^{T}l_t\left(\theta^*\right)=\sum_{t=0}^{T}\nabla^2 l_t\left(\theta^*\right)=\sum_{t=0}^{T}\nabla^2\left[\Omega-0.5\left(\left|\lambda^{-1}\left(\frac{\xi^*}{1-\pi^*}+\alpha^*\sum_{k=0}^{\infty}\pi^{*k}L^k\varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t-f\left(x_t,\beta^*\right)\right)\right|^{v}-\ln\left(\frac{\xi^*}{1-\pi^*}+\alpha^*\sum_{k=0}^{\infty}\pi^{*k}L^k\varepsilon_{t-1}^2\right)\right)\right]$$

But according to lemma two, $l_t(\theta),\nabla\sum_{t=1}^{T}l_t(\theta),\&\nabla^2\sum_{t=1}^{T}l_t(\theta)$ are ergodic. This implies that $l_t\left(\theta^*\right),\nabla\sum_{t=1}^{T}l_t\left(\theta^*\right),\&\nabla^2\sum_{t=1}^{T}l_t\left(\theta^*\right)$ are also ergodic processes. This implies that $E(l_t(\theta)),E(\nabla l_t(\theta)),\&E\left(\nabla^2 l_t(\theta)\right)$ are finite. The fact that $E\left(\nabla^2 l_t(\theta)\right)<\infty$ implies that we can apply the strong law of large numbers on $E\left(\nabla^2 l_t(\theta)\right)$

$$\frac{1}{T}\sum_{t=0}^{T}\nabla^2 l_t\left(\theta^*\right)\to E\left(\nabla^2 l_t(\theta)\right),a.s\Rightarrow\frac{1}{T}\sum_{t=1}^{T}\left[\left|\sum_{t=0}^{T}\nabla^2 l_t\left(\theta^*\right)-E\left(\nabla^2 l_t(\theta)\right)\right|\right]\to 0,a.s$$

$$\text{But }\frac{1}{T}\sum_{t=1}^{T}\left[\left|\sum_{t=0}^{T}\nabla^2 l_t\left(\theta^*\right)-E\left(\nabla^2 l_t(\theta)\right)\right|\right]\to 0,a.s\Rightarrow\frac{1}{T}\sum_{t=1}^{T}\left[\left|\sum_{t=0}^{T}\nabla^2 l_t\left(\theta^*\right)-E\left(\nabla^2 l_t(\theta)\right)\right|\right]\xrightarrow{p}0$$

$$\therefore\frac{1}{T}\sum_{t=1}^{T}\left[\left|\sum_{t=0}^{T}\nabla^2 l_t\left(\theta^*\right)-E\left(\nabla^2 l_t(\theta)\right)\right|\right]\xrightarrow{p}0$$

Let's now look at (ii), Lipchitz continuity. A function f from $S\in\mathrm{R}^n$ is Lipchitz continuous at $x\in S$ if there is a constant C such that (see Posedel, 2005);

$$\left|f(y)-f(x)\right|\le C\left\|y-x\right\|,a.s$$

For any $y\in S$ sufficiently near, C is a random variable bounded almost surely and $\|.\|$ is the Euclidean norm

Here we want to show that $\left|\sum_{t=0}^{T}\nabla^2 l_t(\theta)-\sum_{t=0}^{T}\nabla^2 l_t(\theta^*)\right|\le C\left\|\theta-\theta^*\right\|$

For any f, real valued function, defined and differentiable on the interval $I\in \mathrm{R}$ .If f is bounded on I, then f is a Lipchitz function on I .So, any differentiable function is Lipchitz. One of our assumptions is that the likelihood function is differentiable. This means therefore that $\nabla L_T(\theta)$ is Lipchitz.

$$\therefore\left|\sum_{t=0}^{T}\nabla^2 l_t(\theta)-\sum_{t=0}^{T}\nabla^2 l_t(\theta^*)\right|\le C\left\|\theta-\theta^*\right\|$$

$$\text{Given that } \frac{1}{T}\sum_{t=1}^{T}\left[\left|\sum_{t=0}^{T}\nabla^2 l_t(\theta^*)-E\left(\nabla^2 l_t(\theta)\right)\right|\right]\xrightarrow{p}0,\&\left|\sum_{t=0}^{T}\nabla^2 l_t(\theta)-\sum_{t=0}^{T}\nabla^2 l_t(\theta^*)\right|\le C\left\|\theta-\theta^*\right\|$$

$$\therefore Sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\nabla^2\sum_{t=0}^{T}l_t(\theta^*)-E\left(\nabla^2\sum_{t=0}^{T}l_t(\theta^*)\right)\right|\xrightarrow{p}0$$

Indeed $\nabla^2\sum_{t=0}^{T}l_t(\theta^*)$ obeys the weak uniform law of large numbers, WULLN.

