Minimally Cross-Entropic Conditional Density: A Generalization of the GARCH Model

Zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für
Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)
angenommene

Dissertation

von

Dipl. Wi.-Ing. Matthias Scherer

Tag der mündlichen Prüfung:	11. Mai 2011
Referent:	Prof. Dr. S.T. Rachev
Korreferent:	Prof. Dr. M. Uhrig-Homburg
	2011, Karlsruhe

To Ela

Acknowledgements

I would like to express my deeply felt gratitude to my supervisor Prof. Dr. Svetlozar. T. Rachev and his assistant Dr. Young Shin Kim for their very enthusiastic and inspiring support during my work. I am very grateful for the possibilities they gave me and the trust they had in me. I am also indebted to my co-supervisor Prof. Dr. Marliese Uhrig-Homburg and to Prof. Dr. Frank J. Fabozzi for their support in writing and finishing this dissertation.

Special thanks go to my family and my friends without whom this effort would have been impossible. In particular, I would like to dedicate a thank you to Ms. Elżbieta Żuk and Mr. Simon Notheis for their great suggestions and valuable advices.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

$\mathbb{R}_{>0}$	Set of positive real numbers
$\mathbb{Z}_{>0}$	Set of positive integers
$(X_t)_{t\in T}$	Stochastic process
$\mathcal{F}_t$	Natural filtration
$\mathrm{E}[X]$	Expected value of X
$\mathrm{V}[X]$	Variance of X
$\mathrm{S}[X]$	Skewness of X
$\mathrm{K}[X]$	Kurtosis of X
$\mathrm{Cov}[X,Y]$	Covariance between X and Y
$\mathrm{Corr}[X,Y]$	Correlation between X and Y
$SI(P)$	Self-information of P
$H(P,Q)$	Cross-entropy between P and Q
MLE	Maximum likelihood estimation
QMLE	Quasi-maximum likelihood estimation
OLSE	Ordinary least square estimation
EF	Estimation function
PDF	Probability density function
CDF	Cumulative distribution function
CF	Characteristic function
TS	Family of tempered stable distributions
CTS	Classical tempered stable distributions
EP	Exponential power distribution
SEP	Skewed exponential power distribution
ARMA	Autoregressive moving average
GARCH	General autoregressive conditional heteroskedastic
ARCD	Autoregressive conditional density
MCECD	Minimally cross-entropic conditional density

Introduction

Since Markowitz (1952) introduced Modern Portfolio Theory, the mean-variance framework has been at the core of financial analysis. In particular the seminal Black-Scholes model with its Gaussian assumption restricts the distributional parameter set to mean and variance thus postulating the variance as a measure of risk. Given the stylized facts, that empirical distributions of log-returns in financial time series are generally asymmetric (non-zero skewness) with a significant probability of high losses (leptokurtosis), the assumption of normality has been rejected in numerous applications.

Another generally accepted stylized fact in the finance literature is the volatility clustering. It describes the tendency of large changes to be followed by large changes and small changes to be followed by small changes. The models proposed by Engle (1982) and Bollerslev (1986)–autoregressive conditional heteroskedasticity (ARCH) and generalized ARCH (GARCH)–are recognized as the leading concepts for modeling time-varying volatility in financial time series. This fact is reflected in the unparalleled growth of the GARCH literature, including numerous variants and applications over the past decades. Although the first formal approach to analyze the behavior of speculative prices dates back to Bachelier (1900), it was Mandelbrot's groundbreaking papers (Mandelbrot (1963) and Mandelbrot (1967)) that found clear empirical evidence for changes in the variance over time. With Engle (1982) and Bollerslev (1986) a mathematical formulation of heteroskedasticity was provided which until now has been extended and modified to cover more sophisticated empirical facts.

As a consequence of the rejection of the Gaussian assumption and the time-varying property of certain distributional parameters, generalizations of the GARCH model have been suggested. They consider non-zero skewness as well as leptokurtosis and at the same time allow for time-varying features not only in the mean and variance. It was Hansen (1994) who argued that “there is no reason to assume, in general, that the only features of the conditional distribution which depend upon the conditioning information are the mean and variance”.[1] This led to the introduction of the first GARCH-like approach to conditional density, that is the autoregressive conditional density (ARCD). His original concept is based on a specific distributional assumption, the skewed Student’s t distribution. Parameter dynamics are modeled by independent autoregressions of corresponding moments. Various empirical studies have already been conducted to analyze and test the behavior of the ARCD model, stating the necessity for conditional density models.

In this thesis, we introduce a new, discrete time model for conditional densities which includes the GARCH model as a special case. Our approach resorts to the cross-entropy concept from information theory in order to model the parameter dynamics. The minimally cross-entropic conditional density (MCECD) model overcomes three shortcomings of the classical autoregression-based approach. First, there is a direct link between conditional distribution and parameter dynamics, thereby avoiding the problems associated with moment estimators. For some distributions—such as the stable Paretian distribution—even the first and second moments may not be finite, which makes sample moments unsuitable for parameter inference. Furthermore, there is no optimal estimator for higher moments available, as discussed by Kim and White (2004), leading to numerous alternative ARCD specifications for skewness and kurtosis dynamics as reported by Dark (2010). Second, MCECD consistently models multiple time-varying parameters and accounts for potential inter-dependencies. In ARMA-GARCH, each new observation is interpreted as a driver for both changing mean and volatility at the same time. New facts can, however, only signal a change in one factor. As a result, the use of ARMA-GARCH estimated parameter trajectories for conditional density models is problematic. Finally, MCECD can cope with a non-linear parameter process, thus significantly improving the explanatory power. Higher moments represent a non-linear feature of a random variable but classical autoregression is a linear model even if applied to non-linear estimators, e.g. absolute or squared values.

For skewness and kurtosis analyses, the selection of the underlying distribution is crucial. Suitable candidates can be found in the classes of

[1] The work of Gallant *et al.* (1991) had already promoted the idea of a conditional density.

tempered stable and tempered infinitely divisible distributions. Analogously to the stable Paretian distribution, there is unfortunately no mathematical expression for their density functions available which calls for a FFT-based approximation, implemented in our research papers Scherer *et al.* (2010a). Our analysis relies on these distributional families since they provide sufficient flexibility to describe empirical log-return distributions in financial time series.

This work is organized as follows: Chapter 1 gives a short overview of relevant concepts in statistics and probability theory. The focus is on describing non-Gaussian probability laws and introducing likelihood-based inference methods. Chapter 2 deals with various time series models that are dominating current research. Special concern is paid to ARMA, GARCH, and ARMA-GARCH specifications and followed by a brief discussion of several estimation methods for these models. Following the review of basic econometric theory, we start with the discussion of our contribution. Chapter 3 consists of the definition of our general MCECD model and a discussion of the stationarity property. Chapters 4 and 5 focus on specific applications of the MCECD model with regard to time-varying volatility and skewness. From the theoretical and empirical analyses, we derive strong arguments in favor of relevant MCECD specifications, when compared to existing models, e.g. GARCH, ARMA-GARCH, and ARCD. Finally, we conclude our work with a short overview of the main results and future research potential.

Chapter 1

Probability theory

In this chapter we will review some of the main theoretical concepts which constitute the background to the idea of the MCECD model. In particular, we will focus on non-Gaussian distributional assumptions and the relation between likelihood and cross-entropy with regard to parameter inference.

1.1 Distributions and random variables

In our analysis we assume the probability space $(\mathbb{R}, \wp(\mathbb{R}), P)$, where $\wp(\mathbb{R})$ denotes the Borel set on $\mathbb{R}$. The random variable $X : \mathbb{R} \to \mathbb{R}$ is a $\wp(\mathbb{R})$-measurable function for which $P(X < x)$ is differentiable and invertible. The probability law P is fully determined by either of the three functional expressions: its cumulative distribution function (CDF) which is given by $F_X : \mathbb{R} \to [0, 1],\ F_X(x) = P(X < x)$, its probability density function (PDF) defined as $f_X : \mathbb{R} \to \mathbb{R}_{>0},\ f_X(x) = \frac{dP(X<x)}{dx}$ and its characteristic function (CF) $\phi_X : \mathbb{R} \to \mathbb{C}$, which is the Fourier transform of the PDF

$$\phi_X(u) := \mathrm{E}[e^{iuX}] \ .$$

The inverse formula of the Fourier transform yields

$$f_X(x) = \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-iux} \cdot \phi_X(u) du \ .$$

In statistics a probability law can be characterized based on four features:

- Location
- Scale
- Asymmetry
- Shape

For each of those there are different measures available. We will focus on the statistical moments as measures. The n-th moment is defined as

$$\mathrm{E}[X^n] = \int\limits_{-\infty}^{\infty} x^n f_X(x) dx$$

and the n-th central moment as

$$\mathrm{E}[(X - \mathrm{E}[X])^n] = \int\limits_{-\infty}^{\infty} (x - \mathrm{E}[X])^n f_X(x) dx \ .$$

The first moment $\mathrm{E}[X]$ is called the mean and describes the location of a distribution which is the “center” of the probability mass. If X is a random variable, then it provides information about the average value of the observations according to the “Law of Large Numbers”. This sample mean is calculated as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The variance is the second central moment $\mathrm{V}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$ and it is a measure of how the observations are spread around the mean. The sample moment is given by

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

In order to describe the asymmetry of the probability law, the skewness is most commonly used. It is a rescaled third central moment defined by

$$\mathrm{S}[X] = \frac{\mathrm{E}[(X - \mathrm{E}[X])^3]}{\mathrm{V}[X]^{3/2}}.$$

It is applied to gain information on whether or not the distribution is symmetric around the mean and, in case of asymmetry, in which direction the distribution is skewed. A zero skewness indicates symmetry, a positive skewness means that compared to the left tail, the right tail of the distribution is elongated, and for a negative skewness it is *vice versa.* The tails constitute the endings of the distribution. The corresponding sample moment is

$$\hat{\varsigma} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(s^2)^{3/2}}.$$

The shape of a distribution is determined by the concentration of probability mass in its tails. The corresponding measure is the kurtosis, a rescaled fourth central moment, which is given by

$$\mathrm{K}[X] = \frac{\mathrm{E}[(X - \mathrm{E}[X])^4]}{\mathrm{V}[X]^2}.$$

The Gaussian distribution has a kurtosis value of 3. This is the reference to assess the thickness of the distributional tails. If the kurtosis is above 3, the distribution is called leptokurtic, which means that its tails are heavier than in the normal case and its “peakedness” is higher. This also implies that rare events are more likely than in the normal case. A distribution with a kurtosis below 3 is called platykurtic. Its tails are lighter and it is less peaked compared to the normal distribution. As a result kurtosis is a measure for the probability of extreme events.

It is important to note that without the functional definition mean, variance, skewness and kurtosis alone cannot provide a comprehensive description of a probability law. In case the moments of all orders n are known, the distribution is completely defined. This is a result of the relation between the moment generating function $M_X(u)$ given by

$$M_X(u) = \mathrm{E}[e^{uX}]$$

and the CF $\phi_X(u)$

$$\phi_X(u) = M_{iX}(u) = M_X(iu).$$

In other words, the CF is the moment generating function of iX. Moreover, it holds that

$$\mathrm{E}[X^n] = \frac{d^n M_X(u)}{du^n}|_{u=0},$$

which means that $M_X(u)$ is determined if the moments of all order n are known and so is the CF. This is also the basic principle of parameter inference using moment estimators. Given that we know the functional form $f_\theta(x)$ of the probability law, sample moments can be used to fit parameters to the empirical data.

The cumulant generating function $g_X(u)$ is closely related to $M_X(u)$ and therefore also represents a potential characterization of the probability law. It is defined as the logarithm of the moment generating function

$$g_X(u) = \log(M_X(u)).$$

The cumulant of order n is given by

$$c_n(X) = \frac{d^n g_X(u)}{du^n}|_{u=0}.$$

Despite of its similarities to the moment-generating function, the advantage is that the cumulants directly yield the central moments $\mathrm{E}[(X - \mathrm{E}[X])^n]$

$$\begin{aligned}
\mathrm{E}[(X - \mathrm{E}[X])^2] &= c_2(X) = \mathrm{V}[X] \\
\mathrm{E}[(X - \mathrm{E}[X])^3] &= c_3(X) \\
\mathrm{E}[(X - \mathrm{E}[X])^4] &= c_4(X) + 3c_2^2(X).
\end{aligned}$$

Skewness and kurtosis of a random variable can also be expressed by means of cumulants

$$\begin{aligned}
\mathrm{S}[X] &= \frac{c_3(X)}{c_2(X)^{3/2}} \\
\mathrm{K}[X] &= \frac{c_4(X)}{c_2(X)^2} + 3.
\end{aligned}$$

This result highlights the close connection between parameters and moments.

In statistics there are two important concepts to describe the relation between two different random variables X and Y

- Dependence

- Correlation

The dependence is directly defined by the probability laws of the random variables. Let $f_X(x)$ and $f_Y(y)$ be the PDFs of X and Y respectively, and $f_{X,Y}(x,y)$ be the common PDF of the pair $(X,\ Y)$. In this case the two random variables X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).$$

Alternatively, the condition can be formulated by means of the CDFs of the two random variables

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).$$

The correlation measure is based on the covariance given by

$$\mathrm{Cov}[X,Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])].$$

The variance is hence the covariance of X with itself

$$\mathrm{Cov}[X,X] = \mathrm{V}[X].$$

The correlation is a standardized form of the covariance

$$\mathrm{Corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{V}[X]}\sqrt{\mathrm{V}[Y]}}.$$

Two random variables are called uncorrelated if their correlation is zero

$$\mathrm{Corr}[X,Y] = 0.$$

The relation between dependence and correlation is given by the following statement: If X and Y are independent, then they are also uncorrelated.

This can be easily seen by

$$\begin{aligned}
\operatorname{Cov}[X,Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathrm{E}[X])(y-\mathrm{E}[Y])\ f_{X,Y}(x,y)\ dy\ dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathrm{E}[X])(y-\mathrm{E}[Y])\ f_X(x)\ f_Y(y)\ dy\ dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mathrm{E}[X])\ f_X(x)\ dx(y-\mathrm{E}[Y])\ f_Y(y)\ dy \\
&= 0.
\end{aligned}$$

1.2 Skewness and heavy-tails

Prior to the groundbreaking works of Mandelbrot (1963) and Fama (1963), it was assumed that return distributions follow the normal law. Since the early 1960s a considerable number of empirical studies have documented that this assumption should be rejected.[1] The findings of these studies suggest that return distributions have heavier tails than the normal distribution (i.e., exhibit leptokurtosis) and have non-zero skewness (i.e., are asymmetric). In this section, we will highlight two different generalization techniques of the normal probability law and present specimens for each class.

The Gaussian distribution $N(\mu,\sigma^2)$ with location parameter μ and scale parameter σ is completely determined by its PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right),$$

or its CF

$$\phi_X(u) = \exp\left(iu\mu - \frac{1}{2}\sigma^2u^2\right).$$

The expected value equals the location parameter $\mathrm{E}[X] = \mu$ and the variance equals the squared scale parameter $\mathrm{V}[X] = \sigma^2$. Skewness is always zero and its kurtosis is 3.

[1] For a review of these studies, see Rachev *et al.* (2005).

The first way of generalizing the Gaussian distribution to account for leptokurtosis is to let the exponent 2 vary. This yields the exponential power distribution[2] (EP) with the PDF

$$f_X(x) = \frac{\alpha}{2\sigma\Gamma(1/\alpha)} \exp\left(-\frac{|x-\mu|^\alpha}{\sigma^\alpha}\right),$$

where $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_{>0}$, and $\alpha \in \mathbb{R}_{>0}$ are the parameters for location, scale and shape (kurtosis) respectively. This family includes the Gaussian and the Laplace distribution as special cases for $\alpha = 2$ and $\alpha = 1$. For the parameter range of $\alpha \in (0, 2)$, the distribution has heavier tails than the Gaussian one and for $\alpha \in (2, \infty)$ it has lighter tails. The moments of the EP distributions are: $\mathrm{E}[X] = \mu$, $\mathrm{V}[X] = \sigma^2\Gamma(3/\alpha)/\Gamma(1/\alpha)$, $\mathrm{S}[X] = 0$, and $\mathrm{K}[X] = \Gamma(5/\alpha)\Gamma(1/\alpha)/\Gamma(3/\alpha)^2$.

In order to introduce non-zero skewness, the skewed exponential power distribution (SEP) has been proposed by Zhu and Zinde-Walsh (2009) as a generalization of the exponential power distribution (EP). Given the parameters for location $\mu \in \mathbb{R}$, scale $\sigma > 0$, shape $\alpha > 0$, and skewness $\beta \in (0, 1)$, the PDF of the SEP is

$$f_X(x) = \begin{cases} \frac{1}{\sigma}K(\alpha)\exp\left(-\frac{1}{\alpha}\left|\frac{x-\mu}{2\beta\sigma}\right|^\alpha\right) & : \quad x \le \mu \\ \frac{1}{\sigma}K(\alpha)\exp\left(-\frac{1}{\alpha}\left|\frac{x-\mu}{2(1-\beta)\sigma}\right|^\alpha\right) & : \quad x > \mu, \end{cases} \tag{1.1}$$

where $K(\alpha) = [2\alpha^{1/\alpha}\Gamma(1 + 1/\alpha)]^{-1}$. The corresponding mean and variance can be derived analytically

$$\begin{aligned} \mathrm{E}[X] &= \frac{1}{K(\alpha)}\frac{\alpha\Gamma(2/\alpha)}{\Gamma^2(1/\alpha)}\left[(1-\beta)^2 - \beta^2\right] \\ \mathrm{V}[X] &= \frac{1}{K(\alpha)^2}\frac{\alpha^2\Gamma(3/\alpha)}{\Gamma^3(1/\alpha)}\left[(1-\beta)^3 - \beta^3\right] - E^2[X]. \end{aligned}$$

There are, however, two drawbacks of this approach: the density functions are not differentiable at $x = \mu$ and the moments depend on several distributional parameters. As a result, the modeling of the key features of a probability law, such as location, scale, asymmetry and shape, is cumbersome.

The alternative Gaussian generalizations allow constructing distribu-

[2] Subbotin (1923) first proposed this probability law as the generalized error distribution (GED). Box and Tiao (1973) then introduced the name exponential power distribution.

tions, which are differentiable at all points and offer dedicated parameters to manipulate the four distributional features. The idea for this approach is to extend the CF rather than the PDF. The example most commonly known for this type of generalization is the stable Paretian distribution. It is defined by its characteristic function $\phi(u;\alpha,\beta,C,\mu)$

$$\phi_X(u) = \exp\left(iu\mu - C|u|^{\alpha}(1 - i\beta\,\mathrm{sign}(u)z(u,\alpha))\right), \tag{1.2}$$

where $\mu \in \mathbb{R}$, $C > 0$, $\beta \in [-1,1]$, and $\alpha \in (0,2]$ drive mean, dispersion, skewness and kurtosis, respectively, and

$$z(u,\alpha) := \begin{cases} \tan(\frac{\pi\alpha}{2}) & : \quad \alpha \neq 1 \\ -\frac{2}{\pi}\ln|u| & : \quad \alpha = 1. \end{cases}$$

There are three well-known special cases of the stable law, namely the Cauchy distribution ($\alpha = 1$, $\beta = 0$), the Gaussian distribution ($\alpha = 2$, $\beta = 0$), and the Lévy distribution ($\alpha = 0.5$, $\beta = 1$) for which there exists a closed-form expression of the PDF. In general, the PDF has to be approximated using the Fast Fourier transform, which is a computationally efficient procedure for the Discrete Fourier transform.[3] Its appealing property is the so-called stability property which claims that the sum of rescaled stable random variables with common stability index α, follows again a stable Paretian distribution with stability index α. The drawback of this distribution is that for any $\alpha \leq n$, the expected value $\mathrm{E}[X^n]$ is infinite. This is caused by the thickness of its tails and does not allow for moment modeling without tail truncation.

As a result, Rosiński (2007) introduced the class of tempered stable (TS) distributions, which exhibit thinner tails than the stable Paretian model, but still allow for leptokurtosis. The specimens of this class are defined by the Lévy tupel (γ, σ^2, ν). Applying the Lévy-Khintchine representation in Sato (1999) gives as a result the corresponding CF. The classical tempered stable

[3] FFT-based approximation of the stable Paretian PDF has been suggested by DuMouchel (1975).

(CTS) distribution, for example, is given by

$$\begin{aligned}
\gamma &= m - \int\limits_{|x|>1} x\nu(dx) \\
\sigma^2 &= 0 \\
\nu(dx) &= \Big(C_+ e^{-\lambda_+ x}\mathbf{1}_{x>0} + C_- e^{-\lambda_- x}\mathbf{1}_{x<0}\Big)\,\frac{dx}{|x|^{\alpha+1}}\ .
\end{aligned}$$

where $C_+, C_-, \lambda_+, \lambda_- \in \mathbb{R}_{>0}$, $\alpha \in (0,2)$, $m \in \mathbb{R}$, and $\mathbf{1}_A$ denotes the indicator function. Another representation of the CF is thus

$$\begin{aligned}
\phi_X(u) = \exp\Big\{ & ium - iu\Gamma(1-\alpha)(C_+\lambda_+^{\alpha-1} - C_-\lambda_-^{\alpha-1}) \\
& + C_+\Gamma(-\alpha)\big((\lambda_+ - iu)^\alpha - \lambda_+^\alpha\big) \\
& + C_-\Gamma(-\alpha)\big((\lambda_- + iu)^\alpha - \lambda_-^\alpha\big)\Big\}\ . \qquad (1.3)
\end{aligned}$$

This yields for the cumulants of a CTS distributed random variable

$$\begin{aligned}
c_1(X) &= m \\
c_n(X) &= C_-\ \Gamma(n-\alpha)\lambda_+^{\alpha-n} + (-1)^n C_-\ \Gamma(n-\alpha)\lambda_-^{\alpha-n}, \text{ for } n > 0.
\end{aligned}$$

For $\lambda = \lambda_+ = \lambda_-$, $C_+ = C \cdot \frac{1+\beta}{2}$, and $C_- = C \cdot \frac{1-\beta}{2}$, where $\beta \in (-1,1)$, the CTS turns into an adjusted version of the distribution suggested by Koponen (1995). Its CF takes the form

$$\begin{aligned}
\phi_X(u) = \exp\Big\{ & ium - iu\ \Gamma(1-\alpha)\ C\lambda^{\alpha-1}\beta + C\ \Gamma(-\alpha) \\
& \cdot\Big[\frac{1+\beta}{2}\ ((\lambda - iu)^\alpha - \lambda^\alpha) + \frac{1-\beta}{2}((\lambda + iu)^\alpha - \lambda^\alpha)\Big]\Big\}\ .
\end{aligned}$$

Its advantages result from the parameterization. Each parameter governs one of the important features of a random variable.

$$\begin{aligned}
\mathrm{E}[X] &= m \\
\mathrm{V}[X] &= \Gamma(2-\alpha)\ C\ \lambda^{\alpha-2} \\
\mathrm{S}[X] &= \frac{\Gamma(3-\alpha)\ C\ \lambda^{\alpha-3}\beta}{\mathrm{V}[X]^{3/2}} \\
\mathrm{K}[X] &= \frac{\Gamma(4-\alpha)\ C\ \lambda^{\alpha-4}}{\mathrm{V}[X]^2} + 3
\end{aligned}$$

We can use parameter m for location, C for scale, β for skewness, and α and λ for kurtosis. The standard Koponen model results from solving $\mathrm{V}[X] = 1$

for parameter C_0, which leads to

$$C_0 = \frac{1}{\Gamma(2-\alpha)\ \lambda^{\alpha-2}}.$$

The class of tempered infinitely divisible (TID) distribution introduced in Bianchi *et al.* (2010) stems from an alternative definition of the CF. A representative of this family is the rapidly decreasing tempered stable (RDTS) distribution, defined by the Lévy tupel (γ, σ^2, ν) with

$$\begin{aligned}
\gamma &= m - \int\limits_{|x|>1} x\nu(dx) \\
\sigma^2 &= 0 \\
\nu(dx) &= \left(C_+ e^{-\lambda_+^2 \frac{x^2}{2}} \mathbf{1}_{x>0} + C_- e^{-\lambda_-^2 \frac{x^2}{2}} \mathbf{1}_{x<0} \right) \frac{dx}{|x|^{\alpha+1}}\ , \qquad (1.4)
\end{aligned}$$

where $C_+, C_-, \lambda_+, \lambda_- \in \mathbb{R}_{>0}$, $\alpha \in (0,2)$, and $m \in \mathbb{R}$. Using the Lévy-Khintchine representation, the characteristic function of a RDTS random variable $X \sim \text{RDTS}(\alpha, C_+, C_-, \lambda_+, \lambda_-, m)$ takes the form

$$\begin{aligned}
\phi_X(u) &= \exp\left(ium - iu \int\limits_{|x|>1} x\nu(dx) + \int\limits_{\mathbb{R}} \left(e^{iux} - 1 - iux\mathbf{1}_{|x|<1}\right) \nu(dx) \right) \\
&= \exp\left(ium + \int\limits_{\mathbb{R}} \left(e^{iux} - 1 - iux\right) \nu(dx) \right)\ . \qquad (1.5)
\end{aligned}$$

In Kim *et al.* (2010) this result was reformulated using the confluent hypergeometric function $M(a, b; z)$

$$\phi_X(u) = \exp\Big(ium + C_+ \cdot G(iu; \alpha, \lambda_+) + C_- \cdot G(-iu; \alpha, \lambda_-) \Big), \qquad (1.6)$$

where $G(x; \alpha, \lambda)$ is defined as

$$\begin{aligned}
G(x; \alpha, \lambda) := 2^{-\frac{\alpha}{2}-1}\ \lambda^\alpha\ \Gamma\left(-\frac{\alpha}{2}\right)\ &\left[M\left(-\frac{\alpha}{2}, \frac{1}{2}; \frac{x^2}{2\lambda^2} \right) - 1 \right] + \\
+ 2^{-\frac{\alpha}{2}-\frac{1}{2}}\ \lambda^{\alpha-1}\ x\ \Gamma\left(\frac{1-\alpha}{2}\right)\ &\left[M\left(\frac{1-\alpha}{2}, \frac{3}{2}; \frac{x^2}{2\lambda^2} \right) - 1 \right].
\end{aligned}$$

This yields for the cumulants $c_n(X)$

$$
\begin{aligned}
c_1(X) &= m \\
c_n(X) &= 2^{\frac{n-\alpha-2}{2}} \Gamma\left(\frac{n-2}{\alpha}\right) \left(C_+ \lambda_+^{\alpha-n} + (-1)^n C_- \lambda_-^{\alpha-n}\right) \text{ for } n > 1.
\end{aligned}
$$

There are closed-form expressions available for the mean, variance, skewness, and kurtosis of a RDTS distributed random variable X

$$
\begin{aligned}
\mathrm{E}[X] &= m \\
\mathrm{V}[X] &= 2^{-\alpha/2}\ \Gamma\left(\frac{2-\alpha}{2}\right) \ \left(C_+\ \lambda_+^{\alpha-2} + C_-\ \lambda_-^{\alpha-2}\right) \\
\mathrm{S}[X] &= \frac{2^{\alpha/4+1/2}\ \Gamma\left(\frac{3-\alpha}{2}\right)\ \left(C_+\ \lambda_+^{\alpha-3} - C_-\ \lambda_-^{\alpha-3}\right)}{\mathrm{V}[X]^{3/2}} \\
\mathrm{K}[X] &= \frac{2^{\alpha/2+1}\ \Gamma\left(\frac{4-\alpha}{2}\right)\ \left(C_+\ \lambda_+^{\alpha-4} + C_-\ \lambda_-^{\alpha-4}\right)}{\mathrm{V}[X]^{2}} + 3\ .
\end{aligned}
$$

For a comprehensive definition of the TS and TID class, we refer to Rosiński (2007) and Bianchi *et al.* (2010).

