The University of Auckland

SNM-GARCH: A non-parametric mixture extension to GARCH modelling

Author: Cheng-Jan (Michael) Kao

Supervisor: Senior Lecturer Yong Wang

July 13, 2012

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistics, The University of Auckland.

Abstract

Volatility modelling is critical to modern financial risk management as it provide an accurate measure of the risk faced by investor. Vast amount of research and effort has devoted to measure the volatility of prices in the past century, yet there is no consensus to how the volatility should be modelled and measured. Amongst the enormous varieties of models, the most widely recognised and applied is the ARCH/GARCH family pioneered by Engle in the late 1980s. Due to the size and the potential impact of the result, highly accurate measures are desired. Consider a moderate portfolio of a billion dollar, a 1% error in the estimation of the volatility can result in a risk measure differ by one million dollar. In this paper we propose a flexible class of distribution known as the scale normal mixture for GARCH in which the volatility can be measured accurately and can be applied over a large spectrum of assets with the potential to explain the micro-market structure or driving force of volatility.

Acknowledgement

First I want to thank my parents who brought me to this world to experience life, nurture me as to who I am today, to support every decision I make. It is their support both financially and emotionally to allow me to discover what I am passionate about, mathematics and statisics. Their love and support can not be expressed sufficiently with human language, only the concept of Infinity can best describe my gratitude towards them.

A huge thank to my supervisor Yong Wang, who spent relentless amount of time spent towards helping me completing this thesis, overcome the difficulty of communicate with me while oversea. I would not be qualified as a statistician without his mentoring and the knowledge gained from him.

For the past four and a half years, the department of statistics has not only taught me about statistics, but also about life and to take care of me while as a student. I would thus like to express my greatest appreciation for their work and effort of making this the best department.

I would also like to thank my employer and the team at the Food and Agriculture Organization of the United Nation for providing financial support in the last three months of my thesis allowing me to concentrate solely on this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Profit or return is the main drive or motivation for investment, however, every investment comes with a risk. Risk is the uncertainty associated with the price movement of the investment in the future typically defined as the deviation of the price that differs from its expected value. Volatility is the measure of such risk usually in terms of standard deviation or a bound representing possible loss.

Risk management is at the core of modern financial modeling and analysis. The ability to quantify the risk of an asset or financial instrument determines the value of the asset in accordance with the risk profile of the investor. Vast amount of research and theory has been developed to understand the mechanism of volatility movement in the finance and the economic discipline. From the utility function of individual investor to portfolio theory of large conglomerate; from rational investor of the efficient market to the irrational action of behaviour finance, various theory and hypothesis have been proposed.

Beyond the financial point of view, in recent years, the concern for risk management has risen as a public issue. This is a result that the fall of large financial giants can have a disastrous impact on the society as a whole. Consequently, the Basel Accord have been tightened and now requires the need for banks to hold levels of capital sufficient to cover losses on the bank's trading portfolio over a 10-day holding period in 99% of occasions.

Many attempts has been made since the early 20th century to understand the movement of the price and the volatility of assets over time. The first rigorous study dates back to the 1900 in a thesis entitled "Theorie de la Speculation" by Bachelier (1900), which contains both theoretical work and empirical analysis of the financial market using the Brownian motion. One of the most famous observation made in 1963 by the mathematician Benoit Mandelbrot (1963) was that the unconditional distribution of asset returns have higher peak, while at the same time a heavier tail. Furthermore, he is credited as the first to notice the clustering of volatility which induced interest in formulating time series model of volatility. Following Mandelbrot, the econometrician Robert Engle proposed the ARCH process in 1982 and further generalised by Tim Bollerslev in 1986 which explains the volatility clustering and provide a simple and intuitive framework for time series volatility modelling. Yet, this is just the beginning rather than the end towards understanding the volatility process. Much of the empirical observation such as leverage effect which investor respond differently to positive and negative shocks and the driving force of volatility are just the tip of the ice berg of the mystery of volatility.

The ARCH family and its extension GARCH are the most influential discrete time series model in both the literature with over 25000 citations on the two foundation papers; and application in almost every single asset with a wide range of extensions. In Bollerslev (2008), a glossary of 130 different extension of the GARCH model were presented.

1.1 Research Aim

The aim of this research is to introduce the scale normal mixture in the general context of GARCH and to assess the potential of becoming the standard distribution in volatility models. The literature on univariate GARCH models is quite voluminous, and it is not possible to incorporate all developments and extensions of the original model in this thesis. However, we hope that the work of this text is sufficient to justify the use of the scale normal mixture in practical work.

1.2 Thesis outline

In this paper, the original normal GARCH model is present with its shortfall in capturing extreme risks and outline some of the proposed research for alleviating this problem. The newly proposed scale normal mixture will then follow at the end of the section, with the motivation and theoretical properties described.

We then turn to the theory and numerical method in which the model is based upon in Chapter 3, and a simple demonstration how the model can be fitted to data and the potential to explain the source of volatility. Chapter 4 follows on by describing the details of the algorithm, with the reasoning and ideas behind the initialisation of the model and a new strategy for accounting unobserved conditional variance. The mathematical detail for implementing the model is also presented for the interested audience.

In chapter 5 we conduct a simulation study to examine the performance of competing models when the true distribution is known. Then the Scale normal mixture GARCH (SNM-GARCH) is applied to real data and again assessed with alternative models to determine the practicality.

Finally, the conclusion of the research and possible future research are given in Chapter 7.

Chapter 2

Background

2.1 The Generalized Autoregressive Heteroscedasticity model for volatility

In the ground breaking paper presented in 1982 by Robert Engle (1982) a class of discrete stochastic process called autoregressive conditional heteroscedastic (ARCH) process along with a regression model framework was introduced to estimate the variance of the inflation in the United Kingdom.

Engle pointed out that the traditional econometric model assuming the conditional variance of the financial time series depending only on the information set available in the single previous period is an unrealistic one. He also mentioned that having the conditional variance dependent on other exogenous variable is also unsatisfactory in theory as it requires a specification of the causes of the changing variance which may jointly evolve over time. He expressed the conditional variance of a time series depends only on the realization of the time series itself and coined the process autoregressive conditional heteroscedastic process which we will simply refer to as an ARCH(p) process for brevity.

By Engle's definition, an ARCH(P) process assuming normal error has the following expression

$$x_t = \epsilon_t \sigma_t, \qquad \epsilon_t \sim N(0,1), \tag{2.1}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 \tag{2.2}$$

Where p is the autoregressive order of the ARCH(p) process, α's are the autocorrelation coefficients and ϵ_t is an *iid* normally distributed random variable with zero mean and unit variance. We will refer to Equation 2.1 as the observation equation and Equation 2.2 the volatility equation from herein. Although, the model is simple and intuitive, but large lags are often required to account for the long memory characteristic of financial time series. Tim Bollerslev (1986) generalised this in 1986 to give the GARCH(p, q) process which he shown can close resemble the lag of an ARCH(8) model as GARCH(1, 1) replacing 9 parameters with only 3. He extend the volatility equation from depending solely on the observation to also the volatility itself, and has the following form

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 \tag{2.3}$$

Which is analogous as the ARMA representation to AR and gives a parsimonious expression of the model. Both α and β are required to be positive.

From the expressions, we can see the ARCH/GARCH family model allows the volatility to transit smoothly over time as new information are incorporated. This fact allows us to generate time series with volatility clustering and similar patterns as to typical financial time series. In Figure 2.1 we compare a real financial time series the SNP500; a random sample from the GARCH normal process; and a random sample drawn independently from a normal distribution. From the figure, the volatility clustering can be visually identified in both the SNP500 data and the GARCH Normal process but not the *iid* process. Furthermore, even though the variance of the three time series are equivalent, we can distinguish them with two phenomenon (1) The GARCH normal process has more extreme observations in comparison to the

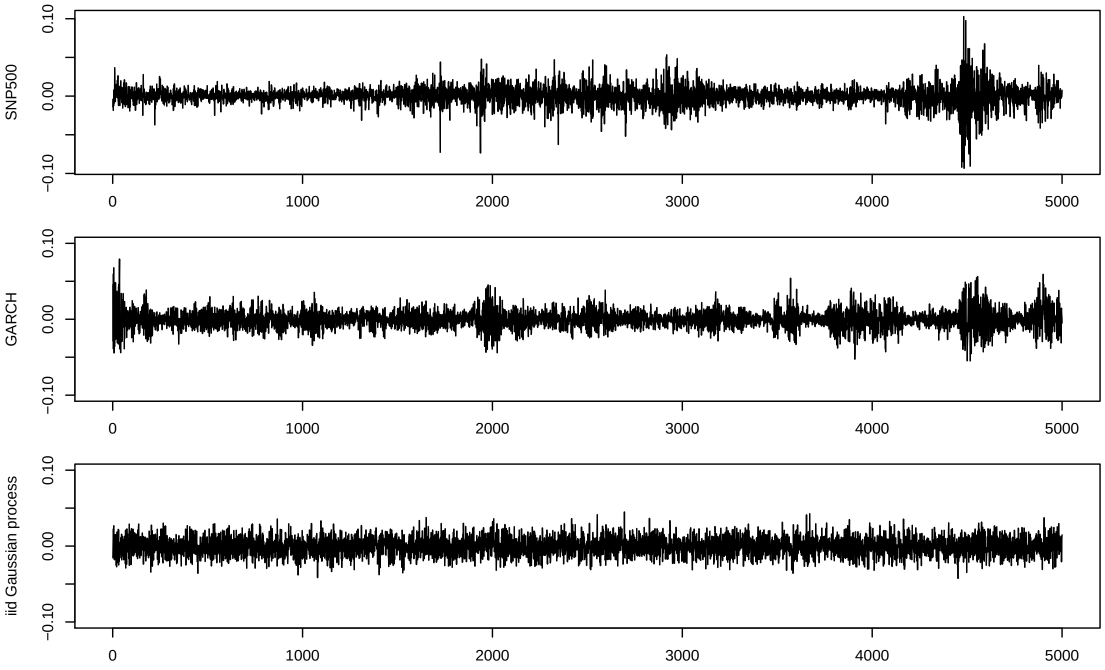

Figure 2.1: In this graph we compare a sample of 5000 observation of the SNP500, random sample from a GARCH(1, 1) process ($\alpha = 0.1, \beta = 0.89$) and a *iid* Gaussian process. All three process have the same variance.

iid Gaussian process. We will shown in the next section that the GARCH process can generate a process which is more leptokurtic than the underlying distribution, (2) Yet, the SNP500 data seems to exhibit more extreme values than the GARCH normal process which suggests that the normal distribution lack the heavy tail of financial times series.

Prior researches by Bera and Higgins (1993) suggest that a GARCH(1, 1) process is capable of representing majority of the financial time series and that GARCH models of higher order are rarely seen in practice .Hansen and Lunde (2005) also pointed out in their comprehensive study of more than 330 ARCH type models that increasing the order of the model does not necessary improve the predictability, rather the ability to capture higher moments of the distribution and the leverage effect is the key to higher performance forecast. Thus for the remainder of the paper we will focus our attention on the GARCH(1, 1) model.

2.1.1 Properties

Studying the properties of the GARCH process is the fastest way to understand the model and characteristics.

Moments

First, by taking the expectation of observation equation in Equation 2.1 we obtain the conditional mean of the process.

$$\mathbb{E}(x_t) = \mathbb{E}[\mathbb{E}(x_t|F_{t-1})] = \mathbb{E}[\sigma_t \mathbb{E}(\epsilon_t)] = 0 \tag{2.4}$$

To understand the higher moments of the GARCH models, it is informative to use the following ARMA representation. Let $\xi_t = x_t^2 - \sigma_t^2$ then we can rewrite observation equation as

$$x_t^2 = \alpha_0 + (\alpha_1 + \beta_1)x_{t-1}^2 + \xi_t - \beta_1 \xi_{t-1}. \tag{2.5}$$

It can be shown that the series ξ_t is a martingale difference series, and that Equation 2.5 is an ARMA representation of the time series x_t^2. Then borrowing from the theory of ARMA, we have the unconditional variance of x_t as

$$\text{VAR}(x_t) = \mathbb{E}(x_t^2) = \frac{\alpha_0}{1 - \alpha_1 - \beta_1} \tag{2.6}$$

given the normality assumption we can then solve for the unconditional kurtosis as

$$\frac{\mathbb{E}(x_t^4)}{[\mathbb{E}(x_t^2)]^2} = \frac{3[1 - (\alpha_1 + \beta_1)^2]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} > 3 \tag{2.7}$$

Where we can see that the kurtosis and hence the tail of the conditional distribution is heavier than that of the normal. This is the reason why in Figure 2.1 a GARCH normal process generates more extreme observation when both process assumes a normal distribution with equal variance.

Persistence

The persistence of a GARCH(1, 1) process is defined as

$$\tau = \alpha_1 + \beta_1 \tag{2.8}$$

It is one of the most important property of the GARCH process, as it characterizes the dependency of the current conditional variance on previous realization. It is equivalent to the exponential decay rate of the autocorrelation function can be used to calculate the unconditional variance. Furthermore, it determines the nature of the stochastic process where

- Stationary GARCH process if $\tau < 1$.
- A unit root process if $\tau = 1$.
- Non-stationary process if $\tau > 1$.

2.1.2 Estimation

To estimate the GARCH model under the maximum likelihood estimation, the following likelihood or equivalently the log-likelihood function must be formulated and maximised.

$$\ell(\mathbf{x}; \sigma_t^2) = \prod_{t=1}^{T} \mathcal{D}(x_t, \sigma_t^2) \tag{2.9}$$

In order to compute Equation 2.9, the conditional variance σ_t^2 of the time series governed by the volatility equation must be calculated first which in turn can then be substituted to evaluate the likelihood. We have rewritten Equation 2.2 as Equation 2.10 below purely in terms of the realisation and the GARCH parameters. An initial value is required as the conditional variance and observations prior to time $t \leq 1$ is unobserved.

$$\sigma_t^2 = \begin{cases} \sigma_1^2 & \text{if } t = 1 \\ \alpha_0 \sum_{i=0}^{t-2} \alpha_1^i + \alpha_1^{t-1}\sigma_1^2 + \beta_1 \sum_{i=0}^{t-2} \alpha_1^i x_{t-i-1}^2 & \text{if } t \neq 1 \end{cases} \tag{2.10}$$

We will discuss the details of how to initialize σ_1^2 in section 4.2 and also propose modification which provides a more consistent and stable estimate.

Under the normality assumption, the likelihood is then

$$\ell(\mathbf{x}; \sigma_t^2) = \sum_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(\frac{-x_t^2}{2\sigma_t^2}\right) \tag{2.11}$$

and is typically optimised using the BHHH algorithm.

2.2 Short coming

Mandelbrot (1963) was the amongst the first to notice the fact that financial time series are typically heavy-tailed and a greater probability mass centering the mode . He proposed the use of stable distribution with the decay parameter $\alpha \in [0, 2]$ and develop a heavy-tailed stochastic process for financial returns. He coined the term “Levy flight” which is a random walk in which the step-lengths has a Pareto distribution.

It was not long after the emergence of the original GARCH model in the 1980s, researchers became aware of the fact that the normal distribution even under the GARCH model which increases the unconditional kurtosis was not a well particularly suitable distribution for volatility modelling. While the GARCH(1, 1) model does capture some degree of leptokurtic nature typically observed in financial time series, Bollerslev (1987), and many others have pointed out the tails are much heavier than what the normal GARCH(1, 1) would predicts and the fact that the normal distribution was unsuitable was well established. Consequently, this led to a number of non-normal distributions proposed in the literature. Notably, the Student’s t-GARCH shown by Bollerslev (1987) and many others has demonstrated that it better captures the observed kurtosis in empirical log-return time series. Nelson (1991) also suggested the family of Generalized Error Distribution implemented by Box and Tiao can also capture the heavy tail nature of the financial time series.

In the following section we will address and compare several distributions which has been proposed and used in the GARCH literature. By relaxing the normality assumption to any univariate distribution with zero mean and unit variance, we are on a path to search for a suitable distribution

$$\epsilon_t = \mathcal{D}(0,1) \tag{2.12}$$

2.3 Prior research for a better distribution

In this section, a selection of the most commonly seen univariate distributions appearing in the GARCH literature are described and compared with one another.

Normal distribution

The standard normal distribution can be expressed as

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2.13}$$

This is the distribution initially implemented by Engle (1982) in his paper Even though he was aware that other distribution were possibly more appropriate.

Student t distribution

Large number of researchers has adopted the student-t and demonstrated that this distribution better captures the excessive kurtosis. Among them are Bollerslev (1987), Bollerslev et al. (1992).

The density of the standardized student t distribution is

$$f(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi(\nu)}\Gamma(\frac{\nu}{2})} \frac{1}{\left(1+\frac{z^2}{\nu}\right)^{\frac{\nu+1}{2}}} \tag{2.14}$$

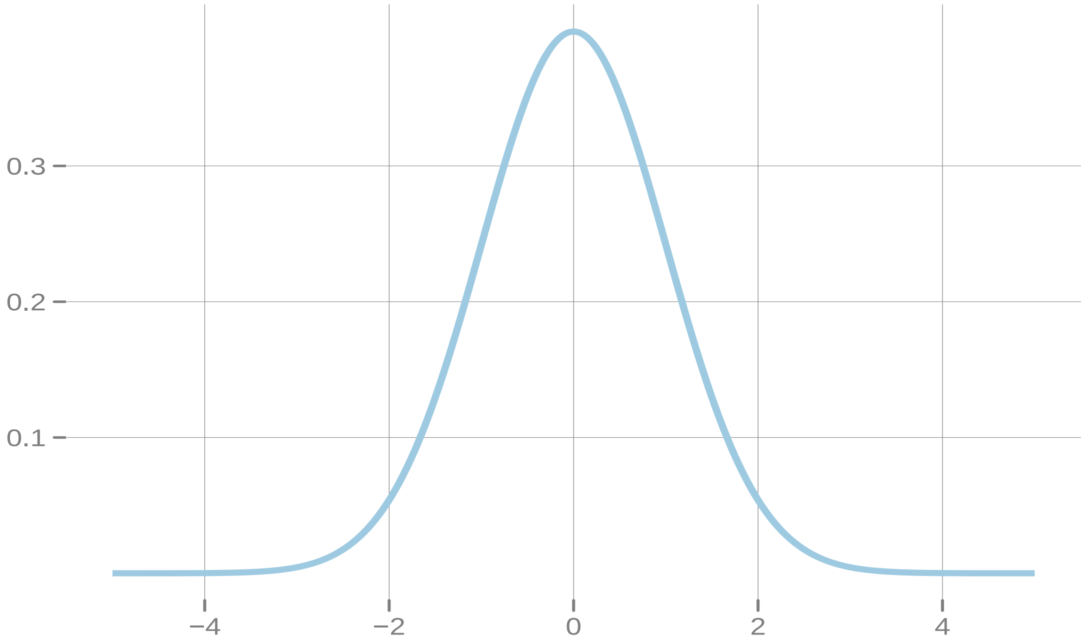

Figure 2.2: The standard Normal Distribution

Where $\nu > 2$ is the shape parameter which manoeuvres the kurtosis of the distribution and approximates the standard normal distribution as it approaches infinity.

Generalized error distribution

Several author such as Harvey (1990) and Nelson (1991) has suggested the family of Generalized Error Distribution (GED), as an alternative to the standard Normal distribution.

The advantage of this family is that the tail of the distribution is much heavier in comparison to the normal distribution yet it also include the Laplace, Normal and Uniform distribution as its special case when the shape parameter ν equals to 1, 2 and ∞ respectively.

