Essays in Modelling and Estimating Value-at-Risk

Yang Yan

A thesis submitted to the Department of Statistics

of the London School of Economics for the degree of

Doctor of Philosophy, London

September 2014

Contents

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I confirm that Chapter 2 was jointly co-authored with Professor Oliver Linton, Section 3.2 and 3.3 was jointly co-authored with Professor Oliver Linton and Dajing Shang, and Section 4.3 was jointly co-authored with Dajing Shang.

Preface

Throughout my 4 year PhD study at LSE, I have received invaluabe encouragement and support from my supervisors, Prof. Oliver Linton and Prof. Qiwei Yao, for which I am very grateful. Oliver has been an inspiring mentor leading me to the world of semiparametric and nonparametric world, and a great personal friend helping me through many difficulties. Qiwei has patiently discussed various statistical questions with me and provided helpful ideas along the way. Thanks Prof. Xiaohong Chen, Prof. Piotr Fryzlewicz and Dr. Matteo Barigozzi for their comments and suggestions about my work.

I would like to thank the faculty and the fellow students in my department, expecially my officemates, who make my four years' stay a great experience. Thanks those who organize and attend the Joint Econometrics and Statistics Workshop in LSE and the Econometrics Workshop in the Economics Department, University of Cambridge. Sharing and discussing innovative ideas with them in those occasions are such pleasant moments.

Finally, sincere gratitude to my family and friends for their great support, especially my husband and my mum who love me unconditionally. And little Alexander Linton and Florence Linton who make my stay at Cambridge so enjoyable and fun.

Abstract

The thesis concerns semiparametric modelling and forecasting Value-at-Risk models, and the applications of these in financial data. Two general classes of semiparametric VaR models are proposed, the first method is introduced by defining some efficient estimators of the risk measures in a semiparametric GARCH model through moment constraints and a quantile estimator based on inverting an empirical likelihood weighted distribution. It is found that the new quantile estimator is uniformly more efficient than the simple empirical quantile and a quantile estimator based on normalized residuals. At the same time, the efficiency gain in error quantile estimation hinges on the efficiency of estimators of the variance parameters. We show that the same conclusion applies to the estimation of conditional Expected Shortfall. The second model proposes a new method to forecast one-period-ahead Value-at-Risk (VaR) in general ARCH(∞) models with possibly heavy-tailed errors. The proposed method is based on least square estimation for the log-transformed model. This method imposes weak moment conditions on the errors. The asymptotic distribution also accounts for the parameter uncertainty in volatility estimation. We test our models against some conventional VaR forecasting methods, and the results demonstrate that our models are among the best in forecasting VaR.

Chapter 1

Introduction

The attention placed on effective risk management in the financial industry has never been greater, especially after the recent financial crisis. The profit-driven industry is aware of the importance of measuring and managing risk properly. The Basel Committee (1996) also requires financial institutions to hold a certain amount of cash against market risk. Value-at-Risk (VaR), as one of the measures of market risk, becomes widely known when JP morgan introduces Riskmetrics (1996) and set an industry standard. As a forward looking estimate, VaR is defined as the maximum potential loss in value of a portfolio of financial instruments with a given confidence level over a certain horizon. It is an important risk measure as portfolio managers are concerned with large potential loss in asset returns.

From an econometric point of view, VaR is a quantile of the conditional distribution of portfolio returns over a certain holding period. VaR forecasts are mostly cast in GARCH type models because financial time series are characterized by conditional heteroskedasticity and heavy-tailed distributions. The method proceeds in two steps: the first is to estimate the conditional volatility and the second is to estimate devolatized residual quantile. This method employs the volatility estimator as a filter to transform the conditional correlated returns into i.i.d. errors, for which vast quantile estimators such as empirical quantile or extreme value theory based quantile can be readily applied. For example, Riskmetrics (1996) employs a GARCH model with normal errors; McNeil and Frey (2000) propose new VaR forecast methods by combining GARCH models with Extreme Value Theory (EVT); Engle (2001) illustrates VaR forecasts in GARCH models with empirical quantiles; Nyström and Skoglund (2004) use GMM-type volatility estimators for the GARCH based VaR forecasts. See Duffie and Pan (1997), Engle and Manganelli (2004) and Gourieroux and Jasiak (2002) for more detailed surveys for VaR forecasts.

Consistency and asymptotic normality have been established under various conditions, see Weiss (1986), Lee and Hansen (1994), Hall and Yao (2003), and Jensen and Rahbek (2006). For the semiparametric models, references can be found in Engle and Gonzalez-Rivera (1991) , Linton (1993) and Drost and Klaassen (1997).

In practice, full parametric methods are very popular, but the commonly used normal distribuion is a flaw, since most of the financial returns have heavy-tails. Fully nonparametric methods,such as Historical Simulation, are easy to implement, but do not provide precise VaR prediction. Semi-parametric method, on the other hand, have been found to perform relatively well. The approach contains a parametric GARCH estimation and a nonparametric standardized residual estimation. It is accurate and at the same time flexible, because there are a rich class of GARCH family models to choose and no specific distribution assumption is required. The approach has been proposed in Pritsker (1997), Hull and White (1998), McNeil and Frey (2000) and Kuestre, Mittnik and Paolella (2006).

The thesis contributes to the semi- and nonparametric work in VaR modelling. Two general classes of semiparametric models have been proposed. Moment constraints are often used to identify and estimate the mean and variance parameters and are however discarded when estimating error quantiles. In order to prevent this efficiency loss in quantile estimation,the first approach is introduced by defining some efficient estimators of the risk measures in a semiparametric GARCH model through moment constraints and a quantile estimator based on inverting an empirical likelihood weighted distribution. It is found that the new quantile estimator is uniformly more efficient than the simple empirical quantile and a quantile estimator based on normalized residuals. At the same time, the efficiency gain in error quantile estimation hinges on the efficiency of estimators of the variance parameters. We show that the same conclusion applies to the estimation of conditional Expected Shortfall.

The second method is a new method to forecast one-period-ahead Value-at-Risk (VaR) in general ARCH(∞) models with possibly heavy-tailed errors. The proposed method is based on least square estimation for the log-transformed model. This method imposes weak moment conditions on the errors. Consequently, it has better prediction performance than commonly used QMLE-based VaR methods in the presence of non-normal errors. In addition, we characterize the asymptotic distribution of the proposed VaR forecast, and this distribution accounts for the uncertainty in volatility estimation.

ARCH/GARCH process is the most popular way to estimate volatility and many surveys have been done regarding to this topic. Bera and Higgins (1993) have a paper introducing properties, estimation and testing of the ARCH process, Bauwens, Laurent and Rombouts (2006) talk about multivariate GARCH models, Bollerslev (2009) provides an encyclopedic reference and Terasvirta (2009) summarises univariate GARCH models. However, no one has focused on semi and nonparametric approach of estimating ARCH/GARCH process. Chapter 2 of the thesis fills the needs by surveying the semi- and nonparametric approaches of ARCH/GARCH estimation. Chapter 3 and 4 propose two different classes of semiparametric approaches of VaR prediction.

Chapter 2 has been published in Journal of Probability and Statistic, Volume 2011, Article ID 906212.

Chapter 2

Semi- and nonparametric (G)ARCH Process

2.1 Introduction

The key properties of financial time series appear to be: (a) Marginal distributions have heavy tails and thin centres (Leptokurtosis); (b) the scale or spread appears to change over time; (c) Return series appear to be almost uncorrelated over time but to be dependent through higher moments. See Mandelbrot (1963) and Fama (1965) for some early discussions. The traditional linear models like the autoregressive moving average class do not capture all these phenomena well. This is the motivation for using nonlinear models. This chapter is about the nonparametric approach.

2.2 The GARCH Model

Stochastic volatility models are of considerable current interest in empirical finance following the seminal work of Engle (1982). Perhaps still the most popular version is Bollerslev's (1986) GARCH(1,1) model in which the conditional variance σ_t^2 of a martingale difference sequence y_t is

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2, \tag{2.1}$$

where the ARCH(1) process corresponds to $\beta = 0$. This model has been extensively studied and generalized in various ways, see the review of Bollerslev, Engle, and Nelson (1994). Following Drost and Nijman (1993), we can give three interpretations to (2.1). The *strong* form GARCH(1,1) process arises when

$$\frac{y_t}{\sigma_t} = \varepsilon_t \tag{2.2}$$

is i.i.d. with mean zero and variance one, where σ_t^2 is defined in (2.1). The most common special case is where ε_t are also standard normal. The *semi-strong* form arises when for ε_t in (2.2)

$$E(\varepsilon_t \,|\mathcal{F}_{t-1}) = 0 \text{ and } E(\varepsilon_t^2 - 1 \,|\mathcal{F}_{t-1}) = 0, \tag{2.3}$$

where $\mathcal{F}_{t-1}$ is the sigma field generated by the entire past history of the y process. Finally, there is a *weak* form in which σ_t^2 is defined as a projection on a certain subspace, so that the actual conditional variance may not coincide with (2.1). The properties of the strong GARCH process are well understood, and under restrictions on the parameters $\theta = (\omega, \beta, \gamma)$ it can be shown to be strictly positive with probability one, to be weakly and/or strictly stationary, and to be geometrically mixing and ergodic. The weaknesses of the model are by now well documented, see Tsay (2007) for example.

2.3 The Univariate Model

There are several different ways in which nonparametric components have been introduced into stochastic volatility models. This work was designed to overcome some of the restrictiveness of the parametric assumptions in Gaussian strong GARCH models.

2.3.1 Error Density

Estimation of the strong GARCH process usually proceeds by specifying that the error density ε_t is standard normal and then maximizing the (conditional on initial values) Gaussian likelihood function. It has been shown that the resulting estimators are consistent and asymptotically normal under a variety of conditions. Quansi-Maximum Likelihood Estimation (QMLE) method proposed in Weiss (1986) and Bollerslev and Wooldridge (1988) shows that the estimators of the parameters obtained by maximizing a likelihood function constructed under the normality assumption can still be consistent even if the true density is not normal. In many cases, there is evidence that the standardized residuals from estimated GARCH models are not normally distributed, especially for high frequency financial time series. Engle and Gonzalez-Rivera (1991) initiated the study of semiparametric models in which ε_t is i.i.d. with some density f that may be non-normal, thus suppose that

$$\begin{aligned} y_t &= \varepsilon_t \sigma_t \\ \sigma_t^2 &= \omega + \beta \sigma_{t-1}^2 + \gamma y_{t-1}^2, \end{aligned}$$

where ε_t is i.i.d. with density f of unknown functional form. There is evidence that the density of the standardized residuals $\varepsilon_t = y_t/\sigma_t$ is non-Gaussian. One can obtain more efficient estimates of the parameters of interest by estimating f nonparametrically. Linton (1993) and Drost and Klaassen (1997) developed kernel based estimates and establish the semiparametric efficiency bounds for estimation of the parameters. In some cases, e.g., if f is symmetric about zero, it is possible to adaptively estimate some parameters, i.e., one can achieve the same asymptotic efficiency as if one knew the error density. In other cases, or for some parameters, it is not possible to adapt, i.e., it is not possible to estimate as efficiently as if f were known. These semiparametric models can readily be applied to deliver value at risk and conditional value at risk measures based on the estimated density.

2.3.2 Functional form of Volatility Function

Another line of work has been to question the specific functional form of the volatility function, since estimation is not robust with respect to its specification. The news impact curve is the relationship between σ_t^2 and $y_{t-1} = y$ holding past values σ_{t-1}^2 constant at some level σ^2. This is an important relationship that describes how new information affects volatility. For the GARCH process, the news impact curve is

$$m(y, \sigma^2) = \omega + \gamma y^2 + \beta \sigma^2. \tag{2.4}$$

It is separable in σ^2, i.e., $\partial m(y, \sigma^2)/\partial \sigma^2$ does not depend on y, it is an even function of news y, i.e., $m(y, \sigma^2) = m(-y, \sigma^2)$, and it is a quadratic function of y with minimum at zero. The evenness property implies that $\text{cov}(y_t^2, y_{t-j}) = 0$ for ε_t with distribution symmetric about zero.

Because of limited liability, we might expect that negative and positive shocks have different effects on the volatility of stock returns, for example. The evenness of the GARCH process news impact curve rules out such 'leverage effects'. Nelson (1991) introduced the Exponential GARCH model to address this issue. Let $h_t = \log \sigma_t^2$ and let $h_t = \omega + \gamma \left[\theta \varepsilon_{t-1} + \delta \left|\varepsilon_{t-1}\right|\right] + \beta h_{t-1}$, where $\varepsilon_t = y_t/\sigma_t$ is i.i.d. with mean zero and variance one. This allows asymmetric effect of past shocks ε_{t-j} on current volatility, i.e., the news impact curve is allowed to be asymmetric. For example, $\text{cov}(y_t^2, y_{t-j}) \neq 0$ even when ε_t is symmetric about zero. An alternative approach to allowing asymmetric news impact curve is the Glosten, Jeganathan and Runkle (1994) model $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 + \delta y_{t-1}^2 1(y_{t-1} < 0)$.

There are many different parametric approaches to modelling the news impact curve and they can give quite different answers in the range of perhaps most interest to practitioners. This motivates a nonparametric approach, because of the greater flexibility in functional form thereby allowed. The nonparametric ARCH literature apparently begins with Pagan and Schwert (1990) and Pagan and Hong (1991). They consider the case where $\sigma_t^2 = \sigma^2(y_{t-1})$, where $\sigma(\cdot)$ is a smooth but unknown function, and the multilag version $\sigma_t^2 = \sigma^2(y_{t-1}, y_{t-2}, \ldots, y_{t-d})$. This allows for a general shape to the news impact curve and nests all the parametric ARCH processes. Under some general conditions on $\sigma(\cdot)$ (for example that $\sigma(\cdot)$ does not grow at a more than quadratic rate in the tails) the process y is geometrically strong mixing. Härdle and Tsybakov (1997) applied local linear fit to estimate the volatility function together with the mean function and derived their joint asymptotic properties. The multivariate extension is given in Härdle, Tsybakov and Yang (1996). Masry and Tjøstheim (1995) also estimate nonparametric ARCH models using the Nadaraya-Watson kernel estimator. Lu and Linton (2006) extended the CLT to processes that are only near epoch dependent. Fan and Yao (1998) have discussed efficiency issues in this model, see also Avramidis (2002). Franke, Neumann, and Stockis (2004) have considered the application of bootstrap for improved inference. In practice, it is necessary to include many lagged variables in $\sigma^2(.)$ to match the dependence found in financial data. The problem with this is that nonparametric estimation of a multi-dimension regression surface suffers from the well-known "curse of dimensionality": the optimal rate of convergence decreases with dimensionality d, see Stone (1980). In addition, it is hard to describe, interpret and understand the estimated regression surface when the dimension is more than two. Furthermore, even for large d this model greatly restricts the dynamics for the variance process since it effectively corresponds to an ARCH(d) model, which is known in the parametric case not to capture the dynamics well. In particular, if the conditional variance is highly persistent, the non-parametric estimator of the conditional variance will provide a poor approximation, as reported in Perron (1998). So not only does this model not capture adequately the time series properties of many datasets, but the statistical properties of the estimators can be poor, and the resulting estimators hard to interpret.

Additive models offer a flexible but parsimonious alternative to nonparametric models, and have been used in many contexts, see Hastie and Tibshirani (1990). Suppose that

$$\sigma_t^2 = c_v + \sum_{j=1}^{d} \sigma_j^2(y_{t-j}) \tag{2.5}$$

for some unknown functions σ_j^2. The functions σ_j^2 are allowed to be of general functional form but only depend on y_{t-j}. This class of processes nests many parametric ARCH models. Again, under growth conditions the process y can be shown to be stationary and geometrically mixing. The functions σ_j^2 can be estimated by special kernel regression techniques, such as the method of marginal integration, see Linton and Nielsen (1995) and Tjøstheim and Auestad (1994). The best achievable rate of convergence for estimates of $\sigma_j^2(.)$ is that of one-dimensional nonparametric regression, see Stone (1985). Masry and Tjøstheim (1995) developed estimators for a class of time series models including (2.5). Yang, Härdle, and Nielsen (1999) proposed an alternative nonlinear ARCH model in which the conditional mean is again additive, but the volatility is multiplicative $\sigma_t^2 = c_v \prod_{j=1}^{d} \sigma_j^2(y_{t-j})$. Kim and Linton (2004) generalized this model to allow for arbitrary [but known] transformations, i.e., $G(\sigma_t^2) = c_v + \sum_{j=1}^{d} \sigma_j^2(y_{t-j})$, where $G(.)$ is a known function like log or level. The typical empirical findings are that the news impact curves have an inverted asymmetric U-shape.

These models address the curse of dimensionality but they are rather restrictive with respect to the amount of information allowed to affect volatility, and in particular do not nest the GARCH(1,1) process. Linton and Mammen (2005) proposed the following model

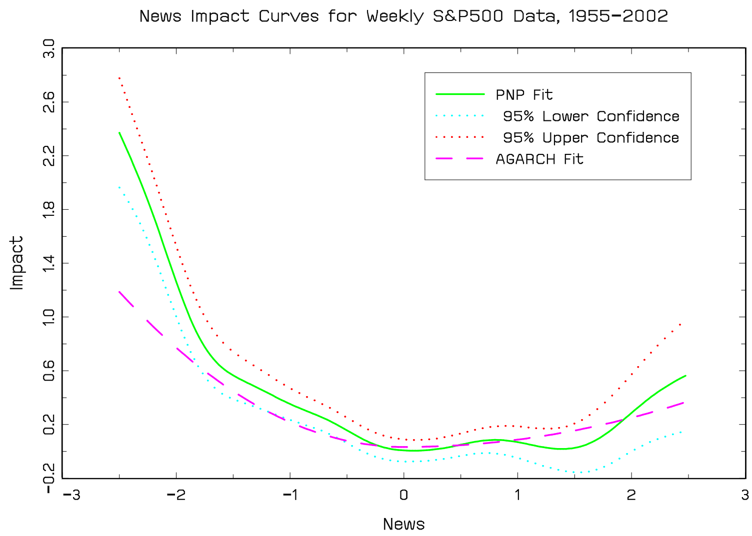
$$\sigma_t^2(\theta, m) = \sum_{j=1}^{\infty} \psi_j(\theta) m(y_{t-j}), \tag{2.6}$$

where $\theta \in \Theta \subset \mathbb{R}^p$ and m is an unknown but smooth function. The coefficients $\psi_j(\theta)$ satisfy at least $\psi_j(\theta) \geq 0$ and $\sum_{j=1}^{\infty} \psi_j(\theta) < \infty$ for all $\theta \in \Theta$. A special case of this model is the Engle and Ng (1993) PNP model where

$$\sigma_t^2 = \beta \sigma_{t-1}^2 + m(y_{t-j}),$$

where $m(.)$ is a smooth but unknown function. This model nests the simple GARCH(1,1) model but permits more general functional form: it allows for an asymmetric leverage effect, and as much dynamics as GARCH(1,1). Estimation methods for these models are based on iterative smoothing. Linton and Mammen (2005) showed that the news impact curves for daily and weekly S&P500 data are quite asymmetric with non-quadratic tails and is not minimal at zero but at some positive return. Below we show their estimator, denoted PNP here, in comparison with a common parametric fit, denoted AGARCH.

Figure 1: News impact curve (PNP v.s. AGARCH)



Yang (2006) introduced a semiparametric index model

$$\sigma_t^2 = g\left(\sum_{j=1}^{\infty} \nu_j(y_{t-j};\theta)\right),$$

where $\nu_j(y;\theta)$ are known functions for each j satisfying some decay condition and g is smooth but unknown. This process nests the GARCH(1,1) when g is the identity, but also the quadratic model considered in Robinson (1991).

Audrino and Bühlmann (2001) proposed their model as $\sigma_t^2 = \Lambda(y_{t-1}, \sigma_{t-1}^2)$ for some smooth but unknown function $\Lambda(.)$, and includes the PNP model as a special case. They proposed an estimation algorithm. However, they did not establish the distribution theory of their estimator, and this may be very difficult to establish due to the generality of the model.

2.3.3 Relationship between Mean and Variance

The above discussion has centered on the evolution of volatility itself, whereas one is often very interested in the mean as well. One might expect that risk and return should be related, Merton (1973). The GARCH-in-Mean process captures this idea, it is

$$y_t = g(\sigma_t^2; b) + \varepsilon_t\sigma_t,$$

for various functional forms of g e.g., linear and log-linear and for some given specification of σ_t^2. Engle, Lilien and Robbins (1987) introduced this model and applied it to the study of the term Structure. Here, b are parameters to be estimated along with the parameters of the error variance. Some authors find small but significant effects. Again, the nonparametric approach is well motivated here on grounds of flexibility. Pagan and Hong (1991) and Pagan and Ullah (1988) considered a case where the conditional variance is nonparametric (with a finite number of lags) but enters in the mean equation linearly or log linearly. Linton and Perron (2002) studied the case where g is nonparametric but σ_t^2 is parametric, for example GARCH. The estimation algorithm was applied to stock index return data. Their estimated g function was non-monotonic for daily S&P500 returns.

2.3.4 Long Memory

Another line of work has argued that conventional models involve a dependence structure that does not fit the data well enough. The GARCH(1,1) process $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2$ is of the form

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j y_{t-j}^2 \tag{2.7}$$

for constants c_j satisfying $c_j = \gamma\beta^{j-1}$, provided the process is weakly stationary, which requires $\gamma + \beta < 1$. These coefficients decay very rapidly so the actual amount of memory is quite limited. There is some empirical evidence on the autocorrelation function of y_t^2 for high frequency returns data that suggests a slower decay rate than would be implied by these coefficients, see Bollerslev and Mikkelson (1996). Long memory models essentially are of the form (2.7) but with slower decay rates. For example, suppose that $c_j = j^{-\theta}$ for some $\theta > 0$. The coefficients satisfy $\sum_{j=1}^{\infty} c_j^2 < \infty$ provided $\theta > 1/2$. Fractional integration (FIGARCH) leads to such an expansion. There is a single parameter called d that determines the memory properties of the series, and

$$(1-L)^d \sigma_t^2 = \omega + \gamma\sigma_{t-1}^2(\varepsilon_{t-1}^2 - 1),$$

where $(1-L)^d$ denotes the fractional differencing operator. When $d = 1$ we have the standard IGARCH model. For $d \neq 1$ we can define the binomial expansion of $(1-L)^{-d}$ in the form given above. See Robinson (1991) and Bollerslev and Mikkelson (1996) for models. The evidence for long memory is often based on sample autocovariances of y_t^2, and this may be questionable when only few moments of y_t exist, see Mikosch and Stărică (2002). See Giraitis (2007) for a nice review.

2.3.5 Locally Stationary Processes

Recently, another criticism of GARCH processes has come to the fore, namely their usual assumption of stationarity. The IGARCH process (where $\beta + \gamma = 1$) is one type of nonstationary GARCH model but it has certain undesirable features like the non-existence of the variance. An alternative approach is to model the coefficients of a GARCH process as changing over time, thus

$$\sigma_t^2 = \omega(x_{tT}) + \beta(x_{tT})\sigma_{t-1}^2 + \gamma(x_{tT})(y_{t-1} - \mu_{t-1})^2,$$

where ω, β, and γ are smooth but otherwise unknown functions of a variable x_{tT}. When $x_{tT} = t/T$, this class of processes is nonstationary but can be viewed as locally stationary along the lines of Dahlhaus (1997), provided the memory is weak, i.e., $\beta(\cdot) + \gamma(\cdot) < 1$. In this way the unconditional variance exists, i.e., $E[\sigma_t^2] < \infty$, but can change slowly over time as can the memory. Dahlhaus and Subba Rao (2006) have recently provided a comprehensive theory of such processes and about inference methods for the ARCH special case. See Spokoiny (2007) for a further review.

Engle and Rangel (2008) propose a special case of this model where the unconditional variance $\sigma^2(t/T) = \omega(t/T)/(1-\beta(t/T)-\gamma(t/T))$ varies over time but the coefficients $\beta(t/T)$ and $\gamma(t/T)$ are assumed to be constant. In this way, we can write $y_t = \sigma(t/T)g_t^{1/2}\varepsilon_t$, where g_t is a unit GARCH(1,1) process representing "high frequency" volatility, while $\sigma^2(t/T)$ is the low-frequency unconditional volatility modelled nonparametrically. Engle and Rangel (2008) also allow for covariates in the low frequency component of volatility.

2.3.6 Continuous Time

Recently there has been much work on nonparametric estimation of continuous time processes, see for example Bosq (1998). Given a complete record of transaction or quote prices, it is natural to model prices in continuous time (e.g., Engle (2000)). This matches with the vast continuous time financial economic arbitrage-free theory based on a frictionless market. Under the standard assumptions that the return process does not allow for arbitrage and has a finite instantaneous mean, the asset price process, as well as smooth transformations thereof, belong to the class of special semi-martingales, as detailed by Back (1991). Under some conditions, the semiparametric GARCH processes we reviewed can approximate such continuous time processes as the sampling interval increases. Work on continuous time is reviewed elsewhere in this volume, so here we just point out that this methodology can be viewed as nonparametric and as a competitor of the discrete time models we outlined above.

2.4 The Multivariate Case

It is important to extend the volatility models to the multivariate framework, as understanding the comovements of different financial returns is also of great interest. The specification of an MGARCH model should be flexible enough to represent the dynamics structure of the conditional variances and covariance matrix and parsimonious enough to deal with the rapid expansion of the parameters when the dimension increases. Semiparametric and nonparametric methods offer an alternative way to the parametric estimation by taking the advantage of not imposing a particular structure on the data. In general we have a vector time series $y_t \in \mathbb{R}^n$, that satisfies

$$y_t = \Sigma_t^{1/2} \varepsilon_t, \tag{2.8}$$

where ε_t is a vector of martingale difference sequences satisfying $E[\varepsilon_t|\mathcal{F}_{t-1}] = 0$ and $E[\varepsilon_t \varepsilon_t^\top - I_n|\mathcal{F}_{t-1}] = 0$, while Σ_t is a symmetric positive definite matrix. In this case, Σ_t is the conditional covariance matrix of y_t given its own history. The usual approach here is to specify a parametric model for Σ_t and perhaps also the marginal density of ε_t. There are many parametric models for Σ_t, and we just mention two recent developments that are particularly useful for large dimensional systems. First, the so-called CCC (constant conditional correlation) (Bollerslev (1990)) models where

$$\Sigma_t = D_t R D_t,$$

where D_t is a diagonal matrix with elements σ_{it}, where σ_{it}^2 follows a univariate parametric GARCH or other specification, while R is an n by n correlation matrix. The second model generalizes this to allow R to vary with time albeit in a restricted parametric way, and is thereby called DCC (dynamic conditional correlation)(Engle (2002)).

2.4.1 Error Density

Hafner and Rombouts (2007) consider a number of semiparametric models where the functional form of the conditional covariance matrix is parametrically specified while the innovation distribution is unspecified i.e., ε_t is i.i.d with density function $f : \mathbb{R}^n \to \mathbb{R}$, where f is of unknown functional form. In the most general case, they treat the multivariate extension of the semiparametric model of Engle and Gonzalez-Rivera (1991). They show that it is not generally possible to adapt, although one can achieve a semiparametric efficiency bound for the identified parameters. The semiparametric estimators are more efficient than the QMLE if the innovation distribution is non-normal. These methods can often deliver efficiency gains but may not be robust to say dependent or time varying ε_t. In practice, the estimated density is quite heavy tailed but close to symmetric for stock returns.

It is also worth mentioning the SNP (SemiNonParametric) method, which was first introduced by Gallant and Tauchen (1989). The fundamental part of the estimating procedure of the conditional density of a stationary multivariate time series relies on the Hermite series expansion, associating with a model selection strategy to determine the appropriate degree of the expansion. The estimator is consistent under some reasonable regularity conditions.

One major issue with the unrestricted semiparametric model is the curse of dimensionality: as n increases the best possible rate at which the error density can be estimated gets worse and worse. In practice, allowing for four or more variables in an unrestricted way is impractical with even enormous sample sizes. This motivates restricted versions of the general model that embody a compromise between flexibility of functional form and reasonable small sample properties of estimation methods.

The first class of models is the family of spherically symmetric densities in which

$$f(x) = g(x^\top x),$$

where $g : \mathbb{R} \to \mathbb{R}$ is an unknown but scalar function. This construction avoids the "curse of dimensionality" problem, and can in principle be applied to very high dimensional systems. This class of distributions is important in finance, since the CAPM is consistent with returns being jointly elliptically symmetric (i.e., spherically symmetric after location and scale transformation), Ingersoll (1984). Hafner and Rombouts (2007) develop estimation methods for parametrically specified Σ_t under this assumption.

Another approach is based on copula functions. By Sklar's theorem, any multivariate distribution can be modelled by the marginal distribution of each individual series and the dependence structure between individual series which is captured by copula functions. A copula itself is a multivariate distribution function with uniform marginals. The joint distribution function of random variables X and Y defined as $F_{X,Y}(x, y) = C(F(x), G(y))$. A bivariate distribution function whose marginals are $F(\cdot)$ and $G(\cdot)$, and $C(\cdot) : [0.1]^2 \to \mathbb{R}$ is the copula function measures the dependency.

Chen and Fan (2006a) proposed a new class of semiparametric copula-based multivariate dynamic models, the so-called SCOMDY models, in which case the conditional mean and the conditional variance of a multivariate time series are specified parametrically, while the multivariate distribution of the standardized innovation are specified semiparametrically as a parametric copula evaluated at nonparametric marginals. The advantage of this method is a very flexible innovation distribution by estimating the univariate marginal distributions nonparametrically and fitting a parametric copula and its circumvention of the "curse of dimensionality". An important class of the SCOMDY models is the semiparametric copula-based multivariate GARCH models, which has the following set up:

$$\begin{aligned} y_{i,t} &= \sigma_{i,t}\varepsilon_{i,t} \\ \sigma^2_{i,t} &= \omega_i + \sum_{j=1}^{p_i}\gamma_{i,j}y^2_{i,t-j} + \sum_{j=1}^{q_i}\beta_{i,j}\sigma^2_{i,t-j}, \end{aligned}$$

where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})^\top$ is a sequence of i.i.d. random vectors with zero mean and unit variance. In this case, the conditional covariance matrix of returns is in the class of the CCC models. The key feature of the SCOMDY is the semiparametric form taken by the joint distribution function $\mathbf{F}_\varepsilon$ of ε_t:

$$\mathbf{F}_\varepsilon(\varepsilon_1, \ldots, \varepsilon_n) = C(F_{\varepsilon,1}(\varepsilon_1), \ldots, F_{\varepsilon,n}(\varepsilon_n); \theta_0), \tag{2.9}$$

where $C(\cdot)$ is a parametrized copula function depended on unknown $\theta \in \Theta \subset \mathbf{R}^m$, and for $i = 1, \ldots, n$, $F_{\varepsilon,i}(\cdot)$ is the marginal distribution function of the innovation which is assumed to be continuous but otherwise unspecified. Many examples of combinations have been introduced in the paper, such as { GARCH(1,1),Normal copula} and { GARCH(1,1), Student's-t copula}. They also construct simple estimators of the parameters. They establish the large sample properties of the estimator under a misspecified parametric copula, showing that both of the estimators of unknown dynamic parameters and the marginal distribution are still consistent while the estimator of the copula dependence parameter will converge in this case. Chen and Fan (2006b) modelled a univariate version of this class of semiparametric models, but their two-step estimators are verified to be inefficient and even biased if the time series has strong tail dependence in the simulation study of Chen, Wu and Yi (2009). The new paper considers the efficient estimation by using a sieve MLE method which is first introduced by Chen, Fan and Tsyrennikov (2006).

Embrechts, McNeil and Strumann (2002) was the most influential paper of the early study of copulas in finance and since then, numerous copula-based models are being introduced and used in financial applications. The copula-GARCH models of Patton (2006a, 2006b) proposed to make the parameter of the copula time varying in a dynamic fashion. Jondeau and Rockinger (2006) modelled daily return series with univariate time-varying skewed Student-t distribution and a Gaussian or

Student-t copula for the dependence. Panchenko (2006) also considered a semiparametric copula-based model applied to risk management. Rodriguez (2007) and Okimoto (2007) proposed the regime-switching copula models for pairs of international stock indices. A recent paper by Chollette, Heinei and Valdesogo (2008) estimated the multivariate regime switching model of copula as an extension of the Pelletier (2006) model to non-Gaussian case.

2.4.2 Conditional Covariance Matrix

Hafner, van Dijk and Franses (2005) proposed a semiparametric approach for the conditional covariance matrix which allows the conditional variance to be modelled parametrically by using any choice of univariate GARCH-type models, while the conditional correlation are estimated by nonparametric methods. The conditional covariance matrix Σ_t is defined as follows:

$$\Sigma_t = D_t R_t D_t \tag{2.10}$$

where D_t is parametrically modelled by any choice of univariate GARCH specification, and R_t is treated nonparametrically as an unknown function of a state variable x_t, thus $R_t = R(x_t)$ for some unknown matrix function $R(.)$. The function $R(.)$ is estimated using kernel methods based ont he rescaled residuals from the initial univariate parametric fits of the GARCH models.

Recently, Hafner and Linton (2009) introduced a multivariate multiplicative volatility model which can be regarded as the multivariate version of the spline-GARCH model of Engle and Rangel (2008). A vector time series y_t takes the form:

$$y_t = H(t/T)^{1/2} G_t^{1/2} \varepsilon_t \tag{2.11}$$

where ε_t is (at least) a strictly stationary unit conditional variance martingale difference sequence. The model allows the slowly varying unconditional variance matrix $H(\cdot)$ to be unknown along with the short run dynamics captured through $G(\cdot)$, which is itself a unit variance multivariate GARCH process, for example the BEKK model

$$G_t = I - AA^\top - BB^\top + AG_{t-1}A^\top + Bu_{t-1}u_{t-1}^\top B^\top,$$

where A, B are parameter matrices and $u_t = G_t^{1/2}\varepsilon_t$.

Feng (2007) proposes an alternative specification call the local dynamic conditional correlation (LDCC) model, where the total covariance matrix is decomposed into a conditional and an unconditional components. The total covariance matrix takes the form:

$$\Sigma_t = D_t^L D_t^C R_t D_t^C D_t^L,$$

where $D_t^L = \text{diag}(\sigma_{it}^L), D_t^C = \text{diag}(\sigma_{it}^C)$ and $R_t = \rho_{ijt}, (i, j = 1, \ldots, n,)$and $(\sigma_{it}^L)^2$ are the local variances, $(\sigma_{it}^C)^2$ are the conditional variances and ρ_{ijt} denote the dynamic correlations. Specifically, $\sigma_{it}^L = \sigma_i^L(t/T)$, while σ_{it}^{2C} follows a parametric unit GARCH type process. As in parametric DCC models one first proceeds by estimating the univariate models and then using standardized residuals to estimate the model for R_t.

2.5 Conclusion

In conclusion, there have been many advances in the application of nonparametric methods to the study of volatility, and many difficult problems have been overcome. These methods have offered new insights into functional form, dependence, tail thickness, and nonstationarity that are fundamental to the behaviour of asset returns. They can be used by themselves to estimate quantities of interest like value at risk. They can also be used as a specification device enabling the practitioner to see with respect to which features of the data their parametric model is a good fit.

Chapter 3

Efficient Estimation of Conditional Risk Measures in a Semiparametric GARCH Model

3.1 Introduction

Many popular time series models specify some parametric or nonparametric structure for the conditional mean and variance. Often, these models are completed by a sequence of i.i.d errors ε_t.[1] For example, many models can be written in the form of $\mathcal{T}(y_t, y_{t-1}, \ldots; \theta) = \varepsilon_t$, where the parametric model $\mathcal{T}(\cdot; \theta)$ is used to remove the temporal dependence structure in y_t so that the error ε_t is i.i.d with certain distribution $F(\cdot)$. Parameters θ and $F(\cdot)$ together define the model. Often one assumes moment conditions on ε_t such as it being mean zero and variance one. These moment constraints are often used to identify and estimate the mean and variance parameters θ but are however often discarded when estimating the error distribution or quantile. Knowledge of the conditional distribution is very important in finance since all financial instruments are more or less pricing or hedging certain sections of the distribution of underlying assets. For example, mean-variance trade-off in portfolio management is concerned with the first and second moments; exotic derivatives are traded for transferring downside risks, which are lower portions of the asset's distribution. Other practical usage of conditional distribution estimation includes the risk-neutral density estimation and Value-at-Risk (VaR) estimation.

[1] There are some notable exceptions to this including Engle and Manganelli (2004).

In this chapter, we consider how best to utilize this conditional information to estimate the distribution $F(\cdot)$, and further the quantiles of ε_t, so that one can construct an efficient estimator for the conditional distribution and hence quantiles of y_{t+1} given $\mathcal{F}_t = \{y_t, y_{t-1}, \ldots, y_0\}$. Besides proposing a VaR estimator, we also introduce Expected Shortfall (ES)

Recently, it has been argued that Value at Risk is not a coherent measure of risk, specifically it can violate the subadditivity axiom of Artzner et al. (1999). Instead the expected shortfall (ES) is an alternative risk measure that does satisfy all of their axioms. ES is defined as the expected return on the portfolio in the worst $100\alpha\%$ of the cases. ES incorporates more information than VaR because ES gives the average loss in the tail below $100\alpha\%$. The estimation of unconditional ES has been considered in Scaillet (2004) and Chen (2008). The recent Basel Committee on Banking Supervision round III has suggested using expected shortfall in place of value at risk, so this measure is likely to gain in prominence in the future.

We consider the following popular AR(p)-GARCH(1,1) model

$$\begin{aligned} y_t &= \sum_{j=1}^{p} \rho_j y_{t-j} + h_t^{1/2}\varepsilon_t \\ h_t &= \omega + \beta h_{t-1} + \gamma u_{t-1}^2, \end{aligned} \tag{3.1}$$

where $u_t = h_t^{1/2}\varepsilon_t$, and $\{\varepsilon_t\}$ is an i.i.d sequence of innovations with mean zero and variance one and p is a finite and known integer. We suppose that ε_t has a density function $f(\cdot)$, which is unknown apart from the two moment conditions:

$$\int x f(x)dx = 0; \int x^2 f(x)dx = 1. \tag{3.2}$$

These moment conditions are standard in parametric settings and identify h_t as the conditional variance of y_t given $\mathcal{F}_{t-1}$. Furthermore, the error density and all the parameters are jointly identified in the semiparametric model. In this case, the conditional Value-at-Risk of y_t given $\mathcal{F}_{t-1}$ and the conditional expected shortfall of y_t given $\mathcal{F}_{t-1}$ are respectively,

$$\xi_t(\alpha) = \sum_{j=1}^{p} \rho_j y_{t-j} + h_t^{1/2} q_\alpha$$

$$\begin{aligned} \chi_t(\alpha) &= E[y_t | y_t \leq \xi_t(\alpha), \mathcal{F}_{t-1}] \\ &= \sum_{j=1}^{p} \rho_j y_{t-j} + h_t^{1/2} E[\varepsilon_t | \varepsilon_t \leq q_\alpha] \end{aligned}$$

$$= \sum_{j=1}^{p} \rho_j y_{t-j} + h_t^{1/2} ES_\alpha,$$

where q_α is the α-quantile of ε_t, while $ES_\alpha = E[\varepsilon_t|\varepsilon_t \leq q_\alpha]$ is the α-expected shortfall of ε_t. In the sequel we assume that $p = 0$ for simplicity of notation. This is quite a common simplification in the literature; the main thrust of our results carry over to the more general p case.

Let $\theta = (\omega, \beta, \gamma)$. The goal of this paper is to estimate the parameters $(\theta, q_\alpha, ES_\alpha)$ efficiently and plug in these efficient estimators to obtain the conditional quantile $\widehat{\xi}_{n,t} = h_t^{1/2}(\widehat{\theta})\widehat{q}_\alpha$ and the conditional expected shortfall $\widehat{\chi}_t(\alpha) = h_t^{1/2}(\widehat{\theta})\widehat{ES}_\alpha$.

Since this model involves both finite dimensional parameters θ and infinite dimensional parameter $f(\cdot)$, we call it a semiparametric model. This chapter constructs an efficient estimator for both θ and the α'th quantile of $f(\cdot)$, q_α, for model (3.1) under moment constraints (3.2). Consequently, the conditional quantile estimator and conditional expected shortfall estimator are efficient.

Estimation of GARCH parameters has a long history. However, there are only limited papers discussing the efficiency issues involved in estimating semiparametric GARCH models. The first attempt is due to Engel and Gonzalez-Rivera (1991), who showed partial success in achieving efficiency via Monte Carlo simulations. In their theoretical work, Linton (1993) and Drost and Klaassen (1997) explained that full adaptive estimation of θ is not possible and showed their efficient estimators for β via a reparamerization. Ling and McAleer (2003) further considers adaptive estimation in nonstationary ARMA-GARCH models.

We complement previous work on GARCH models by providing an efficient estimator for $F(\cdot)$ and thus the quantile of ε_t. It is well known that, in the absence of any auxilliary information about $F(\cdot)$, the empirical distribution function $\widehat{F}(x) = n^{-1}\sum_{t=1}^{n} 1(\varepsilon_t \leq x)$ is semiparametrically efficient. However, $\widehat{F}(x)$ is no longer efficient when moment constraints (3.2) are available, see Bickel et al. (1993). The empirical likelihood (EL) weighted empirical distribution estimator is efficient with the existence of auxiliary information in the form of moments restrictions (3.2). The EL method was initiated by Owen (1990) and extended by Kitamura (1997) to time series. In i.i.d settings, Chen (1996) discovered second order improvement by empirical likelihood weighted kernel density estimation under moment restrictions. Zhao (1996) showed that there are variance gains by empirical likelihood weighted M-estimation when moment restrictions are available. Schick and Wefelmeyer (2002) provide an efficient estimator for the error distribution in nonlinear autoregressive models. However, the proposed estimator has the shortcoming that it is not a distribution itself. Müller et al. (2005) showed that the EL-weighted empirical distribution estimator is efficient in an autoregressive model. In this paper, we use EL weighted distribution estimator to construct estimates of VaR and ES in GARCH models. We show that, the resulting quantile and ES estimators for ε are efficient. Furthermore, the conditional VaR $\xi_t(\alpha)$ and ES estimators $\chi_t(\alpha)$ are asymptotically mixed-normal.