3.4.7 Lemma seven

$$\boxed{\sqrt{T}\left(\tilde{\theta}_{SEM}-\theta_0\right)=-\left[E\left(\nabla^2\sum_{t=0}^{T}l_t(\theta)\right)\right]^{-1}\sqrt{T}\nabla\sum_{t=0}^{T}l_t(\theta_0)}$$

Proof: From lemma 6 above,

$$Sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\left|\nabla^2\sum_{t=0}^{T}l_t(\theta^*)-E\left(\nabla^2\sum_{t=0}^{T}l_t(\theta^*)\right)\right|\xrightarrow{p}0\Rightarrow\left|\nabla^2\sum_{t=0}^{T}l_t(\theta^*)-E\left(\nabla^2\sum_{t=0}^{T}l_t(\theta^*)\right)\right|\to 0$$

Using lemma 5;

$$\nabla\sum_{t=0}^{T}l_t\left(\tilde{\theta}_{SEM}\right)=0=\nabla\sum_{t=0}^{T}l_t(\theta_0)+\left(\tilde{\theta}_{SEM}-\theta_0\right)\nabla\sum_{t=0}^{T}l_t(\theta^*)\Rightarrow 0=\nabla\sum_{t=0}^{T}l_t(\theta_0)+\left(\tilde{\theta}_{SEM}-\theta_0\right)\nabla\sum_{t=0}^{T}l_t(\theta^*)$$

$$\left(\tilde{\theta}_{SEM}-\theta_0\right)\nabla\sum_{t=0}^{T}l_t(\theta^*)=-\nabla\sum_{t=0}^{T}l_t(\theta_0)\Rightarrow\left(\tilde{\theta}_{SEM}-\theta_0\right)=-\left[E\left(\nabla\sum_{t=0}^{T}l_t(\theta_0)\right)\right]^{-1}\nabla\sum_{t=0}^{T}l_t(\theta_0)$$

Multiplying through by $\sqrt{T}$;

$$\sqrt{T}\left(\tilde{\theta}_{SEM}-\theta_0\right)=-\left[E\left(\nabla^2\sum_{t=0}^{T}l_t(\theta)\right)\right]^{-1}\sqrt{T}\nabla\sum_{t=0}^{T}l_t(\theta_0)$$

This technically implies that the asymptotic distribution of $\sqrt{T}\left(\tilde{\theta}_{SEM} - \theta_0\right)$ is determined by the asymptotic distribution of the normalized score; $\nabla \sum_{t=0}^{T} l_t(\theta_0)$

4. THEORETICAL RESULTS AND DISCUSSION

4.1 Introduction

In this section we present the main results of our analysis. Given the assumptions and the lemmas above, we show in this chapter we derive the semi-parametric GARCH (1, 1) estimator and then prove that it is not only consistent but also asymptotically normal. We present these results in the following result and theorems.

Before proceeding any further, it is worth mentioning here that most techniques used in the following proofs are standard when proving consistency and asymptotic normality of a vector of parameters given that the innovations are being allowed to be serially dependent[12].One may use/may have used the same techniques with a different probability law (say the normal distribution, the student t distribution, the gamma distribution among others) and/or with a different volatility models(e.g. the ARCH(2) model, the Exponential GARCH(1,2) etc) altogether (see Posedel, 2005; Holly & Montifort, 2010). This paper has entirely proved the following theorems using the generalised error distribution in the realm of GARCH (1, 1) volatility model given that the innovations are being allowed

[12] Check Posedel(2005) for more information.

to be serially independent. Where specific ideas have been taken from someone/somewhere, relevant referencing has accordingly been made.

4.2 The semi-parametric GARCH (1, 1) estimator

This paper is trying to derive the semi parametric GARCH (1, 1) estimator using quasi-maximum likelihood estimation technique. From equation (2.19) in section 2.2 of chapter two, it was noted that in the spirit of Kullback-Leiber Information Criterion (KLIC) the quasi-maximum likelihood estimator, $\hat{\theta}_{QMLE}$, is given as shown in equation (4.01) below(everything as defined in section 2.2 above);

$$\hat{\theta}_{QMLE} = \arg\min KLIC = \arg\min \sum_{t=1}^{T} \ln\left(\frac{\zeta(\varepsilon)}{\zeta(\varepsilon|\theta)}\right)\zeta(\varepsilon)\partial\varepsilon = E\left[\ln \zeta(\varepsilon|\theta)\right] \tag{4.01}$$

It can be noted here that equation (4.01) implies that the quasi-maximum likelihood estimator is the one that minimises the KLIC. But, minimising the KLIC is the same as maximising the unobserved function.