The standardized generalized error distribution with zero mean and unit variance has the following formula

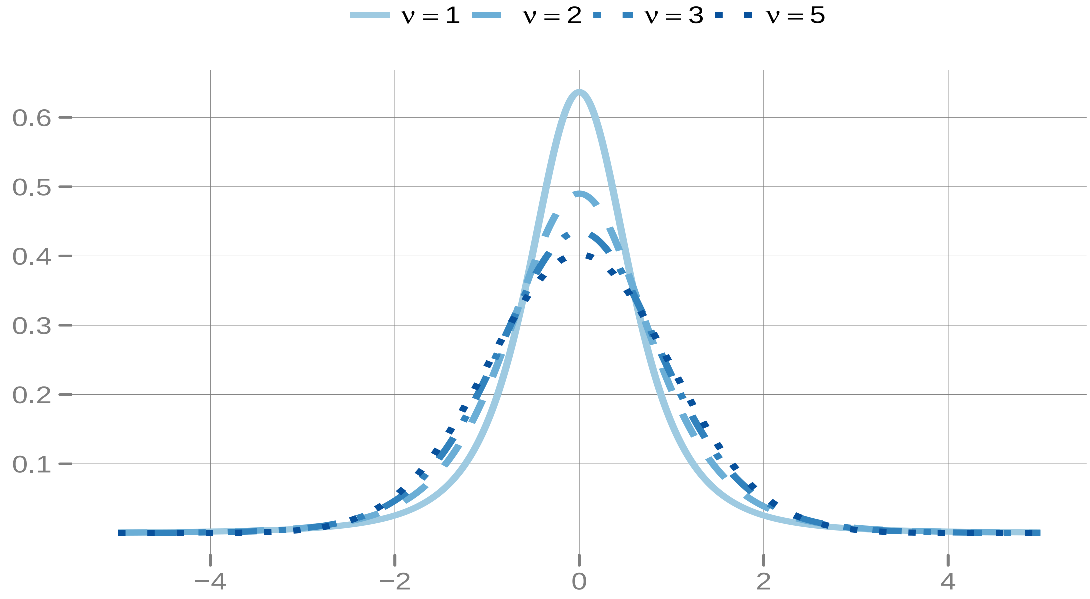

Figure 2.3: Student-t distribution with different degree of freedom.

$$
\begin{aligned}
f(x;\nu) &= \frac{\nu}{\lambda_\nu 2^{1+1/\nu}\Gamma(1/\nu)} e^{-\frac{1}{2}|\frac{z}{\lambda_\nu}|^{\nu}}, \qquad (2.15) \\
\lambda_\nu &= \left(\frac{2^{-2/\nu}\Gamma(\frac{1}{\nu})}{\Gamma(\frac{2}{\nu})} \right)^{1/2}.
\end{aligned}
$$

Where the shape parameter $\nu > 0$ determines the rate of decay of the distribution as shown in Figure 2.4.

Cauchy distribution

The Cauchy-Lorenz distribution is another distribution in statistic which possess a fat tail property. Due to the fact that the variance of the distribution is infinite, the special case standard Cauchy distribution where the location parameter $x_0 = 0$ and the scale parameter $\gamma = 1$ is usually used Inc (2012).

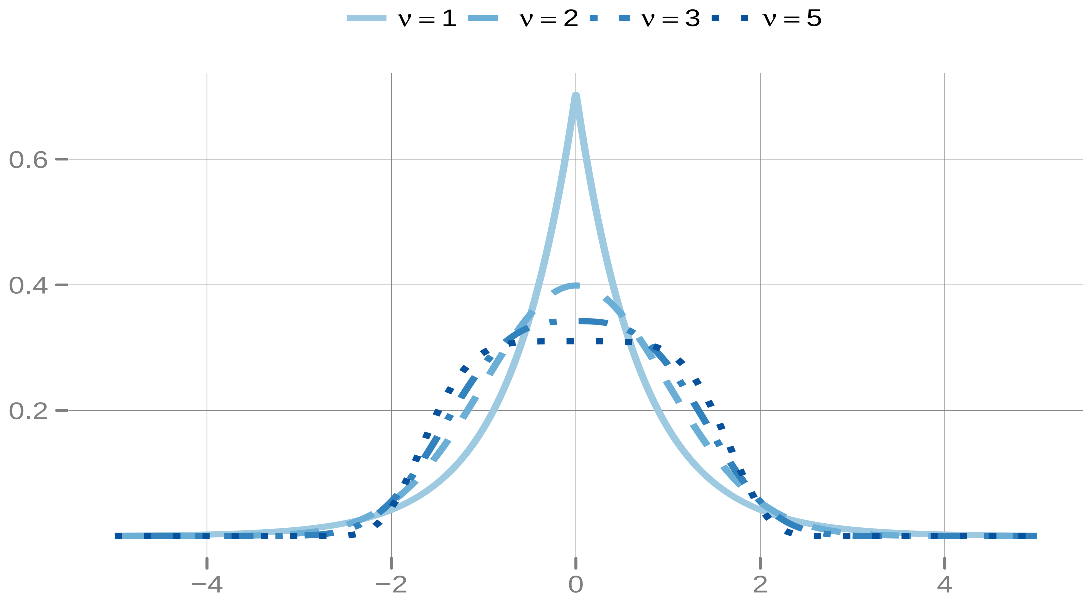

Figure 2.4: Generalised Error Distribution with different ν.

The density function of the standard Cauchy is

$$f(x) = \frac{1}{\pi(1 + x^2)} \tag{2.16}$$

Normal Inverse Gaussian and the Generalized Hyperbolic family

The Normal Inverse Gaussian is a special case of the generalized hyperbolic family, many consider the generalized hyperbolic family to be a strong candidate for volatility modelling due to its theoretical properties and flexibility. Nonetheless, the difficulty of the implementation and estimation restrict the use of this distribution in practice and only the special case Normal Inverse Gaussian is currently available. Barndorff-Nielsen (1978) expressed the density of the generalized hyperbolic distribution as

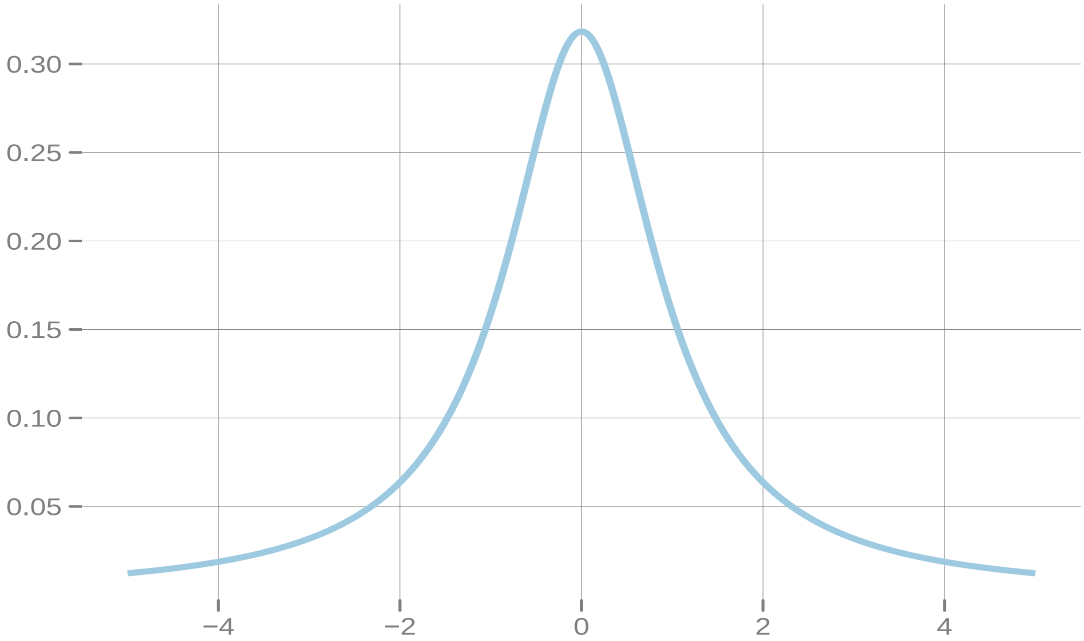

Figure 2.5: Plot of the standard Cauchy distribution.

$$
\begin{aligned}
f(x;\alpha,\beta,\delta,\mu,\lambda) = {} & \frac{\alpha_\lambda(\alpha,\beta,\delta)}{\sqrt{\delta^2+(x-\mu)^2}^{1/2-\lambda}} \\
& \times K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)e^{\beta(x-\mu)}, \qquad (2.17)
\end{aligned}
$$

$$
\alpha_\lambda(\alpha,\beta,\delta) = \frac{\sqrt{\alpha^2-\beta^2}^{\lambda}}{\sqrt{2\pi}\alpha^{\lambda-1/2}\delta^\lambda K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)}. \qquad (2.18)
$$

Where K_ν is the modified Bessel function of third order and index ν. Then the special case, the standardized normal inverse Gaussian can be obtained with the following set of parameters

$$\alpha = \left[\frac{\zeta}{1-\rho^2}\left(1+\rho^2\zeta^2(\frac{K_{\lambda+2}(\zeta)}{\zeta K_{-\lambda}(\zeta)}-\frac{1}{\zeta})/(1-\rho^2)\right)\right]^{\frac{1}{2}}, \tag{2.19}$$

$$\beta = \alpha\rho, \tag{2.20}$$

$$\delta = \frac{\zeta}{\alpha(1-\rho^2)^{\frac{1}{2}}}, \tag{2.21}$$

$$\mu = \frac{\delta^2}{\zeta} - \beta, \tag{2.22}$$

$$\lambda = -0.5. \tag{2.23}$$

We can see now the number of parameter is reduced to only ζ and ρ which controls the peakness and skewness of the distribution respectively.

However, the currently implementation of this distribution in R can give out solutions which have persistence greater than one implying the model is non-stationary. In addition, the implementation frequently runs into numerical error where the Hessian matrix is exactly singular.

Finite Mixture Normal

The use of finite mixture normal distribution was seen in Alexander and Lazar (2006) under the frequentist framework and Ausin and Galeano (2006) using the Bayesian approach and has gain considerable interest in the past decade. The popularity of the distribution arose from the fact that each component can be used to describe certain aspect of the market and provide understanding of the actors. Moreover, the difference in the location between the two component allow the distribution to capture skewness without any transformation. In both cases, a two component mixture distribution were assumed, which gave the following density function

$$\begin{aligned} f(x;\omega_1,\boldsymbol{\mu},\boldsymbol{\sigma}) = & \frac{\omega_1}{\sqrt{2\pi\sigma_1^2}}\exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \\ & + \frac{(1-\omega_1)}{\sqrt{2\pi\sigma_2^2}}\exp\left(\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right) \end{aligned} \tag{2.24}$$

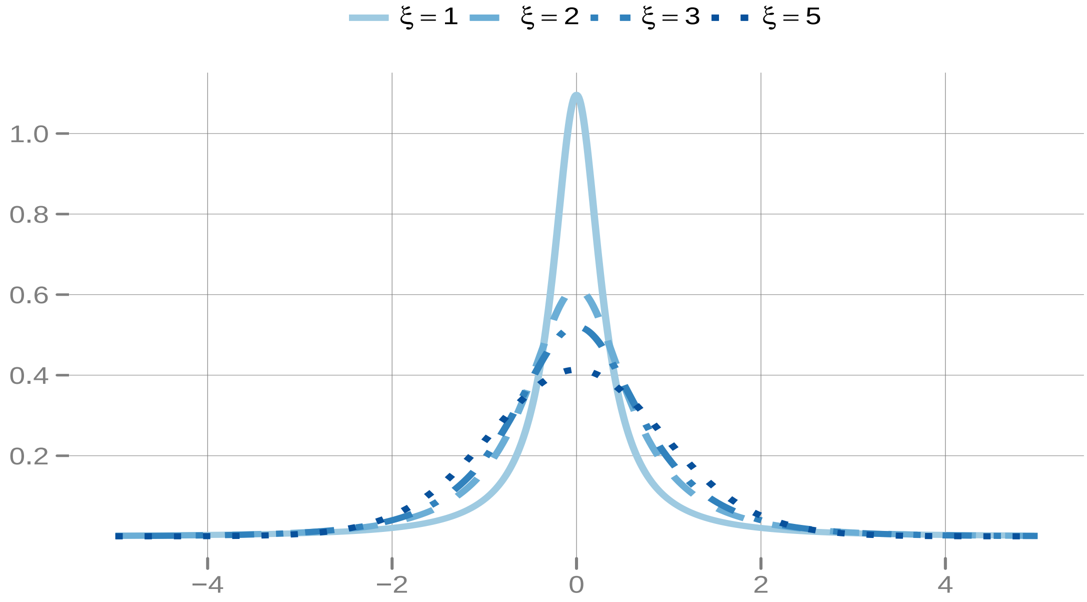

Figure 2.6: Standard Normal Inverse Gaussian with different ζ.

Despite the popular use of the finite mixture distribution in the literature, the dimension of the mixture is unknown for volatility and thus the assumption of two component is inappropriate.

2.4 Scale Normal Mixture

In this paper, we propose the use of the scale normal mixture distribution and extend the work of Chang (2010) to allow one component of the variance to change over time via the GARCH model.

The scale normal mixture is a special class of the mixture distribution family when the component consist of any number of normal distributions each with distinct variance. The use of the scale normal mixture in the financial literature first appeared in Chang (2010) where he modeled the unconditional distribution of financial time series. This research can be seen as an extension of the previous work, where we incorporate the time dimension and allow one component of the variance in the scale normal mixture to be driven by previous realization of the time series.

There are numerous advantage of the scale normal mixture over the finite mixture models. First, it encompass several other distributions when the mixing distribution $G(\theta)$ takes on different form shown below. Second, the dimension of the distribution is unrestricted, which is more appropriate as the number of groups or driving forces of volatility is unknown. Third, there are no assumption made about whether the mixing distribution is discrete or continuous.

The use of normal distribution as component density gives a simple interpretation as it can be completely described by the location, scale parameter and the proportion of each component. In addition, test statistics and asymptotic theory are all well established and can be easily modified for the use of scale normal mixture.

2.4.1 Definition

The distribution is a mixture of normal distribution where the variance has a mixing distribution function $G(\theta)$ and zero mean. It is closely related to Bayesian where the $G(\theta)$ can be interpreted similarly as a prior distribution. The distribution has the following expression

$$\begin{aligned} f(x;G) &= \int_{\Omega} \phi(x;\theta) dG(\theta), \\ &= \int_{0}^{\infty} \frac{1}{(2\pi\theta^2)^{1/2}} \exp\left(\frac{-x^2}{2\theta^2}\right) dG(\theta) \end{aligned} \tag{2.25}$$

Where $\phi(x;\mu,\theta)$ is the normal density, and $G(\theta)$ the mixing distribution of the standard deviation θ. Ω is the feasible region of θ and in this case the range is restricted to the positive real line $\mathbb{R}^+$ for standard deviation.

2.4.2 Special cases

We give a list of distribution which exists as special case of the scale normal mixture. Detailed proofs can be found in Andrews and Mallows (1974), Gneiting (1997) and West (1987).

Normal distribution

It is clear that the normal distribution is a special case if we let the mixing distribution function be:

$$G(\theta) = 1 \tag{2.26}$$

This is a useful fact that we will see in Chapter 4 where the use of this special case allows us to devise an initialization strategy which increase the computation and the robustness of the model.

Standard t-distribution

The widely used t-distribution in the financial literature is also a special case of the scale normal mixture if the mixing distribution function $G(\theta)$ takes the following expression,

$$G(\theta) = \frac{\nu}{2}\frac{(\frac{1}{2}\nu\theta)^{\frac{\nu}{2}-1}}{\Gamma(\frac{\nu}{2})}\exp\left(-\frac{1}{2}\nu\theta\right) \tag{2.27}$$

Which is the chi-squared distribution.

Laplace distribution

This is a special case of the scale normal mixture having the mixing distribution as

$$G(\theta) = \frac{1}{2\theta^2}\exp\left(-\frac{1}{2\theta}\right) \tag{2.28}$$

Logistic distribution

The logistic distribution can also be expressed as a scale normal mixture when

$$G(\theta) = 2\sum_{i=1}^{\infty}(-1)^{i-1}\frac{i^2}{\theta}\exp\left(-\frac{i^2}{2\theta}\right) \tag{2.29}$$

Generalized Error distribution

Proven by West (1987), the generalized error distribution also known as the exponential power family can be represented as a scale normal mixture for the subset $\nu \in [1, 2]$. Although, he did establish the result, the mixing distribution $G(\theta)$ was not given.

Symmetric stable and Cauchy distribution

The stable distribution proposed by Mandelbrot where the α exponent is between zero and two which includes the Cauchy distribution can also be represented by the scale normal mixture with the following mixing distribution

$$G(\theta) = \frac{1}{2}S_{\alpha/2}\left(\frac{\theta}{2}\right) \tag{2.30}$$

Where $S_{\alpha/2}$ is the density of the positive stable distribution with index $\alpha/2$ and the Cauchy distribution is obtained when $\alpha = 2$.

The Symmetric Generalised Hyperbolic Distribution and the Normal Inverse Gaussian

If we consider only the symmetric case of the generalised hyperbolic distribution, then it can be formulated if the mixing distribution is the generalized inverse Gaussian distribution.

$$\begin{aligned} G(\theta) = &\frac{\left(\frac{(\alpha^2-\beta^2)^{1/2}}{\delta}\right)^{\lambda}}{2K_\lambda\left(\delta(\alpha^2-\beta^2)^{1/2}\right)}\theta^{\lambda-1} \\ &\exp\left\{-\frac{1}{2}\left((\alpha^2-\beta^2)\theta + \frac{\delta^2}{\theta}\right)\right\} \end{aligned} \tag{2.31}$$

Where the parameters $(\alpha, \beta, \delta, \lambda)$ are the same as used in the generalized hyperbolic distribution. The class also entails the symmetric normal inverse Gaussian distribution.

Zero inflated distribution

Theoretically this is impossible as the likelihood approaches infinity and the distribution converges to the delta Dirac function if one of the support point θ is zero. Nevertheless, by having a degenerate support point sufficiently close to zero will allow us to estimate a component density which corresponds to the excessive zeroes. Although unsound in theory, a case study in chapter 5 will demonstrate the use of this special case which prove to be an exceptional strength of the scale normal mixture.

2.5 Summary

In this chapter, we give a brief background on the GARCH model, its strength as a simple model to capture volatility clustering and some of the leptokurtic nature of the time series; and the weakness of failing to allocate sufficient probability mass for extreme observations under the normality assumption.

Subsequently, this has lead to the tremendous amount of distribution proposed in the past two decade to over come this short fall. Despite the large number of researches, not one single distribution stands out in becoming the standard for GARCH modeling. One of the main reason arose from the fact that the characteristics of financial time series can vary depending on the type of asset and the market. Whether it be stock exchange, foreign exchange, or the commodity market, different distribution has found to fit particular well to a certain type of asset but not vice versa.

The lack of an accurate evaluation method also contributes to the difficulty of determining the best distribution. The fact that most of these heavy-tail distribution suffices in passing the test suggest that new evaluation method are required.

In theory, the proposed scale normal mixture is a strong candidate as it incorporates basically all the distributions we have introduced in this chapter as its special case. The fact that the normal component forms the basis of this distribution render the ability to easily modify most of tests and theories based on the normal distribution. Furthermore, the nature of decomposing the distribution provides potential to understand the market force.

Chapter 3

Non-Parametric Maximum Likelihood Estimate of Scale Normal Mixture

Even though the use of the mixture distribution in GARCH was not unprecedented and has gain considerable interest in recent years. Nevertheless these researches were exclusively restricted to finite mixture of normal or log-normal distributions and were mostly limited to just two component. In this paper, we propose the use of non-parametric maximum likelihood estimation of the scale normal mixture which posses several advantages over the prior researches.

The study of mixture distribution dates as far back as more than 100 years ago by the well-known biometrician Karl Pearson in 1894. However, given the similarity with the Bayesian paradigm, attention have not been well received until recent years due to the complex nature of the problem and the lack of technology to pursue a solution at a reasonable time frame.