Various quantile estimators have been proposed recently, see Koenker, and Xiao (2009) and Chen, Koenker, and Xiao (2009). For fully nonparametric estimators, see Chen and Tang (2005) and Cai and Wang (2008). However, nonparametric estimators are subject to the curse of dimensionality and thus not widely applicable in practice. Furthermore, these nonparametric quantile estimators are too flexible to capture the stylized fact that financial returns are conditionally heteroskedastic. Given that this time-varying volatility is the key feature of financial time series, historical simulation method would be more advantageous than nonparametric methods in VaR forecasting. In our semiparametric model, the quantile estimator preserves the property of time-varying volatility and allows other aspect of conditional distribution unspecified. Model information is fully explored in the estimation so we gain by providing an efficient solution to conditional quantile estimation. Furthermore, the parametric filter (the GARCH model for volatility) bundle the conditioning set into a one-dimensional volatility so that there is no curse of dimensionality.

To the best of our knowledge, the only paper to address efficient conditional quantile estimation is Komunjer and Vuong (2010). However, their model is different from ours: they consider efficient conditional quantile estimation without moment constraints (3.2). Ai and Chen (2003) provide a very general framework for estimation and efficiency in semiparametric time series models defined through moment restrictions. No doubt some of our results can be replicated by their methodology using the sieve method.

We apply our method to simulated data and daily stock return data. We find superior performance of our forecasting method over some standard alternatives.

We will discuss efficient estimation of θ in section 2 and efficient estimation of q_α in section 3. Once we collect efficient estimators for these parameters, we can construct the conditional quantile estimator $\xi_t(\alpha)$ and ES estimator $\chi_t(\alpha)$ and discuss their asymptotic distribution in section 4. We present our simulation results and empirical applications in section 5. Section 6 concludes with further extensions.

3.2 Efficient estimation of θ

Efficient estimation for semiparametric GARCH models was initially addressed by Engel and Gonzalez-Rivera (1991). Their Monte Carlo evidence showed that their estimation of GARCH parameters cannot fully capture the potential efficiency gain. Linton (1993) considered the ARCH(p) special case of (3.1) with no mean effect and assumed only that the errors were distributed symmetrically about zero. In that case, the error density is not jointly identified along with all the parameters, although the identified subvector is adaptively estimable. Drost and Klaassen (1997) consider a general case that allowed for different identification conditions. They showed that a subvector of the parameters can be adaptively estimated while a remaining parameter cannot be.

We rewrite the volatility model to reflect this. Specifically, now let $h_t = c^2 + ac^2 y_{t-1}^2 + bh_{t-1}$. The finite dimensional parameter in this model $\theta = (c, a, b)^\top \in \Theta \subset \mathbb{R}^3$ is to be partitioned into two parts: $(c, \beta^\top)$ where $\beta = (a, b)^\top \in B$ for the reason that only β is adaptively estimable, see Linton (1993) and Drost and Klaassen (1997). As a result, we can rewrite the volatility as $h_t(\theta) = c^2 g_t(a, b)$, where $g_t(\beta) = 1 + au_{t-1}^2 + bg_{t-1}(\beta)$.

In the sequel we will use the following notations frequently: moment conditions $R_1(\varepsilon) = 1(\varepsilon \leq q_\alpha) - \alpha$, $R_2(\varepsilon) = \left(\varepsilon, \varepsilon^2 - 1\right)^\top$; the Fisher scale score $R_3(\varepsilon) = 1 + \varepsilon \frac{f'(\varepsilon)}{f(\varepsilon)}$ of the error density f; derivatives $G_t(\beta) = \partial \log g_t(\beta)/\partial\beta$, $G(\beta) = E[G_t(\beta)]$, $H_t(\theta) = \partial \log h_t(\theta)/\partial\theta$, $H(\theta) = E[H_t(\theta)]$,

$$G_2(\beta) = E\left[\frac{\partial \log g_t(\beta)}{\partial\beta}\frac{\partial \log g_t(\beta)}{\partial\beta^\top}\right] \quad ; \quad H_2(\theta) = E\left[\frac{\partial \log h_t(\theta)}{\partial\theta}\frac{\partial \log h_t(\theta)}{\partial\theta^\top}\right].$$

When the argument is evaluated at the true value, we use abbreviation: for example, $G = G(\beta_0)$ and $H_t = H_t(\theta_0)$.

The log-likelihood of observations $\{y_1, \ldots, y_n\}$ (given h_0) assuming that f is known is

$$\mathcal{L}(\theta) = \sum_{t=1}^n \log f(c^{-1} g_t^{-1/2}(\beta) y_t) + \log c^{-1} g_t^{-1/2}(\beta).$$

Then the score function in the parametric model at time t as

$$l_t(\theta) = -\frac{1}{2}\left(1 + \varepsilon_t(\theta)\frac{f'(\varepsilon_t(\theta))}{f(\varepsilon_t(\theta))}\right)\frac{\partial \log h_t(\theta)}{\partial\theta}.$$

We now consider the semiparametric model where f is unknown. To see why the parameter θ is not adaptively estimable, we consider the density function $f(x; \eta)$ with a shape parameter $\eta \in \Upsilon$. It is clear from $E[\partial l_t(\theta; \eta)/\partial\eta] \neq 0$ that the estimation of η affects the efficiency of the estimates of θ. If we knew the density function $f(\cdot)$ and are interested in estimating β in presence of the nuisance parameter c, the efficient score function for β is the vector

$$l_{1t}^*(\beta) = -\frac{1}{2}\{G_t(\beta) - G(\beta)\} R_3(\varepsilon_t), \tag{3.3}$$

according to the Convolution Theorem 2.2 in Drost and Klaassen (1997). The density function $f(\cdot)$ is unknown. Drost and Klaassen (1997) showed that introduction of unknown $f(\cdot)$ in presence of unknown c does not change the efficient influence function for β.

We make the following assumptions:

Assumptions A

A1. $c > 0, a \geq 0$ and $b \geq 0$. $E[\ln\{b + ac^2\varepsilon_t^2\}] < 0$.

A2. The density function f satisfies the moment restrictions: $\int xf(x)dx = 0$ and $\int x^2 f(x)dx = 1$; it has finite fourth moment $\int x^4 f(x)dx < \infty$, and $E\varepsilon^4 - 1 - (E\varepsilon^3)^2 \neq 0$.

A3. The density function f is positive and f' is absolutely continuous with

$$||f||_\infty = \sup_{x\in\mathbb{R}} f(x) < \infty, \sup_{x\in\mathbb{R}} |x|f(x) < \infty, \int |x|f(x)dx < \infty.$$

A4. The density function f has positive and finite Fisher information for scale

$$0 < \int (1 + xf'(x)/f(x))^2 f(x)dx < \infty.$$

A5. The density function f for the initial value h_{01} satisfies that, the likelihood ratio for h_{01},

$$l_n(h_{01}) = \log\{f_{\widetilde{\theta}_n}/f_{\theta_n}(h_{01}) \xrightarrow{P_n} 0, \quad as\ n \to \infty$$

where the contiguous parameter sequences $\widetilde{\theta}_n$ and θ_n are defined as in Drost and Klaassen (1997, p199).

Remark. Assumption A.1 ensures the positivity of h_t and the strict stationarity of y_t. Since $E[\ln\{b + ac^2\varepsilon_t^2\}] \leq b + ac^2 - 1$, a sufficient condition for strict stationarity is $b + ac^2 < 1$, see Nelson (1990). A.2 is introduced to make sure that the variance matrix $E[R_2(\varepsilon)R_2(\varepsilon)^\top]$ is invertible A.3 is made because we will need some boundedness of f to make a uniform expansion for the empirical distributions, see section 3. A.4 is typically assumed for efficiency discussion, see for example, Linton (1993) and Drost and Klaassen (1997). A.5 is assumed to obtain the uniform LAN theorem and the Convolution Theorem, as in Drost and Klaassen (1997).

We will suppose that there exists an initial $\sqrt{T}$-consistent estimator of all the parameters, for example the QMLE. The large sample property of GARCH parameters has been studied in different context. For example, Lee and Hansen (1994) and Berkes et. al. (2003) for detailed consistency discussion of Gaussian QMLE, and Weiss (1986) for OLS. Jensen and Rahbek (2004) considered the asymptotic theory of QMLE for nonstationary GARCH models. We have the following result which extends Drost and Klaassen (1997) and Drost, Klaassen, and Werker (1997).

Theorem 1 *Suppose that assumptions A hold. Then there exists an efficient estimator $\widehat{\theta}$ that has the following expansion*

$$\sqrt{n}(\widehat{\theta}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\psi_t(\theta_0)+o_p(1), \tag{3.4}$$

$$\psi_t(\theta_0)=\begin{pmatrix} -\frac{1}{2}E[l_{1t}^{*}l_{1t}^{*\top}]^{-1}\{G_t-G\} & 0 \\ \frac{c_0}{4}G^{\top}E[l_{1t}^{*}l_{1t}^{*\top}]^{-1}\{G_t-G\} & \frac{c_0}{2}(-E\varepsilon^3,1) \end{pmatrix}\begin{pmatrix} R_3(\varepsilon_t) \\ R_2(\varepsilon_t) \end{pmatrix}.$$

Consequently,

$$\sqrt{n}(\widehat{\theta}-\theta_0)\Longrightarrow N(0,\Omega_\theta),$$

$$\Omega_\theta=\begin{pmatrix} E[l_{1t}^{*}l_{1t}^{*\top}]^{-1} & -\frac{c_0}{2}E[l_{1t}^{*}l_{1t}^{*\top}]^{-1}G \\ -\frac{c_0}{2}G^{\top}E[l_{1t}^{*}l_{1t}^{*\top}]^{-1} & \frac{c_0^2}{4}\{E\varepsilon^4-1-(E\varepsilon^3)^2+G^{\top}E[l_{1t}^{*}l_{1t}^{*\top}]^{-1}G\} \end{pmatrix}.$$

For technical reasons, the estimator employed in the theorem makes use of sample splitting, discretization, and trimming in order to facilitate the proof. In practice, none of these devices may be desirable.

We have found the following estimator scheme works well in practice. Suppose that $k(\cdot)$ is a symmetric, second-order kernel function with $\int k(x)dx=1$ and $\int xk(x)dx=0$, and let h and b be positive bandwidths that (in the theory will satisfy $h\to 0, nh^4\to\infty$, $b\to 0, nb^4\to\infty$).

Estimation Algorithm

1. Let $\widehat{\theta}_1=(\widehat{\beta}_1^{\top},\widehat{c}_1)^{\top}$ be an initial $\sqrt{T}$-consistent estimator, for example the QMLE, and compute the residuals $\widehat{\varepsilon}_{1t}=y_t/h_t^{1/2}(\widehat{\theta}_1)$.

2. Update the estimator of β by using the Newton–Raphson method:

$$\widehat{\beta}=\widehat{\beta}_1+\left[\frac{1}{n}\sum\nolimits_{t=1}^{n}\widehat{l}_{1t}^{*}(\widehat{\beta}_1)\widehat{l}_{1t}^{*}(\widehat{\beta}_1)^{\top}\right]^{-1}\frac{1}{n}\sum\nolimits_{t=1}^{n}\widehat{l}_{1t}^{*}(\widehat{\beta}_1)$$

$$\widehat{l}_{1t}^{*}(\widehat{\beta}_1)=-\frac{1}{2}\left[G_t(\widehat{\beta}_1)-\frac{1}{n}\sum\nolimits_{s=1}^{n}G_s(\widehat{\beta}_1)\right]\widehat{R}_3(\widehat{\varepsilon}_{1t}),\quad \widehat{R}_3(x)=1+x\widehat{f}'(x)/\widehat{f}(x)$$

$$\widehat{f}(x)=\frac{1}{nh}\sum\nolimits_{t=1}^{n}k\left(\frac{\widehat{\varepsilon}_{1t}-x}{h}\right)\quad ;\quad \widehat{f}'(x)=-\frac{1}{nb^2}\sum\nolimits_{t=1}^{n}k'\left(\frac{\widehat{\varepsilon}_{1t}-x}{b}\right).$$

3. Denote $\widehat{e}_t = y_t g_t^{-1/2}(\widehat{\beta})$ and the efficient estimator for c is

$$\widehat{c} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \widehat{e}_t^2 - \frac{1}{n}\frac{\sum_{t=1}^{n} \widehat{e}_t^3}{\sum_{t=1}^{n} \widehat{e}_t^2}\sum_{t=1}^{n} \widehat{e}_t}.$$

This procedure can be repeated until some convergence criterion is met, although for most theoretical purposes, one iteration is sufficient.

Remark. It can be shown that the simpler estimator $\widetilde{c} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \widehat{e}_t^2}$ has an asymptotic variance $c_0^2\{E\varepsilon^4 - 1 + G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1}G\}/4$, which is strictly larger than our efficient estimator $\widehat{c}$ unless the error distribution is symmetric, i.e. $E\varepsilon^3 = 0$.

3.3 Efficient estimation of q_α and ES_α

We now turn to the estimation of the quantities of interest. To motivate our theory, we first discuss the estimation of q_α with the availability of true errors, and then discuss what to do in the case of estimation errors.

3.3.1 Quantile estimation with true errors available

In this subsection we estimate the quantile by inverting various distribution estimators. Because the unknown error distribution satisfies condition (3.2), it is desirable to construct distribution estimators that have this property.

The empirical distribution function $\widehat{F}(x) = n^{-1}\sum_{t=1}^{n} 1(\varepsilon_t \le x)$ is commonly used but it does not impose these moment constraints. In practice, a common approach is to recenter the errors. Therefore, we also consider a modified empirical distribution, $\widehat{F}_N(x) = n^{-1}\sum_{t=1}^{n} 1((\varepsilon_t - \widehat{\mu}_\varepsilon)/\widehat{\sigma}_\varepsilon \le x)$, where $\widehat{\mu}_\varepsilon = n^{-1}\sum_{t=1}^{n}\varepsilon_t$ and $\widehat{\sigma}_\varepsilon^2 = n^{-1}\sum_{t=1}^{n}\varepsilon_t^2 - (n^{-1}\sum_{t=1}^{n}\varepsilon_t)^2$. By construction, this distribution estimator satisfies the moment constraints (3.2). It is easy to see that the relationship between $\widehat{F}(x)$ and $\widehat{F}_N(x)$ is $\widehat{F}_N(x) = \widehat{F}(\widehat{\mu}_\varepsilon + x\widehat{\sigma}_\varepsilon)$.

In this paper, we consider a new weighted empirical distribution estimator $\widehat{F}_w(x) = \sum_{t=1}^{n}\widehat{w}_t 1(\varepsilon_t \le x)$, where the empirical likelihood weights $\{\widehat{w}_t\}$ come from the following:

$$\max_{\{w_t\}} \Pi_{t=1}^{n} w_t$$
$$\text{s.t.} \sum_{t=1}^{n} w_t = 1; \sum_{t=1}^{n} w_t\varepsilon_t = 0; \sum_{t=1}^{n} w_t(\varepsilon_t^2 - 1) = 0.$$

By construction, $\widehat{F}_w$ satisfies the moment restrictions.

In the absence of the moment constraints, it is easy to see that $\arg\max_{\{w_t\}}\{\Pi_{t=1}^{n} w_t + \lambda(1 - \sum_{t=1}^{n} w_t)\} = 1/n$. In this case our weighted empirical distribution estimator is the same as $\widehat{F}(x)$. Since the unknown distribution is in the family $\mathcal{P} = \{f(x) : \int xf(x)dx = 0, \int(x^2-1)f(x)dx = 0\}$, we expect $\widehat{F}_w(x)$ to be more efficient by incorporating these moment constraints, Bickel, Klaassen, Ritov, and Wellner (1993). Lemma 1 (appendix) which shows the uniform expansion for the distribution estimators $\widehat{F}(x)$, $\widehat{F}_N(x)$ and $\widehat{F}_w$ confirms our conjecture. It is well-known that $\sqrt{n}(\widehat{F}(x) - F(x)) \Longrightarrow N(0, F(x)(1-F(x)))$. The empirical distribution is the most efficient estimator without any auxiliary information about $F(\cdot)$. This is consistent with our result because $w_t = 1/n$ is the solution to the problem of $\max_{\{w_t\}}\{\Pi_{t=1}^{n} w_t + \lambda(1 - \sum_{t=1}^{n} w_t)\}$.

We obtain an asymptotic expansion for $\widehat{F}_N(x)$ and $\widehat{F}_w(x)$ in the appendix (Lemma 1) and show that:

$$\begin{aligned}
\sqrt{n}(\widehat{F}_N(x) - F(x)) &\Longrightarrow N(0, F(x)(1-F(x)) + C_x) \\
\sqrt{n}(\widehat{F}_w(x) - F(x)) &\Longrightarrow N(0, F(x)(1-F(x)) - A_x^\top B^{-1} A_x).
\end{aligned}$$

We can see that normalization has introduced some additional error; see Durbin (1973). This estimation error has been cumulated and is reflected by the additional term C_x in the asymptotic variance. The sign of C_x function is indeterminate, see the Figure 1 in the appendix. It depends on the density $f(x)$ and the point to be evaluated. For standard normal distribution and student distributions, $C_x \leq 0$, which means, for these two distributions, $\widehat{F}_N(x)$ is more efficient than $\widehat{F}(x)$. In contrast, for mixed normal distribution and Chi-squared distributions, the efficiency ranking depends on the point to be evaluated. On the other hand, weighting the empirical distribution takes into account the information in (3.2), which is reflected in the term $-A_x^\top B^{-1} A_x$. This term can be explained as the projection of $1(\varepsilon \leq x) - F(x)$ onto $R_2(\varepsilon)$. The covariance A_x measures the relevance of moment constraints (3.2) in estimating the distribution function. The information content that helps in estimating unknown $F(x)$ is weak when A_x is small. In case of $A_x = 0$, the moment constraints (3.2) do not have any explanation power at all since $1(\varepsilon \leq x) - F(x)$ and $R_2(\varepsilon)$ is orthogonal. In the appendix we give conditions under which $\widehat{F}_N(x)$ and $\widehat{F}(x)$ can be as efficient as $\widehat{F}_w(x)$.

We now define our quantile and expected shortfall estimators. For an estimated c.d.f., $\widetilde{F}$, let

$$\widetilde{q}_\alpha = \sup\{t : \widetilde{F}(t) \leq \alpha\} = \widetilde{F}^{-1}(\alpha) \quad ; \quad \widetilde{ES}_\alpha = \frac{1}{\alpha}\int_{-\infty}^{\widetilde{q}_\alpha} x d\widetilde{F}(x). \tag{3.5}$$

$\widehat{\theta}_1 = \widehat{q}_\alpha, \widehat{\theta}_2 = \widehat{q}_{N\alpha}, \widehat{\theta}_3 = \widehat{q}_{w\alpha}, \widehat{\theta}_4 = \widehat{ES}_\alpha, \widehat{\theta}_5 = \widehat{ES}_{N\alpha}$, and $\widehat{\theta}_6 = \widehat{ES}_{w\alpha}$ be defined from (3.5) using the $\widehat{F}(x)$, $\widehat{F}_N(x)$, and $\widehat{F}_w(x)$ as required. The next theorem presents the asymptotic distribution of these quantile estimators. Define:

$$
\begin{aligned}
V_1 &= \frac{\alpha(1-\alpha)}{f(q_\alpha)^2} \quad ; \quad V_2 = \frac{\alpha(1-\alpha)}{f(q_\alpha)^2} + \frac{C_{q_\alpha}}{f(q_\alpha)^2} \quad ; \quad V_3 = \frac{\alpha(1-\alpha)}{f(q_\alpha)^2} - \frac{A_{q_\alpha}^\top B^{-1} A_{q_\alpha}}{f(q_\alpha)^2} \\
V_4 &= \alpha^{-2}\mathrm{var}((\varepsilon - q_\alpha)1(\varepsilon \le q_\alpha)) \quad ; \quad V_5 = \alpha^{-2}\mathrm{var}((\varepsilon - q_\alpha)1(\varepsilon \le q_\alpha) - \alpha\varepsilon - \frac{\varepsilon^2}{2}\int_{-\infty}^{q_\alpha} x f(x)dx)) \\
V_6 &= \alpha^{-2}\mathrm{var}((\varepsilon - q_\alpha)1(\varepsilon \le q_\alpha) + R_2^\top(\varepsilon)B^{-1}\int_{-\infty}^{q_\alpha} A_x dx).
\end{aligned}
$$

Theorem 2 *Suppose that assumptions A.1-A.5 hold. The quantile and expected shortfall estimators are asymptotically normal*

$$\sqrt{n}(\widehat{\theta}_j - \theta_j) \Longrightarrow N(0, V_j)$$

for $j = 1, \ldots, 6$, *where* $\theta_1 = \theta_2 = \theta_3 = q_\alpha$ *and* $\theta_4 = \theta_5 = \theta_6 = ES_\alpha$.

Remark. It is clear from the comparison of asymptotic variances that $\widehat{q}_{w\alpha}$, which is based on inverting empirically weighted distribution estimators, is the most efficient one. The same conclusion holds for ES since ES is the aggregation of lower quantiles: $\widetilde{ES}_\alpha = \frac{1}{\alpha}\int_{-\infty}^{\widetilde{q}_\alpha} x d\widetilde{F}(x) = \frac{1}{\alpha}\int_0^\alpha \widetilde{q}_\alpha d\alpha$.

Remark. For improvement in mean squared efficiency, one could consider inverting the smoothed weighted empirical distribution $\widehat{F}_{sw}(x) = \sum_{t=1}^n \widehat{w}_t K(\frac{x-\varepsilon_t}{h})$ with $\widehat{F}_s(x) = n^{-1}\sum_{t=1}^n K(\frac{x-\varepsilon_t}{h})$ being a special case. However, the first order large sample properties will be the same as the unsmoothed one here. The unsmoothed distribution estimators considered in this paper are free from the complication of bandwidth choice.

3.3.2 Quantile estimation with estimated parameters

We now assume that we don't know the true parameters θ, and so we don't observe ε_t. Instead we observe the polluted error, $\varepsilon_t(\theta_n) = y_t/h_t^{1/2}(\theta_n)$, where θ_n is an estimator sequence satisfying $\theta_n - \theta_0 = O_p(n^{-1/2})$. Now we construct an efficient estimator for residual distribution $F(x)$ and then invert to get back the quantile estimator $q_{n\alpha} = F_n^{-1}(\alpha)$. We treat a general class of estimators θ_n for completeness.

Motivated by the efficiency gain shown in Lemma 1, we estimate the quantile by inverting the following distribution function estimator:

$$\widehat{\widehat{F}}_w(x) = \sum_{t=1}^n \widehat{\widehat{w}}_t 1(\varepsilon_t(\theta_n) \le x), \tag{3.6}$$

where $\{\widehat{w}_t\}$ are defined by the solution of the following optimization problem

$$\max_{\{w_t\}} \Pi_{t=1}^n w_t$$
$$\text{s.t. } \sum_{t=1}^n w_t = 1; \sum_{t=1}^n w_t\varepsilon_t(\theta_n) = 0; \sum_{t=1}^n w_t(\varepsilon_t^2(\theta_n) - 1) = 0.$$

For comparison purposes, we also consider the residual empirical distribution estimator $\widehat{\widehat{F}}(x) = n^{-1}\sum_{t=1}^n 1(\varepsilon_t(\theta_n) \leq x)$ and the standardized empirical distribution $\widehat{\widehat{F}}_N(x) = n^{-1}\sum_{t=1}^n 1((\varepsilon_t(\theta_n) - \widehat{\mu}_\varepsilon)/\widehat{\sigma}_\varepsilon \leq x)$, where $\widehat{\mu}_\varepsilon = n^{-1}\sum_{t=1}^n \varepsilon_t(\theta_n)$ and $\widehat{\sigma}_\varepsilon^2 = n^{-1}\sum_{t=1}^n \varepsilon_t^2(\theta_n) - (n^{-1}\sum_{t=1}^n \varepsilon_t(\theta_n))^2$.

Suppose that there is an estimator $\widetilde{\theta}$ that has influence function $\chi_t(\theta_0)$, i.e.

$$\sqrt{n}(\widetilde{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{t=1}^n \chi_t(\theta_0) + o_p(1). \tag{3.7}$$

In the appendix (Lemma 2 and Corollary 3) we derive uniform expansion of the distribution estimators $\widehat{\widehat{F}}(x), \widehat{\widehat{F}}_N(x)$ and $\widehat{\widehat{F}}_w(x)$ based on $\widetilde{\theta}$ and give their asymptotic variances, which depend on the influence function $\chi_t(\theta_0)$. We next explore these asymptotic variances for some widely used estimators (with expansion 3.7). Suppose that

$$\chi_t(\theta_0) = J_t(\theta_0)(\varepsilon_t^2 - 1), \tag{3.8}$$

where $J_t(\theta_0) \in \mathcal{F}_{t-1}$, so that $\chi_t(\theta_0)$ is a martingale difference sequence. Denote $J(\theta_0) = E[J_t(\theta_0)]$. Then the asymptotic variances of the three distribution estimators are respectively:

$$\begin{aligned}
\Omega_{1,J}(x) &= F(x)(1-F(x)) + \frac{[E[\varepsilon^4]-1]x^2 f(x)^2}{4}\{H(\theta_0)^\top J(\theta_0)\}^2 + xf(x)H(\theta_0)^\top J(\theta_0)a_2(x) \\
\Omega_{2,J}(x) &= F(x)(1-F(x)) + \frac{[E[\varepsilon^4]-1]x^2 f(x)^2}{4}\{H(\theta_0)^\top J(\theta_0)\}^2 + xf(x)H(\theta_0)^\top J(\theta_0)a_2(x) \\
&\quad + f(x)^2 + \frac{x^2 f(x)^2[E[\varepsilon^4]-1]}{4} + xf(x)^2 E[\varepsilon^3] + xf(x)a_2(x) + 2f(x)a_1(x) \\
&\quad + xf(x)^2 E[\varepsilon^3]H(\theta_0)^\top J(\theta_0) + \frac{x^2 f(x)^2[E[\varepsilon^4]-1]}{2}H(\theta_0)^\top J(\theta_0) \\
\Omega_{3,J}(x) &= F(x)(1-F(x)) - A_x^\top B^{-1}A_x \\
&\quad + \{E[\varepsilon^4]-1\}\left\{\frac{xf(x)}{2} + \begin{pmatrix} 0 & 1 \end{pmatrix} B^{-1}A_x\right\}^2 \{H(\theta_0)^\top J(\theta_0)\}^2.
\end{aligned}$$

In the special case of the least squares estimator,

$$\chi_t(\theta_0) = H_1(\theta_0)^{-1}h_t(\theta_0)\frac{\partial h_t(\theta_0)}{\partial\theta}(\varepsilon_t^2 - 1),$$

where $H_1(\theta_0) = E[\frac{\partial h_t(\theta_0)}{\partial \theta}\frac{\partial h_t(\theta_0)}{\partial \theta^\top}]$. Denote $H_3(\theta_0) = E[h_t(\theta_0)\frac{\partial h_t(\theta_0)}{\partial \theta}]$, then $J_t(\theta_0) = H_1(\theta_0)^{-1}H_{3t}(\theta_0)$. In the special case of the Gaussian QMLE,

$$\chi_t(\theta_0) = \{H_2(\theta_0)\}^{-1}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial \theta}(\varepsilon_t^2 - 1),$$

then $J_t(\theta_0) = H_2(\theta_0)^{-1}H_t(\theta_0)$. In both cases the asymptotic variance is increased relative to Lemma 1. Since the QMLE residuals $\varepsilon_t(\widetilde{\theta})$ are obtained under the moment condition $n^{-1}\sum_{t=1}^n[H_t(\widetilde{\theta})(\varepsilon_t^2(\widetilde{\theta}) - 1)] = 0$ with probability one, the first moment of $\widehat{\widehat{F}}(x)$ is $\int x d\widehat{\widehat{F}}(x) = n^{-1}\sum_{t=1}^n(\varepsilon_t^2(\widetilde{\theta}) - 1)$, which may not be zero with probability one.

We construct quantile estimators by inverting these distribution estimators. Based on the asymptotic expansion of the distribution estimators in the appendix (Lemma 2), we obtain the asymptotic properties of the Value at Risk and Expected shortfall estimators, which is the main result of the paper. Let $\widehat{\widehat{\theta}}_1 = \widehat{\widehat{q}}_\alpha, \widehat{\widehat{\theta}}_2 = \widehat{\widehat{q}}_{N\alpha}, \widehat{\widehat{\theta}}_3 = \widehat{\widehat{q}}_{w\alpha}, \widehat{\widehat{\theta}}_4 = \widehat{\widehat{ES}}_\alpha, \widehat{\widehat{\theta}}_5 = \widehat{\widehat{ES}}_{N\alpha}$, and $\widehat{\widehat{\theta}}_6 = \widehat{\widehat{ES}}_{w\alpha}$ be defined from (3.5) using the estimated c.d.f.s $\widehat{\widehat{F}}(x)$, $\widehat{\widehat{F}}_N(x)$, and $\widehat{\widehat{F}}_w(x)$ as required (and define correspondingly, $\theta_1 = \theta_2 = \theta_3 = q_\alpha$ and $\theta_4 = \theta_5 = \theta_6 = ES_\alpha$.). Define the asymptotic covariance matrices:

$$\begin{aligned}
\Omega_1 &= \Omega_\alpha = \frac{\alpha(1-\alpha)}{f(q_\alpha)^2} + \frac{q_\alpha^2}{4}[E\varepsilon^4 - 1 - (E\varepsilon^3)^2] + \frac{q_\alpha(a_{2q_\alpha} - a_{1q_\alpha}E\varepsilon^3)}{f(q_\alpha)} \\
\Omega_2 &= \Omega_{N\alpha} = \frac{\alpha(1-\alpha)}{f(q_\alpha)^2} + \frac{C_{q_\alpha}}{f(q_\alpha)^2} + \frac{3q_\alpha^2[E\varepsilon^4 - 1 - (E\varepsilon^3)^2]}{4} + \frac{q_\alpha(a_{2q_\alpha} - a_{1q_\alpha}E\varepsilon^3)}{f(q_\alpha)} \\
\Omega_3 &= \Omega_{w\alpha} = \frac{\alpha(1-\alpha)}{f(q_\alpha)^2} - \frac{A_{q_\alpha}^\top B^{-1}A_{q_\alpha}}{f(q_\alpha)^2} + [\frac{q_\alpha}{2} + \frac{\begin{pmatrix} 0 & 1 \end{pmatrix} B^{-1}A_{q_\alpha}}{f(q_\alpha)}]^2[E\varepsilon^4 - 1 - (E\varepsilon^3)^2] \\
\Omega_4 &= \Omega_{ES} = \alpha^{-2}\mathrm{var}\left((\varepsilon - q_\alpha)1(\varepsilon \le q_\alpha) - \frac{\varepsilon^2 - \varepsilon E\varepsilon^3}{2}\int_{-\infty}^{q_\alpha} xf(x)dx\right) \\
\Omega_5 &= \Omega_{ESN} = \alpha^{-2}\mathrm{var}\left((\varepsilon - q_\alpha)1(\varepsilon \le q_\alpha) - \varepsilon\int_{-\infty}^{q_\alpha}[f(x) - \frac{xf(x)}{2}E\varepsilon^3]dx - \varepsilon^2\int_{-\infty}^{q_\alpha} xf(x)dx\right) \\
\Omega_6 &= \Omega_{ESW} = \alpha^{-2}\mathrm{var}\left((\varepsilon - q_\alpha)1(\varepsilon \le q_\alpha) - (\varepsilon^2 - \varepsilon E\varepsilon^3)\int_{-\infty}^{q_\alpha}[\frac{xf(x)}{2} + \begin{pmatrix} 0 & 1 \end{pmatrix} B^{-1}A_x]dx\right. \\
&\quad \left. + R_2^\top(\varepsilon)B^{-1}\int_{-\infty}^{q_\alpha} A_x dx\right).
\end{aligned}$$

Theorem 3 *Suppose that assumptions A.1-A.5 hold. The quantile and expected shortfall estimators are asymptotically normal:*

$$\sqrt{n}(\widehat{\widehat{\theta}}_j - \theta_j) \Longrightarrow N(0, \Omega_j)$$

for $j = 1, \ldots, 6$.

Remark. For the same reason as given in the discussion about the efficiency of distribution estimators, we can see that $\widehat{\widehat{q}}_{w\alpha}$ is more efficient than $\widehat{\widehat{q}}_{\alpha}$. The same conclusion holds for ES.

Remark. Notice that the asymptotic variances of VaRs and ESs do not contain any functional form of the heteroskedasticity. This is due to the orthogonality in information between estimators for the distribution $F(x)$ and variance estimator for β.

Remark. We can compute consistent standard errors by the obvious plug-in method.

3.4 Efficient estimation of conditional VaR and conditional expected shortfall

We have discussed the asymptotic property of efficient estimators $\widehat{\theta}$ and $\widehat{\widehat{q}}_{w\alpha}$. They are shown to be the best among competitors in terms of smallest asymptotic variances. Both are important ingredients to the conditional quantile estimator $\widehat{\xi}_{n,t}$ as $\widehat{\xi}_{n,t} = h_t^{1/2}(\widehat{\theta})\widehat{\widehat{q}}_{w\alpha}$ and the conditional expected shortfall $\widehat{\chi}_{n,t} = h_t^{1/2}(\widehat{\theta})\widehat{\widehat{ES}}_{w\alpha}$. In this section, we will show that these two quantities are asymptotically mixed normal. Define:

$$\omega_{\xi t} = h_t(\theta_0)\left\{\frac{q_\alpha^2}{4}(G_t^\top - G)E[l_{1t}^* l_{1t}^{*\top}]^{-1}(G_t - G) + \Omega_{w\alpha}\right\},$$

$$\omega_{\chi t} = h_t(\theta_0)\left\{\frac{ES_\alpha^2}{4}H_t^\top \Omega_\theta H_t + ES_\alpha \frac{(-E\varepsilon^3, 1)E\left[\left\{(\varepsilon_t - q_\alpha)1(\varepsilon_t \le q_\alpha) + R_2^\top(\varepsilon_t)C\right\}R_2(\varepsilon_t)\right]}{\alpha} + \Omega_{ESW}\right\}$$

$$C = \int_{-\infty}^{q_\alpha}[\frac{xf(x)}{2} + \begin{pmatrix} 0 & 1 \end{pmatrix} B^{-1}A_x]dx(E\varepsilon^3, -1) + \int_{-\infty}^{q_\alpha} A_x^\top dx B^{-1}.$$

Theorem 4 *Suppose assumptions A.1-A.5 hold. The conditional quantile estimator $\widehat{\xi}_{n,t}$ and conditional quantile estimator $\widehat{\chi}_{n,t}$ are asymptotically mixed normal*

$$\begin{aligned}\sqrt{n}(\widehat{\xi}_{n,t} - \xi_t) &\implies MN(0, \omega_{\xi t})\\ \sqrt{n}(\widehat{\chi}_{n,t} - \chi_t) &\implies MN(0, \omega_{\chi t}),\end{aligned}$$

where the random positive scalars $\omega_{\xi t}$ and $\omega_{\chi t}$ are independent of the underlying normals.

Remark. From the influence functions of $(\widehat{a}, \widehat{b})$ and $\widehat{\widehat{q}}_{w\alpha}$, we can see that they are asymptotically orthogonal. This is anticipated as the parameter (a, b) is adaptively estimated with respect to the error distribution.

Remark. This mixed normal distribution asymptotics is similar to results obtained in Barndorff-Nielsen and Shephard (2002) for estimation of the quadratic variation of Brownian semimartingales, see also Hall and Heyde (1980). It follows that $\sqrt{n}(\widehat{\xi}_{n,t} - \xi_t)/\omega_{\xi t}^{1/2} \Longrightarrow N(0,1)$ and $\sqrt{n}(\widehat{\xi}_{n,t} - \xi_t)/\widehat{\omega}_{\xi t}^{1/2} \Longrightarrow N(0,1)$, where $\widehat{\omega}_{\xi t}$ is a consistent estimator of $\omega_{\xi t}$. Therefore, one can conduct inference about $\xi_{n,t}$ with the usual confidence intervals.

3.5 Numerical Work

In this section we present some numerical evidence. The first part is Monte-Carlo simulation and the second is an empirical application.

3.5.1 Simulations

We follow Drost and Klaassen (1997) to simulate several GARCH (1,1) series from the model (1) with the following parameterizations:

1. $(c, a, b) \in \{(1, 0.3, 0.6), (1, 0.1, 0.8), (1, 0.05, 0.9)\}$;
2. $f(x) \in \{N(0,1), MN(2,-2), L, t(5), t(7), t(9), \chi_6^2, \chi_{12}^2\}$, which are, respectively, referred to the densities of standardized (mean 0 and variance 1) distributions from Normal, Mixed Normal with means $(2,-2)$, Laplace, student distributions with degree of freedom 5, 7 and 9 and chi-squared distribution with 6 and 12 degrees of freedom.

Sample size is set to $n = 500, 1000$. Simulations are carried out 2500 times. We consider the performance of the three distribution estimators and their associated quantile and ES estimators with α being 5% and 1%. We also have the simulation results for small samples $n = 25, 50, 100$, and for IGARCH models with $a + b = 1$. These results are similar to those in this paper and are available upon request.

The criterion for distribution estimator $\widehat{F}(x)$ is the integrated mean squared error (IMSE)

$$IMSE = \int E[\widehat{F}(x) - F(x)]^2 dx$$

and that for quantile and ES estimators ($\widehat{q}_\alpha$ and $\widehat{ES}_\alpha$) is the mean squared error

$$MSE = E[(\widehat{q}_\alpha - q_\alpha)^2]; MSE = E[(\widehat{ES}_\alpha - ES_\alpha)^2].$$

First, we consider the case where the true errors are available. The IMSEs of three distribution function estimators are summarized in Table 1. It is clear form this table that the weighted empirical distribution estimator $\widehat{F}_w(x)$ performs the best in all cases. The relative efficiency of $\widehat{F}_w(x)$ to the unweighted empirical distribution $\widehat{F}(x)$ is very large: it ranged from 50% in case of errors being Laplacian to 72% in the case of Mixed-normal. Figure 2 visualizes this gain by plotting the overlays of simulated distribution estimators with 100 replications. The colored region represents the possible paths of function estimators and it is clear that the magenta area (realizations of $\widehat{F}_w(x)$) is has the smaller width than blue area (realizations of $\widehat{F}(x)$). In order to compare the quantile estimators based on inverting these distribution estimators, we compute their average biases and mean squared errors under different distributional assumptions and in 500 and 1000 sample sizes. The average is taken over 2500 simulations. It is found that $\widehat{q}_{w\alpha}$ performs much better than $\widehat{q}_{\alpha}$ in all cases. This improvement is clearer in the case of $\alpha = 0.05$ than the case of $\alpha = 0.01$. This is because the further to the tail (when a is smaller), the smaller the covariance between $R_2(\varepsilon)$ and $1(\varepsilon \leq x) - \alpha$.

Next, we compare the distribution estimators when the errors are not observable and we use estimated errors from QMLE. Since QMLE is consistent in all above error distribution assumptions, we expect the QMLE residuals will behave close to the true errors, although with some estimation noises. Table 4-6 list the IMSE for distribution estimators under three different parameterizations. We find that, there are efficiency gains by weighting the empirical distribution estimator with empirical likelihoods. Figure 3 visualizes these gains, which vary across the assumptions of true error distributions. Table 7-12 compare the performance of residual quantile estimators. The conclusion is the same: empirical likelihood weighting reduces the variation of quantile estimators. However, these reductions are not of the same magnitude as in i.i.d case. The reason is because we use estimated errors in stead of true errors and the added estimation noise affect the performance of residual based estimators.

Thirdly, we compare different estimators for expected shortfall in the case of iid errors and GARCH residuals. As seen from table 15-18, the same conclusion holds for ES. For sample size n=500 and 1000, the proposed estimator does not do very well in the case of $a = 0.01$, see table 18. This is expected because our efficient estimator (EL-weighted) involves an additional layer of numerical optimization, and for such low quantile/ES, the effective sample size is $n/100$. Therefore we tabulate the results for large sample $n = 10000$,which is the table 19(c). It's clear from table that our proposed VaR and ES estimators outperform other estimators in terms of smaller MSE.(The comparison of the estimators for $\widehat{q}_{0.01}$ and $\widehat{ES}_{0.01}$,when the true errors are available and $\widehat{\widehat{q}}_{0.01}$ and $\widehat{\widehat{ES}}_{0.01}$, when the polluted errors are calculated are provided in table 19(a) and 19(b)).

Finally, we consider the case of distribution and quantile estimation based on efficient residuals: the estimated errors are residuals from efficient estimation of parameter θ_0. As we notice that the performance of these estimators does not change much under different parameterization of θ_0, we only report the results in the case of $c = 1, a = 0.05, b = 0.9$. Table 13 summarizes the performance of quantile estimators for $q_{0.01}$ and $q_{0.05}$, while table 14 reports the true VaR and ES for distribution estimators and Figure 4 visualize the efficiency gains.