1.3 Likelihood and entropy

Parameter inference using MLE goes back to the seminal work of Fisher (1922). Decades later Godambe (1960) proved that the MLE is optimal among all estimating functions regarding efficiency.[4] Compared to other inference methods, such as (generalized) methods of moments (GMM), it does not depend on moment estimators. Given a PDF $f_\theta : \mathbb{R} \to \mathbb{R}_{>0}$ with parameter vector θ and the observation vector x, the MLE parameters can be derived from the first-order optimality of the log-likelihood function under certain smoothness conditions

$$
\frac{\partial \log f_\theta(x)}{\partial \theta} = 0.
$$

This makes MLE especially attractive for applications with non-zero skewness and leptokurtosis, where sample moments might differ significantly from the underlying value.[5]

[4] See Bera and Bilias (2002) for a historical review of parameter estimation.

[5] See chapter 5 for an analysis of the skewness case.

For the inference not to be ill-posed, the number of observations should be greater than or equal to the dimension of the parameter vector. A simple example demonstrates this condition. Given the PDF of a Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$ and one observation x_1, applying the first-order condition leads to the following estimators: $\mu = x_1$ and $\sigma^2 = 0$. A zero variance suggests, however, that the observed process is non-stochastic, which is inconsistent with our assumption. If only one observation is available, only one parameter can be estimated. The remaining components of the vector θ have to be given *ex ante*.

The term entropy originates from thermodynamics and defines a measure for the disorder within a system. Shannon (1948) extended the definition for the use in information theory, where it is a measure of uncertainty associated with a random variable. In a probability space $(\Omega, \wp, P)$, the entropy $H(X)$ of a finite-state $\wp$-measurable random variable X with probabilities $P(X = x_i) = p_i$ for $i = 1, ..., n$ is mathematically speaking

$$H(X) := -\sum_{i=1}^{n} p_i \cdot \log(p_i).$$

The higher the entropy $H(X)$, the higher the disorder, or the lesser the available information. It is particularly relevant that the term $-\log(p_i)$ is referred to as self-information (SI) and is a measure of the information content associated with the outcome of X.

Given an alternative distribution Q defined on the measurable space $(\Omega, \wp)$ and $Q(X = x_i) = q_i$, then the cross-entropy[6] is given by

$$H(P, Q) := -\sum_{i=1}^{n} p_i \cdot \log(q_i).$$

This term is closely related to the Kullback-Leibler (KL) divergence $D(P||Q)$ (also known as relative entropy, Kullback (1959)).

$$D(P||Q) := \sum_{i=1}^{n} p_i \cdot \log\left(\frac{p_i}{q_i}\right).$$

Thus cross-entropy can be decomposed into the entropy and the KL diver-

[6] The concept was first introduced as “inaccuracy” by Kerridge (1961).

gence[7]

$$H(P, Q) := H(P) + D(P||Q).$$

We can observe that cross-entropy minimization against the uniform distribution $P(X = x_i) = \frac{1}{n}$

$$H(P, Q) := -\frac{1}{n} \cdot \sum_{i=1}^{n} \log(q_i).$$

is the equivalent to log-likelihood maximization for the distribution Q. For a non-trivial distribution P, the minimum cross-entropy can be interpreted as a weighted MLE, where P determines the importance of the observations x_i.[8] On the other hand, the KL divergence is a measure of distance between two distributions. Hence, an alternative view is that minimizing the cross-entropy minimizes the difference between the theoretical *a priori* probability model P and the empirical *a posteriori* Q.

Maximum entropy (ME) and minimum cross-entropy (MCE) are already an integral part of several important concepts and applications. Jaynes (1957) introduced the principle of maximum entropy, which is applied for parameter inference in the empirical likelihood method by Owen (1988). The principle of minimum discrimination information (MDI) by Kullback (1959)—sometimes also called principle of minimum cross-entropy—is with particular relevance to our approach. MDI postulates that given new facts, a new distribution should be chosen which is as close (regarding KL divergence) as possible to the original distribution, so that the information gain by new data is as small as possible. Under the assumption that P is known and fixed, the minimization only affects the measure Q. If, moreover, a functional form for the distribution of Q is given, then the method optimizes the parameter vector θ of Q. The results of the cross-entropy minimization in this case equal the ones from minimizing the KL divergence

$$\underset{\theta}{\operatorname{argmin}}\, H(P, Q(\theta)) = \underset{\theta}{\operatorname{argmin}}(H(P) + D(P||Q(\theta))) = \underset{\theta}{\operatorname{argmin}}(D(P||Q(\theta))).$$

For our model, we will apply the cross-entropy minimization to describe the parameter dynamics for the conditional density. Briefly speaking,

[7] See Kannappan (1972) and Sharma and Taneja (1974) for a common characterization of entropy, cross-entropy, and KL divergence measure.

[8] See Bera and Bilias (2002) for an overview of the link between minimum cross-entropy and maximum likelihood.

MCECD is defined as the likelihood-based alternative for ARCD, just as MLE is the likelihood-based alternative for GMM.

Chapter 2

Econometric models

Our subsequent analyses are based on time series models rather than factor models. A time series is a sequence of data points in our case historical observation of financial log-returns. Hence, the explanatory power of the presented models solely arises from the inherent sample data. A factor model, however, considers additional information from selected factors, e.g. economic variables, related time series.

This chapter deals with time series analysis after having introduced a discrete-time stochastic process. We will present well-known time series models such as ARMA, GARCH, and ARMA-GARCH. With special focus paid to the heteroskedasticity, we will outline various specifications dealing with different aspects of this phenomenon. Finally, we will take a glance at the parameter inference methods related to time series theory.

2.1 Stochastic processes

Stochastic processes form the foundation for modeling financial time series. Therefore, we take a closer look at this concept and review the theoretical background. Let $(\mathbb{R}, \wp(\mathbb{R}), P)$ be the probability space, where $\wp(\mathbb{R})$ denotes the Borel σ-algebra. Then a stochastic process $(X_t)_t$ is defined by the following functional relation

$$X : T \times \mathbb{R} \to \mathbb{R},$$

such that for every $t \in T$, X_t is $\wp(\mathbb{R})$-measurable, which means X_t is a random variable. $X(\bullet, x)$ is called the trajectory of a stochastic process and describes its path over time for a certain realization $x \in \mathbb{R}$.

Financial time series generally consist of a set of data points based on a certain frequency, such as daily, weekly, or monthly. Hence, the stochastic processes we consider in the following are in discrete time, as opposed to continuous time. Mathematically this translates into $T \in \mathbb{Z}_{>0}$ or $T \in \mathbb{Z}$ depending on whether or not the history is finite or infinite. In practical application the time horizon is of course always limited. In order to describe the history of observations from the time series, the mathematical concept of a natural filtration is used. A natural filtration $\mathcal{F}_t$ for a discrete time process $(X_t)_{t\in\mathbb{Z}_{>0}}$ is the σ-algebra

$$\mathcal{F}_t = \mathcal{F}_t^X = \wp\big(\{X_s | s \leq t\}\big).$$

The dynamics of a stochastic process are determined by the family of finite dimensional distributions on X. For all partitions $\{t_1, ..., t_n\}$ of T with $t_i \in T$ and $n \in \mathbb{N}$ arbitrary, this family is given by $P(X_{t_1} < x_1, ..., X_{t_n} < x_n)$, where $x_i \in \mathbb{R}$. Hence, a time-discrete stochastic process can be identified using data samples of the matching frequency. One of the most important properties of a stochastic process is defined based on this family of finite dimensional distributions: the strict stationarity. $(X_t)_t$ is strictly stationary if the distributions are invariant to time shifts

$$P(X_{t_1} < x_1, ..., X_{t_n} < x_n) = P(X_{t_1+k} < x_1, ..., X_{t_n+k} < x_n) \text{ , for all } k \in \mathbb{Z}. \tag{2.1}$$

In practice it is often cumbersome to test for strict stationarity because the distributions in equation (2.1) are not known *a priori*. The alternative concept of weak stationarity is in this sense much easier to apply. Let $(X_t)_t$ be a time series and $s < t$, then $(X_t)_t$ is weakly stationary if and only if

$$\begin{aligned} \mathrm{E}[X_t] &= \mu \\ \mathrm{Cov}[s, t] &= \mathrm{Cov}[t - s]. \end{aligned}$$

These conditions focus on the first two moments. The mean is required to be constant over time and the autocovariance depends only on the time difference, not on the specific points in time. It is obvious that weak stationarity can be tested calculating the sample moments for the given time series. Under the assumption of normality, the two definitions of stationarity coincide. This relation exists due to the fact that the Gaussian distribution basically models mean and variance. If we consider, however, leptokurtosis and non-zero skewness, the concept of strict stationarity is more relevant. If a time series is stationary, this means that past observations can be used to estimate the dynamics of future trajectories determined by the underlying distributional law which is constant over time.

A simple example of a time-discrete stochastic process is the Gaussian white noise process. In this case every random variable y_t is standard normal distributed $\mathrm{N}(0,1)$ and for $s < t$, y_s and y_t are independent random variables. As a result, the process is independent and identically distributed (i.i.d.) and thus also strictly stationary. A white noise process $(y_t)_t$ has a zero mean and constant variance

$$\begin{aligned}\mathrm{E}[y_t] &= 0\\ \mathrm{V}[y_t] &= \sigma^2,\end{aligned}$$

and for $s < t$ the random variables y_s and y_t are uncorrelated

$$\mathrm{E}[y_s \cdot y_t] = 0.$$

In econometrics, time series models are applied to describe financial returns. The return ${}_sR_t$ of an asset with price process $(S_t)_t$ between two points in time s and t with $s < t$ is defined by

$${}_sR_t := \frac{S_t - S_s}{S_s}.$$

This term implies that the asset is only traded at s and t. At today's stock markets, most assets are, however, traded almost continuously in time which demands for a continuous return process. One definition in this context is the spot return r_t^S at time t. It is derived by decreasing the time span $t - s$ of the average return to zero

$$\begin{aligned}r_t^S &= \lim_{s \to t} \frac{{}_sR_t}{t-s}\\ &= \frac{dS_t/dt}{S_t} = d\log(S_t).\end{aligned}$$

Using the spot return, the continuous return between time s and t is given by

$${}_sr_t = \int\limits_s^t r_u^S du = \log(S_t) - \log(S_s) = \log({}_sR_t).$$

Due to its defining expression, ${}_s r_t$ is often called the log-return. Since we use a discrete and equidistant time approximation in our analysis, we consider the log-return

$$r_t = \log(S_t) - \log(S_{t-1}).$$

In order to construct a time series model for log-returns based on the white noise process, it is necessary to adjust mean and variance. This yields the following dynamics

$$r_t = \mu + \sigma \cdot \epsilon_t,$$

where $(\epsilon_t)_t$ is white noise. Introducing time-varying mean μ and variance σ^2 leads us to the well-known ARMA, GARCH, and ARMA-GARCH models.

2.2 The ARMA model

The moving average (MA) model of order q is constructed from a weighted sum of the preceding realizations of the error process $(e_t)_t$

$$y_t = c + \sum_{j=1}^{q} a_j \ e_{t-j} + e_t,$$

where $c, a_i \in \mathbb{R}$ and $q \in \mathbb{N}$. The first-order MA process has a constant expected value

$$\mathrm{E}[y_t] = c + a_1 \cdot \mathrm{E}[e_{t-1}] + \mathrm{E}[e_t] = c,$$

and a constant variance

$$\mathrm{V}[y_t] = (1 + a_1^2)\sigma^2.$$

The autocovariance is

$$\begin{aligned}\mathrm{Cov}[y_t, y_{t-j}] &= \mathrm{E}[(y_t - \mathrm{E}[y_t])(y_{t-j} - \mathrm{E}[y_{t-j}])] \\ &= \mathrm{E}[(e_t + a_1 \ e_{t-1})(e_{t-j} + a_1 \ e_{t-j-1}].\end{aligned}$$

For $j = 1$ this yields

$$\mathrm{Cov}[y_t, y_{t-1}] = a_1 \cdot \sigma^2,$$

and for $j > 1$ it equals zero. From these results, we can draw the conclusion that a MA model is always weakly stationary. The form of the covariance is characteristic for the MA process of finite order q because it models only a finite number of correlations to past observations. Metaphorically speaking, the MA process has only a finite memory.

In opposition to the MA process, the autoregressive (AR) model has an infinite memory structure. AR(1) can be viewed as MA of infinite order. The dynamics for AR(p) are given by

$$y_t = c + \sum_{i=1}^{p} b_i \; y_{t-i} + e_t,$$

where $c, b_i \in \mathbb{R}$ and $p \in \mathbb{N}$. The AR(1) is stationary for $|b_1| < 1$, whereas for $|b_1| \geq 1$ the innovations accumulate rather than die out. A stationary AR(1) has the mean and variance

$$\begin{aligned} \mathrm{E}[y_t] &= \frac{c}{1 - b_1} \\ \mathrm{V}[y_t] &= \frac{\sigma^2}{1 - b_1^2}. \end{aligned}$$

The covariance is given by

$$\mathrm{Cov}[y_t, y_{t-j}] = \sigma^2 \cdot \frac{b_1^j}{1 - b_1^2}.$$

From this formula, we can observe that the correlation exponentially decays with increasing j because $|b_1| < 1$. This covariance structure highlights the infinite memory property because the correlation is non-zero for all j.

The autoregressive moving average (ARMA) model combines the features from AR and MA. It enables us to model infinite memory and at the same time emphasizes more recent observations. ARMA models can account for trends in the mean of the underlying data. For $c, a_i, b_i \in \mathbb{R}$ and $p, q \in \mathbb{N}$ the ARMA(p,q) model follows the dynamics

$$y_t = c + \sum_{i=1}^{p} b_i \; y_{t-i} + \sum_{j=1}^{q} a_j \; e_{t-j} + e_t,$$

where $(e_t)_t$ is an i.i.d. error process. p defines the order of the autoregressive part and the parameters b_i are the coefficients of the regression. Similarly, q

determines the order of the moving average part, whereas the a_i specify the weights for the moving average. The ARMA model is a conditional mean model. The conditional expected value using the natural filtration $\mathcal{F}_t$ of the error process $(e_t)_t$ yields

$$\mathrm{E}[y_t|\mathcal{F}_{t-1}] = \mu_t = c + \sum_{i=1}^{p} b_i \; y_{t-i} + \sum_{j=1}^{q} a_j \; e_{t-j}.$$

The ARMA process is stationary if and only if the roots of the equation

$$1 - \sum_{i=1}^{p} b_i \; z^i = 0$$

lie inside the unit circle. Non-stationary ARMA is called ARIMA, autoregressive integrated moving average model.

Under the assumption of the Gaussian distribution $e_t \sim \mathrm{N}(0, \sigma^2)$, the return process $(y_t)_t$ is also normally distributed with time-varying mean μ_t: $y_t \sim \mathrm{N}(\mu_t, \sigma^2)$. Subsequently, we present models with a focus on the scale parameter rather than the location parameter.

2.3 The GARCH model

In this paragraph we will tackle the phenomenon of conditional volatility and the corresponding time series models. Mandelbrot in his seminal papers (Mandelbrot (1963) and Mandelbrot (1967)) found clear empirical evidence for changes in the variance over time. This behavior is called heteroskedasticity as opposed to homoskedasticity. First approaches to model conditional volatility were implemented by exponential smoothing over the squared log-return process $(r_t)_t^2$. In his path-breaking work, Engle (1982) introduced the autoregressive conditional heteroskedastic model of order $q \in \mathbb{N}$ (ARCH(q)), which describes the volatility dynamics of a process $(y_t)_t$ by the following equation

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \; \sigma_{t-i}^2 \; \epsilon_{t-i}^2,$$

where $\alpha_0, \alpha_i \in \mathbb{R}_{\geq 0}$.

This model was generalized by Bollerslev (1986), who transferred the idea of ARMA to the volatility case and hence suggested the generalized

ARCH (GARCH(p,q))

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \; \sigma_{t-i}^2 \; \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \; \sigma_{t-j}^2.$$

Assuming $(\epsilon_t)_t$ to be white noise, yields $(y_t)_t$, $y_t = \mu + \sigma_t \epsilon_t$ for the log-return process. The conditional distribution of y_t on $\mathcal{F}_{t-1}$ is hence $\mathrm{N}(\mu, \sigma_t^2)$. The stationarity condition[1] for a GARCH(p,q) process is

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1.$$

In econometrics, the GARCH(1,1) is the most commonly used specification due to the small number of parameters. This helps avoiding overfitting and keeps inference computationally efficient. In this work we use GARCH as a synonym for GARCH(1,1) and restrict our analysis to this special case.

Christoffersen and Jacobs (2004) use a more general formulation of the GARCH model based on the "News Impact Function"[2] $g(\epsilon_{t-1})$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \cdot \sigma_{t-1}^2 \; g(\epsilon_{t-1}) + \beta_1 \cdot \sigma_{t-1}^2.$$

Depending on the specific choice for the news impact function $g(\bullet)$, different models can be derived. $g(z) = z^2$ yields the basic GARCH(1,1). The task of the news impact function is to transform the observations, or equivalently the innovations, into market signals. The resulting values mirror the information that the market associates with the historical values. This way, empirical findings in time series analysis can be incorporated into the GARCH framework.

Choosing $g(z) = (z - \delta)^2$ as the news impact function leads us to the N-GARCH proposed by Engle and Ng (1993) which is based on the volatility dynamics

$$\sigma_t^2 = \alpha_0 + \alpha_1 \cdot \sigma_{t-1}^2 (\epsilon_{t-1} - \delta)^2 + \beta_1 \; \sigma_{t-1}^2.$$

The GJR-GARCH which Glosten *et al.* (1993) introduced in their article is

[1] See Bougerol and Picard (1992) for a discussion of strict stationarity of GARCH models.

[2] Pagan and Schwert (1990), Engle and Ng (1993), Ding *et al.* (1993), and Hentschel (1995) have already considered a news impact function.

derived by setting $g(z) = (z^2 + \kappa\ z^2\ \mathbf{1}_{z>0})$

$$\sigma_t^2 = \alpha_0 + \alpha_1\ \sigma_{t-1}^2\ \epsilon_{t-1}^2 + \alpha_1\ \kappa\ \sigma_{t-1}^2\ \epsilon_{t-1}^2 \mathbf{1}_{e_{t-1}>0} + \beta_1\ \sigma_{t-1}^2.$$

Both models resort to an asymmetric news impact function. In particular, positive observations have a different effect on the volatility compared to negative ones. Black (1976) analyzes this effect and derives clear empirical evidence that after negative log-return volatility increases significantly more than after positive log-returns. This finding is commonly known as the “leverage effect”. A possible explanation for this effect is that bad news increases the fear of even further losses. This leads to an increase in trading activities and also to an increase of volatility. In contrast, positive news might be interpreted as confirmation of the current strategy and, as such, have a calming effect. From a technical point of view, the difference between N-GARCH and GJR-GARCH is that N-GARCH uses a shift in its news impact function, whereas GJR-GARCH applies a tilting. Moreover, Christoffersen and Jacobs (2004) argue, based on an empirical analysis, that the N-GARCH model is the specification to use in option pricing applications due to its superior forecasting quality.

There is a close link between GARCH specification and the autocovariance of the absolute log-returns. For basic GARCH it holds that

$$\mathrm{Cov}[|e_t|^2, |e_{t-k}|^2] = \left(\alpha_1 + \frac{\alpha_1^2\beta_1}{1 - 2\alpha_1\beta_1 - \beta_1^2}\right)(\alpha_1 + \beta_1)^{k-1},$$

if $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$ and e_t has a finite fourth moment. Consequently Ding *et al.* (1993) scrutinize the autocorrelation of the absolute log-returns in an empirical study. They modified the exponent α of $|e_t|$ and calculated

$$\mathrm{Cov}[|e_t|^\alpha, |e_{t-k}|^\alpha].$$

According to their empirical findings, the correlation is highest for an exponent of $1 < \alpha < 2$. This has led to the introduction of the power-ARCH model. It generalize GARCH in a way that exponents in the volatility dynamics can differ from 2. Alternatively, it can be viewed as an application of the Box-Cox transform[3]

$$\sigma_t^\alpha = \alpha_0 + \alpha_1\ \sigma_{t-1}^\alpha\ |\epsilon_{t-1}|^\alpha + \beta_1\ \sigma_{t-1}^\alpha.$$

[3] See Box and Cox (1964).

Mittnik *et al.* (2002) apply the power-ARCH model to solve the issues under the stable Paretian assumption. Since the variance of a stable Paretian distributed random variable is generally not finite, GARCH does not lead to stationary time series models.

Nelson (1991) introduced the E-GARCH model as an alternative to Bollerslev's GARCH. This concept applies regression to the natural logarithm of the volatility $\log(\sigma_t^2)$

$$\log(\sigma_t^2) = \alpha_0 + \alpha_1 \ (|\epsilon_{t-1}| + \gamma \ \epsilon_{t-1}) + \beta_1 \ \log(\sigma_{t-1}^2).$$

As a consequence, it allows for the less restrictive feasible set of $\alpha_0, \alpha_1, \beta_1 \in \mathbb{R}$. According to the author, this approach is closer to the definition of an ARMA process than the GARCH model.

Of course there is a variety of GARCH models we do not list here. The interested reader is referred to Bera and Higgins (1993), Duan (1997), and Christoffersen and Jacobs (2004) for a more comprehensive overview of GARCH-like models and empirical studies based on financial time series.

2.4 The ARMA-GARCH model

Resulting from the success of the ARMA and GARCH models, various authors have used the combined ARMA-GARCH approach to analyze time series with conditional mean and volatility. With our notation from paragraph 2.1, a process $(y_t)_t$ is of the ARMA-GARCH type, if for all $t \in T$ $\mathrm{E}[y_t|\mathcal{F}_{t-1}] = \mu_t$ and $\mathrm{V}[y_t|\mathcal{F}_{t-1}] = \sigma_t^2$ with

$$\mu_t = c + \sum_{i=1}^{p} b_i \ y_{t-j} + \sum_{j=1}^{q} a_i \ \sigma_{t-i} \ \epsilon_{t-i}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^2 \ \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2.$$

Under the assumption that the innovation process $(\epsilon_t)_t$ is i.i.d. standard normally distributed, y_t is also normally distributed with $y_t \sim \mathrm{N}(\mu_t, \sigma_t^2)$ conditional on the historical information $\mathcal{F}_{t-1}$. In this case ARMA-GARCH equals a conditional density model in which potentially all parameters are time-varying. The updating $\theta_t|\mathcal{F}_{t-1}$ of a parameter θ_t is implemented by use of a regression approach. The terms $\sigma_{t-i} \ \epsilon_{t-i}$ and $\sigma_{t-i}^2 \ \epsilon_{t-i}^2$ can be interpreted as moment estimators for mean and variance based on a single observation.

Typical properties of such an approach, based on autoregression and moment estimation, are:

- The parameter updating $\theta_t|\mathcal{F}_{t-1}$ is independent of the distributional assumption.
- There is no inter-dependence between different parameter processes μ_t and σ_t.
- The parameter updating is linear in the current and historical estimators.

2.5 Model inference

In this section we review three important concepts of parameter estimation for time series models. First, we will take a look at the ordinary least square (OLS) estimation,[4] then the MLE, and finally, the Quasi-MLE (QMLE) method.

Let us assume a basic regression model of the form

$$y_t = b \cdot x_t + e_t,$$

where x_t is a deterministic factor and $(e_t)_t$ is i.i.d. with mean 0 and variance σ^2. Moreover, we consider the residual sum of squares (RSS) as the loss function

$$RSS = \sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T} (y_t - b\ x_t)^2.$$

The estimator $\hat{b}$ of parameter b is called the ordinary least square (OLS) estimator if it minimizes the RSS

$$\hat{b} = \underset{b}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - b\ x_t)^2.$$

This means that the parameter b is chosen in a way that minimizes the loss function, that is the accumulated squared error. The following calculations show how the OLS estimator can be derived using the first-order optimality

[4] See Hamilton (1994) for a comprehensive view on OLS estimation.

condition

$$\frac{\partial \sum_{t=1}^{T}(y_t - bx_t)^2}{\partial b}\bigg|_{b=\hat{b}} = 2\sum_{t=1}^{T}(y_t - bx_t)\cdot(-x_t) = 0.$$

Solving for $\hat{b}$ yields

$$\hat{b} = \frac{\sum_{t=1}^{T} y_t x_t}{\sum_{t=1}^{T} x_t x_t}.$$

Under some regularity assumptions the OLS estimator is an unbiased minimum-variance (MVU) estimator.[5] The OLS estimator can also be applied for parameter inference in a time series model with autoregression

$$y_t = b \cdot y_{t-1} + e_t.$$

Let $|b| < 1$, that is y_t is stationary, and furthermore e_t an i.i.d. sequence with mean zero, constant variance, and finite fourth moment, then the OLS for b is

$$\hat{b}_T = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}.$$

Although this estimator is generally biased, it can be shown that the distribution $F_T(x)$ of the estimation error compared to the real value b

$$\sqrt{T}(\hat{b}_T - b)$$

asymptotically converges to the normal distribution $\mathrm{N}(0, 1-b^2)$

$$\lim_{T\to\infty} F_T(x) = F^N_{0,1-b^2}(x),$$

which means that the bias vanishes asymptotically. This type of convergence is called *convergence in distribution* and denoted as

$$\sqrt{T}(\hat{b}_T - b) \xrightarrow{L} \mathrm{N}(0, 1-b^2).$$

From Hamilton (1994) we also know that under certain regularity assumptions, e.g. the Gaussian distribution, the OLSE and the MLE are equivalent. The advantage of the OLS method is that it can be applied without a full distributional assumption.

[5] See section 5.1.2 for a definition of MVUE.

For more advanced time series models such as the GARCH model, there exists no OLSE. Thus, the MLE is the preferred method for parameter inference. Since the distribution of y_t is often unknown, we focus on the likelihood function for the error process $e_t = y_t - c$. In the original model from Bollerslev e_t is normally distributed with variance σ_t^2, which yields for the log-likelihood of the GARCH model

$$\begin{aligned} L(\theta) &= \sum_{t=1}^{T} \log f_\theta(y_t) \\ &= -\frac{T}{2}\,\log(2\pi) - \frac{1}{2}\,\sum_{t=1}^{T} \log(\sigma_t^2) - \frac{1}{2}\,\sum_{t=1}^{T} \left(\frac{y_t - c}{\sigma_t}\right)^2, \end{aligned}$$

with the parameter vector $\theta = (c, \alpha_0, \alpha_1, \beta_1)$ and the volatility process

$$\sigma_t^2 = \alpha_0 + \alpha_1 (y_{t-1} - c)^2 + \beta_1 \sigma_{t-1}^2.$$

The iterative formula for the volatility dynamics is

$$\sigma_t^2 = \alpha_0 \sum_{s=1}^{t-1} \beta_1^{s-1} + \alpha_1 \sum_{s=1}^{t-1} \beta_1^{s-1} (y_{t-s} - c)^2 + \beta_1^t \sigma_0^2,$$

where σ_0^2 is the initial volatility value. The optimal parameters are derived using the first-order optimality

$$\frac{\partial L(\theta)}{\partial \theta_i}\bigg|_{\theta_i = \hat{\theta}_i} = 0.$$

This results in the equation systems

$$\begin{aligned} \frac{\partial L(\theta)}{\partial c} &= -\frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \cdot \frac{\partial \sigma_t^2}{\partial c} - \frac{1}{2} \sum_{t=1}^{T} \left[2 \frac{e_t}{\sigma_t^2} - \frac{e_t^2}{\sigma_t^4} \cdot \frac{\partial \sigma_t^2}{\partial c} \right] = 0 \\ \frac{\partial L(\theta)}{\partial \alpha_0} &= -\frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \alpha_0} - \frac{1}{2} \sum_{t=1}^{T} - \frac{e_t^2}{\sigma_t^4} \frac{\partial \sigma_t^2}{\partial \alpha_0} = 0 \\ \frac{\partial L(\theta)}{\partial \alpha_1} &= -\frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \cdot \frac{\partial \sigma_t^2}{\partial \alpha_1} - \frac{1}{2} \sum_{t=1}^{T} - \frac{e_t^2}{\sigma_t^4} \cdot \frac{\partial \sigma_t^2}{\partial \alpha_1} = 0 \\ \frac{\partial L(\theta)}{\partial \beta_1} &= -\frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \cdot \frac{\partial \sigma_t^2}{\partial \beta_1} - \frac{1}{2} \sum_{t=1}^{T} - \frac{e_t^2}{\sigma_t^4} \cdot \frac{\partial \sigma_t^2}{\partial \beta_1} = 0. \end{aligned}$$

The partial derivatives of σ_t^2 can be calculated as

$$\begin{aligned}
\frac{\partial\sigma_t^2}{\partial c} &= -\alpha_1 \cdot \sum_{s=1}^{t-1} 2\beta_1^{s-1} e_{t-s} \\
\frac{\partial\sigma_t^2}{\partial\alpha_0} &= \sum_{s=1}^{t-1} \beta_1^{s-1} \\
\frac{\partial\sigma_t^2}{\partial\alpha_1} &= \sum_{s=1}^{t-1} \beta_1^{s-1} e_{t-s}^2 \\
\frac{\partial\sigma_t^2}{\partial\beta_1} &= \alpha_0 \sum_{s=1}^{t-1} (s-1)\beta_1^{s-2} + \alpha_1 \sum_{s=1}^{t-1} (s-1)\beta_1^{s-2} e_t^2 + t\beta_1^{t-1}\sigma_0^2.
\end{aligned}$$

Before we focus on the non-Gaussian case, we will introduce the term "consistent estimator". Let $(X_T)_T$ be a sequence of random variables. The sequence is said to "converge in probability" to c if for every $\epsilon > 0$ and every $\delta > 0$ there exists a value N such that, for all $T \geq N$

$$P(|X_T - c| > \delta) < \epsilon.$$

The notation for convergence in probability is

$$X_T \xrightarrow{p} c.$$

An estimator $\hat{b}_T$ is called *consistent* if the sequence of estimators $(\hat{b}_T)_T$ converges in probability to the real value b

$$\hat{b}_T \xrightarrow{p} b.$$

Roughly speaking, the probability that the estimator $\hat{b}_T$ assymptotically converges to b is one.