The mixture distribution has found a broad range of application in statistics and many mathematical science disciplines due to its flexibility to adapt to different situation and strength to reveal group heterogeneity. These same reason also forms the motivation for us to test the applicability of the mixture distribution in volatility models. The vast amount and high degree of variability in the nature of financial time series requires a very general framework to capture the uncertainty structure; the salient nature of the mixture distribution to decompose the distribution can also serve as a tool to provide insight into the market composition through its mixture decomposition representation.

In this section we describe the necessary theory and numerical method to estimate a mixture distribution under the non-parametric maximum likelihood estimate (NPMLE) framework.

3.1 Fundamental Theory of Non-Parametric Maximum Likelihood Estimator

The density function of a mixture distribution has the general expression of

$$f(x;G) = \int_{\Omega} f(x;\theta)\, dG(\theta), \tag{3.1}$$

where $f(x;\theta)$ can be any likelihood kernel, $G(\theta)$ the mixing distribution function also known as the latent distribution of θ and Ω the feasible region of θ.

If we assume that a random sample was drawn from Equation 3.1,then we can formulate the likelihood of G as

$$l(G) = \sum_{i=1}^{n} \log\left\{\int_{\Omega} f(x_i;\theta)\, dG(\theta)\right\}. \tag{3.2}$$

Consider the problem of maximizing the objective function in Equation 3.2 then the fundamental theorem of mixture NPMLE states:

1. The solution exists and is discrete.

2. The gradient function contains all the information required to obtain a maximum likelihood estimator.

3. The gradient of the support points of the solution are zero.

4. The solution is unique

In short, the fundamental NPMLE mixture theorem not only guarantees the existence and uniqueness of the solution $\hat{G}$. Furthermore, the solution $\hat{G}$ will maximizes the $l(G)$ among all distribution functions that are defined on Ω with discrete support points θ_m and mixing proportion ω_m less than the size of distinctive sample D.

Given a finite sample, then under the fundamental NPMLE mixture theorem, we can then rewrite Equation 3.1 as

$$f(x; \boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \omega_j f(x; \theta_j), \tag{3.3}$$

and log-likelihood function 3.2 as

$$l(\boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{i=1}^{D} n_D \log \left\{ \sum_{j=1}^{m} \omega_j f(x_i; \theta_j) \right\}. \tag{3.4}$$

Where m is the number of mixture no greater than the distinct sample size D with n_D being the count of each unique set; and ω_j being the mixing proportion of each component density j characterised by a unique support point θ_j.

The $\hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\theta}}$ are the non-parametric maximum likelihood estimates equivalent to $\hat{G}$, and is a consistent estimator of $G(\theta)$ which converges in distribution.

We can thus rewrite the density function of the scale normal mixture in Equation 2.25 as

$$f(x; \boldsymbol{\omega}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \omega_j \left\{ \frac{1}{\sqrt{2\pi\theta_j^2}} \exp\left(\frac{-x^2}{2\theta_j^2} \right) \right\} \tag{3.5}$$

Where $\theta_j > 0$ is the unique standard deviation of the component density and the sum of the weight ω_j must be equal to one.

3.2 Numerical method for Non-Parametric Maximum likelihood Estimator

The idea of finding a NPMLE of a mixture distribution began in 1950, appearing in an abstract by Robbins (1950), with significant theoretical development later follow from Kiefer and Wolfowitz (1956) in 1956 whom laid out the theoretical foundation leading to the development of the field.

Despite early studies, the use of mixture models in both academic and practice were confined to finite mixture models. This is largely due to the lack of numerical methods for the computation of NPMLE mixture model, the field underwent 20 years of silence before the existence of any algorithm became available. The resurface of mixture model like the Bayesian paradigm were a result of advancement in technology and modern computing power. Nevertheless, contrary to Bayesian the fact that explicit formulas for parameter estimation are typically not available resulted in a huge literature on the estimation methodology while no consensus exists among the community.

In 2007, Wang (2007) introduced the stable and fast gradient-based Constrained Newton Method (CNM) providing non-parametric estimator of a mixture distribution. The speed of this algorithm is the reason that this study is possible, the algorithm has a speed up factor of $7 \sim 3500$. Under traditional methods, the simulation study and data analysis in the second half of the thesis would literally take month if not years to carry out.

Here in the remainder of the section, we give a brief account of the Constrained Newton Method. Readers with strong interest in the theoretical details are referred to Wang (2007).

3.2.1 The Constrained Newton Method

The main obstacle in computing any NPMLE of a mixture distribution in contrast to standard optimisation problem and finite mixture models lies in the nature of the indefinite number of dimensions of the objective function, the log-likelihood function.

Two key ingredients namely the mixing proportions ω_j and the support points θ_j in Equation 3.4 of the mixture distribution are crucial in identifying a solution. Explaining how these are computed will be the focus of the remaining chapter.

Search and allocate support points

We rely on part two of the fundamental theorem of NPMLE to search for possible support points and to determine whether a set of solution is the maximum likelihood estimate.

Let us start with the current solution G_0, and propose a new solution G_1. Then define a set of intermediate solution as

$$G_\alpha = (1-\alpha)G_0 + \alpha G_1, \quad \alpha \in [0,1] \tag{3.6}$$

Then we can compute the likelihood along this path of intermediate solution and take the derivative at $\alpha = 0$ to obtain the directional derivative from G_0 to G_1 which has the following form.

$$\begin{aligned} d(G_1; G_0) &\equiv \left. \frac{\partial l\{(1-\alpha)G_0 + \alpha G_1\}}{\partial \alpha} \right|_{\alpha=0} \\ &= \sum_{i=1}^{n} \frac{f(x_i; G_1)}{f(x_i; G_0)} - n. \end{aligned} \tag{3.7}$$

If the directional derivative is greater than zero, it implies that all likelihood computed along the path are greater than the likelihood of the current solution G_0 and thus it can not be the maximum likelihood estimate.

This is sufficient to determine whether a solution is a maximum likelihood estimate, but in order to search for possible support points let us consider the special case where G_1 is a single point and denote this ϑ. Then according to G.Lindsay (1995) the gradient function is defined to be

$$d(\vartheta; G_0) = \sum_{i=1}^{n} \frac{f(x_i; \vartheta)}{f(x_i; G_0)} - n. \tag{3.8}$$

We can then identify and add the new support point by including the point $\vartheta \in \Omega$ which gives the largest value of the gradient function.

The gradient function characterizes the NPMLE $\hat{G}$, owing to the celebrated *general equivalence theorem* which states:

$$\hat{G} \text{ maximizes } l(G) \Leftrightarrow \hat{G} \text{ minimizes } \sup_{\vartheta}\{d(\vartheta; G)\} \Leftrightarrow \sup_{\vartheta}\{d(\vartheta; \hat{G})\} = 0. \tag{3.9}$$

The theorem can be interpreted as the given solution is the maximum likelihood estimate if no new support point can be added to increase the likelihood and that each support point is the local maximum of the gradient function.

Computing mixing proportions

Given any likelihood kernel and its support points, we formulate log-likelihood function as in Equation 3.2. Let ∇ denote the vector of first-derivative operators and ∇^2 for the matrix of second-derivative operators, with respect to the mixing proportion ω only. Then the first and second order derivative of the likelihood with respect to ω can be expressed as:

$$\nabla l = \mathbf{F}^T\mathbf{1}, \tag{3.10a}$$

$$\nabla^2 l = -\mathbf{F}^T\mathbf{F}, \tag{3.10b}$$

Where $\mathbf{F}$ is the $n \times m$ log-likelihood matrix for each observation evaluate at each support point.

$$\mathbf{F} = \begin{pmatrix} \frac{f(x_1;\theta_1)}{f(x_1;\boldsymbol{\omega},\boldsymbol{\theta})} & \cdots & \frac{f(x_1;\theta_m)}{f(x_1;\boldsymbol{\omega},\boldsymbol{\theta})} \\ \vdots & \ddots & \vdots \\ \frac{f(x_n;\theta_1)}{f(x_n;\boldsymbol{\omega},\boldsymbol{\theta})} & \cdots & \frac{f(x_n;\theta_m)}{f(x_n;\boldsymbol{\omega},\boldsymbol{\theta})} \end{pmatrix} \tag{3.11}$$

Denote ω' as the update vector of ω for possible solution, $\eta = \omega' - \omega$ and $\mathbf{Q}(\boldsymbol{\omega'}|\boldsymbol{\omega},\boldsymbol{\theta}) = l(\boldsymbol{\omega},\boldsymbol{\theta}) - l(\boldsymbol{\omega'},\boldsymbol{\theta})$. By expanding $l(\boldsymbol{\omega'},\boldsymbol{\theta})$ in the Taylor series about ω to the second order and substitute Equation 3.10 we yield the expression for the change of the log-density function with respect to ω as:

$$\begin{aligned}\mathbf{Q}(\boldsymbol{\omega'}|\boldsymbol{\omega},\boldsymbol{\theta}) &\equiv \mathbf{1}^T\mathbf{F}\boldsymbol{\eta} + \frac{1}{2}\boldsymbol{\eta}^T\mathbf{F}^T\mathbf{F}\boldsymbol{\eta} && (3.12)\\ &= \frac{1}{2}\|\mathbf{F}\boldsymbol{\eta} - \vec{1}\|^2 - \frac{n}{2} && (3.13)\\ &= \frac{1}{2}\|\mathbf{F}\boldsymbol{\omega'} - \vec{2}\|^2 - \frac{n}{2} && (3.14)\end{aligned}$$

Where $\vec{2} = (2,\ldots,2)^T$ and $\|.\|$ denotes the L_2-norm. Thus, maximizing Equation 3.12 in the neighbourhood of ω can be approximated by the following linear regression problem with inequality constraints:

$$\min_{\boldsymbol{\omega'}}\|\mathbf{F}\boldsymbol{\omega'} - \vec{2}\|^2, \qquad \text{subject to } \boldsymbol{\omega'}^T\vec{1} = 1, \boldsymbol{\omega'} \geq \vec{0}. \qquad (3.15)$$

After several modifications, Wang (2010) settled with the method from Dax (1990) which transform this problem to a least squares problem with only non-negativity constraints:

$$\min_{\tilde{\boldsymbol{\omega}}}\|\mathbf{C}\tilde{\boldsymbol{\omega}}\|^2 + |\tilde{\boldsymbol{\omega}}^T\vec{1} - 1|^2, \quad \text{subject to } \tilde{\boldsymbol{\omega}} \geq \vec{0}. \qquad (3.16)$$

Where $\mathbf{C} = \mathbf{F} - \mathbf{2}$ and states that the solution of Equation 3.15 can be obtained as $\boldsymbol{\omega'} = \tilde{\boldsymbol{\omega}}/\tilde{\boldsymbol{\omega}}^T\vec{1}$ when the solution to Equation 3.16 $\tilde{\boldsymbol{\omega}}$ is given. This method is superior to previous implementation in the sense that no control parameter is required, and the result given is algebraically exact rather than a limiting case.

3.2.2 Illustration

Below we give a demonstration of some of the mixtures when fitted to a sample of real data. Four data sets from different markets were chosen to give a fair representation of the characteristics of financial time series.

The four data sets chosen were: NASDAQ the second largest market indices in the world. which contains 2,711 listings and a market capitalization of 4.5 trillion; the exchange rate between the Euro and the US dollar, the two largest currency traded; the stock price of Google and the Australian coal price from the commodity market.

Presented in Figure 3.1 are the mixing distribution of four standardized scale normal mixture distribution, we can see that all the distribution has a peak higher than the normal distribution ($\phi(0) \approx 0.4$) and the fitted density of the coal price has an extraordinary peak at zero which we will follow up with mode detail as a case study in the empirical analysis subsection 6.3.3. Furthermore, we can see in Figure 3.2 that the tail of all four distribution are much greater than the normal distribution ($\phi(4) \approx 1e{-}4$). Both figures confirms the finding of Mandelbrot (1963) that the density is higher at the peak and possess heavy tail when compared to the normal distribution.

NASDAQ		EUR/USD		Google		Coal Price	
$\hat{\theta}$	$\hat{\omega}$	$\hat{\theta}$	$\hat{\omega}$	$\hat{\theta}$	$\hat{\omega}$	$\hat{\theta}$	$\hat{\omega}$
0.1801	0.0069	0.9593	0.9934	0.0156	0.006	0.0001	0.3994
0.5047	0.1095	3.5438	0.0066	0.5326	0.4168	1.2869	0.6006
1.0045	0.8681			1.034	0.5544		
2.4768	0.0155			3.546	0.0228		

Table 3.1: Table of mixing distribution of each mixture distribution

So how does the mixture decomposition help us understand the market? Take the NASDAQ composite index which comprises 2,711 stocks for example, we can think of this as a clustering exercise where certain group of stocks are more or less volatile than the other clusters. From the mixture we can suggest that 87% of the stock behave like the normal distribution, and approximately 11% of them have a very small volatility while the remaining 1.5% are very risky. Epps and Epps (1976) also had a similar interpretation arose from the behavioural nature of investor, where each component represent a different of group traders or actors in the market having different investment motivation or perception of return and risk. Ball and Torous (1983) had a slightly different interpretation, where they asserted that each component to represent different market state. For example, under the two component mixture, one of the component can represent the state under normal condition while the other component can represent unusual market state such as financial crisis which may depend on the arrival of new and relevant information in the market. However, this statement does not seem very nature since the change of market condition is already incorporate in the GARCH model, the mixture itself does not change rapidly over time.

3.3 Summary

In this section we touched on the fundamental theorem of non-parametric maximum likelihood estimation of mixture distribution. This theorem forms the basis for proving the existence of the solution and the framework for a numerical method under the nonparametric likelihood estimation.

We also shown how to formulate the scale normal mixture distribution under the NPMLE setting and the relationship of the theorem to the Constrained Newton Method of Yong.

Several illustration based on real data were given, and demonstrate the use of the scale normal mixture GARCH. This solidifies our position that the mixture representation can give us an understanding of the nature of risk in the market. An other characteristic lack by the GARCH model where only the evolution mechanism of the volatility is addressed but not the driving source of volatility.

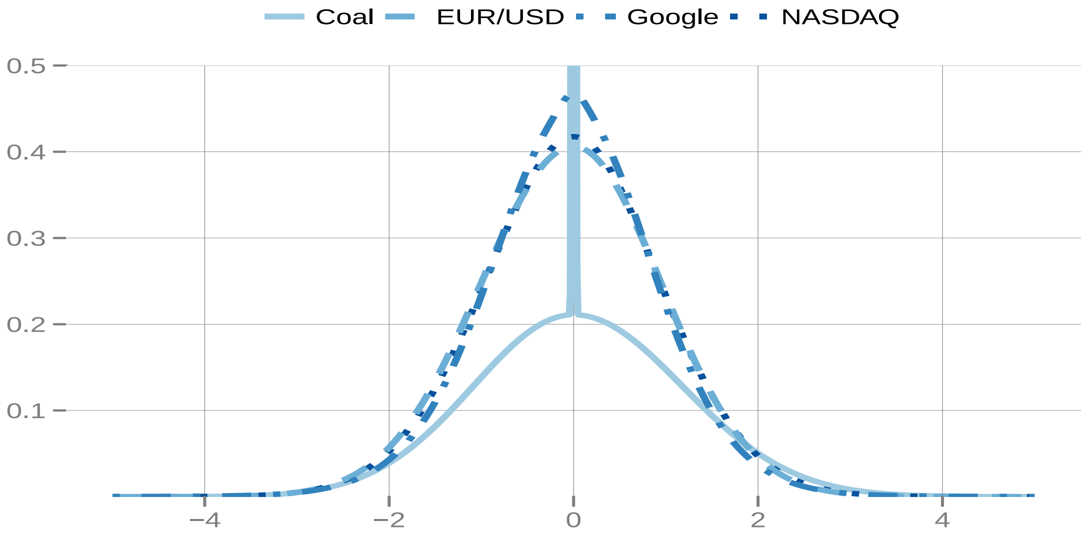

Figure 3.1: Scale Normal Mixture Distribution with different mixing distribution.

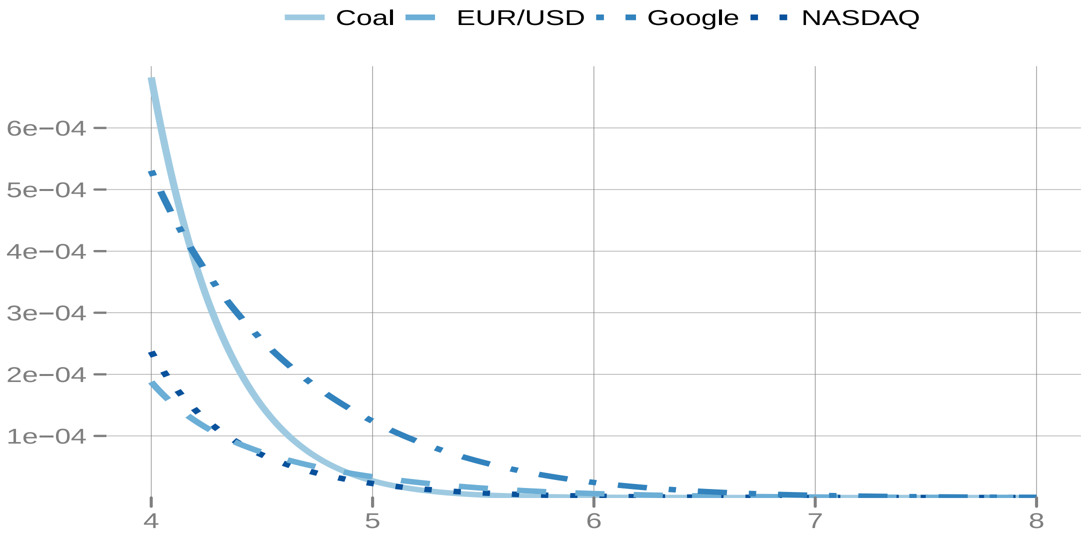

Figure 3.2: Tail of Scale Normal Mixture Distribution with different mixing distribution.

Chapter 4

The Semi-parametric SNM-GARCH

This chapter is devoted to the details and the implementation of the SNM-GARCH, with the necessary mathematics and the background on the CNM-MS algorithm described.

The sections will begin by incorporating skewness in the proposing distribution to account for its common presence in financial time series. This is then followed by a section on the essential requirement for the integration of the scale normal mixture into the GARCH framework along with a new initialisation strategy. Finally, we describe the CNM-MS algorithm for the semi-parametric model with implementation details and finally end with a pseudo code diagram.

4.1 Capturing skewness

Another commonly feature regarding higher moments of financial time series distribution is the presence of skewness. As pointed out in Chapter 1, the ability to ensnare the higher moments of the time series is the key to producing accurate forecast.Many different type of transformation and skewness parametization exist in literature, here two common general skewness parametization are considered.

4.1.1 Cumulative distribution function parametrization

Azzalini (1995) proposed the class of skewed normal density, also applicable to all continuous distribution which are symmetry about zero in 1985. The method involves transforming the density by the respective distribution function, which yields the skewed density preserving many theoretical properties

$$f(x;\lambda) = 2f(x)F(x\lambda) \tag{4.1}$$

Where $f(x)$ and $F(x)$ denotes the density and distribution function respectively. The skewness parameter λ can be any real number and governs the degree of skewness, the transformation retains the original distribution when $\lambda = 0$. However, the use of the distribution function makes this form of parametrization difficult to implement in certain circumstances. In addition, the moments and the cumulative distribution function does not come readily in analytical form.

4.1.2 Scale parametrization

The second parametization in consideration belongs to Fernandez and Steel (1998) which can be applied to all continuous, symmetric, uni-modal distribution centered at zero. The underlying idea is to introduce inverse scale factors in the positive and the negative orthant and achieve the skewness by allocate mass from one side of the mode to the other.

$$f(x|\gamma) = \frac{2}{\gamma + \gamma^{-1}} \left[f(x\gamma)H(-x) + f(x\gamma^{-1})H(x) \right], \tag{4.2}$$

Where $f(x)$ is the symmetrical distribution of interest; $0 < \gamma < \infty$ is the skewness parameter which governs the symmetry of the transformed distribution; and $H(.)$ the Heaviside function. However, in contrast to the cumulative distribution function parametization, the symmetric case of the distribution is obtained when $\gamma = 1$. Note, we have implemented the half-maximum convention of the Heaviside function where $H(0) = 0.5$.