Therefore, using lemma 1, the semi parametric GARCH (1, 1) estimator, $\tilde{\theta}_{SEM}$ is the one that maximises the unobserved likelihood function as given in equation (4.02). That is,

$$\tilde{\theta}_{SEM} = \arg\max \sum_{t=1}^{T} l_{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_T}}\left(y_{t_1}, y_{t_2}, \ldots, y_{t_T}; \theta\right)$$

$$= \arg\max \sum_{t=1}^{T}\left[\Omega - 0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t - f(x_t, \beta)\right)\right|^{v} - \ln\left(\frac{\xi}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2\right)\right)\right] \tag{4.02}$$

So the semi-parametric GARCH (1, 1) under serially dependent innovations is given as;

$$\tilde{\theta}_{SEM} = \arg\max \sum_{t=1}^{T}\left[\Omega - 0.5\left(\left|\lambda^{-1}\left(\frac{\xi}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2\right)^{-\frac{1}{2}}\left(y_t - f(x_t, \beta)\right)\right|^{v} - \ln\left(\frac{\xi}{1-\pi} + \alpha\sum_{k=0}^{\infty}\pi^k L^k \varepsilon_{t-1}^2\right)\right)\right]$$

4.3 Theorem One

Given the above assumptions and lemmas,

$$\boxed{\tilde{\theta}_{SEM} \to \theta_0, a.s \Rightarrow \tilde{\theta}_{SEM} = \theta_0 + o_p(1)}$$

Remark: This theorem is saying that, given all the assumptions outlined above and all the lemmas proved above, the estimator is consistent. In other words, the estimator converges almost surely to the true population parameter as we increase the sample size indefinitely.

Proof: Before we prove this, let's look at some basic mathematical concepts necessary in this section. An adherent point (also known as closure point or point of closure or contact point) of subset A of a topological space X, is a point x in X such that every open set containing x contains at least one point of A (Hansen, 2004).Any compact set has an adherent point (Stout, 1974; Hansen, 2006) Now consider the finite series of our estimator $\left(\tilde{\theta}_{SEM}\right)$ defined on Θ. Since Θ is compact by assumption, there exists an adherent point. Let this adherent point be $\left(\theta_{0(T)}\right)$. There exists a sub-sequence of estimators $\left(\tilde{\theta}_{\psi(SEM)}\right)$ such that $\left(\tilde{\theta}_{\psi(SEM)}\right)\to\left(\theta_{0(T)}\right), a.s$ where $\psi(SEM)$ is an increasing injective function. From lemma 1 above, we can see that;

$$\tilde{\theta}_{SEM} = \arg\max_{\forall\theta\in\Theta}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}(0)}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right) = \arg\max\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right)$$

$$\Rightarrow \tilde{\theta}_{SEM} = \arg\max\frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right)$$

$$\therefore \forall\theta\in\Theta, \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right) \le \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\tilde{\theta}_{SEM}\right)$$

$$\Rightarrow \forall\theta\in\Theta, \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right) \le \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\tilde{\theta}_{\psi(SEM)}\right)$$

$$\Rightarrow \forall\theta\in\Theta, \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right) \le \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\tilde{\theta}_{\psi(SEM)}\right).$$

$$\text{But } \left(\hat{\theta}_{\psi(T)}\right)\to\left(\theta_{0(T)}\right), a.s \;\&\; \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right) \to E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right)\right]$$

$$\Rightarrow \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right) \to E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)\right], a.s,$$

$$\& \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\hat{\theta}_{\psi(SEM)}\right) \to E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_{0(T)}\right)\right]$$

$$\text{Hence, } \forall\theta\in\Theta, \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right) \le \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{t_T}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\tilde{\theta}_{\psi(SEM)}\right),$$

$$\text{becomes}: \forall \theta \in \Theta, E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)\right] \le E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_{0(T)}\right)\right], a.s$$

$$\text{By definition of } \theta_0, \forall \theta \in \Theta, \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right) \le \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right),$$

$$and; \left(\tilde{\theta}_{\psi(SEM)}\right) \to \left(\theta_{0(T)}\right), a.s \;\&\; \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right) \to E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right)\right]$$