Table 1. Integrated Mean Squared Error ($\times 10^{-3}$) of Distribution Function Estimators

	$n = 500$			$n = 1000$		
	$\widehat{F}(x)$	$\widehat{F}_N(x)$	$\widehat{F}_w(x)$	$\widehat{F}(x)$	$\widehat{F}_N(x)$	$\widehat{F}_w(x)$
N	0.3365	0.1199	0.1212	0.1616	0.0580	0.0583
MN	0.3286	0.1412	0.0916	0.1622	0.0687	0.0462
L	0.3313	0.2188	0.1603	0.1692	0.1092	0.0810
$t(5)$	0.3419	0.2157	0.1635	0.1657	0.1055	0.0797
$t(7)$	0.3255	0.1594	0.1458	0.1695	0.0791	0.0708
$t(9)$	0.3336	0.1439	0.1361	0.1664	0.0730	0.0687
χ^2_6	0.3308	0.1479	0.1217	0.1692	0.0721	0.0605
χ^2_{12}	0.3297	0.1335	0.1213	0.1692	0.0654	0.0595

Table 2. Comparison of quantile estimators for $q_{0.01}$ (true errors are available)

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{q}_\alpha$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{q}_\alpha$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{q}_\alpha$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{q}_\alpha$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$
N	-2.1	-0.8	3.7	27.1	19.4	22.2	-0.7	-1.1	0.6	14.1	9.9	11.1
MN	-1.8	-2.5	-0.5	6.3	5.8	5.6	-1.3	-0.9	-0.4	3.3	3.1	2.9
L	-3.7	3.3	17.9	94.3	60.9	68.6	11.9	9.4	11.8	47.8	31.6	31.9
$t(5)$	25.1	25.3	40.7	102.5	67.5	72.8	8.4	11.4	22.2	50.4	34.1	35.1
$t(7)$	8.4	10.5	21.1	65.4	45.3	50.9	-1.7	0.1	4.9	34.9	23.1	23.6
$t(9)$	10.5	9.3	15.4	56.1	36.7	41.8	11.4	8.1	8.9	30.9	20.0	21.3
χ^2_6	-3.4	0.4	-1.7	1.7	3.7	1.7	-1.0	0.5	-0.4	0.9	1.8	0.8
χ^2_{12}	-4.6	-3.5	-3.1	4.9	6.0	4.8	-2.8	-0.9	-1.7	2.5	3.1	2.3

Table 3. Comparison of quantile estimators for $q_{0.05}$ (true errors are available)

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{q}_{\alpha}$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{q}_{\alpha}$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{q}_{\alpha}$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{q}_{\alpha}$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$
N	-3.6	-1.6	-1.0	8.9	4.2	4.4	-1.2	-0.6	0.1	4.5	2.0	2.0
MN	-1.4	0.0	-0.1	2.4	2.0	1.4	-0.8	-1.1	0.0	1.3	1.0	0.7
L	0.8	4.6	7.9	18.2	8.8	8.8	2.6	2.7	4.6	9.4	4.9	4.6
$t(5)$	0.7	5.0	12.1	13.6	9.0	8.1	-1.4	2.7	4.9	6.9	4.8	3.9
$t(7)$	-2.2	1.3	5.3	12.2	6.3	6.7	0.6	2.9	3.5	6.3	3.3	3.2
$t(9)$	-2.6	0.2	1.9	11.7	6.0	6.0	3.5	2.1	3.1	5.8	2.9	2.8
χ^2_6	-1.4	1.3	0.4	1.5	2.1	1.1	-2.3	0.2	-1.0	0.7	1.1	0.6
χ^2_{12}	-1.7	-0.7	1.1	2.8	2.4	1.9	-2.0	-1.2	-0.8	1.4	1.3	1.0

Table 4. Integrated Mean Squared Error ($\times 10^{-3}$), $c = 1, a = 0.3, b = 0.6$.

	$n = 500$			$n = 1000$		
	$\widehat{\widehat{F}}(x)$	$\widehat{\widehat{F}}_N(x)$	$\widehat{\widehat{F}}_w(x)$	$\widehat{\widehat{F}}(x)$	$\widehat{\widehat{F}}_N(x)$	$\widehat{\widehat{F}}_w(x)$
N	0.3018	0.1196	0.1193	0.1498	0.0595	0.0595
MN	0.2954	0.1445	0.0940	0.1508	0.0726	0.0468
L	0.3249	0.2188	0.1699	0.1653	0.1089	0.0848
$t(5)$	0.3551	0.2069	0.1859	0.1751	0.1031	0.0951
$t(7)$	0.3211	0.1550	0.1490	0.1631	0.0794	0.0762
$t(9)$	0.3176	0.1428	0.1389	0.1599	0.0717	0.0703
χ^2_6	0.3999	0.1434	0.1548	0.2080	0.0739	0.0808
χ^2_{12}	0.3415	0.1302	0.1333	0.1745	0.0664	0.0678

Table 5. Integrated Mean Squared Error ($\times 10^{-3}$), $c = 1, a = 0.1, b = 0.8$.

	$n = 500$			$n = 1000$		
	$\widehat{\widehat{F}}(x)$	$\widehat{\widehat{F}}_N(x)$	$\widehat{\widehat{F}}_w(x)$	$\widehat{\widehat{F}}(x)$	$\widehat{\widehat{F}}_N(x)$	$\widehat{\widehat{F}}_w(x)$
N	0.3023	0.1195	0.1193	0.1499	0.0594	0.0595
MN	0.2960	0.1442	0.0937	0.1511	0.0726	0.0468
L	0.3250	0.2185	0.1700	0.1655	0.1088	0.0848
$t(5)$	0.3591	0.2067	0.1891	0.1763	0.1029	0.0962
$t(7)$	0.3222	0.1550	0.1495	0.1635	0.0792	0.0764
$t(9)$	0.3187	0.1428	0.1396	0.1600	0.0715	0.0702
χ^2_6	0.4014	0.1430	0.1547	0.2076	0.0738	0.0807
χ^2_{12}	0.3422	0.1301	0.1335	0.1740	0.0663	0.0672

Table 6. Integrated Mean Squared Error ($\times 10^{-3}$), $c = 1, a = 0.05, b = 0.9$.

	$n = 500$			$n = 1000$		
	$\widehat{\widehat{F}}(x)$	$\widehat{\widehat{F}}_N(x)$	$\widehat{\widehat{F}}_w(x)$	$\widehat{\widehat{F}}(x)$	$\widehat{\widehat{F}}_N(x)$	$\widehat{\widehat{F}}_w(x)$
N	0.3026	0.1193	0.1191	0.1500	0.0594	0.0595
MN	0.2964	0.1440	0.0937	0.1511	0.0727	0.0468
L	0.3255	0.2182	0.1700	0.1656	0.1088	0.0848
$t(5)$	0.3607	0.2069	0.1902	0.1769	0.1034	0.0968
$t(7)$	0.3232	0.1557	0.1502	0.1636	0.0791	0.0763
$t(9)$	0.3187	0.1428	0.1393	0.1602	0.0715	0.0702
χ^2_6	0.4021	0.1425	0.1548	0.2079	0.0737	0.0808
χ^2_{12}	0.3432	0.1299	0.1336	0.1741	0.0663	0.0673

Table 7. Comparison of quantile estimators for $q_{0.01}$, with $c = 1, a = 0.05, b = 0.9$

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$
N	-2.6	-5.1	-3.5	21.2	18.6	20.6	-2.5	-4.9	-4.3	10.8	9.6	10.0
MN	-2.3	-3.9	-3.8	5.9	6.1	5.8	-0.8	-2.9	-1.8	3.0	3.2	3.0
L	0.8	-3.7	-1.8	61.2	54.0	58.1	3.3	1.2	1.1	35.1	31.0	29.8
$t(5)$	28.4	23.5	22.5	73.8	65.5	66.8	9.2	9.0	12.1	36.9	33.1	32.5
$t(7)$	5.5	1.3	1.2	48.9	43.1	45.7	1.4	0.3	-0.3	25.7	22.6	22.2
$t(9)$	1.0	-0.4	1.8	39.8	34.8	38.9	1.2	-1.2	-1.2	20.2	18.0	18.9
χ^2_6	27.3	21.4	25.8	7.0	4.5	5.9	13.5	10.5	13.3	3.4	2.1	2.8
χ^2_{12}	13.9	10.0	13.4	8.7	6.5	7.4	8.5	5.5	7.8	4.1	3.1	3.4

Table 8. Comparison of quantile estimators for $q_{0.01}$, with $c = 1, a = 0.3, b = 0.6$

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$
N	-2.3	-4.3	-2.4	21.5	18.9	20.7	-1.6	-4.0	-3.8	11.0	9.9	10.3
MN	-2.4	-3.6	-3.7	5.8	6.1	5.7	-0.2	-2.2	-1.4	3.0	3.3	2.9
L	1.1	-3.1	-0.9	60.7	53.5	57.3	1.7	-0.2	0.4	34.8	30.5	29.7
$t(5)$	33.7	28.6	28.3	70.5	62.8	67.4	9.0	8.2	11.9	36.3	32.7	32.1
$t(7)$	4.5	0.8	-0.1	48.2	42.4	44.4	1.8	1.0	1.5	26.0	23.0	23.2
$t(9)$	2.1	1.5	3.2	40.0	35.1	38.8	0.6	-1.3	-1.9	20.4	18.2	18.8
χ^2_6	31.6	25.7	29.8	7.6	5.0	6.3	14.6	11.9	14.1	3.4	2.2	2.8
χ^2_{12}	16.2	12.6	15.8	8.7	6.6	7.5	9.8	7.2	9.1	4.2	3.2	3.5

Table 9. Comparison of quantile estimators for $q_{0.01}$, with $c = 1, a = 0.1, b = 0.8$

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_\alpha$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$
N	-3.1	-5.5	-3.6	21.4	18.9	20.5	-3.1	-4.0	-2.2	11.2	10.1	10.5
MN	-2.5	-4.0	-3.9	5.8	6.0	5.7	-1.5	-3.2	-2.1	2.9	3.2	2.8
L	-0.0	-3.8	-0.8	60.8	53.6	58.0	-0.9	-1.2	2.3	37.0	33.1	33.0
$t(5)$	29.7	25.7	22.1	71.7	64.0	67.4	9.2	9.1	14.0	37.0	33.3	34.4
$t(7)$	4.3	0.7	-0.1	48.4	42.7	45.4	4.7	3.6	4.9	26.4	23.5	23.3
$t(9)$	0.7	-0.1	2.3	39.9	35.0	38.4	3.3	1.1	1.2	20.3	18.2	18.7
χ^2_6	29.7	23.7	28.5	7.3	4.7	6.1	16.6	13.0	15.5	3.3	2.1	2.7
χ^2_{12}	15.3	11.6	14.8	8.7	6.6	7.6	8.7	5.8	8.0	4.4	3.2	3.6

Table 10. Comparison of quantile estimators for $q_{0.05}$, with $c = 1, a = 0.3, b = 0.6$

	$n=500$						$n=1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$
N	0.3	0.9	2.2	6.2	4.2	4.3	-0.7	-1.6	-0.6	3.0	2.1	2.1
MN	0.6	-1.6	0.1	1.8	2.1	1.5	-0.0	-1.0	-0.2	0.8	1.0	0.7
L	9.3	9.8	13.0	14.1	9.7	9.3	4.0	3.0	4.2	6.8	4.4	4.1
$t(5)$	19.5	15.9	18.4	12.3	8.6	8.4	8.2	7.1	9.0	6.1	4.5	4.4
$t(7)$	4.9	5.0	8.2	9.5	6.3	6.3	6.2	5.0	5.7	5.0	3.4	3.4
$t(9)$	5.5	4.6	6.6	8.4	5.7	5.8	4.5	3.1	4.1	4.6	3.1	3.0
χ^2_6	11.7	8.4	10.8	4.4	2.1	2.7	6.7	3.9	5.7	2.2	1.1	1.4
χ^2_{12}	7.0	2.4	5.5	4.7	2.4	2.7	4.1	2.2	3.6	2.3	1.3	1.4

Table 11. Comparison of quantile estimators for $q_{0.05}$, with $c=1, a=0.1, b=0.8$

	$n=500$						$n=1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$
N	3.8	2.0	3.6	6.1	4.1	4.2	1.2	0.6	1.4	3.1	2.1	2.1
MN	-0.0	-1.2	-0.7	1.8	2.1	1.5	-0.4	-1.9	-0.9	0.9	1.1	0.7
L	14.4	11.7	13.9	13.7	9.0	8.7	6.4	6.6	8.2	7.3	4.9	4.6
$t(5)$	18.9	16.3	18.5	12.1	8.7	8.6	8.8	9.1	10.5	6.6	4.9	4.8
$t(7)$	6.0	4.9	8.0	9.2	6.3	6.4	4.8	4.2	5.3	4.9	3.3	3.2
$t(9)$	5.1	4.0	5.7	8.6	5.7	5.8	6.6	4.9	5.4	4.2	2.9	3.0
χ^2_6	11.9	7.5	10.8	4.5	2.3	2.8	7.9	4.4	6.6	2.2	1.0	1.3
χ^2_{12}	8.3	5.8	7.6	4.7	2.5	2.9	4.9	2.2	3.9	2.3	1.3	1.4

Table 12. Comparison of quantile estimators for $q_{0.05}$, with $c=1, a=0.05, b=0.9$

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$
N	1.2	1.3	3.3	6.1	4.2	4.3	-0.3	-1.4	-0.3	3.0	2.1	2.1
MN	1.0	-1.6	-0.1	1.8	2.0	1.5	0.1	-1.0	-0.5	0.8	1.0	0.7
L	8.9	9.0	12.1	14.1	9.7	9.2	4.1	2.9	4.1	6.9	4.4	4.2
$t(5)$	16.1	13.2	15.4	12.8	8.8	8.9	8.0	7.0	8.8	6.3	4.6	4.5
$t(7)$	4.2	3.8	7.1	9.4	6.2	6.2	5.7	4.2	5.5	5.1	3.4	3.4
$t(9)$	4.5	3.2	5.5	8.5	5.6	5.7	5.1	3.5	4.7	4.6	3.1	3.1
χ^2_6	11.2	7.8	10.1	4.5	2.1	2.7	6.0	3.2	5.1	2.2	1.1	1.4
χ^2_{12}	6.7	1.8	4.9	4.8	2.4	2.8	3.7	1.7	3.2	2.4	1.3	1.4

Table 13. Comparison of quantile estimators, with $c = 1, a = 0.05, b = 0.9, n = 1000, s = 500$

	$q_{0.01}$						$q_{0.05}$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$
N	-2.5	-4.9	-4.3	10.8	9.6	10.0	-0.3	-1.4	-0.3	3.0	2.1	2.1
MN	-0.8	-2.9	-1.8	3.0	3.2	3.0	0.1	-1.0	-0.5	0.8	1.0	0.7
L	3.3	1.2	1.1	35.1	31.0	29.8	4.1	2.9	4.1	6.9	4.4	4.2
$t(5)$	9.2	9.0	12.1	36.9	33.1	32.5	8.0	7.0	8.8	6.3	4.6	4.5
$t(7)$	1.4	0.3	-0.3	25.7	22.6	22.2	5.7	4.2	5.5	5.1	3.4	3.4
$t(9)$	1.2	-1.2	-1.2	20.2	18.0	18.9	5.1	3.5	4.7	4.6	3.1	3.1
χ^2_6	13.5	10.5	13.3	3.4	2.1	2.8	6.0	3.2	5.1	2.2	1.1	1.4
χ^2_{12}	8.5	5.5	7.8	4.1	3.1	3.4	3.7	1.7	3.2	2.4	1.3	1.4

Table 14. True VaRs and Expected Shortfalls for standardized distributions

	$q_{0.01}$	$q_{0.05}$	$ES_{0.01}$	$ES_{0.05}$
N	-2.3263	-1.6449	-2.6655	-2.0626
MN	-1.8129	-1.4676	-1.977	-1.679
L	-2.7662	-1.6282	-3.4734	-2.3352
$t(5)$	-2.6065	-1.5608	-3.4487	-2.2388
$t(7)$	-2.5337	-1.6012	-3.1863	-2.193
$t(9)$	-2.4883	-1.6167	-3.0524	-2.1643
χ^2_6	-1.4803	-1.2600	-1.5475	-1.3932
χ^2_{12}	-1.7207	-1.3827	-1.8472	-1.5880

Table 15. Comparison of estimators for $ES_{0.05}$ (true errors are available)

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$
N	9.7	8.4	4.2	12.3	5.9	6.5	4.4	3.5	1.6	5.9	2.8	3.0
MN	5.1	3.6	2.3	3.0	2.8	2.4	1.7	2.3	1.2	1.5	1.4	1.1
L	17.6	13.9	8.4	38.3	14.3	12.7	6.3	5.6	3.1	20.1	7.8	6.5
$t(5)$	12.7	8.2	-0.0	45.7	17.3	15.6	3.3	3.1	-3.9	21.8	9.3	7.6
$t(7)$	9.0	8.9	3.8	29.5	12.5	12.3	5.5	5.0	1.9	14.4	6.0	5.8
$t(9)$	9.7	5.9	1.4	23.3	10.2	10.4	5.9	5.4	3.4	11.9	5.2	4.9
χ^2_6	3.6	2.3	2.7	1.1	2.6	1.1	1.2	-0.3	0.6	0.5	1.3	0.5
χ^2_{12}	5.3	3.7	3.4	2.8	3.0	2.3	2.3	1.6	1.3	1.3	1.5	1.1

Table 16. Comparison of estimators for $ES_{0.01}$ (true errors are available)

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$
N	29.4	29.4	22.4	40.7	32.0	55.7	19.7	19.0	18.7	19.9	15.8	21.4
MN	13.6	14.8	10.3	9.7	9.0	20.2	9.0	9.2	8.8	4.8	4.8	7.2
L	65.3	67.7	64.0	195.3	129.8	166.4	27.2	25.4	18.6	97.8	64.6	70.4
$t(5)$	69.6	69.1	70.2	331.8	213.1	209.7	20.8	23.6	22.7	180.3	114.6	102.6
$t(7)$	59.4	59.3	55.2	179.4	123.0	147.3	18.6	22.3	25.4	94.8	64.6	65.5
$t(9)$	43.0	45.4	44.5	132.8	94.8	119.3	29.5	30.3	27.2	62.6	45.5	51.6
χ^2_6	8.9	4.0	7.5	1.4	4.0	9.5	4.6	2.6	5.1	0.7	2.0	2.5
χ^2_{12}	14.6	13.6	15.1	5.6	7.1	15.5	7.1	5.4	6.4	2.7	3.5	4.9

Table 17. Comparison of estimators for $ES_{0.05}$, with $c = 1, a = 0.05, b = 0.9$

	$n = 500$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$
N	10.1	13.6	11.8	8.3	6.3	6.5	3.3	4.4	3.5	3.9	2.9	3.0
MN	4.0	5.2	5.2	2.3	2.7	2.3	1.4	2.7	1.9	1.1	1.4	1.1
L	8.3	11.6	8.3	20.8	14.5	12.6	5.6	6.5	3.8	10.9	7.7	6.5
$t(5)$	8.2	9.4	5.6	23.6	16.7	14.5	1.9	3.9	3.3	11.7	8.4	7.2
$t(7)$	11.2	12.8	8.9	16.8	12.0	11.5	5.7	8.0	7.5	8.2	5.9	5.5
$t(9)$	9.3	12.3	10.3	14.6	10.3	10.2	3.6	4.2	3.1	7.1	5.1	4.9
χ^2_6	-18.9	-14.7	-18.0	5.0	2.7	3.9	-11.9	-9.7	-11.6	2.6	1.4	1.9
χ^2_{12}	-12.2	-6.7	-10.3	5.7	3.3	4.1	-4.1	-2.4	-3.8	2.6	1.6	1.8

Table 18. Comparison of estimators for $ES_{0.01}$, with $c = 1, a = 0.05, b = 0.9$

	$n=500$						$n=1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$
N	38.8	41.6	40.9	37.4	34.7	55.4	22.8	25.4	20.4	17.5	16.3	20.5
MN	19.0	20.7	16.8	9.3	9.5	19.7	9.8	11.9	11.7	4.7	4.9	7.2
L	102.6	107.8	101.2	134.9	127.5	144.2	48.7	51.3	49.5	69.0	64.0	67.9
$t(5)$	91.7	98.4	100.9	219.9	204.3	214.4	51.9	52.6	51.1	115.1	109.2	109.6
$t(7)$	82.6	87.5	86.4	130.3	122.5	141.9	39.1	40.7	42.5	67.8	63.0	64.5
$t(9)$	69.5	71.4	65.7	99.7	93.6	115.8	39.4	42.1	43.8	51.3	48.8	53.0
χ^2_6	-39.6	-33.5	-38.9	9.3	6.5	15.6	-22.2	-19.1	-20.7	4.2	2.8	5.7
χ^2_{12}	-12.4	-8.4	-13.6	10.2	8.0	19.3	-9.4	-6.3	-7.5	5.0	3.9	6.6

Table 19. Comparison of estimators for $q_{0.01}$ and $ES_{0.01}$. (a) when the true errors are available, $n=1000$

	$n=1000$						$n=1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{q}_\alpha$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{q}_\alpha$	$\widehat{q}_{N\alpha}$	$\widehat{q}_{w\alpha}$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$	$\widehat{ES}$	$\widehat{ES}_N$	$\widehat{ES}_w$
N	-0.7	-1.1	0.6	14.1	9.9	11.1	19.7	19.0	18.7	19.9	15.8	21.4
MN	-1.3	-0.9	-0.4	3.3	3.1	2.9	9.0	9.2	8.8	4.8	4.8	7.2
L	11.9	9.4	11.8	47.8	31.6	31.9	27.2	25.4	18.6	97.8	64.6	70.4
$t(5)$	8.4	11.4	22.2	50.4	34.1	35.1	20.8	23.6	22.7	180.3	114.6	102.6
$t(7)$	-1.7	0.1	4.9	34.9	23.1	23.6	18.6	22.3	25.4	94.8	64.6	65.5
$t(9)$	11.4	8.1	8.9	30.9	20.0	21.3	29.5	30.3	27.2	62.6	45.5	51.6
χ^2_6	-1.0	0.5	-0.4	0.9	1.8	0.8	4.6	2.6	5.1	0.7	2.0	2.5
χ^2_{12}	-2.8	-0.9	-1.7	2.5	3.1	2.3	7.1	5.4	6.4	2.7	3.5	4.9

(b) with $c=1, a=0.05, b=0.9, n=1000$

	$n = 1000$						$n = 1000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$
N	-2.5	-4.9	-4.3	10.8	9.6	10.0	22.8	25.4	20.4	17.5	16.3	20.5
MN	-0.8	-2.9	-1.8	3.0	3.2	3.0	9.8	11.9	11.7	4.7	4.9	7.2
L	3.3	1.2	1.1	35.1	31.0	29.8	48.7	51.3	49.5	69.0	64.0	67.9
$t(5)$	9.2	9.0	12.1	36.9	33.1	32.5	51.9	52.6	51.1	115.1	109.2	109.6
$t(7)$	1.4	0.3	-0.3	25.7	22.6	22.2	39.1	40.7	42.5	67.8	63.0	64.5
$t(9)$	1.2	-1.2	-1.2	20.2	18.0	18.9	39.4	42.1	43.8	51.3	48.8	53.0
χ^2_6	13.5	10.5	13.3	3.4	2.1	2.8	-22.2	-19.1	-20.7	4.2	2.8	5.7
χ^2_{12}	8.5	5.5	7.8	4.1	3.1	3.4	-9.4	-6.3	-7.5	5.0	3.9	6.6

(c) with $c = 1, a = 0.05, b = 0.9, n = 10000$

	$n = 10000$						$n = 10000$					
	$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$			$Bias(\times 10^{-3})$			$MSE(\times 10^{-3})$		
	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{q}}_{\alpha}$	$\widehat{\widehat{q}}_{N\alpha}$	$\widehat{\widehat{q}}_{w\alpha}$	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$	$\widehat{\widehat{ES}}$	$\widehat{\widehat{ES}}_N$	$\widehat{\widehat{ES}}_w$
N	0.9	0.6	0.8	1.1	1.0	1.0	1.6	1.7	1.4	1.8	1.7	1.7
MN	0.3	0.1	0.2	0.3	0.3	0.3	0.5	0.6	0.4	0.4	0.5	0.4
L	0.0	0.2	0.8	3.8	3.4	3.2	3.6	4.2	4.6	7.2	6.8	6.6
$t(5)$	1.0	1.3	2.2	4.0	3.5	3.3	4.8	5.0	5.3	12.3	11.7	10.9
$t(7)$	-0.1	-0.2	0.5	2.7	2.5	2.4	4.3	4.4	3.8	7.1	6.7	6.6
$t(9)$	-0.6	-0.4	-0.0	2.2	2.0	2.0	1.6	1.9	1.8	5.1	4.8	4.7
χ^2_6	1.4	1.4	1.4	0.3	0.2	0.2	-2.6	-2.3	-2.6	0.3	0.2	0.3
χ^2_{12}	0.5	0.1	0.3	0.4	0.3	0.3	-1.4	-1.2	-1.4	0.5	0.4	0.4

Figure 2 : Efficiency comparison, "Empirical CDF" v.s. "Normalized CDF"

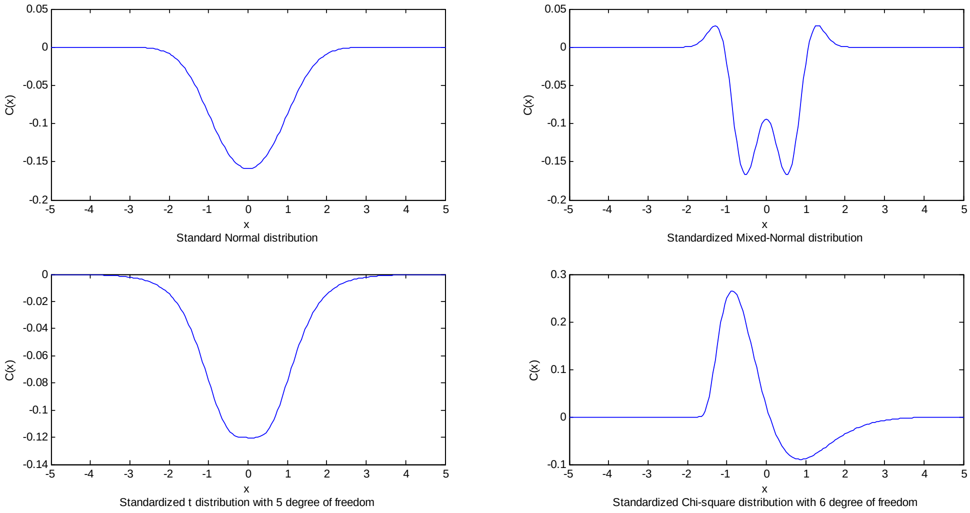


Figure 3: Overlay of two estimates using iid errors,"Empirical CDF" v.s. "EL-weighted CDF"

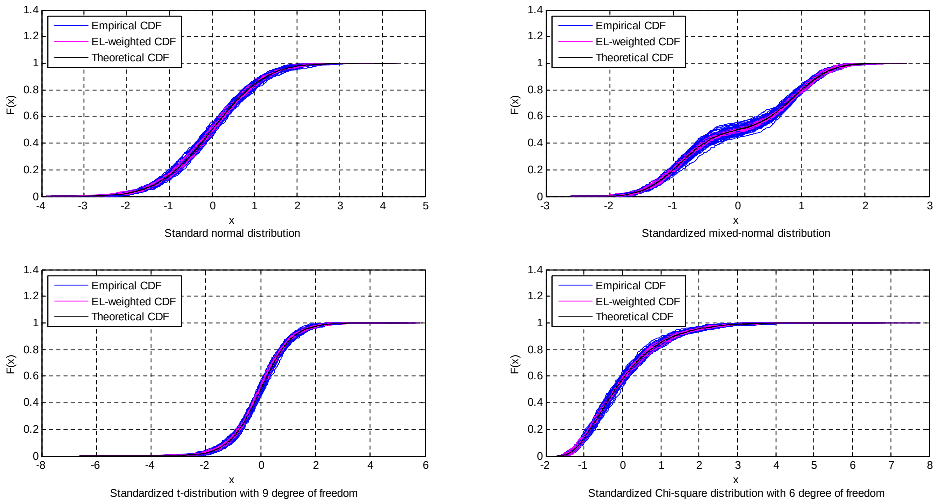
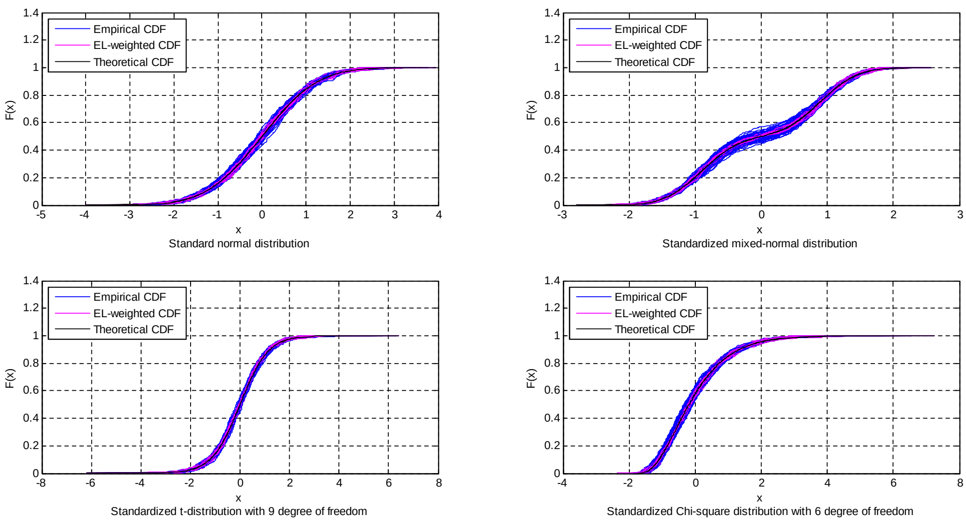


Figure 4: Overlay of two estimates using GARCH errors,"Empirical CDF" v.s. "EL-weighted CDF"



3.6 Empirical Work

Finally, we investigate whether our new proposed conditional VaR and ES method have good forecasting ability by comparing them with other conventional methods using both index and company data.

VaR, summarizing in a single number all the risk of a portfolio, is the most widely used risk measure in financial industry, even though it violates the subadditivity axiom of Artzner et. al. (1999) when the tails are super fat and it can not capture the risk of extreme movements on the tails. A number of alternative risk measures have been proposed to overcome the problem of lack of subadditivity in the VaR and also provide more information about the tail shape. Expected shortfall is one of the most popular alternative risk measures whose torch has recently been taken up by the Basel Committee on Banking Supervision.

3.6.1 Descriptive Statistics

The four datasets that we use are S&P 500(01/01/2000−31/12/2012,CRSP), MSCIworld (01/01/1970−29/01/2013, Datastream), MSCI Emerging Market (01/01/1988−29/01/2013, Datastream) and Microsoft Corporate (14/03/1986 − 31/12/2012, CRSP). Table 20 gives the descriptive statistics for these four datasets, the Ljung-Box test for autocorrelation and the KS test for normality.

Table 20. data summary statistics

	S&P 500	**MSCIworld**	**MSCIEM**	**MSFT**
Mean	1.5961e-04	2.4420e-04	3.6224e-04	0.0011
Standard deviation	0.0135	0.0087	0.0117	0.0227
Min	-0.0900	-0.1036	-0.0999	-0.3012
Max	0.1151	0.0910	0.1007	0.1957
Skewness	0.0364	-0.4591	-0.5726	-0.1171
Kurtosis	10.3762	14.1965	10.7623	13.1984

3.6.2 Backtesting VaR

Violation Ratio

The main purpose of the empirical study is to see how our model performs in forecasting risk. This can be done by backtesting various VaR and ES methods. Backtesting evaluates VaR forecasts by checking how a VaR forecast model perform over a certain period of time. The number of the observations that are used to forecast the risk is called the estimation window, W_E and the data sample over which risk is forecast is called the testing window, W_T. In our empirical study, we choose 1000 observations as our estimation window. (Figure 9,10 and 11). We later use a technique called violation ratio (VR) to judge the quality of the VaR forecasts. If the actual return on a particular day exceeds the VaR forecast, we say that the VaR limit is being violated. The VR is defined by the observed number of violations over the expected number of violations.

$$VR = (observed\ no\ of\ violation\ /\ expected\ no\ of\ violation)$$

If the VaR forecast of our model is accurate, the violation ratio is expected to be equal to 1. A useful rule of thumb is that if the VR is between 0.8 and 1.2, the model is considered to be a good forecast. If VR<0.8 means that the model underestimate risk while if VR>1.2 means that the model overestimate risk.

Backtesting fundamental models

First of all, we investigate the backtest performance of five fundamental models, including EWMA,MA, HS,GARCH(1,1), and our model GARCH-ELW with the significance level equals 0.95 and 0.99. The common assumption of EWMA, MA and GARCH(1,1) is that the standardized residuals are normally distributed with mean 0 and variance 1. HS is a nonparametric method and the new model that we proposed, GARCH-ELW, relax the distribution assumption of the standardized residual.

Model 1 (MA): One of the simplest volatility forecast methods is the moving average (MA) method, which puts equal weight on all the past observations. The conditional variance process is

$$h_{t+1} = \frac{1}{W_E}\sum_{i=1}^{W_E} u_{t-i+1}^2,$$

where W_E is the length of the estimation window). Hence, the conditional Value-at-Risk of return given $\mathcal{F}_{t-1}$ is,

$$\xi_{t+1}(\alpha) = \sum_{j=1}^{p} \rho_j y_{t+1-j} + h_{t+1}^{1/2} z_\alpha,$$

where $z_\alpha = \Phi^{-1}(\alpha)$ is the standard normal quantile. The model is very simple but the equal weighted assumption is not realistic as a model of volatility.

Model 2 (EWMA): The basic structure of the conditional variance process is a restricted IGARCH(1,1)

$$h_{t+1} = (1-\lambda)u_t^2 + \lambda h_t.$$

The conditional Value-at-Risk of return series given $\mathcal{F}_{t-1}$ is

$$\xi_{t+1}(\alpha) = \sum_{j=1}^{p} \rho_j y_{t+1-j} + h_{t+1}^{1/2} z_\alpha.$$

An EWMA is similar to MA although the EWMA places relatively more weight on recent observations than on observation in the distant past. The attractiveness of the RiskMetrics model is that there is no parameter to be estimated, λ is fixed at 0.94 for daily data and 0.97 for monthly data and it's easy to extend to multivariate setting. However, the disadvantage is also that the parameters are not estimated and the model process collapses to zero eventually.

Model 3 (HS): Historical Simulation is a nonparametric method based on the assumption that the history will happen again.

Model 4 (GARCH-N(1,1)): Fully parametric methods provide a natural method to compute VaR. We take GARCH(1,1) for simplicity.

$$h_{t+1} = \omega + \beta h_t + \gamma u_t^2$$

$$\xi_{t+1}(\alpha) = \sum_{j=1}^{p} \rho_j y_{t+1-j} + h_{t+1}^{1/2} z_\alpha.$$

Model 5 (GARCH-ELW): Under the model specification, the conditional variance is modelled by GARCH(1,1) and the conditional quantile is estimated by the empirical likelihood. The conditional Value-at-Risk of return series given $\mathcal{F}_{t-1}$ is,

$$\xi_t(\alpha) = \sum_{j=1}^{p} \rho_j y_{t-j} + h_t^{1/2} q_\alpha.$$

The probability of losses exceeding VaR, α, must be specified, with the most common probability level being 1% and 5%. Test results are in Table 21 and 22, and backtest VaR plot are provided in Figure 5 and 6 (for S&P 500 data only, others are provided upon request), with α equals 1% and 5% respectively. Overall, the models do much better when we choose 0.95 significant level, which probably shows that most of the fundamental and simple models are inadequate in forecasting risk in the extreme tail, such as in the 0.99 significant level case. However, our model is the best candidate among all in the extreme case. Interestingly, the models represent better forecasting ability on the individual stock data than on the index data. The worst performance of all the models happen when using the S&P 500 data.

Table 21: backtesting VaR (MA, EWMA, HS, GARCH(1,1) and GARCH-ELW) ($\alpha = 0.01$)

$\alpha = 0.01$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
MA	2.8647	1.9821	2.0025	1.3546
EWMA	2.3358	1.7585	2.1108	1.3199
HS	1.8510	1.3519	1.1546	1.3025
GARCH-N(1,1)	2.1155	1.5247	1.6598	1.1289
GARCH-EL	1.4103	1.1079	0.7938	0.9726

Table 22: backtesting VaR (MA, EWMA, HS, GARCH(1,1) and GARCH-ELW) ($\alpha = 0.05$)

$\alpha = 0.05$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
MA	1.0489	0.9209	0.9706	0.7016
EWMA	1.2252	1.0632	1.1077	0.8163
HS	1.0930	1.0876	1.0319	1.0247
GARCH-N(1,1)	1.0930	0.9961	1.0211	0.7746
GARCH-ELW	1.0137	0.9473	0.8840	0.9691

Figure 5: Backtesting VaR (fundamental model, $\alpha = 0.01$)

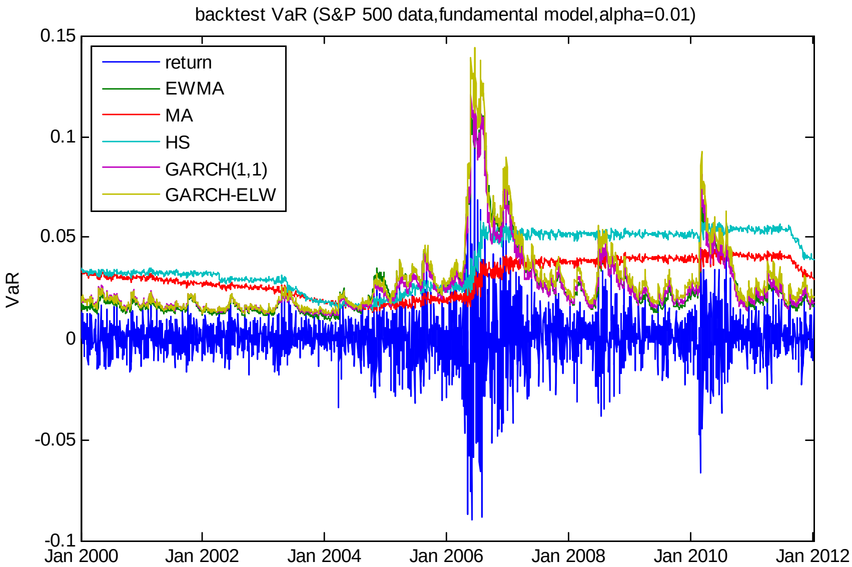
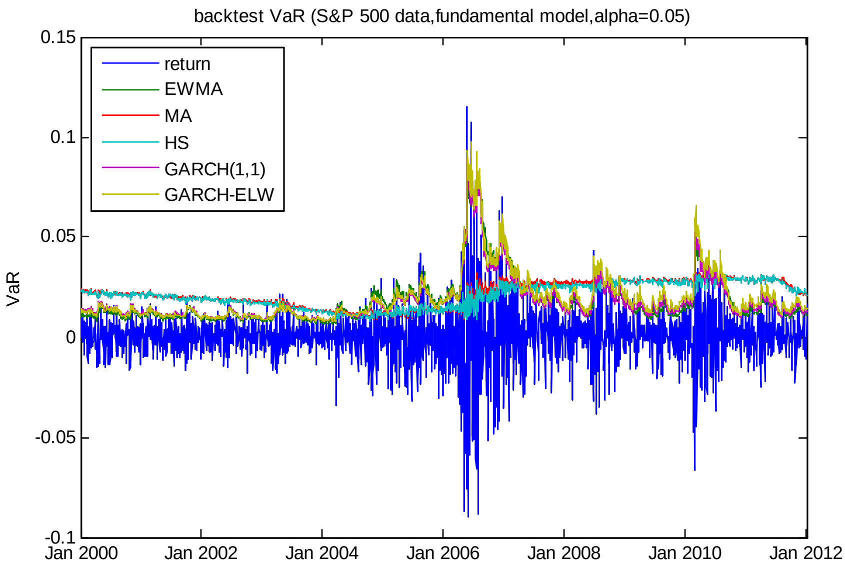


Figure 6: Backtest VaR (fundamental model, $\alpha = 0.05$)



Backtesting extended models

In the financial literature, it is often found that positive and negative shocks to returns have different impact on conditional volatility. Several extensions of the GARCH model aim at accommodating the asymmetry in the response. These include the GJR-GARCH model of Glosten, Jagannathan and Runkle (1993), EGARCH model of Nelson (1991) and the asymmetric GARCH model of Engle and Ng (1993), all documenting that large positive and negative unexpected shock lead to an increase of the conditional volatility, although the negative innovation with the similar magnitude lead to larger increase.

The GJR model is a GARCH variant that includes leverage terms for modeling asymmetric volatility clustering. In the GJR formulation, large negative changes are more likely to be clustered than positive changes.

Hence, we further examine if adding in asymmetric information both in the conditional variance process and the standardized residual help to improve risk forecasting. The models that we use in the analysis are GARCH-N, GARCH-T,GARCH-ELW,GARCH-GJR-N, GARCH-GJR-T and GARCH-GJR-ELW. All the conditional models are GARCH(1,1) and GARCH (1,1)-GJR with N, T and ELW representing normal, student-t and no specific distribution assumptions of the innovations. The conditional variance process for GARCH(1,1)-GJR is,

$$h_{t+1} = \omega + \beta h_t + \gamma u_t^2 + \xi I(u_t < 0) u_t^2$$

The leverage coefficients are applied to negative shocks which give the negative shock more weight in the conditional variance process than a similar positive shock. Similarly, test results are provided in Table 23 and 24, while Figure 7 and 8 are graphic plot of the backtesting performance.

We make some further comments on Tables 23 and 24. First of all, the results are similar as when testing the fundamental models in that the models do better with $\alpha = 0.05$. Secondly, the semiparametric methods that we proposed (both the GARCH-ELW and GARCH-GJR-ELW) seems to be the best models in both cases, although GARCH-N and GARCH-GJR-N are doing well when we choose a smaller significant level. Furthermore, adding the asymmetric term in the conditional variance does not have obvious impact on improving forecast. The performance of GARCH-ELW and GARCH-GJR-ELW are quite similar. Finally, using the normal distributed standardized residuals normally underestimate risk while the fat-tail student-t distribution generally overestimate risk, especially in the case when $\alpha = 0.01$, which is consistent with the conventional literature.