Both parametization has its merit, we have decided to implement the scale form of skewness for the following reasons. (1) The implementation of this type of parametization under the CNM where the derivatives are required is exact as we do not need to compute the error function numerically; (2) By scale the positive and the negative orthant the mode at zero is preserve. If this property does not hold then it is possible for the resulting mixture distribution to be multi-modal (3) It offers an interpretation of the skewness parameter γ via Equation 4.3 below

$$\frac{P(x \geq 0|\gamma)}{P(x < 0|\gamma)} = \gamma^2 \tag{4.3}$$

From which we can see that the skewness parameter controls the allocation of mass to each side of the mode. This same equation also implies that the reciprocal of the skewness parameter will give the mirror distribution around the zero. That is

$$f(x|\lambda) = f(-x|1/\lambda) \tag{4.4}$$

Another nice property of this transformation is that the moments of the distribution are readily available as obtained by Steel and Fernandez. Provided the integral exists, then the moments of the distribution after the skewness transformation can be computed by

$$\mathbb{E}(x^r|\gamma) = M_r \frac{\gamma^{r+1} + \frac{(-1)^r}{\gamma^{r+1}}}{\gamma + \frac{1}{\gamma}}, \tag{4.5}$$

Where

$$M_r = 2\int_0^\infty s^r f(s)\, ds. \tag{4.6}$$

Which will prove to be extremely useful in the GARCH setting demonstrated in the later part of the chapter where we will use the moments to ensure the distribution is zero meaned with unit variance.

To capture the skewness in the data, the scale normal mixture is transformed by Equation 4.2 to give the expression of skewed scale normal mixture distribution which will be used in the proposing GARCH model

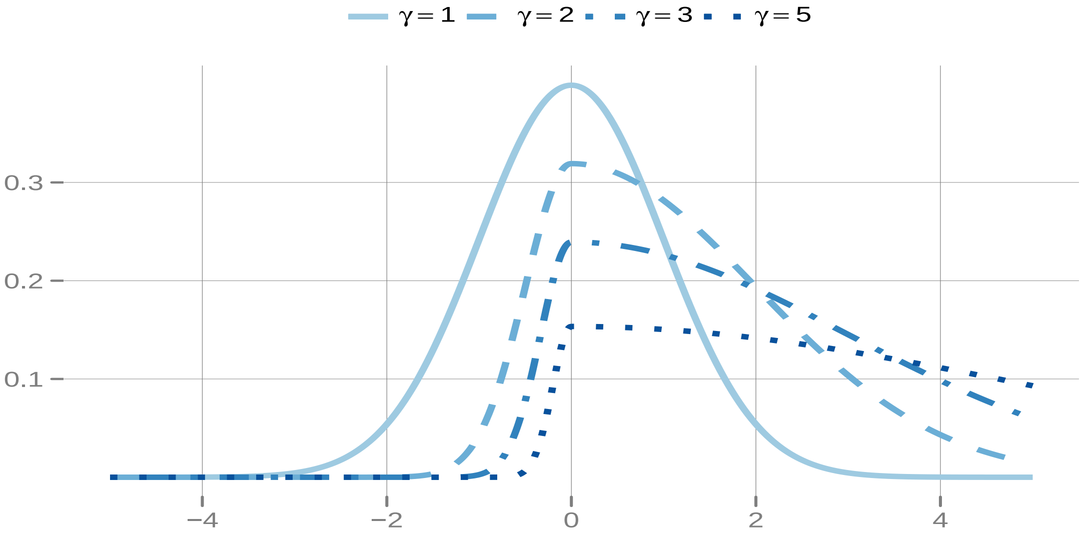

Figure 4.1: Skewed Normal Distribution with different γ

$$f(x;\boldsymbol{\omega},\boldsymbol{\theta},\gamma)=\sum_{j=1}^{m}\omega_j\frac{2}{\gamma+\gamma^{-1}}\frac{1}{\sqrt{2\pi\theta_j^2}} \times \exp\left(\frac{-[x(\gamma H(-x)+\gamma^{-1}H(x))]^2}{2\theta_j^2}\right) \tag{4.7}$$

In Figure 4.1, depicted are the normal distribution and the skewed normal distributions transformed with $\lambda = 1, 2, 3,$ and 5. The mode at zero is preserved, and the transformation shifts the mass on the left towards the right. The transformation not only alters the skewness, but also the kurtosis and we can see the tail on the right hand side are becoming much much more heavier as we increase λ.

4.2 Estimating σ_1^2

As briefly mentioned in chapter one, since the conditional variance and observation prior to the first observation are not observed they need to be initialised and estimated.

Current practice commonly use the first few observation to calculate the conditional variance and then discard them Tsay (2002). Others take the unconditional variance of the data as the initial value in Fiorentini et al. (1996). In this section we propose and compare several different initialisation strategy.

Unconditional expectation

In the work of Fiorentini et al. (1996), also implemented in the fGarch package Wuertz et al. (2012), the initial value σ_1^2 is given as

$$\sigma_1^2 = \frac{1}{T}\sum_{t=1}^{T} x_t^2 \tag{4.8}$$

which provide a consistent estimate of the unconditional variance and a proxy to the conditional variance. However, this can be badly fit and biased if the initial conditional variance is much greater or smaller than the unconditional variance.

In our implementation of the GARCH model, we have allow two different method of estimating the initial variance.

Back filter

In this method, we back estimate the initial variance by inverting the GARCH equation and use the latest set of parameter to obtain the expected variance at $t = 1$ given by the following equation.

$$\mathbb{E}[\sigma_1^2|\boldsymbol{\beta}] = \frac{\mathbb{E}[\sigma_2^2|\sigma_3^2, \cdots, \sigma_T^2, \boldsymbol{\beta}] - \alpha_0 - \beta_1 x_1^2}{\alpha_1} \tag{4.9}$$

Where $\mathbb{E}[\sigma_2^2|\sigma_3^2, \cdots, \sigma_T^2, \boldsymbol{\beta}]$ is the second conditional standard deviation given the last set of parameters. Thus we obtain an initial conditional variance which is consistent with the volatility equation and the first observation x_1.

However, the drawback of this method lies with the fact that the likelihood is not guaranteed to increase monotonically with each iteration and may result in the optimisation algorithm to diverge.

Local estimation

The second method makes assumption that the variation of the volatility in the very short run is relatively constant, and place a common volatility for the first k observation which is to be estimated. This result in the new set of volatility equation as:

$$\sigma_t^2 = \begin{cases} \sigma_c^2 & \text{if } t \leq k \\ \alpha_0 \sum_{i=0}^{t-k-1} \alpha_1^i + \alpha_1^{t-k} \sigma_c^2 + \beta_1 \sum_{i=0}^{t-k-1} \alpha_1^i x_{t-i-1}^2 & \text{if } t > k \end{cases} \tag{4.10}$$

Comparing this to Equation 2.10, we can see this merely replace the index 1 with k. The motivation for this arose from the fact that if we try to estimate the expected variance at $t = 1$, we result in an estimate that is largely driven by the initial observation as the time series can be very noisy. Although squaring the first observation x_1 give a consistent estimator of the expected variance but it is too volatile and thus inappropriate for statistical inference. Given the estimate of the persistence is generally very close to one, it is plausible to assume that the volatility is close to constant over a small period of 5 observations. This greatly improves the fit as we can see in Figure 4.2 and also a much more stable initialisation regardless where we start the times series sample.

Figure 4.2 shows the fit of the three initialisation method outlined above on a sample of the SNP500 data. We can clearly see that the typical assumption of using the unconditional standard deviation may be unsuitable as it is not time invariant and the fit can be biased if the initial volatility severely is different to the unconditional variance.

We have adopted the local estimation as our default method due to its numerical stability and the invariant property regardless whether the sample starts.

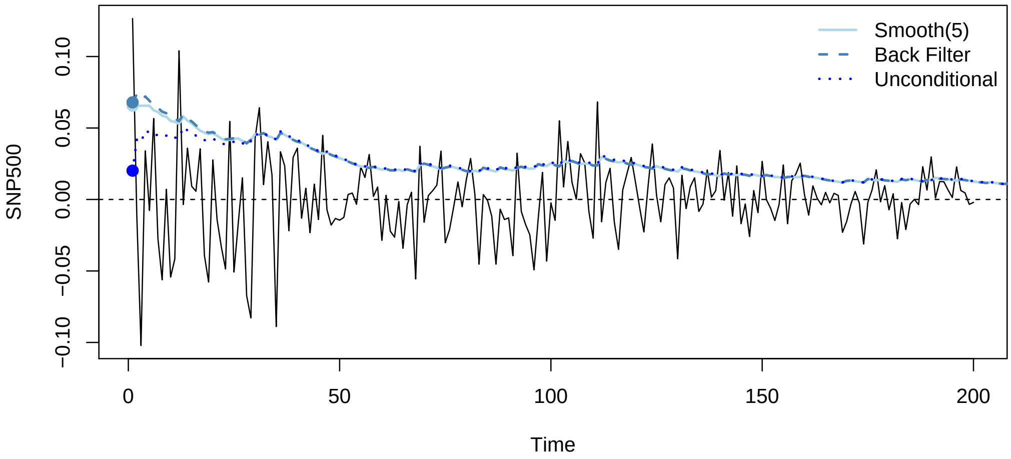

Figure 4.2: Demonstration of different initialisation strategy.

4.3 Model and further requirements

Assuming that the model has a skewed scale normal distribution from Equation 4.7, then we can formulate the likelihood of each observation of a SNM-GARCH(1, 1) model as:

$$f(x_t|\boldsymbol{\omega}, \boldsymbol{\theta}, \sigma_t, \gamma) = \sum_{i=1}^{m} \omega_i \frac{2}{\gamma + \gamma^{-1}} \frac{1}{\sqrt{2\pi\theta_i^2\sigma_t^2}} \exp\left[\frac{x_t[\gamma H(-x_t) + \gamma^{-1}H(x_t)]}{2\theta_i\sigma_t}\right]^2 \tag{4.11}$$

Where σ_t^2 is computed using the local estimation method in Equation 4.10 and can be interpreted as the time varying variance component, while the support points $\boldsymbol{\theta}$ are the non-time varying variance component.

Several further restrictions are required to integrate the distribution into the GARCH model. In order to avoid identification problem between the time varying variance σ_t^2 and the non-time varying variance θ, the distribution of ϵ_t in Equation 2.12 must be zero mean and unit variance.

Based on Equation 4.6, we can first solve the integral for the first two moments

$$
\begin{aligned}
M_1 &= 2\int_0^\infty x\, f(x;\boldsymbol{\omega},\boldsymbol{\theta})\, dx \\
&= 2\int_0^\infty x \times \sum_{i=1}^m \omega_i \left[\frac{1}{\sqrt{2\pi\theta_i^2}} \exp\left\{\frac{-x^2}{2\theta_i^2}\right\}\right] dx \\
&= \sum_{i=1}^m \omega_i \frac{2}{\sqrt{2\pi\theta_i^2}} \int_0^\infty x \exp\left\{\frac{-x^2}{2\theta_i^2}\right\} dx \\
&= \sum_{i=1}^m \omega_i \frac{2}{\sqrt{2\pi\theta_i^2}} \left[\frac{2\theta_i^2\Gamma(1)}{2}\right] \\
&= \sum_{i=1}^m \omega_i \sqrt{\frac{2\theta_i^2}{\pi}} \qquad (4.12)
\end{aligned}
$$

and

$$
\begin{aligned}
M_2 &= 2\int_0^\infty x^2\, f(x;\boldsymbol{\omega},\boldsymbol{\theta})\, dx \\
&= 2\int_0^\infty x^2 \times \sum_{i=1}^m \omega_i \left[\frac{1}{\sqrt{2\pi\theta_i^2}} \exp\left\{\frac{-x^2}{2\theta_i^2}\right\}\right] dx \\
&= \sum_{i=1}^m \omega_i \frac{2}{\sqrt{2\pi\theta_i^2}} \int_0^\infty x^2 \exp\left\{\frac{-x^2}{2\theta_i^2}\right\} dx \\
&= \sum_{i=1}^m \omega_i \frac{2}{\sqrt{2\pi\theta_i^2}} \left[\frac{\theta_i^2}{2} \times \sqrt{2\pi\theta_i^2}\right] \\
&= \sum_{i=1}^m \omega_i \theta_i^2 \qquad (4.13)
\end{aligned}
$$

Then we substitute the solution into Equation 4.5 to obtain the moments of the skewed scale normal mixture

$$
\mathbb{E}[x_t|\boldsymbol{\omega},\boldsymbol{\theta},\gamma] = \left(\sum_{i=1}^m \omega_i \sqrt{\frac{2\theta_i}{\pi}}\right) \frac{\gamma^2-\gamma^{-2}}{\gamma+\gamma^{-1}} \qquad (4.14)
$$

$$\mathbb{E}[x_t^2|\boldsymbol{\omega}, \boldsymbol{\theta}, \gamma] = \left(\sum_{i=1}^{m} \omega_i \theta_i^2\right) \frac{\gamma^3 + \gamma^{-3}}{\gamma + \gamma^{-1}} \tag{4.15}$$

The condition for GARCH estimation is fulfilled if the moment criteria $\mathbb{E}[x_t|\boldsymbol{\omega}, \boldsymbol{\theta}, \gamma] = 0$ and $\mathbb{E}[x_t^2|\boldsymbol{\omega}, \boldsymbol{\theta}, \gamma] = 1$ are satisfied.

4.4 The CNM-MS algorithm

In 2010, three algorithm namely CNM-AP, CNM-PL and CNM-MS adopting different strategy were proposed by Wang (2010) for fitting semi-parametric mixture models. The CNM-AP alternates the optimisation of the parameter and the mixture separately; while the CNM-PL maximises the profile likelihood and the CNM-MS modifies the mixture before optimise the likelihood globally.

The CNM-MS was chosen in this research due to its speed and suitability of the project outlined in the same paper containing numerical studies with other available algorithms. The CNM-MS utilises the CNM outlined in section 3.2 for modifying the non-parametric mixture distribution at every iteration while a further optimisation step is conducted to ensure global maximum.

The algorithm can be described as:

Step 1:

Set i = 0, and initialise the parameter $\beta_0 \in \mathbb{R}^r$ and mixing distribution G_0.

Step 2:

Use one step of the CNM to allocate new support points and solve for mixing proportion then update $(\boldsymbol{\omega}_s, \boldsymbol{\theta}_s)$ to $(\boldsymbol{\omega}_s^+, \boldsymbol{\theta}_s^+)$.

Step 3:

Use an optimization algorithm (BFGS) to update the parameters $(\boldsymbol{\omega}_s^+, \boldsymbol{\theta}_s^+, \boldsymbol{\beta}_s)$ to $(\boldsymbol{\omega}_{s+1}, \boldsymbol{\theta}_{s+1}, \boldsymbol{\beta}_{s+1})$ which maximises the likelihood function.

Step 4:

set i = i + 1 and stop if converged.

In brief, at every iteration the CNM is run to ensure that (1) the support set is constantly modified, with both an expansion step to add new support points with greater than zero gradient and a contraction step to remove support points with zero masses; (2) to ensure that the mixing distribution is an interior set corresponding to the probability simplex and to secure the use of the unconstrained optimisation in step three. The BFGS is then used to locate the global maximum of the likelihood with the given dimension of the mixture distribution.

4.5 Initialize $\boldsymbol{\beta}_0$, G_0 and the search space

The initial value of most optimization algorithm has the potential to affect the speed of convergence and the ability to locate the global maximum, therefore the choice of the initial value should be carefully considered. In contrast to the EM algorithm typically used for finite mixture, the solution of the CNM algorithm is fairly robust to the selection of the initial parameters. Nevertheless, given the flat likelihood surface of GARCH models, it is generally wise to give initial values that are sensible and close to the final solution.

GARCH parameter and mixing distribution

We have initialise the parameters β_0 as the solution obtained by fitting a GARCH model assuming a standard normal distribution which is also the initial value of the mixture G_0. This is different to other methods such as the EM algorithm in the literature where a large number of support points were given initially, then were collapsed to give the final solution.

The reasoning for the growing strategy oppose to the pruning strategy is based on the idea that we start off from the special case of the scale normal mixture then optimise the model if the assumption is invalid. This allow us to increase the likelihood function monotonically while also having the normal distribution always as the special case if the likelihood can not be increased. In addition, the solution of the scale normal mixture should be sufficiently close to the solution under the normality assumption and thus increase the speed of the algorithm.

In addition to the initial value, the stopping criteria is also crucial to the solution of the optimisation. Currently, we have adopted a tolerance of $1e^{-8}$ which in typical case gives a 4 decimal place accuracy for the parameters estimated.

Search for grid points in the feasible region

In order to find the support points, the range of the feasible region Ω of the support points θ must be given so that the algorithm can compute the gradient over the range to identify potential support points and determine whether a solution is the NPMLE estimate. More precisely, the range of the feasible region and the number of grid points are required. The large the number of grid points, the greater the chance of allocating the global solution, however it also increases the computation burden.

In the case of the scale normal mixture, the support points which are the standard deviation of the distribution must clearly be positive. Furthermore, the support points must be strictly greater than zero in which the likelihood would otherwise be infinite.

We have not chosen the points in the feasible region uniformly, rather an exponential grid is used where the support points are more frequent when close to zero then the interval size increases as it tends to infinity. The exact form of the grid point is $2^{(k_1 \sim k_2)}$ where k_2 is the largest integer where the expression will not evaluate as infinite. This can be obtained in R as `(.Machine$double.max.exp - 1)`. Instead of arbitrarily choosing a very small value for k_1 as the lower bound, we define the minumum of the search grid as the square root of the smallest observation omitting zero divided by three. This allows us to avoid numerical problem of infinite likelihood, while at the same time allowing us to nest a zero inflated model as a special case if there are excessive number of zeroes present in the data. By default, we have used 5000 grid points over the range of zero to infinite.

This way, we can speed up the computation without having to search too much in the part of the feasible region with very unlikely support points, yet still allow the algorithm to search the whole sample space.

4.6 Implementation

To set up and estimate the SNM-GARCH, the log-density matrix L which is equivalent to the log of Equation 3.11 without the normalising constant $f(x; \boldsymbol{\pi}, \boldsymbol{\theta})$ and its derivative with respect to all parameters and support points are required. These are used in the BFGS to maximise the global optimisation and to calculate the gradient function for locating the support points.

Each element of the matrix $L_{t,i}$ has the following expression

$$\begin{aligned} f(x_t; \omega, \theta_i, \sigma_t, \gamma) = & \frac{2}{\gamma + \gamma^{-1}} \frac{1}{\sqrt{2\pi\theta_i^2\sigma_t^2}} \\ & \exp\left\{ \frac{-x_t^2[\gamma^2 H(-x_t) + \gamma^{-2} H(x_t)]}{2\theta_i^2\sigma_t^2} \right\} \end{aligned} \tag{4.16}$$

taking the log of the matrix which give us,

$$\begin{aligned} \ell(x_t; \omega, \theta_i, \sigma_t, \gamma) = & \log(2) - \frac{1}{2}\log(2\pi) - \log(\gamma + \gamma^{-1}) - \log(\theta_i) \\ & - \log(\sigma_t) - \left[\frac{x_t[\gamma H(-x_t) + \gamma^{-1} H(x_t)]}{2\theta_i\sigma_t} \right]^2 \end{aligned} \tag{4.17}$$

Then the respective derivatives of each element to each different parameters are:

Non-time-varying variance:

$$\frac{\partial \ell}{\partial \theta_i} = -\frac{1}{\theta_i} + \frac{[x_t(\gamma H(-x_t) + \gamma^{-1} H(x_t))]^2}{\theta_i^3\sigma_t^2} \tag{4.18}$$

Time-varying variance:

$$\frac{\partial \ell}{\partial \sigma_t} = -\frac{1}{\sigma_t} + \frac{[x_t(\gamma H(-x_t) + \gamma^{-1} H(x_t))]^2}{\theta_i^2 \sigma_t^3} \tag{4.19}$$

Skewness:

$$\frac{\partial \ell}{\partial \gamma} = \frac{1-\gamma^{-2}}{\gamma+\gamma^{-1}} - \frac{[x_t(\gamma H(-x_t) - \gamma^{-3} H(x_t))]^2}{\theta_i^2 \sigma_t^2} \tag{4.20}$$

GARCH parameter:

The derivatives of the log-density with respect to the GARCH parameters $\boldsymbol{\beta} = (\alpha_0, \alpha_1, \beta_1, \sigma_c)$ can be simplified using the product rule.