$$\Rightarrow \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right) \to E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)\right], a.s,$$

$$\& \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\tilde{\theta}_{\psi(SEM)}\right) \to E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_{0(T)}\right)\right]$$

$$\therefore \forall \theta \in \Theta, \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right) \le \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right),$$

$$\text{can be written as}; \forall \theta \in \Theta, E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta\right)\right] \le E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)\right], a.s$$

$$\theta = \theta_{0(T)} \Rightarrow \forall \theta \in \Theta, \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_{0(T)}\right) \le \frac{1}{T}\sum_{t=1}^{T} l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)$$

$$\Rightarrow \forall \theta \in \Theta, E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_{0(T)}\right)\right] \le E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)\right], a.s$$

But we just showed that,

$\forall \theta \in \Theta, E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)\right] \le E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_{0(T)}\right)\right], a.s,$ and now we have shown that

$\forall \theta \in \Theta, E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_{0(T)}\right)\right] \le E\left[l_{Y_{t_1},Y_{t_2},\ldots,Y_{tT}}\left(y_{t_1}, y_{t_2},\ldots,y_{t_T};\theta_0\right)\right], a.s \Rightarrow \theta_{0(T)} = \theta_0$

$. \text{Since } \left(\tilde{\theta}_{\psi(SEM)}\right) \to \left(\theta_{0(T)}\right), a.s \Rightarrow \tilde{\theta}_{SEM} \to \theta_0, a.s$

But almost sure convergence implies convergence in probability (see Rao, 1973; Stout, 1974). This technically implies that;

$$\tilde{\theta}_{SEM} = \theta_0 + o_p(1), i.e. \forall \varepsilon > 0, \lim_{T\to\infty} \text{Prob}\left(\left|\tilde{\theta}_{SEM} - \theta_0\right| > \varepsilon\right) = 0$$

This proves that, $\tilde{\theta}_{SEM} \to \theta, a.s \Rightarrow \tilde{\theta}_{SEM} = \theta_0 + o_p(1)$. This technically means that $\tilde{\theta}_{SEM}$ is consistent.

4.4 Theorem Two

Given the above assumptions and lemmas,

$$\boxed{\tilde{\theta}_{SEM} \xrightarrow{d} N\left(\theta_0, \left[E\left(\nabla^2 \sum_{t=1}^{T} l_t(\theta)\right)\right]^{-1} \text{var}\left(\nabla \sum_{t=1}^{T} l_t(\theta)\right)\left(\left[E\left(\nabla^2 \sum_{t=1}^{T} l_t(\theta)\right)\right]'\right)^{-1}\right)}$$

Remark: This theorem is saying that, given all the assumptions outlined above and all the lemmas proved above, the estimator is asymptotically normal. In other words, the estimator converges in distribution to the normal distribution as we increase the sample size indefinitely.

Proof: First we note that $\text{var}\left(\sqrt{T}\nabla \sum_{t=1}^{T} l_t(\theta)\right) \leq \infty$ .This is because $\text{var}\left(\nabla l_t(\theta)\right) \leq \infty$ as we have already shown above that $\nabla l_t(\theta)$ is a function of ε process that is strictly stationary and ergodic which its first and second moments exist. This means that we can apply the martingale central limit theorem;

$$\left[\text{var}\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right)\right]^{0.5}\left[\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right) - E\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right)\right] \xrightarrow{d} N(0,1)$$

But $E\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right) = 0$, since (θ_0) maximizes the score function globally. This implies

$$\text{that}\left[\text{var}\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right)\right]^{0.5}\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right) \xrightarrow{d} N(0,1)$$

From lemma seven above, $\sqrt{T}\left(\tilde{\theta}_{SEM} - \theta_0\right) = -\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta_0)\right)\right]^{-1} \sqrt{T}\nabla \sum_{t=0}^{T} l_t(\theta_0)$

$$\Rightarrow \left[\text{var}\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right)\right]^{0.5}\sqrt{T}\left(\tilde{\theta}_{SEM} - \theta_0\right) = -\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]^{-1}\left[\text{var}\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right)\right]^{0.5}\sqrt{T}\nabla \sum_{t=0}^{T} l_t(\theta_0)$$

$$\Rightarrow \sqrt{T}\left(\tilde{\theta}_{SEM} - \theta_0\right) = -\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]^{-1}\left[\text{var}\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right)\right]^{0.5} \times \left[\left[\text{var}\left(\sqrt{T}\sum_{t=1}^{T}\nabla l_t(\theta)\right)\right]^{-0.5}\sqrt{T}\nabla \sum_{t=0}^{T} l_t(\theta_0)\right]$$