We will now apply the consistency concept to the parameter estimation for GARCH models. From Bollerslev and Wooldridge (1992) we know that Gaussian log-likelihood functions can be used for parameter inference even if the error terms e_t do not follow a Gaussian law. This method is named QMLE and leads to consistent estimators provided that the innovation process $\epsilon_t = \frac{e_t}{\sigma_t}$ satisfies the following standardization conditions

$$\begin{aligned} \mathrm{E}[\epsilon_t] &= 0 \\ \mathrm{V}[\epsilon_t] &= 1. \end{aligned}$$

It can be shown that, under certain regularity conditions, the estimation error $\sqrt{T}(\hat{\theta}_t - \theta)$ is asymptotically normal distributed with probability law[6]

$$\sqrt{T}(\hat{\theta}_t - \theta) \xrightarrow{L} \mathrm{N}(0, \Sigma^2),$$

which means that the bias is asymptotically zero. The variance of the estimator is, however, not optimal unless the innovation process is Gaussian distributed. Nevertheless, QMLE allows for efficient parameter inference even in the non-Gaussian case, which is the basis for, e.g. inference of TS and TID GARCH models.

[6] See Hamilton (1994).

Chapter 3

The MCECD model

The MCECD model combines the conditional density with the minimum cross-entropy. Before presenting the formal definition of our model, we will take a closer look at these concepts. In particular, we would like to outline how conditional density can be interpreted as a generalization of the mean-variance framework. For the minimum cross-entropy we illustrate its link to existing models using a simple example.

3.1 Conditional density

Since Markowitz (1952) published his seminal portfolio selection framework, the mean-variance approach has been at the core of financial analysis. Preferences of market participants are often summarized by the first two moments. The mean represents the expected return, whereas the variance is considered as a measure of risk. The market price of risk in a Black-Scholes model is calculated as the ratio of excess return and volatility

$$\lambda = \frac{\mu - r}{\sigma},$$

where r denotes the risk-free return. This result stems from the underlying Brownian motion, which makes use of the Gaussian assumption. In such a model μ and σ are the only parameters and thus represent the only mean to model market preferences.

Since Mandelbrot (1963) and Fama (1963) a considerable number of empirical studies have documented that the assumption, that return distributions can be characterized by a normal distribution, should be rejected. The findings of these studies suggest that return distributions have heavier tails than the normal distribution (i.e., exhibit leptokurtosis) and have non-zero skewness (i.e., are asymmetric). Today alternative distributions, such as the stable Paretian distribution or the family of tempered stable distributions, fill this gap between theoretical models and empirical findings. As a result, additional parameters are available to describe the market preferences. The introduction of the value-at-risk (VaR) and conditional value-at-risk (CVaR)[1] in risk management has been motivated by the fact that volatility alone is not an appropriate risk measure. Especially the CVaR as the conditional expected value of the tail underlines the necessity to focus on various features of the probability law.

The particular value of conditional density lies in its ability to exploit all information provided by the empirical distribution function. It is equivalent to modeling the moments of all orders simultaneously because the moment generating function $M_X(t)$ and the density $f_X(x)$ contain equivalent information about the probability law. The work of Gallant *et al.* (1991) has already introduced the concept of a conditional density. Hansen (1994) with his ARCD model suggested that potentially all parameters under a specific distributional assumption are conditional on the historical observations. Based on this idea, the changes in market preferences over time are implicitly modeled by the parameter dynamics. Adjusted risk perception measured by e.g. CVaR is thus not only dependent on heteroskedasticity, but also on shifts in the asymmetry of the applied distribution.

3.2 Minimum cross-entropy

In this paragraph we will outline the principles of cross-entropy minimization regarding our application as a parameter updating method. The prerequisite is that we observe a time series $(x_t)_{t\in\mathbb{Z}}$. Each new information x_t should be used to adapt the parameters of our distributional assumption $f_\theta(x)$, also giving weight to the history $\{x_s|s<t\}$. The challenge is that we do not know when θ changes. Our goal is to derive the following functional G based on a cross-entropy approach

$$\theta_t = G(\{x_s|s<t\}).$$

In order to illustrate the method, we restrict ourselves to the case where $f_\theta(x)$ is the Gaussian distribution and the mean $\theta=\mu$ is the only relevant factor. It is commonly known that in the Gaussian case MLE equals MVUE

[1] Also known as expected tail loss (ETL).

for the mean and based on a single observation x it results in $\hat{\mu} = x$. The corresponding SI is $-\log(f_\mu(x))$.

Since we do not know when the mean changes, we assume that the event occurs with a fixed probability $p \in [0, 1]$. In case of a change at time $s-1$, the new parameter equals the mean estimator given the latest observation $\mu_s = x_{s-1}$. At a fixed time t there are only two scenarios possible: the parameter μ_t is either x_{t-1} with probability p or μ_{t-1} with probability $q = 1 - p$. The corresponding probability law, a Bernoulli distribution, is called the scenario distribution. Due to the recursive structure of this dynamics, we can observe that the conditional probability $P(\mu_t = x_s|\{x_s|s < t\})$ that the parameter at time t equals a historic observation x_s with $s < t$ is geometrically distributed

$$P(\mu_t = x_s|\{x_s|s < t\}) = q^{t-s-1} \cdot p.$$

From MLE, ME, and MCE we can derive that given a set of observations $x = (x_1, ..., x_n)$ with $n \in \mathbb{N}$

$$\begin{aligned}\hat{\mu} &= \underset{\mu}{\operatorname{argmax}} \ \sum_{i=1}^{n} \log(f_\mu(x_i)) \\ &= \underset{\mu}{\operatorname{argmin}} - \sum_{i=1}^{n} \log(f_\mu(x_i)) = \underset{\mu}{\operatorname{argmin}} - \sum_{i=1}^{n} \frac{1}{n} \log(f_\mu(x_i)).\end{aligned}$$

The uniform distribution $1/n$ accounts for the fact that in standard MLE there is no information available on whether or not a certain observation is more relevant for the parameter estimate than another one. Consequently the available SI of x_i are equally weighted. The MLE can be rewritten as

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \ E[-\log(f_\mu(X))].$$

Back to our example, the probability with which the parameter μ_t equals a certain observation x_s is described by a geometric distribution. Thus, we can directly apply this probability law to the expected value calculation

$$\begin{aligned}\hat{\mu}_t &= \underset{\mu}{\operatorname{argmin}} \ E[-\log(f_\mu(X))] = \underset{\mu}{\operatorname{argmin}} - \sum_{s=-\infty}^{t-1} q^{t-s-1} \ p \ \log(f_\mu(x_s)) \\ &= G(\{x_s|s < t\}).\end{aligned}$$

This also defines the investigated updating functional. Transforming the summation variable and using the first-order derivative regarding μ yields

$$\sum_{s=1}^{\infty} q^{s-1}\ p\ \frac{x_{t-s}-\mu}{\sigma^2}\Big|_{\mu=\hat{\mu}_t} = 0.$$

After some basic calculation, we derive the following formula for the estimator $\hat{\mu}_t$

$$\hat{\mu}_t = \sum_{s=1}^{\infty} q^{s-1}\ p \cdot x_{t-s},$$

or as a recursive formula

$$\hat{\mu}_t = p \cdot x_{t-1} + q \cdot \hat{\mu}_{t-1}.$$

The parameter dynamics coincide with the exponential smoothing. This example proves that there is a close connection between parameter updating based on minimum cross-entropy and autoregression. The specific form depends on the distributional assumptions concerning the observations and the scenarios. Now we introduce a new time series model based on minimally cross-entropic parameter updating. It generalizes the one presented here in the way that it allows for more complex scenarios.

3.3 The MCECD definition

In this section we will introduce our MCECD model for a financial return series. We assume a stochastic process $\epsilon : T \times \mathbb{R} \to \mathbb{R}$ with natural filtration $\mathcal{F}_t = \mathcal{F}_t^{\epsilon} = \wp(\{\epsilon_s | s \leq t\})$.[2] In our model, the conditional density will only depend on the history of the process $(\epsilon_t)_t$ and hence on its natural filtration. We also assume that the CDF $F_\theta : \mathbb{R} \to [0, 1]$ contains the Gaussian as a special case $\theta = \theta_{Norm}$.

Remark 3.3.1. *We use the notation* (υ, ω_{-i}) *to refer to a vector of the form*

$$(\upsilon, \omega_{-i}) = (\omega_1, ..., \omega_{i-1}, \upsilon, \omega_{i+1}, ..., \omega_m).$$

[2] $\wp(\bullet)$ denotes the σ-algebra.

Definition 3.3.2. ***(MCECD Model)*** *Given a white noise process $(\epsilon_t)_{t\in\mathbb{N}_{>0}}$ with $\epsilon_t \sim N(0,1)$, the CDF $F_\theta : \mathbb{R} \to [0,1]$ with m-dimensional parameter vector $\theta = (\theta_1, ..., \theta_m) \in \Theta$, we can define the return process r_t as a transformed white noise process*

$$r_t = F_{\theta_t}^{-1}(F_{\theta_{Norm}}(\epsilon_t)) \ . \tag{3.1}$$

The time-varying parameters $\theta_t = (\theta_{t,1}, ..., \theta_{t,m})$ can be derived by component, minimizing the m-dimensional cross-entropy process $(H_t(\theta))_{t\in\mathbb{N}_{>0}}$ with $H_t(\theta) = (H_t^1(\theta), ..., H_t^m(\theta))$

$$\theta_{t,i} = \underset{\xi\in\Theta_i}{\operatorname{argmin}} -H_t^i(\xi, \theta_{t,-i}). \tag{3.2}$$

The dynamics of the i-th component ($i \in \{1, ..., m\}$) of the cross-entropy process $(H_t(\theta))_t$ follow the equations

$$\begin{aligned} H_t^i(\theta) &:= \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_{t-1})) + \beta_i \cdot H_{t-1}^i(\theta) \qquad (3.3)\\ H_1^i(\theta) &:= \log(f_\theta(x_{0,i})), \end{aligned}$$

where $\bar{x} = (\bar{x}_1, ..., \bar{x}_m) \in \mathbb{R}^m$ and $x_0 = (x_{0,1}, ..., x_{0,m}) \in \mathbb{R}^m$ are m-dimensional constants, β_i is defined by $\beta_i := 1 - \alpha_0 - \alpha_i$, and the α_i satisfy for all $i \in \{0, ..., m+1\}$

$$\alpha_i \geq 0 \quad \textit{and} \quad \sum_{i=0}^{m+1} \alpha_i = 1. \tag{3.4}$$

The vector $\alpha = (\alpha_0, ..., \alpha_{m+1})$ can be interpreted as a discrete probability measure. α_0 is the probability that the parameters are time-invariant. For $i \in \{1, ..., m\}$, α_i is the likelihood that the current observation r_{t-1} signals a change in parameter i. α_{m+1} stands for the probability that the parameters in t equal the ones in $t-1$. The m-dimensional x_0 determines the starting points of the parameter processes, whereas $\bar{x}$ defines average parameter values associated with the probability α_0. From definition 3.3.2 we see that parameter dynamics in the MCECD are derived from a minimum cross-entropy expression, which is equivalent to a weighted MLE. Since the distributional assumption is used in the cross-entropy term, there is a close link between parameter dynamics and probability law. MLE inherently accounts for dependencies in the parameter structure and that is why we expect an equivalent characteristic for the MCECD model. Later on, we will explicitly analyze the multiple parameter case for time-varying mean and volatility. The subsequent proposition indicates how the recursive definition of the cross-entropy process can be reformulated in an iterative way.

Proposition 3.3.3. *Let $(H_t(\theta))_{t\in\mathbb{N}_{>0}}$ be a general cross-entropy process from definition 3.3.2, then for each $t \in \mathbb{N}_{>1}$ the following iterative formula holds for every component $i \in \{1, ..., m\}$*

$$\begin{aligned} H_t^i(\theta) = {\beta_i}^{t-1} \cdot \log(f_\theta(x_{0,i})) + \sum_{s=1}^{t-1} {\beta_i}^{s-1} \cdot \alpha_0 \log(f_\theta(\bar{x}_i)) \qquad (3.5) \\ + \sum_{s=1}^{t-1} {\beta_i}^{s-1} \cdot \alpha_i \log(f_\theta(r_{t-s})). \end{aligned}$$

Proof. Proof by induction. See appendix A.1. □

The parameter dynamics in equation (3.2) rely only on past observations. This suggests that the parameter vector θ_t is only dependent on the innovations ϵ_s with $s < t$. The following proposition formalizes this statement.

Proposition 3.3.4. *Given the MCECD model from definition 3.3.2 with the innovation process $(\epsilon_t)_{t\in\mathbb{N}_{>0}}$ and its natural filtration $\mathcal{F}_t$, then the cross-entropy process $H_t^i(\theta)$ is predictable, that means $H_t^i(\theta)$ is $\mathcal{F}_{t-1}$-measurable for all $i \in \{1, ..., m\}$ and $\theta \in \Theta$.*

Proof. See appendix A.2. □

The MCECD models from definition 3.3.2 can be applied for arbitrary combinations of time-varying parameters. In order to specify a distinct model, we introduce the following nomenclature:

Remark 3.3.5. *The names of the moments, that are modeled as time-varying by the MCECD model, are used as prefixes. A Vola-MCECD model denotes a model with the volatility parameter as the only time-varying parameter. Analogously, in a Mean-Vola-MCECD, only the parameters corresponding to the first two moments are modeled as time-varying, and in a Skew-MCECD, only the skewness parameter is time-varying.*

In chapter 4 we will define and analyze the Vola-MCECD and Mean-Vola-MCECD models more thoroughly, whereas chapter 5 is dedicated to the Skew-MCECD and Vola-Skew-MCECD models.

3.4 Stationarity

A key feature of models for financial time series is stationarity which claims, briefly speaking, that future returns follow the same distributional law as past returns. Although, in the context of MCECD, the conditional density function is time-dependent, the unconditional probability function is stationary. This origins from the fact that both the innovation process $(\epsilon_t)_{t\in\mathbb{Z}}$ and the parameter process $(\theta_t)_{t\in\mathbb{Z}}$ are stationary. As white noise satisfies this condition by definition, we focus on the $(\theta_t)_{t\in\mathbb{Z}}$ in the remaining part of this section.

Lemma 3.4.1. *Given a return series $(r_t)_{t\in\mathbb{Z}}$, $\beta > 0$ and a PDF $f_\theta(x)$ such that all r_t induce a positive value independent of θ*

$$f_\theta(r_t) > 0,$$

then the weighted geometric series S_∞

$$S_\infty = \sum_{k=0}^{\infty} \beta^k \cdot \log(f_\theta(r_{t-k-1})) \tag{3.6}$$

is absolute convergent, if and only if $\beta < 1$.

Proof. See appendix A.3. □

For $\beta = 0$, the convergence is trivial. Subsequently, we assume that $f_\theta(r_t) > 0$ is always satisfied and hence $\log(f_\theta(r_t))$ is finite.

Based on the convergence property in proposition 3.4.1, we define a MCECD process with infinite history, the unconditional MCECD, analogous to Nelson (1990).

Definition 3.4.2. ***(Unconditional MCECD)*** *Let $F_{\theta_t}(x)$, ϵ_t, r_t, and θ_t be as given in definition 3.3.2, but with infinite history $t \in \mathbb{Z}$. Then the unconditional MCECD is completely specified by the following equation system for its cross-entropy process*

$$_{-\infty}H_t^i(\theta) = \sum_{s=1}^{\infty} {\beta_i}^{s-1} \cdot \alpha_0 \log(f_\theta(\bar{x}_i)) + \sum_{s=1}^{\infty} {\beta_i}^{s-1} \cdot \alpha_i \log(f_\theta(r_{t-s})). \tag{3.7}$$

The results of propositions 3.3.4 and 3.4.1 lead us directly to the following proposition.

Proposition 3.4.3. *Given the unconditional MCECD model from definition 3.4.2, then the resulting m-dimensional parameter process $(\theta_t)_t$ is (strictly) stationary.*

Proof. See appendix A.4. □

From this proposition we can draw the conclusion that the return process generated by MCECD according to equation (3.1) is stationary.

Chapter 4

Conditional volatility

In this chapter we will analyze models for heteroskedasticity based on minimum cross-entropy. Our aim is to show how the MCECD model fits into the existing research in this field. GARCH and ARMA-GARCH models have proven to be very successful regarding time-varying volatility. First, we will focus on the volatility as the only conditional parameter and then we extend our analysis to cover simultaneous modeling of conditional mean and volatility.

4.1 The Vola-MCECD model

4.1.1 Vola-MCECD and GARCH

MCECD generalizes the seminal GARCH framework. In this section we will resort to a special MCECD model, the Vola-MCECD, where $m = 1$ and θ_t is the volatility parameter, and we will show the equivalence of Vola-MCECD and GARCH. Therefore we need the following assumption.

Assumption A1 *The volatility is the only time-varying parameter $\theta_t = \sigma_t$ and the conditional distribution is Gaussian $r_t \sim N(\mu, \sigma_t^2)$ with PDF $f_{\mu,\sigma_t}(x)$.*

The resulting conditional Vola-MCECD model takes the form

$$\begin{aligned}
\sigma_t &= \underset{\sigma}{\operatorname{argmin}} -H_t(\sigma) \\
H_t(\sigma) &= \alpha_0 \cdot \log(f_{\mu,\sigma}(\bar{x})) + \alpha_1 \cdot \log(f_{\mu,\sigma}(r_{t-1})) + \alpha_2 \cdot H_{t-1}(\sigma) \\
H_1(\sigma) &= \log(f_{\mu,\sigma}(x_0)) \ ,
\end{aligned}$$

where $H_t(\sigma) = H_t^1(\sigma)$ and $\bar{x}$ and x_0 are scalars. The unconditional Vola-

MCECD model is defined analogously to definition 3.4.2.

Proposition 4.1.1. *Given a Vola-MCECD model which satisfies assumption A1, then there exists an equivalent GARCH model with specification*

$$\begin{aligned}\sigma_t^2 &= \tilde{\alpha}_0 + \alpha_1 \cdot e_{t-1}^2 + \alpha_2 \cdot \sigma_{t-1}^2 \\ \sigma_1^2 &= \sigma_0^2,\end{aligned}$$

where $(e_t)_{t\in\mathbb{N}_{>0}}$ *with* $e_t := r_t - \mu$ *is the excess return process* $e_t \sim N(0, \sigma_t^2)$ *and* $\tilde{\alpha}_0 := (1 - \alpha_1 - \alpha_2) \cdot \bar{\sigma}^2$*, where* $\bar{\sigma}^2 := (\bar{x} - \mu)^2$ *and* $\sigma_0^2 := (x_0 - \mu)^2$*. Both models govern the same volatility process*

$$\sigma_t^{MCECD} = \sigma_t^{GARCH} \; \forall t \in \mathbb{N}_{>0}$$

Proof. See appendix A.5. □

Remark 4.1.2. *The equivalence of the two models should, of course, also be reflected in equivalent stationarity conditions. From Nelson (1990) we know that the GARCH model is stationary if and only if* $\alpha_1 + \alpha_2 < 1$ *given that* $\tilde{\alpha}_0 > 0$*. For the Vola-MCECD model we know that* $\alpha_0 + \alpha_1 + \alpha_2 = 1$*. From proposition 4.1.1 we can easily see that* $\tilde{\alpha}_0 > 0$ *implies* $\alpha_0 > 0$*. Hence a positive* $\tilde{\alpha}_0$ *leads to* $\alpha_1 + \alpha_2 = 1 - \alpha_0 < 1$*, which is exactly the stationarity condition presented in Nelson (1990).*

Note that our result is based on the Gaussian distribution (see assumption A1). Researchers as well as practitioners, however, use a variety of distributions in order to account for special features of the return data. In section 2.5 we reviewed Bollerslev's QMLE which states that even if the assumption of normality is violated, the normal distribution can be used for inference of GARCH parameters. Given proposition 4.1.1, QMLE is also applicable to the special case of Vola-MCECD. Since one of our objectives is to show that MCECD provides a link between parameter process and distributional assumption, we will nevertheless analyze the non-Gaussian case more thoroughly in the next section.

4.1.2 Non-Gaussian models

Log-returns of financial time series display leptokurtosis and non-zero skewness. One way to account for these features is to use a stable Paretian distribution. Mittnik *et al.* (2002) deal with the stationarity issue of the GARCH model under this specific distributional assumption. They propose a solution within the empirically relevant parameter range using the power-ARCH specification. This highlights one of the drawbacks related to autoregression models: the classical approach does not link the distributional assumption and the parameter dynamics. Instead, GARCH-like models rely on moment estimators.

In order to demonstrate the effects of the distributional assumption in the MCECD model, we consider two distributions which account for both leptokurtosis and non-zero skewness: the skewed exponential power distribution (SEP), a generalization of the exponential power distribution (EP), and the α-stable distribution $S_\alpha(C, \beta, \mu)$. For the SEP, we will derive an explicit Vola-MCECD model and outline how it differs from the Gaussian case. For $S_\alpha(C, \beta, \mu)$, we will explore the induced parameter process by means of numerical analysis, due to the lack of a closed-form expression for its PDF.

In order to compare the MCECD approach to the classical autoregression, we will introduce the term "linear autoregressive" parameter process, which resembles the GARCH concept.

Definition 4.1.3. *Given a parameter process $(\theta_t)_{t\in\mathbb{Z}}$ of the MCECD model from definition 3.3.2. Then the i-th component of the parameter process is called linear autoregressive, if $\theta_{t,i}$ follows the equations*

$$\begin{aligned}\theta_{t,i}^{\gamma} &= \alpha_0 \cdot \bar{\theta}_i^{\gamma} + \alpha_i \cdot g_{\theta_{-i}}(r_{t-1}) + \beta_i \cdot \theta_{t-1,i}^{\gamma} \qquad (4.1)\\ \theta_{1,i}^{\gamma} &= \theta_{0,i}^{\gamma},\end{aligned}$$

where $g_{\theta_{-i}}(x)$ is an estimator for parameter θ_i based on the observation x, γ is a real-valued exponent, and $\bar{\theta}$ and θ_0 are m-dimensional parameter vectors.

A simple induction over time leads us to the iterative formula for a linear autoregressive parameter process

$$\theta_{t,i}^{\gamma} = \beta_i^{t-1} \cdot \theta_{0,i}^{\gamma} + \alpha_0 \sum_{s=1}^{t-1} \beta_i^{s-1} \cdot \bar{\theta}_i^{\gamma} + \alpha_i \sum_{s=1}^{t-1} \beta_i^{s-1} \cdot g_{\theta_{-i}}(r_{t-s}). \qquad (4.2)$$

Our findings in proposition 4.1.1 suggest that the volatility process under Gaussian assumption is linear autoregressive. This raises the question of which feature the underlying distribution must possess so that the corresponding volatility process is linear autoregressive. For our analysis, we have chosen to restrict the set of probability laws to those which satisfy a standardization condition: if $f_\theta(x)$ is a PDF based on a random variable X

with location parameter $\theta_1 = \mu$ and scale parameter $\theta_2 = \sigma$, then for the standardized random variable $\frac{X-\mu}{\sigma}$ with parameter vector θ^{Std} it holds that

$$f_\theta(x) = \frac{1}{\sigma} \cdot f_{\theta^{Std}} \left(\frac{x-\mu}{\sigma} \right) = \frac{1}{\sigma} \cdot f \left(\frac{x-\mu}{\sigma} \right), \tag{4.3}$$

For the cross-entropy process in a Vola-MCECD this implies

$$\begin{aligned} H_t(\sigma) &= \alpha_0 \cdot \log \left[\frac{1}{\sigma} f \left(\frac{\bar{x}-\mu}{\sigma} \right) \right] \\ &\quad + \alpha_1 \cdot \log \left[\frac{1}{\sigma} f \left(\frac{r_{t-1}-\mu}{\sigma} \right) \right] + \alpha_2 \cdot H_{t-1}(\sigma) \\ H_1(\sigma) &= \log \left[\frac{1}{\sigma} f \left(\frac{x_0-\mu}{\sigma} \right) \right]. \end{aligned}$$

Furthermore, the first derivative of the log-density function with respect to the scale parameter σ is

$$\frac{\partial \log \left[\frac{1}{\sigma} f(\frac{x-\mu}{\sigma}) \right]}{\partial \sigma} = -\frac{1}{\sigma} - \frac{f'(\frac{x-\mu}{\sigma})}{f(\frac{x-\mu}{\sigma})} \cdot \frac{x-\mu}{\sigma^2},$$

where $f'(\frac{x-\mu}{\sigma})$ denotes the first derivative of f.

In order to obtain the parameter process, we will take a look at the first-order optimality for the cross-entropy minimization

$$\sigma_t = \underset{\sigma}{\operatorname{argmin}} -H_t(\sigma),$$

given the iterative formula for the cross-entropy

$$\begin{aligned} \frac{\partial H_t(\sigma)}{\partial \sigma}\Big|_{\sigma_t} &= {\alpha_2}^{t-1} \cdot \left(-\frac{1}{\sigma} - \frac{f'(\frac{x_0-\mu}{\sigma})}{f(\frac{x_0-\mu}{\sigma})} \cdot \frac{x_0-\mu}{\sigma^2} \right) \\ &\quad + \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \cdot \alpha_0 \cdot \left(-\frac{1}{\sigma} - \frac{f'(\frac{\bar{x}-\mu}{\sigma})}{f(\frac{\bar{x}-\mu}{\sigma})} \cdot \frac{\bar{x}-\mu}{\sigma^2} \right) \\ &\quad + \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \cdot \alpha_1 \cdot \left(-\frac{1}{\sigma} - \frac{f'(\frac{r_{t-s}-\mu}{\sigma})}{f(\frac{r_{t-s}-\mu}{\sigma})} \cdot \frac{r_{t-s}-\mu}{\sigma^2} \right) \\ &= 0. \end{aligned} \tag{4.4}$$

With this equation, we can formulate a distributional condition for linear autoregressive volatility processes.