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{\partial \ell}{\partial \sigma_t} \times \frac{\partial \sigma_t}{\partial \boldsymbol{\beta}}$$

The parameter space reduces to $\boldsymbol{\beta} = (\alpha_0, \alpha_1, \beta_1)$ if the back estimation method is used, and the derivative for $t = 1$ are all zero. The following derivatives are defined for the smoothing window initialization method.

If $t <= k$,

$$\frac{\partial \sigma_t}{\partial \alpha_0} = \frac{\partial}{\partial \alpha_0}\sigma_c = 0 \tag{4.21}$$

$$\frac{\partial \sigma_t}{\partial \alpha_1} = \frac{\partial}{\partial \alpha_1}\sigma_c = 0 \tag{4.22}$$

$$\frac{\partial \sigma_t}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\sigma_c = 0 \tag{4.23}$$

$$\frac{\partial \sigma_t}{\partial \sigma_c} = \frac{\partial}{\partial \sigma_c}\sigma_c = 1 \tag{4.24}$$

If $t > k$,

$$\begin{aligned}
\frac{\partial \sigma_t}{\partial \alpha_0} &= \frac{\partial}{\partial \alpha_0} \alpha_0 \sum_{i=0}^{t-k-1} \alpha_1^i + \alpha_1^{t-k}\sigma_c^2 + \beta_1 \sum_{i=0}^{t-k-1} \alpha_1^i x_{t-i-1}^2 \\
&= \frac{1}{\sigma_t^2} \sum_{i=0}^{t-2} \alpha_1^i && (4.25) \\
\frac{\partial \sigma_t}{\partial \alpha_1} &= \frac{\partial}{\partial \alpha_1} \alpha_0 \sum_{i=0}^{t-k-1} \alpha_1^i + \alpha_1^{t-k}\sigma_c^2 + \beta_1 \sum_{i=0}^{t-k-1} \alpha_1^i x_{t-i-1}^2 \\
&= \frac{1}{\sigma_t^2} \left(\alpha_0 \sum_{i=0}^{t-2} i\alpha_1^{i-1} + (t-1)\alpha_1^{t-2}\sigma_t^2 + \beta_1 \sum_{i=0}^{t-2} i\alpha_1^{i-1} x_{t-i-1}^2 \right) && (4.26) \\
\frac{\partial \sigma_t}{\partial \beta_1} &= \frac{\partial}{\partial \beta_1} \alpha_1 \sum_{i=0}^{t-k-1} \alpha_1^i + \alpha_1^{t-k}\sigma_c^2 + \beta_1 \sum_{i=0}^{t-k-1} \alpha_1^i x_{t-i-1}^2 \\
&= \frac{1}{\sigma_t^2} \sum_{i=0}^{t-2} \alpha_1^i x_{t-i-1}^2 && (4.27) \\
\frac{\partial \sigma_t}{\partial \sigma_c} &= \frac{\partial}{\partial \sigma_c} \alpha_1 \sum_{i=0}^{t-k-1} \alpha_1^i + \alpha_1^{t-k}\sigma_c^2 + \beta_1 \sum_{i=0}^{t-k-1} \alpha_1^i x_{t-i-1}^2 \\
&= \frac{2\alpha_1^{t-1}}{\sigma_t} && (4.28)
\end{aligned}$$

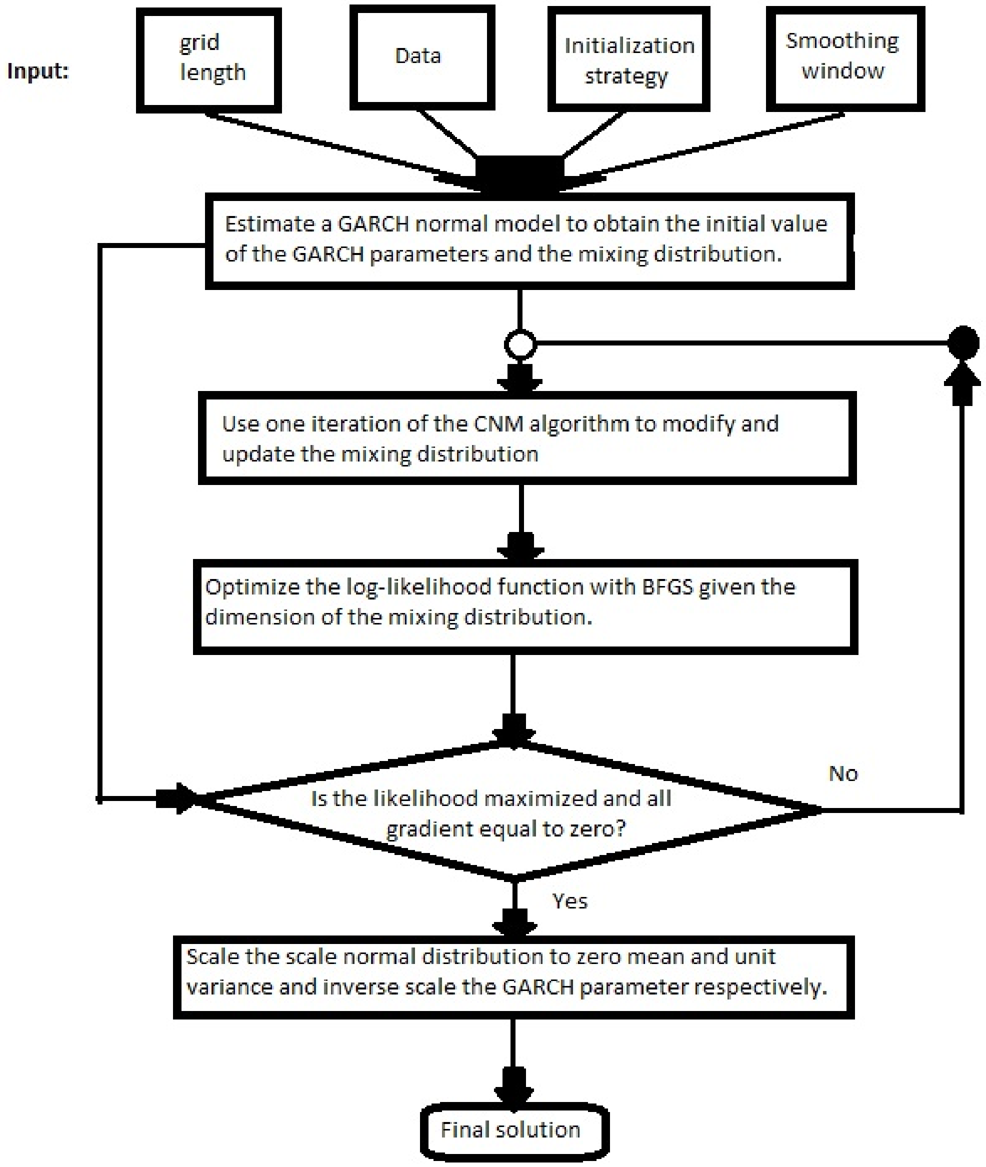
Input:
grid length
Data
Initialization strategy
Smoothing window
Estimate a GARCH normal model to obtain the initial value of the GARCH parameters and the mixing distribution.
Use one iteration of the CNM algorithm to modify and update the mixing distribution
Optimize the log-likelihood function with BFGS given the dimension of the mixing distribution.
Is the likelihood maximized and all gradient equal to zero?
No
Yes
Scale the scale normal distribution to zero mean and unit variance and inverse scale the GARCH parameter respectively.
Final solution

Figure 4.3: The pseudo code diagram for fitting a SNM-GARCH

4.7 Summary

This chapter began by extending the scale normal mixture to incorporate skewness which is commonly observed in financial time series. Although the scale parametization has many desirable properties, it would however be of theoretical interest to solve for a skewness transformation in which the non-symmetrical stable and the normal inverse Gaussian as its special case.

We also proposed new solutions to the problem of initializing the conditional variance which are unobserved at the beginning of the time series. Although the current implementation seems to work well, but the fact that we are making the constant variance assumption does not seem very elegant in theory. Future work may attempt to resolve the numerical obstacle of the back estimation, and extend the back propagation to the full length of the time series.

A lengthy description was given to the initialization and the mathematical details of the SNM-GARH model. We discuss how the choice of the initial value of the mixing distribution and the GARCH parameter are made in order to speed up the computation and also always have the normal distribution as special case. The selection of the search space or the feasible region of the support points was also discussed, where we balanced the trade-off between obtaining an exact solution with an incredibly large number of support points or a sufficient solution with several magnitude of speed-up.

The chapter should provide both the background and the mathematical details required for implementing a SNM-GARCH model in-house.

Chapter 5

Simulation Study

In this chapter we conduct a simulation to study the applicability of the scale normal mixture. The aim of the simulation is to see how well each distribution performs in recovering the true underlying density. Given it is almost impossible in practice to actually retrieve the data generating distribution, we would like to have a model that gives us a good approximation to the true distribution regardless of the form it takes.

Although we have shown that almost all distribution used in the literature can be a special case of the scale normal mixture, however the flexibility of the distribution may result in over-fitting to the data. This is one of the motivation to see whether this problem exists, second we would also want to compare how well the distributions performs relative to each of other distributions.

5.1 Study design

In order to assess how each model recovers under different distribution assumption, we first generate a GARCH process assuming a particular distribution then fit GARCH model of various distribution assumption. Then we calculate several distance statistics to evaluate the similarity of the fitted distribution and the true distribution. The process of the simulation is

1. Generate a GARCH process of size n assuming a specific distribution

from the following list

- Standard Normal distribution
- Student t distribution with $\nu \in (2, 100]$.
- Generalised Error distribution with $\nu \in [1.5, 2.5]$.
- Normal Inverse Gaussian distribution with $\zeta \in [0.1, 100]$.

We have not simulate from a scale normal mixture distribution because the number of parameters is indefinite and can be large. Thus it is hard to determine the range of each parameter which will be feasible for financial data.

2. Fit all the GARCH model with other competing distribution to the simulated data.

3. Compute the following distance statistics between the true distribution and the fitted distribution.

   - Hellinger's distance
   - Kullback-Leibner Divergence
   - Mean integrated squared error

4. Check if all the statistics were computed, if not then repeat from step one.

All the simulation are repeated one hundred times, and the mean of each statistics are returned.

Given the varying degree of complexity between distributions, we also simulate the GARCH process under 5 different sample size (50, 100, 200, 500, 1000) to assess the robustness of the model. This would also allows to understand which model is most suitable when we have data of various sample size. In particular for low frequency data where the number of observation is small.

5.2 Evaluation Criteria

Three different statistic listed below are adopted to examine the models.

Hellinger's distance

The Hellinger distance is a statistical measure of similarity between two distribution from the f-divergence family, and it can be calculated as:

$$D_{hd}(f||g) \overset{def}{=} \frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right) dx \tag{5.1}$$

The Hellinger's distance give more weights to the tails in comparison to the Kullback-Leibner Divergence below and thus it would be more suitable for risk management.

Kullback-Leibner Divergence

The Kullback-Leibner divergence, also known as *relative entropy* is a commonly used measure of similarity between two distribution also from the f-divergence family.

$$D_{kld}(f||g) \overset{def}{=} \int f(x)\log\left(\frac{f(x)}{g(x)}\right) dx \tag{5.2}$$

This is exactly the quantity AIC aims to estimate. Since the function is weighted by the true density, thus it puts more weight on regions closer to the mode. However, since the integral can sometimes be infinite we have only integrate over the range $[-20, 20]$ to get the difference that made practical matter, and that so the mean can be computed. Although a newer measure known as Shannon-Jensen divergence based on the Jensen's inequality and the Shannon entropy which is symmetric and closely related to the Kullback-Leibner possessing several favourable features is available. We were unable to implement before the deadline but it may have been a better measure to avoid numerical problems we have been suffering from the Kullback-Leibner Divergence.

Mean Integrated Square Error

Although the mean integrated square error is more commonly used in density estimation, we have also adopted it for our assessment as it has a different cost function

$$D_{mise}(f||g) \overset{def}{=} \mathbb{E} \int (f(x) - g(x))^2 \, dx \tag{5.3}$$

All the measures above are computed using numerical integration, if any of the integrals above fail to be computed for a single iteration, then it will be repeated until all measures are calculated successfully. As a result, the simulation should only be viewed as a pilot study since we do not know the reason for the failure of the numerical integration. This could potentially ignore certain special cases and generate bias in the result.

5.3 Results

Each subsection below presents the fit of other distribution when the underlying GARCH process are generated from a particular distribution. The original distribution is not fitted.

Assuming Standard Normal distribution

Figure 5.1 show the results of fitting other distributions to a normal-GARCH process. From the plot, we can observe that the scale normal mixture outperforms the generalised error distribution and the normal inverse Gaussian by a small margin in large sample however the normal inverse Gaussian performs much better in very small sample ($\leq$ 100). On the other hand, the standard t distribution performs poorly overall. This is an unsurprising result as the normal distribution is a special case of the generalised error distribution, normal inverse Gaussian and the scale normal mixture, but only a limiting case of the standard t. The reason for the small competitive advantage of the scale normal mixture may have been the fact that the standard normal distribution is used as the initial value. From the figure,

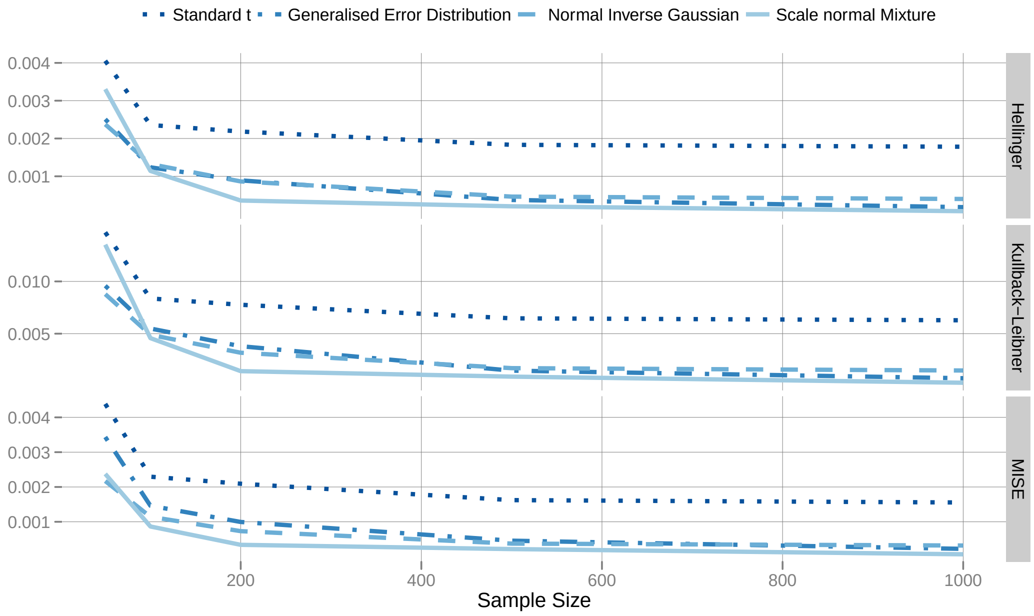

Figure 5.1: Recover statistic of different distribution assuming a normal distribution

we can also see a sample size of 400 is sufficient for most distribution to achieve its minimum dissimilarity, increasing the sample size from then does not reduce the distance much further.

Assuming Standard t distribution ($\nu = 2 \sim 100$)

Depicted in Figure 5.2 is the simulation result of assuming a standard t distribution with shape parameter $\nu \in (2, 100]$ which can be heavy tailed or close to a normal distribution. Again, the Scale normal mixture beats other competing model by a small margin (difficult to see from the figure but the value of the scale normal mixture is about $2 \sim 5$ times smaller than the other distribution). However the normal distribution is more robust for sample size ≤ 100.

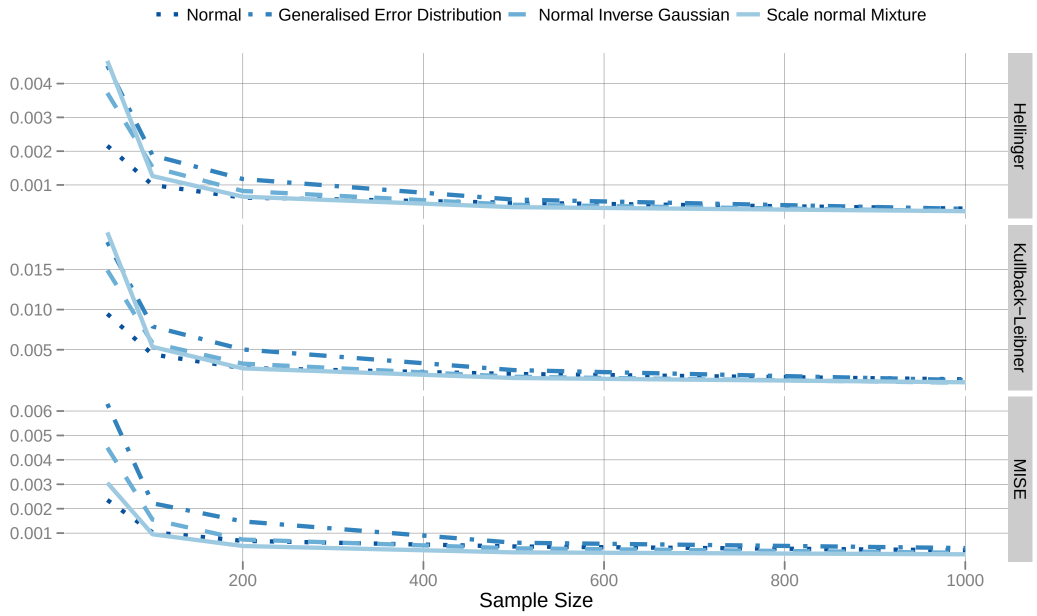

Figure 5.2: Recover statistic of different distribution assuming a Standard t distribution

Assuming Generalised Error distribution ($\nu = 1.5 \sim 2.5$)

The result when the underlying assumption is the generalised error distribution is given in Figure 5.3. Similar results were obtained where the scale normal mixture has a small edge over the other models over most sample size (≥ 100) and the standard t having a poor recovery.

Assuming Normal Inverse Gaussian distribution ($\zeta = 0.5 \sim 100$)

Lastly, we present the final case where we assume a normal inverse gaussian distribution in Figure 5.4. The story repeates with the normal distribution again showing its strength in robustness under small sample while being over taken by the scale normal mixture in larger samples.