Table 23: backtesting VaR (GARCH-N, GARCH-T,GARCH-ELW,GARCH-GJR-N, GARCH-GJR-T and GARCH-GJR-ELW) ($\alpha = 0.01$)

$\alpha = 0.01$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
GARCH-N	2.2036	1.5247	1.6598	1.1289
GARCH-T	0.7933	0.6607	0.6855	0.3473
GARCH-ELW	1.6307	1.1079	0.7938	0.9726
GARCH-GJR-N	2.2036	1.4840	1.7319	1.1636
GARCH-GJR-T	0.8814	0.7624	0.6495	0.3994
GARCH-GJR-ELW	1.4544	1.1994	0.9562	0.9552

Table 24: backtesting VaR (GARCH-N, GARCH-T,GARCH-ELW,GARCH-GJR-N, GARCH-GJR-T and GARCH-GJR-ELW) ($\alpha = 0.05$)

$\alpha = 0.05$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
GARCH-N	1.0930	0.9961	1.0211	0.7746
GARCH-T	0.8814	0.7298	0.6855	0.4203
GARCH-ELW	1.0137	0.9473	0.8840	0.9691
GARCH-GJR-N	1.1635	1.0510	1.0139	0.7850
GARCH-GJR-T	0.9784	0.7684	0.7252	0.4272
GARCH-GJR-ELW	1.0489	1.0144	0.9165	0.9726

Figure 7:Backtesting VaR (extended model, $\alpha = 0.01$)

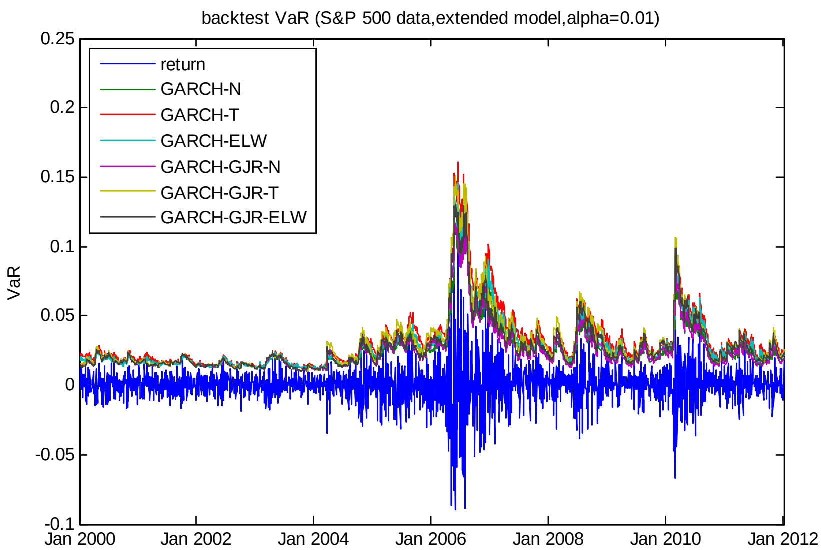
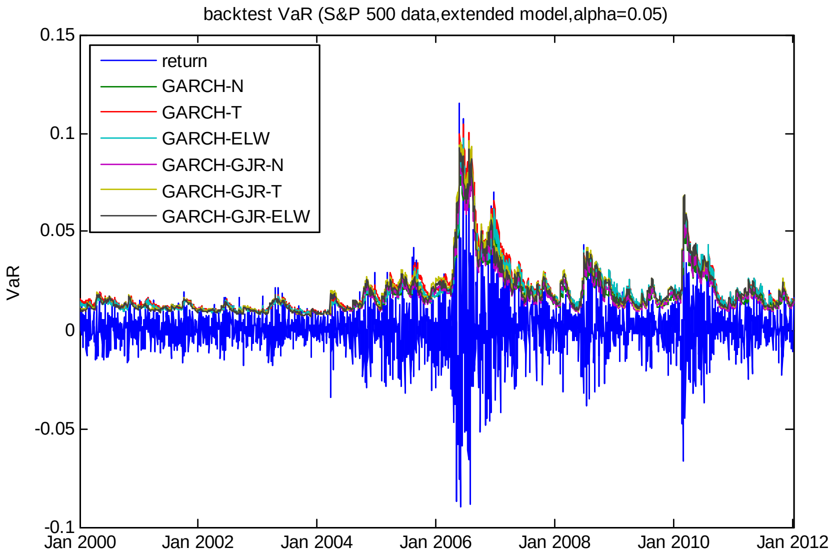


Figure 8: Backtesting VaR (extended model, $\alpha = 0.05$)



3.6.3 Backtesting ES

Average normalized shortfall

It's harder to backtest the Expected shortfall as we are testing an expectation rather than a quantile. Fortunately, average normalized shortfall, a methodology which is analogous to the use of violation ratios for VaR is able to help us backtest ES. The test procedure is sketched below. For days when VaR is violated, normalized shortfall (NS) is calculated as

$$NS_t = \frac{y_t}{ES_t},$$

where ES_t is the observed ES on day t. From the definition of ES, the expected y_t— given VaR is violated — is:

$$\frac{E(y_t \mid y_t < -VaR_t)}{ES_t} = 1.$$

The Null hypothesis: average NS ($\overline{NS}$)= 1. The test result is reported in Table 27. The average NS in our model is 0.9895, which is the nearest to 1.

Backtesting fundamental models

We backtest ES by using the methodology provided in the previous paragraph. The backtesting ES results are showing in Table 25 and 26 and Figure 9 and 10. From the tables, we are confident to say that our proposed model is the best in terms of the NS ratio, while the difference between different models and datasets are not so distinguished as in the case of backtesting VaR.

Table 25:backtesting ES (MA, EWMA, HS, GARCH(1,1) and GARCH-ELW) ($\alpha = 0.01$)

$\alpha = 0.01$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
MA	1.3849	1.5703	1.5571	1.2494
EWMA	1.0730	1.1488	1.2390	1.1917
HS	1.1408	1.1565	1.1639	0.9952
GARCH(1,1)	1.0735	1.1676	1.2798	1.1986
GARCH-ELW	0.9983	1.0703	1.1813	0.9933

Table 26: backtesting ES (MA, EWMA, HS, GARCH(1,1) and GARCH-ELW) ($\alpha = 0.05$)

$\alpha = 0.05$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
MA	1.3949	1.2463	1.5077	1.1947
EWMA	1.1101	1.1364	1.2698	1.1182
HS	1.1622	1.0494	1.1767	1.0280
GARCH(1,1)	1.1182	1.1331	1.2190	1.1012
GARCH-ELW	1.0385	1.0581	1.1034	0.9884

Figure 9: Backtesting ES (fundamental model, $\alpha = 0.01$)

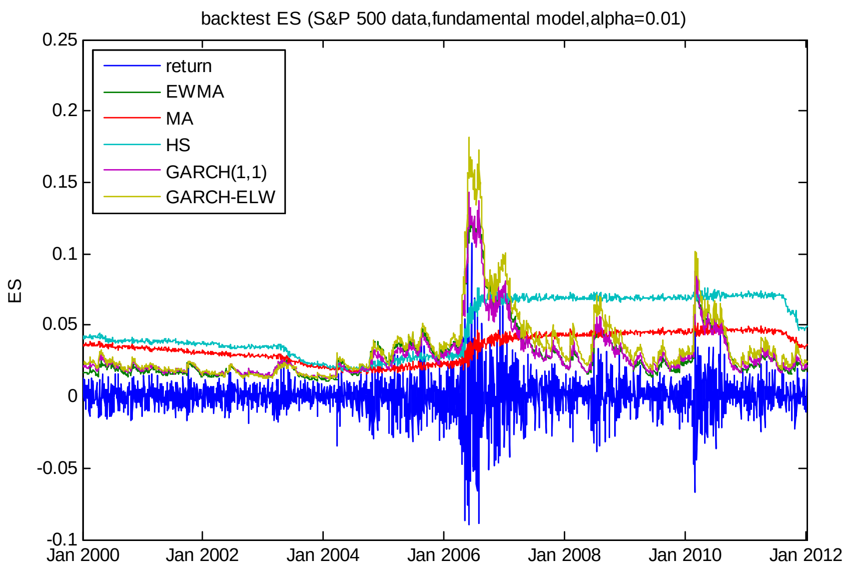
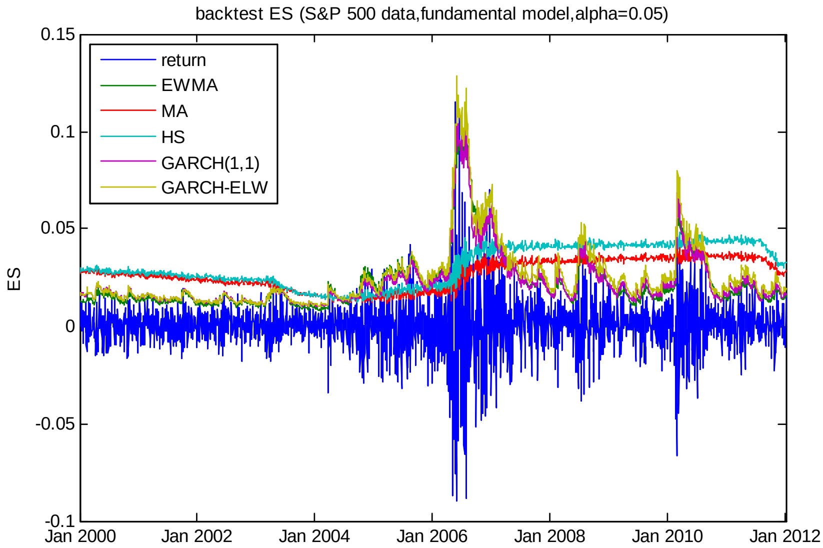


Figure 10:Backtesting ES (fundamental model, $\alpha = 0.05$)



Backtesting extended models

Similar findings in the extended cases, where we show the backtest ES results in Table 27 and 28 and Figure 11 and 12, that the difference in the ES backtesting regarding to different models and datasets are not as big as backtesting VaR. All the models have good forecasting ability according to the above tables so it's very hard to choose a better one among these models.

Table 27: backtesting ES (GARCH-N, GARCH-T,GARCH-ELW,GARCH-GJR-N, GARCH-GJR-T and GARCH-GJR-ELW) ($\alpha = 0.01$)

$\alpha = 0.01$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
GARCH-N	1.0648	1.1676	1.2798	1.1986
GARCH-T	1.0103	1.1008	1.1275	1.0967
GARCH-ELW	0.9827	1.0703	1.1813	0.9933
GARCH-GJR-N	1.0668	1.1393	1.1689	1.1943
GARCH-GJR-T	1.0273	1.0584	1.0986	1.0626
GARCH-GJR-ELW	0.9816	1.0349	1.0965	0.9969

Table 28: backtesting ES (GARCH-N, GARCH-T,GARCH-ELW,GARCH-GJR-N, GARCH-GJR-T and GARCH-GJR-ELW) ($\alpha = 0.05$)

$\alpha = 0.05$	**S&P 500**	**MSCI world**	**MSCI EM**	**MSFT**
GARCH-N	1.1182	1.1331	1.2190	1.1012
GARCH-T	0.9902	1.0243	1.0979	0.9913
GARCH-ELW	1.0385	1.0581	1.1034	0.9884
GARCH-GJR-N	1.1083	1.0857	1.1557	1.0988
GARCH-GJR-T	1.0032	1.0036	1.0447	1.0027
GARCH-GJR-ELW	1.0424	1.0252	1.0604	0.9883

Figure 11: Backtesting ES (extended model, $\alpha = 0.01$)

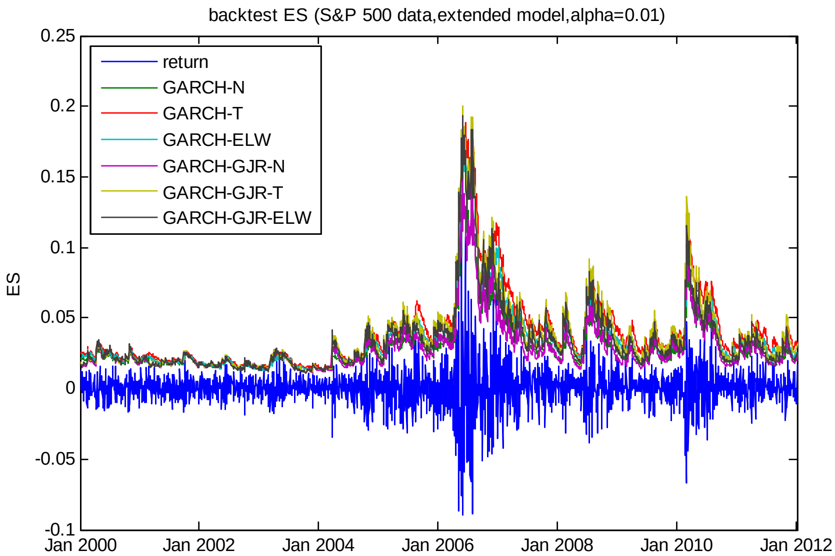
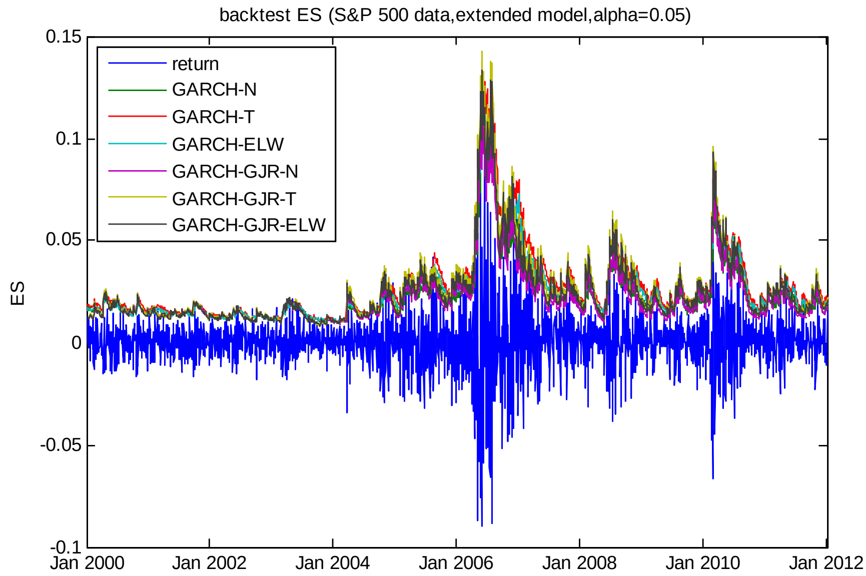


Figure 12: Backtesting ES (extended model, $\alpha = 0.05$)



3.7 Conclusion and Extension

This paper proposes and investigates new efficient conditional VaR and ES estimators in a semiparametric GARCH model. These proposed estimators for risk measures fully exploit the moment information which has been previously ignored in constructing innovation distribution estimators. We show they can achieve large efficiency improvement and quantify this magnitude in Monte Carlo simulations. At the same time, we present the asymptotic theory for one period ahead VaR and ES forecasts. The theory can be used as guidance as to constructing confidence intervals for point risk measure forecasts.

Even though we consider a simple GARCH(1,1) model in this paper, the efficient estimation method for both variance parameters and error quantile can be used for more complicated parametric volatility models. For example, one could consider GARCH with leverage effects or GARCH in mean models. Although the efficiency gain hinges on the efficiency of volatility estimators in theory, our MonteCarlo experiments show that this impact on efficiency improvement is quantitatively small.

Sometimes unconditional Value-at-Risk is also of interest to risk managers. Then the question in the current GARCH(1,1) context is whether we have efficiency gains from integrating the conditional VaR versus unconditional. This question is to be addressed in a separate paper.

3.8 Appendix

3.8.1 Proofs

Proof of Theorem 1. Given θ and the observations $\{h_0, y_1, \ldots, y_n\}$, then the log likelihood is

$$\mathcal{L}(\theta) = \sum_{t=1}^{n}[\log f(c^{-1}g_t^{-1/2}(\beta)y_t) + \log c^{-1}g_t^{-1/2}(\beta)].$$

Now we can write the conditional score at time t as

$$l_t(\theta) = -(1+\varepsilon_t(\theta)\frac{f'(\varepsilon_t(\theta))}{f(\varepsilon_t(\theta))})\begin{pmatrix} \frac{1}{2g_t(\beta)}\frac{\partial g_t(\beta)}{\partial \beta} \\ \frac{1}{c} \end{pmatrix}.$$

Then, according to Drost and Klaassen (1997), the efficient score and information matrix for β are

$$\begin{aligned} l_{1t}^*(\beta_0) &= -\frac{1}{2}\{G_t(\beta_0) - G(\beta_0)\}(1+\varepsilon_t\frac{f'(\varepsilon_t)}{f(\varepsilon_t)}) \\ E[l_{1t}^*(\beta_0)l_{1t}^*(\beta_0)^\top] &= \frac{E[R_3(\varepsilon)^2]}{4}\{G_2(\beta_0) - G(\beta_0)G(\beta_0)^\top\}, \end{aligned}$$

and

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} E[l_{1t}^*(\beta_0)l_{1t}^*(\beta_0)^\top]^{-1}l_{1t}^*(\beta_0) + o_p(1).$$

Next, as the efficient estimator for c is $\widehat{c} = \sqrt{\frac{1}{n}\sum_{t=1}^n \widehat{e}_t^2 - \frac{1}{n}\frac{\sum_{t=1}^n \widehat{e}_t^3}{\sum_{t=1}^n \widehat{e}_t^2}\sum_{t=1}^n \widehat{e}_t}$. Using the delta method, we can see

$$\begin{aligned} \widehat{e}_t - e_t &= y_t[g_t^{-1/2}(\widehat{\beta}) - g_t^{-1/2}(\beta)] = -\frac{1}{2}\frac{e_t}{g_t(\beta_0)}\frac{\partial g_t(\beta_0)}{\partial \beta^\top}(\widehat{\beta} - \beta_0) + o_p(\frac{1}{n}) \\ \widehat{e}_t^2 - e_t^2 &= y_t^2[g_t^{-1}(\widehat{\beta}) - g_t^{-1}(\beta)] = -\frac{e_t^2}{g_t(\beta_0)}\frac{\partial g_t(\beta_0)}{\partial \beta^\top}(\widehat{\beta} - \beta_0) + o_p(\frac{1}{n}), \end{aligned}$$

consequently,

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^n \widehat{e}_t - \frac{1}{n}\sum_{t=1}^n e_t &= -\frac{1}{2}E[\frac{e_t}{g_t(\beta_0)}\frac{\partial g_t(\beta_0)}{\partial \beta^\top}](\widehat{\beta} - \beta_0) + o_p(n^{-1/2}) \\ \frac{1}{n}\sum_{t=1}^n \widehat{e}_t^2 - \frac{1}{n}\sum_{t=1}^n e_t^2 &= -E[\frac{e_t^2}{g_t(\beta_0)}\frac{\partial g_t(\beta_0)}{\partial \beta^\top}](\widehat{\beta} - \beta_0) + o_p(n^{-1/2}), \end{aligned}$$

as a result, by LLN and Ergodic Theorem,

$$\sqrt{n}(\widehat{c} - c_0)$$

$$
\begin{aligned}
&= \sqrt{n}(\sqrt{\frac{1}{n}\sum\nolimits_{t=1}^{n}\widehat{e}_t^2 - \frac{1}{n}\frac{\sum_{t=1}^{n}\widehat{e}_t^3}{\sum_{t=1}^{n}\widehat{e}_t^2}\sum\nolimits_{t=1}^{n}\widehat{e}_t} - \sqrt{\frac{1}{n}\sum\nolimits_{t=1}^{n}c_0^2\varepsilon_t^2} + \sqrt{\frac{1}{n}\sum\nolimits_{t=1}^{n}c_0^2\varepsilon_t^2} - c_0) \\
&= \frac{1}{2c_0}\{-c_0^2 G^\top\sqrt{n}(\widehat{\beta}-\beta_0) - c_0^2\frac{E\varepsilon_t^3}{E\varepsilon_t^2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t + c_0^2\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\varepsilon_t^2-1)\} + o_p(1) \\
&= \frac{c_0}{2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{(\varepsilon_t^2-1) - \varepsilon_t E\varepsilon^3 - G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1} l_{1t}^*\} + o_p(1).
\end{aligned}
$$

Since $E[(\varepsilon_t^2-1)l_{1t}^*] = 0$ and $E[\varepsilon_t l_{1t}^*] = 0$, we have

$$
\Omega_c = \frac{c_0^2}{4}\{E\varepsilon^4 - 1 - (E\varepsilon^3)^2 + G^\top E[l_{1t}^* {l_{1t}^*}^\top]^{-1} G\}.
$$

We can thus conclude that

$$
\begin{aligned}
&\sqrt{n}\begin{pmatrix}\widehat{\beta}-\beta_0 \\ \widehat{c}-c_0\end{pmatrix} \\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{pmatrix} -\frac{1}{2}E[l_{1t}^* l_{1t}^{*\top}]^{-1}\{G_t - G\} & 0 \\ \frac{c_0}{4}G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1}\{G_t - G\} & \frac{c_0}{2}(-E\varepsilon^3, 1)\end{pmatrix}\begin{pmatrix} R_3(\varepsilon_t) \\ R_2(\varepsilon_t)\end{pmatrix} + o_p(1)
\end{aligned}
$$

and

$$
\Omega_\theta = \begin{pmatrix} E[l_{1t}^* l_{1t}^{*\top}]^{-1} & -\frac{c_0}{2}E[l_{1t}^* l_{1t}^{*\top}]^{-1}G \\ -\frac{c_0}{2}G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1} & \frac{c_0^2}{4}\{E\varepsilon^4 - 1 - (E\varepsilon^3)^2 + G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1}G\}\end{pmatrix}.
$$

■

Lemma 1. *Suppose that assumptions A.2-A.4 hold. Then $\widehat{F}_N(x)$ and $\widehat{F}_w(x)$ have the following expansion:*

$$
\begin{aligned}
\sup_{x\in\mathbb{R}}\left|\widehat{F}_N(x) - F(x) - \frac{1}{n}\sum_{t=1}^{n}\{1(\varepsilon_t \le x) - F(x)\} - f(x)\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t - \frac{xf(x)}{2}\frac{1}{n}\sum_{t=1}^{n}(\varepsilon_t^2-1)\right| &= o_p(n^{-1/2}) \\
\sup_{x\in\mathbb{R}}\left|\widehat{F}_w(x) - F(x) - \frac{1}{n}\sum_{t=1}^{n}\{1(\varepsilon_t \le x) - F(x)\} + \frac{1}{n}\sum_{t=1}^{n}A_x^\top B^{-1}R_2(\varepsilon_t)\right| &= o_p(n^{-1/2}).
\end{aligned}
$$

Consequently, the process $\sqrt{n}(\widehat{F}_N - F)$ converges weakly to a zero-mean Gaussian process $\mathcal{Z}_N$ with covariance function Ω_N and the process $\sqrt{n}(\widehat{F}_w - F)$ converges weakly to a zero-mean Gaussian process $\mathcal{Z}_w$ with covariance function Ω_w, where:

$$
\begin{aligned}
\Omega_N(x,x') &= \operatorname{cov}(\mathcal{Z}_N(x), \mathcal{Z}_N(x')) \\
&= E\Big[[1(\varepsilon \le x) - F(x) + f(x)\varepsilon + \frac{xf(x)}{2}(\varepsilon^2-1)] \\
&\quad \times [1(\varepsilon \le x') - F(x') + f(x')\varepsilon + \frac{x'f(x')}{2}(\varepsilon^2-1)]\Big]
\end{aligned}
$$

$$\begin{aligned}\Omega_w(x,x') &= \operatorname{cov}\left(\mathcal{Z}_w(x),\mathcal{Z}_w(x')\right)\\ &= E\left[[1(\varepsilon\le x)-F(x)-A_x^\top B^{-1}R_2(\varepsilon)][1(\varepsilon\le x')-F(x')-A_{x'}^\top B^{-1}R_2(\varepsilon)]\right].\end{aligned}$$

Where we define the following quantities:

$$A_x = E[R_2(\varepsilon)1(\varepsilon\le x)] \quad ; \quad B = E[R_2(\varepsilon)R_2(\varepsilon)^\top];$$

$$C_x = f(x)^2\left\{\frac{E[\varepsilon^4]-1}{4}x^2 + xE[\varepsilon^3]+1\right\} + f(x)\left\{2E[\varepsilon 1(\varepsilon\le x)] + xE[(\varepsilon^2-1)1(\varepsilon\le x)]\right\},$$

Proof of Lemma 1. We follow the proof of Theorem 4.1 in Koul and Ling (2006) closely. Define the empirical process

$$\nu_n(x,z_1,z_2) = \frac{1}{\sqrt{n}}\sum_{t=1}^n\{1(\varepsilon_t\le z_1+xz_2) - E[1(\varepsilon_t\le z_1+xz_2)]\}.$$

For any $z=(z_1,z_2)\in\mathbb{R}^2$, let $|z|=|z_1|\vee|z_2|$. In $\mathbb{R}^2$, we define a pseudo-metric

$$d_c(x,y) = \sup_{|z|\le c}|F(x(1+z_1)+z_2) - F(y(1+z_1)+z_2)|^{1/2}, (x,y)\in\mathbb{R}^2, c>0.$$

Let $\mathcal{N}(\delta,c)$ be the cardinality of the minimal δ-net and let

$$\mathcal{I}(c) = \int_0^1\{\ln\mathcal{N}(u,c)\}^{1/2}du$$

According to Theorem 4.1 in Koul and Ling (2006), assumptions imply that $\mathcal{I}(c)<\infty$ for any $c\in[0,1)$. This combines with Koul and Ossiander (1994) show that the following stochastic equicontinuity condition holds:

$$\sup_{x\in\mathbb{R},|z_1|\le Cn^{-1/2},|z_2-1|\le Cn^{-1/2}}|\nu_n(x,z_1,z_2)-\nu_n(x,0,1)| = o_p(1).$$

As a result,

$$\nu_n(x,z_1,z_2) = \nu_n(x,0,1)+\nu_n(x,z_1,z_2)-\nu_n(x,0,1) = \nu_n(x,0,1)+o_p(1).$$

By LLN, we know that

$$\frac{1}{n}\sum_{t=1}^n\varepsilon_t = o_p(1), \quad \sqrt{\frac{1}{n}\sum_{t=1}^n\varepsilon_t^2 - (\frac{1}{n}\sum\nolimits_{t=1}^n\varepsilon_t)^2} = o_p(1).$$

Therefore, $\widehat{F}_N(x)$ can be expanded as, uniformly in $x \in \mathbb{R}$,

$$
\begin{aligned}
& \sqrt{n}(\widehat{F}_N(x) - F(x)) \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left\{ 1\left(\frac{\varepsilon_t - \widehat{\mu}_\varepsilon}{\widehat{\sigma}_\varepsilon} \le x \right) - F(x) \right\} \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{ 1\left(\varepsilon_t \le \widehat{\mu}_\varepsilon + x\widehat{\sigma}_\varepsilon \right) - F\left(\widehat{\mu}_\varepsilon + x\widehat{\sigma}_\varepsilon \right) \} + \sqrt{n}\{F(\widehat{\mu}_\varepsilon + x\widehat{\sigma}_\varepsilon) - F(x)\} \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} + o_p(1) + \sqrt{n}\{F(\widehat{\mu}_\varepsilon + x\widehat{\sigma}_\varepsilon) - F(x)\} \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} + f(x)\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t + \sqrt{n}xf(x) \left(\sqrt{\frac{1}{n}\sum_{t=1}^{n} \varepsilon_t^2 - (\widehat{\mu}_\varepsilon)^2} - 1 \right) + o_p(1) \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} + f(x)\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t + \sqrt{n}\frac{xf(x)}{2} \left(\frac{1}{n}\sum_{t=1}^{n} (\varepsilon_t^2 - 1) - (\widehat{\mu}_\varepsilon)^2 \right) + o_p(1) \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} + f(x)\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t + \frac{xf(x)}{2}\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\varepsilon_t^2 - 1) + o_p(1).
\end{aligned}
$$

We know from Owen (2001) that

$$
\widehat{w}_t = \frac{1}{n}\frac{1}{1+\lambda_n' R_2(\varepsilon_t)}; \lambda_n = B^{-1}(\frac{1}{n}\sum_{t=1}^{n} R_2(\varepsilon_t)) + o_p(n^{-1/2}).
$$

Consequently, uniformly in $x \in \mathbb{R}$,

$$
\begin{aligned}
& \sqrt{n}(\widehat{F}_w(x) - F(x)) \\
= & \sqrt{n}(\sum_{t=1}^{n} \widehat{w}_t 1(\varepsilon_t \le x) - \frac{1}{n}\sum_{t=1}^{n} 1(\varepsilon_t \le x) + \frac{1}{n}\sum_{t=1}^{n} 1(\varepsilon_t \le x) - F(x)) \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{[n\widehat{w}_t - 1]1(\varepsilon_t \le x)\} + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} \\
= & -\lambda_n^\top \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{R_2(\varepsilon_t)1(\varepsilon_t \le x)\} + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} + o_p(n^{-1/2}) \\
= & -\frac{1}{\sqrt{n}} \sum_{t=1}^{n} R_2(\varepsilon_t)^\top B^{-1} \frac{1}{n} \sum_{t=1}^{n} \{R_2(\varepsilon_t)1(\varepsilon_t \le x)\} + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} + o_p(n^{-1/2}) \\
= & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{1(\varepsilon_t \le x) - F(x)\} - \frac{1}{\sqrt{n}} \sum_{t=1}^{n} A_x B^{-1} R_2(\varepsilon_t) + o_p(n^{-1/2}),
\end{aligned}
$$

where the last equality holds because of ergodic theorem: $n^{-1} \sum_{t=1}^{n} \{R_2(\varepsilon_t)1(\varepsilon_t \le x)\} = A_x + o_p(1)$.

■

COROLLARY 1. *Denote* $E[\varepsilon 1(\varepsilon \le x)] = a_1(x)$ *and* $E[(\varepsilon^2 - 1)1(\varepsilon \le x)] = a_2(x)$. $\widehat{F}(x)$ *is* *asymptotically less efficient than* $\widehat{F}_w(x)$*, and* $\widehat{F}(x)$ *achieves the efficiency bound iff*

$$a_1(x) = a_2(x) = 0.$$

$\widehat{F}_N(x)$ *is asymptotically less efficient than* $\widehat{F}_w(x)$ *as* $C_x \geq -A_x^\top B^{-1} A_x$*.* $\widehat{F}_N(x)$ *achieves the efficiency bound iff*

$$x = \frac{2(a_2(x) - a_1(x)E[\varepsilon^3])}{a_1(x)(E[\varepsilon^4]-1) - a_2(x)E[\varepsilon^3]} \quad ; \quad f(x) = -\frac{2a_1(x) + xa_2(x)}{\frac{E[\varepsilon^4]-1}{2}x^2 + 2xE[\varepsilon^3] + 2}.$$

Proof of Corollary 1. Notice that

$$A_x^\top B^{-1} A_x = \frac{\{E[\varepsilon^4]-1\}\{a_1(x) - a_2(x)\frac{E[\varepsilon^3]}{E[\varepsilon^4]-1}\}^2}{E[\varepsilon^4] - 1 - E[\varepsilon^3]^2} + \frac{a_2^2(x)}{E[\varepsilon^4]-1},$$

and under the moment condition (2), $E[\varepsilon^4] - 1 = \mathrm{Var}(\varepsilon^2) \geq 0$ and

$$\begin{aligned} E[\varepsilon^4] - 1 - E[\varepsilon^3]^2 &= \{E[\varepsilon^4]-1\}\{1 - \frac{E[\varepsilon^3]^2}{E[\varepsilon^4]-1}\} \\ &= \{E[\varepsilon^4]-1\}\{1 - \mathrm{corr}(\varepsilon, \varepsilon^2)^2\} \\ &\geq 0 \end{aligned}$$

so $A_x^\top B^{-1} A_x \geq 0$ and $A_x^\top B^{-1} A_x = 0 \Leftrightarrow a_1(x) = a_2(x) = 0$. As for the asymptotical efficiency comparison between $\widehat{F}_N(x)$ and $\widehat{F}_w(x)$, we have

$$\begin{aligned} & C_x + A_x^\top B^{-1} A_x \\ = \; & f(x)^2\{\frac{E[\varepsilon^4]-1}{4}x^2 + xE[\varepsilon^3] + 1\} + f(x)\{2a_1(x) + xa_2(x)\} \\ & + \frac{\{E[\varepsilon^4]-1\}\{a_1(x) - a_2(x)\frac{E[\varepsilon^3]}{E[\varepsilon^4]-1}\}^2}{E[\varepsilon^4]-1-E[\varepsilon^3]^2} + \frac{a_2^2(x)}{E[\varepsilon^4]-1} \\ = \; & f(x)^2\{\frac{E[\varepsilon^4]-1}{4}x^2 + xE[\varepsilon^3] + 1\} + f(x)\{2a_1(x) + xa_2(x)\} \\ & + \frac{\{E[\varepsilon^4]-1\}a_1(x) - 2a_1(x)a_2(x)E[\varepsilon^3] + a_2^2(x)}{E[\varepsilon^4]-1-E[\varepsilon^3]^2} \\ = \; & \{\frac{E[\varepsilon^4]-1}{4}x^2 + xE[\varepsilon^3] + 1\}\{f(x) + \frac{2a_1(x) + xa_2(x)}{\frac{E[\varepsilon^4]-1}{2}x^2 + 2xE[\varepsilon^3] + 2}\}^2 \\ & + \frac{\{x[a_1(x)(E[\varepsilon^4]-1) - a_2(x)E[\varepsilon^3]] + 2(a_1(x)E[\varepsilon^3] - a_2(x))\}^2}{4\{E[\varepsilon^4]-1-E[\varepsilon^3]^2\}\{\frac{E[\varepsilon^4]-1}{4}x^2 + xE[\varepsilon^3] + 1\}}, \end{aligned}$$

additionally

$$\frac{E[\varepsilon^4]-1}{4}x^2 + xE[\varepsilon^3] + 1 = \frac{E[\varepsilon^4]-1}{4}\{x + \frac{2E[\varepsilon^3]}{E[\varepsilon^4]-1}\}^2 + \frac{E[\varepsilon^4]-1-E[\varepsilon^3]^2}{E[\varepsilon^4]-1} \geq 0,$$

so we can conclude $C_x \geq -A_x^\top B^{-1} A_x$, and $C_x = -A_x^\top B^{-1} A_x$ if and only if

$$x = \frac{2(a_2(x) - a_1(x)E[\varepsilon^3])}{a_1(x)(E[\varepsilon^4]-1) - a_2(x)E[\varepsilon^3]}; f(x) = -\frac{2a_1(x) + xa_2(x)}{\frac{E[\varepsilon^4]-1}{2}x^2 + 2xE[\varepsilon^3] + 2}.$$

■

LEMMA 2. *Suppose assumptions A.1-A.4 hold and there is an estimator $\widetilde{\theta}$ that has influence function $\chi_t(\theta_0)$,then the following expansion for distribution estimators based on $\widetilde{\theta}$ is*

$$\sup_{x\in\mathbb{R}} \left| \widehat{\widehat{F}}(x) - F(x) - \frac{1}{n}\sum_{t=1}^{n}\{1(\varepsilon_t \leq x) - F(x)\} - \frac{xf(x)}{2}H(\theta_0)^\top \frac{1}{n}\sum_{t=1}^{n}\chi_t(\theta_0) \right| = o_p(n^{-1/2})$$

$$\sup_{x\in\mathbb{R}} \left| \widehat{\widehat{F}}_N(x) - F(x) - \frac{1}{n}\sum_{t=1}^{n}\{1(\varepsilon_t \leq x) - F(x)\} - \frac{xf(x)}{2}H(\theta_0)^\top \frac{1}{n}\sum_{t=1}^{n}\chi_t(\theta_0) - f(x)\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t \right.$$
$$\left. -\frac{xf(x)}{2}\frac{1}{n}\sum_{t=1}^{n}\{\varepsilon_t^2 - 1\} \right| = o_p(n^{-1/2})$$

$$\sup_{x\in\mathbb{R}} \left| \widehat{\widehat{F}}_w(x) - F(x) - \frac{1}{n}\sum_{t=1}^{n}\{1(\varepsilon_t \leq x) - F(x)\} - \left\{\frac{xf(x)}{2} + e_2^\top B^{-1}A_x\right\} H(\theta_0)^\top \frac{1}{n}\sum_{t=1}^{n}\chi_t(\theta_0) \right.$$
$$\left. +\frac{1}{n}\sum_{t=1}^{n} A_x^\top B^{-1} R_2(\varepsilon_t) \right| = o_p(n^{-1/2}),$$

where $e_2^\top = (0,1)$.

Proof of Lemma 2. By Taylor expansion,

$$\begin{aligned}\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}} - 1 &= \frac{1}{2}\frac{\partial \log h_t(\theta)}{\partial \theta^\top}(\widetilde{\theta} - \theta_0) \\ &+ \frac{1}{4}(\widetilde{\theta} - \theta_0)^\top [\frac{1}{h_t(\theta)}\frac{\partial^2 h_t(\theta_1)}{\partial\theta\partial\theta^\top} - \frac{1}{2}\frac{\partial \log h_t(\theta_1)}{\partial\theta}\frac{\partial \log h_t(\theta_1)}{\partial\theta^\top}](\widetilde{\theta} - \theta_0),\end{aligned}$$

where θ_1 lies in between θ_0 and $\widetilde{\theta}$. Since $\widetilde{\theta} - \theta_0 = \frac{1}{n}\sum_{t=1}^{n}\chi_t(\theta_0) + o_p(\frac{1}{\sqrt{n}})$, and $E_{\theta_0}\sup_{\theta\in U_{\theta_0}} ||\frac{\partial \log h_t(\theta)}{\partial\theta}||^2 < \infty$,which is due to Example 3.1 in Koul and Ling (2006), we have

$\sum_{t=1}^{n}\left(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}-1\right)^2 = o_p(1)$. This implies

$$\sup_{1\le t\le n}\left|\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}-1\right| = o_p(1).$$

Using the same empirical process argument as in lemma , Lemma 4.1 in Koul and Ling (2006), and the fact that it is clear that, uniformly in $x\in\mathbb{R}$,

$$\begin{aligned}
&\sqrt{n}(\widehat{\widehat{F}}(x)-F(x))\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t(\widetilde{\theta})\le x)-F(x)\}\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}x)-F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}x)+F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}x)-F(x)\}\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}x)-F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}x)\}+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}x)-F(x)\}\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}x)-F(x)\}+o_p(1)\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{xf(x)}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)+o_p(1)\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}+\frac{1}{n}\sum_{t=1}^{n}\frac{xf(x)}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\chi_t(\theta_0)+o_p(1)\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}+\frac{xf(x)}{2}H(\theta_0)^\top\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\chi_t(\theta_0)+o_p(1),
\end{aligned}$$

where the last equation holds because of the ergodicity theorem $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\frac{\partial\log h_t(\theta_0)}{\partial\theta}=E[\frac{\partial\log h_t(\theta_0)}{\partial\theta}]$.