Proposition 4.1.4. *Consider a distribution with PDF $f_{\theta_t}(x)$ that satisfies equation (4.3) and is differentiable on $x \in \mathbb{R}\backslash\{\mu\}$. For this distribution, let $(\sigma_t)_{t\in\mathbb{N}_{>0}}$ be the volatility process from a Vola-MCECD model. Then $(\sigma_t)_{t\in\mathbb{N}_{>0}}$ is linear autoregressive if and only if for $x \neq \mu = 0$ it holds that*

$$-\frac{f'_{\theta_t^{Std}}(x)\cdot x}{f_{\theta_t^{Std}}(x)} = k(\operatorname{sign}(x), \theta_t^{Std}) \cdot x^{\gamma}, \tag{4.5}$$

where $k(\operatorname{sign}(x), \theta_t^{Std})$ is a function independent of σ_t and γ is a real-valued exponent.

Proof. The proposition follows directly from equation (4.4) because the equation can be solved with a linear autoregressive form for σ_t as given in equation (4.2), if and only if the ratio $-f'(\frac{x-\mu}{\sigma})/f(\frac{x-\mu}{\sigma})$ is ceteris paribus piecewise proportional to $\left(\frac{x-\mu}{\sigma}\right)^{\gamma-1}$ in the intervals $x<\mu$ and $x>\mu$. □

Remark 4.1.5. *The condition for a linear autoregressive volatility process in equation (4.5) can be reformulated as*

$$\frac{df_{\theta_t^{Std}}(x)/f_{\theta_t^{Std}}(x)}{dx/x} = -k(\operatorname{sign}(x), \theta_t^{Std}) \cdot x^{\gamma}.$$

The term on the left-hand side of the equation resembles the definition of the elasticity. In the following we refer to this ratio as the elasticity of the PDF.

We will now exemplify the rule for linear autoregressive parameter processes by scrutinizing two non-Gaussian distributions: the SEP and the $S_\alpha(\beta, C, \mu)$. For the SEP, we will resort to the characterization by Zhu and Zinde-Walsh (2009). Given the parameters for location $\mu \in \mathbb{R}$, scale $\sigma > 0$, shape $\alpha > 0$, and skewness $\beta \in (0,1)$, the PDF of the SEP is

$$f_{SEP}(x;\alpha,\sigma,\beta,\mu) = \begin{cases} \frac{1}{\sigma}K(\alpha)\exp\left(-\frac{1}{\alpha}\left|\frac{x-\mu}{2\beta\sigma}\right|^{\alpha}\right) & : \quad x \leq \mu \\ \frac{1}{\sigma}K(\alpha)\exp\left(-\frac{1}{\alpha}\left|\frac{x-\mu}{2(1-\beta)\sigma}\right|^{\alpha}\right) & : \quad x > \mu, \end{cases}$$

where $K(\alpha) = [2\alpha^{1/\alpha}\Gamma(1+1/\alpha)]^{-1}$. By definition, the PDF satisfies the standardization condition in equation (4.3). Hence, proposition 4.1.4 applies and we compute the first derivative f' of the PDF with $\mu = 0$ and $\sigma = 1$

$$f'_{SEP}(x;\alpha,1,\beta,0) = \begin{cases} -\frac{K(\alpha)}{(2\beta)^{\alpha}}\exp\left(-\frac{1}{\alpha}\left|\frac{x}{2\beta}\right|^{\alpha}\right)\cdot|x|^{\alpha-1} & : \quad x<0 \\ -\frac{K(\alpha)}{(2(1-\beta))^{\alpha}}\exp\left(-\frac{1}{\alpha}\left|\frac{x}{2(1-\beta)}\right|^{\alpha}\right)\cdot|x|^{\alpha-1} & : \quad x>0. \end{cases}$$

Due to the absolute value function, the PDF is not differentiable at $x = \mu = 0$. The elasticity of the PDF satisfies equation (4.5)

$$\frac{f'_{SEP}(x;\alpha,1,\beta,0)\cdot x}{f_{SEP}(x;\alpha,1,\beta,0)} = \begin{cases} -\frac{1}{(2\beta)^\alpha}\,|x|^\alpha & : \quad x<0 \\ -\frac{1}{(2(1-\beta))^\alpha}\cdot|x|^\alpha & : \quad x>0, \end{cases} \tag{4.6}$$

and that is why the volatility process σ_t is of a linear autoregressive type. To obtain an explicit formula for the volatility process, we insert (4.6) in the first-order optimality in (4.4). Solving for σ_t then yields

$$\begin{aligned} \sigma_t^\alpha = \alpha_2^{t-1}\cdot\sigma_0^\alpha + \alpha_0 \sum_{s=1}^{t-1} \alpha_2^{s-1}\cdot\bar{\sigma}^\alpha \\ + \alpha_1 \sum_{s=1}^{t-1} \alpha_2^{s-1}\cdot k(\operatorname{sign}(r_{t-s}-\mu),\beta)\cdot|r_{t-s}-\mu|^\alpha, \end{aligned} \tag{4.7}$$

with $\sigma_0^\alpha := k(\operatorname{sign}(x_0-\mu),\beta)\cdot|x_0-\mu|^\alpha$ and $\bar{\sigma}^\alpha := k(\operatorname{sign}(\bar{x}-\mu),\beta)\cdot|\bar{x}-\mu|^\alpha$. Note that the special case $x=\mu$ is also covered in this formula. Apart from the different exponent compared to the classical GARCH model, the equation contains a scaling term for the variance estimator

$$k(\operatorname{sign}(x-\mu),\beta) := \begin{cases} (2\cdot\beta)^{-\alpha} & : \quad x<\mu \\ (2\cdot(1-\beta))^{-\alpha} & : \quad x>\mu. \end{cases}$$

The value of $k(\operatorname{sign}(x-\mu),\beta)$ at $x=\mu$ can be arbitrary because $|x-\mu|^\alpha = 0$. For the SEP based Vola-MCECD, the volatility effect (change in conditional volatility caused by the latest observation r_{t-1}) depends on the skewness β of the underlying distribution. For example, $\beta > 0.5$ implies a negative skewness and the impact of a positive excess return $e_t = r_t - \mu > 0$ on the volatility is higher compared to $e_t < 0$. These characteristics directly stem from the ML inference with a skewed distribution. As the probability mass is not spread symmetrically around the mean, the ML variance estimators also differ with the sign of the excess return.

Consequently, the skewness of a distribution has an inverted, yet much smaller impact on the volatility estimator than the empirically observed leverage effect. In order to enable our model to reproduce this empirical finding, we can modify the cross-entropy scenarios. For example, using the adjusted observation $\tilde{r}_t := r_t - \delta$ ($\delta \in \mathbb{R}$) for the cross-entropy process, the Vola-MCECD—analogously to the N-GARCH—can account for the leverage effect.

A look at the volatility formula (4.7) for the SEP driven Vola-MCECD model reveals its close relation to the power-ARCH model proposed by Ding *et al.* (1993) and applied by Mittnik *et al.* (2002) in the stable Paretian case. In fact, for zero-skewness ($\beta = 0.5$) we obtain the exact power-ARCH dynamics. Therefore, the question arises as to whether a Vola-MCECD model based on a stable Paretian distribution yields the same parameter process as in (4.7).

The stable Paretian distribution is defined by its characteristic function $\phi(t; \alpha, \beta, C, \mu)$ as we have outlined in section 1.2. Although stable Paretian distributions have, in general, infinite variance, we can model the dispersion of the distribution by its scale parameter C. In fact, $\sqrt{2C}$ equals σ if $\alpha = 2$ (the Gaussian case). For our following argumentation, we will use dispersion and volatility process as synonyms. It is common knowledge that the stable Paretian PDF $f(x; \alpha, \beta, C, \mu)$ fulfills the standardization condition in (4.3), but does not have a closed-form expression. In order to apply proposition 4.1.4, we need to analyze the elasticity of the PDF numerically. The log-transformation of equation (4.5) yields

$$\log \left| \frac{f'_{\theta_t^{Std}}(x) \cdot x}{f_{\theta_t^{Std}}(x)} \right| = \log \left| k(\text{sign}(x), \theta_t^{Std}) \right| + \gamma \cdot \log |x| \, .$$

If the log-elasticity of the PDF is a linear function of $\log(x)$, we know that the parameter process is linear autoregressive.

The log-elasticity of a stable Paretian distribution is non-linear in $\log(x)$ except for the Gaussian case $\alpha = 2$ as shown in figure 4.1. Therefore the dispersion parameter process of a stable Paretian driven Vola-MCECD model is not linear autoregressive and hence the power-ARCH model does not accurately describe the volatility process for a stable Paretian model.

Applying the non-Gaussian assumption to the volatility dynamics, we can make three key observations. First, if the MLE inference is applicable for the assumed probability law, then parameter processes exist and are uniquely defined. Second, in the MCECD approach inter-dependences between parameters are model-inherent. This also emphasizes the need to specify all parameters correctly; for example, to estimate the volatility in the SEP driven MCECD, one needs a good estimator for skewness. Third, optimal MCECD parameter processes, even for volatility, can be non-linear,

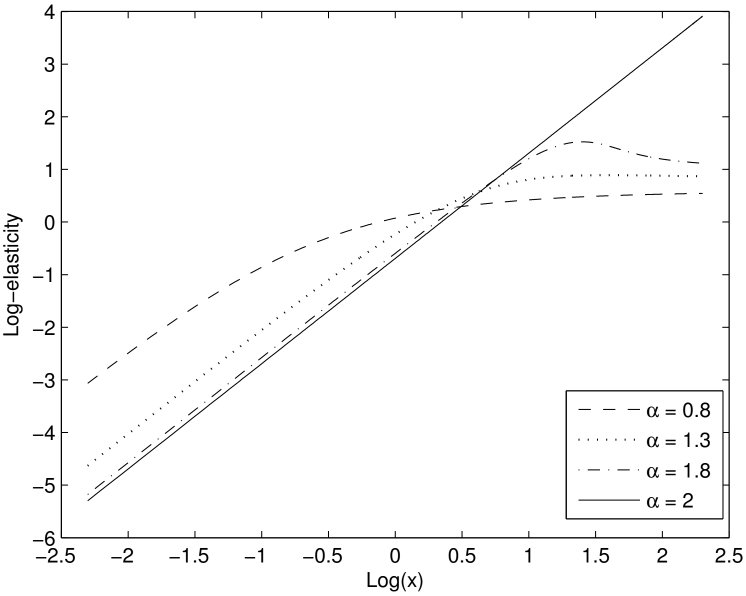

Figure 4.1: Log-elasticity of a stable Paretian distribution with parameters $\beta = 0$, $C = 1$, and $\mu = 0$.

as shown in the stable Paretian case.

4.2 The Mean-Vola-MCECD model

Optimal parameter trajectories are, in general, dependent on each other. That is why in this section we will analyze a model with conditional mean and volatility. Consistent with our nomenclature, the corresponding model is called Mean-Vola-MCECD. To guarantee traceability, we will assume:

Assumption A2 *The mean and the volatility are the only time-varying parameters $\theta_t = (\mu_t, \sigma_t)$ and the conditional distribution is Gaussian $r_t \sim N(\mu_t, \sigma_t^2)$.*

The resulting conditional Mean-Vola-MCECD model is of the form

$$\mu_t = \underset{\xi}{\operatorname{argmin}} -H_t^1(\xi, \sigma_t)$$
$$\sigma_t = \underset{\xi}{\operatorname{argmin}} -H_t^2(\mu_t, \xi),$$

with the cross-entropy processes for the mean component

$$\begin{aligned}
&H_t^1(\mu,\sigma) = \alpha_0 \cdot \log(f_{\mu,\sigma}(\bar{x}_1)) + \alpha_1 \cdot \log(f_{\mu,\sigma}(r_{t-1})) + (\alpha_2+\alpha_3) \cdot H_{t-1}^1(\mu,\sigma)\\
&H_1^1(\mu,\sigma) = \log(f_{\mu,\sigma}(x_{0,1}))
\end{aligned}$$

and for the volatility component

$$\begin{aligned}
&H_t^2(\mu,\sigma) = \alpha_0 \cdot \log(f_{\mu,\sigma}(\bar{x}_2)) + \alpha_2 \cdot \log(f_{\mu,\sigma}(r_{t-1})) + (\alpha_1+\alpha_3) \cdot H_{t-1}^2(\mu,\sigma)\\
&H_1^2(\mu,\sigma) = \log(f_{\mu,\sigma}(x_{0,2})).
\end{aligned}$$

$\bar{x}$ and x_0 are two-dimensional vectors. The unconditional Mean-Vola-MCECD model can be specified analogously to definition 3.4.2. For the Gaussian case, closed-form expressions of parameter dynamics θ_t are available.

Proposition 4.2.1. *Given a MCECD model which satisfies assumption A2, then the mean process μ_t follows the equations*

$$\begin{aligned}
\mu_t &= \alpha_0 \cdot \bar{x}_1 + \alpha_1 \cdot r_{t-1} + (\alpha_2+\alpha_3) \cdot \mu_{t-1}\\
\mu_1 &= x_{0,1},
\end{aligned}$$

and the volatility process σ_t follows

$$\begin{aligned}
\sigma_t^2(\mu_t) &= \alpha_0 \cdot (\bar{x}_2-\mu_t)^2 + \alpha_2 \cdot (r_{t-1}-\mu_t)^2 + (\alpha_1+\alpha_3) \cdot \sigma_{t-1}^2\\
\sigma_1^2(\mu_t) &= (x_{0,2}-\mu_t)^2.
\end{aligned}$$

Proof. See appendix A.6. □

In the Mean-Vola-MCECD model, the volatility dynamics given in proposition 4.2.1 are dependent on the estimator of the mean. The model inherently accounts for inter-dependencies in the parameter structure, even when multiple parameters are time-varying. The empirical results in the next section also emphasize the strength of Mean-Vola-MCECD when analyzing the trajectories of parameter processes.

4.3 Simulation and empirical results

This section deals with an empirical comparison of Mean-Vola-MCECD and its autoregression-based alternative, the ARMA-GARCH process. The dynamics of ARMA-GARCH are given by

$$\begin{aligned} r_t &= a \cdot e_{t-1} + b \cdot r_{t-1} + c + e_t \\ r_1 &= c, \end{aligned} \tag{4.8}$$

where $e_t = \sigma_t \cdot \epsilon_t$ and σ_t is modeled by standard GARCH

$$\begin{aligned} \sigma_t^2 &= \alpha_0 + \alpha_1 \cdot e_{t-1}^2 + \beta_1 \cdot \sigma_{t-1}^2 \\ \sigma_1^2 &= \sigma_0^2. \end{aligned}$$

We will analyze the models along three dimensions: (1) simultaneous modeling of time-varying mean and volatility, (2) distinguishing time-varying from time-invariant trajectories and (3) forecasting properties. Concerning goodness-of-fit, we will apply the Kolmogorov-Smirnov (KS) test, the Anderson-Darling (AD) statistic, and the Cramér-van Mises (CvM) statistic. They measure general fit (KS, CvM) and tail fit (AD, AD^2) as well as the biggest distance (KS, AD) and average distance (AD^2, CvM). For inference, we will use Bollerslev's QMLE method, whereby the innovation process is governed by the Koponen distribution (see section 1.2).

In order to test the modeling of conditional moments, we employ simulated Gaussian log-returns with time-varying mean and volatility. For the remaining analyses we will resort to daily log-returns of U.S. stock indices and several individual U.S. stocks from the Dow Jones. For the goodness-of-fit tests, the different time windows always end at 06/25/2009. This means that a 10-year time span starts at 06/26/1999 and ends at 06/25/2009, an 8-year time span starts at 06/26/2001 and ends at 06/25/2009, and so on. Backtesting is performed based on log-returns between 06/26/2008 and 06/24/2009, using a shifting time window of 9 years of historical data for model calibration. The time window has been chosen in such a way that it includes the Dotcom Collapse in April 2000 and the U.S. financial crisis that began in September 2008.

4.3.1 Time-varying mean and volatility

We generate a conditional density process $(r_t)_{t\in\mathbb{N}_{>0}}$ based on the Gaussian distribution $r_t \sim \mathrm{N}(\mu_t, \sigma_t^2)$, where mean and volatility are time-varying. The parameter processes are independent of r_t, but instead derived from two uniformly distributed processes $(p_t^\mu)_{t\in\mathbb{N}_{>0}}$ and $(p_t^\sigma)_{t\in\mathbb{N}_{>0}}$ using the following

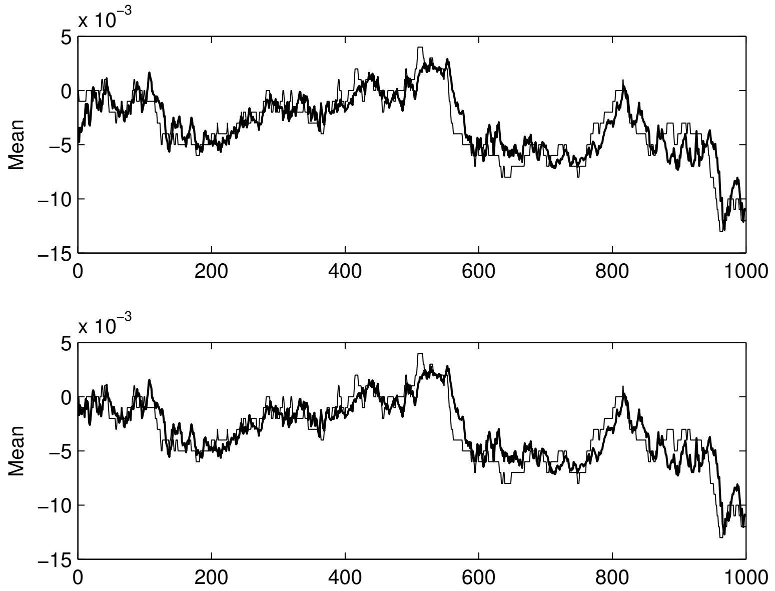

Figure 4.2: Conditional mean trajectories of simulated data compared to corresponding trajectories (bold lines) of Mean-Vola-MCECD (top chart) and ARMA-GARCH (bottom chart) based on Gaussian distribution

specifications

$$\mu_t = \begin{cases} \mu_{t-1} + 0.001 & : \quad p^{\mu}_{t-1} \geq 0.9 \\ \mu_{t-1} - 0.001 & : \quad p^{\mu}_{t-1} \leq 0.1 \end{cases}$$

$$\sigma_t = \begin{cases} \sigma_{t-1} \cdot 1.08 & : \quad p^{\sigma}_{t-1} \geq 0.75 \\ \sigma_{t-1} \cdot 0.925 & : \quad p^{\sigma}_{t-1} \leq 0.25. \end{cases}$$

	KS test	p-value	AD	AD^2	CvM
Mean-Vola-MCECD	0	0.98651	0.08704	0.2751	0.02900
ARMA-GARCH	0	0.95672	0.10329	0.3571	0.03731

Table 4.1: Goodness-of-fit results for Mean-Vola-MCECD and ARMA-GARCH model on simulated data

Figures 4.2 and 4.3 show that both models, ARMA-GARCH and Mean-Vola-MCECD are suitable for modeling time series with conditional mean

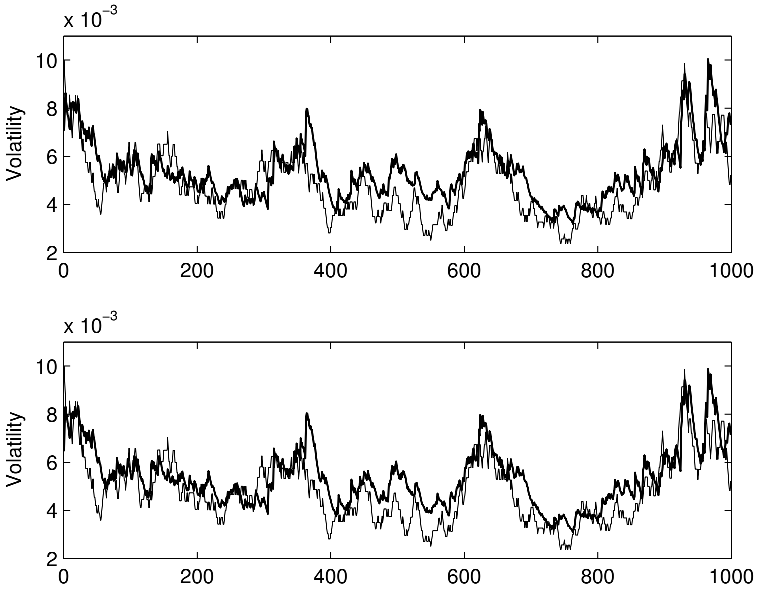

Figure 4.3: Conditional volatility trajectories of simulated data compared to corresponding trajectories (bold lines) of Mean-Vola-MCECD (top chart) and ARMA-GARCH (bottom chart) based on Gaussian distribution

and conditional volatility. Their approximation quality for the parameter trajectories is similar. This finding is supported by the goodness-of-fit analysis shown in table 4.1. Both models yield an equivalent overall as well as tail fit.

4.3.2 The time-varying property

In this paragraph we examine Mean-Vola-MCECD and ARMA-GARCH models when applied to empirical stock index returns. Although, in general, both models can cope with time-varying mean and volatility, the parameter estimates for Mean-Vola-MCECD from table B.1 suggest that the mean of the S&P 500 index log-returns is time-invariant and positive. This result contrasts with the ARMA-GARCH estimates in table B.1. The ARMA parameters clearly suggest a time-varying component in the mean. Figure 4.4 illustrates the time-dependency of the conditional mean. In figure 4.5, we can see the relative deviation ρ of the conditional volatility trajectories

$$\rho = \frac{\sigma_t^{GARCH} - \sigma_t^{MCECD}}{\sigma_t^{GARCH}}.$$

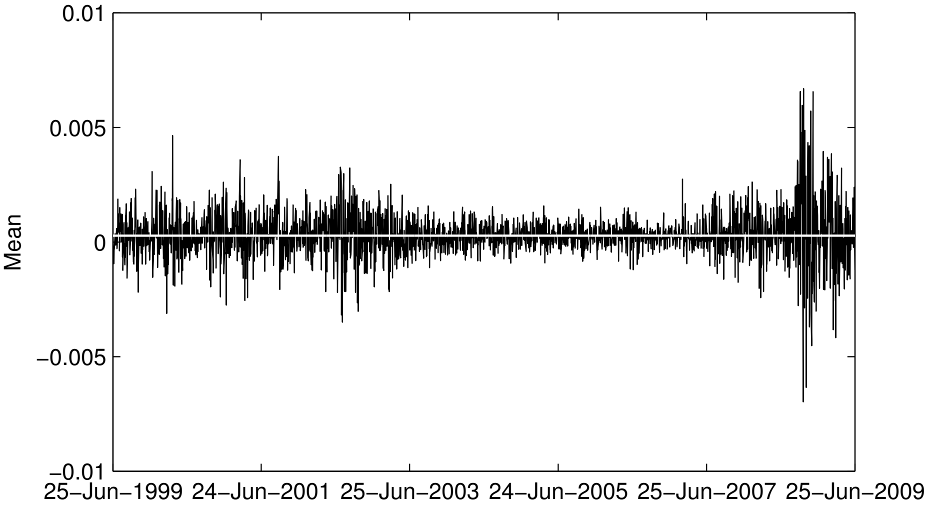

Figure 4.4: Conditional mean trajectories of ARMA-GARCH (black) and Mean-Vola-MCECD (white) for 10 years daily log-return data of S&P 500 index based on Koponen distribution

Based on the above, we conclude that the volatility estimators are equivalent for the S&P 500 index data.

Furthermore, the goodness-of-fit results in table B.4 speak in favor of the Mean-Vola-MCECD model, hence the ARMA-GARCH results might be misleading when it comes to time-invariant mean. Another way to see this is to look at the performance of a pure GARCH model with non-zero mean. Since the GARCH model also yields a better fit, we deduce that the data are characterized by a time-invariant mean. The Mean-Vola-MCECD indicates whether a parameter process is time-invariant or not. Therefore, it might be the preferred choice to obtain reliable parameter trajectories for the conditional density.

Tables B.2 and B.3 for parameter estimates as well as B.5 and B.6 for goodness-of-fit results suggest that our findings for the S&P 500 index data are also valid for the Dow Jones and Nasdaq 100 indices.

4.3.3 Quality of one-day forecasting

In a first step, we will resort to classical VaR backtesting in order to evaluate the one-day forecasting quality of both models. We will apply the

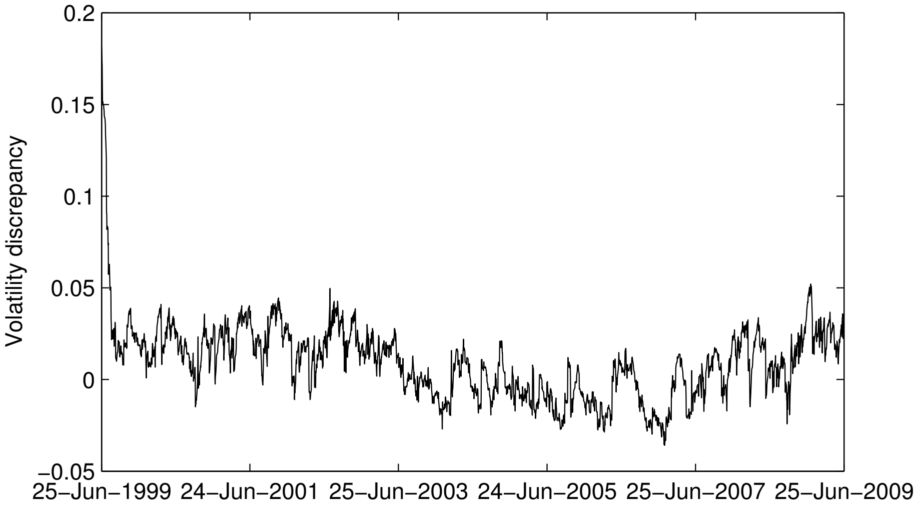

Figure 4.5: Relative deviation of conditional volatility trajectories of ARMA-GARCH and Mean-Vola-MCECD for 10 years daily log-return data of S&P 500 index based on Koponen distribution

Kupiec test[1] and the Lopez statistic[2] for confidence levels 0.01 and 0.05. Both statistics focus on the left tail of the return distribution. The Kupiec statistic measures the frequency of exceeding over the specified quantile, whereas the Lopez statistics considers even the distance to the quantile.[3]

According to the results in table 4.2, there is no statistical evidence for an improved forecasting quality of Mean-Vola-MCECD. The strength of Mean-Vola-MCECD is to model multiple parameters and hence the whole CDF more accurately. VaR, however, evaluates only one point of the distribution. In order to judge the out-of-sample goodness-of-fit for the conditional CDF, we need a holistic approach. Under the distributional assumption $F_\theta(x)$, we can define for the log-return process $(r_t)_t$ and the derived parameter process $(\theta_t)_t$

$$y_t := F_{\theta_t}(r_t). \tag{4.9}$$

If F_{θ_t} describes the log-return distribution over time, then y_t is uniformly distributed. Hence the forecasting quality for the conditional CDF can be

[1] See Kupiec (1995).

[2] See Lopez (1998).

[3] See Chernobai *et al.* (2007) for a comprehensive view on VaR backtesting.

		0.01 quantile		0.05 quantile	
Data	Model	Kupiec	Lopez	Kupiec	Lopez
S&P 500	MCECD	3	5.353	19	32.921
	ARMA-GARCH	3	5.311	20	33.666
	GARCH	2	4.146	19	31.640
Dow Jones	MCECD	1	1.2646	24	31.8987
	ARMA-GARCH	1	1.2469	24	32.0214
	GARCH	1	1.2057	23	30.0478
Nasdaq 100	MCECD	4	11.5732	17	34.0926
	ARMA-GARCH	3	10.9023	19	36.5957
	GARCH	4	11.5892	18	35.1163

Table 4.2: One-day VaR backtesting results for U.S. stock indices from 06/26/2008 to 06/24/2009 based on 0.01 and 0.05 confidence levels using a shifting time window for parameter inference

assessed by analyzing the empirical distribution of y_t.