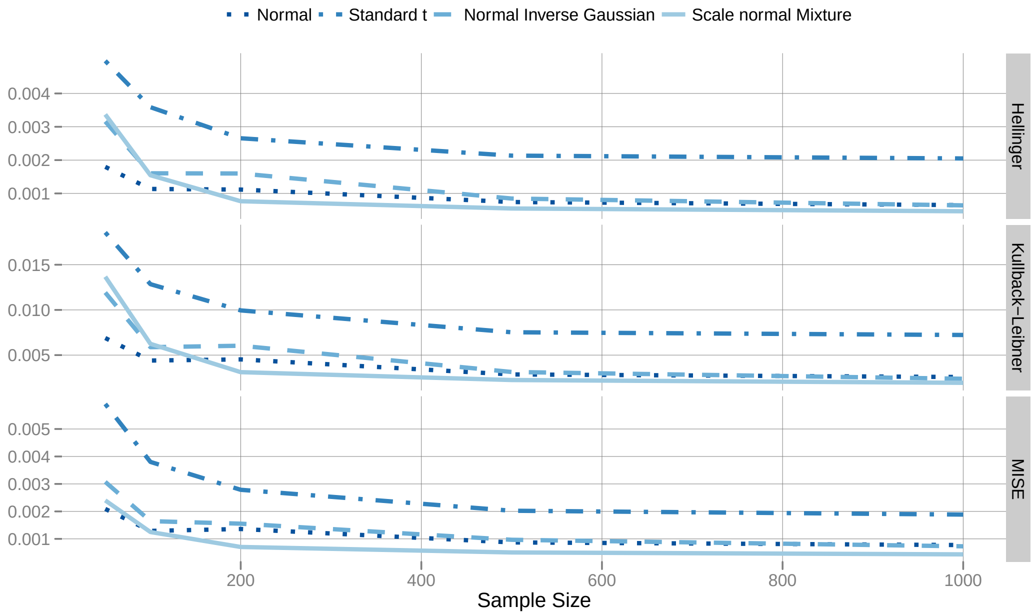

Figure 5.3: Recover statistic given Generalised error distribution

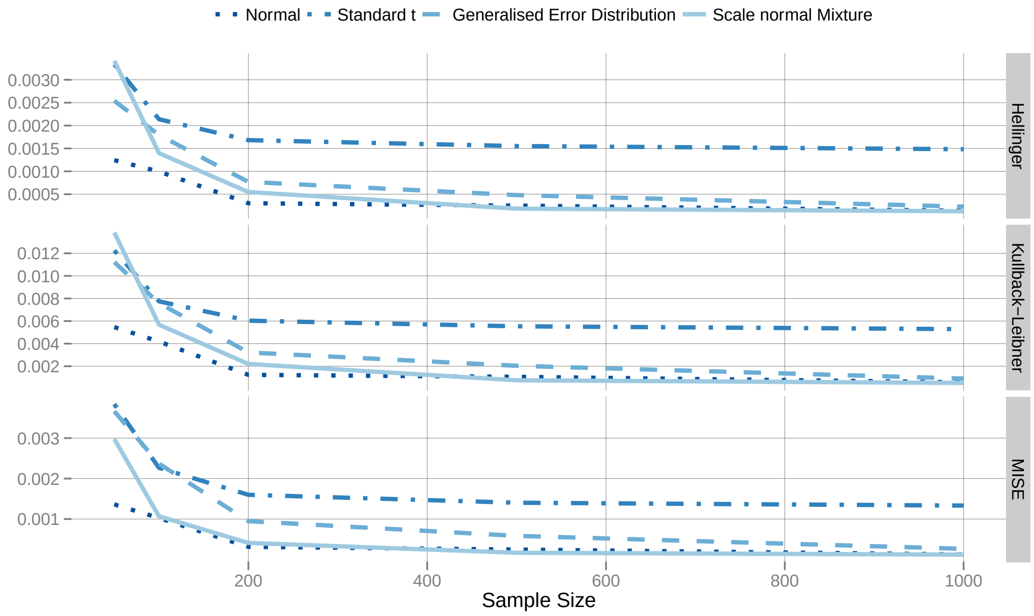

Figure 5.4: Recover statistic under Normal Inverse Gaussian assumption

5.4 Summary

From the simulation study, we can see that the scale normal mixture has a small edge over all the distribution considered in terms of recovering and approximate the true distribution as close as possible. The case of over-fitting seems to only occur for sample size less than one hundred but converges rapidly when the sample size increase to greater than two hundred; this is also generally true for other distribution as well. On the other hand, the normal distribution dominates and showing its robust characteristic when the sample size is small (≤ 100).

Nevertheless, none of the evaluation criteria really focus on the fit of the tail and thus it would be mis-leading if we ignore this fact and use the normal distribution as default for small sample. What would be more favourable is to tailor a measure which incorporates the risk of the observation as the cost function. For example we can devise a measure which weights the difference by how extreme the observation is. This way, we would be able to align the assessment more closely to the risk actually faced by the investors.

Finally, although the t-distribution in general being a poor choice shown in the simulation study, we should not use this as evidence to reject the use of the distribution. The reasons are (1) We have used selection criteria which are sound in theory, but may not be in the interest of practical work; (2) Only a handful of distribution have been selected, and a thorough study is required for such strong statement1; (3) The study only assess the flexibility and the generality of distributions rather than testing whether a distribution is suitable for GARCH modelling.

Chapter 6

Empirical Analysis

The performance of volatility models are in general hard to assess due to the nature that the volatility itself is not directly observable and thus render it impossible to compare the predicted value against the future observation then compute a statistic in typical forecast setting. This along with the fact of lacking a strong test are the main reason why it is difficult to make any conclusive statement which distribution is the best.

In order to assess the model under this setting, we will first compare how the model fit to the data. The commonly used log-likelihood, AIC and BIC are reported. We then evaluate the one-step ahead prediction by computing a predicted likelihood which can bee understand as the likelihood of the future observation under current model. The greater this value is, the more robust the model and fits the future observation well. The interval forecast methodology of Christoffersen (1998) is adopted to examine the forecast of the tail probability. More specifically, whether the model is capable of assigning the correct coverage probability to the tail and the degree of efficiency of incorporating past information.

These statistics and tests are computed for all the models and bench marked with other competing distributions. We will compare the performance of the scale normal mixture with three distributions and its skewed counter parts a total of seven distributions made available by the package fGarch Wuertz et al. (2012).

- Standard normal distribution
- Standard student t
- Standard generalized error distribution
- Skewed standard normal distribution
- Skewed standard student t
- Skewed standard generalized error distribution
- Standardized normal inverse Gaussian distribution

In this chapter, three financial times series are chosen to demonstrate the use of the SNM-GARCH and its performance with other competing distributions. The data set chosen were the Standard and Poors 500, the exchange rate between Deutsche Mark and the British pound and the Australian Coal price. The choice of data try to give a comprehensive view of financial time series where they represent different market, different asset, different observation frequency and also geographical space.

6.1 Demean the time series

Before modelling the volatility, the time series must first be demean to ensure that there are no autocorrelation in the first moment. The *auto.arima* function from the forecast package with contributions from Slava Razbash and Schmidt (2012) is used to model the conditional mean of the financial time series via an ARIMA model with the best AIC.

6.2 Assessment

To assess the distributions in a more general setting, we measure both the fit of the model, and also the predictability of the model base on the interval forecast framework.

In order to avoid selection bias of the data, we have use a window sliding approach also adopted by Kuester et al. (2006). At every step, all competing models are built on a sample of the data, with the statistics and prediction computed. We then slide the window one observation to the next set of sample for which we will build another set of models, and this is repeated to the end of the time series. The fit and test statistics are then summed or averaged depending on the nature of the statistic to give the final overall performance statistic. A window size of one thousand is used for all the case studies.

6.2.1 In sample fit

The typical model fitting statistics such as the log-likelihood, AIC, and BIC are computed to assess how well the model fit to the sub-sample of the data. The average of the statistics are reported to give a broad view of the fit across different time frames.

6.2.2 Out of sample forecast

At every step we also make a one-step ahead forecast of the volatility and use it to compute likelihood of the next observation.

We have defined the n-step ahead predicted likelihood as:

$$\begin{aligned} Pll^{t}_{t+n} &= f(x_{t+n}|\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}) \\ &= f(x_{t+n}|\sigma^2_{t+n}) \end{aligned} \tag{6.1}$$

Where σ^2_{t+n} is the n-step forecast of the conditional variance given the current model and information set at time t. The predicted likelihood is then summed up across all predictions to give a measure which is the likelihood of observation under the historical model.

We have also implemented the value at risk framework herein refer to as VaR for brevity, which is widely used in application and also formed the basis for the basil standard set out by the basil committee. It is designed to assess the potential monetary loss faced at a given time and target probability λ by an asset holder. The one step value at risk faced by an investor can be calculated by the following expression:

$$VaR_{t+1}^{\lambda} = -\inf_{x}\{x \in \mathbb{R} : P(x_{t+1} \leq x|\mathcal{F}_t) \geq \lambda\}, \quad 0 < \lambda < 1 \tag{6.2}$$

Although the VaR is a scalar value, it is in fact an interval forecast since it represent the minimum loss with the given probability λ which has a maximum loss of infinity.

In the paper of Christoffersen (1998), he introduced a general framework to determine whether the sequence of interval forecast produced were satisfactory. This is then implemented by Kuester et al. (2006) specifically for VaR.

We first define the violation as:

$$H_t = I(x_t < -VaR_t^{\lambda}), \tag{6.3}$$

Where $I(.)$ is the indicator function, and H_t is a binary sequence representing whether the observed value exceeded the VaR. If the sequence of VaR forecasts is efficient in incorporate all past information, then the following expression should be satisfied.

$$E[H_t|\mathcal{F}_{t-1}] = \lambda, \tag{6.4a}$$

$$H_t|\mathcal{F}_{t-1} \overset{iid}{\sim} \text{Bernoulli}(\lambda), \quad t = 1, 2, \ldots, T \tag{6.4b}$$

Which states that the expected proportion of violation should be the same as the target coverage probability and that the violation sequence H_t are independent Bernoulli random variables.

Three tests were set out in Christoffersen's frame work to evaluating the interval forecast. The unconditional coverage probability in Equation 6.4a is first tested, then the independence test of Equation 6.4b is evaluate and combined with the first test to form the final test.

Test of unconditional coverage

If a given sequence of VaR forecast is correct, we can test whether the proportion of violation is different to the target probability by the following likelihood ratio test:

$$LR_{uc} = 2[\ell(\hat{\lambda}; H_1, H_2, \ldots, H_T) - \ell(\lambda; H_1, H_2, \ldots, H_T)] \overset{asy}{\sim} \chi^2_1, \tag{6.5}$$

Where $\ell(.)$ is the log binomial likelihood while $\hat{\lambda}$ and λ are the proportion of violation and the target probability respectively.

Test of independence

The test of unconditional coverage is usually a simple to pass by most GARCH models and that is why stronger tests are desired. In Christoffersen (1998)'s paper, he test the independence between the violation sequence as a binary first order Markov chain with transition probability matrix

$$\Pi = \begin{pmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{pmatrix}, \quad \pi_{ij} = P(H_t = j | H_{t-1} = i), \tag{6.6}$$

Then the approximate binomial joint likelihood of the violation conditional on the first observation is:

$$\ell(\hat{\Pi}; H_2, H_3, \ldots, H_T | H_1) = \pi_{00}^{n_{00}} \pi_{01}^{n_{01}} \pi_{10}^{n_{10}} \pi_{11}^{n_{11}}, \tag{6.7}$$

Where n_{ij} is the count of each state transition to another, and π_{ij} is the maximum likelihood estimate of the transitional probability given as

$$\hat{\pi}_{01} = \frac{n_{01}}{n_{00} + n_{01}} \quad \text{and} \quad \hat{\pi}_{11} = \frac{n_{11}}{n_{10} + n_{11}} \tag{6.8}$$

Under the assumption that the sequence is independent, then the transition probability should be:

$$\pi_{01} = \pi_{11}$$
$$\pi_{00} = \pi_{10}$$

and have the conditional binomial joint likelihood as

$$\ell(\Pi; H_2, H_3, \ldots, H_T|H_1) = \pi_{00}^{n_{00}+n_{10}} \pi_{01}^{n_{01}+n_{11}}, \tag{6.9}$$

As a result, we can test the independence again with another likelihood ratio test given by

$$LR_{ind} = 2[\ell(\hat{\Pi}; H_2, \ldots, H_T|H_1) - \ell(\Pi; H_2, \ldots, H_T|H_1)] \overset{asy}{\sim} \chi_1^2 \tag{6.10}$$

Test of conditional coverage

The two test above can be combined to form the test for overall conditional coverage as suggested by Christoffersen as:

$$LR_{cc} = LR_{uc} + LR_{ind} \overset{asy}{\sim} \chi_2^2 \tag{6.11}$$

The tests and statistics described in this section will be used to assess the model fit given the sample size is available. We will use three different type of financial asset to illustrate the competence of the SNM-GARCH model.

6.3 Case studies

6.3.1 SNP 500

Standard and Poors 500 also known as the SNP500 is one of the most widely used and scrutinised financial time series for volatility modelling. We have taken the sample from 1st of January 2005 to 5th of July 2012. The data can be downloaded in R using the *get.hist.quote* function from the tseries package Trapletti and Hornik (2012).

Shown below in Table 6.1 are the unconditional moments of the time series. We can see that the there are some minor skewness present in the data, furthermore the excess kurtosis suggests that the normal distribution is far from suitable for this exercise.

Mean	Variance	Skewness	Kurtosis
8.562405e-05	0.01408255	-0.3086132	10.45502

Table 6.1: Moment statistics of the SNP500 data starting on the 1st of January 2005

From the plot of the data in Figure 6.2 we can see four possible of clusters of volatility. Where the volatility is significantly different to the initial two years of data. The volatility is even more extreme towards late 2008 and early 2009 where the highest value is 0.1255 or about nine times of the unconditional standard deviation. Clearly from the data we can see a model which assumes constant variance is an unrealistic one.

Examining the fit statistics, we can see that all the numbers unanimously favours the skewed generalised error distribution which has the best log-likelihood, AIC, and BIC. Only the normal inverse Gaussian distribution beats the generalized error distribution in the predicted likelihood category.

Furthermore, we can see that almost all the statistics for the skewed distribution are better than its symmetric counter-part. A strong indication where the skewness is required and the incorporation of the skewness significantly improves the fit.

Given the flexibility of the scale normal mixture, we were surprised by the result of the fit statistic where it only exceeded the distribution which

Normal	Standard-t	Ged	S.Normal	S.Standard-t	S.Ged	S.Nig	S.Mixture Normal
2983.07	3004.29	3004.91	2997.97	3015.98	**3020.93**	3017.86	3013.70
-5960.13	-6000.58	-6001.82	-5987.94	-6021.96	**-6031.85**	-6025.73	-6015.41
-5943.50	-5978.40	-5979.64	-5965.76	-5994.23	**-6004.13**	-5998.01	-5982.14
261.38	278.27	240.06	264.13	265.49	268.64	**285.66**	271.82

Table 6.2: Fit statistic of SNP500 data, the statistics are log-likelihood, AIC, BIC and predicted likelihood in each with respective row.

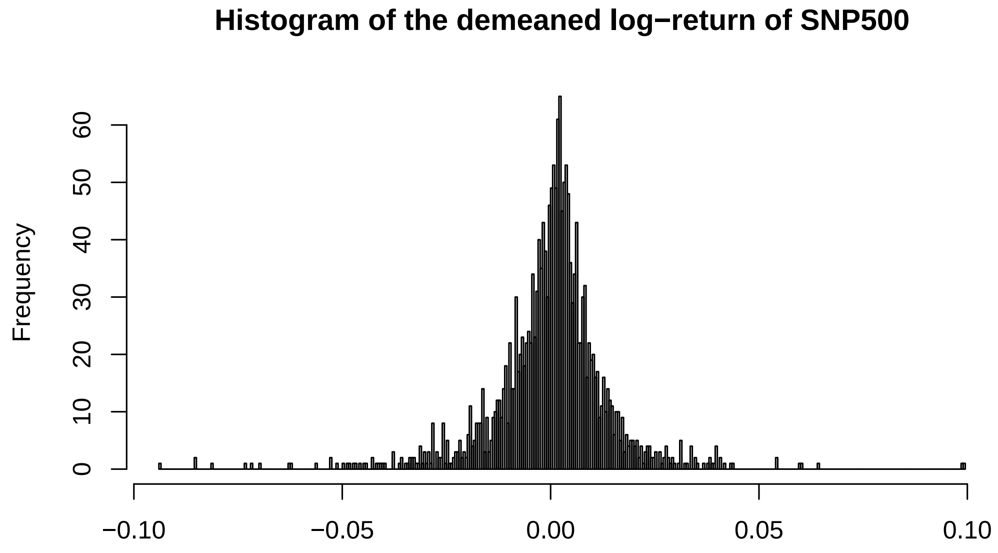

Figure 6.1: The histogram of the SNP 500 stock index between January 3rd 2005 to July 5th of 2012

does not account for skewness. The predicted likelihood however suggests that the prediction of the scale normal mixture was robust and reliable only surpassed by the normal inverse Gaussian and the symmetric standard t.

If we examine the unconditional distribution in Figure 6.1, we can see that the distribution does have a sharp shape very like the Laplace or the generalized error distribution with the shape parameter $\nu \approx 1$. Although this shape is obtainable by the scale normal mixture, but it requires a large number of grid point very close to zero and a extremely fine grid is also required over the feasible region. This lead to the result where our model fail to capture the peak close to zero, nevertheless this can be improved by increase the grid points but the computation burden would also increase exponentially. Furthermore, the non parametric nature of the method may not be able to capture this peak without a large and sufficient number of observations close to zero.

Now we turn our attention to the forecasted value-at-risk and the likeli-

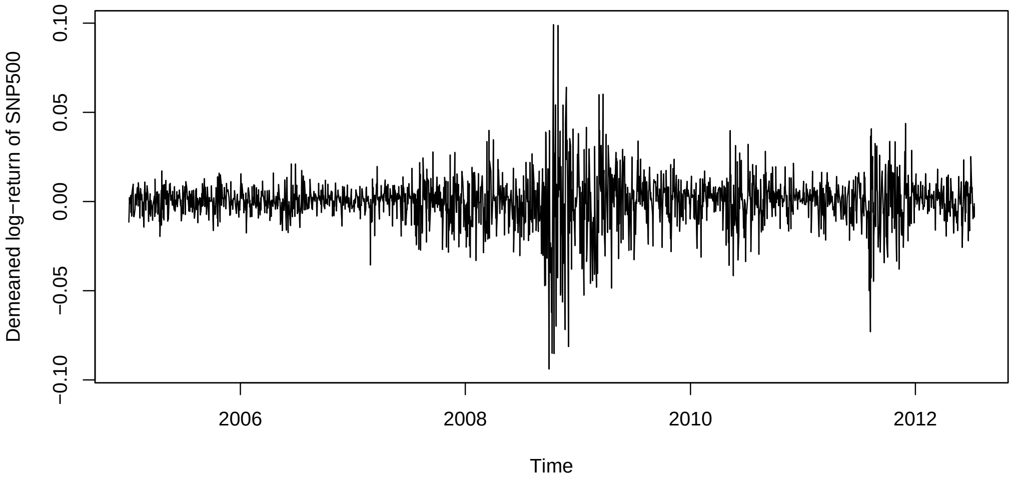

Figure 6.2: Figure depicting the return of the SNP 500 stock index between January 3rd 2005 to July 5th of 2012

hood ratio test in Table 6.3. All the proposing distribution appears to be capable of capturing the tail sufficiently even at the 1% level except the normal distribution. The normal inverse Gaussian model however fails the conditional coverage test when the independence and the unconditional coverage are tested altogether. Nevertheless, the test does not provide any insight as to which model whom pass the test is better. It is although possible that we can extend the tail probability as far as we like where only one model survives all the test, but the result may not be valid without a very large sample size and may subject to sampling variation.

Overall, the generalized error distribution appears to be the natural choice for modelling this particular data set from both the exploratory analysis and the test statistics. Yet, we believe that the scale normal mixture can obtain similar if not better result when we increase the granularity of the search space and a larger sample size.