$$= -\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]^{-1} \sqrt{T}\nabla \sum_{t=0}^{T} l_t(\theta_0)$$

$$\Rightarrow \text{var}\left[\sqrt{T}\left(\tilde{\theta}_{SEM} - \theta_0\right)\right] = \text{var}\left[-\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]^{-1} \sqrt{T}\nabla \sum_{t=0}^{T} l_t(\theta_0)\right]$$

$$= \left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]^{-1} \text{var}\left[\sqrt{T}\nabla \sum_{t=0}^{T} l_t(\theta_0)\right]\left(\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]^{-1}\right)'$$

$$= \left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]^{-1} \text{var}\left[\sqrt{T}\nabla \sum_{t=0}^{T} l_t(\theta_0)\right]\left(\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta)\right)\right]'\right)^{-1}$$

$$= T\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]^{-1} \text{var}\left[\nabla \sum_{t=0}^{T} l_t(\theta_0) \right] \left(\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]' \right)^{-1}$$

$$\therefore T \,\text{var}\left(\tilde{\theta}_{SEM}\right) = T\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]^{-1} \text{var}\left[\nabla \sum_{t=0}^{T} l_t(\theta_0) \right] \left(\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]' \right)^{-1}$$

$$\Rightarrow \text{var}\left(\tilde{\theta}_{SEM}\right) = \left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]^{-1} \text{var}\left[\nabla \sum_{t=0}^{T} l_t(\theta_0) \right] \left(\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]' \right)^{-1}$$

$$\text{and } E\left[\sqrt{T}\left(\tilde{\theta}_{SEM} - \theta_0\right)\right] = E\left[-\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]^{-1} \sqrt{T} \nabla \sum_{t=0}^{T} l_t(\theta_0) \right]$$

$$\left[\sqrt{T}\left(E(\tilde{\theta}_{SEM}) - \theta_0\right)\right] = \left[-\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]^{-1} \sqrt{T} E\left[\nabla \sum_{t=0}^{T} l_t(\theta_0) \right] \right]$$

$$= \left[-\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]^{-1} \sqrt{T} \left[\nabla \sum_{t=0}^{T} E\left(l_t(\theta_0)\right) \right] \right]$$

$$= -\left[E\left(\nabla^2 \sum_{t=0}^{T} l_t(\theta) \right) \right]^{-1} \times \mathbf{0}, \text{since} \left[\nabla \sum_{t=0}^{T} E\left(l_t(\theta_0)\right) \right] = \mathbf{0}$$

$$\Rightarrow \left[\sqrt{T}\left(E(\tilde{\theta}_{SEM}) - \theta_0\right)\right] = \mathbf{0}$$

$$\Rightarrow \left(E(\tilde{\theta}_{SEM}) - \theta_0\right) = \mathbf{0}$$

$$\Rightarrow E(\tilde{\theta}_{SEM}) = \theta_0$$

$$\Rightarrow \tilde{\theta}_{SEM} \xrightarrow{d} N\left(\theta_0, \left[E\left(\nabla^2 \sum_{t=1}^{T} l_t(\theta) \right) \right]^{-1} \text{var}\left(\nabla \sum_{t=1}^{T} l_t(\theta) \right) \left(\left[E\left(\nabla^2 \sum_{t=1}^{T} l_t(\theta) \right) \right]' \right)^{-1} \right)$$

This proves that;

$$\tilde{\theta}_{SEM} \xrightarrow{d} N\left(\theta_0, \left[E\left(\nabla^2 \sum_{t=1}^{T} l_t(\theta) \right) \right]^{-1} \text{var}\left(\nabla \sum_{t=1}^{T} l_t(\theta) \right) \left(\left[E\left(\nabla^2 \sum_{t=1}^{T} l_t(\theta) \right) \right]' \right)^{-1} \right)$$

i.e $\tilde{\theta}_{SEM}$ is asymptotically normal

4.5 Chapter summary

In this chapter, we have shown that the semi-parametric GARCH (1, 1) estimator is consistent. This means that as the sample size increases indefinitely the estimator converges to the true population parameter. It should be noticed here that the estimator has been found to converge almost surely to the true population parameter. What this means is that, as we increase the sample size indefinitely the estimator hits the true

population parameter i.e. approaches the true population parameter with 100 percent probability. This is a very powerful result for it means that with a large sample at hand one is assured that the estimator they get is the same as the true population parameter.