Table B.7 suggests that for the three stock indices investigated, Mean-Vola-MCECD leads to a better approximation of forecasted CDFs. The difference is even more pronounced for the three individual U.S. stocks. Hence, Mean-Vola-MCECD is a more suitable approach for conditional CDF forecasting, yielding both a better tail and overall fit compared to ARMA-GARCH. For application in portfolio and risk management, we expect Mean-Vola-MCECD to lead to better backtesting results, when more advanced criteria such as the expected tail loss (ETL) or spectral risk measures are applied.

Chapter 5

Conditional skewness

One of the well established stylized facts in financial modeling states that the empirical distributions of log-return time series are skewed and leptokurtic. Therefore there are various models which have been introduced that resort to an alternative non-Gaussian assumption. With the help of the stable Paretian distribution and the classes of tempered stable and tempered infinitely divisible distributions, it is possible to achieve a sufficient goodness-of-fit when applied to financial time series models.

Gallant *et al.* (1991) and Hansen (1994) have promoted the idea of conditional density. This principle not only relates to time-varying mean and volatility but also the effects that skewness and kurtosis might depend on the conditioning information. Given a specific likelihood model, each parameter has the potential to evolve in time. Most risk and performance measures such as the VaR, CVaR, or the STARR ratio strongly depend on the left-tail of the return distribution. Since the skewness defines the asymmetry of the distribution, it has a significant impact on the shape of the tails. As a result, it is crucial to model skewness as accurately as possible also considering changes over time. Chen *et al.* (2001) have already stressed the importance of conditional skewness for market crash prediction. In order to derive their results, they applied a factor model using trading volumes. In the following analysis we will focus on conditional skewness of financial log-returns based on time series models where the only conditioning information available are historical observations.

5.1 Skewness models

Before we introduce our model for conditional skewness, we will take a closer look at the properties of the skewness. Given a random variable X with CDF $F_\theta(x)$, it is commonly known that mean and volatility can be adjusted by rescaling and shifting X

$$Y = a \cdot X + b.$$

In this case the new mean is $E[Y] = E[X] + b$ and the new variance is $V[Y] = a^2 \cdot V[X]$. Mean and variance can therefore be characterized as linear properties of a random variable. See figures 5.1 and 5.2 for an illustration of this conclusion.

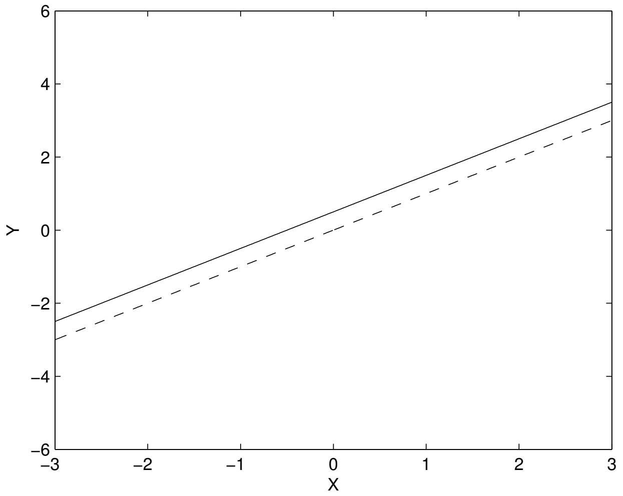

Figure 5.1: Effects of a mean transformation on a random variable for parameter $b = 0.5$

Changing the skewness of a random variable is a non-linear transformation as we can see from figure 5.3. Its functional form is highly dependent on the probability law of X. Using the CDF, a skewness transform can be defined by

$$Y = F_{\theta^*}(F_\theta^{-1}(X)),$$

where θ^* and θ are the parameter sets inducing the new and the old skewness values *ceteris paribus*.

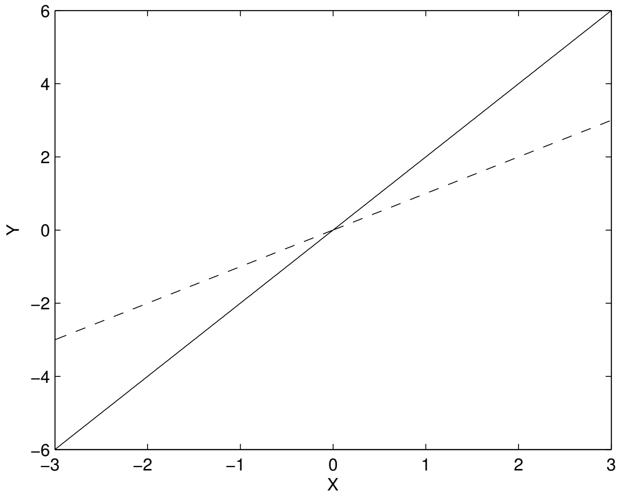

Figure 5.2: Effects of a variance transformation on a random variable for parameter $a = 2$

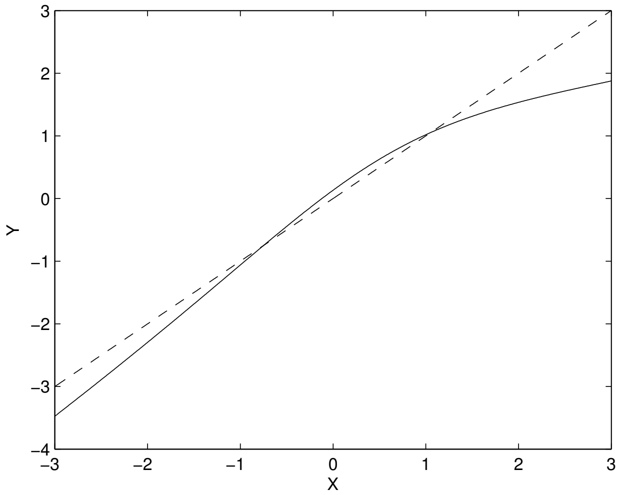

Figure 5.3: Effects of a skewness transformation on a standard Koponen ($\alpha = 0.5$, $\beta = 0$, $\lambda = 1.7$) distributed random variable for parameter $\beta^* = -0.5$

5.1.1 The ARCD model

Based on the concept of time-varying parameters, Hansen (1994) suggested the autoregressive conditional density (ARCD) model as a generalization of Engle's ARCH model. In his approach the skewed Student's t distribution is the likelihood model for the log-returns. The parameter processes are derived from separate regressions for each variable based on the excess return e_t and the squared excess return e_t^2. In order to account for the bounded parameter range of the shape parameter, his model applies a logistic transformation to the regression result. His goal was to keep the parameter dynamics independent of the distributional assumption. This results, however, in the drawback that the underlying regression and the logistic transformation seem to be arbitrary and might not coincide with the parameter logic given by the probability law.

Harvey and Siddique (1999) proposed the "GARCH with skewness" (GARCHS) which models time-varying moments based on autoregressive equations. e_t^2 and e_t^3 are used as the volatility and skewness estimator respectively. This leads to the following equations for the conditional volatility σ_t and skewness s_t

$$\begin{aligned} \sigma_t^2 &= \beta_0 + \beta_1 \sigma_t^2 + \beta_2 e_t^2 \\ s_t &= \gamma_0 + \gamma_1 s_t + \gamma_2 e_t^3. \end{aligned}$$

Moreover, they select the non-central t distribution as the likelihood model. The moments of this distribution can be expressed as functions of the distributional variables. Thus, it is possible to derive the parameter dynamics directly from the conditional moments. The advantage of this method is that parameter process and distribution are closely tied, and follow the same logic. Apart from that, the fact that the skewness estimator is based on e_t^3 reflects the connection between skewness and the third central moment. Since Harvey and Siddique (1999) also pursue the concept of conditional parameters, GARCHS can be interpreted as a conditional density model. Subsequently, we use the term "ARCD" for all approaches modeling time-varying parameters with the help of autoregressive parameter dynamics. Hence ARCD also includes the GARCHS model.

In order to compare the ARCD approaches to our MCECD model for skewness, we resort to the autoregressive models discussed in Dark (2010). For both models, we assume that the innovation process follows an adjusted

Koponen distribution because it has a dedicated skewness parameter β which means that mean, variance and kurtosis are independent of β

$$\begin{aligned} \mathrm{E}[X] &= m \\ \mathrm{V}[X] &= \Gamma(2-\alpha)\cdot C\lambda^{\alpha-2} \\ \mathrm{S}[X] &= \frac{\Gamma(3-\alpha)\cdot C\lambda^{\alpha-3}\beta}{\mathrm{V}[X]^{3/2}} \\ \mathrm{K}[X] &= \frac{\Gamma(4-\alpha)\cdot C\lambda^{\alpha-4}}{\mathrm{V}[X]^{2}}+3. \end{aligned} \tag{5.1}$$

We define the moment dynamics as in Harvey and Siddique (1999) and derive the conditional parameters using the equations (5.1). Given the standardized excess return $\epsilon_t = e_t/\sigma_t$ with $\mathrm{V}[\epsilon_t] = 1$ the conditional β_t follows

$$\beta_t = \frac{\mathrm{S}[\epsilon_t]}{\Gamma(3-\alpha)\ C\ \lambda^{\alpha-3}},$$

where the conditional skewness s_t is used as the estimator for $\mathrm{S}[\epsilon_t]$ and the parameter C is defined by the condition $\mathrm{V}[\epsilon_t] = 1$

$$C = \frac{\mathrm{V}[X]}{\Gamma(2-\alpha)\cdot\lambda^{\alpha-2}}.$$

Consequently the ARCD model can be described by

$$r_t = \mu_t + \sigma_t \cdot F_{\beta_t}(F^{-1}_{\theta_{Norm}}(\epsilon_t)), \tag{5.2}$$

where $(\epsilon_t)_t$ is white noise, $F^{-1}_{\theta_{Norm}}(x)$ is the inverse of the Gaussian CDF, and $F_\beta(x)$ is the standard Koponen CDF. The parameters dynamics follow an autoregressive approach

$$\begin{aligned} \mu_t &= \alpha_0 + \alpha_1\ r_{t-1} + \alpha_2\ \mu_{t-1} \\ \sigma_t^2 &= \beta_0 + \beta_1\ (r_{t-1}-\mu_{t-1})^2 + \beta_2\ \sigma_{t-1}^2 \\ s_t &= \gamma_0 + \gamma_1\ \left(\frac{r_{t-1}-\mu_{t-1}}{\sigma_{t-1}}\right)^3 + \gamma_3\ s_{t-1}. \end{aligned} \tag{5.3}$$

Equations for higher moments such as conditional kurtosis k_t can be defined analogously.

5.1.2 Skewness estimation

In statistics there are several methods to infer the distributional parameters of a random variable X from a finite data sample $(x_1, x_2, ..., x_n)$. Out of these we consider the method of moments (MM) and the estimating function (EF) approach.[1] Both methods differ in the definition of "optimal" estimation due to a different assessment of the estimation quality, that is the loss function.

In MM the parameters are inferred by a comparison of the distributional moments with the sample moments. An optimal moment estimator is provided by the minimum variance unbiased estimator (MVUE). Given a function of the sample data $\delta(x_1, ..., x_n) = \hat{m}$, the parameter vector is derived by solving the following equation for θ

$$\hat{m} = m(\theta).$$

$\delta(x_1, ..., x_n)$ is unbiased if its expected value equals the true parameter θ or, in other words, in average the error of the estimator is zero

$$\mathrm{E}[\delta(x_1, ..., x_n)] = \mathrm{E}[\hat{m}] = m.$$

The minimum-variance property is satisfied if

$$\mathrm{V}[\delta(x_1, ..., x_n)] \leq \mathrm{V}[\delta^*(x_1, ..., x_n)],$$

for all unbiased estimators δ^*. The idea behind the MVUE concept is to obtain estimators that yield in average the true value and at the same time vary only minimally around this average. For the mean and the variance of a random variable the MVUEs are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

It is, however, important to note that the structurally related estimators for

[1] We refer to Bera and Bilias (2002) for a discussion and a synthesis of the different estimation techniques.

skewness and kurtosis

$$\begin{aligned}\hat{s} &= \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(X_i-\hat{\mu})^3}{\hat{\sigma}^3}\\ \hat{k} &= \frac{(n+1)\ n}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(X_i-\hat{\mu})^4}{\hat{\sigma}^4}.\end{aligned}$$

are in general biased although the estimators for the third and fourth central moment have no bias. That is why $\hat{s}$ and $\hat{k}$ are not MVUEs. As a matter of fact there are no MVUEs for skewness or kurtosis available. Kim and White (2004) have been analyzing alternative moment estimators and conclude that there are no generally optimal estimators for skewness or kurtosis available.

The second concept, the EF, goes back to the works of Durbin (1960) and Godambe (1960). Instead of calculating sample moments first, the method uses the so-called estimating function $g(x, \theta)$ based on sample data and parameters. The parameter vector θ is directly derived by solving the equation

$$g(x, \theta) = 0.$$

Optimal estimation is no longer defined for the estimator, but for the EF. As a consequence, concepts such as unbiasedness or minimum-variance are applied to EF. $g(x, \theta)$ is thus unbiased if

$$\mathrm{E}[g(x, \theta)] = 0.$$

Instead of the minimum-variance condition, Godambe (1960) suggests applying the efficiency criterion

$$\mathrm{V}\left[\frac{g}{\partial g/\partial\theta}\right] = \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher information on the PDF $f(x, \theta)$ defined by

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log(f(x, \theta))\right)^2\right].$$

The condition can be reformulated as the minimum-variance of a standardized g

$$\mathrm{V}\left[\frac{g}{\partial g/\partial\theta}\right] \leq \mathrm{V}\left[\frac{g^*}{\partial g^*/\partial\theta}\right].$$

The standardization serves two goals: The variance $\mathrm{V}[g]$ should be as small as possible and at the same time a deviation from the true parameter $\frac{\partial g}{\partial\theta}$ should lead as far away from zero as possible in order to yield a good discriminatory power. In this sense the efficiency criterion is more restrictive than the minimum-variance. Godambe (1960) concluded that the first derivative of the PDF $\frac{\partial}{\partial\theta}f(x,\theta)$ is the optimal estimating function which translates into the optimality of the MLE approach. In case there exists a dedicated skewness θ_s or kurtosis θ_k parameter, these can be derived by solving

$$\frac{\partial}{\partial\theta}f(x,\theta) = 0.$$

In order to illustrate the difference between the MM and the EF approach for skewness estimation, we compare the parameter estimators based on the standard Koponen distribution. In time series analysis, there is only one observation for one time period available. Thus, the conditional skewness is based on one data point only. We will use this prerequisite throughout the following analysis.

Given the skewness of a standard Koponen distributed random variable

$$\mathrm{S}[X] = \Gamma(3-\alpha)\ C\ \lambda^{\alpha-3}\beta = \frac{\Gamma(3-\alpha)\ \lambda^{\alpha-3}\beta}{\Gamma(2-\alpha)\ \lambda^{\alpha-2}} = \frac{2-\alpha}{\lambda}\beta,$$

and the skewness estimator X^3 for a single observation, the MM method yields for β_{MM}

$$\beta_{MM} = \frac{\lambda}{2-\alpha}X^3.$$

Due to the restriction $\beta_{MM} \in [-1,1]$, we define

$$\beta_{MM} = \min\left\{\max\left\{\frac{\lambda}{2-\alpha}X^3, -1\right\}, 1\right\}.$$

The ML estimator for β_{ML} is derived numerically from

$$\frac{\partial \log(f_\beta(X))}{\partial\beta}\bigg|_{\beta=\beta_{ML}} = 0.$$

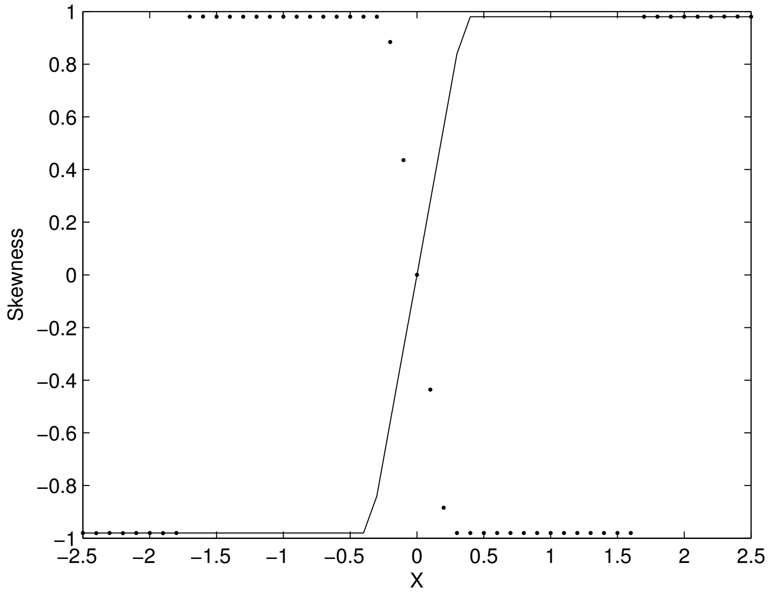

Figure 5.4: Comparison of skewness estimators using MM (line) and MLE (dots) for the standard Koponen distribution ($\alpha = 1.5$, $\lambda = 1.4$) based on a single observation

Figures 5.4 and 5.5 support the hypothesis that the MM method is not appropriate for conditional skewness. The MM estimator has the opposite sign of the ML estimator which satisfies the efficiency and unbiasedness criteria for EFs. Figure 5.5 also shows that this result is independent of the assumed probability law. The differing shape of the SEP based estimator is due to the fact that β_{SEP} drives both skewness and kurtosis. To ensure comparability of the parameter trajectories, we transform the β variable of the SEP distribution in a way that its feasible set is $(-1, 1)$ instead of $(0, 1)$ and that -1 indicates negative skewness and 1 positive skewness. The corresponding equation is hence

$$\beta := 1 - 2 \cdot \beta^*, \tag{5.4}$$

where β^* is the original parameter from the PDF given in equation (1.1). Note that for both the Koponen and SEP case we assume $\mathrm{E}[X] = 0$ and $\mathrm{V}[X] = 1$.

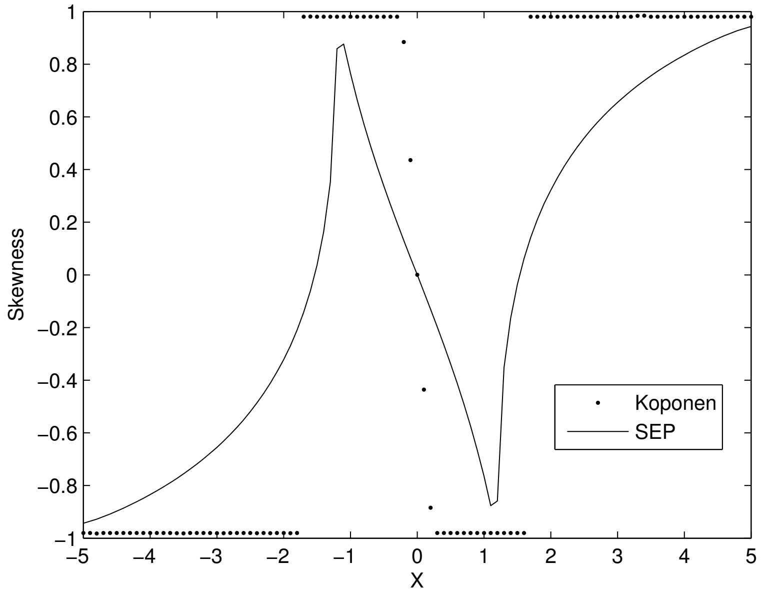

Figure 5.5: Comparison of ML skewness estimators for the standard Koponen ($\alpha = 1.4$, $\lambda = 1.4$) and the standard SEP ($\alpha = 1.4$) distribution based on a single observation

5.2 The Skew-MCECD model

5.2.1 The definition

The Skew-MCECD model focuses on the skewness parameter β_t as the only time-varying parameter in the conditional density. The corresponding dynamics are

$$\beta_t = \underset{\xi}{\operatorname{argmin}} -H_t(\xi),$$

where, according to the general MCECD model, the cross-entropy follows the equations

$$\begin{aligned} H_t(\beta) &= \alpha_0 \cdot \log(f_\beta(\bar{x})) + \alpha_1 \cdot \log(f_\beta(r_{t-1})) + \alpha_2 \cdot H_{t-1}(\beta) \qquad (5.5) \\ H_1(\beta) &= \log(f_\beta(x_0)) \end{aligned}$$

with $H_t(\beta) = H_t^1(\beta)$ and $\bar{x}$ and x_0 being scalars.

5.2.2 Explicit Skew-MCECD dynamics

As a next step, we will derive a closed-form expression for the skewness parameter dynamics of a specific Skew-MCECD model. Assumption A3 provides the necessary foundation for our analysis.

Assumption A3 *The skewness is the only time-varying parameter $\theta_t = \beta_t$ and the conditional distribution is of the SEP type $r_t \sim$ SEP$(\alpha, \beta_t, \sigma, \mu)$, with kurtosis parameter $\alpha = 1$.*

Since the SEP distribution is piecewise defined for $x \leq \mu$ and $x > \mu$, we need to discriminate between non-positive excess returns $r_t \leq \mu$ and positive excess returns $r_t > \mu$. For this purpose, we introduce the index sets I_t^- and I_t^+ to split the historical returns accordingly.

$$\begin{aligned} I_t^- &:= \{s \in \mathbb{N}_{>0} | r_{t-s} \leq \mu\} \\ I_t^+ &:= \{s \in \mathbb{N}_{>0} | r_{t-s} > \mu\}. \end{aligned}$$

Proposition 5.2.1. *Given a Skew-MCECD model which satisfies assumption A3, the dynamics for the skewness parameter can be derived explicitly*

$$\beta_t = \begin{cases} \frac{M_t^- - \sqrt{M_t^- \cdot M_t^+}}{M_t^- - M_t^+} & : \quad M_t^+ \neq M_t^- \\ 0.5 & : \quad M_t^+ = M_t^-, \end{cases} \tag{5.6}$$

where M_t^- and M_t^+ are defined by

$$\begin{aligned} M_t^- \quad &:= \quad \alpha_2^{t-1} \left| \frac{x_0 - \mu}{2\sigma} \right| \cdot \mathbf{1}_{x_0 \leq \mu} + \sum_{s=1}^{t-1} \alpha_2^{s-1} \alpha_0 \left| \frac{\bar{x} - \mu}{2\sigma} \right| \cdot \mathbf{1}_{\bar{x} \leq \mu} \qquad (5.7) \\ &\quad + \sum_{s \in I_t^-} \alpha_2^{s-1} \alpha_1 \left| \frac{r_{t-s} - \mu}{2\sigma} \right| \\ M_t^+ \quad &:= \quad \alpha_2^{t-1} \left| \frac{x_0 - \mu}{2\sigma} \right| \cdot \mathbf{1}_{x_0 > \mu} + \sum_{s=1}^{t-1} \alpha_2^{s-1} \alpha_0 \left| \frac{\bar{x} - \mu}{2\sigma} \right| \cdot \mathbf{1}_{\bar{x} > \mu} \\ &\quad + \sum_{s \in I_t^+} \alpha_2^{s-1} \alpha_1 \left| \frac{r_{t-s} - \mu}{2\sigma} \right|. \end{aligned}$$

Proof. See appendix A.7. □

From definition 4.1.3 we deduce that the Skew-MCECD model based on SEP distribution is not of the linear autoregressive type. Equation (5.7) suggests that in the constant value $\bar{x}$ and the starting value x_0 vanish for either M_t^- or M_t^+. This creates a certain asymmetry in the equations. In order to avoid this effect, we can model the constant scenario by $\bar{\beta}$ and the starting scenario by β_0. As a result we need to determine two tuples $(\bar{x}^-, \bar{x}^+)$ and (x_0^-, x_0^+) which induce the skewness parameters $\bar{\beta}$ and β_0 respectively

$$\begin{aligned}
\bar{\beta} &= \operatorname*{argmin}_{\beta} \left[\log(f_{SEP}(\bar{x}^-; \alpha, \sigma, \beta, \mu)) + \log(f_{SEP}(\bar{x}^+; \alpha, \sigma, \beta, \mu))\right] \\
\beta_0 &= \operatorname*{argmin}_{\beta} \left[\log(f_{SEP}(x_0^-; \alpha, \sigma, \beta, \mu)) + \log(f_{SEP}(x_0^+; \alpha, \sigma, \beta, \mu))\right],
\end{aligned}$$

where $\bar{x}^-, x_0^- < \mu$ and $\bar{x}^+, x_0^+ > \mu$. The first-order condition then yields

$$\frac{\partial \left[\log(f_{SEP}(x^-; \alpha, \sigma, \beta, \mu)) + \log(f_{SEP}(x^+; \alpha, \sigma, \beta, \mu))\right]}{\partial \beta}\Big|_{\beta} = 0.$$

Further calculations lead us to

$$\left|\frac{x^- - \mu}{2\sigma}\right| \cdot \beta^{-2} - \left|\frac{x^+ - \mu}{2\sigma}\right| \cdot (1-\beta)^{-2} = 0.$$

With the definitions of $x^- < \mu$ and $x^+ > \mu$, we get

$$x^- = -(x^+ - \mu) \cdot \frac{\beta^2}{(1-\beta)^2} + \mu.$$

If we set the distance of x^+ and μ to 1, the equations for x^- and x^+ are

$$\begin{aligned}
x^- &= \mu - \frac{\beta^2}{(1-\beta)^2} \\
x^+ &= \mu + 1,
\end{aligned}$$

which can be used both for $\bar{\beta}$ and β_0.

The drawback of the SEP driven Skew-MCECD is that β drives not only the skewness, but also other moments of the distribution. Hence it is impossible to model skewness as a stand-alone feature.

5.3 The Vola-Skew-MCECD model

Assuming that the volatility and the skewness are the only time-varying parameters $\theta_t = (\sigma_t, \beta_t)$ results in the so-named Vola-Skew-MCECD model.

It is a special case of the MCECD framework and follows the equations

$$\sigma_t = \underset{\xi}{\operatorname{argmin}} -H_t^1(\xi, \beta_t),$$
$$\beta_t = \underset{\xi}{\operatorname{argmin}} -H_t^2(\sigma_t, \xi),$$

with the cross-entropy process for the volatility component

$$H_t^1(\sigma, \beta) = \alpha_0 \cdot \log(f_{\sigma,\beta}(\bar{x}_1)) + \alpha_1 \cdot \log(f_{\sigma,\beta}(r_{t-1})) + (\alpha_2 + \alpha_3) \cdot H_{t-1}^1(\sigma, \beta)$$
$$H_1^1(\sigma, \beta) = \log(f_{\sigma,\beta}(x_{0,1}))$$

and for the skewness component

$$H_t^2(\sigma, \beta) = \alpha_0 \cdot \log(f_{\sigma,\beta}(\bar{x}_2)) + \alpha_2 \cdot \log(f_{\sigma,\beta}(r_{t-1})) + (\alpha_1 + \alpha_3) \cdot H_{t-1}^2(\mu, \sigma)$$
$$H_1^2(\sigma, \beta) = \log(f_{\sigma,\beta}(x_{0,2})).$$

$\bar{x} = (\bar{x}_1, \bar{x}_2)$ and $x_0 = (x_{0,1}, x_{0,2})$ are two-dimensional vectors.

The Koponen distribution does not result in closed-form expressions for the parameter vector process $(\theta_t)_t$. This drives computational complexity in practical applications because for every period t we need to simultaneously search for two optimal cross-entropies $H_t^1(\sigma, \beta)$ and $H_t^2(\sigma, \beta)$. Moreover, these calculations have to be carried out numerically. In order to reduce the computational complexity of the optimization, we make the following simplifying assumption.