The next is to show the asymptotic expansion for $\widehat{\widehat{F}}_N(x)$. Since $\varepsilon_t(\widetilde{\theta})=\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t$, the renormalized empirical distribution estimator can be shown, uniformly in $x\in\mathbb{R}$:

$$\begin{aligned}
&\sqrt{n}(\widehat{\widehat{F}}_N(x)-F(x))\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1\left(\frac{\varepsilon_t(\widetilde{\theta})-\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t(\widetilde{\theta})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2(\widetilde{\theta})-(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t(\widetilde{\theta}))^2}}\le x\right)-F(x)\}\\
=\ & \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t+\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2-(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t(\widetilde{\theta}))^2x)}
\end{aligned}$$

$$
\begin{aligned}
&-F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t+\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2-(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t(\widetilde{\theta}))^2}x)\\
&+F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t+\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2-(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t(\widetilde{\theta}))^2}x)-F(x)\}\\
=\ &\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t\\
&+\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2(\widetilde{\theta})-(\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t(\widetilde{\theta}))^2}x)-F(x)\}+o_p(1)
\end{aligned}
$$

where the last equation used empirical process approximation and

Now given that $\sqrt{n}(\widetilde{\theta}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\chi_t(\theta_0)+o_p(1)$, we know that $\sqrt{\frac{1}{n}\sum\limits_{t=1}^{n}\varepsilon_t^2(\widetilde{\theta})-(\frac{1}{n}\sum\limits_{t=1}^{n}\varepsilon_t(\widetilde{\theta}))^2}$ is of the same order as $\sqrt{\frac{1}{n}\sum\limits_{t=1}^{n}\varepsilon_t^2(\widetilde{\theta})}$, which is due to the fact that $(\frac{1}{n}\sum\limits_{t=1}^{n}\varepsilon_t(\widetilde{\theta}))^2$ is of higher order than $\frac{1}{n}\sum\limits_{t=1}^{n}\varepsilon_t^2(\widetilde{\theta})$. As a result,

$$
\begin{aligned}
&F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t+\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2}x)-F(x)\\
=\ &f(x)\{\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}-1\}\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t\\
&+f(x)\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t\\
&+xf(x)\{\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}-1\}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2}\\
&+xf(x)\{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2}-1\}+O_p(\frac{1}{n})\\
=\ &I_{1t}+I_{2t}+I_{3t}+I_{4t},
\end{aligned}
$$

where

$$
\begin{aligned}
I_{1t} &= f(x)\{\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^{\top}}(\widetilde{\theta}-\theta_0)\}\frac{1}{n}\sum_{t=1}^{n}[1-\frac{1}{\sqrt{h_t(\theta_0)}}\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^{\top}}(\widetilde{\theta}-\theta_0)]\varepsilon_t\\
I_{2t} &= f(x)\frac{1}{n}\sum_{t=1}^{n}[1-\frac{1}{\sqrt{h_t(\theta_0)}}\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^{\top}}(\widetilde{\theta}-\theta_0)]\varepsilon_t\\
I_{3t} &= xf(x)\{\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^{\top}}(\widetilde{\theta}-\theta_0)\}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2}
\end{aligned}
$$

$$I_{4t} = xf(x)\{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2} - 1\}.$$

It is easy to see that

$$\begin{aligned}
\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}} - 1 &= \frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0) + O_p(\frac{1}{n}) \\
\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}} - 1 &= -\frac{1}{\sqrt{h_t(\theta_0)}}\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0) + O_p(\frac{1}{n}) \\
\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})} - 1 &= -\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0) + O_p(\frac{1}{n})
\end{aligned}$$

so now the four components can be rewritten as:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1}^{n} I_{1t} \\
= &\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{f(x)\{\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)\}\frac{1}{n}\sum_{t=1}^{n}[1-\frac{1}{\sqrt{h_t(\theta_0)}}\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)]\varepsilon_t\} \\
= &\frac{f(x)}{2}\{\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t - \frac{1}{n}\sum_{t=1}^{n}\frac{1}{2h_t^{3/2}(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)\varepsilon_t\}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{t=1}^{n} I_{2t} &= f(x)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}[1-\frac{1}{\sqrt{h_t(\theta_0)}}\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)]\varepsilon_t \\
\frac{1}{\sqrt{n}}\sum_{t=1}^{n} I_{3t} &= xf(x)\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)\} \\
\frac{1}{\sqrt{n}}\sum_{t=1}^{n} I_{4t} &= xf(x)\sqrt{n}\{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2} - 1\}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{F(\sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\frac{1}{n}\sum_{t=1}^{n}\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t + \sqrt{\frac{h_t(\widetilde{\theta})}{h_t(\theta_0)}}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}\varepsilon_t^2}x) - F(x)\} \\
= &f(x)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}[1-\frac{1}{2h_t^{3/2}(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)]\varepsilon_t + xf(x)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0) \\
&+\frac{xf(x)}{2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{\varepsilon_t^2 - 1\} + o_p(1) \\
= &f(x)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{[1-\frac{1}{2h_t^{3/2}(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)]\varepsilon_t
\end{aligned}$$

$$+x\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)+\frac{x}{2}[\varepsilon_t^2-1]\}+o_p(1),$$

and by CLT and LLN,

$$\begin{aligned}\sqrt{n}(\widetilde{\theta}-\theta_0) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\chi_t(\theta_0)+o_p(1)\\ \frac{1}{n}\sum_{t=1}^{n}\frac{1}{2h_t^{3/2}(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta}\varepsilon_t &= o_p(1)\end{aligned}$$

we have

$$\frac{1}{\sqrt{n}}f(x)\sum_{t=1}^{n}\{\frac{1}{\sqrt{h_t(\theta_0)}}\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)\varepsilon_t\}=o_p(1).$$

Therefore, uniformly in $x\in\mathbb{R}$,

$$\begin{aligned}&\sqrt{n}(\widehat{\widehat{F}}_N(x)-F(x))\\ =\;& n^{-1/2}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}\\ &+f(x)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{\varepsilon_t+x\frac{1}{2h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)+\frac{x}{2}[\varepsilon_t^2-1]\}+o_p(1).\end{aligned}$$

Since we know that

$$\widehat{\widehat{w}}_t=\frac{1}{n}\frac{1}{1+\widehat{\lambda}_n'R_2(\varepsilon_t(\widetilde{\theta}))};\widehat{\lambda}_n=B_n^{-1}(\frac{1}{n}\sum_{t=1}^{n}R_2(\varepsilon_t(\widetilde{\theta})))+o_p(n^{-1/2})$$

where $B_n=\frac{1}{n}\sum_{t=1}^{n}R_2(\varepsilon_t(\widetilde{\theta}))R_2(\varepsilon_t(\widetilde{\theta}))^\top$. Therefore,

$$\begin{aligned}&\sqrt{n}(\widehat{\widehat{F}}_w(x)-F(x))\\ =\;&\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{n\widehat{\widehat{w}}_t1(\varepsilon_t(\widetilde{\theta})\le x)-F(x)\}\\ =\;&\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{n\widehat{\widehat{w}}_t1(\varepsilon_t(\widetilde{\theta})\le x)-1(\varepsilon_t(\widetilde{\theta})\le x)+1(\varepsilon_t(\widetilde{\theta})\le x)-F(x)\}\\ =\;&\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{[n\widehat{\widehat{w}}_t-1]1(\varepsilon_t(\widetilde{\theta})\le x)\}+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t(\widetilde{\theta})\le x)-F(x)\}\\ =\;&I_5+\sqrt{n}(\widehat{\widehat{F}}(x)-F(x)).\end{aligned}$$

Define $\varepsilon_t^*=\varepsilon_t+\frac{\varepsilon_t}{2}\frac{\partial\log h_t(\theta)}{\partial\theta^\top}(\widetilde{\theta}-\theta)$. From the $\sqrt{n}$-consistency of $\widetilde{\theta}$ and $E[\frac{\varepsilon_t}{2}\frac{\partial\log h_t(\theta)}{\partial\theta^\top}]=0$, we can see that

$$\sum_{t=1}^{n}\left(\varepsilon_t(\widetilde{\theta})-\varepsilon_t\right)^2=\sum_{t=1}^{n}\left(\frac{\varepsilon_t}{2}\frac{\partial\log h_t(\theta)}{\partial\theta^\top}(\widetilde{\theta}-\theta)+O_p((\widetilde{\theta}-\theta)^2)\right)^2=o_p(1),$$

which implies that $\max_{1\le t\le n}|\varepsilon_t(\widetilde{\theta})-\varepsilon_t|=o_p(1)$.

This means residuals $\varepsilon_t(\widetilde{\theta})=\sqrt{\frac{h_t(\theta_0)}{h_t(\widetilde{\theta})}}\varepsilon_t$ are uniformly close to ε_t. Therefore for the weights $\widehat{\widehat{w}}_t=\frac{1}{n}\frac{1}{1+\widehat{\lambda}_n' R_2(\varepsilon_t(\widetilde{\theta}))}$, define $B_n=\frac{1}{n}\sum_{t=1}^{n}R_2(\varepsilon_t(\widetilde{\theta}))R_2(\varepsilon_t(\widetilde{\theta}))^\top$, and we can see

$$\begin{aligned}\widehat{\lambda}_n &= B_n^{-1}(\frac{1}{n}\sum_{t=1}^{n}R_2(\varepsilon_t(\widetilde{\theta})))+o_p(n^{-1/2})\\ &= B^{-1}\frac{1}{n}\sum_{t=1}^{n}[\begin{pmatrix}\varepsilon_t\\ \varepsilon_t^2-1\end{pmatrix}-\frac{1}{2}\begin{pmatrix}\varepsilon_t\\ 2\varepsilon_t^2\end{pmatrix}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)]+o_p(n^{-1/2})\end{aligned}$$

so

$$\begin{aligned}\sqrt{n}\widehat{\lambda}_n &= B^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}[\begin{pmatrix}\varepsilon_t\\ \varepsilon_t^2-1\end{pmatrix}-\frac{1}{2}\begin{pmatrix}\varepsilon_t\\ 2\varepsilon_t^2\end{pmatrix}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)]+o_p(1)\\ &= B^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}R_2(\varepsilon_t)-\frac{1}{2}B^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{pmatrix}\varepsilon_t\\ 2\varepsilon_t^2\end{pmatrix}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)+o_p(1).\end{aligned}$$

Hence,

$$\begin{aligned}I_5 &= -\sqrt{n}\widehat{\lambda}_n^\top\frac{1}{n}\sum_{t=1}^{n}\{R_2(\varepsilon_t(\widetilde{\theta}))1(\varepsilon_t(\widetilde{\theta})\le x)\}\\ &= -\{\frac{1}{\sqrt{n}}\sum_{t=1}^{n}R_2(\varepsilon_t)^\top B^{-1}A_x-\frac{1}{2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{pmatrix}\varepsilon_t & 2\varepsilon_t^2\end{pmatrix}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)B^{-1}A_x\}+o_p(1),\end{aligned}$$

where the last equality holds because of ergodic theorem: $n^{-1}\sum_{t=1}^{n}\{R_2(\varepsilon_t(\widetilde{\theta}))1(\varepsilon_t(\widetilde{\theta})\le x)\}=A_x+o_p(1)$. So, uniformly in $x\in\mathbb{R}$,

$$\begin{aligned}&\sqrt{n}(\widehat{\widehat{F}}_w(x)-F(x))\\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{[n\widehat{\widehat{w}}_t-1]1(\varepsilon_t(\widetilde{\theta})\le x)\}+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t(\widetilde{\theta})\le x)-F(x)\}\\ &= -\frac{1}{\sqrt{n}}\sum_{t=1}^{n}A_x^\top B^{-1}R_2(\varepsilon_t)+\frac{1}{2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{pmatrix}\varepsilon_t & 2\varepsilon_t^2\end{pmatrix}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)B^{-1}A_x\\ &+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}+\frac{xf(x)}{2}H(\theta_0)^\top\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\chi_t(\theta_0)+o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{1(\varepsilon_t\le x)-F(x)\}+\{\frac{xf(x)}{2}+e_2^\top B^{-1}A_x\}H(\theta_0)^\top\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\chi_t(\theta_0)\end{aligned}$$

$$-\frac{1}{\sqrt{n}}\sum_{t=1}^{n} A_x^\top B^{-1} R_2(\varepsilon_t) + o_p(1),$$

because

$$\begin{aligned}
& \frac{1}{2}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\begin{pmatrix} \varepsilon_t & 2\varepsilon_t^2 \end{pmatrix}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}(\widetilde{\theta}-\theta_0)B^{-1}A_x \\
= & \frac{1}{2}\frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} \varepsilon_t & 2\varepsilon_t^2 \end{pmatrix}\frac{1}{h_t(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top}\sqrt{n}(\widetilde{\theta}-\theta_0)B^{-1}A_x \\
= & e_2^\intercal B^{-1}A_x H(\theta_0)^\top \sqrt{n}(\widetilde{\theta}-\theta_0) + o_p(1) \\
= & \frac{a_2(x)-a_1(x)E[\varepsilon^3]}{E[\varepsilon^4]-1-E[\varepsilon^3]^2}H(\theta_0)^\top \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\chi_t(\theta_0) + o_p(1).
\end{aligned}$$

■

Proof of Theorem 2. Lemma 1 and the Proposition 1 of Gill (1989) imply the results regarding VaR. Notice that, for any consistent distribution function estimator $\widetilde{F}(x)$ with associated quantile estimator $\widetilde{q}_\alpha = \widetilde{F}^{-1}(\alpha)$, the expected shortfall can be expressed as

$$\alpha\widetilde{ES}_\alpha = \int_{-\infty}^{\widetilde{q}_\alpha} x d\widetilde{F}(x) = \widetilde{q}_\alpha\widetilde{F}(\widetilde{q}_\alpha) - \int_{-\infty}^{\widetilde{q}_\alpha}\widetilde{F}(x)dx = \alpha\widetilde{q}_\alpha - \int_{-\infty}^{\widetilde{q}_\alpha}\widetilde{F}(x)dx$$

we can see that

$$\begin{aligned}
& \alpha(\widetilde{ES}_\alpha - ES_\alpha) \\
= & \int_{-\infty}^{\widetilde{q}_\alpha} x d\widetilde{F}(x) - \int_{-\infty}^{q_\alpha} x dF(x) \\
= & \alpha\widetilde{q}_\alpha - \int_{-\infty}^{\widetilde{q}_\alpha}\widetilde{F}(x)dx - \alpha q_\alpha + \int_{-\infty}^{q_\alpha} F(x)dx \\
= & \int_{-\infty}^{q_\alpha}(F(x)-\widetilde{F}(x))dx + \alpha(\widetilde{q}_\alpha - q_\alpha) - \int_{q_\alpha}^{\widetilde{q}_\alpha}\widetilde{F}(x)dx \\
= & \int_{-\infty}^{q_\alpha}(F(x)-\widetilde{F}(x))dx + (\widetilde{q}_\alpha - q_\alpha)(\alpha - \widetilde{F}(\overline{\widetilde{q}}_\alpha)) \\
= & \int_{-\infty}^{q_\alpha}(F(x)-\widetilde{F}(x))dx + o_p(n^{-1/2}).
\end{aligned}$$

As a result:

$$\begin{aligned}
& \alpha(\widehat{ES}_\alpha - ES_\alpha) \\
= & \int_{-\infty}^{q_\alpha}(F(x)-\widehat{F}(x))dx + o_p(n^{-1/2}) \\
= & \int_{-\infty}^{q_\alpha} F(x)dx - \frac{1}{n}\sum_{t=1}^{n}(q_\alpha - \varepsilon_t)1(\varepsilon_t \le q_\alpha) + o_p(n^{-1/2})
\end{aligned}$$

$$
\begin{aligned}
&\alpha(\widehat{ES}_{N\alpha} - ES_\alpha) \\
= & \int_{-\infty}^{q_\alpha} (F(x) - \widehat{F}_N(x))dx + o_p(n^{-1/2}) \\
= & -\frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{q_\alpha}\left\{1(\varepsilon_t \le x) - F(x) + f(x)\varepsilon_t + \frac{xf(x)}{2}(\varepsilon_t^2 - 1)\right\}dx + o_p(n^{-1/2}) \\
= & -\frac{1}{n}\sum_{t=1}^{n}\left\{\int_{-\infty}^{q_\alpha} 1(\varepsilon_t \le x)dx - \int_{-\infty}^{q_\alpha} F(x)dx + \varepsilon_t \int_{-\infty}^{q_\alpha} f(x)dx + (\varepsilon_t^2 - 1)\int_{-\infty}^{q_\alpha}\frac{xf(x)}{2}dx\right\} + o_p(n^{-1/2}) \\
= & -\frac{1}{n}\sum_{t=1}^{n}\left\{(q_\alpha - \varepsilon_t)1(\varepsilon_t \le q_\alpha) - \int_{-\infty}^{q_\alpha} F(x)dx + \alpha\varepsilon_t + \frac{\varepsilon_t^2 - 1}{2}\int_{-\infty}^{q_\alpha} xf(x)dx\right\} + o_p(n^{-1/2})
\end{aligned}
$$

$$
\begin{aligned}
&\alpha(\widehat{ES}_{w\alpha} - ES_\alpha) \\
= & \int_{-\infty}^{q_\alpha} (F(x) - \widehat{F}_w(x))dx + o_p(n^{-1/2}) \\
= & -\frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{q_\alpha}\left\{1(\varepsilon_t \le x) - F(x) - A_x^\top B^{-1}R_2(\varepsilon_t)\right\}dx + o_p(n^{-1/2}) \\
= & -\frac{1}{n}\sum_{t=1}^{n}\left\{\int_{-\infty}^{q_\alpha} 1(\varepsilon_t \le x)dx - \int_{-\infty}^{q_\alpha} F(x)dx - \int_{-\infty}^{q_\alpha} A_x^\top B^{-1}R_2(\varepsilon_t)dx\right\} + o_p(n^{-1/2}) \\
= & -\frac{1}{n}\sum_{t=1}^{n}\left\{(q_\alpha - \varepsilon_t)1(\varepsilon_t \le q_\alpha) - \int_{-\infty}^{q_\alpha} F(x)dx - R_2^\top(\varepsilon_t)B^{-1}\int_{-\infty}^{q_\alpha} A_x dx\right\} + o_p(n^{-1/2}).
\end{aligned}
$$

■

COROLLARY 3. *Suppose that the semiparametric efficient estimator that has influence function $\chi_t(\theta_0) = \psi_t(\theta_0)$ is used. Then, the process $\sqrt{n}(\widehat{\widehat{F}} - F)$ converges weakly to a zero-mean Gaussian process $\mathcal{Z}$ with covariance function Ω, the process $\sqrt{n}(\widehat{\widehat{F}}_N - F)$ converges weakly to a zero-mean Gaussian process $\mathcal{Z}_{\widehat{N}}$ with covariance function $\Omega_{\widehat{N}}$, and the process $\sqrt{n}(\widehat{\widehat{F}}_w - F)$ converges weakly to a zero-mean Gaussian process $\mathcal{Z}_{\widehat{w}}$ with covariance function $\Omega_{\widehat{w}}$, where:*

$$
\begin{aligned}
\Omega(x, x') &= \operatorname{cov}(\mathcal{Z}(x), \mathcal{Z}(x')) \\
&= E\Big[[1(\varepsilon \le x) - F(x) + \frac{xf(x)}{2}(\varepsilon^2 - 1 - \varepsilon E\varepsilon^3)] \\
&\quad \times [1(\varepsilon \le x') - F(x') + \frac{x'f(x')}{2}(\varepsilon^2 - 1 - \varepsilon E\varepsilon^3)]\Big] \\
\Omega_{\widehat{N}}(x, x') &= \operatorname{cov}\left(\mathcal{Z}_{\widehat{N}}(x), \mathcal{Z}_{\widehat{N}}(x')\right) \\
&= E\Big[[1(\varepsilon \le x) - F(x) + [f(x) - \frac{xf(x)}{2}E\varepsilon^3]\varepsilon + xf(x)(\varepsilon^2 - 1)] \\
&\quad \cdot [1(\varepsilon \le x') - F(x') + [f(x') - \frac{x'f(x')}{2}E\varepsilon^3]\varepsilon + x'f(x')(\varepsilon^2 - 1)]\Big] \\
\Omega_{\widehat{w}}(x, x') &= \operatorname{cov}(\mathcal{Z}_{\widehat{w}}(x), \mathcal{Z}_{\widehat{w}}(x')) \\
&= E\Big[[1(\varepsilon \le x) - F(x) + \{\frac{xf(x)}{2} + e_2^\intercal B^{-1}A_x\}(\varepsilon^2 - 1 - \varepsilon E\varepsilon^3) - A_x^\top B^{-1}R_2(\varepsilon)]
\end{aligned}
$$

$$\cdot[1(\varepsilon \le x') - F(x') + \{\frac{x'f(x')}{2} + e_2^\intercal B^{-1}A_{x'}\}(\varepsilon^2 - 1 - \varepsilon E\varepsilon^3) - A_{x'}^\top B^{-1}R_2(\varepsilon)]\Big].$$

Denote $E[\varepsilon 1(\varepsilon \le x)] = a_1(x)$ and $E[(\varepsilon^2 - 1)1(\varepsilon \le x)] = a_2(x)$. The pointwise asymptotic variances are $\Omega_j(x)$, where:

$$\begin{aligned}
\Omega_1(x) &= F(x)(1-F(x)) + \frac{x^2 f(x)^2[E\varepsilon^4 - 1 - (E\varepsilon^3)^2]}{4} + xf(x)(a_2(x) - a_1(x)E\varepsilon^3) \\
\Omega_2(x) &= F(x)(1-F(x)) + C_x + \frac{3x^2 f(x)^2[E\varepsilon^4 - 1 - (E\varepsilon^3)^2]}{4} + xf(x)(a_2(x) - a_1(x)E\varepsilon^3) \\
\Omega_3(x) &= F(x)(1-F(x)) - A_x^\top B^{-1}A_x + \left\{\frac{xf(x)}{2} + e_2^\intercal B^{-1}A_x\right\}^2 [E\varepsilon^4 - 1 - (E\varepsilon^3)^2].
\end{aligned}$$

It can be shown that $\Omega_1(x) - \Omega_3(x) = a_1^2(x) \ge 0$. As a result, $\hat{\hat{F}}_w(x)$ is uniformly more efficient than $\hat{\hat{F}}(x)$ and they are equally efficient at x where $E[\varepsilon 1(\varepsilon \le x)] = 0$. It can also be shown that $\Omega_2(x) \ge \Omega_3(x)$.

Proof of Corollary 3. Since $H_t(\theta_0) = \binom{G_t(\theta_0)}{2/c}$, we know that

$$H(\theta_0)^\top \begin{pmatrix} -\frac{1}{2}E[l_{1t}^* l_{1t}^{*\top}]^{-1}\{G_t - G\} & 0 \\ \frac{c_0}{4}G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1}\{G_t - G\} & \frac{c_0}{2}(-E\varepsilon^3, 1) \end{pmatrix} \begin{pmatrix} R_3(\varepsilon_t) \\ R_2(\varepsilon_t) \end{pmatrix} = \varepsilon_t^2 - 1 - \varepsilon_t E\varepsilon^3.$$

Plug

$$\chi_t(\theta_0) = \begin{pmatrix} -\frac{1}{2}E[l_{1t}^* l_{1t}^{*\top}]^{-1}\{G_t - G\} & 0 \\ \frac{c_0}{4}G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1}\{G_t - G\} & \frac{c_0}{2}(-E\varepsilon^3, 1) \end{pmatrix} \begin{pmatrix} R_3(\varepsilon_t) \\ R_2(\varepsilon_t) \end{pmatrix}$$

into the expressions in lemma 2 and get, uniformly in $x \in \mathbb{R}$,

$$\begin{aligned}
\hat{\hat{F}}(x) - F(x) &= \frac{1}{n}\sum_{t=1}^n \{1(\varepsilon_t \le x) - F(x) + \frac{xf(x)}{2}[\varepsilon_t^2 - 1 - \varepsilon_t E\varepsilon^3]\} + o_p(n^{-1/2}) \\
\hat{\hat{F}}_N(x) - F(x) &= \frac{1}{n}\sum_{t=1}^n \{1(\varepsilon_t \le x) - F(x) + [f(x) - \frac{xf(x)}{2}E\varepsilon^3]\varepsilon_t + xf(x)[\varepsilon_t^2 - 1]\} + o_p(n^{-1/2}) \\
\hat{\hat{F}}_w(x) - F(x) &= \frac{1}{n}\sum_{t=1}^n \{1(\varepsilon_t \le x) - F(x)\} + \{\frac{xf(x)}{2} + e_2^\intercal B^{-1}A_x\}[\varepsilon_t^2 - 1 - \varepsilon_t E\varepsilon^3] \\
&\quad - \frac{1}{n}\sum_{t=1}^n A_x^\top B^{-1}R_2(\varepsilon_t) + o_p(n^{-1/2}).
\end{aligned}$$

Due to moment constraints (3.2), the following holds:

$$\begin{aligned}
E[(1 + \varepsilon\frac{f'(\varepsilon)}{f(\varepsilon)})1(\varepsilon &\le x)] = F(x) + \int_{-\infty}^x \varepsilon df(\varepsilon) = xf(x) \\
E[\varepsilon\frac{f'(\varepsilon)}{f(\varepsilon)}] &= \int \varepsilon df(\varepsilon) = -1
\end{aligned}$$

$$
\begin{aligned}
E[\varepsilon^2 \frac{f'(\varepsilon)}{f(\varepsilon)}] &= \int \varepsilon^2 df(\varepsilon) = 0 \\
E[\varepsilon^3 \frac{f'(\varepsilon)}{f(\varepsilon)}] &= \int \varepsilon^3 df(\varepsilon) = -3.
\end{aligned}
$$

These equations and CLT show that they have asymptotic variance as follows:

$$
\begin{aligned}
\Omega_1 &= F(x)(1-F(x)) + \frac{x^2 f(x)^2}{4}[E\varepsilon^4 - 1 - (E\varepsilon^3)^2] + xf(x)(a_2(x) - a_1(x)E\varepsilon^3) \\
\Omega_2 &= F(x)(1-F(x)) + C_x + \frac{3x^2 f(x)^2 [E\varepsilon^4 - 1 - (E\varepsilon^3)^2]}{4} + xf(x)(a_2(x) - a_1(x)E\varepsilon^3) \\
\Omega_3 &= F(x)(1-F(x)) - A_x^\top B^{-1} A_x + \{\frac{xf(x)}{2} + e_2^\intercal B^{-1} A_x\}^2 [E\varepsilon^4 - 1 - (E\varepsilon^3)^2].
\end{aligned}
$$

■

Proof of Theorem 3. Lemma 2 and the Proposition 1 of Gill (1989) imply above results for VaR. Similar to the proof of corollary 2, we have

$$
\begin{aligned}
& \alpha(\widehat{\widehat{ES}}_\alpha - ES_\alpha) \\
= & \int_{-\infty}^{\hat{\hat{q}}_\alpha} x d\hat{\hat{F}}(x) - \int_{-\infty}^{q_\alpha} x dF(x) \\
= & \int_{-\infty}^{q_\alpha} (F(x) - \hat{\hat{F}}(x))dx + \alpha(\hat{\hat{q}}_\alpha - q_\alpha) - \int_{q_\alpha}^{\hat{\hat{q}}_\alpha} \hat{\hat{F}}(x)dx \\
= & \int_{-\infty}^{q_\alpha} (F(x) - \hat{\hat{F}}(x))dx + \alpha(\hat{\hat{q}}_\alpha - q_\alpha) - (\hat{\hat{q}}_\alpha - q_\alpha)\hat{\hat{F}}(\hat{\hat{q}}_\alpha) \\
= & \int_{-\infty}^{q_\alpha} (F(x) - \hat{\hat{F}}(x))dx + o_p(n^{-1/2}).
\end{aligned}
$$

Then the theorem holds because of the following:

$$
\begin{aligned}
& \int_{-\infty}^{q_\alpha} (F(x) - \hat{\hat{F}}(x))dx \\
= & -\frac{1}{n}\sum_{t=1}^{n} \int_{-\infty}^{q_\alpha} (1(\varepsilon_t \le x) - F(x) + \frac{xf(x)}{2}(\varepsilon_t^2 - 1 - \varepsilon_t E\varepsilon^3))dx \\
= & -\frac{1}{n}\sum_{t=1}^{n} \left\{(q_\alpha - \varepsilon_t)1(\varepsilon_t \le q_\alpha) - \int_{-\infty}^{q_\alpha} F(x)dx + (\varepsilon_t^2 - 1 - \varepsilon_t E\varepsilon^3)\int_{-\infty}^{q_\alpha} \frac{xf(x)}{2}dx\right\},
\end{aligned}
$$

$$
\begin{aligned}
& \int_{-\infty}^{q_\alpha} (F(x) - \hat{\hat{F}}_N(x))dx \\
= & -\frac{1}{n}\sum_{t=1}^{n} \int_{-\infty}^{q_\alpha} \left\{1(\varepsilon_t \le x) - F(x) + [f(x) - \frac{xf(x)}{2}E\varepsilon^3]\varepsilon_t + xf(x)(\varepsilon_t^2 - 1)\right\} dx
\end{aligned}
$$

$$= -\frac{1}{n}\sum_{t=1}^{n}\left\{\begin{array}{c}(q_\alpha-\varepsilon_t)1(\varepsilon_t\le q_\alpha)-\int_{-\infty}^{q_\alpha}F(x)dx+\varepsilon_t\int_{-\infty}^{q_\alpha}[f(x)-\frac{xf(x)}{2}E\varepsilon^3]dx\\ +(\varepsilon_t^2-1)\int_{-\infty}^{q_\alpha}xf(x)dx\end{array}\right\},$$

$$\begin{aligned}
&\int_{-\infty}^{q_\alpha}(F(x)-\widehat{\widehat{F}}_w(x))dx\\
=&-\frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{q_\alpha}\left\{1(\varepsilon_t\le x)-F(x)+[\frac{xf(x)}{2}+e_2^{\intercal}B^{-1}A_x](\varepsilon_t^2-1-\varepsilon_tE\varepsilon^3)-A_x^{\top}B^{-1}R_2(\varepsilon_t)\right\}dx\\
=&-\frac{1}{n}\sum_{t=1}^{n}\left\{(q_\alpha-\varepsilon_t)1(\varepsilon_t\le q_\alpha)-\int_{-\infty}^{q_\alpha}F(x)dx\right.\\
&\left.+(\varepsilon_t^2-1-\varepsilon_tE\varepsilon^3)\int_{-\infty}^{q_\alpha}[\frac{xf(x)}{2}+e_2^{\intercal}B^{-1}A_x]dx-R_2^{\top}(\varepsilon_t)B^{-1}\int_{-\infty}^{q_\alpha}A_xdx\right\}.
\end{aligned}$$

■

Proof of Theorem 4. Since $R_1(\varepsilon)=1(\varepsilon\le q_\alpha)-\alpha$, $R_2(\varepsilon)=(\varepsilon,\varepsilon^2-1)^{\top}$, $R_3(\varepsilon)=1+\varepsilon\frac{f'(\varepsilon)}{f(\varepsilon)}$, and $R(\varepsilon)=(R_1(\varepsilon),R_2(\varepsilon)^{\top},R_3(\varepsilon))^{\top}$. It is seen that

$$E[R_1(\varepsilon_t)|\mathcal{F}_{t-1}]=E[R_2(\varepsilon_t)|\mathcal{F}_{t-1}]=E[R_3(\varepsilon_t)|\mathcal{F}_{t-1}]=0,$$

which implies that $\{Z_s\}$ is Martingale Difference Series. From Theorem 2, we have

$$\sqrt{n}\begin{pmatrix}\widehat{\theta}-\theta_0\\ \widehat{\widehat{q}}_{w\alpha}-q_\alpha\end{pmatrix}=\frac{1}{\sqrt{n}}\sum_{s=1}^{n}Z_s+o_p(1),$$

$$\begin{aligned}
Z_s&=\Psi_sR(\varepsilon_s)\\
\Psi_t&=\begin{pmatrix}0 & 0 & -\frac{1}{2}E[l_{1t}^*l_{1t}^{*\top}]^{-1}\{G_t-G\}\\ 0 & \frac{c_0}{2}(-E\varepsilon^3,1) & \frac{c_0}{4}G^{\top}E[l_{1t}^*l_{1t}^{*\top}]^{-1}\{G_t-G\}\\ \frac{-1}{f(q_\alpha)} & \frac{A_{q_\alpha}^{\top}B^{-1}}{f(q_\alpha)}-[\frac{q_\alpha}{2}+\frac{\begin{pmatrix}0&1\end{pmatrix}B^{-1}A_{q_\alpha}}{f(q_\alpha)}](-E\varepsilon^3,1) & 0\end{pmatrix}.
\end{aligned}$$

Since: $E[R_1(\varepsilon)R_2(\varepsilon)]=A_{q_\alpha}$, $E[R_1(\varepsilon)R_3(\varepsilon)]=q_\alpha f(q_\alpha)$, and $E[R_2(\varepsilon)R_3(\varepsilon)]=\begin{pmatrix}0 & -2\end{pmatrix}^{\top}$, we have

$$\begin{aligned}
&\Omega_Z\\
=&E[\Psi_sR(\varepsilon_s)R(\varepsilon_s)^{\top}\Psi_s^{\top}]
\end{aligned}$$

$$
\begin{aligned}
&= E[\Psi_s \begin{pmatrix} \alpha(1-\alpha) & A_{q_\alpha}^\top & q_\alpha f(q_\alpha) \\ A_{q_\alpha} & B & \begin{pmatrix} 0 & -2 \end{pmatrix}^\top \\ q_\alpha f(q_\alpha) & \begin{pmatrix} 0 & -2 \end{pmatrix} & E[(1+\varepsilon f'/f)^2] \end{pmatrix} \Psi_s^\top] \\
&= \begin{pmatrix} E[l_{1t}^* l_{1t}^{*\top}]^{-1} & -\frac{c_0}{2} E[l_{1t}^* l_{1t}^{*\top}]^{-1} G & 0 \\ -\frac{c_0}{2} G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1} & \frac{c_0^2[E\varepsilon^4-1-(E\varepsilon^3)^2]+G^\top E[l_{1t}^* l_{1t}^{*\top}]^{-1} G}{4} & \frac{c_0[E\varepsilon^4-1-(E\varepsilon^3)^2][\frac{q_\alpha}{2}+\frac{e_2^\top B^{-1} A_{q_\alpha}}{f(q_\alpha)}]}{2} \\ 0 & \frac{c_0[E\varepsilon^4-1-(E\varepsilon^3)^2][\frac{q_\alpha}{2}+\frac{e_2^\top B^{-1} A_{q_\alpha}}{f(q_\alpha)}]}{2} & \Omega_{\widehat{\widehat{q}}_{w\alpha}} \end{pmatrix}.
\end{aligned}
$$

Due to Taylor expansion,

$$
\begin{aligned}
& \sqrt{n}(\widehat{\xi}_{n,t} - \xi_t) \\
&= \sqrt{n}(h_t^{1/2}(\widehat{\theta})\widehat{\widehat{q}}_{w\alpha} - h_t^{1/2}(\theta_0)\widehat{\widehat{q}}_{w\alpha} + h_t^{1/2}(\theta_0)\widehat{\widehat{q}}_{w\alpha} - h_t^{1/2}(\theta_0)q_\alpha) \\
&= \sqrt{n}\frac{q_\alpha}{2h_t^{1/2}(\theta_0)}\frac{\partial h_t^{1/2}(\theta_0)}{\partial\theta^\top}(\widehat{\theta}-\theta_0) + h_t^{1/2}(\theta_0)\sqrt{n}(\widehat{\widehat{q}}_{w\alpha} - q_\alpha) + o_p(1) \\
&= \sqrt{n}\begin{pmatrix} \frac{q_\alpha}{2h_t^{1/2}(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top} & h_t^{1/2}(\theta_0) \end{pmatrix} \begin{pmatrix} \widehat{\theta}-\theta_0 \\ \widehat{\widehat{q}}_{w\alpha} - q_\alpha \end{pmatrix} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{s=1}^n W_t^\top Z_s + o_p(1),
\end{aligned}
$$

where $W_t = \begin{pmatrix} \frac{q_\alpha}{2h_t^{1/2}(\theta_0)}\frac{\partial h_t(\theta_0)}{\partial\theta^\top} & h_t^{1/2}(\theta_0) \end{pmatrix}^\top$. Denote $X_{ns} = n^{-1/2}W_t^\top Z_s$, it follows that

$$
\sum_{s=1}^n X_{ns}^2 = W_t^\top \frac{1}{n}\sum_{s=1}^n Z_s Z_s^\top W_t \to^p W_t^\top \Omega_Z W_t.
$$

From Martingale Central Limit theorem, we can see that

$$
\frac{\sum_{s=1}^n X_{ns}}{\sqrt{\sum_{s=1}^n X_{ns}^2}} \Longrightarrow N(0,1)
$$

and $\sqrt{n}(\widehat{\xi}_{n,t} - \xi_t) \Longrightarrow N(0, \omega_{\xi t})$, where

$$
\begin{aligned}
\omega_{\xi t} &= W_t^\top \Omega_Z W_t \\
&= h_t(\theta_0)\{\frac{q_\alpha^2}{4}(G_t^\top - G)E[l_{1t}^* l_{1t}^{*\top}]^{-1}(G_t - G) \\
&\quad + \frac{\alpha(1-\alpha) - A_{q_\alpha}^\top B^{-1} A_{q_\alpha}}{f(q_\alpha)^2} + [q_\alpha + \frac{e_2^\top B^{-1} A_{q_\alpha}}{f(q_\alpha)}]^2[E\varepsilon^4 - 1 - (E\varepsilon^3)^2]\}.
\end{aligned}
$$

Denote a truncated version of h_{n+1}as

$$h_{n+1}^{*}=\frac{c^{2}}{1-b}+ac^{2}\sum_{j=1}^{m}b^{j-1}y_{n+1-j}^{2}$$

where the truncation order is $m=\log n$. As a result, the approximation error is of order $o_{p}(1)$:

$$h_{n+1}-h_{n+1}^{*}=ac^{2}\sum_{j=m+1}^{\infty}b^{j-1}y_{n+1-j}^{2}=O_{p}(b^{m})$$

Similarly, we can show that $\frac{\partial h_{n+1}}{\partial\beta}-\frac{\partial h_{n+1}^{*}}{\partial\beta}=O_{p}(b^{m})$. Consequently, $W_{n+1}-W_{n+1}^{*}=O_{p}(b^{m})$.

At the same time, we have the following truncation approximation

$$\begin{aligned}\frac{1}{n}\sum_{s=1}^{n}Z_{s} &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-m}\Psi_{s}R(\varepsilon_{s})+\sqrt{\frac{m-1}{n}}\frac{1}{\sqrt{m-1}}\sum_{t=n-m+1}^{n}\Psi_{s}R(\varepsilon_{s})\\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n-m}\Psi_{s}R(\varepsilon_{s})+o_{p}(1)\end{aligned}$$

As $\{\varepsilon_{s}\}$ is iid sequence, we can draw the conclusion that $W_{n+1}\rightarrow^{p}W_{n+1}^{*}\perp\frac{1}{\sqrt{n}}\sum_{t=1}^{n-m}\Psi_{s}R(\varepsilon_{s})$.

The above argument applies to $\widehat{\chi}_{n,t}=h_{t}^{1/2}(\widehat{\theta})\widehat{\widehat{ES}}_{w\alpha}$ as:

$$\begin{aligned}&\sqrt{n}(\widehat{\chi}_{n,t}-\chi_{t})\\ =\ &\sqrt{n}(h_{t}^{1/2}(\widehat{\theta})\widehat{\widehat{ES}}_{w\alpha}-h_{t}^{1/2}(\theta_{0})\widehat{\widehat{ES}}_{w\alpha}+h_{t}^{1/2}(\theta_{0})\widehat{\widehat{ES}}_{w\alpha}-h_{t}^{1/2}(\theta_{0})ES_{\alpha})\end{aligned}$$

$$\widehat{\widehat{ES}}_{w\alpha}-ES_{\alpha}=\frac{1}{n}\sum_{t=1}^{n}\left\{\frac{1}{\alpha}(\varepsilon_{t}-q_{\alpha})1(\varepsilon_{t}\leq q_{\alpha})+\frac{1}{\alpha}\int_{-\infty}^{q_{\alpha}}F(x)dx+\frac{1}{\alpha}C^{\top}R_{2}(\varepsilon_{t})\right\}+o_{p}(n^{-1/2}),$$

where $C=\int_{-\infty}^{q_{\alpha}}[\frac{xf(x)}{2}+e_{2}^{\intercal}B^{-1}A_{x}]dx(E\varepsilon^{3},-1)+\int_{-\infty}^{q_{\alpha}}A_{x}^{\top}dxB^{-1}$. Notice that,

$$\operatorname{cov}(\widehat{\widehat{ES}}_{w\alpha}-ES_{\alpha},\widehat{\theta}-\theta_{0})=\begin{pmatrix}0\\ \frac{c_{0}(-E\varepsilon^{3},1)}{2\alpha}E\left[\left\{(\varepsilon_{t}-q_{\alpha})1(\varepsilon_{t}\leq q_{\alpha})+C^{\top}R_{2}(\varepsilon_{t})\right\}R_{2}(\varepsilon_{t})\right]\end{pmatrix},$$

so the conclusion regarding $\widehat{\chi}_{n,t}$ holds.

■

3.8.2 Second Step Updates by Newton Methods

The consistent estimators of θ (coefficient for GARCH process) that we use for the empirical study is from the first step QMLE. The reason that we do not go further to update them by the Newton-

Raphson method is because the people in the industry would perfer to work with simpler model than something that needs to use complex calculations. However, we still sketch the details if the updated is needed in practice.

In order to make sure the convergence is on the right direction hence provides a better estimator, the variable step length algorithm is used, which is similar as the BHHH methods. (BHHH is an acronym of the four originators: Berndt, B. Hall, R. Hall, and Jerry Hausman). The algorithms are iterative, defininng a sequence of approximatiosn, given by,

$$\widehat{\beta}_{k+1} = \widehat{\beta}_k + \lambda_k \left[\frac{1}{n} \sum\nolimits_{t=1}^{n} \widehat{l}^*_{kt}(\widehat{\beta}_k) \widehat{l}^*_{kt}(\widehat{\beta}_k)^\top \right]^{-1} \frac{1}{n} \sum\nolimits_{t=1}^{n} \widehat{l}^*_{kt}(\widehat{\beta}_k),$$

where $\widehat{\beta}_k$ is the parameter estimate at step k, and λ_k is a parameter (called step size) which partly determines the particular algorithm. For the BHHH algorithm , λ_k is determined by calculations within a given iterative step, involving a line-search until a point $\widehat{\beta}_{k+1}$ is found satisfying certain criteria. More information will be provided upon request.

Chapter 4

Semiparametric Value-at-Risk Forecasts for ARCH(∞) Models

4.1 Introduction

Despite the large empirical literature on the two-stage VaR estimation, there is rather sparse literature investigating the sampling properties of the proposed procedures. The statistical properties of the proposed VaR estimator is important because confidence intervals of the conditional VaR are very useful in setting up prudent capital reserve requirements for banks and conservative trading limits for traders. Christoffersen and Gonçalves (2005) give the following example to illustrate the importance of these confidence bands. Suppose a portfolio manager has a point estimate for the VaR of 13% and is capped with a VaR of up to 15% of current capital. If this is the only information available, the 13% point forecast indicates the portfolio is safe. Now suppose the manager is given the confidence band of 10%-16% for the VaR, he may decide to rebalance the portfolio. The major difficulty in exploring the large sample theory of two-stage VaR lies in the parameter uncertainty in volatility estimation. This parameter uncertainty complicates interval estimation of VaR since VaR estimation is based on the devolatized residuals instead of the true errors.

This chapter proposes a new VaR forecast method that is robust to heavy-tailed errors and to provide a complete asymptotic theory that acknowledges parameter uncertainty in volatility estimation. The proposed forecasts methods allow for a wide class of error distributions, including heavy-tailed ones. The existence of heavy-tailed financial time series is well documented and has recently received great attention. For example, Mittnik and Rachev (2000) and Rachev (2003) show that even after GARCH filtering, some residual time series are still heavy-tailed and far from normal.[1]

The existence of heavy-tailed errors poses challenges to volatility estimation. Many volatility estimators require at least a finite fourth moment.[2] For example, Weiss (1986) proves the consistency and asymptotic normality of the quasi-maximum likelihood estimator (QMLE) in the linear ARCH(q) model under fourth order moment assumptions on the ARCH process. Lee and Hansen (1994) weaken the moment conditions to existence of the fourth moment of the errors for GARCH processes.[3] In the presence of heavy-tailed errors, Hall and Yao (2003) show that the QMLE for parametric GARCH models suffers from complex limit distributions and slow convergence speed. The distribution of the subsequent quantile estimator based on the devolatized residuals therefore possesses unknown properties. This causes additional problems for the VaR estimators.