We have also proved in this study that the semi-parametric GARCH (1, 1) estimator is asymptotically normal. This means that as one increases the sample size indefinitely the estimator converges, in distribution, to the normal distribution. This implies that with large sample size, the usual statistical inferences can be done on the estimator. It should also be mentioned here that the estimator has been found to have convergence rate that is the same as the parametric convergence rate, $\sqrt{T}$ .This is a very powerful result for it means the derived semi-parametric GARCH (1, 1) estimator converges at a highest speed possible. We have been able to derive an estimator that has a very high convergence speed using a flexible approach under serially dependent innovations

5. CONCLUSIONS AND THEORETICAL IMPLICATIONS

5.1 Summary

The main objective in this paper was to derive a semi-parametric GARCH (1, 1) estimator, under realistic assumptions that are in line with financial data empirical regularities. Specifically, the study aimed at proving that the derived semi-parametric GARCH (1, 1) estimator is not only consistent but also asymptotically normally distributed. To avoid running a risk of getting inconsistent estimators the study assumed that the innovations are serially dependent and have a probability distribution that belongs to the generalised error distribution. However, the assumption that the innovations are serially dependent brought three technical problems.

Firstly, we could not split the joint probability distribution into a product of marginal distributions as is normally done. The paper got around this problem, however, by splitting the joint distribution into a product of conditional probability densities. This was possible since the innovations were assumed to be martingale differences. Having done this, we applied the quasi-maximum likelihood estimation technique to derive the estimator.

Secondly, we could not use the cerebrated weak law of large numbers to prove consistency of the derived estimator. This is because the weak law of large numbers works on assumption that the innovations are independent, an assumption that this study relaxed. Luckily though, we were able to prove consistency by using the strong law of large numbers since it does not require the i.i.d assumption. Since strong convergence implies weak convergence, we were then able to show that the derived estimator converges in probability to the true parameter. It should be emphasized again here that the estimator has been found to converge almost surely to the true population parameter. What this means is that, as we explained above, as we increase the sample size indefinitely the estimator hits the true population parameter i.e. approaches the true population parameter with 100 percent probability. This is a very powerful result for it means that with a large sample at hand one is assured that the estimator they get is the same as the true population parameter.

Thirdly, we could not apply the cerebrated Linde-berg central limit theorem to prove that the derived semi-parametric GARCH (1, 1) estimator is asymptotically normally distributed. This was because the Linde-berg central limit theorem works on the assumption that the innovations are independent, an assumption that this study relaxed. However, we managed to show that the derived estimator is asymptotically normal by using the martingale central limit theorem which does not require the innovations to be

independent. Use of the martingale central limit theorem was possible since the innovations in this study were assumed to be martingale differences.

In a nutshell, the study has derived the semi-parametric GARCH (1, 1) estimator under serially dependent innovations using the quasi-maximum likelihood estimation technique. The derived estimator has then been shown to have nice asymptotic properties i.e. consistency and asymptotic normality.

5.2 Implications

It can be noticed here that the estimator we have derived here is better than the parametric estimator of GARCH (1, 1). This is because, despite both having the same (the highest) convergence rate of $\sqrt{T}$, the semi parametric GARCH (1, 1) estimator here has been derived based on realistic assumptions (i.e. non normal and serially dependent innovations) of the behaviour of time series financial data. Similarly it can be seen that the estimator derived in this study is better than the non-parametric GARCH (1, 1) estimators that have so far been proposed in literature. This is because the semi parametric estimator here has a higher convergence rate, $\sqrt{T}$, compared to the convergence rate of non-parametric GARCH (1,1) estimators proposed in literature, $T^{\frac{1}{5}}$ .In the same vein, the semi parametric estimator derived in this study is also better than the semi parametric GARCH (1,1) estimators already proposed in the literature. This is simply because, the estimator in this study has been derived under realistic assumptions of time series financial data than the existing semi parametric GARCH (1, 1) estimators that have been proposed so far, as we stated in the problem statement above. The study therefore offers an estimator that is not only realistic (i.e. based on assumptions that are in line with empirical financial data regularities) but also an estimator that has nice asymptotic properties (i.e. consistency(almost sure convergence) and asymptotic normality with the higher possible convergence rate, $\sqrt{T}$).The GARCH(1,1) estimator derived in this study therefore seems to be the theoretically best estimator in the class of existing estimators in the literature.