Assumption A4 *The volatility estimator is independent of the conditional skewness and based on the Gaussian assumption. The skewness parameter is defined by a standard Koponen model.*

Figure 5.6 illustrates the approximation of assumption A4. The volatility estimator becomes asymmetric with non-zero skewness. Negative observations hence imply a lower volatility compared to positive observations. Since the impact of the skewness is relatively small and its result is overcompensated by the “leverage effect”, it seems reasonable to neglect the asymmetry in the estimator and use the zero-skewness version instead.

Another consequence of assumption A4 is that parameter inference for the volatility dynamics is based on the Gaussian distribution. Recalling Bollerslev’s QMLE method, we know that the GARCH estimators under the normality assumption are consistent even if the underlying distribution

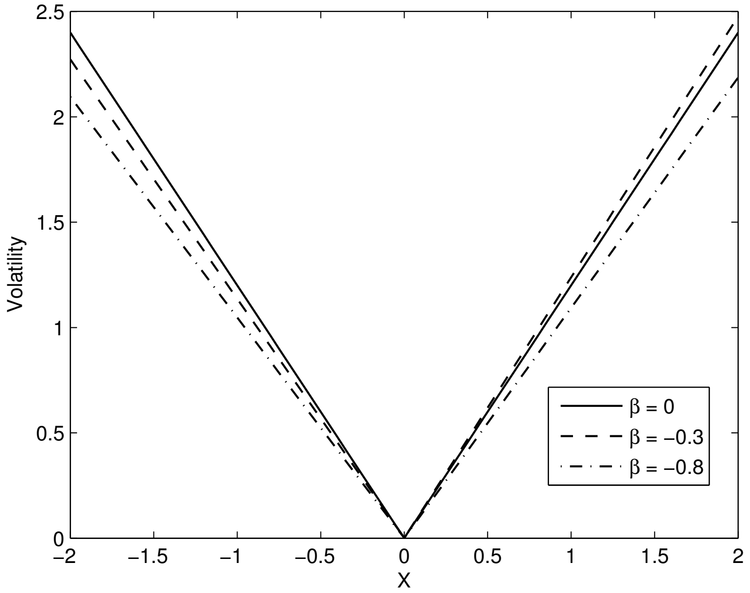

Figure 5.6: Effects of skewness parameter on volatility estimator for standard Koponen distribution ($\alpha = 0.5$, $\lambda = 1.7$) based on a single observation

is non-Gaussian. Thus, we can justify the use of the GARCH model as a consistent approximation.

Summing up, assumption A4 changes the approach from a parallel to a sequential two-step method. First, the volatility process $(\sigma_t)_t$ is estimated using the dynamics given by proposition 4.1.1

$$\begin{aligned}
&\sigma_t^2 = \alpha_0 \cdot (\bar{x}_1 - \mu)^2 + \alpha_1 \cdot (r_{t-1} - \mu)^2 + (\alpha_2 + \alpha_3) \cdot \sigma_{t-1}^2 \\
&\sigma_1^2 = (x_{0,1} - \mu)^2.
\end{aligned}$$

Afterwards a Skew-MCECD model is applied to the standardized log-return process $(\frac{r_t}{\sigma_t})_t$. Since the conditional skewness process is based on the Koponen assumption, there is no explicit form of the parameter dynamics β_t available. The two-step procedure makes parameter inference less cumbersome because we can use a sequential method where the first step, the GARCH inference, can be carried out highly efficiently.

5.4 Simulation and empirical results

In this section we show three main results for the MCECD models for conditional skewness: (1) Skew-MCECD is capable of distinguishing time-varying from time-invariant trajectories, (2) the skewness of daily log-return data for U.S. stock indices varies over time, and (3) the skewness trajectory from Skew-MCECD is more informative compared to the one from the ARCD model. Goodness-of-fit is again measured my means of the KS test as well as the AD and CvM statistics.

In order to test the modeling of conditional skewness, we employ simulated log-returns with time-varying skewness, but constant mean, volatility, and kurtosis. The analyses (2) and (3) are run on daily log-returns of U.S. stock indices from 06/26/2005 till 06/26/2009. Our selection includes the U.S. financial crisis in September 2008.

5.4.1 The time-varying property

In section 4.3 we have already shown the capability of MCECD to evaluate the time-varying property of mean and volatility. This paragraph extends this result to cover conditional skewness. We will simulate two different data sets of 1000 innovations from a standard Koponen distribution with parameters $\alpha = 0.5$, $\lambda = 1.7$, $C = C_0$, and $m = 0$. Assuming zero mean and constant volatility $\sigma = 0.01$ yields $r_t = \sigma \cdot \epsilon_t$. With a kurtosis of $\mathrm{K}[r_t] = 4.2976 > 3$, the log-returns possess the empirically observed leptokurtosis.

The first sample A possesses a time-varying skewness which is generated according to regime switching dynamics. Let $(p_t^\beta)_{t\in\mathbb{N}_{>0}}$ be an uniformly distributed random process with $p_t^\beta \sim \mathrm{U}(0,1)$ and $\beta = (-0.6, -0.4, -0.2, -0.1, 0)$ the set of potential skewness parameters ranked from 1 to 5. Then the variable k_t indicates the β valid at time t and is defined by the following specifications

$$k_t = \begin{cases} k_{t-1} + 1 & : \quad p_{t-1}^\beta > 0.95 \\ k_{t-1} - 1 & : \quad p_{t-1}^\beta < 0.05. \\ k_{t-1} & : \quad \text{else.} \end{cases}$$

We enforce the condition $1 \leq k_t^* \leq 5$ by $k_t^* = \max\{\min\{k_t, 5\}, 1\}$.

For sample A, we estimate the Skew-MCECD parameters and compare goodness-of-fit statistics as well as the resulting skewness trajectories and the scenario probability α_1 to a constant skewness model. From table 5.1, we can notice that the Skew-MCECD successfully detects the time-varying property of the conditional skewness: $\alpha_1 > 0$. Furthermore, the probability that the skewness remains unchanged is around $\alpha_2 \approx 90\%$ which corresponds to our data generation scheme. Table 5.2 supports the fact that, in comparison to an unconditional density model with constant skewness, the Skew-MCECD provides an improved goodness-of-fit for the tail as well as for the overall distribution. The estimated trajectory of parameter β_t in figure 5.7 again highlights that Skew-MCECD can cope with time-varying skewness. We can also see that the quality of the estimates is not comparable to the conditional mean nor the volatility case. This effect is caused by the non-linearity of the skewness property. The example shows that general trends and relevant parameter areas can be identified by the Skew-MCECD model.

Method	α	β	λ	$\bar{x}$	α_0	α_1	α_2
Skew-MCECD	0.5	-	1.7	0.3127	0.045	0.0450	0.91
Constant skewness	0.5	-0.4689	1.7	-	-	-	-

Table 5.1: Parameter estimates for simulated data with time-varying skewness

Method	KS test	p-value	AD	AD^2	CvM
Skew-MCECD	0	0.9523	0.0744	0.3148	0.0430
Constant skewness	0	0.8651	0.0712	0.3463	0.0550

Table 5.2: Goodness-of-fit results for simulated data with time-varying skewness

The data in reference sample B has constant, negative skewness with $\beta = -0.3825$. Again we analyze the performance of the Skew-MCECD model with focus on the scenario probability α_1. According to the parameter estimates in table 5.3, the likelihood for a change in conditional skewness is $\alpha_1 = 0.1\%$. Given the approximations used for the inference procedure, e.g. discretization of the feasible parameter sets, this might not be significant. From the fact that the parameter trajectory in figure 5.8 is almost constant, we can deduce that the Skew-MCECD model detects the time-invariant feature for the given data sample. The deviation from the constant skewness is negligible. Finally the goodness-of-fit results are summarized in table 5.4.

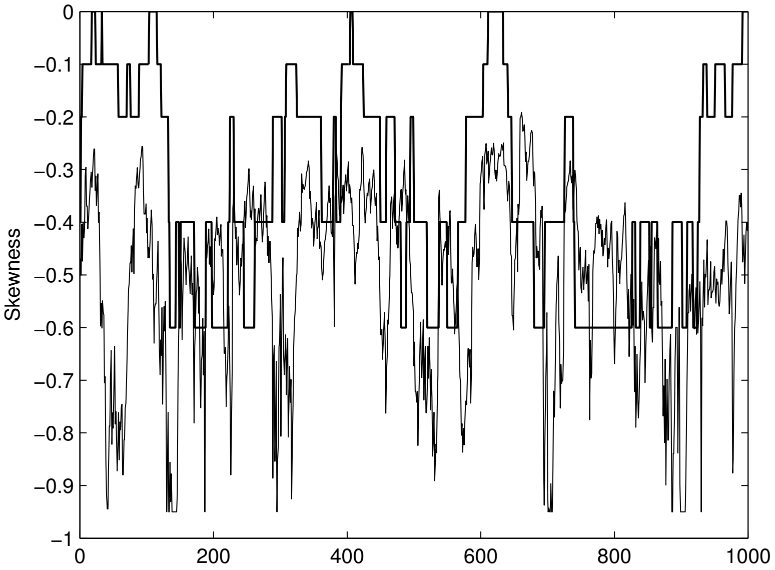

Figure 5.7: Comparison of underlying (top line) and estimated (bottom line) trajectories for simulated, time-varying skewness based on standard Koponen distribution ($\alpha = 0.5$, $\lambda = 1.7$)

To sum up, the analysis based on two simulated data samples strengthens our findings from chapter 4. The MCECD is capable of assessing the time-varying property of parameter processes. Its performance in the conditional skewness case is, however, not as strong as for the mean and volatility due to the increased complexity of the estimation.

Method	α	β	λ	$\bar{x}$	α_0	α_1	α_2
Skew-MCECD	0.5	-	1.7	0.2608	0.0230	0.001	0.976
Constant skewness	0.5	-0.3825	1.7	-	-	-	-

Table 5.3: Parameter estimates for simulated data with constant skewness

5.4.2 Empirical skewness

Now we will analyze daily log-return data of three U.S. stock indices using a Vola-Skew-MCECD. We will compare our results to the basic GARCH model and the ARCD approach with regard to goodness-of-fit and parameter trajectories. This way we can evaluate the explanatory power

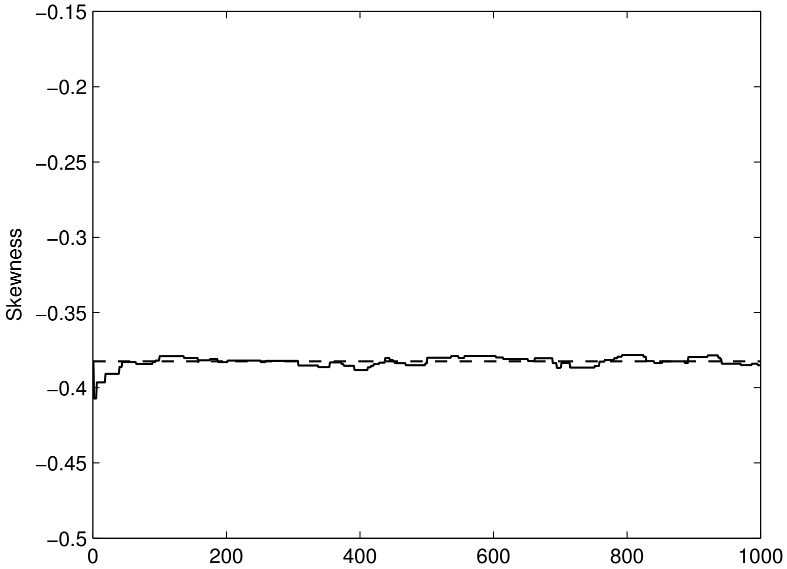

Figure 5.8: Comparison of underlying (dashed) and estimated (line) trajectories for constant skewness based on standard Koponen distribution ($\alpha = 0.5$, $\beta = -0.3825$, $\lambda = 1.7$)

Method	KS test	p-value	AD	AD^2	CvM
Skew-MCECD	0	0.0940	0.0818	2.3725	0.4070
Constant skewness	0	0.0954	0.0816	2.3707	0.4068

Table 5.4: Goodness-of-fit results for simulated data with constant skewness

of the respective models and their ability to describe parameter processes appropriately.

Given the return data r_t, we use an ARCD specification with constant mean $\mu_t = c$ and constant kurtosis $k_t = k_{const}$

$$\sigma_t^2 = \beta_0 + \beta_1\ (r_{t-1} - c)^2 + \beta_2\ \sigma_{t-1}^2$$
$$s_t = \gamma_0 + \gamma_1\ \left(\frac{r_{t-1} - c}{\sigma_{t-1}}\right)^3 + \gamma_3\ s_{t-1}.$$

The innovations $\frac{r_t - c}{\sigma_t} \sim \text{Koponen}(\alpha, C_0, \beta_t, \lambda, 0)$ are assumed to be standard Koponen distributed. To reduce the number of parameters to be estimated and hence the complexity of the inference, we assume $\alpha = 0.5$ and $\lambda = 1.7$

which induces leptokurtic shape of the distribution (kurtosis is 4.2976).

Table B.8 shows that all three models provide a similar goodness-of-fit. This holds true for the overall fit as well as for the tail fit. From this point of view there is a strong argument in favor of the GARCH model because it requires two parameters less than the others. This not only simplifies calculations but also reduces the risk for overfitting. As GARCH implies constant skewness, this result would suggest unconditional skewness models.

The parameter estimates presented in table B.9, however, indicate that the skewness is time-varying. We have seen for the Skew-MCECD model that it is capable of distinguishing between time-varying and time-invariant conditional skewness. A positive probability $\alpha_1 \approx 0.1$ for a parameter change is thus a good indicator for a time-varying moment. That is why the estimators speak in favor of time-varying skewness. This contradiction to our findings from the goodness-of-fit analysis might be caused by our simplifying approximations. Fixing the symmetric volatility estimator as well as some of the distributional parameters could decrease the quality of the fit for the conditional skewness model. In order to get a better understanding of the estimated conditional skewness effects, we take a look at the induced skewness trajectories. Knowing that for a Koponen model the parameter β_t and the conditional skewness s_t are proportional, we will compare the results from ARCD and Vola-Skew-MCECD based on the β_t trajectories.

Figure 5.9 shows the implied skewness of the ARCD model. We can observe two different structures. For the S&P 500 and the Dow Jones, the skewness trajectories are reverting. This is a direct result of the negative autoregression parameter γ_2. In this market model, the sign of the skewness toggles from positive to negative and vice versa. It suggests that the preferences change substantially from one day to another. This is, however, a questionable result since we expect that the market changes gradually over time, as it is the case for the volatility. The figure for the Nasdaq 100 results from a positive γ_2 and a negative γ_1. For this sample it stands clear that the skewness varies around its constant alternative. Nevertheless the changes are still extreme, resulting in high peaks. Although the trajectory differs from the first two, it still is not in line with the general idea of conditional density which suggests a gradual change of parameters based on a gradual stream of information.

The time-varying skewness implied by the Vola-Skew-MCECD model in figure 5.10 conforms to the concept of conditional parameter updating. The skewness changes step-by-step with every new piece of information. Although it is in general negative, it varies around the constant alternative suggested by the GARCH model with Koponen distributed innovations. The trajectory for S&P 500 data is almost constant which, according to our previous analysis for the Skew-MCECD model, implies the constant parameter case. For Dow Jones and Nasdaq 100 the Vola-Skew-MCECD models yields a significantly positive α_1 and thus a time-varying skewness.

For possible interpretations of the conditional skewness, we have summarized the three charts for price, volatility and skewness of the daily Dow Jones log-returns in figure 5.11. Without an explicit correlation analysis it is obvious that the conditional skewness is less negative in times where the market is stable and rising, whereas it plummets in times of crisis, especially in the pre-crisis time in the beginning of 2008. With regard to our findings from the simulated conditional skewness in figure 5.7, it is important to note that the estimated absolute skewness may be less informative than the direction of the parameter changes. This means that the estimator does not always reflect the proper value of the parameter, but it imitates the movements. Conditional skewness can consequently provide an additional indicator for market preferences. This finding coincides with the results from Chen *et al.* (2001) who used conditional skewness in the forecast of market crashes.

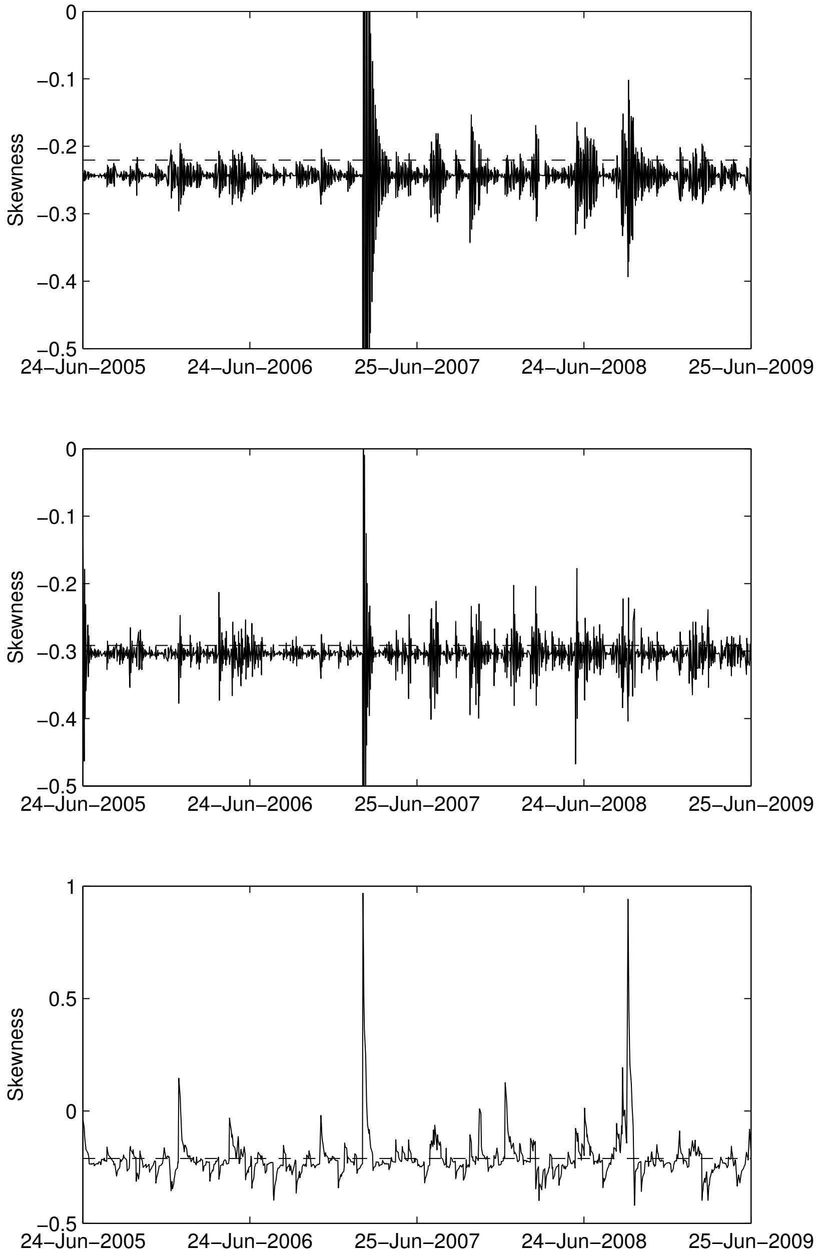

Figure 5.9: Time-varying skewness parameters (solid line) of ARCD models based on Koponen distribution for daily log-return data of S&P 500 (top), Dow Jones (middle), and Nasdaq 100 (bottom) compared to constant skewness parameter (dashed line)

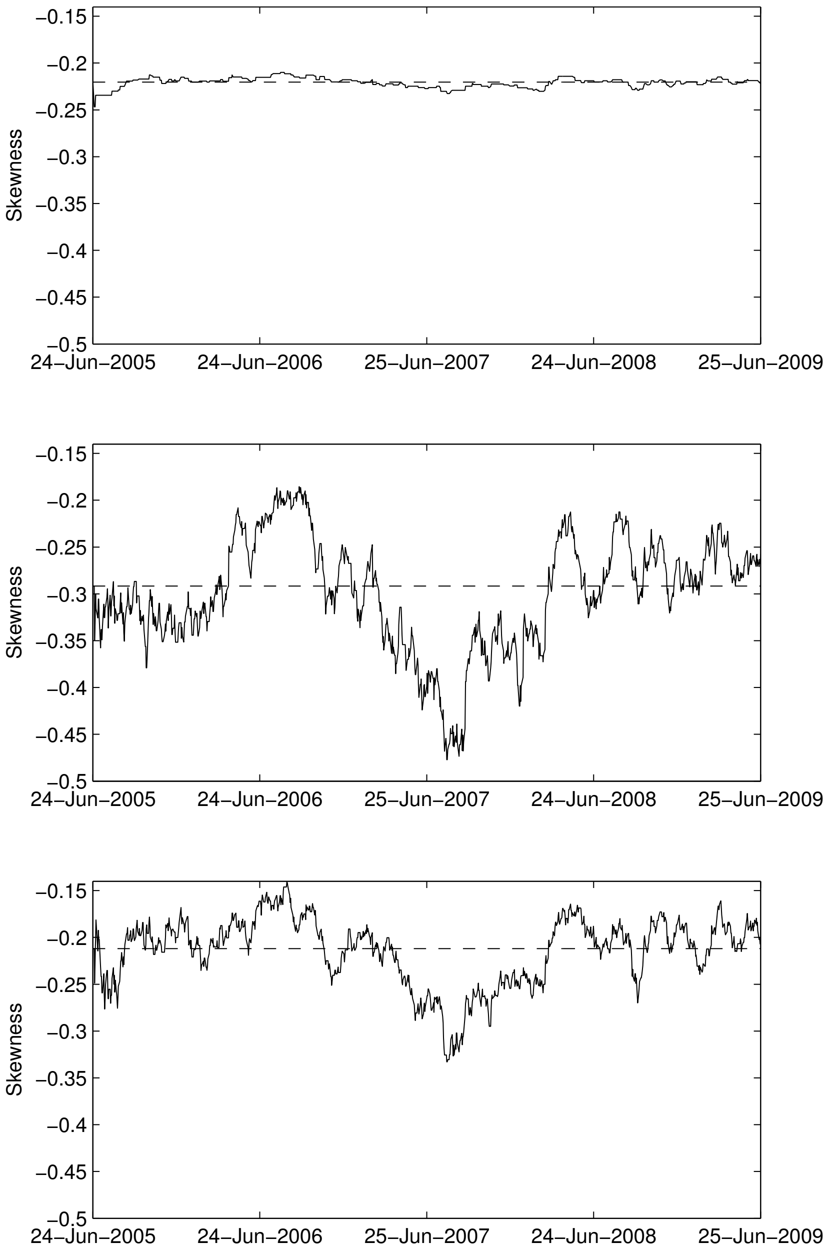

Figure 5.10: Time-varying skewness parameters (solid line) of Vola-Skew-MCECD models based on Koponen distribution for daily log-return data of S&P 500 (top), Dow Jones (middle), and Nasdaq 100 (bottom) compared to constant skewness parameter (dashed line)

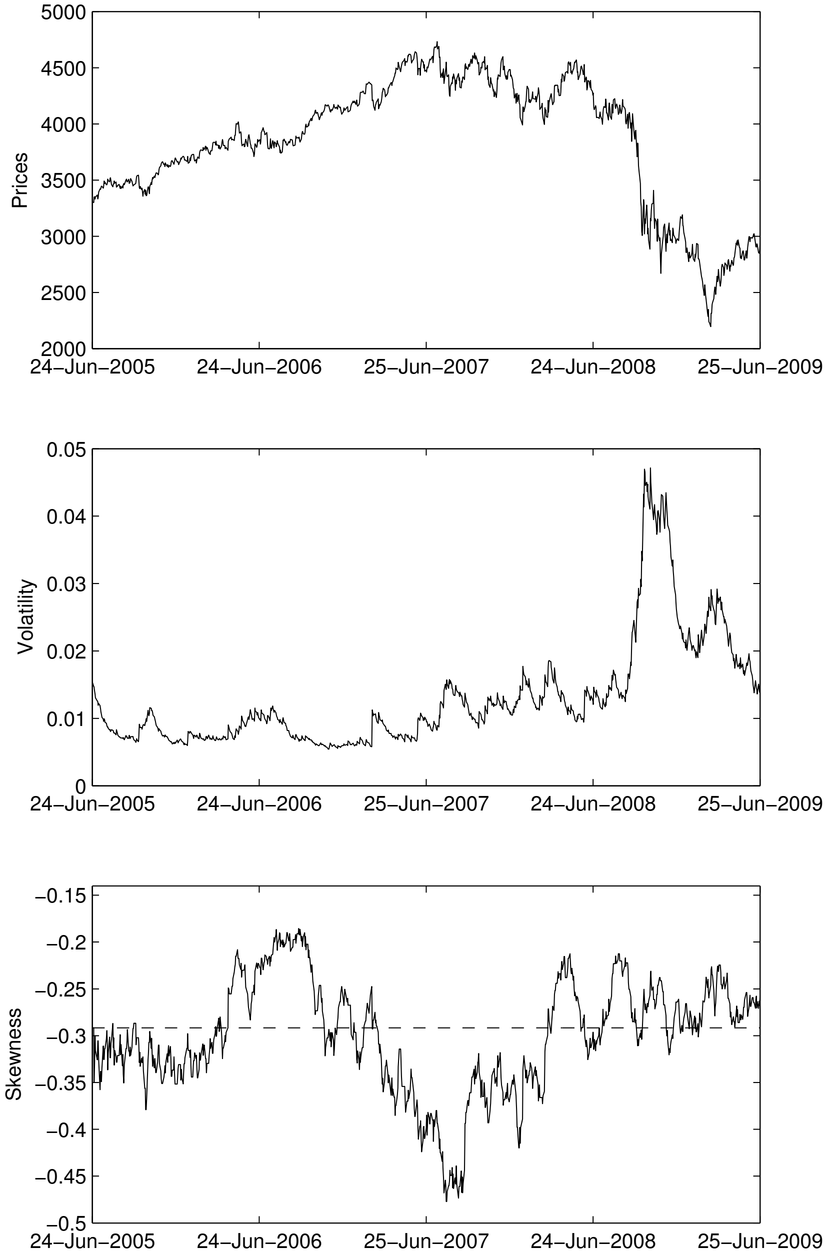

Figure 5.11: Price-volatility-skewness triplet of Vola-Skew-MCECD model based on Koponen distribution for daily log-return data of Dow Jones

Conclusion

In this dissertation we introduce a new time series model, the minimally cross-entropic conditional density model. Our approach to the research can be defined as a generalization of the seminal GARCH model. We show that MCECD can overcome drawbacks associated with an autoregressive approach. In particular, MCECD establishes a strong link between distributional assumption and parameter dynamics, thus accounting for dependencies in the parameter structure. Furthermore, it does not rely on moment estimators, resolving inference problems for distributions with infinite moments, such as stable Paretian.

In the realm of non-Gaussian theory, we show that MCECD includes not only the GARCH model, but also the power-ARCH model as a special case and that induced parameter dynamics can be non-linear, even for the volatility process. Concerning skewness estimation, our analyses suggest that MLE as an efficient estimator is preferable to moment estimators which always introduce a bias. Especially for the case of a single data point, we highlight the weaknesses of modeling based sample moments. Moreover, we formulate a conditional skewness model and derive the explicit, non-linear parameter dynamics for the Laplace distribution.