In the first step of the two-stage VaR forecasts, this paper proposes new volatility estimators to safeguard against heavy-tailed errors. The proposed volatility estimators employ least squares methods for log-transformed data. This volatility estimator asks for fewer moment conditions, thus allowing for a wider range of error distributions than QMLE. The reason is that the transformed errors, $\log(\varepsilon_t^2)$, have much thinner tails than original errors ε_t. Additionally, after the transformation, the regression problem becomes homoskedastic in stead of heteroskedastic.

In the second step, empirical quantile and extreme-value-theory based quantile are investigated. It is found that the parameter uncertainty changes the asymptotic variance of empirical quantile estimator.

The chapterr is organized as follows. First we discuss the model and the proposed forecast method in section 2. Then the asymptotic theory is presented in section 3. Extreme Value Theory based methods are discussed in section 4. Simulations and empirical studies are provided in section 5 and 6. Section 7 comes the conclusion and the potential extension.

4.2 The model and Value-at-Risk forecasts

This approach addresses Value-at-Risk forecast in a semiparametric multiplicative model

$$y_t = h_t^{1/2}(\beta)\varepsilon_t, \tag{4.1}$$

[1] See Bollerslev et al. (1992), Nolan (2001), Rachev and Fabozzi (2005) and Tsay (2005) for additional evidence.

[2] The consistency property holds under weaker moment conditions; it is the asymptotic normality that requires finite fourth moment. See Berkes and Horvath (2004) for general discussion.

[3] For more recent volatility estimators, see Härdle and Tsybakov (1997) for nonparametric models and Yang(2006) for semiparametric models.

where $\{\varepsilon_t\}$ is an i.i.d sequence of errors with unknown density $f(\cdot)$. The conditional scale of y_t conditioning on $\mathcal{F}_{t-1}$ is $h_t(\beta)$, and ε_t is independent of $\mathcal{F}_{t-1}$. We assume a general ARCH(∞) parametric structure for h_t

$$h_t(\beta) = G(c_0 + \sum_{j=1}^{\infty} c_j(\beta)\psi(y_{t-j}, \varepsilon_{t-j})), \tag{4.2}$$

where $G(\cdot)$ and $\psi(\cdot)$ are known functions, while c_0 and c_j are unknown finite-dimensional parameters. The link function $G(\cdot)$ is positive and invertible. Additionally,

$$c_0 > 0\ ; c_j(\beta) > 0,\ j \geq 1;\ \sup_{\beta} |c_j(\beta)| \leq c\rho^j, \rho < 1. \tag{4.3}$$

The structure (1)-(3) is very flexible and encompasses a wide classes of conditional volatility models. Examples are as follows[4]:

Example 1 *GARCH (Engle (1982), Bollerslev (1986)):* $h_t = c + \sum_{i=1}^{p} a_i y_{t-i}^2 + \sum_{j=1}^{q} b_j h_{t-j}$

$$G(x) = x; c_0 = c/(1 - \sum_{j=1}^{q} b_j); c_j = \sum_{i=1, i\leq j}^{p} a_i G^{j-i}(1,1), G = (\begin{matrix} b_1 & ... & b_q \\ I_{q-1} & ... & 0 \end{matrix}); \psi(x) = x^2$$

Example 2 *linear GARCH (Koenker and Xiao (2009)):* $h_t = c + \sum_{i=1}^{p} a_i |y_{t-i}| + \sum_{j=1}^{q} b_j h_{t-j}$

$$G(x) = x; c_0 = c/(1 - \sum_{j=1}^{q} b_j); c_j = \sum_{i=1, i\leq j}^{p} a_i G^{j-i}(1,1), G = (\begin{matrix} b_1 & ... & b_q \\ I_{q-1} & ... & 0 \end{matrix}); \psi(x) = |x|$$

Example 3 *leverage GARCH (Glosten, Jagannathan and Runkle (1993)):* $h_t = c + ay_{t-1}^2[1 + \gamma 1_{(y_{t-1}<0)}] + bh_{t-1}$

$$G(x) = x; c_0 = c; c_j = ab^{j-1}; \psi(x) = x^2 + \gamma x^2 1_{(x<0)}$$

Example 4 *GARCH-in-Mean model (Drost and Klaasens(1997)):* $h_t = c + ay_{t-1}^2 + bh_{t-1}; \varepsilon_t = \mu + \sigma\xi_t$

$$G(x) = x; c_0 = c/(1-b); c_j = ab^{j-1}; \psi(x) = x^2$$

Example 5 *IGARCH(1,1):* $h_t = c + (1-b)y_{t-1}^2 + bh_{t-1}$

$$G(x) = x; c_0 = c/(1-b); c_j = (1-b)b^{j-1}; \psi(x) = x^2$$

[4] Bollerslev(2008) has a comprehensive list of parametric GARCH models.

Example 6 *Quadratic GARCH (QGARCH) Sentana (1995)* $h_t = c + ay_{t-1}^2 + bh_{t-1} + \phi y_{t-1}$

$$G(x) = x; c_0 = c/(1-b); c_j = b^{j-1}; \psi(x) = ax^2 + \phi x$$

Example 7 *Stable GARCH(SGARCH/NGARCH) Liu and Brorsen(1995) Higgins and Bera(1992)* $h_t^\delta = c + ay_{t-1}^\delta + bh_{t-1}^\delta$

$$G(x) = x^{1/\delta}; c_0 = c; c_j = ab^{j-1}; \psi(x) = x^\delta$$

Example 8 *EGARCH(Nelson 1991). TGARCH(Zakoian 1994). NAGARCH(Engle and Ng(1993)), FGARCH(Hentschel(1995))*

$$\frac{h_t^{\lambda/2} - 1}{\lambda} = \varpi + ah_{t-1}^{\lambda/2} f^v(\varepsilon) + b\frac{h_{t-1}^{\lambda/2} - 1}{\lambda}$$

$$G(x) = x^{1/\lambda}; c_0 = \lambda\varpi + 1; c_j = ab^{j-1}; \psi(y, \varepsilon) = (\frac{y}{\varepsilon})^\lambda f^v(\varepsilon)$$

Our model is similar to the ARCH(∞) model considered in Robinson and Zaffaroni (2005). Robinson and Zaffaroni (2005) treats general QMLE estimation based on the Gaussian distribution, which is sensitive to distribution assumptions, especially heavy-tailed errors. Zaffaroni (2009) studies the Whittle estimation of EGARCH based on log squared returns. Similar log treatment appears in Kim, Shephard and Chib (1998) and further transformations are available. Weiss (1986) considered the least squares estimator based on the following

$$\begin{aligned} y_t^2 &= h_t(\beta) + \epsilon_t \\ \epsilon_t &= h_t(\beta)(\varepsilon_t^2 - 1), \end{aligned}$$

where ϵ_t is conditional heteroskedastic. The large sample theory requires the existence of 8th moment of y_t. Bose and Mukherjee (2003) considered a two stage least squares estimator for GARCH models but their first stage estimator still assumes strong moment conditions.

For all the above parametric GARCH models, the standard estimation method is QMLE, based on the assumption of normality of ε. Baillie and Bollerslev (1989) uses $student - t$ distribution to allow for more flexible parametric family of distributions. However, Drost and Klaassen (1997) and Newey and Steigerwald (1997) argues that QMLE based on nonnormal distributions generally fails to be consistent if the true distribution is different.

Various parametric GARCH models are proposed to accommodate different regularities found in financial data. Observe that model (1) is not changed when the unknown parameter β_0 and innovation ε_t are replaced by $\beta_0 c$ and innovation ε_t/c, respectively, for some positive c. Therefore, scale normalization is needed to make identification possible. For example, two common normalizations in volatility estimation literature are $E[\varepsilon_t^2] = 1$ and $Median[\varepsilon_t^2] = 1$.

In this paper we assume $E[\log(\varepsilon_t^2)] = 0$ for reasons to be explained later in this section. Note that we can always rewrite model 1 to have errors that satisfy $E[\log(\varepsilon_t^2)] = 0$: in case $E[\log(\varepsilon_t^2)] = K$, model 1 can be rewritten as $y_t = h_t^{1/2}(\beta_0)\exp(K/2)\varepsilon_t \exp(-K/2)$, where $E[\log((\varepsilon_t \exp(-K/2))^2)] = 0$. Furthermore, the nuisance parameter K doesn't affect the estimation of Value-at-Risk since $Q(y_{n+1}|\mathcal{F}_n) = h_{n+1}^{1/2}(\beta_0)q_\alpha = h_{n+1}^{1/2}(\beta_0)\exp(K/2)q_\alpha \exp(-K/2)$.

Given time series sample $\{y_t\}_{t=1}^n$, our forecast target is the parameter $h_{n+1}^{1/2}(\beta_0)q_\alpha$. This forecast is a combination of the volatility forecast $h_{n+1}^{1/2}(\widehat{\beta})$ and the error quantile estimation $\widehat{q}_\alpha$. Most of the literature employs QMLE for $\widehat{\beta}$ but this paper proposes a new estimation method for β_0. Given sample $\{y_0, y_1, ...y_t\}$, our Value-at-Risk forecasts proceeds with the following two steps:

(i) From the transformation $\log y_t^2 = \log h_t + \log \varepsilon_t^2$, we consider the least square estimation problem of

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{t=1}^{n}[\log y_t^2 - \log h_t(\beta)]^2. \tag{4.4}$$

(ii) Estimate the unconditional quantile $\widehat{q}_\alpha$ from the standardized residuals $\widehat{\varepsilon}_t = y_t/h_t^{1/2}(\widehat{\beta})$ from the following minimization problem

$$\widehat{q}_\alpha = \arg\min_{q} \frac{1}{n}\sum_{t=1}^{n}\{|\widehat{\varepsilon}_t - q| + (2\alpha - 1)(\widehat{\varepsilon}_t - q)\}. \tag{4.5}$$

(iii) The one-step ahead conditional VaR prediction at time n is

$$h_{n+1}^{1/2}(\widehat{\beta})\widehat{q}_\alpha. \tag{4.6}$$

We can see that this one-period-ahead VaR prediction involves volatility prediction $h_{n+1}^{1/2}(\widehat{\beta})$ and quantile estimation $\widehat{q}_\alpha$. The variation from this VaR prediction is from the sampling variation of $\widehat{\beta}$ and $\widehat{q}_\alpha$. Since we don't know the true parameters (β_0, q_α) and use estimators $(\widehat{\beta}, \widehat{q}_\alpha)$ in the prediction, this parameter uncertainty is the major source of prediction variation. As a result, we will study the joint distribution of $(\widehat{\beta}, \widehat{q}_\alpha)$ in the following section.

The volatility estimator in (2) doesn't rely on the assumption of a particular error distribution so it offers robustness to both error distribution misspecification and existence of heavy-tailed errors. After log transformation, the new model has error term $\log \varepsilon_t^2$ that is homoskedastic; the moment condition used for estimation is only $E[\log \varepsilon_t^2]^2 < \infty$. In few cases like multiplicative ARCH(p) models, one can even obtain a closed form solution to (i). In many cases, we have to solve the nonlinear optimization problem of (i). This objective function is smooth and has continuous derivatives for most specifications of $h_t(\beta)$. This feature offers a numerical advantage over the LAD optimization problem, which is common in robust statistics. We postpone our discussion regarding this moment condition to the next section.

One could instead do nonlinear quantile regression based on the following

$$\min_{\beta, q_\alpha} \frac{1}{n} \sum_{t=1}^{n} l_\alpha(y_t - h_t^{1/2}(\beta) q_\alpha), \tag{4.7}$$

where $l_\alpha(x) = |x| + (2\alpha - 1)x$. However, there are three issues: first, β and q_α are not jointly identifiable; second, the above nonlinear quantile estimation is hard to solve numerically; third, and more importantly, this one-step estimation method overlooks the moment conditions in (2), thus is less efficient. More intuitively, the model specifies the conditional quantile in a way that part of the finite dimensional parameters β does not vary with quantile level α. In contrast, this one-step regression does not assume β is constant across all quantile level α.

4.3 Asymptotic theory

Denote $\theta = (\beta', q_\alpha)'$ and

$$l_t(\beta) = [\log y_t^2 - \log h_t(\beta)] \frac{1}{h_t(\beta)} \frac{\partial h_t(\beta)}{\partial \beta}.$$

Before we show the asymptotic theory for the estimators $(\widehat{\beta}, \widehat{q}_\alpha)$, we discuss the following assumptions. For some $r' > r > 1$,

Assumptions B

- B1 *The process $\{y_t\}$ is strictly stationary and absolute regular with mixing coefficients ϱ_j such that $\sum_{j=1}^{\infty} j^{1/(r-1)} \varrho_j < \infty$.*
- B2 ε_t is i.i.d with continuously differentiable density $f(\cdot)$, $E[\log \varepsilon_t^2] = 0$, $E[\log \varepsilon_t^2]^{2r} < \infty$ and ε_t has αth quantile $q_\alpha < \infty$.
- B3 $rank(E[\frac{1}{h_t^2} \frac{\partial h_t}{\partial \beta} \frac{\partial h_t}{\partial \beta'}]) = \dim(\beta)$
- B4 $E[\sup_\beta |\frac{1}{h_t(\beta)} \frac{\partial h_t(\beta)}{\partial \beta}|^{2r'}] < \infty, E \sup_\beta |\frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r'} < \infty, E \sup_\beta |\frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'}|^{2r'} < \infty, E[\sup_\beta |\log \frac{h_t}{h_t(\beta)}|^{2r'r/(r'-r)}] < \infty$

B5 $E[\sup_\beta |h_t^{1/2}(\beta)|] < \infty$.

For assumption B1 to hold, a sufficient condition is that the mixing coefficients decay exponentially. A variety of time series models have been verified to be geometrically ergodic, which implies exponentially decaying mixing coefficients. For nonlinear homoskedastic autoregressive models, relevant results have been obtained in Bhattacharya and Lee (1995) and Lee (1997). Cline and Pu (1999), Cline and Pu (2004) and Liebscher (2005) have extended above results for ARCH type of heteroskedasticity. Recently, Carrasco and Chen (2002), Francq and Zakoïan (2006) and Kristensen (2007) have shown similar results hold for a large family of GARCH models.

The assumption $E[\log \varepsilon_t^2]^{2r} < \infty$ is considerably weaker than $E[\varepsilon_t^4] < \infty$, which is commonly assumed for volatility estimation. One sufficient condition is that the density is bounded on a compact set containing zero and the tails decay fast enough. To see this, without loss of generality, we consider the case that density is bounded in $\varepsilon \in [-1, 1]$ by K, i.e., $f(\varepsilon) \leq K$. Notice that any density must decay faster than x^{-1} at infinity. Suppose that $f(\varepsilon) = O(\varepsilon^{-a})$ as $\varepsilon \to \infty$ for some $a > 1$. As a result, we can see

$$\begin{aligned} E[(\log(\varepsilon_t^2))^{2r}] &= \int_{-\infty}^{-1} [\log(\varepsilon^2)]^{2r} f(\varepsilon) d\varepsilon + \int_{-1}^{1} [\log(\varepsilon^2)]^{2r} f(\varepsilon) d\varepsilon + \int_{1}^{\infty} [\log(\varepsilon^2)]^{2r} f(\varepsilon) d\varepsilon \\ &\leq 2C \cdot |\int_{\infty}^{0} \frac{t^{2r}}{2} e^{t/2} [e^{-at/2}] dt| + K \cdot \int_{-1}^{1} [\log(\varepsilon^2)]^{2r} d\varepsilon \\ &< \infty, \end{aligned}$$

where the last inequality holds because $\int_{-1}^{1} [\log(\varepsilon^2)]^{2r} d\varepsilon < \infty$.[5]

The condition $E[(\log(\varepsilon_t^2))^2] < \infty$ is satisfied by most commonly used distributions. For example, stable distributions, t distributions, etc. Table 29 reports that $E[(\log(\varepsilon_t^2))^2] < \infty$ can serve as a legitimate assumption for various t distributions:

Table 29. A comparison of moment conditions

	$t(1)$	$t(2)$	$t(3)$	$t(4)$	$t(5)$	$t(500)$
$E[(\varepsilon_t^4)]$	∞	∞	∞	∞	26	3
$E[(\log(\varepsilon_t^2))^2]$	9.89	7.07	6.69	6.58	6.55	6.55

Based on averages of 10^6 simulated samples from each distribution

For robustness consideration, one could assume instead $E[\log(\varepsilon_t^2)|\mathcal{F}_{t-1}] = 0$ and $E[1_{\{\varepsilon_t \leq q_\alpha\}} -$

[5] This is because $E[\log \varepsilon_t^2]^{2r} < \infty$ holds for uniformly distributed random variables ε. To see this, consider the moment generating function(MGF) for $\log \varepsilon^2$: $E[\exp\{t \cdot [\log \varepsilon^2]\}] = E[\varepsilon^{2t}]$, which is finite for all $t > -1/2$. Therefore the MGF of $\log \varepsilon^2$ exists and all moments exist.

$\alpha)|\mathcal{F}_{t-1}] = 0$. Our volatility estimation is still valid in this case although the asymptotic variance of the volatility estimator would be different. The condition $E[1_{\{\varepsilon_t \leq q_\alpha\}} - \alpha|\mathcal{F}_{t-1}] = 0$ is also called quantile independence assumption, see Manski (1988) and Chaudhuri (1997).

The assumption B4 has been verified by many authors for various GARCH models.

The extra information of i.i.d assumptions on ε_t can be exploited by constructing a more efficient estimator in the following additional steps. Let

$$\widehat{f}(\varepsilon) = \frac{1}{nh}\sum_{t=1}^{n} k(\frac{\varepsilon - \widehat{\varepsilon}_t}{h}),$$

where the estimated residuals $\widehat{\varepsilon}_t = y_t/h_t^{1/2}(\widehat{\beta})$. Then let

$$\widetilde{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{t=1}^{n}\{\log \widehat{f}(\frac{y_t}{h_t^{1/2}(\beta)}) - \log h_t^{1/2}(\beta)\}.$$

We do not pursue this here.

The following theorem characterizes the joint distribution of $\widehat{\beta}$ and $\widehat{q}_\alpha$. Let:

$$Q = \begin{pmatrix} -E[\frac{1}{h_t^2}\frac{\partial h_t}{\partial \beta}\frac{\partial h_t}{\partial \beta'}] & 0 \\ \frac{q_\alpha f(q_\alpha)}{2}E[\frac{1}{h_t}\frac{\partial h_t}{\partial \beta'}] & f(q_\alpha) \end{pmatrix}$$

$$\Omega = \begin{pmatrix} \Omega_\beta & \Omega_{\beta q} \\ \Omega'_{\beta q} & \Omega_q \end{pmatrix}$$

$$\begin{aligned}
\Omega_\beta &= E[\log \varepsilon_t^2]^2 \cdot E[\frac{1}{h_t^2}\frac{\partial h_t}{\partial \beta}\frac{\partial h_t}{\partial \beta'}]^{-1} \\
\Omega_{\beta q} &= -E[\frac{1}{h_t}\frac{\partial h_t}{\partial \beta}]E[\frac{1}{h_t^2}\frac{\partial h_t}{\partial \beta}\frac{\partial h_t}{\partial \beta'}]^{-1}\{\frac{q_\alpha E[\log \varepsilon_t^2]^2}{2} + \frac{E[\log \varepsilon_t^2(1\{\varepsilon_t \leq q_\alpha\} - \alpha)]}{f(q_\alpha)}\} \\
\Omega_q &= \frac{\alpha(1-\alpha)}{f^2(q_\alpha)} + \frac{q_\alpha E[\log \varepsilon_t^2(1\{\varepsilon_t \leq q_\alpha\} - \alpha)]}{f(q_\alpha)}E[\frac{1}{h_t}\frac{\partial h_t}{\partial \beta'}]E[\frac{1}{h_t^2}\frac{\partial h_t}{\partial \beta}\frac{\partial h_t}{\partial \beta'}]^{-1}E[\frac{1}{h_t}\frac{\partial h_t}{\partial \beta}] \\
&\quad + \frac{q_\alpha^2 E[\log \varepsilon_t^2]^2}{4}E[\frac{1}{h_t}\frac{\partial h_t}{\partial \beta'}]E[\frac{1}{h_t^2}\frac{\partial h_t}{\partial \beta}\frac{\partial h_t}{\partial \beta'}]^{-1}E[\frac{1}{h_t}\frac{\partial h_t}{\partial \beta}].
\end{aligned}$$

Theorem 5 *Suppose that Assumptions B1-5 hold. Then,*

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Q^{-1}\begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial \beta}\log \varepsilon_t^2 \\ 1\{\varepsilon_t \leq q_\alpha\} - \alpha \end{pmatrix} + o_p(1)$$

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to^d N(0, \Omega).$$

$$\sqrt{n}(\widehat{\theta}-\theta_0) \to^d N(0,\Omega).$$

Instead of considering least square estimation (LSE) as in (2), one can consider least absolute estimation (LAD) for the log-transformed model. This LAD type estimator has been proposed by Peng and Yao (2003) and further extended to semi-strong GARCH case by Linton et. al. (2009). Here, we compare their relative efficiency by computing the ratio of the two asymptotic variances

$$R_\Sigma = \frac{\Omega_{\beta_{LAD}}}{\Omega_{\beta_{LSE}}} = \frac{1}{E[(\log(\varepsilon_t^2))^2][f(1)+f(-1)]^2}.$$

To make the comparison clearer, we report the value of R_Σ for t distributions with various degree of freedoms:

Table 30. Relative deficiency ratio $R_\Sigma = \Omega_{\beta_{LAD}}/\Omega_{\beta_{LSE}}$

$t(1)$	$t(2)$	$t(3)$	$t(4)$	$t(5)$	$t(500)$
0.9979	0.9551	0.8737	0.8246	0.7919	0.6543

From the above table we can see that, for t distributions, the asymptotic variance of LSE is larger than that of LAD. The heavier the distribution is, the more efficient LAD is. In general, this efficiency comparison depends on the variance of $\log(\varepsilon_t^2)$ and the error density at ± 1. However, because the objective function in (i) is smooth, while that for LAD is nonsmooth, the LSE method might have numerical advantage over LAD. Furthermore, in terms of calculating confidence bands for the proposed volatility estimator or VaR estimator, the variance of the LSE is easier to estimate than that of LAD, because $\widehat{\Omega}_{\beta_{LAD}}$ involves density estimation at ± 1.

Let

$$x_{n+1} = \left(\frac{q_\alpha}{2}\frac{1}{h_{n+1}^{1/2}}\frac{\partial h_{n+1}}{\partial \beta'}, \sqrt{h_{n+1}}\right)'.$$

Theorem 6 *Suppose that Assumptions B1-5 hold. Then,*

$$(x_{n+1}'\Omega x_{n+1})^{-1/2}\sqrt{n}(h_{n+1}^{1/2}(\widehat{\beta})\widehat{q}_\alpha - h_{n+1}^{1/2}q_\alpha) \to^d N(0,1).$$

For example, in GARCH(1,1) models, the one period ahead forecast can be written as $h_{n+1} = \frac{c}{1-b} + a\sum_{j=0}^{\infty} b^j y_{n-j}^2$ and

$$\frac{\partial h_{n+1}}{\partial \beta} = \left(\frac{1}{1-b}, \sum_{j=0}^{\infty} b^j y_{n-j}^2, \frac{c}{(1-b)^2} + a\sum_{j=0}^{\infty} j b^{j-1} y_{n-j}^2\right)'$$

From theorem 6, we can construct the predictive confidence intervals for VaR. Specifically, we employ the moment counterparts:

$$
\begin{aligned}
\widehat{h}_{n+1} &= \frac{\widehat{c}}{1-\widehat{b}} + \widehat{a}\sum_{j=0}^{n} \widehat{b}^j y_{n-j}^2 \\
\widehat{\frac{\partial h_{n+1}}{\partial \beta'}} &= \left(\frac{1}{1-\widehat{b}}, \sum_{j=0}^{n} \widehat{b}^j y_{n-j}^2, \frac{\widehat{c}}{(1-\widehat{b})^2} + \widehat{a}\sum_{j=0}^{n} j\widehat{b}^{j-1} y_{n-j}^2 \right)' \\
\widehat{x}_{n+1} &= \left(\frac{\widehat{q}_\alpha}{2} \frac{1}{\widehat{h}_{n+1}^{1/2}} \widehat{\frac{\partial h_{n+1}}{\partial \beta'}}, \sqrt{\widehat{h}_{n+1}} \right)' \\
\widehat{\Omega} &= \begin{pmatrix} \widehat{\Omega}_\beta & \widehat{\Omega}_{\beta q} \\ \widehat{\Omega}'_{\beta q} & \widehat{\Omega}_q \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\Omega}_\beta &= \frac{1}{n}\sum_{t=1}^{n} [\log \widehat{\varepsilon}_t^2]^2 \cdot \{ \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t^2} \frac{\partial \widehat{h}_t}{\partial \beta} \frac{\partial \widehat{h}_t}{\partial \beta'} \}^{-1} \\
\widehat{\Omega}_{\beta q} &= -\frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t} \frac{\partial \widehat{h}_t}{\partial \beta} \{ \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t^2} \frac{\partial \widehat{h}_t}{\partial \beta} \frac{\partial \widehat{h}_t}{\partial \beta'} \}^{-1} \{ \frac{1}{n}\sum_{t=1}^{n} \frac{\widehat{q}_\alpha [\log \widehat{\varepsilon}_t^2]^2}{2} + \frac{1}{n}\sum_{t=1}^{n} \frac{\log \widehat{\varepsilon}_t^2 (1\{\widehat{\varepsilon}_t \le \widehat{q}_\alpha\} - \alpha)}{\widehat{f}(\widehat{q}_\alpha)} \} \\
\widehat{\Omega}_q &= \frac{\alpha(1-\alpha)}{\widehat{f}^2(\widehat{q}_\alpha)} + \frac{1}{n}\sum_{t=1}^{n} \frac{\widehat{q}_\alpha \log \widehat{\varepsilon}_t^2 (1\{\widehat{\varepsilon}_t \le \widehat{q}_\alpha\} - \alpha)}{f(\widehat{q}_\alpha)} \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t} \frac{\partial \widehat{h}_t}{\partial \beta'} \{ \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t^2} \frac{\partial \widehat{h}_t}{\partial \beta} \frac{\partial \widehat{h}_t}{\partial \beta'} \}^{-1} \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t} \frac{\partial \widehat{h}_t}{\partial \beta} \\
&\quad + \frac{1}{n}\sum_{t=1}^{n} \frac{\widehat{q}_\alpha^2 [\log \widehat{\varepsilon}_t^2]^2}{4} \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t} \frac{\partial \widehat{h}_t}{\partial \beta'} \{ \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t^2} \frac{\partial \widehat{h}_t}{\partial \beta} \frac{\partial \widehat{h}_t}{\partial \beta'} \}^{-1} \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\widehat{h}_t} \frac{\partial \widehat{h}_t}{\partial \beta}.
\end{aligned}
$$

The confidence interval with level α_0 for $h_{n+1}^{1/2}(\widehat{\beta})\widehat{q}_\alpha$ based on the asymptotic theory above is

$$
I_{n+1}^{\alpha_0} = (h_{n+1}^{1/2}(\widehat{\beta})\widehat{q}_\alpha - z_{\alpha_0}\sqrt{\widehat{x}'_{n+1}\widehat{\Omega}\widehat{x}_{n+1}/T}, h_{n+1}^{1/2}(\widehat{\beta})\widehat{q}_\alpha + z_{\alpha_0}\sqrt{\widehat{x}'_{n+1}\widehat{\Omega}\widehat{x}_{n+1}/T}),
$$

where z_{α_0} solves $\Pr(|N(0,1)| \le z_{\alpha_0}) = \alpha_0$.

4.4 Extreme Value Theory

Instead of approximate the error distribution by the empirical distribution of estimated errors as in (ii), we may employ extreme value theory to estimate the quantile q_α. This allows for extrapolation outside the range covered by the sample.

Suppose that $F(\cdot)$ is a heavy-tailed distribution in the sense that the tail distribution has a polynomial representation

$$
\lim_{\lambda \to \infty} \frac{1-F(\lambda x)}{1-F(\lambda)} = x^{-\gamma}, x > 0, \gamma > 0. \tag{4.8}
$$

Heavy-tails in market return distributions also have some behavioral origins (investor excessive optimism or pessimism leading to large market moves). Examples of heavy-tailed distributions are Pareto-like distributions, such as Pareto, Cauchy, Student-t, Burr and Stable distributions with exponent less than two. For any $\delta > \gamma$, the expectation $E[|X|^\delta]$ is infinite. Empirical studies frequently encounter time series with $\gamma \in (3,5)$, see for instance Embrechts et al. (1997, Page 330). This polynomial representation can be reexpressed as $1-F(x) = a(x)x^{-\gamma}$, where $\lim_{\lambda\to\infty} \frac{a(\lambda x)}{a(\lambda)} = 1$. For simplicity, we assume

$$\overline{F}(x) = 1 - F(x) = cx^{-\gamma}. \tag{4.9}$$

Note that for $\gamma > 0$, the choice of the scale $a(x)$ does not make a difference asymptotically.

There are two main methods for extreme values, Block Maxima and Threshold Exceedances.

4.4.1 Block Maxima

The main idea is we can divide the total observations of an i.i.d series into m block of size n, so then we have m block maxima and the true distribution of these maximas can be approximated by the Generalized Extreme Value (GEV) Distribution as long as total observation is large enough. The parameters can be then estimated by the Maximum likelihood. The addition assumption for this method to work is to require the underlying distribution of the data to be in the domain of attraction of an extreme value distribution. But the downside of it is the approach is wasteful of data.

Definition 1 *The distribution function of the (standard) GEV distribution is given by:*

$$F_\phi(x) = \begin{cases} \exp(-(1+\phi x)^{-1/\phi}), & \phi \neq 0 \\ \exp(-e^{-x}), & \phi = 0, \end{cases}$$

where $1 + \phi x > 0$

A location and scale parameter can be added in as $F_{\phi,\mu,\sigma}(x) = F_\phi((x-\mu)/\sigma)$, and μ and σ are location and scale parameters respectively.

4.4.2 Threshold Exceedances

A more efficient and practical method to estimate the extreme value is the so called Threshold Exceedance method. The interest centres to estimate is the tail index γ,which can be done by using the Generalized Pareto Distribution method (GPD) and the Hill method.

GPD Method

Definition 2 *Excess Distribution over the Threshold: Let $\{X\}$ be an random variable with distribution function F. The Excess Distribution over the threshold u is defined as*

$$F_u(x) = P(X - u \leq x \mid X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}.$$

The excess distribution can be fitted into the Generalized Pareto Distribution (GPD) .

Definition 3 *Generalized Pareto Distribution (GPD): The df of the GPD is given by*

$$G_{\phi,\omega}(x) = \begin{cases} 1 - (1 + \phi x/\omega)^{-1/\phi}, & \phi \neq 0 \\ 1 - \exp(-x/\omega), & \phi = 0, \end{cases}$$

where ϕ and ω are shape and scale parameters.

In the heavy-tail case, $\phi > 0$, and $G_{\phi,\omega}$ is the ordinary Pareto distribution with $\phi = 1/\gamma$, where γ is the tail index as stated above. Solving maximization of log-likelihood function yields a GPD model $G_{\widehat{\phi},\widehat{\omega}}$ for the excess distribution function, and hence we can get the tail index easily. However, how to choose the threshold u is a real difficulty and we will illustrate the graphic method in the later empirical study.

Hill Method

The Hill approach is a well-known method for estimating the tail thickness parameterss of heavy-tailed distribution. We take one-step further to estimate the EVT based VaR using the estimated residual $\{\widehat{\varepsilon}_t\}$:

1. First, take the transformation $\eta_t = -\widehat{\varepsilon}_t$ and take the k_T largest order statistics $\{\eta_{T,T-t}\}_{t=1}^{k_T}$ from $\eta_{T,1} \leq \cdots \leq \eta_{T,T}$; Consider the censored data $\{1_{\{\eta_t > \eta_{T,T-k_T}\}}, \max\{\eta_t, \eta_{T,T-k_T}\}\}_{t=1}^{T}$ with the following log-likelihood function

$$L(\gamma, c) = \sum_{t=1}^{T} [1_{\{\eta_t > \eta_{T,T-k_T}\}} \log(c\gamma\eta_t^{-\gamma-1}) + 1_{\{\eta_t \leq \eta_{T,T-k_T}\}} \log(1 - c\eta_{T,T-k_T}^{-\gamma})].$$

2. The maximum likelihood estimator $(\widehat{\gamma}, \widehat{c})$ is

$$\widehat{\gamma} = \{\frac{1}{k_T} \sum_{t=1}^{k_T} \log \frac{\eta_{T,T-t+1}}{\eta_{T,T-k_T}}\}^{-1} \quad ; \quad \widehat{c} = \frac{k_T}{T} \eta_{T,T-k_T}^{\widehat{\gamma}}$$

3. The residual quantile estimator is

$$\widehat{q}_\alpha = -(T\frac{\alpha}{k_T}\eta_{T,T-k_T}^{-\widehat{\gamma}})^{-1/\widehat{\gamma}} = -\eta_{T,T-k_T}(\frac{k_T}{\alpha T})^{\frac{1}{k_T}\sum_{t=1}^{k_T}\log\frac{\eta_{T,T-t+1}}{\eta_{T,T-k_T}}}$$

4. The conditional VaR estimator is given by

$$\widehat{\xi}_\alpha(\mathbf{x}) = \widehat{\sigma}(\mathbf{x})\widehat{q}_\alpha,$$

where volatility is estimated by the LSE method as defined in the previous section.

Choosing k_T encounter the same difficult as choosing the threshold parameter in the GPD methods. We will illustrate the graphic method in the empirical study as well.

EVT Asymptotic Theory

Denote $U(x)$ as the inverse function of $1/(1-F(x))$. Suppose there exists a function $A(t) \to 0$, as $t \to \infty$, such that

$$\lim_{t\to\infty}\frac{U(tx)/U(t) - x^{1/\gamma}}{A(t)} = x^{1/\gamma}\frac{x^\rho - 1}{\rho}$$

for some $\rho < 0$ and all $x > 0$.

Before showing the asymptotic theory for the conditional quantile $\widehat{q}_\alpha$, we list out some assumptions. Further details can be found in Hill (2013).

Assumptions C

C1 *Smoothness and Moments*

a. Let $\{\Im_t\}_{t\in\mathbb{Z}}$ be a sequence of σ-field that do not depend on θ and define $\mathcal{F}:= \sigma(\cup_{t\in\mathbb{Z}}\Im_t)$. $x_t(\theta)$ lies on a probability measure spaece $(\Omega, \mathcal{F}, P)$ and is $\Im_t-$measurable

b. $x_t(\theta)$ is stationary, ergodic and thrice continuously differentiable with $\Im_t-$measurable stationary and ergodic derivatives $g_t(\theta)$ and $h_t(\theta)$.

c. Each $\omega_t(\theta) \in \{x_t(\theta), g_{i,t}(\theta), h_{i,j,t}(\theta)\}$ is govened by a non-degenerate distribuion that is absolutely continuous with respect to Lebesgue measure, with uniformly bounded derivatives: $\sup_{\theta\in\Theta}\sup_{a\in\mathbb{R}}\|(\partial/\partial\theta)P(\omega_t(\theta)\le a)\| < \infty$ and $\sup_{\theta\in\Theta}\sup_{a\in\mathbb{R}}\|(\partial/\partial a)P(\omega_t(\theta)\le a)\| < \infty$. Further $E(\sup_{\theta\in\Theta}|\omega_t(\theta)|^\iota < \infty$ for some tiny $\iota > 0$

d. $\inf_{\theta\in\Theta} x_t(\theta) \succeq \delta$ *a.s.* for some $\delta > 0$

C2 Regular Variation and Fractile Bound

a. There exists a neighborhood $\aleph_0(\delta)$ such that

$$\lim_{a\to\infty} \sup_{\theta\in\aleph_0(\delta)} \left| \frac{a^{\gamma(\theta)}}{L(a,\theta)} P(x_t(\theta) > a) - 1 \right| = 0$$

The tail component $L(a,\theta)$ is slowly varying with reaminder in a,uniformly on Θ. Moreover, the tail index $\gamma(\theta)$ is locally bounded $inf_{\theta\in\aleph_0(\delta)}\gamma(\theta) \succeq 0$ and $sup_{\theta\in\aleph_0(\delta)}\gamma(\theta) < \infty$, and is twice differentiable with bounded derivatives and a Lipschitz first derivatives$\|(\partial/\partial\theta)\gamma(\theta)\| < \infty$, $\left\|(\partial/\partial\theta)^2\gamma(\theta)\right\| < \infty$, and $\left\|(\partial/\partial\theta)\gamma(\theta) - (\partial/\partial\theta)\gamma(\widetilde{\theta})\right\| \leq K \left\|\theta - \widetilde{\theta}\right\|$ for each $\theta, \widetilde{\theta} \in \aleph_0(\delta)$

b. $k_T \to \infty$ and $k_T = o(T/\ln(T))$.

C3 Mixing Condition

$\aleph_0(\delta)$ be the neighborhood of θ^0 defined in Assumption 7.a. Then $x_t(\theta)$ is a β-mixing for each $\theta \in \aleph_0(\delta)$ with summable coefficients.

C4 Plug in

The plug-in estimator must be satisfied: there exists a unique point $\theta^0 \in \Theta$ such that $k_T^{1/2}\ln(T)(\widehat{\theta}_T - \theta^0) = o_p(1)$

C5 $k_T \to \infty, k_T/T \to 0, \sqrt{k_T}A(T/k_T) \to 0, \frac{1}{A(T/k_T)T^{1/4}} \to 0, \log(\frac{k_T}{\alpha T})/\sqrt{k_T} \to 0$

Define the asymptotic variance

$$\sigma_{KT} = E(k_T^{1/2}(\widehat{\gamma}^{-1} - \gamma^{-1})^2$$

where $\widehat{\gamma} = \{\frac{1}{k_T}\sum_{t=1}^{k_T} \log \frac{\eta_{T,T-t+1}}{\eta_{T,T-k_T}}\}^{-1}$

Theorem 7 *Suppose that Assumptions C1-C5 hold. Then, $T \to \infty$,* ***we have***

$$\frac{\sqrt{k_T}\log((\widehat{q}_\alpha)/q_\alpha)}{\sigma_{KT}|\log(\frac{k_T}{T\alpha})|} \to^d N(0,1),$$

where q_α is the real residual quantile, and the other notations are consistent as above.

Remark: A nonparametric estimator of the asymptotic variance was proposed in Hill (2010). Under regular condition, the estimator is consistent,

$$\frac{\widehat{\sigma}^2_{KT}}{\sigma^2_{KT}} \to^p 1.$$

4.5 Simulations

A small simulation study is used to illustrate the accuracy of our proposed VaR forecasting model under the heavy-tailed situation. The data are generated by the following process

$$
\begin{aligned}
y_t &= h_t^{1/2}\varepsilon_t \\
h_t &= c + a y_{t-1}^2 + b h_{t-1} \\
\varepsilon_t &\sim {}^{iid}F(x)
\end{aligned}
$$

where we fix $c = 0.02$. The distributions we choose in the simulation study are Student-t distribution and skew-t distribution.