	$\lambda = 5\%$			$\lambda = 1\%$		
	LR$_{uc}$	LR$_{ind}$	LR$_{cc}$	LR$_{uc}$	LR$_{ind}$	LR$_{cc}$
Normal	0.406	0.319	0.431	**0.001**	**0.001**	**0**
Student t	0.328	0.247	0.317	0.716	0.547	0.781
Generalized Error Distribution	0.406	0.319	0.431	0.495	0.388	0.546
S. Normal	0.818	0.721	0.914	0.061	**0.045**	**0.023**
S. Student t	0.818	0.721	0.914	0.506	0.459	0.609
S. Generalized Error Distribution	0.817	0.817	0.948	0.152	0.147	0.126
S. Normal Inverse Gaussian	0.817	0.817	0.948	0.064	0.063	**0.032**
S. Scale Normal Mixture	0.939	0.884	0.986	0.321	0.251	0.316

Table 6.3: Likelihood Ratio Test of VaR for SNP500 data, with the tests statistics being rejected at the 5% level in bold

6.3.2 Deutsche Mark/British Pound exchange rate DEM/GBP

This data set is a well accepted benchmark data for GARCH modelling suggested by Fiorentini et al. (1996) and is used by Wurtz et al. (raft) and many others. The time series contains 1975 daily observation start from January 2, 1984 to December 31, 1991 provided by the fGarch package.

Table 6.4 shows the presence of skewness and the heavy tail again. Examining the histogram in Figure 6.4, we can see the skewness is more profound than the SNP500 data, and the data seems to have a peak at zero as well but it is not as sharp as in the previous case. Comparing the time series plot for the exchange rate and the SNP500 data, we observe the typical volatility clustering of the financial time series but the SNP500 seems to deviate much more significantly during volatile periods such as the one in between 2008 and 2009.

Mean	Variance	Skewness	Kurtosis
-0.016427	0.221130	-0.249325	3.620941

Table 6.4: Moment statistics of the DEM/GBP data

Given that we have 1975 observations, this leaves us 975 observation to build our model and compute prediction and fit statistics for a window size of one thousand.

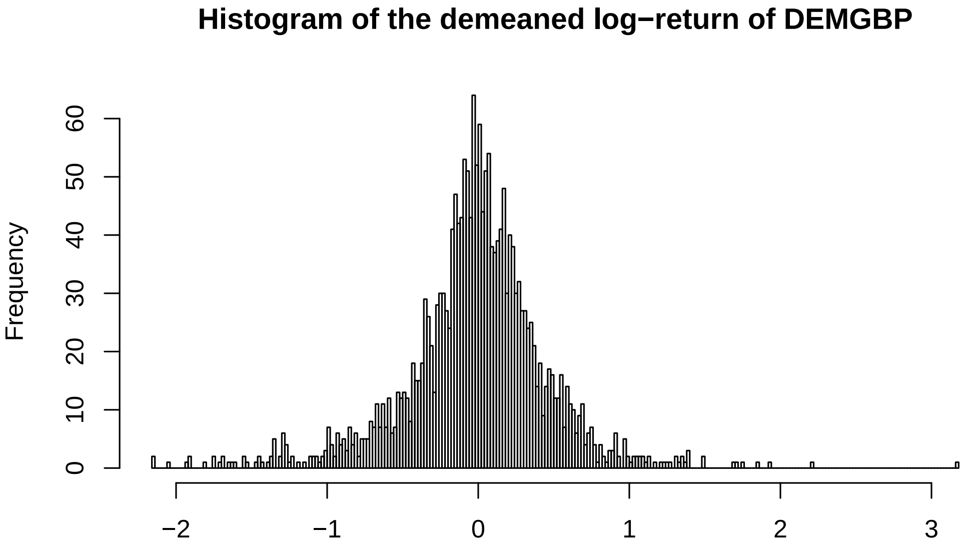

Figure 6.3: The Deutsch Mark and British Pounds exchange rate data between January 2, 1984, to December 31, 1991

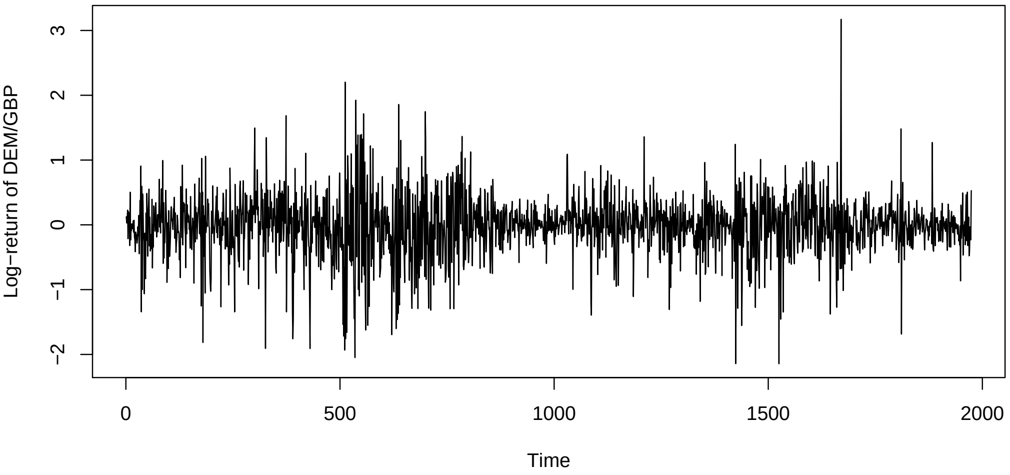

Figure 6.4: Figure display the return of the exchange rate between Deutsch Mark and British Pounds over the period of January 2, 1984, to December 31, 1991

Normal	Student-t	Ged	S.Normal	S.Student-t	S.Ged	S.Normal.Mix
-559.63	-501.74	-508.91	-555.47	-498.86	-507.10	**-496.77**
1125.26	1011.47	1025.83	1118.93	1007.73	1024.21	**1005.55**
1142.03	**1033.82**	1048.18	1141.28	1035.67	1052.15	1039.08
297.98	**333.06**	273.65	298.16	274.24	268.31	316.30

Table 6.5: Fit statistic of DEM/GBP data, the statistics are log-likelihood, AIC, BIC and predicted likelihood in each with respective row.

Table 6.5 reports the average of the fit statistic over the 975 models built, the result for the normal inverse Gaussian was not reported due to the fact that the model failed to fit half of the time. The table shows that the scale normal mixture fits the data much better in contrast to the SNP500 data and was the best performer in the log-likelihood and AIC and the runner up for the predicted likelihood. The standard t distribution also showing signs of a good fit with the best BIC and predicted likelihood. This suggests that the grid point we have devised works fine in this case, however not particularly suited for the SNP500. Again, the fit statistic is in general better for skewed distribution but the BIC seems to prefer model which does not account for skewness. Interestingly, the dominant generalized error distribution performed extremely badly only defeated the normal distribution.

Looking at the likelihood tests for the prediction of the VaR. In general, none of the model in Table 6.6 are rejected at the 5% nor the 1% level except the normal distribution. This is not a shocking result and is the reason for the large amount of literature proposed for GARCH. The lack of a standard and strong test is the main problem in determining a best distribution and model for volatility modelling.

Although we were unable to establish the best model overall, nevertheless the two case study demonstrated the difference in the behaviour of the financial time series can result in different model selection. Ironically, the generalized error distribution which easily outperformed other competing model in the SNP500 data happens to be the worst model when fitted to the DEM/GBP exchange rate data.

	$\lambda = 5\%$			$\lambda = 1\%$		
	LR$_{uc}$	LR$_{ind}$	LR$_{cc}$	LR$_{uc}$	LR$_{ind}$	LR$_{cc}$
Normal	0.314	0.073	0.121	0.034	0.018	**0.007**
Student t	0.737	0.382	0.645	0.198	0.066	0.081
Generalized Error Distribution	0.314	0.190	0.254	0.198	0.066	0.081
S. Normal	0.187	0.099	0.108	0.065	0.030	**0.018**
S. Student t	0.393	0.251	0.359	0.198	0.066	0.081
S. Generalized Error Distribution	0.245	0.139	0.171	0.318	0.085	0.138
S. Scale Normal Mixture	0.965	0.341	0.635	0.198	0.066	0.081

Table 6.6: Likelihood Ratio Test of 5% VaR for DEM/GBP data, with the best performing model statistic in bold

6.3.3 Coal Price

This data comprises of 363 monthly observations of the coal price. Given the size of the data, we are unable to replicate the same methodology to assess our model as we did in the two previous case study. However this exercise serves as an important example due to the nature that the number of zero is excessive. In addition, it is one of the most important and influential commodity used in household and particularly commercially for power generation.

Zero returns are not atypical in financial time series, especially for those that are not actively traded. In the work of Alexander and Lazar (2006), they have assumed they are missing data and removed them from the data set, we think this is extremely inappropriate to discard data without examine whether they are missing value or true data. We will use the coal price to demonstrate how the excessive zeroes can be dealt in a natural way using the scale normal mixture without torturing our data.

Looking at the table of unconditional moments, we can see this data actually behave quite like the SNP500, other than the fact that it has a much lower mean and a slightly higher variance. One of the possible explanation for the lower mean is due to the higher efficiency of the SNP500 market having more traders at any single time.

From Figure 6.6, the activity seems to be low prior to 2005 with only spikes once now and then which contributed to the excessive amount of zeros

Mean	Variance	Skewness	Kurtosis
0.002832903	0.05417312	-0.195945	10.3917

Table 6.7: Moment of the Coal data since January 1980 to December 2011

in the data. The reason for this is unclear, the market may have been traded under restricted market condition. However, since 2005 the market became more volatile than in the previous period and looks like the typical figures we have seen for other markets. The phenomenon is clear when we examine the histogram where the frequency of other observations are dwarfed by the abundance of zero.

In Figure 3.1, we can see that the fitted density of the coal has a spike at zero. The reason for this is that one of the mixture has a support point arbitrary close to zero to account for the excessive number of zero in the data. One possible scenario for this is that if the asset is not trade frequently, then it is possible that the price will not change over consecutive periods. This type of behaviour is clearly not possible to capture with other competing models we have examine, and the likelihood value between the models are extremely large.

Without being able to assign the correct probability to the abundant zeroes, a model will allocate excessive mass to other parts of the density and thus over estimate the probability of values of return.

The log-likelihood is shown in Table 6.8, where the log-likelihood of the scale normal mixture is almost double of the second best model. Although we were unable to utilize the interval forecast evaluation, but we suspect that all other models are likely to be rejected as the tail probability assigned is over estimated and would be greater than what is observed in the empirical data.

Normal	Student-t	Ged	S.Normal	S.Student-t	S.Ged	S.Nig	S.Normal.Mix
640.6025	948.0280	785.5237	751.8446	937.0532	785.5237	751.8446	1834.825

Table 6.8: log-likelihood of GARCH models fitted to the Australian coal price

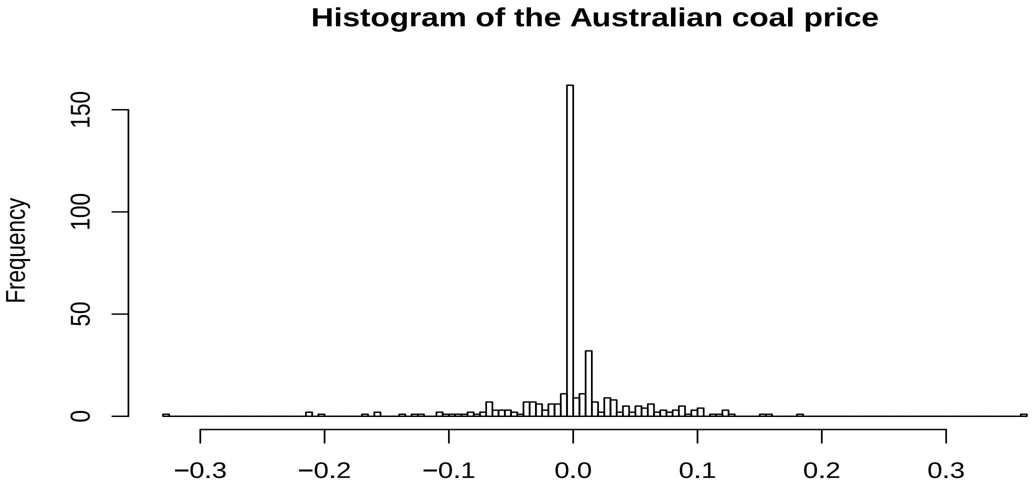

Figure 6.5: Histogram of the Australian coal price.

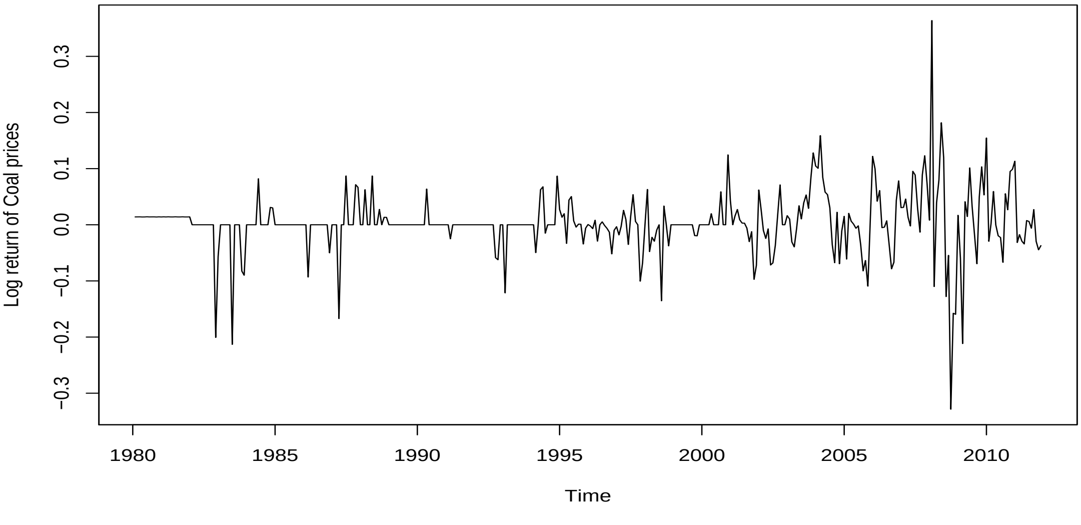

Figure 6.6: Figure display the return of the Australian coal price over the past 30 years.

6.4 Summary

This section examines three data set from different market which illustrates how the characteristics and behaviour of financial time series can differ and lead to different model selection.

The SNP500 data was best fitted by the generalized error distribution, which ironically turns out to be the worst model for the exchange rate series. Even though we believe the scale normal mixture can achieve a better result when given more data and greater computing power for finer grid. Whereas the scale normal mixture and the student t were the preferred model for the DEM/GBP exchange rate data. The disagreement is natural as the source and the driving force of volatility may be different between markets and assets. The Australian coal price is not the only price series which exhibits excessive zeroes and no literature existed to account for this type of behaviour. The scale normal mixture is the only distribution where we can conclude and make a firm statement to be the only distribution considered applicable to these data.

These result suggest that the scale normal mixture is promising as supported by theory.

Chapter 7

Discussion, Conclusion and Future Works

In this paper, the scale normal mixture distribution is introduced for modelling volatility and the applicability of the proposition was examined in detail.

7.1 Discussion

Why the Scale Normal Mixture?

We have seen the strong theoretical properties in the introductory and background section where it is shown to embrace almost all the distribution in the literature as its special case. Rendering it a strong candidate to replace all current distribution and become the standard distribution for GARCH or the more general volatility modelling setting.

The simulation study demonstrated the flexibility of the distribution where it is capable and best at recovering the true underlying distribution regardless the type or family of distribution.

We further illustrated the ability in which the mixture representation can reveal group heterogeneity or uncover the source of volatility which lacking by the GARCH model where only the macro mechanism is addressed but not the micro market structure. Finally, a strong case study was presented where the scale normal distribution is clearly the only distribution to capture data with excessive zeroes. All these elements put forward a strong argument for use the scale normal mixture distribution.

Computation complexity

Despite the result were not as spectacular when applied to empirical data in particular the case of the SNP500 data, it is can be improved. Two reason underlie this problem (1) The grid point supplied may have not been fine enough to encapsulate the values sufficiently close to zero; (2) The sample size may need to be increased so there is sufficient observations to give more support points around zero to give the sharp continuous decay shape of Laplace type data. We have not been able to use a finer grid and observe the improvement due to the fact that the grid size is already at the maximum capacity supported by the RAM of my computer. we believe with additional research, fine tuning of the algorithm and greater computation power the desirable result supported by the theory can be achieved.

The scale normal mixture was made feasible for application and analysis due to the CNM algorithm. Nevertheless, the current implementation for the SNM-GARCH in R can still take quite some time usually between 30 seconds to a minute to fit a sample of 1000 observations. Which is not comparable to the fGarch package with most of the implementation code are based in FORTRAN.

7.2 Future work

Improve the grid of the feasible region

As mentioned in the discussion, it would be desirable to implement a set of grid point which is capable of capturing sharp modes. This is also related to the zero inflation model, rather than having a support point sufficiently close to zero to account for abundance of zero exact zero support point would be more favoured.

Skewness

Although the parametization of Fernandez and Steel is very flexible and easy to implement, a skewness based on the mixing distribution which also incorporate the non-symmetric normal inverse Gaussian and the generalized hyperbolic family would be highly desired.

Stronger tests and aligned measure

The evaluation criteria implemented in the simulation study were of theoretical interest but (1) Most of them suffers from numerical problem and it is possible for the measure to be infinite (2) Does not focus on the tail probability which is more important for practical purposes. For numerical stability and to avoid infinity we promote the use of the Jensen-Shannon Divergence which is more numerically stable and always have a finite value. Moreover, a measure with a quadratic cost function putting more weight on the tails and extreme observations may be better aligned with the need of risk management.

In order to establish any formal result, stronger tests are required. This problematic scenario was illustrated in the empirical study where all the distribution passes all the test even at the 1% level and no conclusive statement can made.

Extend to higher order and extension of GARCH model

The focus of this paper was to examine the distribution solely. There are hundreds of GARCH extension accounting for different phenomenons in empirical data, from the work of this paper we believe other extension of the GARCH model will also benefit from the flexibility of the scale normal mixture.

7.3 Conclusion

The work presented in this thesis proved that the scale normal mixture is a promising distribution for volatility. It has the peakness and heavy tail property observed by Mandelbrot, and the potential to unify all the proposed distribution in the GARCH literature. This will avoid researchers trying to find a best distribution when working with different markets and assets, a single distribution will suffice. With further research and improvements, we believe that fruitful result is promised.

Appendix A

Source Codes

In this appendix, we give the relevant codes for implementing the SNM-GARCH outlined in Chapter 4 for reference.