In order to assess the modeling quality of the MCECD, we compare our model to the generally known ARMA-GARCH along three dimensions: goodness-of-fit, forecasting quality, and induced trajectories. Our empirical analysis shows that Mean-Vola-MCECD leads to a slightly improved forecasting quality based on daily return data from U.S. stocks and U.S. stock indices. An advantage of the MCECD based model is that it requires fewer parameters, thus reducing the risk of overfitting. Its most striking feature is the capability to detect if a parameter process is time-varying. MCECD also results in a more accurate estimation of the underlying parameter process while ARMA-GARCH tends to explain the noise. For conditional skewness models, our findings suggest that the skewness of U.S. stock index data varies over time. The probability for a change in conditional skewness is significant. Concerning the goodness-of-fit statistics, conditional skewness models do not outperform traditional time series approaches. The contribution of MCECD is to provide a skewness trajectory as an additional indicator for market preference modeling. Since existing literature suggests that conditional skewness can be used as an indicator for market crash prediction, our model gives way to further research based on time series implied trajectories.

This research on the conditional density model has a great potential for further investigation in portfolio and risk management. It is highly interesting to see a multivariate MCECD using a copula model in order to capture the dependence structure in the portfolio. By means of this model, it is possible to test whether or not the dependence parameters vary over time. Concerning risk management, the conditional density approach offers the possibility to calculate time-varying risk measures such as Value-at-Risk (VaR) or Conditional Value-at-Risk (CVaR). The idea of exploiting the complete distributional information has already been very successful in risk management with the spectral and distortion risk measures.

Bibliography

Bachelier, L. (1900). Théorie de la spéculation. *Annáles de l'Ecole Normale Supérieur*, *17*, 21–86.

Bera, A. K. and Bilias, Y. (2002). The MM, ME, ML, EL, EF and GMM approaches to estimation: a synthesis. *Journal of Econometrics*, *107*(1–2), 51–86.

Bera, A. K. and Higgins, M. L. (1993). ARCH models: Properties, estimation and testing. *Journal of Economic Surveys*, *7*(4), 305–366.

Bianchi, M. L., Rachev, S. T., Kim, Y. S., and Fabozzi, F. J. (2010). Tempered infinitely divisible distributions and processes. *Theory of Probability and Its Applications, to appear*.

Black, F. (1976). Studies of stock price volatility changes. Working paper, Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Bollerslev, T. and Wooldridge, J. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews*, *11*(2), 143–172.

Bougerol, P. and Picard, N. (1992). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics*, *52*(1–2), 115–127.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis (Wiley Classics Library)*. Wiley-Interscience.

Chen, J., Hong, H., and Stein, J. C. (2001). Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics*, *61*(3), 345–381.

Chernobai, A. S., Rachev, S. T., and Fabozzi, F. J. (2007). *Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis*. Wiley Publishing.

Christoffersen, P. and Jacobs, K. (2004). Which GARCH model for option valuation? *Management Science*, *50*(9), 1204–1221.

Dark, J. G. (2010). Estimation of time varying skewness and kurtosis with an application to value at risk. *Studies in Nonlinear Dynamics & Econometrics*, *14*(2).

Ding, Z., Granger, C. W., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, *1*(1), 83–106.

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

DuMouchel, W. H. (1975). Stable distributions in statistical inference: 2. Information from stably distributed samples. *Journal of the American Statistical Association*, *70*(350), 386–393.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *22*(1), 139–153.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007.

Engle, R. F. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, *48*(5), 1749–78.

Fama, E. F. (1963). Mandelbrot and the stable Paretian hypothesis. *Journal of Business*, *36*, 420–425.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, 309–368.

Gallant, A., Hsieh, D., and Tauchen, G. (1991). On fitting a recalcitrant series: The pound/dollar exchange rate, 1974-83. *Nonparametric and Semiparametric Methods in Econometrics and Statistics, Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, 199–240.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, *31*(4), 1208–1211.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, 1 ed.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, *35*(3), 705–730.

Harvey, C. R. and Siddique, A. (1999). Autoregressive conditional skewness. *The Journal of Financial and Quantitative Analysis*, *34*(4), 465–487.

Hentschel, L. (1995). All in the family nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics*, *39*(1), 71–104.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review Online Archive (Prola)*, *106*(4), 620–630.

Kannappan, P. (1972). On shannon's entropy, directed divergence and inaccuracy. *Probability Theory and Related Fields*, *22*, 95–100.

Kerridge, D. F. (1961). Inaccuracy and inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, *23*(1), 184–194.

Kim, T.-H. and White, H. (2004). On more robust estimation of skewness and kurtosis. *Finance Research Letters*, *1*(1), 56–73.

Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2010). Tempered stable and tempered infinitely divisible GARCH models. *Journal of Banking & Finance, to appear*.

Koponen, I. (1995). Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process. *Phys. Rev. E*, *52*(1), 1197–1199.

Kullback, S. (1959). *Information theory and statistics*. John Willey & Sons, New York.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. Finance and Economics Discussion Series 95-24, Board of Governors of the Federal Reserve System (U.S.).

Lopez, J. A. (1998). Methods for evaluating value-at-risk estimates. Research Paper 9802, Federal Reserve Bank of New York.

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business*, *36*(4), 394–419.

Mandelbrot, B. (1967). The variation of some other speculative prices. *The Journal of Business*, *40*(4), 393–413.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Mittnik, S., Paolella, M. S., and Rachev, S. T. (2002). Stationarity of stable power-GARCH processes. *Journal of Econometrics*, *106*(1), 97–107.

Nelson, D. B. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, *6*(3), 318–334.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347–370.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*(2), 237–249.

Pagan, A. R. and Schwert, G. W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, *45*(1–2), 267–290.

Rachev, S. T., Fabozzi, F. J., and Menn, C. (2005). *Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing*. John Wiley & Sons.

Rosiński, J. (2007). Tempering stable processes. *Stochastic Processes and Their Applications*, *117*(6), 677–707.

Sato, K. (1999). *Lévy processes and infinitely divisible distributions*. Cambridge University Press.

Scherer, M., Kim, Y. S., Rachev, S. T., and Fabozzi, F. J. (2010a). A FFT-based approximation of tempered stable and tempered infinitely divisible distributions. Technical report, Chair of Econometrics, Statistics and Mathematical Finance School of Economics and Business Engineering University of Karlsruhe (http://www.statistik.uni-karlsruhe.de/download/).

Scherer, M., Kim, Y. S., Rachev, S. T., and Fabozzi, F. J. (2010b). Minimally cross-entropic conditional density: a generalization of GARCH. Technical report, Chair of Econometrics, Statistics and Mathematical Finance School of Economics and Business Engineering University of Karlsruhe (http://www.statistik.uni-karlsruhe.de/download/).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*, 379–423,623–656.

Sharma, B. D. and Taneja, I. J. (1974). On axiomatic characterization of information-theoretic measures. *Journal of Statistical Physics*, *10*, 337–346.

Subbotin, M. T. (1923). On the law of frequency of error. *Matematicheskii Sbornik*, *31*(2), 296–301.

Zhu, D. and Zinde-Walsh, V. (2009). Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics*, *148*(1), 86–99.

Appendices

Appendix A

Proofs

A.1 The iterative formula

We prove this proposition by means of induction. The base case $t = 2$ directly follows from definition 3.3.2

$$\begin{aligned} H_2^i(\theta) &= {\beta_i}^1 \cdot \log(f_\theta(x_{0,i})) + {\beta_i}^0 \cdot \alpha_0 \log(f_\theta(\bar{x}_i)) + {\beta_i}^0 \cdot \alpha_i \log(f_\theta(r_1)) \\ &= \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_1)) + \beta_i \cdot \log(f_\theta(x_{0,i})) \\ &= \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_1)) + \beta_i \cdot H_1^i(\theta) \end{aligned}$$

For the inductive step, we assume that there exists a $t \in \mathbb{N}_{>0}$ for which the equation (3.5) holds and write

$$\begin{aligned} H_{t+1}^i(\theta) &= \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_{t-1})) + \beta_i \cdot H_t^i(\theta) \\ &= \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_{t-1})) \\ &\quad + \beta_i \cdot \Big[{\beta_i}^{t-1} \cdot \log(f_\theta(x_{0,i})) + \sum_{s=1}^{t-1} {\beta_i}^{s-1} \cdot \alpha_0 \log(f_\theta(\bar{x}_i)) \\ &\quad + \sum_{s=1}^{t-1} {\beta_i}^{s-1} \cdot \alpha_i \log(f_\theta(r_{t-s}))\Big]. \end{aligned}$$

Now we expand the equation and get

$$\begin{aligned}H^i_{t+1}(\theta) = \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_{t-1})) + {\beta_i}^t \cdot \log(f_\theta(x_{0,i}))\\ + \sum_{s=2}^{t} {\beta_i}^{s-1} \cdot \alpha_0 \log(f_\theta(\bar{x}_i)) + \sum_{s=2}^{t} {\beta_i}^{s-1} \cdot \alpha_i \log(f_\theta(r_{t-s}))\\ = {\beta_i}^t \cdot \log(f_\theta(x_{0,i})) + \sum_{s=1}^{t} {\beta_i}^{s-1} \cdot \alpha_0 \log(f_\theta(\bar{x}_i))\\ + \sum_{s=1}^{t} {\beta_i}^{s-1} \cdot \alpha_i \log(f_\theta(r_{t-s})).\end{aligned}$$

Base case and inductive step together prove the iterative formula in (3.5).

A.2 Predictability

In order to prove the predictability of the cross-entropy process, we need to show that $H^i_t(\theta)$ is a deterministic function of the past innovations $\epsilon_{t-1}, ..., \epsilon_1$. We do so by induction over time t. With equation (3.3) the base case $t = 2$ results in

$$H^i_2(\theta) = \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_1)) + \beta_i \cdot H^i_1(\theta).$$

If we substitute r_1 and $H^i_1(\theta)$ by their defining terms we get

$$H^i_2(\theta) = \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(F^{-1}_{\theta_1}(F_{\theta_{Norm}}(\epsilon_1)))) + \beta_i \cdot \log(f_\theta(x_{0,i})).$$

Since $\bar{x}_i$, $x_{0,i}$, and α_i are deterministic, it is left to show that the term $\log(f_\theta(F^{-1}_{\theta_1}(F_{\theta_{Norm}}(\epsilon_1))))$ is $\mathcal{F}_1$-measurable. The defining equation system

$$\theta_{1,i} = \operatorname*{argmin}_{\xi \in \Theta_i} -H^i_1(\xi, \theta_{t,-i})$$

reveals that θ_1 is deterministic. Furthermore, we know that θ_{Norm} is deterministic and hence we conclude that the randomness of $\log(f_\theta(F^{-1}_{\theta_1}(F_{\theta_{Norm}}(\epsilon_1))))$ is only driven by ϵ_1, which is—by definition of the filtration—$\mathcal{F}_1$-measurable, thus proving the predictability for the base case.

For the inductive step, we assume that $H^i_t(\theta)$ is $\mathcal{F}_{t-1}$-measurable. In analogy to the base case we conclude from

$$H^i_{t+1}(\theta) = \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(F^{-1}_{\theta_t}(F_{\theta_{Norm}}(\epsilon_t)))) + \beta_i \cdot H^i_t(\theta)$$

that $H^i_{t+1}(\theta)$ is $\mathcal{F}_t$-measurable if $\log(f_\theta(F^{-1}_{\theta_t}(F_{\theta_{Norm}}(\epsilon_t))))$ is $\mathcal{F}_t$-measurable. Since for every component i it holds that

$$\theta_{t,i} = \operatorname*{argmin}_{\xi \in \Theta_i} -H^i_t(\xi, \theta_{t,-i}),$$

and, by assumption, that $H^i_t(\theta)$ is $\mathcal{F}_{t-1}$-measurable, we know that θ_t is also $\mathcal{F}_{t-1}$-measurable. Consequently θ_t is also $\mathcal{F}_t$-measurable. By definition of the filtration, ϵ_t is $\mathcal{F}_t$-measurable. From this it follows directly that the log-term as a deterministic function of $\mathcal{F}_t$-measurable random variables is $\mathcal{F}_t$-measurable and thus that $H^i_{t+1}(\theta)$ is predictable.

A.3 Convergence of weighted geometric series

Since we assume that $f_\theta(r_t) > 0$, and hence $-\infty < \log(f_\theta(r_{t-k-1})) < \infty$ we can define

$$C_{max} := \sup_{k \geq 0} \Big| \log(f_\theta(r_{t-k-1})) \Big| > 0$$

This yields

$$\Big| \beta^k \cdot \log(f_\theta(r_{t-k-1})) \Big| = \beta^k \cdot \Big| \log(f_\theta(r_{t-k-1})) \Big| \leq \beta^k \cdot C_{max}$$

Moreover, we know that if $0 < \beta < 1$, then the geometric series

$$\sum_{k=0}^{\infty} \beta^k = \frac{1}{1-\beta}$$

converges. Hence we conclude with the comparison test for absolute convergence of series that if $0 < \beta < 1$, then the weighted geometric series in (3.6) converges as well.

For the reverse implication, we prove that if $\beta \geq 1$ then (3.6) diverges. According to the n-th term test, the series does not converge if

$$\lim_{k \to \infty} \beta^k \cdot \log(f_\theta(r_{t-k-1})) \neq 0.$$

However, the term $\log(f_\theta(r_{t-k-1}))$ does not converge to 0. This is because there exists an infinite sequence $(l) = (l_1, l_2, ...) \in \mathbb{N}_{>0}^{\infty}$ with $\log(f_\theta(r_{t-l_i-1})) < 0$ and $\liminf_{k\to\infty} \log(f_\theta(r_{t-k-1})) < 0$. Since $\beta^k \geq 1$, we know that the squence $\beta^k \cdot \log(f_\theta(r_{t-k-1}))$ does not converge to 0 which completes the proof.

A.4 Stationarity

Before proving the proposition, we introduce the following notations:

Remark A.4.1. *For $s, t \in \mathbb{Z}$*

a) A subsequence of a time series, which contains the values from s to t:

$$_s(a)_t := (a_s, ..., a_t),$$

b) The value of a time series at t given the sequence started at s with value $a_s = a_0$:

$$_s a_t := a_t\big|_{a_s=a_0}.$$

The proof will be divided into two parts. First, we show that the PDF $g^i_{t+k}(h;\theta)$ of the cross-entropy process $_{-\infty}H^i_{t+k}(\theta)$ for value h at time $t+k$ is independent of the time shift k. The second part proves that the condition for strict stationarity is satisfied in the unconditional MCECD model.

We know that the cross-entropy process $_{-\infty}H^i_t(\theta)$ with PDF $g^i_t(h;\theta)$ and CDF $G^i_t(h;\theta)$ is strictly stationary if and only if the joint distribution is invariant over time.

$$G^i(h_{t_1}, ..., h_{t_u}; \theta) = G^i(h_{t_1+k}, ..., h_{t_u+k}; \theta),$$

where $t_1 < ... < t_u \in \mathbb{Z}$ is a arbitrary set of selected time points, and $k \in \mathbb{N}_{>0}$ is the time shift parameter.

Part I: With lemma 3.4.1, the unconditional cross-entropy process converges for every $t \in \mathbb{Z}$ if and only if all $\beta_i < 1$. On the other hand, with condition (3.4) $\beta_i = 1$ implies $\alpha_0 = 0$ and $\alpha_1 = 0$. This directly yields $_{-\infty}H^i_t(\theta) = 0$. We conclude that the unconditional cross-entropy process converges for every arbitrary selection of α_i which satisfies the non-negativity and standardization condition in (3.4). Due to proposition 3.3.4, $_{\infty}H^i_t(\theta)$ is $_{-\infty}\mathcal{F}_{t-1}$-predictable, where the filtration is defined by

${}_{-\infty}\mathcal{F}_t = \sigma\big(\{\epsilon_s | s \in \mathbb{Z} \text{ and } s \leq t\}\big)$. Conditioning the cross-entropy process at time t by the innovation path ${}_{-\infty}(\epsilon)_{t-1} = y$ yields a deterministic term

$$ {}_{-\infty}H_t^i(\theta)\Big|_{-\infty(\epsilon)_{t-1}}. $$

With the law of the total probability, the PDF for the cross-entropy value h at time t equals the integral over all PDF values of the innovation paths ${}_{-\infty}(\epsilon)_{t-1} = y$ leading to ${}_{-\infty}H_t^i(\theta) = h$. The set of all these innovation paths y will be denoted by $Y_t^i(\theta, h) = \{y \in \mathbb{R}^\infty | {}_{-\infty}H_t^i(\theta)\big|_{-\infty(\epsilon)_{t-1}=y} = h\}$

$$ g_t^i(h;\theta) = \int\limits_{y \in Y_t^i(\theta,h)} f_{t-1}(y)dy, $$

where $f_{t-1}(y)$ is the joint PDF of an infinite history white noise process at time $t-1$. From this we conclude

$$ f_{t-1}(y) = \prod_{s=-\infty}^{t-1} f_{\epsilon_s}(y_s) = \prod_{s=-\infty}^{t-1} f_{\epsilon_1}(y_s) = \prod_{s=-\infty}^{t-1} f_{\epsilon_{s+1}}(y_s) = f_t(y). $$

Since y is infinitely dimensional, it holds that if the innovation path ${}_{-\infty}(\epsilon)_{t-1} = y$ leads to ${}_{-\infty}H_t^i(\theta) = h$, then ${}_{-\infty}(\epsilon)_{t-2} = y$ leads to ${}_{-\infty}H_{t-1}^i(\theta) = h$. In other words, the term

$$ {}_{-\infty}H_t^i(\theta)\Big|_{-\infty(\epsilon)_{t-1}=y} $$

does not depend on t ceteris paribus. The inherent condition for this independence is that all parameters of the MCECD model—α_i, θ_0, and $\bar{\theta}$—and the distributional assumption are time-invariant. This yields, by definition, $Y_t^i(\theta, h) = Y_{t-1}^i(\theta, h)$. Moreover,

$$ g_t^i(h;\theta) = \int\limits_{y \in Y_t^i(\theta,h)} f_{t-1}(y)dy = \int\limits_{y \in Y_{t-1}^i(\theta,h)} f_{t-2}(y)dy = g_{t-1}^i(h;\theta), $$

which proves that $g_{t+k}^i(h;\theta)$ is independent of k.

Part II: With the law of the total probability for continuous random variables, we can write

$$\begin{aligned}
&G^i(h_{t_1}, ..., h_{t_u}; \theta) \\
&= \int\limits_X \int\limits_Y G^i(h_{t_1}, ..., h_{t_u}; \theta | {}_{-\infty}H^i_{t_1-1}(\theta) = x,\ {}_{t_1-1}(\epsilon)_{t_u-1} = y) \\
&\qquad\qquad \cdot g^i_{(t_1-1)}(x; \theta) \cdot f_{(t_u-1)}(y)\, dy\, dx.
\end{aligned}$$

If the cross-entropy ${}_{-\infty}H^i_{t_1-1}(\theta)$ one period before t_1 and the innovation path from $t_1 - 1$ till $t_u - 1$ are known, then the successive cross-entropy value ${}_{-\infty}H^i_{t_j}(\theta)$ with $j \in 1, ..., u$ are deterministic functions due to proposition 3.3.4. This yields

$$G^i(h_{t_1}, ..., h_{t_u}; \theta | \bullet) = \begin{cases} 1 & : \quad {}_{-\infty}H^i_{t_j}(\theta) \leq h_{t_j} \text{ for } j \in 1, ..., u \\ 0 & : \quad \text{else.} \end{cases}$$

The innovations process is assumed to be white noise and hence strictly stationary. Thus, its PDF is invariant over time

$$f_{(t_u-1)}(y) = f_{(t_u-1+k)}(y).$$

From Part I we also know that the distribution of ${}_{-\infty}H^i_{t_j}(\theta)$ is time-invariant

$$g^i_{(t_1-1)}(h; \theta) = g^i_{(t_1-1+k)}(h; \theta).$$

The following calculations conclude the proof for the stationarity of ${}_{-\infty}H^i_t(\theta)$

$$\begin{aligned}
&G^i(h_{t_1}, ..., h_{t_u}; \theta) \\
&= \int\limits_X \int\limits_Y G^i(h_{t_1}, ..., h_{t_u}; \theta | {}_{-\infty}H^i_{t_1-1}(\theta) = x,\ {}_{t_1-1}(\epsilon)_{t_u-1} = y) \\
&\qquad\qquad \cdot g^i_{(t_1-1)}(x; \theta) \cdot f_{(t_u-1)}(y)\, dy\, dx \\
&= \int\limits_X \int\limits_Y G^i(h_{t_1+k}, ..., h_{t_u+k}; \theta | {}_{-\infty}H^i_{t_1-1+k}(\theta) = x,\ {}_{t_1-1+k}(\epsilon)_{t_u-1+k} = y) \\
&\qquad\qquad \cdot g^i_{(t_1-1+k)}(x; \theta) \cdot f_{(t_u-1+k)}(y)\, dy\, dx \\
&= G^i(h_{t_1+k}, ..., h_{t_u+k}; \theta).
\end{aligned}$$

From equation (3.2), we know that the relation between the cross-entropy ${}_{-\infty}H^i_t(\theta)$ and the optimal parameter vector $\theta_{t,i}$ is deterministic. Moreover, it is also independent of t. Hence we conclude that the optimal parameter process $(\theta_t)_t$ of an unconditional MCECD model—as a time-invariant, deterministic transform of the (strictly) stationary cross-entropy process—is strictly stationary.

A.5 Equivalence of Vola-MCECD and GARCH

Under the assumption of only one ($m = 1$) time-varying parameter $\theta_t = \sigma_t$ we can rewrite equation (3.3):

$$\begin{aligned} H_t(\sigma) &= \alpha_0 \cdot \log(f_{\mu,\sigma}(\bar{x})) + \alpha_1 \cdot \log(f_{\mu,\sigma}(r_{t-1})) + \alpha_2 \cdot H_{t-1}(\sigma) \\ H_1(\sigma) &= \log(f_{\mu,\sigma}(x_0)), \end{aligned}$$

where $\bar{x}$ and x_0 are scalars and $f_{\mu,\sigma}(x)$ represents the PDF of the normal distribution. Using the iterative formula in (3.5) yields for $t \in \mathbb{N}_{>1}$

$$\begin{aligned} H_t(\sigma) = {\beta_1}^{t-1} \cdot \log(f_{\mu,\sigma}(x_0)) &+ \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_0 \log(f_{\mu,\sigma}(\bar{x})) \\ &+ \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_1 \log(f_{\mu,\sigma}(r_{t-s})), \end{aligned}$$

where $\beta_1 = \alpha_2$. Furthermore, the log-likelihood of the normal distribution $\mathrm{N}(\mu, \sigma^2)$ can be derived explicitly

$$\log(f_{\mu,\sigma}(x)) = -0.5 \log(2\pi) - \log(\sigma) - 0.5 \cdot \frac{(x-\mu)^2}{\sigma^2}. \tag{A.1}$$

We prove the proposition by applying the iterative formula in (3.5) to the definition of the optimal parameter process in (3.2)

$$\sigma_t = \operatorname*{argmin}_{\sigma} -H_t(\sigma).$$

The first-order optimality for $t > 1$ leads to

$$\begin{aligned}\frac{\partial H_t(\sigma)}{\partial\sigma}\Big|_{\sigma_t} &= {\alpha_2}^{t-1}\frac{\partial\log(f_{\mu,\sigma}(x_0))}{\partial\sigma}\Big|_{\sigma_t}\\ &\quad+\sum_{s=1}^{t-1}{\alpha_2}^{s-1}\alpha_0\frac{\partial\log(f_{\mu,\sigma}(\bar{x}))}{\partial\sigma}\Big|_{\sigma_t}\\ &\quad+\sum_{s=1}^{t-1}{\alpha_2}^{s-1}\alpha_1\frac{\log(f_{\mu,\sigma}(r_{t-s}))}{\partial\sigma}\Big|_{\sigma_t}\\ &= 0.\end{aligned}$$

The first derivative of the Gaussian log-likelihood function with respect to the variance parameter σ is

$$\frac{\partial\log(f_{\mu,\sigma}(x))}{\partial\sigma}\Big|_{\sigma_t} = -\frac{1}{\sigma_t}+\frac{(x-\mu)^2}{\sigma_t^3},$$

which yields

$$\begin{aligned}\frac{\partial H_t(\sigma)}{\partial\sigma}\Big|_{\sigma_t} &= {\alpha_2}^{t-1}\cdot\Big(-\frac{1}{\sigma_t}+\frac{(x_0-\mu)^2}{\sigma_t^3}\Big)\\ &\quad+\sum_{s=1}^{t-1}{\alpha_2}^{s-1}\cdot\alpha_0\cdot\Big(-\frac{1}{\sigma_t}+\frac{(\bar{x}-\mu)^2}{\sigma_t^3}\Big)\\ &\quad+\sum_{s=1}^{t-1}{\alpha_2}^{s-1}\cdot\alpha_1\cdot\Big(-\frac{1}{\sigma_t}+\frac{(r_{t-s}-\mu)^2}{\sigma_t^3}\Big)\\ &= 0.\end{aligned}$$

After basic calculations, we derive

$$\begin{aligned}&\sigma_t^2\Big(\alpha_0\sum_{s=1}^{t-1}{\alpha_2}^{s-1}+{\alpha_2}^{t-1}+\sum_{s=1}^{t-1}{\alpha_2}^{s-1}\alpha_1\Big) \qquad\qquad (A.2)\\ &=\sum_{s=1}^{t-1}{\alpha_2}^{s-1}\alpha_0(\bar{x}-\mu)^2+\sum_{s=1}^{t-1}{\alpha_2}^{s-1}\alpha_1(r_{t-s}-\mu)^2+{\alpha_2}^{t-1}(x_0-\mu)^2.\end{aligned}$$

With shifted summation limits and the formula for the geometric series, we simplify the term in the first brackets

$$\begin{aligned}
&\alpha_0 \sum_{s=1}^{t-1} {\alpha_2}^{s-1} + {\alpha_2}^{t-1} + \alpha_1 \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \\
&= \alpha_0 \sum_{s=0}^{t-2} {\alpha_2}^{s} + {\alpha_2}^{t-1} + \alpha_1 \sum_{s=0}^{t-2} {\alpha_2}^{s} \\
&= \alpha_0 \frac{1-\alpha_2^{t-1}}{1-\alpha_2} + {\alpha_2}^{t-1} + \alpha_1 \frac{1-\alpha_2^{t-1}}{1-\alpha_2}.
\end{aligned}$$

Due to $\alpha_0 + \alpha_1 + \alpha_2 = 1$, it holds that

$$\begin{aligned}
&\alpha_0 \sum_{s=1}^{t-1} {\alpha_2}^{s-1} + {\alpha_2}^{t-1} + \alpha_1 \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \\
&= \frac{\alpha_0 + \alpha_1}{\alpha_0 + \alpha_1} + \frac{\alpha_2^{t-1}(1-\alpha_0-\alpha_1-\alpha_2)}{1-\alpha_2} \\
&= 1 + 0 = 1.
\end{aligned}$$

With this result, we can rewrite equation (A.2)

$$\begin{aligned}
\sigma_t^2 &= \alpha_0 \cdot \Big((\bar{x}-\mu)^2 + \alpha_2 \cdot \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot (\bar{x}-\mu)^2 \Big) \\
&\quad + \alpha_1 \cdot \Big((r_{t-1}-\mu)^2 + \alpha_2 \cdot \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot (r_{t-s-1}-\mu)^2 \Big) \\
&\quad + \alpha_2 \cdot {\alpha_2}^{t-2} \cdot (x_0-\mu)^2 \\
&= \alpha_0 \cdot (\bar{x}-\mu)^2 + \alpha_1 \cdot (r_{t-1}-\mu)^2 \\
&\quad + \alpha_2 \cdot \Big[\sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_0 (\bar{x}-\mu)^2 \\
&\quad + \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_1 (r_{t-s-1}-\mu)^2 + {\alpha_2}^{t-2} \cdot (x_0-\mu)^2 \Big].
\end{aligned} \tag{A.3}$$

σ_{t-1} can as well be calculated using equation (A.2)

$$\begin{aligned}\sigma_{t-1}^2 =& \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_0(\bar{x}-\mu)^2 \\ &+ \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_1(r_{t-s-1}-\mu)^2 + {\alpha_2}^{t-2} \cdot (x_0-\mu)^2.\end{aligned} \tag{A.4}$$

Inserting equation (A.4) in (A.3) concludes the proof

$$\begin{aligned}\sigma_t^2 &= \alpha_0 \cdot (\bar{x}-\mu)^2 + \alpha_1 \cdot (r_{t-1}-\mu)^2 + \alpha_2 \cdot \sigma_{t-1}^2 \\ &= \tilde{\alpha}_0 + \alpha_1 \cdot \epsilon_{t-1}^2 + \alpha_2 \cdot \sigma_{t-1}^2.\end{aligned}$$

For $t = 1$, it holds

$$\frac{\partial H_1(\sigma)}{\partial \sigma}\Big|_{\sigma_1} = -\frac{1}{\sigma_1} + \frac{(x_0-\mu)^2}{\sigma_1^3} = 0,$$

which directly yields

$$\sigma_1^2 = (x_0-\mu)^2.$$

A.6 Explicit Mean-Vola-MCECD dynamics

Under the assumption of time-varying mean and volatility ($m = 2$, $\theta_t = (\mu_t, \sigma_t)$), we can rewrite equation (3.3)

$$\begin{aligned}&H_t^1(\mu,\sigma) = \alpha_0 \cdot \log(f_{\mu,\sigma}(\bar{x}_1)) + \alpha_1 \cdot \log(f_{\mu,\sigma}(r_{t-1})) + (\alpha_2+\alpha_3) \cdot H_{t-1}^1(\mu,\sigma) \\ &H_1^1(\mu,\sigma) = \log(f_{\mu,\sigma}(x_{0,1}))\end{aligned}$$

and

$$\begin{aligned}&H_t^2(\mu,\sigma) = \alpha_0 \cdot \log(f_{\mu,\sigma}(\bar{x}_2)) + \alpha_2 \cdot \log(f_{\mu,\sigma}(r_{t-1})) + (\alpha_1+\alpha_3) \cdot H_{t-1}^2(\mu,\sigma) \\ &H_1^2(\mu,\sigma) = \log(f_{\mu,\sigma}(x_{0,2})),\end{aligned}$$

where $\bar{x}$ and x_0 are two-dimensional vectors and $f_{\mu,\sigma}(x)$ represents the PDF of the normal distribution.