4.5.1 $F(x)$: Student-t Distribution

For each model, we simulate 200,400 and 800 times. We compute prediction errors: $h_{n+1}(\widehat{\beta})\widehat{q}_\alpha - h_{n+1}^{1/2}q_\alpha$ for all simulations and note APE as the median of absolute prediction error. In addition, we compute the bias of estimators for GARCH parameter b. The simulation results with different parameters are in the following tables:

Table 31: Median of Absolute Prediction Error and Bias of $\widehat{b}$ ($a = 0.4, b = 0.5$)

		APE			bias of $\widehat{b}$		
$v = 444$		QMLE	Log-LSE	Log-LAD	QMLE	Log-LSE	Log-LAD
	n=200	**0.0395**	0.0510	0.0525	**-0.0385**	-0.0667	-0.0853
	n=400	**0.0285**	0.0359	0.0368	**-0.0163**	-0.0271	-0.0370
	n=800	**0.0194**	0.0251	0.0257	**-0.0083**	-0.0138	-0.0162
$v = 3$							
	n=200	1.3676	0.8624	**0.7820**	-0.0171	**-0.0072**	-0.0666
	n=400	1.5326	**0.5914**	0.6409	-0.0187	**-0.0047**	-0.0810
	n=800	4.2580	**0.4131**	0.7568	-0.0927	**-0.0024**	-0.1535

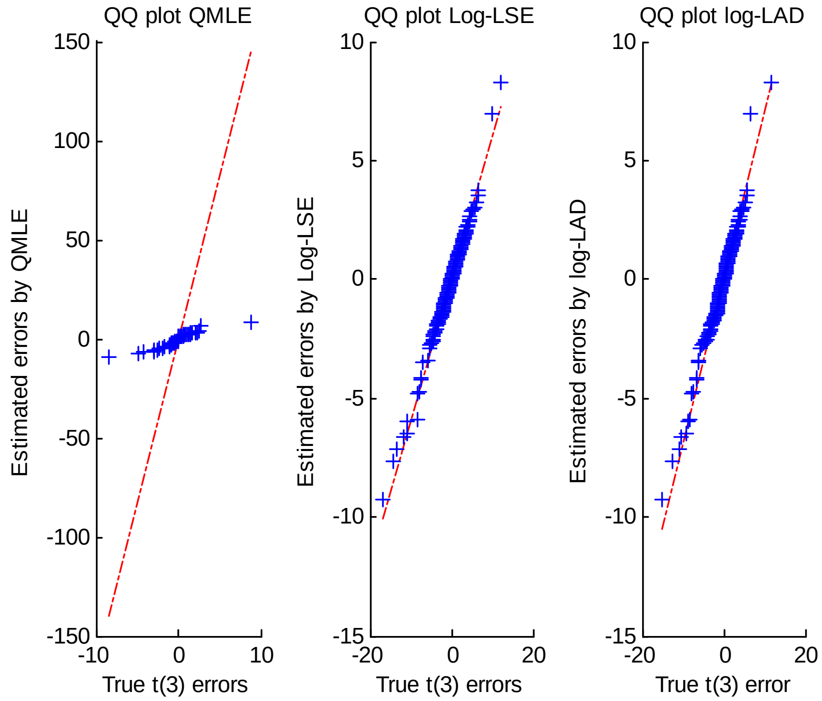
Table 32: Median of Absolute Prediction Error and Bias of $\widehat{b}$ ($a = 0.8, b = 0.1$)

	APE			bias of $\widehat{b}$		
$v = 444$	QMLE	Log-LSE	Log-LAD	QMLE	Log-LSE	Log-LAD
n=200	**0.0267**	0.0329	0.0340	**-0.0045**	0.0247	0.0170
n=400	**0.0194**	0.0228	0.0233	**-0.0072**	0.0118	0.0083
n=800	**0.0140**	0.0165	0.0169	-0.0073	0.0041	**0.0039**
$v = 3$						
n=200	0.2289	**0.1072**	0.1123	0.0386	0.0076	**0.0067**
n=400	0.2272	**0.0769**	0.0830	0.0658	0.0039	**0.0021**
n=800	0.2045	**0.0452**	0.0480	0.0614	**0.0011**	**0.0011**

Table 33:Median of Absolute Prediction Error and Bias of $\widehat{b}$ ($a = 0.1, b = 0.8$)

	APE			bias of $\widehat{b}$		
$v = 444$	QMLE	Log-LSE	Log-LAD	QMLE	Log-LSE	Log-LAD
n=200	**0.0481**	0.0700	0.0701	**-0.1550**	-0.3914	-0.4045
n=400	**0.0338**	0.0483	0.0421	**-0.0851**	-0.2936	-0.2890
n=800	**0.0261**	0.0374	0.0372	**-0.0239**	-0.1766	-0.1834
$v = 3$						
n=200	**0.3654**	0.3966	0.4186	**-0.0609**	-0.0792	-0.1056
n=400	0.3068	**0.2304**	0.2606	-0.0337	**-0.0179**	-0.0230
n=800	0.2674	**0.1480**	0.1553	-0.0118	**-0.0110**	-0.0123

Figure 13: QQ plot of the true student-t (3) distribution by QMLE, log-LSE and log-LAD methods



4.5.2 $F(x)$: Skewness-t Distribution (Hansen, 1994)

Table 34: Median of Absolute Prediction Error and Bias of $\widehat{b}$ ($a = 0.4, b = 0.5, lambda = 0.7, alpha = 0.05$)

	APE			bias of $\widehat{b}$		
$v = 15$	QMLE	Log-LSE	Log-LAD	QMLE	Log-LSE	Log-LAD
n=200	0.6304	**0.1542**	0.9865	-0.0805	**0.1272**	-0.3030
n=400	**0.0351**	0.0526	0.0625	0.0833	0.0960	**0.0682**
n=800	**0.0040**	0.0210	0.0316	0.0532	**-0.0103**	0.0405
$v = 3$						
n=200	**0.0159**	0.0279	0.0386	**-0.0150**	-0.1804	-0.2823
n=400	**0.0103**	0.0297	0.0216	0.0607	0.0145	**-0.0111**
n=800	0.0070	**0.0045**	0.0037	**-0.0445**	-0.0682	-0.0919

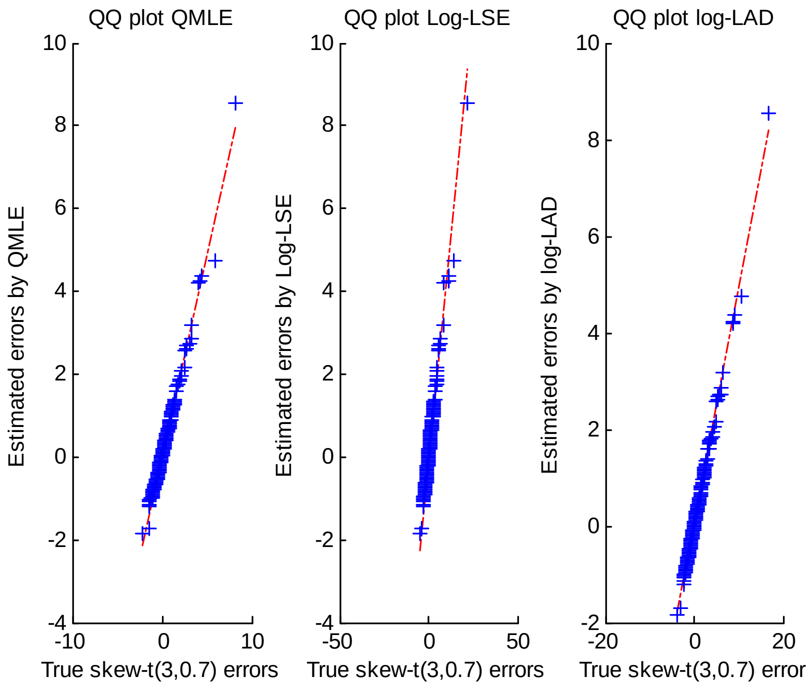
Table 35: Median of Absolute Prediction Error and Bias of $\widehat{b}$ ($a = 0.1, b = 0.8, lambda = 0.7, alpha = 0.05$)

		APE			bias of $\widehat{b}$		
$v = 3$		QMLE	Log-LSE	Log-LAD	QMLE	Log-LSE	Log-LAD
	n=200	0.0117	0.0057	**0.0006**	0.0759	**0.0527**	0.0626
	n=400	0.0319	0.0104	**0.0046**	-0.3167	-0.8000	**-0.1606**
	n=800	0.1512	0.0312	**0.0184**	-0.1611	-0.1533	**-0.0491**
$v = 15$							
	n=200	0.0399	0.0257	**0.0064**	-0.1465	**-0.1252**	-0.1733
	n=400	0.0281	0.0613	**0.0183**	**0.0490**	-0.5630	0.0791
	n=800	0.0085	**0.0008**	0.0039	-0.0464	**-0.0168**	-0.0583

Table 36:Median of Absolute Prediction Error and Bias of $\widehat{b}$ ($\alpha = 0.8, b = 0.1, lambda = 0.7, alpha = 0.05$)

		APE			bias of $\widehat{b}$		
$v = 3$		QMLE	Log-LSE	Log-LAD	QMLE	Log-LSE	Log-LAD
	n=200	0.0545	**0.0008**	0.0044	0.2503	**0.0171**	-0.1000
	n=400	0.0077	**0.0030**	0.0064	**-0.0535**	0.2118	0.3026
	n=800	**0.0144**	0.0205	0.0173	**-0.0749**	0.2243	0.1953
$v = 15$							
	n=200	0.0279	**0.0185**	0.0290	0.1042	0.0667	**0.0244**
	n=400	**0.0112**	0.0354	0.0307	-0.0380	**-0.0366**	-0.0790
	n=800	0.0091	**0.0020**	0.0032	-0.0862	**-0.0511**	-0.0635

Figure 14: QQ plot of true skew-t(3,0.7) distribution by QMLE, log-LSE and log-LAD methods



From the QQ plot of the true t(3) and skew-t(3,0.7) distribution, we can see that the performance of the three methods are quite similar when the errors are skew-t distributed, while log-LSE and log-LAD are significant better when the errors follow t distribution.

4.6 Empirical Study

Finally, we investigate whether our new proposed conditional VaR methods have good forecasting ability by comparing them with other conventional methods using index, individual company and exchange rate data. The advantage of the model is that it can be used in the situation with potential heavy-tailed errors. The datasets that we use here are MSCI(Emerging Market), S&P 500, IBM and GBP/USD exchange rate.

4.6.1 Descriptive Statistics

The four datasets that we use for our study are:

Table 37: Datasets

datasets	period	source
MSCI (Emerging Market)	01/01/1988-31/12/2013	Datastream
S&P 500	01/01/1990 – 31/12/2013	CRSP
IBM	01/01/2010 – 31/12/2013	CRSP
GBP/USD Exchange Rate	01/01/2010 – 31/12/2013	Federal Reserve Bank

Following table gives the descriptive statistics for the above datasets, the Ljung-Box test for autocorrelation and the KS test for normality.

Table 38: Descriptive Statistics of the datasets

	MSCI (EM)	S&P	IBM	FX
Mean	0	0.0004	0.0004	0
Standard deviation	0.0116	0.0115	0.0111	0.0053
Min	-0.0999	-0.0900	-0.0708	-0.0164
Max	0.1007	0.1151	0.0490	0.0235
Skewness	-0.5689	-0.0534	-0.6641	0.1538
Kurtosis	10.8039	11.6135	7.8300	3.4748

In all datasets the Ljung-Box test of returns 20-lags has p-value 0, In all datasets the Ljung-Box test of squared returns 20 lags has p-value 0, In all datasets, the Kolmogorov-Smirnov test for normality has p-value 1.

4.6.2 Models

There are totally eight models to be used in our empirical part. Model 1-7 are applying in the comparison of the forecasting performance and Model 8 represents the extreme value theory forecasting method.

Model 1-4 are the fundamental models that we used in the empirical study of last chapter, naming MA, EWMA, HS, GARCH(1,1).

Model 5 (YLS-our model): Under the model specification, the conditional variance is modelled by GARCH(1,1) and the conditional quantile is estimated by the empirical likelihood. The conditional Value-at-Risk of return series given $\mathcal{F}_{t-1}$ is,

$$\xi_t(\alpha) = \sum_{j=1}^{p} \rho_j y_{t-j} + h_t^{1/2} q_\alpha.$$

The probability of losses exceeding VaR, α, must be specified, with the most common probability level being 1% and 5%.

Model 6 (NCTPARCH): NCTPARCH denotes the noncentral-t distribution which is proposed by Krause and Paolella (2014). The evolution of the conditional variance is modeled flexibly by the APARCH model proposed by Ding et al. (1993) and the Value-at-Risk is

$$\xi_{t+1}(\alpha) = \sum_{j=1}^{p} \rho_j y_{t+1-j} + h_{t+1}^{1/2} q_\alpha.$$

ϵ_α,which represents the standardized error is an i.i.d noncentral-t distribution.

Model 7 (TWMIX): TWMIX is the time varying normal-mixture-GARCH type of models. The conditional distribution of the standardized error is assumed to be a mixed normal distribution with zero mean,

$$\epsilon_t \mid \mathcal{F}_{t-1} \sim MixNormal\ (\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}_t)$$

where $\boldsymbol{\lambda}$ is the vector of the mixing weights, $\boldsymbol{\mu}$ is the vector of location coefficient and $\boldsymbol{\sigma}_t$ is the vector of scale parameters.

Model 8(YLS-EVT): The extreme value theory VaR proposed in the chapter and the empirical results will be showing seperately in the next section.

$$\xi_t(\alpha) = \sum_{j=1}^{p} \rho_j y_{t-j} + h_t^{1/2} q_\alpha.$$

where the conditional quantile is estimated by extreme value theory.

4.6.3 Forecasting Performance

In the paper, we use three methods to evaluate the VaR model forecasing ability, including graph methods, violation ratio method and the White's reality check test.

VaR Forecasting

The following are the one-day VaR prediction using different data. We use the rolling window method with a 250 estimation window length and 0.01 significant level.

Figure 15: One day ahead Value-at-Risk prediction using EWMA and MA methods

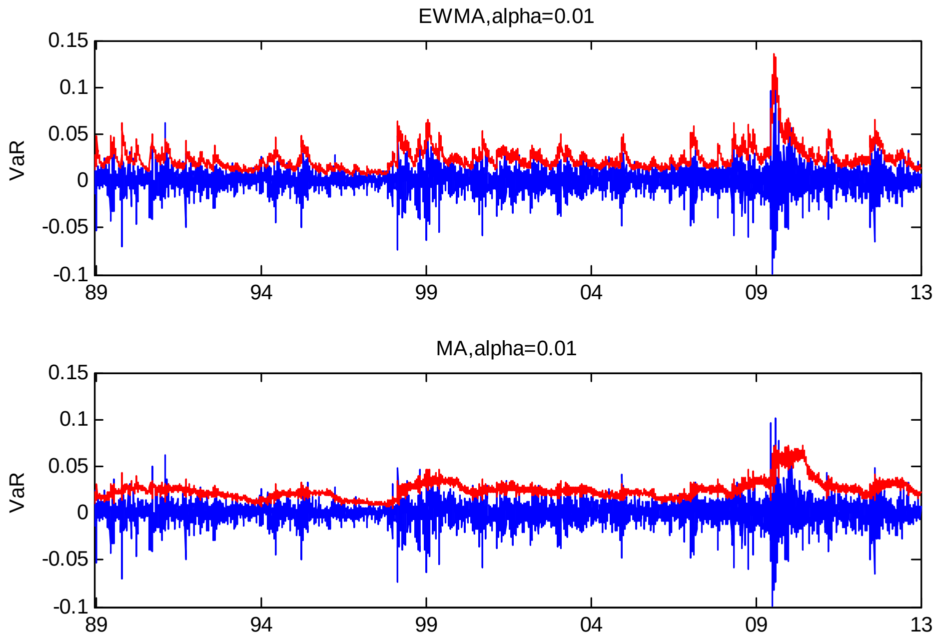


Figure 16: One day ahead Value-at-Risk prediction using HS and GARCH(1,1) methods

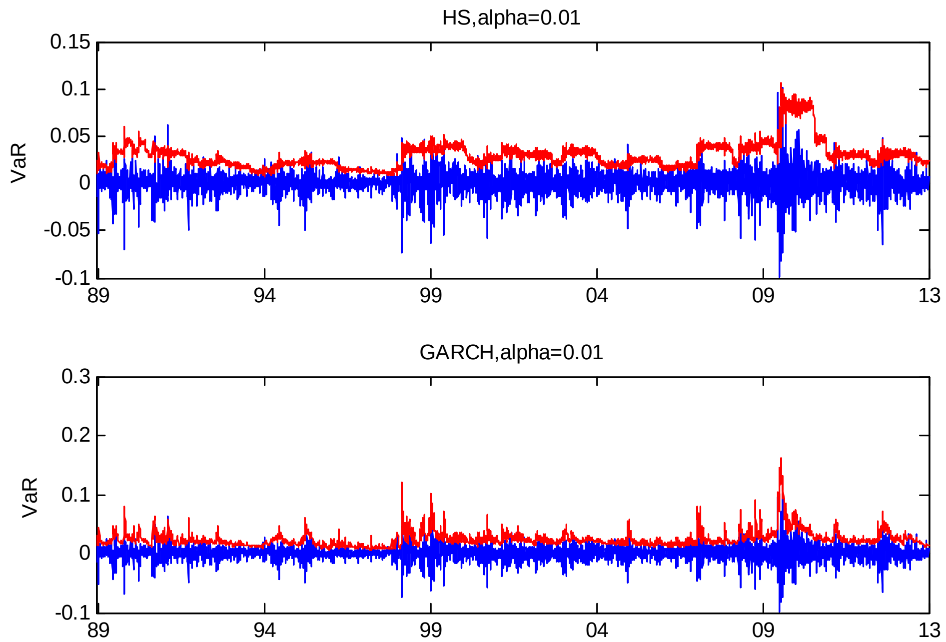


Figure 17:One day ahead Value-at-Risk prediction using YLS and NCTPARCH methods

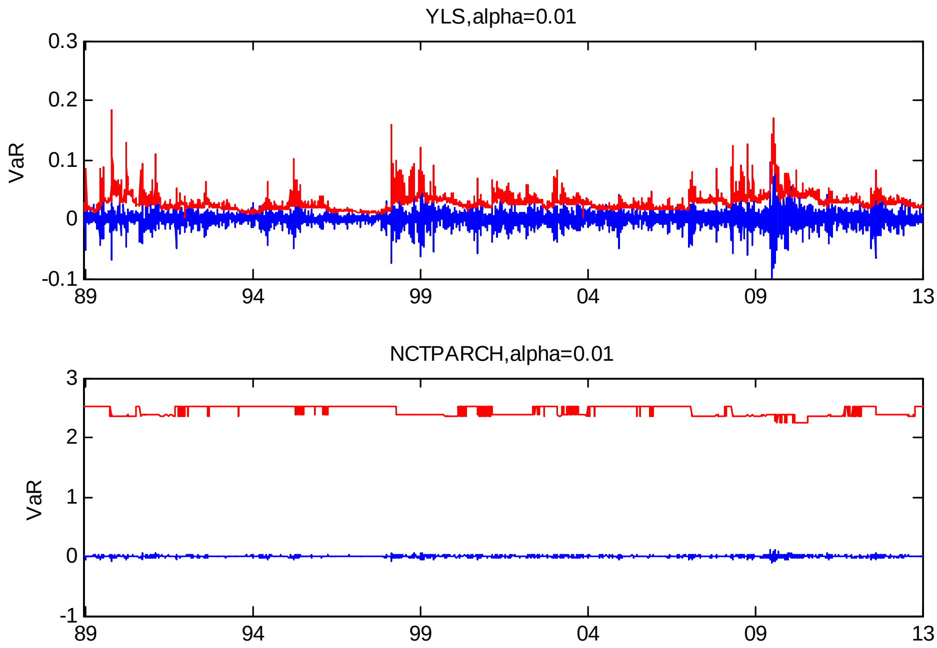
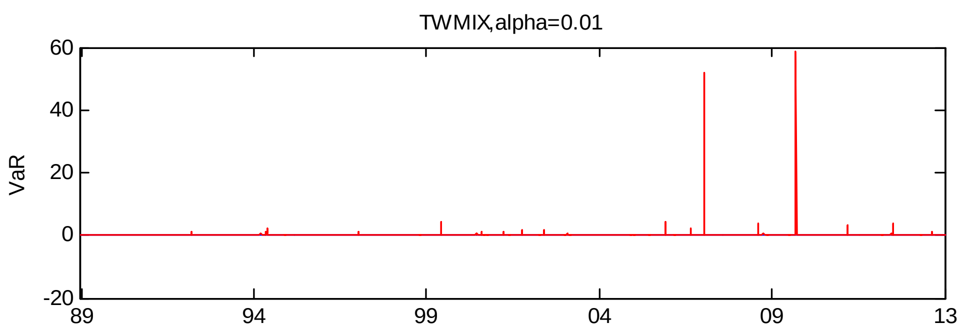


Figure 18: One day ahead Value-at-Risk prediction using TWMIX method



By simply observing the graphs, we may think EWMA, GARCH(1,1) and our method (YLS) have better forecasting ability.

Violation Ratio

We use Violation Ratio, Bernoulli Coverage Test and Independent Test to evaluate the performance of the VaR forecasting models:

Table 39: Violation Ratio (MSCI(EM) and S&P)

Model/data	MSCI(Emerging Market)		S&P 500	
	Violation Ratio	Volatility	Violation Ratio	Volatility
1.EWMA	2.0974	0.0123	1.8969	0.0132
2.MA	2.3117	0.0097	1.9486	0.0112
3.HS	1.3778	0.0140	1.4140	0.0151
4.GARCH(1,1)	1.8983	0.0129	1.9314	0.0132
5.YLS	**1.3472**	0.0140	**1.3968**	0.0128
6.NCTPARCH	0	0.0790	0	0.0765
7.TWMIX	1.7606	0.1673	1.3623	0.2731

Table 40: Bernoulli Coverage Test and Independent Test of MSCI(EM)

Model/data	MSCI (EM)			
	Bernoulli Coverage Test		Independent Test	
	Test Statistics	p-value	Test Statistics	p-value
1.EWMA	60.3845	0.0000	2.7295	**0.0985**
2.MA	Inf	0	33.0133	0.0000
3.HS	8.4264	0.0037	13.4784	0.0002
4.GARCH(1,1)	42.1380	0.0000	0.9942	**0.3187**
5.YLS	7.1744	0.0074	4.2847	**0.0385**
6.NCTPARCH	131.2976	0	NaN	NaN
7.TVMIX	31.1184	0.0000	0.0003	**0.9857**

Table 41: Bernoulli Coverage Test and Independent Test of S&P

Model/data	S&P 500			
	Bernoulli Coverage Test		Independent Test	
	Test Statistics	p-value	Test Statistics	p-value
1.EWMA	37.2988	0.0000	7.5755	0.0059
2.MA	41.2773	0.0000	12.6035	0.0004
3.HS	8.8981	0.0029	7.3055	0.0069
4.GARCH(1,1)	39.9329	0.0000	2.8533	**0.0912**
5.YLS	8.2093	0.0042	2.2018	**0.1378**
6.NCTPARCH	116.5638	0	NaN	NaN
7.TXMIX	6.9070	0.0086	2.3991	**0.1214**

We can see from Table 39 that the best model for both MSCI (EM) and S&P data are YLS, which has a closer violation ratio to the range 0.8-1.2. From table 40 and 41, we can see that for 0.01 significant level, both of the GARCH(1,1) and YLS pass the Bernoulli Coverage test and Independent test, while for 0.05 significan level, only GARCH(1,1) pass both tests.

Both of the MSCI(EM) and S&P data have more than 6000 observations , now we closely examine a shorter period of those datasets and also the IBM and Exchange Rate data which has a much shorter period. The reason is that we would like to test the methods in a much stable economics condition. The following are the subperiod of the data that we choose.

Table 42: Subperiod datasets

datasets	period	source
MSCI (Emerging Market)	01/01/2010 – 31/12/2013	Datastream
S&P 500	01/01/2010 – 31/12/2013	CRSP
IBM	01/01/2010 – 31/12/2013	CRSP
GBP/USD Exchange Rate	01/01/2010 – 31/12/2013	Federal Reserve Bank

Table 43: Violation Ratio of a shorter period (MSCI(EM), S&P)

Model	MSCI(EM) (sub)		S&P (sub)	
	Violation Ratio	Volatility	Violation Ratio	Volatility
1.EWMA	1.3889	0.0088	2.2487	0.0103
2.MA	2.0202	0.0057	1.9841	0.0064
3.HS	**1.2626**	0.0054	**0.9259**	0.0097
4.GARCH(1,1)	2.0202	0.0085	2.2487	0.0101
5.YLS	**1.2626**	0.0085	**0.9259**	0.0113
6.NCTPARCH	0	0.0756	0	0.0754
7.TWMIX	1.3889	0.3056	**1.0582**	0.0939

Table 44: Violation Ratio of a shorter period (IBM and Exchange Rate)

Model	IBM		Exchange Rate	
	Violation Ratio	Volatility	Violation Ratio	Volatility
1.EWMA	1.3245	0.0076	**1.0638**	0.0021
2.MA	1.8543	0.0028	**1.0638**	0.0014
3.HS	1.7219	0.0027	1.3298	0.0010
4.GARCH(1,1)	1.5894	0.0066	**1.0638**	0.0017
5.YLS	**1.1921**	0.0093	**1.1968**	0.0010
6.NCTPARCH	0	0.0761	0	0.0002
7.TWMIX	1.5894	0.5209	**1.0638**	0.0155

Table 45: Bernoulli Coverage Test and Independent Test of MSCI(EM) subperiod

Model/data	MSCI (EM) (sub)			
	Bernoulli Coverage Test		Independent Test	
	Test Statistics	p-value	Test Statistics	p-value
1.EWMA	1.0792	**0.2989**	0.3103	**0.5775**
2.MA	6.4259	0.0112	4.3081	0.0379
3.HS	0.5094	**0.4754**	8.0633	0.0045
4.GARCH(1,1)	6.4259	0.0112	4.3081	0.0379
5.YLS	0.5094	**0.4754**	8.0633	0.0045
6.NCTPARCH	15.9197	0.0001	NaN	NaN
7.TVMIX	1.0792	0.2989	2.1978	0.1382

Table 46: Bernoulli Coverage Test and Independent Test of S&P 500 subperiod

Model/data	S&P 500 (sub)			
	Bernoulli Coverage Test		Independent Test	
	Test Statistics	p-value	Test Statistics	p-value
1.EWMA	8.7912	0.0030	0.7833	**0.3761**
2.MA	5.7496	0.0165	1.0861	**0.2973**
3.HS	0.0430	**0.8358**	0.1310	**0.7174**
4.GARCH(1,1)	8.7912	0.0030	0.7833	**0.3761**
5.YLS	0.0430	**0.8358**	0.1310	**0.7174**
6.NCTPARCH	15.1961	0.0001	NaN	NaN
7.TWMIX	0.0254	**0.8734**	0.1714	**0.6789**

Table 47: Bernoulli Coverage Test and Independent Test of IBM

Model/data	IBM			
	Bernoulli Coverage Test		Independent Test	
	Test Statistics	p-value	Test Statistics	p-value
1.EWMA	0.7288	**0.3933**	0.2688	**0.6041**
2.MA	4.4461	0.0350	5.1464	0.0233
3.HS	3.2683	0.0706	5.7255	0.0167
4.GARCH(1,1)	2.2472	**0.1339**	0.3882	**0.5333**
5.YLS	0.2650	**0.6067**	0.2175	**0.6410**
6.NCTPARCH	15.1760	0.0001	NaN	NaN
7.TVMIX	2.2472	**0.1339**	0.3882	**0.5333**

Table 48: Bernoulli Coverage Test and Independent Test of Exchange Rate

model/data	Exchange Rate			
	Bernoulli Coverage Test		Independent Test	
	Test Statistics	p-value	Test Statistics	p-value
1.EWMA	0.0303	**0.8618**	0.1723	**0.6781**
2.MA	0.0303	**0.8618**	0.1723	**0.6781**
3.HS	0.7486	**0.3869**	0.2699	**0.6034**
4.GARCH(1,1)	0.0303	**0.8618**	0.1723	**0.6781**
5.YLS	0.2768	**0.5988**	0.2183	**0.6403**
6.NCTPARCH	15.1157	0.0001	NaN	NaN
7.TWMIX	0.0303	**0.8618**	0.1723	**0.6781**

From table 43 and 44, we can see that YLS is the best model in terms of the results of the violation ratio. The violation ratios of YLS of different data are all within the satisfactory range. In addition, we also find that exchange rate data is quite indifference to the different methods. All the methods have a violation ration in the range except the Historical Simulation.

From table 45-48, YLS still performs best for the Bernoulli Coverage Test and Independent Test. For 5% significant level, YLS pass both test for almost all the data except the Independent Test for MSCI(EM) subperiod data (with a p-value 0.0045). Seperately, for MSCI(EM) subperiod data, we can see that EWMA passes both test. For S&P subperiod data, HS is the one who passes both tests besides YLS. For IBM data, EWMA, GARCH(1,1), YLS and TVMIX all pass the Bernoulli Coverage Test and Independent Test. For exchange rate data, all of the methods pass both tests except NCTPARCH.

So we can conclude that YLS is the best methods evaluating by violation ratio criteria.

White's Reality Check (RC)

When we evaluate a model's forecasting performance, it is very important to check that the satisfactory results obtained are due to the model's actual forecasting ability, not because of chance. As mentioned in White's paper (2000), the problem of data snooping may occur when the researchers use the same dataset more than once for purpose of inference or model selection. He proposes a simple test called "Reality Check" to identify the issue. Diebold and Mariano (1995) constructs standard approach to compare the predictive performance of models in pairs. The White's test extends the method to a joint test, with the null hypothesis being that the best model is no better than the benchmark.

$$H_0 : \max_{k=1,2,...m} E(f) <= 0$$

where $f_{k,t} = L_{t,m} - L_{t.0}$, the difference of the loss between the alternative and the benchmark. The test requires to choose the loss function at the first place and in our paper, we use the following two loss functions:

1. Mean Square Errors(MSE):

$$l_{i,t} = (x_t - VaR_{j,t})^2$$

2. Absolute Error:

$$l_{i,t} = |x_t - VaR_{j,t}|$$

The following are the results of the reality check. In the analysis, we choose YLS as the benchmark and all the other models as the alternatives. We can't reject the null hypothesis by using different data and different loss fucntion. The result reassure that the satisfactory forecasting performance of our model is due to the actual good forecasting ability.

Table 49: White's Reality Check

data/loss function	Mean Square Error	Absolute Error
	Reality Check p-value	Reality Check p-value
MSCI (EM)	0.4900	0.1830
S&P	0.5250	0.2940
MSCI (EM) (sub)	0.5040	0.3910
S&P (sub)	0.5090	0.2700
IBM	0.7350	0.5660
FX	0.7820	0.2980

Extreme Value Theory Value-at-Risk (YLS-EVT)

Events such as market crashes or cases of individual financial distress regularly point out the potential effects of fat tails in unconditional return distributions. Empirical research in finance aims at a careful modeling of such extreme events and at the same time provides a basis for financial risk management.

Estimation of the tail thickness parameter is the subject of a large and active literature. Koedijk, Schafgans and de Vries (1990), Hols and de Vries (1991) and Wagner and Marsh (2005) showed the advantages of modeling fat-tailed distributions of exchange rate changes. Stock returns are known to have heavy tails following the work of Osborne (1959), Mandelbrot (1963), Fama(1965, 1976) and Markowitz (1991). The approach begins with choosing an estimator for the tail index parameter , the most common being the Hill estimator. The appeal of this estimator derives from its conceptual and computational simplicity.

In practice, the Threshold Exceedance methods are superior to the Block Maxima due to it's capacity of using all the data in the extreme in the sense that they exceed a certain high designated level.

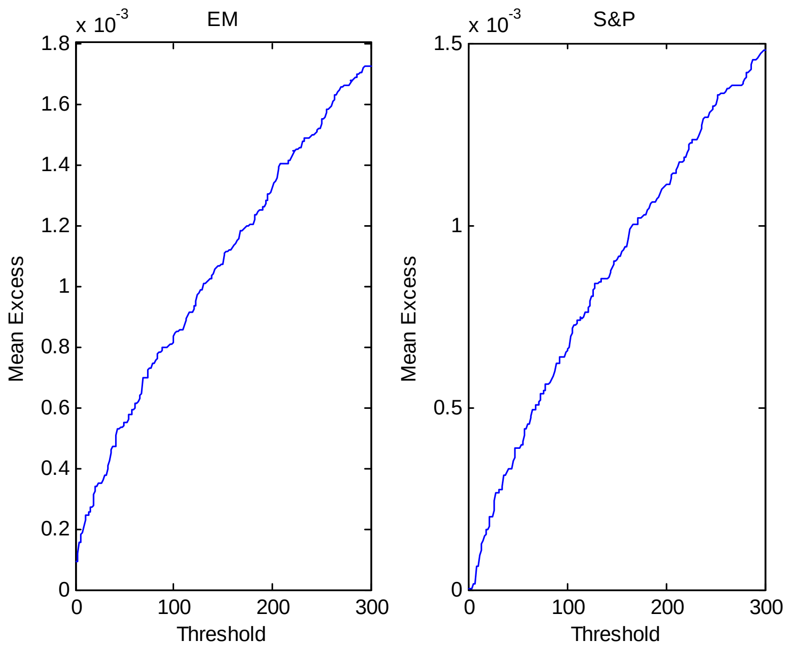
GDP Method First, we need to define the mean excess function.

Definition 4 *Mean Excess Function: The mean excess function of an random variable X with finite mean is:*

$$m(u) = E(X - u \mid X > u)$$

where u is the chosen threshold.

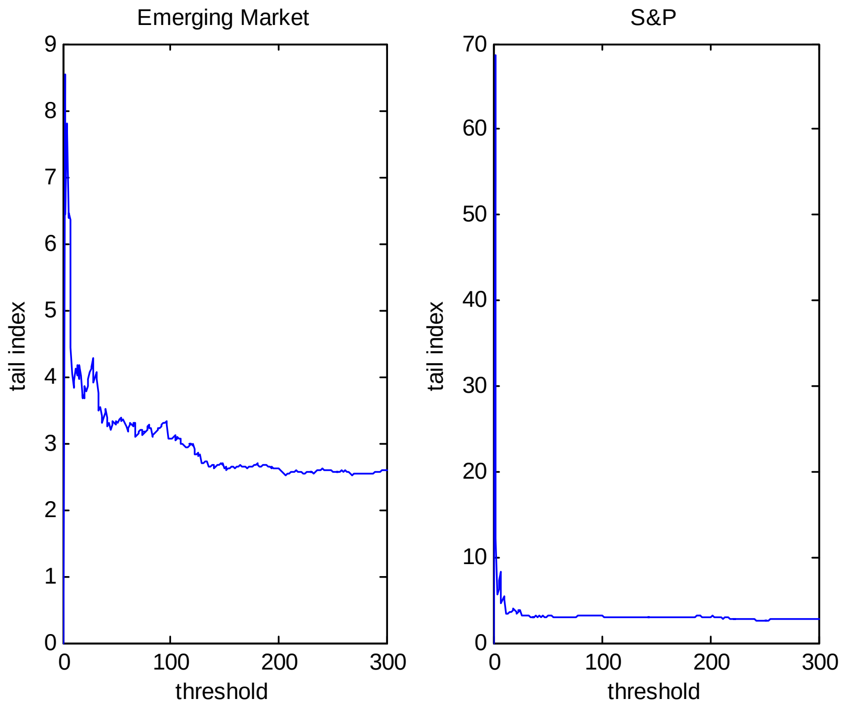
Figure 19: Sample Mean Excess Function Plot



Generally speaking, the mean excess function is linear with a higher threshold $\nu \geq u$, and this property can be used as a diagnostic when data follows a GPD model for the excess distribution.A linear upward trend indicate a GPD model with a positive shape parameter ($\phi > 0$), a linear downward trend is a GPD model with negative shape parameter ($\phi < 0$), and a horizontal line means that the shape parameter is zero ($\phi < 0$, and a exponential excess distribution in this case).

From the Figure 19, we can see that the mean excess function of both datasets are quite 'linear' over the entire period and the upward trends demonstrate the positive shape parameter ($\phi > 0$) to both datasets. However, it is not so easy to decide the threshold level from the plots.

Hill Method Figure 20: Hill Plot



From the Hill plot for all possible values of threshold, we can see that the tail thicknesses (γ) for the EM data and S&P data are stable at around (2.6,3) and (3,4), suggesting ϕ,the shape paramete is (0.33, 0.38) and (0.25, 0.33).Both can be interpreted as the infinite-kurtosis model for the data.

EVT backtesting In order to compare the VaR forecasting performance between EVT-based quantile models and the normal quantile models, we choose the S&P 500 data from 2007 to 2010, totally 1008 observations. This period represents the most volatile time during the last financial crisis. The models that we choose in the study can be decomposed into two parts: the methods estimating the conditional volatility and the methods modelling the conditional quantile. Here are the details:

Table 50:

Model	method of conditional volatility	method ofconditional quantile
GARCH-Hill	GARCH	Hill approach
YLS-Hill	YLS (the method proposed in the paper)	Hill approach
YLS-Normal	YLS	Normal quantile methods

And here shows the violation ratio of the three models:

Table 51: Backtesting Result (Violation Ratio)

Model	α=0.05		α=0.01		α=0.005	
	Violation Ratio	Volatility	Violation Ratio	Volatility	Violation Ratio	Volatility
GARCH-Hill	8.3951	0.0043	2.7778	0.0804	0.3086	0.4300
YLS-Hill	8.7963	0.0038	8.7963	0.0325	**1.2346**	0.1740
YLS-Normal	**0.8642**	0.0130	**1.8519**	0.0163	1.8519	0.0198

We can see that YLS-Normal is the best model when the quantile level is 5% and 1%. While when it goes to more extreme quantile, 0.5% for instance, YLS-Hill is the best methods.

4.7 Conclusion

The paper first proposed an alternative method to estimate GARCH parameters and hence the Value at Risk. The least square estimation based method imposes weak moment conditions on the errors and consequently, it has better prediction performance than commonly used QMLE-based VaR methods in the presence of non-normal errors. An EVT-VaR model is also introduced by applying the log-transformation in the GARCH estimation and EVT approach in the conditional quantile model. Asymptotic theory of both methods are provided.

Expected shortfall (ES) is an alternative risk measure and a proper discussion of ES by using these approaches could be a potential extention.

4.8 Appendix

4.8.1 Lemmas

Lemma 1 $E[[\log y_t^2 - \log h_t(\beta)]\frac{\partial \log h_t(\beta)}{\partial\beta}]^{2r} < \infty$

Proof. Notice that

$$\sup_\beta |[\log\varepsilon_t^2 + \log h_t - \log h_t(\beta)]\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r} \le \sup_\beta |[\log\varepsilon_t^2 + \log h_t - \log h_t(\beta)]|^{2r} \sup_\beta |\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}.$$

Therefore,

$$\begin{aligned}
& E[[\log y_t^2 - \log h_t(\beta)]\frac{\partial \log h_t(\beta)}{\partial\beta}]^{2r} \\
< \ & E[\sup_\beta |[\log y_t^2 - \log h_t(\beta)]\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}] \\
= \ & E[\sup_\beta |[\log\varepsilon_t^2 + \log h_t - \log h_t(\beta)]\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}] \\
= \ & E[\sup_\beta |(\log\varepsilon_t^2)\frac{\partial \log h_t(\beta)}{\partial\beta} + (\log\frac{h_t}{h_t(\beta)})\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}] \\
\le \ & \{[E\sup_\beta |\log\varepsilon_t^2 \frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}]^{1/2r} + [E\sup_\beta |\frac{\partial \log h_t(\beta)}{\partial\beta}\log\frac{h_t}{h_t(\beta)}|^{2r}]^{1/2r}\}^{2r} \\
= \ & I_1 + I_2
\end{aligned}$$

the last inequality is due to Minkowski's Inequality. The rest pof this proof is to show that $I_1 < \infty$ and $I_2 < \infty$.