This is the function which is core to implement the CNM algorithm which computes the derivatives outlined in section 4.6. The function can be compiled for speed up.

```
logd.mgarch <- function(xt, beta, pt, which){
  initMethod = attr(xt, "initialisation")
  if(initMethod == "Smooth"){
    sigma1 = beta[5]
    k = attr(xt, "smoothWindow")
  } else {
    sigma1 = attr(xt, "sigma1")
    k = 1
  }
  T <- length(xt)
  lpt <- length(pt)
  lb <- length(beta)
  dl <- vector("list", length = 4)
  names(dl) <- c("ld", "db", "dt")
```

```
## Calculate the conditional variance
betaSum <- c(filter(xt[c(k:(T - 1))]^2, beta[2], "recursive"))
  sigma.t <-
      sqrt(c(rep(sigma1^2, k),
             beta[1] * (1 - cumprod(rep.int(beta[2], T - k)))/
             (1 - beta[2]) +
             cumprod(rep.int(beta[2], T - k)) * sigma1^2 +
             beta[3] * betaSum))

## Calculate the density
if(which[1] == 1){
  dl$ld = log(2) - log(beta[4] + 1/beta[4]) -
    0.5 * log(2 * pi) - log(outer(sigma.t, pt)) -
    ((xt * beta[4])^2 * Heaviside(-xt) +
     (xt / beta[4])^2 * Heaviside(xt))/
      (2 * outer(sigma.t^2, pt^2))
}

## Calculate the derivatives
if(which[2] == 1){

  ## Piece wise Analytical derivative
  dldsigma = -1/sigma.t +
    ((xt * beta[4])^2 * Heaviside(-xt) +
     (xt / beta[4])^2 * Heaviside(xt))/outer(sigma.t^3, pt^2)

  convFilter <- 1:(T - k - 1) * beta[2]^(0:(T - k - 2))
  betaSum2 <- c(0, beta[3] *
                convolve(convFilter, rev(xt[c(k:(T - 1))]^2),
                         type = "open")[1:(T - k - 1)])

  ## beta[1] - Mean of the conditional variance equation
```

```
if(initMethod == "Smooth"){
  b11 = rep(0, k)
} else if(initMethod == "BackFilter"){
  b11 = 0
} else if(initMethod == "Unconditional"){
  b11 = (1 - beta[2] - beta[3])^-1
}
dsigmadalpha0 <-
  c(b11, cumsum(beta[2]^(0:(T - k - 1))))


## beta[2] - Coefficient for lagged variance
if(initMethod == "Smooth"){
  b21 = rep(0, k)
} else if(initMethod == "BackFilter"){
  b21 = 0
} else if(initMethod == "Unconditional"){
  b21 = (1 - beta[2] - beta[3])^-2
}
dsigmadalpha1 <-
  c(b21, ((beta[1] *
           cumsum(c(0:(T - k - 1)*beta[2]^(-1:(T - k - 2))))) +
    1:(T - k) * beta[2]^(0:(T - k - 1)) * sigma1^2 + betaSum2))


## beta[3] - Coefficient for lagged sqaured observation
if(initMethod == "Smooth"){
  b31 = rep(0, k)
} else if(initMethod == "BackFilter"){
  b31 = 0
} else if(initMethod == "Unconditional"){
  b31 = (1 - beta[2] - beta[3])^-2
}
```

```
dsigmadbeta1 <- c(b31, betaSum)

## beta[4] - Skewness parameter
dldgamma = -((1 - 1/beta[4]^2)/(beta[4] + 1/beta[4])) +
  (xt^2 / beta[4]^3 * Heaviside(xt) -
   xt^2 * beta[4] * Heaviside(-xt))/outer(sigma.t^2, pt^2)

## beta[5] - Initial variance
if(initMethod == "Smooth"){
  dsigmadsigmac <- c(rep(1, k),
                     2 * sigma1 * beta[2]^(1:(T - k)))
}


## Combine everything
if(initMethod == "Smooth"){
dl$db <- array(c(dldsigma * c(dsigmadalpha0/
                     c(rep(1, k), 2 * sigma.t[-c(1:k)])),
                 dldsigma * c(dsigmadalpha1/
                     c(rep(1, k), 2 * sigma.t[-c(1:k)])),
                 dldsigma * c(dsigmadbeta1/
                     c(rep(1, k), 2 * sigma.t[-c(1:k)])),
                 dldgamma,
                 dldsigma * c(dsigmadsigmac/
                     c(rep(1, k), 2 * sigma.t[-c(1:k)]))),
               dim = c(T, lpt, lb))
} else {
dl$db <- array(c(dldsigma * c(dsigmadalpha0/
                     c(rep(1, k), 2 * sigma.t[-c(1:k)])),
                 dldsigma * c(dsigmadalpha1/
                     c(rep(1, k), 2 * sigma.t[-c(1:k)])),
                 dldsigma * c(dsigmadbeta1/
                     c(rep(1, k), 2 * sigma.t[-c(1:k)])),
                     dldgamma), dim = c(T, lpt, lb))
    }
  }
  if(which[3] == 1){
    dl$dt = -1/matrix(rep(pt, each = T), nc = lpt) +
      ((xt * beta[4])^2 * Heaviside(-xt) +
       (xt / beta[4])^2 * Heaviside(xt))/outer(sigma.t^2, pt^3)
  }
  dl
}
logd.mgarch <- cmpfun(logd.mgarch)
```

This function is used to create the class of mgarch for SNM-GARCH model. The method for intialising the time series and the smoothing window is set by this function.

```
## Function for converting different class of time series to mgarch
##
as.mgarch <- function(x, window = 5, init = c("Smooth", "BackFilter",
                      "Unconditional")){
    init = match.arg(init)
    mgts <- as.numeric(data.matrix(x))
    class(mgts) = "mgarch"
    attr(mgts, "initialisation") = init
    attr(mgts, "sigma1") = abs(x[1]) # For back estimate
    attr(mgts, "smoothWindow") = window # For smooth initial
    mgts
}
```

This function estimates and initialise the GARCH parameters β_0 and the mixing distribution G_0 as normal described in section 4.5

```
initial.mgarch <- function(x, beta = NULL, mix = NULL, kmax = NULL){
  initMethod = attr(x, "initialisation")
  if(is.null(beta)){
    gf <- try(garchFit(data = as.numeric(x), trace = FALSE,
                       include.mean = FALSE, cond.dist = "snorm"))
    if(initMethod == "Smooth"){
      if(!(inherits(gf, "try-error"))){
        cgf <- coef(gf)
        beta <- c(cgf["omega"], cgf["beta1"], cgf["alpha1"],
                  cgf["skew"], gf@sigma.t[1])
      } else {
        beta <- c(1e-5, 0.8, 0.1, 1, sd(x))
      }
    } else {
      if(!(inherits(gf, "try-error"))){
        cgf <- coef(gf)
        beta <- c(cgf["omega"], cgf["beta1"], cgf["alpha1"],
                  cgf["skew"])
      } else {
        beta <- c(1e-5, 0.8, 0.1, 1)
      }
    }
  } else {
    beta = beta
  }
  if(initMethod == "Smooth"){
    names(beta) <- c("omega", "beta1", "alpha1", "xi", "sigma1")
  } else {
    names(beta) <- c("omega", "beta1", "alpha1", "xi")
  }
  if(is.null(mix)){
    mix <- disc(1, 1)
  } else {
    mix = mix
  }
  list(beta = beta, mix = mix)
}
```

The bound of grid point k_1 and k_2 of section 4.5 is set by the following function.

```
## Function to determine the grid
gridpoints.mgarch <- function(x, beta, grid){
  2^seq(sqrt(min(abs(x[x != 0]))/3)^(1/2), .Machine$double.max.exp,
        length = grid)
}
```

There two functions were modified from the original garchSpec in the fGarch package to allow simulation of the normal inverse Gaussian and the scale normal mixture distribution.

```
garchSpec <- function(model = list(), presample = NULL,
                      cond.dist = c("norm", "ged", "std", "cauchy",
                                    "snorm", "sged", "smnorm", "sstd",
                                    "snig", "ssmnorm"),
                      rseed = NULL){
  cond.dist = match.arg(cond.dist)
  skew = list(norm = NULL, ged = NULL, std = NULL, smnorm = NULL,
    snorm = 0.9, sged = 0.9, sstd = 0.9, snig = 0, ssmnorm = 0.9)
  shape = list(norm = NULL, ged = 2, std = 4, smnorm = disc(1, 1),
    sged = 2, sstd = 4, snig = 4, ssmnorm = disc(1, 1))
  control = list(omega = 1e-06, alpha = 0.1, gamma = NULL,
    beta = 0.8, mu = NULL, ar = NULL, ma = NULL, delta = 2,
    skew = skew[[cond.dist]], shape = shape[[cond.dist]])
  control[names(model)] <- model
  model <- control
  if (sum(c(model$alpha, model$beta)) > 1)
    warnings("sum(alpha)+sum(beta)>1")
  order.ar = length(model$ar)
  order.ma = length(model$ma)
  order.alpha = length(model$alpha)
  if (sum(model$beta) == 0) {
    order.beta = 0
  }
  else {
    order.beta = length(model$beta)
  }
  if (order.ar == 0 && order.ma == 0) {
    formula.mean = ""
  }
  if (order.ar > 0 && order.ma == 0) {
    formula.mean = paste("ar(", as.character(order.ar), ")",
      sep = "")
  }
  if (order.ar == 0 && order.ma > 0) {
    formula.mean = paste("ma(", as.character(order.ma), ")",
      sep = "")
  }
  if (order.ar > 0 && order.ma > 0) {
    formula.mean = paste("arma(", as.character(order.ar),
      ", ", as.character(order.ma), ")", sep = "")
  }
  formula.var = "garch"
  if (order.beta == 0)
    formula.var = "arch"
  if (!is.null(model$gamma) != 0)
```

```
    formula.var = "aparch"
  if (model$delta != 2)
    formula.var = "aparch"
  if (order.beta == 0) {
    formula.var = paste(formula.var, "(", as.character(order.alpha),
      ")", sep = "")
  }
  else {
    formula.var = paste(formula.var, "(", as.character(order.alpha),
      ", ", as.character(order.beta), ")", sep = "")
  }
  if (formula.mean == "") {
    formula = as.formula(paste("~", formula.var))
  }
  else {
    formula = as.formula(paste("~", formula.mean, "+", formula.var))
  }
  if (is.null(model$mu))
    model$mu = 0
  if (is.null(model$ar))
    model$ar = 0
  if (is.null(model$ma))
    model$ma = 0
  if (is.null(model$gamma))
    model$gamma = rep(0, times = order.alpha)
  if (is.null(rseed)) {
    rseed = 0
  }
  else {
    set.seed(rseed)
  }
  order.max = max(order.ar, order.ma, order.alpha, order.beta)
  iterate = TRUE
```

```
if (!is.matrix(presample)) {
  if (is.null(presample)) {
    iterate = FALSE
    n.start = order.max
  }
  else {
    n.start = presample
  }
  z = rnorm(n = n.start)
  h = rep(model$omega/(1 - sum(model$alpha) - sum(model$beta)),
    times = n.start)
  y = rep(model$mu/(1 - sum(model$ar)), times = n.start)
}
else {
  z = presample[, 1]
  h = presample[, 2]
  y = presample[, 3]
}
presample = cbind(z, h, y)
if (iterate) {
  n.iterate = length(z) - order.max
  deltainv = 1/model$delta
  for (i in n.iterate:1) {
    h[i] = model$omega + sum(model$alpha * (abs(abs(y[i +
       (1:order.alpha)]) - model$gamma *
       y[i + (1:order.alpha)])^model$delta)) +
         sum(model$beta * h[i + (1:order.beta)])
    y[i] = model$mu + sum(model$ar * y[i + (1:order.ar)]) +
      sum(model$ma * (h[i + (1:order.ma)]^deltainv)) +
        h[i]^deltainv * z[i]
  }
}
new("fGARCHSPEC", call = match.call(), formula = formula,
      model = list(omega = model$omega, alpha = model$alpha,
        gamma = model$gamma, beta = model$beta, mu = model$mu,
        ar = model$ar, ma = model$ma, delta = model$delta,
        skew = model$skew, shape = model$shape),
      presample = as.matrix(presample),
      distribution = as.character(cond.dist), rseed = as.numeric(rseed))
}

garchSim <- function (spec = garchSpec(), n = 100, n.start = 100,
                      extended = FALSE){
  stopifnot(class(spec) == "fGARCHSPEC")
  model = spec@model
  if (spec@rseed != 0)
    set.seed(spec@rseed)
  n = n + n.start
  if (spec@distribution == "norm")
    z = rnorm(n)
  if (spec@distribution == "ged")
    z = rged(n, nu = model$shape)
  if (spec@distribution == "std")
    z = rstd(n, nu = model$shape)
  if (spec@distribution == "smnorm")
    z = rsmnorm(n, mix = model$shape)
  if (spec@distribution == "snorm")
    z = rsnorm(n, xi = model$skew)
  if (spec@distribution == "sged")
    z = rsged(n, nu = model$shape, xi = model$skew)
  if (spec@distribution == "sstd")
    z = rsstd(n, nu = model$shape, xi = model$skew)
  if (spec@distribution == "snig")
    z = rsnig(n, zeta = model$shape, rho = model$skew)
  if (spec@distribution == "ssmnorm")
    z = rssmnorm(n, mix = model$shape, xi = model$skew)
```

```
delta = model$delta
z = c(rev(spec@presample[, 1]), z)
h = c(rev(spec@presample[, 2]), rep(NA, times = n))
y = c(rev(spec@presample[, 3]), rep(NA, times = n))
m = length(spec@presample[, 1])
names(z) = names(h) = names(y) = NULL
mu = model$mu
ar = model$ar
ma = model$ma
omega = model$omega
alpha = model$alpha
gamma = model$gamma
beta = model$beta
deltainv = 1/delta
order.ar = length(ar)
order.ma = length(ma)
order.alpha = length(alpha)
order.beta = length(beta)
eps = h^deltainv * z
for (i in (m + 1):(n + m)) {
  h[i] = omega + sum(alpha * (abs(eps[i - (1:order.alpha)]) -
     gamma * (eps[i - (1:order.alpha)]))^delta) +
       sum(beta * h[i - (1:order.beta)])
  eps[i] = h[i]^deltainv * z[i]
  y[i] = mu + sum(ar * y[i - (1:order.ar)]) + sum(ma *
     eps[i - (1:order.ma)]) + eps[i]
}
data = cbind(z = z[(m + 1):(n + m)],
             sigma = h[(m + 1):(n + m)]^deltainv,
             y = y[(m + 1):(n + m)])
rownames(data) = as.character(1:n)
data = data[-(1:n.start), ]
from <- timeDate(format(Sys.time(), format = "%Y-%m-%d")) -
```

```
    NROW(data) * 24 * 3600
  charvec <- timeSequence(from = from, length.out = NROW(data))
  ans <- timeSeries(data = data[, c(3, 2, 1)], charvec = charvec)
  colnames(ans) <- c("garch", "sigma", "eps")
  ans <- if (extended)
    ans
  else ans[, "garch"]
  attr(ans, "control") <- list(garchSpec = spec)
  ans
}
```

We have also implement all the distribution function (d, p, q, r) and also include a m function for computing the momnets of the skewed scale normal mixture.

```
erf <- function(x) 2 * pnorm(x * sqrt(2)) - 1

ierf <- function(x){
    qnorm((1 + x)/2)/sqrt(2)
}

dsmnorm <- function(x, varmix = disc(1, 1)){
    n = length(x)
    n.mix = length(varmix$pt)
    d = 1/sqrt(2 * pi * matrix(rep(varmix$pt, each = n), nc = n.mix)) *
        exp(-0.5 * x^2/matrix(rep(varmix$pt, each = n), nc = n.mix))
    drop(d %*% varmix$pr)
}

dssmnorm <- function(x, varmix = disc(1, 1), xi = 1){
    z = x * (xi * Heaviside(-x) + 1/xi * Heaviside(x))
    (2/(xi + 1/xi)) * dsmnorm(z, varmix = varmix)
```

```
}

psmnorm <- function(q, varmix = disc(1, 1)){
  z = outer(q, 1/sqrt(2 * varmix$pt^2))
  d = 0.5 * (1 + erf(z))
  drop(d %*% varmix$pr)
}

pssmnorm <- function(q, varmix = disc(1, 1), xi = 1){
  g = 2/(xi + 1/xi)
  H = xi * Heaviside(-q) + 1/xi * Heaviside(q)
  P = g * 1/sqrt(2) * (erf(outer(q * H, 1/sqrt(2 * varmix$pt^2)))/
        sqrt(2 * H^2) + 1/sqrt(2 * xi^2))
  drop(P %*% varmix$pr)
}

qsmnorm <- function(p, varmix = disc(1, 1)){
  outer(ierf(2 * p - 1), sqrt(2 * varmix$pt^2)) %*% varmix$pr
}

qssmnorm <- function(p, varmix = disc(1, 1), xi = 1){
  prob = double(length(p))
  for(i in 1:length(p)){
    prob[i] = uniroot(function(x) pssmnorm(x, varmix = varmix, xi = xi)
               - p[i], interval = c(-20, 20))$root
  }
  prob
}

rsmnorm <- function(n, mix = disc(1, 1)){
    ## Generate a sample of length N from U(0, 1) to determine which
    ## distribution should the sample drawn from
    w <- runif(n)
    ## Select the right component distribution and generate the random
    ## sample
    breaks = cumsum(mix$pr)

    m <- double(n)
    for(i in 1:n){
        m[i] = sum(w[i] > breaks) + 1
    }
    sim <- rnorm(n, mean = 0, sd = mix$pt[m])
    sim
}

mssmnorm <- function(varmix = disc(1, 1), xi = 1, moments = 1){
  if(moments == 1){
    m = (sqrt(2 * varmix$pt/pi) * (xi^2 - xi^-2)/(xi + xi^-1)) %*%
       varmix$pr
  } else if(moments == 2){
    m = (varmix$pt^2 * (xi^3 + xi^-3)/(xi + xi^-1)) %*% varmix$pr
  }
  drop(m)
}
```

Bibliography

Alexander, C. and Lazar, E. (2006). Normal mixture garch(1, 1): Applications to exchange rate modelling. *Journal of Applied Econometrics.*

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distribution. *Journal of the Royal Statistical Society.*

Ausin, M. C. and Galeano, P. (2006). Bayesian estimation of the gaussian mixture garch model. *Computational Statistics and Data Analysis.*

Azzalini, A. (1995). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics.*

Bachelier, L. (1900). Theorie de la speculation. *Annales Scientifiques.*

Ball, C. A. and Torous, W. N. (1983). A simplified jump process for common stock return. *Journal of Financial and Quantitative Analysis*, 18(1).

Barndorff-Nielsen, O. E. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandanavian Journal of Statistics.*

Bera, A. K. and Higgins, M. L. (1993). Arch models: Properties, estimation and testing. *Journal of Economic Survey.*

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics.*

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics.*

Bollerslev, T. (2008). Glossary to arch (garch). *Center fore Research in Econometric Analysis of Time Series.*

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). Arch modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics.*

Chang, K.-H. (2010). Semiparametric scale mixtures of normal distributions. *Auckland University Masters thesis.*

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review.*

Dax, A. (1990). The smallest point of a polytope. *Journal of Optimization Theory and Applications*, 64(2).

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica.*

Epps, T. W. and Epps, M. L. (1976). The stochastic dependence of security price changes and transaction volumnes: Implication for the mixture of distributions hypothesis. *Econometrica*, 44(2).

Fernandez, C. and Steel, M. F. J. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association.*

Fiorentini, G., Calzolari, G., and Panattoni, L. (1996). Analytic derivatives and the computation of garch estimates. *Journal of Applied Econometrics.*

G.Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications.* Institute of Mathematical Statistics and the American Statistical Association.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation.*

Hansen, P. R. and Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a garch(1,1)? *Journal of Applied Econometrics.*

Harvey, A. C. (1990). *The Econometric Analysis of Time Series.* MIT press.

Inc, S. I. (2012). *SAS/ETS Users Guide, Version 8.* SAS Institute Inc.

Kiefer, J. and Wolfowitz, J. (1956). Cosistency of the maximum-likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics.*

Kuester, K., Mittnik, S., and Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics.*

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business.*

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica.*

Robbins, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution. *Annals of Mathematical Statistics.*

Trapletti, A. and Hornik, K. (2012). *tseries: Time Series Analysis and Computational Finance.* R package version 0.10-28.

Tsay, R. S. (2002). *Analysis of Financial Time Series.* Wiley Series in Probability and Statistics.

Wang, Y. (2007). On fact computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of Royal Statistical Society.*

Wang, Y. (2010). Maximum likelihood computation for fitting semiparametric mixture models. *Journal of Statistical Computing.*

West, M. (1987). On scale mixtures of normal distribution. *Biometrika.*

with contributions from Slava Razbash, R. J. H. and Schmidt, D. (2012). *forecast: Forecasting functions for time series and linear models.* R package version 3.21.

Wuertz, D., with contribution from Michal Miklovic, Y. C., Boudt, C., Chausse, P., et al. (2012). *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. R package version 2110.80.

Wurtz, D., Chalabi, Y., and Lukson, L. (Draft). Parameter estimation of arma models with garch/aparch errors an r and splus software implementation. *Journal of Statistical Software*.