5.3 Suggestions for further studies

Just like any study, there is possibility of extending this study. Specifically, further studies could focus on testing the efficiency of the semi-parametric GARCH (1, 1) estimator that has been proposed in this study. This could be done by checking whether the variance of the semi-parametric GARCH (1, 1) estimator proposed in this study achieves its Cramer-Rao lower bound. Further research can also focus on empirical performance of the proposed estimator. This could be done by using Monte Carlo simulations and/or applying the estimator to real financial data and then comparing its performance to the parametric and non-parametric counterparts on the basis of mean square errors (MSE).Lastly; further research can focus on finding the implications of relaxing the assumptions made in this study. The results of this study are good as claimed in this study as long as the assumptions made in this study (e.g. the ergodicity assumption, the martingale difference innovations assumptions among others) hold. But a question still remains as to what happens if one/more of these assumptions are not holding. For instance, one would try to examine what would happen if the innovations

happen to have a probability law that is not in the generalized error distribution family or indeed what happens if the innovations are not ergodic.

REFERENCES

Amemiya, T. (1985). *Advanced Econometrics.* Cambridge: Havard University Press.

Andersen, T. G. (1996). GMM estimation of stochastic volatility models:A monte Carlo Study. *Journal of Business and Economic Statistics,48*, 328-352.

Avran, F. (1988). Weak Convergence of the variations,iterated integrals and Doleans-Dade exponentials of sequences of semi-martingales. *Anals of probability 16*, 246-250.

Baillie, R. T., & Bollerslev, T. (1987). The message in Daily Exchange Rates: A conditional Variance Tale. *Econometrica*.

Bollerslev. (1987). A conditional Heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics,69*, 542-547.

Bollerslev, T. (1986). Generalized Autoregressive Heteroskedasticity. *Journal of Econometrics*, 307-327.

Bollersslev, J., & Woodridge, J. (1992). Quasi maximum likelihood estimation and inference in dynamic models with time varying covariances. *Econometric Reviews,11(2)*, 143-172.

Buhlman, P., & McNeil, A. J. (2000). *Non-parametric GARCH Models.* Zurich, Switzerland: Seminar Fur Statistik:CH-8092.

Cameron, C. A., & Trivedi, P. (2005). *Microeconometrics:Methods and Applications.* New York,USA: Cambrdge University Press.

Choi, E. J. (2004). Estimation of Stochastic Volatility Models by Simulated Maximum Likelihood Method. *University of Waterloo*.

Chung, S. S. (2012). A class of non-parametric volatility models:Application to financial time series. *Journal of Econometrics*.

Dahl, C. M., & Levine, M. (2010). Non-parametric estimation of volatility models under serially dependent innovations. *Econometrica*.

Davidson, J. (2000). *Econometric Theory.* Blackwel: Oxford University Press.

Davidson, R., & Mackinnon, J. (1993). *Estimation and inference in Econometrics.* London: Oxford University Press.

Drost, F. C., & Klasssen, C. (1996). Efficient estimation in Semi-parametric GARCH Models. *Discussion paper;vol(1996-38),Tilburg*.

Duan, J. (1997). Augmented GARCH(p,q) process and its diffusion limit. *Journal of Econometrics,79(1)*, 97-127.

Engle, R. (1982). Autoregressive Conditional Heteroskedasticity with estimation of the variance of U.K inflation. *Econometrica*, 987-1008.

Engle, R. F., & Gonzale-Rivera, G. (October,1991). Semiparametric ARCH Models. *Journal of Business and Economic Statistics, 9(4)*, 345-359.

Engle, R. F., & Ng, V. (1993). Measuring and testing the impact of news on volatility. *Journal of finance,48*, 1747-1778.

Fan, & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting:variable bandwidth and spatial adaptation. *Journal of Royal Statistical Society,B,57*, 371-394.

Fan, J., & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika,85*, 645-660.

Gallant, A. R., & Hsieh, D. (1989). Fitting a Recalcitrant series:The Pound/Dollar Exchange Rate,1974-83. *Econometrica*.

Geweke, J. (1986). Modelling Persistence in Conditional Variances: A Comment. *Econometric Review,5*, 56-61.

Glosten, L. R., & Runkle, D. (1993). On the relation between the expected value and volatility of the nominal excess return of stocks. *Journal of Finance,48*, 1779-1801.

Gourieroux, C., & Trognon, A. (1984). Pseudo Maximum Likelihood Methods:Theory,52(1). *Econometrica*, 681-700.

Haafner, C. M. (2003). Analytical quasi maximum likelihood inference in BEKK-GARCH models. *Econometric Institution,Erasmus University,Rotterdam*.