Since $\log\varepsilon_t^2$ is independent of $\frac{\partial \log h_t(\beta)}{\partial\beta}$, it follows that

$$\begin{aligned}
I_1 &= \{E[\sup_\beta |\log\varepsilon_t^2 \frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}]\}^{1/2r} \\
&= \{E[(\log\varepsilon_t^2)^{2r}]E[\sup_\beta |\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}]\}^{1/2r} < \infty
\end{aligned}$$

on the other hand, because of Holder's Inequality,

$$\begin{aligned}
E[\sup_\beta |\frac{\partial \log h_t(\beta)}{\partial\beta}\log\frac{h_t}{h_t(\beta)}|^{2r}] &\le E[\sup_\beta |\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r}\sup_\beta |\log\frac{h_t}{h_t(\beta)}|^{2r}] \\
&\le \{E[\sup_\beta |\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2rp}]\}^{1/p}\{E[\sup_\beta |\log\frac{h_t}{h_t(\beta)}|^{2rp/(p-1)}]\}^{1-1/p} \\
&= \{E[\sup_\beta |\frac{\partial \log h_t(\beta)}{\partial\beta}|^{2r'}]\}^{r/r'}\{E[\sup_\beta |\log\frac{h_t}{h_t(\beta)}|^{2r'r/(r'-r)}]\}^{1-r/r'} \\
&< \infty
\end{aligned}$$

where $r' = rp$ and $p > 1$, so

$$I_2 = [E \sup_\beta |\frac{\partial \log h_t(\beta)}{\partial \beta} \log \frac{h_t}{h_t(\beta)}|^{2r}]^{1/2r} < \infty$$

■

Lemma 2 $E[\sup_{|\beta-\beta_1|<\delta} |l_t(\beta) - l_t(\beta_1)|^{2r}] \leq C\delta, C < \infty$

Proof. First, the derivative is

$$\begin{aligned} \frac{\partial l_t(\beta)}{\partial \beta} &= [\log \varepsilon_t^2 + \log \frac{h_t}{h_t(\beta)}][\frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'} - \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}] \\ & -\frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'} \end{aligned}$$

We want to show that

$$E[\sup_\beta |\frac{\partial l_t(\beta)}{\partial \beta}|^{2r}] < \infty$$

$$\begin{aligned} & \sup_\beta |\frac{\partial l_t(\beta)}{\partial \beta}|^{2r} \\ = & \sup_\beta |[\log \varepsilon_t^2 + \log \frac{h_t}{h_t(\beta)}][\frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'} - \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}] - \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r} \\ \leq & \sup_\beta |[\log \varepsilon_t^2 + \log \frac{h_t}{h_t(\beta)}][\frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'} - \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}]|^{2r} + \sup_\beta |\frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r} \\ \leq & \sup_\beta |[\log \varepsilon_t^2 + \log \frac{h_t}{h_t(\beta)}] \frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'}|^{2r} + \sup_\beta |[\log \varepsilon_t^2 + \log \frac{h_t}{h_t(\beta)}] \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r} \\ & + \sup_\beta |\frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r} \end{aligned}$$

so

$$\begin{aligned} & E \sup_\beta |\frac{\partial l_t(\beta)}{\partial \beta}|^{2r} \\ \leq & E \sup_\beta |\log \varepsilon_t^2 \frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'}|^{2r} + E \sup_\beta |\log \frac{h_t}{h_t(\beta)} \frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'}|^{2r} + E \sup_\beta |\log \varepsilon_t^2 \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r} \\ & + E \sup_\beta |\log \frac{h_t}{h_t(\beta)} \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r} + E \sup_\beta |\frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}|^{2r} \\ = & I_3 + I_4 + I_5 + I_6 + I_7 \end{aligned}$$

it's easy to see

$$\begin{aligned} I_3 &= E\sup_\beta |\log\varepsilon_t^2 \frac{1}{h_t(\beta)}\frac{\partial^2 h_t(\beta)}{\partial\beta\partial\beta'}|^{2r} = E|\log\varepsilon_t^2|^{2r} E\sup_\beta |\frac{1}{h_t(\beta)}\frac{\partial^2 h_t(\beta)}{\partial\beta\partial\beta'}|^{2r} \\ I_5 &= E\sup_\beta |\log\varepsilon_t^2 \frac{1}{h_t^2(\beta)}\frac{\partial h_t(\beta)}{\partial\beta}\frac{\partial h_t(\beta)}{\partial\beta'}|^{2r} = E|\log\varepsilon_t^2|^{2r} E\sup_\beta |\frac{1}{h_t^2(\beta)}\frac{\partial h_t(\beta)}{\partial\beta}\frac{\partial h_t(\beta)}{\partial\beta'}|^{2r} \\ I_7 &= E\sup_\beta |\frac{1}{h_t^2(\beta)}\frac{\partial h_t(\beta)}{\partial\beta}\frac{\partial h_t(\beta)}{\partial\beta'}|^{2r} \end{aligned}$$

we can show

$$\begin{aligned} I_4 &= E\sup_\beta |\log\frac{h_t}{h_t(\beta)}\frac{1}{h_t(\beta)}\frac{\partial^2 h_t(\beta)}{\partial\beta\partial\beta'}|^{2r} \leq E[\sup_\beta |\frac{1}{h_t(\beta)}\frac{\partial^2 h_t(\beta)}{\partial\beta\partial\beta'}|^{2r}\sup_\beta |\log\frac{h_t}{h_t(\beta)}|^{2r}] \\ &\leq \{E[\sup_\beta |\frac{1}{h_t(\beta)}\frac{\partial^2 h_t(\beta)}{\partial\beta\partial\beta'}|^{2rp}]\}^{1/p}\{E[\sup_\beta |\log\frac{h_t}{h_t(\beta)}|^{2rp/(p-1)}]\}^{1-1/p} \\ &= \{E[\sup_\beta |\frac{1}{h_t(\beta)}\frac{\partial^2 h_t(\beta)}{\partial\beta\partial\beta'}|^{2r'}]\}^{r/r'}\{E[\sup_\beta |\log\frac{h_t}{h_t(\beta)}|^{2r'r/(r'-r)}]\}^{1-r/r'} < \infty \end{aligned}$$

the same applies to

$$\begin{aligned} I_6 &= E\sup_\beta |\log\frac{h_t}{h_t(\beta)}\frac{1}{h_t^2(\beta)}\frac{\partial h_t(\beta)}{\partial\beta}\frac{\partial h_t(\beta)}{\partial\beta'}|^{2r} \\ &\leq \{E[\sup_\beta |\frac{1}{h_t^2(\beta)}\frac{\partial h_t(\beta)}{\partial\beta}\frac{\partial h_t(\beta)}{\partial\beta'}|^{2r'}]\}^{r/r'}\{E[\sup_\beta |\log\frac{h_t}{h_t(\beta)}|^{2r'r/(r'-r)}]\}^{1-r/r'} < \infty \end{aligned}$$

■

Lemma 3 $E[\sup_{\theta:|\theta-\theta_1|<\delta}|h_t^{1/2}(\beta)q_\alpha - h_t^{1/2}(\beta_1)q_{\alpha 1}|] \leq C\delta, C < \infty$

Proof. First, the derivative of $h_t^{1/2}(\beta)q$ is

$$\frac{\partial h_t^{1/2}(\beta)q}{\partial\beta} = \frac{q}{2h_t^{1/2}(\beta)}\frac{\partial h_t(\beta)}{\partial\beta}; \frac{\partial h_t^{1/2}(\beta)q}{\partial q} = h_t^{1/2}(\beta)$$

We want to show that

$$E[\sup_\theta |\frac{\partial h_t^{1/2}(\beta)q}{\partial\beta}|] < \infty; E[\sup_\theta |\frac{\partial h_t^{1/2}(\beta)q}{\partial q}|] < \infty$$

that is

$$\begin{aligned} E[\sup_\theta |\frac{q}{2h_t^{1/2}(\beta)}\frac{\partial h_t(\beta)}{\partial\beta}|] &< \infty \\ E[\sup_\theta |h_t^{1/2}(\beta)|] &< \infty \end{aligned}$$

■

4.8.2 Proofs

Proof of Theorem 5:. The proof follows Hansen(2006). We consider the stacked moment condition

$$m_t(\theta) = \begin{pmatrix} l_t(\beta) \\ 1\{\varepsilon_t(\beta) \leq q_\alpha\} - \alpha \end{pmatrix}$$

So it follows that

$$\begin{aligned} m_t &= \begin{pmatrix} \log \varepsilon_t^2 \frac{1}{h_t} \frac{\partial h_t}{\partial \beta} \\ 1\{\varepsilon_t \leq q_\alpha\} - \alpha \end{pmatrix} \\ E[l_t(\beta)] &= E[\log \frac{h_t}{h_t(\beta)} \frac{1}{h_t(\beta)} \frac{\partial h_t(\beta)}{\partial \beta}] \\ \frac{\partial E[l_t(\beta)]}{\partial \beta} &= E\{\log \frac{h_t}{h_t(\beta)} [\frac{1}{h_t(\beta)} \frac{\partial^2 h_t(\beta)}{\partial \beta \partial \beta'} - \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}] - \frac{1}{h_t^2(\beta)} \frac{\partial h_t(\beta)}{\partial \beta} \frac{\partial h_t(\beta)}{\partial \beta'}\} \end{aligned}$$

and

$$\begin{aligned} l_t &= \log \varepsilon_t^2 \cdot \frac{1}{h_t} \frac{\partial h_t}{\partial \beta} \\ E[l_t l_t'] &= E[\log \varepsilon_t^2]^2 \cdot E[\frac{1}{h_t^2} \frac{\partial h_t}{\partial \beta} \frac{\partial h_t}{\partial \beta'}] \\ l_\beta &= \frac{\partial E[l_t(\beta)]}{\partial \beta}|_{\beta=\beta_0} = -E[\frac{1}{h_t^2} \frac{\partial h_t}{\partial \beta} \frac{\partial h_t}{\partial \beta'}] \\ E[l_\beta^{-1} l_t l_t' l_\beta^{-1}] &= E[\log \varepsilon_t^2]^2 \cdot \{E[\frac{1}{h_t^2} \frac{\partial h_t}{\partial \beta} \frac{\partial h_t}{\partial \beta'}]\}^{-1} \end{aligned}$$

It has been shown by lemmas 1-4 that

$$E \sup_{\theta_1 : |\theta - \theta_1| < \delta} |m_t(\theta) - m_t(\theta_1)|^{2r} \leq O(1)\delta$$

this and Andrews(1994, theorem 5) ensure that $\int_0^1 \sqrt{\mathcal{H}(u, \mathcal{F}, \mathcal{L}_{2r}(P))} du < \infty$, where $\mathcal{H}(u, \mathcal{F}, \mathcal{L}_{2r}(P))$ denotes the entropy with bracketing with respect to $\mathcal{L}_{2r}(P)$ norm. By Doukhan et al.(1995) and lemma 1 in Hansen(2006), we know the following weak convergence regarding the score functions:

$$\sqrt{T}\{\overline{m}_T(\theta) - m(\theta)\} \Rightarrow \mathcal{S}(\theta)$$

where $\mathcal{S}(\theta)$ is a centered Gaussian process over $\theta \in \Theta$ and " $\Rightarrow$ " means weak convergence of empirical process $\overline{m}_T(\cdot)$ indexed by $\theta \in \Theta$.

In view of Taylor expansion,

$$0 = m(\theta_0) = m(\widehat{\theta}) + Q(\theta_0 - \widehat{\theta}) + o_p(T^{-1/2}),$$

where $Q = \frac{\partial}{\partial\theta}E[m_t(\theta_0)]$. Therefore, we have the following

$$\sqrt{T}\{\widehat{\theta}-\theta_0\} = Q^{-1}\sqrt{T}\{m(\widehat{\theta})-\overline{m}_T(\widehat{\theta})-[m(\theta_0)-\overline{m}_T(\theta_0)]\}+Q^{-1}\sqrt{T}\overline{m}_T(\widehat{\theta})-Q^{-1}\sqrt{T}\overline{m}_T(\theta_0)+o_p(1),$$

and we have used the fact that $\sqrt{T}\overline{m}_T(\widehat{\theta}) = o_p(1)$, and central limit theorem. We know that first, $\sqrt{T}\overline{m}_T(\widehat{\theta}) = o_p(1)$ holds trivially; second, by consistency of $\widehat{\theta}$ and stochastic equicontinuity of $\overline{m}_T(\cdot)$, $m(\widehat{\theta})-\overline{m}_T(\widehat{\theta})-[m(\theta_0)-\overline{m}_T(\theta_0)] = o_p(1/\sqrt{T})$; thirdly, CLT: $-Q^{-1}\sqrt{T}\overline{m}_T(\theta_0) \to^d N(0,\Omega_\theta)$. The proof of stochastic equicontinuity is based on above entropy condition, indicator function is a IV class defined in (5.3) and theorem 5 of Andrews (1994). In consequence,

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Q^{-1}\begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial\beta}\log\varepsilon_t^2 \\ 1\{\varepsilon_t \le q_\alpha\} - \alpha \end{pmatrix} + o_p(1).$$

■

Proof of Theorem 6. Denote a truncated verion of h_{n+1}as

$$h^*_{n+1} = G(c_0 + \sum_{j=1}^{m} c_j(\beta)\psi(y_{n+1-j})),$$

where the reuncation order is $m = \log n$. As a result, the approximation error is of order $o_p(1)$:

$$\begin{aligned} h_{n+1} - h^*_{n+1} &= G(c_0 + \sum_{j=1}^{\infty} c_j(\beta)\psi(y_{n+1-j})) - G(c_0 + \sum_{j=1}^{m} c_j(\beta)\psi(y_{n+1-j})) \\ &= g(c_0 + \sum_{j=1}^{m} c_j(\beta)\psi(y_{n+1-j})) \sum_{j=m+1}^{\infty} c_j(\beta)\psi(y_{n+1-j}) + O_p(||\sum_{j=m+1}^{\infty} c_j(\beta)\psi(y_{n+1-j})||^2) \\ &= O_p(b^m). \end{aligned}$$

Similarly, we can show that $\frac{\partial h_{n+1}}{\partial\beta} - \frac{\partial h^*_{n+1}}{\partial\beta} = O_p(b^m)$. Consequently, $x_{n+1} - x^*_{n+1} = O_p(b^m)$.

At the same time, we have the following trancation approximation

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{t=1}^{n} Q^{-1}\begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial\beta}\log\varepsilon_t^2 \\ 1\{\varepsilon_t \le q_\alpha\} - \alpha \end{pmatrix} \\ = & \frac{1}{\sqrt{n}}\sum_{t=1}^{n-m} Q^{-1}\begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial\beta}\log\varepsilon_t^2 \\ 1\{\varepsilon_t \le q_\alpha\} - \alpha \end{pmatrix} + \sqrt{\frac{m-1}{n}}\frac{1}{\sqrt{m-1}}\sum_{t=n-m+1}^{n} Q^{-1}\begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial\beta}\log\varepsilon_t^2 \\ 1\{\varepsilon_t \le q_\alpha\} - \alpha \end{pmatrix} \end{aligned}$$

$$= \frac{1}{\sqrt{n}} \sum_{t=1}^{n-m} Q^{-1} \begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial \beta} \log \varepsilon_t^2 \\ 1\{\varepsilon_t \le q_\alpha\} - \alpha \end{pmatrix} + o_p(1).$$

Combining above results, we can say that, conditional on information prior to time $n-m$,

$$x_{n+1}^{*\prime} \frac{1}{\sqrt{n}} \sum_{t=1}^{n-m} Q^{-1} \begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial \beta} \log \varepsilon_t^2 \\ 1\{\varepsilon_t \le q_\alpha\} - \alpha \end{pmatrix} \to^d N(0, x_{n+1}^{*\prime} \Omega x_{n+1}^{*}).$$

Consequently,

$$\begin{aligned}
& (x_{n+1}' \Omega x_{n+1})^{-1/2} \sqrt{n} (h_{n+1}^{1/2}(\widehat{\beta}) \widehat{q}_\alpha - h_{n+1}^{1/2} q_\alpha) \\
= \; & (x_{n+1}' \Omega x_{n+1})^{-1/2} \sqrt{n} (h_{n+1}^{1/2}(\widehat{\beta}) \widehat{q}_\alpha - h_{n+1}^{1/2}(\widehat{\beta}) q_\alpha + h_{n+1}^{1/2}(\widehat{\beta}) q_\alpha - h_{n+1}^{1/2} q_\alpha) \\
= \; & (x_{n+1}' \Omega x_{n+1})^{-1/2} \sqrt{n} h_{n+1}^{1/2} (\widehat{q}_\alpha - q_\alpha) + (x_{n+1}' \Omega x_{n+1})^{-1/2} \sqrt{n} \frac{1}{2 h_{n+1}^{1/2}} \frac{\partial h_{n+1}}{\partial \beta'} (\widehat{\beta} - \beta) q_\alpha + o_p(1) \\
= \; & (x_{n+1}' \Omega x_{n+1})^{-1/2} x_{n+1}' \sqrt{n} \begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{q}_\alpha - q_\alpha \end{pmatrix} + o_p(1) \\
= \; & (x_{n+1}^{*\prime} \Omega x_{n+1}^{*})^{-1/2} x_{n+1}^{*\prime} \frac{1}{\sqrt{n}} \sum_{t=1}^{n-m} Q^{-1} \begin{pmatrix} \frac{1}{h_t}\frac{\partial h_t}{\partial \beta} \log \varepsilon_t^2 \\ 1\{\varepsilon_t \le q_\alpha\} - \alpha \end{pmatrix} + o_p(1) \\
\to \; & {}^d N(0,1).
\end{aligned}$$

■

Proof of Theorem 7. Under Assumption 1-4 in Hill (2013) paper, Hill's (1975) estimator with estimated parameters has the following distribution

$$k_T^{1/2} (\widehat{\gamma}^{-1}(\widehat{\theta}_T) - \gamma^{-1}) / \sigma_{KT} \to^d N(0,1),$$

where $\sigma_{KT} = E(k_T^{1/2}(\widehat{\gamma}^{-1} - \gamma^{-1})^2$ is the MSE.

The residual quantile estimator in our paper is $\widehat{q}_\alpha = -(T \frac{\alpha}{k_T} \eta_{T,T-k_T}^{-\widehat{\gamma}})^{-1/\widehat{\gamma}} = -\eta_{T,T-k_T} (\frac{k_T}{\alpha T})^{\frac{1}{k_T} \sum_{t=1}^{k_T} \log \frac{\eta_{T,T-t+1}}{\eta_{T,T-k_T}}}$. Hence,

$$\begin{aligned}
\widehat{q}_\alpha &= -(T \frac{\alpha}{k_T} \eta_{T,T-k_T}^{-\widehat{\gamma}})^{-1/\widehat{\gamma}} \\
\log \widehat{q}_\alpha &= (\frac{1}{\widehat{\gamma}}) \log(T \frac{\alpha}{k_T} \eta_{T,T-k_T}^{-\widehat{\gamma}}) \\
&= \frac{1}{\widehat{\gamma}} (\log(\frac{Ta}{K_T}) - \widehat{\gamma} \log(\eta_{T,T-k_T}).
\end{aligned}$$

Assume $k_T \to \infty, k_T/T \to 0, \sqrt{k_T} A(T/k_T) \to 0, \frac{1}{A(T/k_T)T^{1/4}} \to 0, \log(\frac{k_T}{\alpha T})/\sqrt{k_T} \to 0$ as $T \to \infty$,

$$k_T^{1/2}(\widehat{\gamma}^{-1} - \gamma^{-1})/\sigma_{KT} \to^d N(0,1)$$

$$k_T^{1/2}(\log(\frac{Ta}{K_T})\widehat{\gamma}^{-1} - \log(\frac{Ta}{K_T})\gamma^{-1})/\sigma_{KT} \to^d N(0, (\log(\frac{Ta}{K_T}))^2)$$

$$k_T^{1/2}\{[\log(\frac{Ta}{K_T})\widehat{\gamma}^{-1} - \log(\eta_{T,T-k_T})] - [\log(\frac{Ta}{K_T})\gamma^{-1} - \log(\eta_{T,T-k_T})]\}/\sigma_{KT} \to^d N(-\log(\eta_{T,T-k_T}), (\log(\frac{Ta}{K_T}))^2)$$

Folllowing the biased reduction results in Gomes and Fegueiredo (2003), Gomes an Pestana (2005) and Beirland et al. (2006), we obtain the biased corrected Hill Estimator

$$K_T^{1/2}(\log(\widehat{q}_\alpha) - \log(q_\alpha))/\sigma_{KT} \to^d N(0, (\log(\frac{Ta}{K_T}))^2).$$

Hence, the conditional quantile follows

$$\frac{\sqrt{k_T}\log((\widehat{q}_\alpha)/q_\alpha)}{\sigma_{KT}|\log(\frac{k_T}{T\alpha})|} \to^d N(0,1).$$

■

Bibliography

[1] Ai, C. and X. Chen, (2003). "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, Econometric Society, vol. 71(6), pages 1795-1843

[2] Audrino,F. and Buhlmann. P : "Tree-structured GARCH models," *Journal of The Royal Statistical Society*, 63, 727-744.

[3] Avramidis,P. (2002). "Local maximum likelihood estimation of volatility function", *Manuscript*, LSE.

[4] Artzner, P.F., Delbaenm J, M. Eber, and D. Heath (1999). "Coherent measures of risk". *Mathematical Finance* 9, 203-228.

[5] Barndorff-Nielsen, O. E. and N. Shephard (2002). "Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models". *Journal of the Royal Statistical Society*, Series B 64, 253-280.

[6] Basel Committee on Banking Supervision (2012). http://www.bis.org/publ/bcbs219.pdf

[7] Back.K(1991), "Asset Pricing for General Processes," *Journal of Mathematical Economics*, 20, 371-395.

[8] Baillie, R.T., Bollerslev, T., (1989). "The Message in Daily Exchange Rates: a Conditional Variance Tale".*Journal of Business and Economic Statistics* 7, 297–305.

[9] Bauwens, L., Laurent, S., and Rombouts, J.V.K (2006), "Multivariate GARCH models: A survey", *Journal of Applied Econometrics*, 21, 79-109

[10] Berkes, I., Horvăth, L. and Kokoszka, P. (2003). "GARCH Processes: Structure and Estimation". *Bernoulli.* 9(2), 201-227.

[11] Bhattacharya,R and Lee,C (1995),"On Geometric Ergodicity of Nonlinear Autoregressive Models", *Statistics and Probability Letters*, Vol 22,Issue 4, 311-315

[12] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). "Efficient and Adaptive Estimation for Semiparametric Models". *Johns Hopkins University Press*, Baltimore. Reprinted (1998), Springer, New York.

[13] Bose,A. and Mukherjee, K (2003), "Estimating the ARCH Parameters by Solving Linear Equations", *Journal of Time Series Analysis*, Vol24,Issue 2,127-136

[14] Bollerslev.T (1986): "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics* 31, 307-327.

[15] Bollerslev.T (1990), "Modeling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Model," *Review of Economics and Statistics*, 72, 498-505.

[16] Bollerslev.T , Engle.R.F and Nelson,D.B (1994), "ARCH Models," in *Handbook of Econometrics*, volume IV, eds. R. F. Engle and D. L. McFadden, Elsevier Science, 2959-3038.

[17] Bosq,D,(1998) "Nonparametric Statistics for Stochastic Processes*". Estimation and Prediction*. Springer, Berlin.

[18] Cai and Wang (2008). "Nonparametric Methods for Estimating Conditional Value-at-risk and Expected Shortfall". *Journal of Econometrics*, Volume 147, Issue 1, 120-130

[19] Carrasco.M. and Chen,X (2002), "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17-39.

[20] Chan, Deng, Peng and Xia (2007), "Interval Estimation of Value-at-risk Based on GARCH Models with Heavy-tailed Innovations". *Journal of Econometrics*, vol. 137, issue 2, 556-576

[21] Chaudhuri, P (1997), "On Average Derivative Quantile Regression", *The Annals of Statistics.*

[22] Chen, S.X. (1996) "Empirical Likelihood Confidence Intervals for Nonparametric Density Estimation". *Biometrika* 83 , 329-341.

[23] Chen, S.X. (2008)." Nonparametric Estimation of Expected Shortfall". *Journal of Financial Econometrics* 1, 87-107.

[24] Chen, S.X. and C. Y. Tang (2005)."Nonparametric Inference of Value at Risk for dependent Financial Returns", *Journal of Financial Econometircs,* 3, 227-255.

[25] Chen, X and Y. Fan (2006), "Estimation and Model Selection of Semiparametric Copula-based Multivariate Dynamic Models under Copula Misspecification", *Journal of Econometrics,* 2006, Vol. 135, 125-154

[26] Chen, X., Xiao,Z., and R. Koenker (2009) "Copula-Based Nonlinear Quantile Autoregression", *The Econometrics Journal*, 12, S50-S67

[27] Chen,J., Peng,L. and Zhao.Y (2009) "Empirical Likelihood Based Confidence Intervals for Copulas", *Journal of Multivariate Analysis*, 100, 137-151.

[28] Chen,S. and Huang,T. (2007), "Nonparametric Estimation of Copula Functions for Dependence Modelling", *The Canadian Journal of Statistics*, 35, 265-282

[29] Chen,X and Fan,Y (2006a), "Estimation and Model Selection of Semiparametric Copula-based Multivariate Dynamic Models under Copula Misspecification." *Journal of Econometrics*, 135, 125-154.

[30] Chen,X and Fan,Y (2006b), "Semiparametric Estimation of Copula-based Time Series Models." *Journal of Econometrics*, 130, 307-335.

[31] Chen,X and Fan,Y and Tsyrennikov,V.,(2006), "Efficient Estimation of Semiparametric Multivariate Copula Models", *Journal of the American Statistical Association*, 101, 1228-1240.

[32] Chen,X., Wu,W and Yi,Y (2009), "Efficient Estimation of Copula-based Semiparametric Markov models", *The Annals of Statistics*, 37 (6B), 4214-4253.

[33] Chollette, L., Heinen,A and Valdesogo,AI (2009), "Modelling International Financial Returns with a Multivariate Regime Switching Copula", *Journal of Financial Econometrics*, Vol7(4),437-480

[34] Christoffersen,P and Gonçalves,S., "Estimation Risk in Financial Risk Management" (2005). *Journal of Risk* 7(3),1-28

[35] Cline,D.B.H and Pu.H.H (1999),"Geometric Ergodicity of Nonlinear Time Series", *Statistica Sinica* 9, 1103-1118

[36] Cline,D.B.H and Pu.H.H (2004),"Stability and the Lyapounov Exponent of Threshold AR-ARCH Models", *The Annuals of Applied Probability*, Vol 14, No. 1920-1949

[37] Dahlhaus,R and Subba Rao, S(2006), "Statistical Inference for Time-varying ARCH Processes". *The Annals of Statistics*, 34, 1075-1114.

[38] Drost,F.C., and Klaassen C.A.J (1997): "Efficient Estimation in Semiparametric GARCH models," *Journal of Econometrics* 81, 193-221.

[39] Drost, F.C.A. J. Klaassen and Bas Werker (1997) Aadaptive estimaiton in time series models. *The Annals of Statistics* Vol. 25, No. 2, 786-817

[40] Duffie, D. and J. Pan (1997), "An Overview of Value at Risk", *Journal of Derivatives*, 4,7-49.

[41] Durbin, J. (1973). "Weak convergence of the sample distribution when parameters are estimated". *The Annals of Statistics* 1, 279-290.

[42] Embrechts,P.,McNeil,A. and Straumann,D (2002), "Correlation and Dependence Properties in Risk Management: Properties and Pitfalls," *Risk management: Value at Risk and Beyong, edited by Dempster. M., Alan, M and Dempster. H, Cambridge Press.*

[43] Engle, R. F.(2001) ,"Financial Econometrics - A New Discipline With New Methods." *Journal of Econometrics*. V100 pp53-56

[44] Engel, R. F. and Gonzalez-Rivera (1991) "Semiparametric ARCH Models", *Journal of Business and Economic Statistics*, 9:345-360.

[45] Engle, R. F. and Manganelli (2004), "CAViaR: Conditional Value at Risk by Quantile Regression". *Journal of Business and Economic Statistics* 22, 367-81.

[46] Engle,R.F. and Ng.V (1993). "Measuring and Testing the Impact of News on Volatility", *Journal of Finance* 48, pp1749-1777.

[47] Engle,R.F (2002), "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models", *Journal of Business & Economic, Statistics*, 20, No. 3 (Jul. 2002), pp. 339-350

[48] Engle,R.F (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica* 50, 987-1008.

[49] Engle,R.F (2000), "The Econometrics of Ultra-High Frequency Data," *Econometrica* 68, 1-22

[50] Engle,R.F and Gonzalez-Rivera, G (1991): "Semiparametric ARCH models," *Journal of Business and Economic Statistics* 9: 345-359.

[51] Engle,R.F., Lilien,D.M., and Robins,R.P (1987), "Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model," *Econometrica*, 19, 3-29.

[52] Engle,R.F. and Ng, V.K, (1993) “Measuring and Testing the impact of news on volatility,” *The Journal of Finance* XLVIII, 1749-1778.

[53] Engle,R.F. and Rangel (2008): “The spline GARCH model for low-frequency volatility and its global macroeconomic causes,”.*Review of Financial Studies*.21(3),1187-1222

[54] Embrechts.P, Klüppelberg.C and Mikosch T (1997), "Modelling Extremal Events-for Insurance and Finance", *Springer*

[55] Engle,R.F."Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of U.K. Inflation" (1982). *Econometrica* 50, 987-1008

[56] Engle,R.F (2001),"GARCH101: The Use of ARCH/GARCH Models in Applied Econometrics". *Journal of Economic Perspectives*, Vol 15,No 4, Page 157-169

[57] Engle, Robert F & Ng, Victor K, (1993). " Measuring and Testing the Impact of News on Volatility," *Journal of Finance*, vol. 48(5), 1749-78

[58] Gallant,A.R. and Tauchen,G (1989). "Seminonparametric estimation of conditionally constrained heterogeneous processes: asset pricing applications.", *Econometrica*, 57, 1091-1120

[59] Genest,C., Ghoudi,K. and Rivest,L.P (1995). "A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions." *Biometrika,* 82, 534-552.

[60] Giraitis,L, Leipus, R, and Surgailis, D. (2009) “ARCH(∞) and Long Memory Properties,” *Handbook of Financial Time Series*, 71-84

[61] Gill, R. D. (1989). "Non- and Semi-parametric Maximum Likelihood Estimators and the Von-Mises method (Part I). *Scand. J. Statist.* 16, 97-128.

[62] Gourieroux,C. and Jasik, J. "Financial Econometrics: Problems, Models and Methods". (2001),*Princeton University Press.*

[63] Glosten,L., Jegannathan, R. and Runke,D. "Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stock". *Journal of Finance*,48, 1779-1801

[64] Fama,E.F (1965). The behavior of stock market prices. *Journal of Business* 38, 34–105.

[65] Fan,J and Yao,Q (1998): “Efficient estimation of conditional variance functions in Stochastic Regression,” *Biometrika* 85, 645-660.

[66] Feng,Y (2007), "A local dynamic conditional correlation model." *MPRA paper,* No. 1592

[67] Fermanian,J.D., Radulovic,D. and Wegkamp,M (2004), " Weak convergence of empirical copula processes." *Bernoulli*, 10, 847-860.

[68] Fermanian,J.D. and Scaillet,O (2003), "Nonparametric estimation of copulas for time series", *Journal of Risk,* 5(4), 25-54.

[69] Francq.C, and Zakoïan (2006), "Mixing Properties of a General Class of GARCH(1,1) Models without Moment Assumptions on the Observed Process", *Econometric Theory*, 22, 815-834

[70] Franke,J., Neumann,M.H., and Stockis,J (2004) Bootstrapping nonparametric estimators of the volatility function. *Journal of Econometrics* 118, No.1-2, 189-218

[71] Hall, P and C. Heyde (1980). "Martingale Limit Theory and Its Application". *New York: Academic Press.*

[72] Hall, P. and Yao, Q. (2003). "Inference in ARCH and GARCH Models with Heavy-tailed Errors". *Econometrica*. 71, 285-317.

[73] Hansen. B.E (2006). "Interval Forecast and Parameter Uncertainty", *Journal of Econometrics*,135, 377-398

[74] Hall,P and Yao,Q (2003). "Inference in ARCH and GARCH Models with Fat-tailed Errors". *Econometrica,* 71, No.1, 285-317

[75] Hafner,C.M. and Linton,O.B (2010), "Efficient Estimation of a Multivariate Multiplicative Volatility Model", *Journal of Econometrics*, 159, 1,55-73

[76] Hafner,C.M. and Rombouts,J.V.K (2007). "Semiparametric Multivariate Volatility Models". *Econometric Theory*, 23, 251-280

[77] Hafner,C.M. , Dijk,D.V. and Franses,P.H. (2005), "Semi-parametric Modelling of Correlation Dynamics." in *Advanced in Econometrics*, ed, by Fomby. T, Hill.C and Terrell. D, 20/A, 59-103. Amsterdam: Elsevier Sciences

[78] Hardle,W. and Tsybakov,A (1997),"Local polynomial estimators of the volatility function," *Journal of Econometrics*, 81, 223-242.

[79] Hentschel, L., (1995). "All in the Family: Nesting Symmetric and Asymmetric GARCH models". *Journal of Financial Economics*,39, 71–104

[80] Hill,J.B (2010). "On Tail Index Estimation for Dependent Heterogeneous Data", *Econometric Theory*,26,1398-1436

[81] Hill,J.B (2013)."Tail Index Estimation for a Filtered Dependent Time Series", *University of North Carolina (UNC) at Chapel Hill Working paper*

[82] Hols, Martien C A B & de Vries, Casper G, (1991). The Limiting Distribution of Extremal Exchange Rate Returns", *Journal of Applied Econometrics*, vol. 6(3), pages 287-302.

[83] Hodgson.D, Linton, O.B and Vorkink.K, "Testing the CAPM efficiently under elliptical symmetry: A Semiparametric Approach ". *Journal of Applied Econometrics* 17, 617-639

[84] Jensen. S.T. and Rahbek, A 2004. Asymptotic Normality of the QMLE Estimator of ARCH in the Nonstationary Case, *Econometrica*, 72(2), pages 641-646.

[85] Ingersoll, J.E (1984). "Theory of Financial Decision Making", *Rowman & Littlefield Pub Inc*

[86] Jondeau,E. and Rockinger,M (2006), "The Copula-GARCH model of conditional dependencies: An international stock market application", *Journal of International Money and Finance*, 25, 827-853.

[87] Kitamura, Y. (1997). "Empirical likelihood methods for weakly dependent processes'. *the Annals of Statistics*, 25, 2084-2102.

[88] Kim,W and Linton,O.B (2004): "A Local Instrumental Variable Estimation method for Generalized Additive Volatility Models," *Econometric Theory* 20, 1094-1139.

[89] Kim.S, Shephard.N and Chib,S. (1998), "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models", *Review of Economic Studies*,65, 361-393

[90] Koedijk.K.G, Schafgans. M.A. and de Vries.C.G (1990). "The tail index of exchange rate returns", *Journal of International Economics 29*, pp93-108

[91] Koenker and Xiao (2009). "Conditional Quantile Estimation for GARCH Models," *Boston College Working Papers in Economics 725,* Boston College Department of Economics.

[92] Komunjer, Ivana, 2005. "Quasi-maximum Likelihood Estimation for Conditional Quantiles," *Journal of Econometrics*, Elsevier, vol. 128(1), pages 137-164,

[93] Komunjer, I., Vuong, Q.(2010), "Semiparametric Efficiency Bound in Time-series Models for Conditional Quantiles". *Econometric Theory,26,02.383-405*

[94] Koul, H. and Ling, S. (2006). "Fitting an Error Distribution in some Heteroscedastic Time Series Models". *Annals of Statistics* 34, 994-1012.

[95] Koul, H. and Ossiander, M. (1994). "Weak Convergence of Randomly Weighted Dependent Residual Empiricals with Applications to Autoregression".*Annals of Statistics,* 22 540–562.

[96] Krause,J and Paolella, M.S. (2014), "A Fast, Accurate Method for Value at Risk and Expected Shortfall", *Working Paper Series*

[97] Kuester K., Mittnik S., Paolella M. ,S(2006)). "Value-at-Risk Prediction: A Comparison of Alternative Strategies". J*ournal of Financial Econometrics* 4:53–89.

[98] Kristensen.D,(2007),"Uniform Ergodicity of a Class of Markov Chains with Applications to Time Series Models". *Columbia University*, 10

[99] Lee, S.-W. and Hansen, B.E. (1994). "Asymptotic Theory for the GARCH (1,1) Quasimaximum Likelihood Estimator". *Econometric Theory,* 10, 29-52.

[100] Lee,C (1997), "On the Functional Central Limit Theorm for a Class of Nonlinear Autoregressive Processes", *Comm. Korean Math Soc* 12,No.4 1033-1038

[101] Lee,S.W and Hansen,B.E (1994), "Asymptotic Theory for the GARCH(1,1) Quansi-Maximum Likelihood Estimator". *Econometric Theory* 10, 29-53

[102] Linton,O.B (1993). "Adaptive Estimation in ARCH Models", *Econometric Theory* (1993) 9, 539-569*Econometric Theory*

[103] Linton,O.B., Pan,J and Wang,H (2009). "Estimation for a Non-stationary Semi-strong GARCH(1,1) Model with Heavy-tailed Errors", ,26,1-28

[104] Linton,O.B, and Mammen,E (2005) "Estimating Semiparametric ARCH(∞) Models by Kernel Smoothing Methods**,"** *Econometrica* 73, 771-836.

[105] Linton,O.B, and Nielsen,J.P (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Using Marginal Integration", *Biometrika*,82, 93-100.

[106] Linton,O.B, and Perron.B (2003), "The Shape of the Risk Premium: Evidence from a Semiparametric Generalized Autoregressive Conditional Heteroskedasticity Model," *Journal of Business & Economic Statistics* 354-367.

[107] Ling, S. and McAleer, M. (2003) , "Adaptive Estimation in Nonstationry ARMA Models with GARCH Noises". *Annals of Statistics*, 31, 642-674.

[108] Liebscher, E. (2005), "Towards a Unified Approach for Providing Geometric Ergodicity and Mixing Properties of Nonlinear Autoregressive Processes", *Journal of Time Series Analysis* 26, 669 – 689.

[109] Liu, S., Brorsen, B.W., 1995. "Maximum Likelihood Estimation of a GARCH-stable model". *Journal of Applied Econometrics,* 10, 273–285.

[110] Lu,Z and Linton,O.B (2007): "Local Linear Fitting under Near Epoch Dependence," *Econometric Theory,*23(1), 37-70

[111] Mammen,E., Linton,O.B.and Nielsen,J.P(1999), "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions", *Annals of Statistics,* **27**, 1443-1490.

[112] Mandelbrot,B (1963). "The Variation of Certain Speculative Prices". *Journal of Business* 36, 394-419.

[113] Masry,E., and Tjϕstheim.D (1995): "Nonparametric estimation and identification of nonlinear ARCH time series: strong convergence and asymptotic normality," *Econometric Theory* 11, 258-289.

[114] Manski,C.F. (1998), "Analog Estimation Methods in Econometrics", *New York: Chapman and Hall*

[115] McNail,A.J and Frey R. "Estimation of Tail-related Risk Masures for Heteroscedastic Financial Time Series: an Extreme Value Approach". Journal of Empirical Finance 7:271-300

[116] Merton,R.C (1973), "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41, 867-887.

[117] Mikosch,T. and Starica,C (2000): "Limit Theory for the Sample Autocorrelations and Extremes of a GARCH(1,1) Process," *Annals of Statistics* 28, 1427-1451.

[118] Mittnik,S and Rachev,S.T, "Stable Paretian Models in Finance". (2000), *Wiley, New York.*

[119] Müller, U.U., Schick,A., and Wefelmeyer.W (2005). "Weighted Residual-based Density Estimators for Nonlinear Autoregressive Models". *Statistics Sinica*, 15, 177-195.

[120] Nelson, D. B. (1991). "Conditional Heteroskedasticity in Asset Returns: A New Approach", *Econometrica* 59: 347-370.

[121] Newey,W.K. and Steigerwald,D.G (1997),"Asymptotic Bias for Quasi-Maximum-Likelihood Estimators in Conditional Heteroskedasticity Models", *Econometrica*, Vol65,No,3,587-599

[122] Nelson,D.B. (1990). "Stationarity and Persistence in the GARCH(1,1) model". *Econometric Theory* 6, 318-34.

[123] Newey,W.K. and Steigerwald,D.G (1997), "Asymptotic Bias for Quasi-Maximum-Likelihood Estimators in Conditional Heteroskedasticity Models". *Econometrica* 65, 3, 587-599.

[124] Nyström, K.,and Skoglund, J. "Efficient Filtering of Financial Time Series and Extreme Value at Risk". Journal of Risk, Vol.7, No.2, pp 63-84

[125] Owen, A. (1990). "Empirical Likelihood Ratio Confidence Regions", *Annals of Statistics* 18 (1),. 90–120.

[126] Owen,A (2001). "Empirical Likelihood". Chapman and Hall.

[127] Oakes,D (1994), "Multivariate Survival Distributions." *Journal of Nonparametric Statistics.* 3, 343-354.

[128] Okimoto,T (2007), "New evidence of asymmetric dependence structures in international equity markets", *Journal of Financial and Quantitative Analysis,43,03,787-815*

[129] Pagan,A.R., and Schwert,G.W, (1990): "Alternative models for conditional stock volatility," *Journal of Econometrics* 45, 267-290.

[130] Pagan,A.R., and Hong,Y.S, (1991): "Nonparametric Estimation and the Risk Premium," in W. Barnett, J. Powell, and G.E. Tauchen (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press.

[131] Patton,A (2006a), "Estimation of multivariate models for time series of possibly different lengths," *Journal of Applied Econometrics*, 21(2), 147-173.

[132] Patton,A (2006b), "Modelling asymmetric exchange rate dependence," *International Economic Review*, 47(2), 527-556.

[133] Pelletier,D. (2006), "Regime-switching for dynamic correlation", *Journal of Econometrics*, 131, 445-473.

[134] Peter M Robinson & Paolo Zaffaroni, 2005. "Pseudo-Maximum Likelihood Estimation of ARCH(∞) Models," STICERD - Econometrics Paper Series /2005/495, LSE.

[135] Peng, L. and Yao, Q. (2003). Least absolute deviations estimation for ARCH and GARCH models. Biometrika, 90, 967-975.

[136] Robinson,P.M, (1991): “Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression,” *Journal of Econometrics* 47, 67-84.

[137] Robinson,P.M. and Zaffaroni,P. (2006): “Pseudo-maximum likelihood estimation of ARCH(∞) models,” *Annals of Statistics* 34, 1049-1074.

[138] Rodriguez,J. (2007), "Measuring finanicial contagion: A copula approach", *Journal of Empirical Finance*, 14, 401-423.

[139] Rachev,S.T.(2003), "Handbook of Heavy Tailed Distribution in Finance ",*North Holland*

[140] Scaillet, O. (2004)." Nonparametric estimation and sensitivity analysis of expected shortfall". *Mathematical Finance* 14, 115-129.

[141] Schick, A. and Wefelmeyer, W. (2002). "Estimating the innovation distribution in nonlinear autoregressive models". *Ann. Inst. Statist. Math.* 54 245–260.

[142] Sentana, E,(1995)"Quadratic ARCH models" , *Review of Economic Studies*, 62, 639–661.

[143] Spokoiny,V. (2009), “Varying coefficient GARCH Models,” *Handbook of Financial Time Series*, 169-185

[144] Stone,C.J (1980): “Optimal rates of convergence for nonparametric estimators,” *Annals of Statistics* 8, 1348-1360.

[145] Stone,C.J (1985): “Additive regression and other nonparametric models,” *Annals of Statistics* 13, 685-705.

[146] Terasvirta,T., Tjϕstheim,D and Granger,C.W.J (1994). "Aspects of Modelling Nonlinear Time SeriesAspects of Modelling Nonlinear Time Series", in *The Handbook of Econometrics*, vol. IV, eds. D.L. McFadden and R.F. Engle, 2919-2960, Amsterdam: Elsevier.

[147] Tjϕstheim,D., and Auestad,B.H (1994), "Nonparametric identification of nonlinear time series: projections". *Journal of American Statistical Association, 89,* 1398-1409.

[148] Tong,H, (1990) "Nonlinear Time Series Analysis: A dynamical systems Approach", *Oxford University Press*, Oxford.

[149] Weiss, A. (1986). "Asymptotic Theory for ARCH models: Estimation and Testing". *Econometric Theory*, 2, 107-131.

[150] Wiener, N. (1958). "Nonlinear problems in random theory".*MIT Press*, Cambridge.

[151] White,H (2000), "A Reality Check for Data Snooping", *Econometrica,* Vol. 68, No.5, 1097-1126

[152] Wu,G and Xiao,Z. (2002): "A generalized partially linear model for asymmetric volatility," *Journal of Empirical Finance* 9, 287-319.

[153] Xiao,ZJ and Koenker,R.(2009) "Conditional Quntile Estimation and Inference for GARCH Models", *Journal of American Statistical Association*, Vol104, No.448:1696-1712

[154] Yan.Y, Shang,DJ and Linton, O.B (2013), "Efficient Estimation of Conditional Risk Measures in a Semiparametric GARCH Model", Cemmap Working Paper CWP25/12

[155] Yang,L. (2006): "A Semiparametric GARCH Model for foreign exchange volatility," *Journal of Econometrics* 130,2, 365-384.

[156] Yang,L., Hurdle,W and Nielsen,J.P (1999): "Nonparametric Autoregression with Multiplicative Volatility and Additive Mean," *Journal of Time Series Analysis* 20, 579-604.

[157] Zangari.P. "RiskMetrics Technical Document", 1996

[158] Zakoian,J.M, (1994), "Threshold Heteroskedasticity Models", *Journal of Economic Dynamics and Control,* 15, 931-955

[159] Zaffaroni,P, (2009), "Whittle Estimation of EGARCH and other Exponential Volatility Models", *Journal of Econometrics*, Vol, 151, 190-200

[160] Zhang, B. (1995) "M-estimation and quantile estimation in the presence of auxiliary information". *Journal of Statistical Planning and Inference* 44:77 94

[161] Ziegelmann,F, (2002),. Nonparametric estimation of volatility functions: the local exponential estimator. *Econometric Theory 18,* 985-992.