Using the iterative formula in (3.5) yields for $t \in \mathbb{N}_{>1}$

$$\begin{aligned} H_t^1(\mu,\sigma) = {\beta_1}^{t-1} &\cdot \log(f_{\mu,\sigma}(x_{0,1})) \\ &+ \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_0 \log(f_{\mu,\sigma}(\bar{x}_1)) \\ &+ \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_1 \log(f_{\mu,\sigma}(r_{t-s})) \end{aligned}$$

and

$$\begin{aligned} H_t^2(\mu,\sigma) = {\beta_1}^{t-1} &\cdot \log(f_{\mu,\sigma}(x_{0,2})) \\ &+ \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_0 \log(f_{\mu,\sigma}(\bar{x}_2)) \\ &+ \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_2 \log(f_{\mu,\sigma}(r_{t-s})), \end{aligned}$$

where $\beta_1 = \alpha_2 + \alpha_3$ and $\beta_2 = \alpha_1 + \alpha_3$.

We prove the proposition in analogy to the proof in appendix A.5 by applying the iterative formula in (3.5) to the definition of the optimal parameter process in (3.2)

$$\begin{aligned} \mu_t &= \underset{\mu}{\operatorname{argmin}} -H_t^1(\mu,\sigma_t) \\ \sigma_t &= \underset{\sigma}{\operatorname{argmin}} -H_t^2(\mu_t,\sigma) \end{aligned} \tag{A.5}$$

The first-order optimality for the mean leads to

$$\begin{aligned} \frac{\partial H_t^1(\mu,\sigma)}{\partial \mu}\Big|_{\mu_t} &= {\beta_1}^{t-1} \frac{\partial \log(f_{\mu,\sigma}(x_{0,1}))}{\partial \mu}\Big|_{\mu_t} \\ &\quad + \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_0 \frac{\partial \log(f_{\mu,\sigma}(\bar{x}_1))}{\partial \mu}\Big|_{\mu_t} \\ &\quad + \sum_{s=1}^{t-1} {\beta_1}^{s-1} \alpha_1 \frac{\log(f_{\mu,\sigma}(r_{t-s}))}{\partial \mu}\Big|_{\mu_t} \\ &= 0. \end{aligned}$$

The partial derivative of the Gaussian log-likelihood function with respect to the mean parameter yields for $t > 1$

$$\begin{aligned}\frac{\partial H_t^1(\mu,\sigma)}{\partial \mu}\Big|_{\mu_t} &= {\beta_1}^{t-1}\Big(-\frac{x_{0,1}-\mu_t}{\sigma^2}\Big) \\ &\quad + \sum_{s=1}^{t-1} {\beta_1}^{s-1}\alpha_0 \cdot \Big(-\frac{\bar{x}_1-\mu_t}{\sigma^2}\Big) \\ &\quad + \sum_{s=1}^{t-1} {\beta_1}^{s-1}\alpha_1 \cdot \Big(-\frac{r_{t-s}-\mu_t}{\sigma^2}\Big) \\ &= 0.\end{aligned}$$

Since the log-returns are random, it holds $\sigma > 0$. After basic calculations, we derive

$$\begin{aligned}&\mu_t \cdot \Big(\alpha_0 \sum_{s=1}^{t-1} {\beta_1}^{s-1} + {\beta_1}^{t-1} + \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_1\Big) \\ &= \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_0 \cdot \bar{x}_1 + \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_1 \cdot r_{t-s} + {\beta_1}^{t-1} \cdot x_{0,1}.\end{aligned}$$

In analogy to the proof for GARCH equivalence,we know that from equation (3.4) it follows

$$\alpha_0 \sum_{s=1}^{t-1} {\beta_1}^{s-1} + {\beta_1}^{t-1} + \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_1 = 1.$$

Hence, the conditional mean is

$$\begin{aligned}\mu_t &= \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_0 \cdot \bar{x}_1 + \sum_{s=1}^{t-1} {\beta_1}^{s-1} \cdot \alpha_1 \cdot r_{t-s} + {\beta_1}^{t-1} \cdot x_{0,1} \\ &= \alpha_0 \cdot \bar{x}_1 + \alpha_1 \cdot r_{t-1} \\ &\quad + \beta_1 \cdot \left(\sum_{s=1}^{t-2} {\beta_1}^{s-1} \cdot \alpha_0 \cdot \bar{x}_1 + \sum_{s=1}^{t-2} {\beta_1}^{s-1} \cdot \alpha_1 \cdot r_{t-s} + {\beta_1}^{t-2} \cdot x_{0,1}\right),\end{aligned}$$

or written as a recursion

$$\mu_t = \alpha_0 \cdot \bar{x}_1 + \alpha_1 \cdot r_{t-1} + \beta_1 \cdot \mu_{t-1}.$$

For $t = 1$, the first-order optimality

$$\frac{\partial H_1^1(\mu,\sigma)}{\partial \mu}\Big|_{\mu_1} = -\frac{x_{0,1} - \mu_1}{\sigma^2} = 0$$

has the solution

$$\mu_1 = x_{0,1}.$$

Furthermore, we know from the proof in appendix A.5 that the solution to the first-order optimality with respect to the variance parameter σ

$$\frac{\partial H_t^2(\mu,\sigma)}{\partial \sigma}\Big|_{\sigma_t} = 0$$

is given by

$$\begin{aligned}\sigma_t^2 &= \alpha_0 \cdot (\bar{x}_2 - \mu)^2 + \alpha_1 \cdot (r_{t-1} - \mu)^2 + \alpha_2 \cdot \sigma_{t-1}^2 \\ \sigma_1^2 &= (x_{0,2} - \mu)^2.\end{aligned}$$

From equation (A.5), it follows that σ_t is contingent on μ_t. Hence in the Mean-Vola-MCECD with time-varying mean parameter μ_t, the optimal volatility process is

$$\begin{aligned}\sigma_t^2(\mu_t) &= \alpha_0 \cdot (\bar{x}_2 - \mu_t)^2 + \alpha_1 \cdot (r_{t-1} - \mu_t)^2 + \alpha_2 \cdot \sigma_{t-1}^2(\mu_t) \\ \sigma_1^2(\mu_t) &= (x_{0,2} - \mu_t)^2.\end{aligned}$$

A.7 Explicit Skew-MCECD dynamics

In order to obtain the dynamics of the skewness parameter, we start from the defining equations for the cross-entropy process of the Skew-MCECD model given in equation (5.5). The optimal parameter process induces minimum cross-entropy and hence it can be derived by

$$\beta_t = \underset{\xi}{\operatorname{argmin}} -H_t(\xi).$$

The first-order optimality with respect to the skewness parameter β based on the PDF $f_\beta(x) = f_{SEP}(x; \alpha, \sigma, \beta, \mu)$ leads to

$$\begin{aligned}\frac{\partial H_t(\beta)}{\partial \beta}\Big|_{\beta=\beta_t} &= \alpha_2^{t-1} \cdot \frac{\partial \log(f_\beta(x_0))}{\partial \beta}\Big|_{\beta=\beta_t} \\ &+ \sum_{s=1}^{t-1} \alpha_2^{s-1}\alpha_0 \cdot \frac{\partial \log(f_\beta(\bar{x}))}{\partial \beta}\Big|_{\beta=\beta_t} \\ &+ \sum_{s=1}^{t-1} \alpha_2^{s-1}\alpha_1 \cdot \frac{\partial \log(f_\beta(r_{t-s}))}{\partial \beta}\Big|_{\beta=\beta_t}.\end{aligned}$$

Since $f_\beta(x)$ is piecewise defined for $x \le \mu$ and $x > \mu$, we split the sum over the historical log-returns into two parts and use the index sets I_t^- and I_t^+ for the summation. This yields

$$\begin{aligned}\frac{\partial H_t(\beta)}{\partial \beta}\Big|_{\beta=\beta_t} &= \alpha_2^{t-1} \cdot \frac{\partial \log(f_\beta(x_0))}{\partial \beta}\Big|_{\beta=\beta_t} \\ &+ \sum_{s=1}^{t-1} \alpha_2^{s-1}\alpha_0 \cdot \frac{\partial \log(f_\beta(\bar{x}))}{\partial \beta}\Big|_{\beta=\beta_t} \\ &+ \sum_{s\in I_t^-} \alpha_2^{s-1}\alpha_1 \cdot \frac{\partial \log(f_\beta(r_{t-s}))}{\partial \beta}\Big|_{\beta=\beta_t} \\ &+ \sum_{s\in I_t^+} \alpha_2^{s-1}\alpha_1 \cdot \frac{\partial \log(f_\beta(r_{t-s}))}{\partial \beta}\Big|_{\beta=\beta_t} \\ &= 0\end{aligned}$$

The first derivative of the log-likelihood function of the SEP distribution with $\alpha = 1$ with respect to β is given by

$$\frac{\partial \log(f_\beta(x))}{\partial \beta} = \begin{cases} \left|\frac{x-\mu}{2\sigma}\right| \cdot \beta^{-2} & : \quad x \le \mu \\ -\left|\frac{x-\mu}{2\sigma}\right| \cdot (1-\beta)^{-2} & : \quad x > \mu. \end{cases}$$

For the optimality equation, this means

$$\begin{aligned}0 = \alpha_2^{t-1} \cdot &\left(\left| \frac{x_0 - \mu}{2\sigma} \right| \beta_t^{-2} \cdot \mathbf{1}_{x_0 \le \mu} - \left| \frac{x_0 - \mu}{2\sigma} \right| (1 - \beta_t)^{-2} \cdot \mathbf{1}_{x_0 > \mu} \right) \\ &+ \sum_{s=1}^{t-1} \alpha_2^{s-1} \alpha_0 \cdot \left(\left| \frac{\bar{x} - \mu}{2\sigma} \right| \beta_t^{-2} \cdot \mathbf{1}_{\bar{x} \le \mu} - \left| \frac{\bar{x} - \mu}{2\sigma} \right| (1 - \beta_t)^{-2} \cdot \mathbf{1}_{\bar{x} > \mu} \right) \\ &+ \sum_{s \in I_t^-} \alpha_2^{s-1} \alpha_1 \cdot \left| \frac{r_{t-s} - \mu}{2\sigma} \right| \beta_t^{-2} \\ &- \sum_{s \in I_t^+} \alpha_2^{s-1} \alpha_1 \cdot \left| \frac{r_{t-s} - \mu}{2\sigma} \right| (1 - \beta_t)^{-2},\end{aligned}$$

where $\mathbf{1}_{x \in A}$ is the indicator function. Sorting the terms by β_t and $(1 - \beta_t)$ yields

$$\begin{aligned}0 = &- (1 - \beta_t)^2 \cdot \Big(\alpha_2^{t-1} \cdot \left| \frac{x_0 - \mu}{2\sigma} \right| \cdot \mathbf{1}_{x_0 \le \mu} \\ &+ \sum_{s=1}^{t-1} \alpha_2^{s-1} \alpha_0 \cdot \left| \frac{\bar{x} - \mu}{2\sigma} \right| \cdot \mathbf{1}_{\bar{x} \le \mu} + \sum_{s \in I_t^-} \alpha_2^{s-1} \alpha_1 \cdot \left| \frac{r_{t-s} - \mu}{2\sigma} \right| \Big) \\ &+ \beta_t^2 \cdot \Big(\alpha_2^{t-1} \cdot \left| \frac{x_0 - \mu}{2\sigma} \right| \cdot \mathbf{1}_{x_0 > \mu} \\ &+ \sum_{s=1}^{t-1} \alpha_2^{s-1} \alpha_0 \cdot \left| \frac{\bar{x} - \mu}{2\sigma} \right| \cdot \mathbf{1}_{\bar{x} > \mu} + \sum_{s \in I_t^+} \alpha_2^{s-1} \alpha_1 \cdot \left| \frac{r_{t-s} - \mu}{2\sigma} \right| \Big).\end{aligned}$$

Applying the defining expressions for M_t^- and M_t^+, we get the following quadratic equation for β_t

$$M_t^- - 2M_t^- \beta_t + \beta_t^2 \cdot (M_t^- - M_t^+) = 0.$$

Under the condition $M_t^- \neq M_t^+$, there are two solutions to such an equation

$$\beta_t = \frac{M_t^-}{M_t^- - M_t^+} \pm \sqrt{\left(\frac{M_t^-}{M_t^- - M_t^+} \right)^2 - \frac{M_t^-}{M_t^- - M_t^+}} \,,$$

which can be rewritten as

$$\beta_t = \frac{M_t^- \mp \sqrt{M_t^- \cdot M_t^+}}{M_t^- - M_t^+} \,.$$

The parameter range for β_t is by definition of the SEP distribution restricted to the interval $(0, 1)$. Given the relations for the geometric mean

$$\min\{M_t^-; M_t^+\} < \sqrt{M_t^- \cdot M_t^+} < \max\{M_t^-; M_t^+\},$$

there is only one solution left

$$0 < \beta_t = \frac{M_t^- - \sqrt{M_t^- \cdot M_t^+}}{M_t^- - M_t^+} < 1 \ .$$

The case that $M_t^- = M_t^+$ implies that the observations are symmetric around μ. The quadratic equation then becomes linear with the solution

$$\beta_t = \frac{1}{2},$$

which is equivalent to a zero skewness.

Appendix B

Tables

Model	Years	Koponen parameters			Model parameters					
		α	β	λ	c	a	b	α_0	α_1	β_1
(MCECD)					$(\bar{\mu})$	$(\bar{\sigma})$		(α_1)	(α_2)	(α_3)
Mean-Vola-MCECD	10	0.5	-0.2160	1.9467	0.00028	0.0114		0	0.0721	0.9180
	8	0.5	-0.2324	1.9665	0.00033	0.0110		0	0.0719	0.9181
	6	0.5	-0.2460	1.6739	0.00038	0.0101		0	0.0673	0.9227
	4	0.5	-0.1944	1.4070	0.00036	0.0101		0	0.0724	0.9159
ARMA-GARCH	10	0.5	-0.2276	1.84969	0.00032	0.0994	-0.1566	1.016E-6	0.0722	0.9225
	8	0.5	-0.2901	1.87581	0.00016	-0.5924	0.5089	9.841E-7	0.0711	0.9222
	6	0.5	-0.3053	1.64002	0.00019	-0.6139	0.5179	1.023E-6	0.0692	0.9212
	4	0.5	-0.2390	1.38241	0.00022	-0.5584	0.4333	1.382E-6	0.0897	0.9034
GARCH	10	0.5	-0.2110	1.8956	0.00027			1.043E-6	0.0725	0.9220
	8	0.5	-0.2273	1.9317	0.00032			9.616E-7	0.0704	0.9232
	6	0.5	-0.2434	1.6412	0.00038			1.023E-6	0.0696	0.9208
	4	0.5	-0.1804	1.3017	0.00036			1.388E-6	0.0896	0.9036

Table B.1: MLE parameter estimates for Mean-Vola-MCECD, ARMA-GARCH, and GARCH models on daily S&P 500 log-return data ending at 06/25/2009

Model	Years	Koponen parameters			Model parameters					
		α	β	λ	c	a	b	α_0	α_1	β_1
(MCECD)					$(\bar{\mu})$	$(\bar{\sigma})$		(α_1)	(α_2)	(α_3)
Mean-Vola-MCECD	10	1.9000	-0.9237	0.3000	0.00036	0.0115		0	0.0807	0.9067
	8	1.9000	-0.9214	0.3000	0.00047	0.0119		0	0.0794	0.9106
	6	0.6634	-0.3335	2.0000	0.00047	0.0100		0	0.0647	0.9250
	4	0.5000	-0.3259	2.0000	0.00044	0.0108		0	0.0719	0.9179
ARMA-GARCH	10	1.8773	-0.9800	0.3000	0.00003	-0.9564	0.9304	1.589E-6	0.0861	0.9044
	8	1.8803	-0.9800	0.3000	0.00010	-0.8469	0.7968	1.356E-6	0.0832	0.9089
	6	0.5760	-0.4177	2.0000	0.00009	-0.8712	0.8118	1.063E-6	0.0697	0.9212
	4	0.5000	-0.4111	1.9342	0.00015	-0.7445	0.6497	1.292E-6	0.0855	0.9085
GARCH	10	1.8841	-0.8250	0.3000	0.00037			1.588E-6	0.0850	0.9055
	8	1.9000	-0.9334	0.3000	0.00046			1.295E-6	0.0802	0.9123
	6	0.5939	-0.3184	2.0000	0.00047			1.048E-6	0.0692	0.9220
	4	0.5000	-0.3150	1.9056	0.00042			1.297E-6	0.0850	0.9091

Table B.2: MLE parameter estimates for Mean-Vola-MCECD, ARMA-GARCH, and GARCH models on daily Dow Jones log-return data ending at 06/25/2009

Model	Years	Koponen parameters			Model parameters					
		α	β	λ	c	a	b	α_0	α_1	β_1
(MCECD)					$(\bar{\mu})$	$(\bar{\sigma})$		(α_1)	(α_2)	(α_3)
Mean-Vola-MCECD	10	1.8870	-0.1202	0.4621	0.00052	0.0153		0	0.0652	0.9249
	8	1.8804	-0.2453	0.3867	0.00046	0.0152		0	0.0629	0.9272
	6	0.6770	-0.2590	2.0000	0.00047	0.0139		0	0.0604	0.9296
	4	0.5000	-0.2293	1.9614	0.00051	0.0128		0	0.0647	0.9240
ARMA-GARCH	10	1.9000	-0.3488	0.3793	0.00020	-0.6345	0.5664	9.876E-7	0.0603	0.9383
	8	1.8939	-0.4243	0.3436	0.00018	-0.6502	0.5865	9.887E-7	0.0537	0.9431
	6	0.6929	-0.3341	2.0000	0.00026	-0.5052	0.4333	1.487E-6	0.0572	0.9352
	4	0.5000	-0.2985	1.8586	0.00023	-0.6182	0.5433	2.061E-6	0.0774	0.9151
GARCH	10	1.9000	-0.0966	0.3396	0.00048			9.7218E-7	0.0600	0.9387
	8	1.9000	-0.2509	0.3132	0.00043			9.8243E-7	0.0534	0.9435
	6	0.6942	-0.2671	2.0000	0.00046			1.4831E-6	0.0574	0.9351
	4	0.5000	-0.2194	1.8167	0.00049			2.0615E-6	0.0771	0.9155

Table B.3: MLE parameter estimates for Mean-Vola-MCECD, ARMA-GARCH, and GARCH models on daily Nasdaq 100 log-return data ending at 06/25/2009

Data	Years	Model	KS test	p-value	AD	AD^2	CvM
S&P 500	10	ARMA-GARCH	0	0.0328	0.1667	2.5991	0.4337
		Mean-Vola-MCECD	0	0.0933	0.1690	2.3865	0.3932
		GARCH	0	0.0864	0.1861	2.2785	0.3855
	8	ARMA-GARCH	0	0.0318	0.2530	2.8068	0.4693
		Mean-Vola-MCECD	0	0.0664	0.2485	2.2448	0.3750
		GARCH	0	0.0497	0.2770	2.1321	0.3627
	6	ARMA-GARCH	0	0.1078	0.2357	1.7363	0.2829
		Mean-Vola-MCECD	0	0.1404	0.2395	1.4132	0.2240
		GARCH	0	0.1647	0.2329	1.3737	0.2186
	4	ARMA-GARCH	0	0.1633	0.2005	1.3792	0.2214
		Mean-Vola-MCECD	0	0.2837	0.2198	1.1220	0.1584
		GARCH	0	0.2860	0.1687	0.9637	0.1491

Table B.4: Goodness-of-fit results for the S&P 500 index based on daily log-return data ending at 06/25/2009

Data	Years	Model	KS test	p-value	AD	AD^2	CvM
Dow Jones	10	ARMA-GARCH	1	0.0028	0.0806	3.9408	0.7003
		Mean-Vola-MCECD	0	0.0428	0.0625	2.4999	0.4416
		GARCH	0	0.0275	0.0679	2.6892	0.4881
	8	ARMA-GARCH	1	0.0173	0.0784	3.0054	0.5214
		Mean-Vola-MCECD	0	0.0300	0.0711	2.3447	0.4247
		GARCH	1	0.0220	0.0735	2.4405	0.4492
	6	ARMA-GARCH	0	0.2934	0.1355	1.0866	0.1817
		Mean-Vola-MCECD	0	0.4229	0.1686	0.8373	0.1348
		GARCH	0	0.4133	0.1589	0.8125	0.1370
	4	ARMA-GARCH	0	0.5644	0.1677	0.8683	0.1502
		Mean-Vola-MCECD	0	0.8089	0.2029	0.6604	0.1003
		GARCH	0	0.6583	0.1761	0.6107	0.1057

Table B.5: Goodness-of-fit results for the Dow Jones index based on daily log-return data ending at 06/25/2009

Data	Years	Model	KS test	p-value	AD	AD^2	CvM
Nasdaq 100	10	ARMA-GARCH	0	0.0274	0.0677	2.4576	0.3585
		Mean-Vola-MCECD	0	0.0927	0.0758	2.4656	0.3259
		GARCH	0	0.0427	0.0612	2.1054	0.3248
	8	ARMA-GARCH	0	0.0498	0.0786	2.1468	0.3096
		Mean-Vola-MCECD	0	0.0634	0.0671	1.9064	0.2838
		GARCH	0	0.0837	0.0636	1.7768	0.2692
	6	ARMA-GARCH	0	0.3191	0.1447	0.9276	0.1320
		Mean-Vola-MCECD	0	0.3718	0.1325	0.8151	0.1198
		GARCH	0	0.4485	0.1471	0.8243	0.1193
	4	ARMA-GARCH	0	0.3636	0.1389	0.7459	0.1172
		Mean-Vola-MCECD	0	0.7549	0.1857	0.7085	0.0987
		GARCH	0	0.6484	0.1256	0.6088	0.0973

Table B.6: Goodness-of-fit results for the Nasdaq 100 index based on daily log-return data ending at 06/25/2009

Data	Model	KS test	p-value	AD	AD^2	CvM
S&P 500	Mean-Vola-MCECD	0	0.5925	0.1870	1.3137	0.1528
	ARMA-GARCH	0	0.4656	0.1877	1.4686	0.1836
DJA	Mean-Vola-MCECD	0	0.2204	0.2431	2.1057	0.2898
	ARMA-GARCH	0	0.0798	0.2423	2.7859	0.4555
Nasdaq 100	Mean-Vola-MCECD	0	0.4574	0.1519	0.6804	0.1016
	ARMA-GARCH	0	0.4363	0.1488	0.8035	0.1170
Bank of America	Mean-Vola-MCECD	0	0.6055	0.2463	1.2235	0.1425
	ARMA-GARCH	0	0.3672	0.2302	1.7748	0.2392
ExxonMobile	Mean-Vola-MCECD	0	0.8172	0.2707	0.8798	0.0864
	ARMA-GARCH	0	0.1351	0.2568	1.8586	0.3097
General Electric	Mean-Vola-MCECD	0	0.0351	0.3077	2.6751	0.2842
	ARMA-GARCH	1	0.0225	0.3346	3.3420	0.3729

Table B.7: One-day CDF forecasting results for U.S. stocks and U.S. stock indices based on daily log-return data from 06/26/2008 to 06/24/2009 using a shifting time window for parameter inference

Data	Model	KS test	p-value	AD	AD^2	CvM
S&P 500	Vola-Skew-MCECD	0	0.0389	0.3370	1.6236	0.2973
	GARCH	0	0.0376	0.3362	1.6251	0.2975
	ARCD	0	0.0812	0.4911	1.4546	0.2425
Dow Jones	Vola-Skew-MCECD	0	0.7894	0.1432	0.6096	0.1044
	GARCH	0	0.8143	0.1381	0.5942	0.0999
	ARCD	0	0.5883	0.1341	0.9092	0.1560
Nasdaq 100	Vola-Skew-MCECD	0	0.7485	0.1081	0.5954	0.0906
	GARCH	0	0.8022	0.1111	0.5935	0.0905
	ARCD	0	0.7352	0.1282	0.6629	0.1002

Table B.8: Goodness-of-fit results for U.S. stock indices based on daily log-return data from 06/26/2005 to 06/26/2009 based on a standard Koponen distribution with parameters $\alpha = 0.5$ and $\lambda = 1.7$. Vola-Skew-MCECD results are based on the two-step approximation

Data	Model	Volatility parameters				Skewness parameters		
		c	β_0	β_1	β_2	γ_0	γ_1	γ_2
(MCECD)						$(\bar{x})$	(α_1)	(α_2)
S&P 500	Vola-Skew-MCECD	3.63E-4	1.40E-6	0.0899	0.9032	0.1559	0.0020	0.9780
	GARCH	3.63E-4	1.40E-6	0.0899	0.9032	-0.2204	-	-
	ARCD	3.73E-4	1.19E-6	0.0895	0.8997	-0.4083	0.0018	-0.9017
Dow Jones	Vola-Skew-MCECD	4.25E-4	1.29E-6	0.0850	0.9092	0.2032	0.0100	0.9820
	GARCH	4.25E-4	1.29E-6	0.0850	0.9092	-0.2916	-	-
	ARCD	4.99E-4	1.43E-6	0.0813	0.9088	-0.4228	0.0032	-0.6178
Nasdaq 100	Vola-Skew-MCECD	4.94E-4	2.07E-6	0.0773	0.9153	0.1501	0.0090	0.9800
	GARCH	4.94E-4	2.07E-6	0.0773	0.9153	-0.2119	-	-
	ARCD	5.02E-4	2.29E-6	0.0745	0.9153	-0.0438	-0.0085	0.7834

Table B.9: MLE parameter estimates for Vola-Skew-MCECD, GARCH, and ARCD models on daily U.S. stock index data from 06/26/2005 to 06/26/2009 based on a standard Koponen distribution with parameters $\alpha = 0.5$ and $\lambda = 1.7$. Vola-Skew-MCECD results are based on the two-step approximation