Hafner, C. M., & Rombonts, J. (2002). Semiparametric multivariate GARCH models . *Discussion paper,2002/XX,CORE*.

Hansen, B. C. (2006). *Econometrics.* New York: Cambridge University Press.

Hansen, P., & Heyde, C. (1980). *Martingale limit theory and its applications.* New York: Academic Press.

Hansen, R. P., & Lunde, A. (2001). A comparison of volatility models:Does anything beat GARCH(1,1)? *Centre for analytica finance:University of AARHUS*.

Hentshel, L. (1995). All in the family: Nesting Symetric and Asymetric GARCH Models. *Journal of Financial Economics,39*, 71-104.

Herwartz, H. (2004). ConditionalHeteroskedasticity.In H. Lutkepoh, & M. Kratzig (Eds.) *Themes in Modern Econometrics*, pp. 197-220.

Higging, M. L., & Bera, A. (1992). A Class of non-linear ARCH Models. *International Economic Review*.

Holly. (2009). *Modelling Risk using fourth order Pseudo Maximum Likelihood Methods.* University of Lausanne, Institute of Healthy Economics.

Holly, A., & Montifort, A. (2010). Fourth Order Pseudo Maximum Likelihood Methods. *Econometrica*.

Holly, A., & Pentsak, Y. (2004). *Maximum Likelihhod Estimation of the Conditional mean E(Y|X) for Skewed Dependent variables in Fourth-Parameter families of Distributions.* Technical Report, University of Lausanne, Institute of Healthy Economics and Management.

Hood, W., & Koopman, T. (1953). The estimation of simultaneous linear economic relationships. *Econometric Methods*.

Ibragimov, R., & Philips, P. (2010). Regression asymptotics using martingale convergence. *Yale University press*.

Kouassi, E. (2015). Consistency of Pseudomaximum likelihood estimation in ARCH(1) under dependent innovations. *Working paper*.

Linton, O., & Mammen, E. (May, 2003). Estimating Semi-parametric ARCH(∞) nodels by kernel smoothing methods. *Discussion paper,No:EM/03/453*.

McCullagh, P. (1994). Exponential mixtures and quadratic exponential families. *Biometrika, 81*(4), 721-729.

Nadaraya, E. (1964). "On Estimating Regression".The Theory of Probability and its Applications. *Econometrica*, 141-2.

Nielsen, B. (1978). *Information and exponential families in statiatical theory.* New York: Wiley.

Pantula, S. G. (1986). Modelling Persistence in Conditional Variances: A comment. *Econometric Review,5*, 71-74.

Posedel, P. (2005). Properties and estimation of GARCH(1,1) model. *Metodoloski Zvezki, 2*, 243-257.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: John Willey & Sons.

Rossi, E. (2004, march). A note on GARCH models. *Working paper*.

Sentana, C. (1995). Quadratic ARCH Models. *Review of Economic Studies,62(4)*, 639-661.

Sherphard, N. (2008). Statistical aspects of ARCH and Stochastic Volatility. In D. R. Cox, D. Hinkly, & O. Barndorff (Eds) *Time series models in Econometrics, Finance and other fields.Monographs on statistics and Applied probability,65*, pp. 1-65.

Sousi, P. (2013). *Advanced Probability.* New York: Cambridge University Press.

Stout, W. F. (1974). *Almost Sure Convergence.* New York: Academic Press.

Su, L., Ullah, A., & Mashra, S. (2011). Non-parametric and semi-parametric volatility models:specification,estimation and testing. *Econometrica*.

Tapia, R. A., & Thompson, J. (n.d.). Nonparametric Probability Density Estimation.

Taylor, S. (1986). *Modelling Financial Time Series.* John Wiley & Sons.

Tsay, R. S. (2010). *Analysis of time series.* Willey.

Tschiernig, R. (2004). Non-parametric Econometrics. In H. Lutkepoh, & M. Kratzig (Eds) *Themes in Econometrics*, pp. 243-289.

Watson, G. (1964). Smooth Regression Analysis. *The Indian Journal of Statistics*, 359-372.

Weiss, A. A. (1986). Asymptotic Theory for ARCH Models: Estimation and Testing. *Econometric Theory,2*, 107-131.

Williams, D. (1991). *Probability with martingales.* New York: Cambridge University Press.

Yang, L., & Song, Q. (2012). Efficient Semi-parametric GARCH Modelling of Financial Volatility. *Statistica Sinica, 22*, 249-270.

Zakoian, M. J. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control,18*, 931-955.