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

Matthias Scherer

Department of Statistics, Econometrics and Mathematical Finance, School of Economics and Business Engineering, University of Karlsruhe and KIT

Kollegium am Schloss, Bau II, 20.12, R210, Postfach 6980, D-76128, Karlsruhe, Germany

E-mail: matthias.scherer@alumni.uni-karlsruhe.de

Svetlozar T. Rachev

Chair-Professor, Chair of Statistics, Econometrics and Mathematical Finance, School of Economics and Business Engineering, University of Karlsruhe and KIT, and Department of Statistics and Applied Probability, University of California, Santa Barbara, and Chief-Scientist, FinAnalytica INC

Kollegium am Schloss, Bau II, 20.12, R210, Postfach 6980, D-76128, Karlsruhe, Germany

Tel.: +49(0721)608 - 7535

Fax.: +49(0721)608 - 3811

E-mail: rachev@statistik.uni-karlsruhe.de

Young Shin Kim

Department of Statistics, Econometrics and Mathematical Finance, School of Economics and Business Engineering, University of Karlsruhe and KIT

Kollegium am Schloss, Bau II, 20.12, R210, Postfach 6980, D-76128, Karlsruhe, Germany

E-mail: aaron.kim@statistik.uni-karlsruhe.de

Frank J. Fabozzi

Professor in the Practice of Finance, Yale School of Management

135 Prospect Street, New Haven, CT 06511 USA

E-mail: frank.fabozzi@yale.edu

Abstract Over the past decades the stylized fact of time-varying volatility in financial series has gained significant attention amongst scholars as well as practitioners. Within this context, the GARCH model has been exceptionally successful and numerous publications have revealed the empirical relevance of the model. In this paper, we introduce a new model: the minimally cross-entropic conditional density (MCECD) model which is a generalization of the GARCH(1,1) model. It is so-named because the parameter updating method is based on cross-entropy minimization rather than autoregression. Our approach is capable of explaining a conditional density, where potentially all parameters are time-varying.

JEL classifications C32, C58, G17.

Keywords General autoregressive conditional heteroskedasticity; autoregressive conditional density; cross-entropy; exponential power distribution; stable Paretian distribution; tempered stable distribution; tempered infinitely divisible distribution;

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

Introduction

Today volatility clustering is a generally accepted stylized fact in the finance literature. 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 heteroscedasticity (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 pathbreaking 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 up to date has been extended and modified to cover more sophisticated empirical facts.[1]

One way to generalize the GARCH model is to consider not only conditional volatility, but also conditional higher moments. Hansen (1994) argues 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".[2] As a consequence Hansen (1994) introduced the first GARCH-like approach to conditional density, 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.[3]

In this paper, we introduce a new model for conditional densities which includes the GARCH model as a special case.[4] 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.[5] 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 consequence, the use of ARMA-GARCH estimated parameter trajectories for conditional density models is problematic. Finally, MCECD can cope with a non-linear parameter process, 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).

Our paper is divided into three part. We first outline important concepts of parameter estimation and information theory. This is meant as a short overview of relevant terms and their relation, rather than a comprehensive summary. Then we introduce the general MCECD model and analyze special cases for conditional mean and volatility. Our focus is especially on the link between distributional assumption and parameter process, as well as the simultaneous modeling of multiple time-varying parameters. Our goal is to highlight the advantages of MCECD compared to GARCH-like models. We conclude the paper with an empirical comparison of the MCECD model we propose and the battle-tested ARMA-GARCH model with respect to explanatory power, goodness-of-fit, and forecasting quality.

1 Background

In this section, we explain the central concepts for the MCECD model. We highlight some important facts concerning the maximum likelihood estimation (MLE), present an intuitive definition of entropy and cross-entropy terms, and point out their relevance for our model.

Parameter inference using MLE goes back to the seminal work of Fisher (1922). Decades later Godambe (1960) proved that of all estimating functions the MLE is optimal with respect to efficiency.[6] Compared to other inference methods, such as (generalized) methods of moments (GMM), it does not depend on moment estimators. Given a probability density function (PDF) $f_\theta : \mathbb{R} \to [0, 1]$ with parameter vector θ and the observation vector x, the optimal 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.

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 estimated parameters: $\mu = x_1$ 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

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

Hence the higher the entropy $H(X)$ is, the higher the disorder, or the lesser the available information.

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

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

Note that this term is closely related to the Kullback-Leibner (KL) divergence (also known as relative entropy, Kullback (1959)). From this definition, we can see 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 maximation 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.[7] 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^{Start}.

Maximum entropy and minimum cross-entropy are already an integral part of several important concepts and applications. Jaynes (1957) introduced the principle of maximum entropy, which is applied in the empirical likelihood method by Owen (1988) for parameter inference. Closest to our approach is the principle of minimum discrimination information (MDI) by Kullback (1959)—sometimes also called principle of minimum cross-entropy (MCE). MDI postulates that given new facts, a new distribution should be chosen which is as close (KL divergence) as possible to the original distribution, so that the information gain by new data is as small as possible. For our model, we will apply the cross-entropy minimization to describe the parameter dynamics for the conditional density. In this sense, MCECD is defined as the likelihood-based alternative for ARCD, just as MLE is the likelihood-based alternative for GMM.

2 The MCECD model

2.1 The definition

In the following, we introduce our MCECD model for a financial return series. We assume a probability space $(\mathbb{R}, \wp(\mathbb{R}), P)$, where $\wp(\mathbb{R})$ denotes the Borel σ-algebra. Furthermore, there exists a stochastic process $\epsilon : T \times \mathbb{R} \to \mathbb{R}$ inducing the natural filtration $\mathcal{F}_t = \mathcal{F}_t^{\epsilon} = \wp\big(\{\epsilon_s | s \leq t\}\big)$.[8] 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 cumulative distribution

function (CDF) $F_\theta : \mathbb{R} \to [0,1]$ contains the Gaussian as a special case $\theta = \theta_{Norm}$.

Remark 2.1 *We use the notation (υ, ω_{-i}) to refer to a vector of the form*

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

Definition 2.2 ***(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*

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

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 $H_t(\theta) = (H_t^1(\theta), ..., H_t^m(\theta))$

$$\theta_{t,i} = \underset{\xi\in\Theta_i}{\operatorname{argmin}} -H_t^i(\xi, \theta_{t,-i}). \tag{2}$$

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

$$\begin{aligned} H_t^i(\theta) &:= \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_{t-1})) + \beta_i \cdot H_{t-1}^i(\theta) \\ H_1^i(\theta) &:= \log(f_\theta(x_{0,i})) \,, \end{aligned} \tag{3}$$

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

$$\alpha_i \leq 0 \quad \textit{and} \quad \sum_{i=0}^{m+1} \alpha_i = 1. \tag{4}$$

The α_i 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 defintion 2.2 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 explicitly analyze the multiple parameter case for time-varying mean and volatility.

Proposition 2.3 *Let $(H_t(\theta))_{t\in\mathbb{N}_{>0}}$ be a general cross-entropy process defined in 2.2, then for each $t \in \mathbb{N}_{>1}$ the following iterative formula holds for every component $i \in \{1, ..., m\}$*

$$H_t^i(\theta) = {\beta_i}^{t-1} \cdot \log(f_\theta(x_{0,i})) + \sum_{s=1}^{t-1} {\beta_i}^{s-1} \cdot \alpha_0 \log(f_\theta(\bar{x}_i)) + \sum_{s=1}^{t-1} {\beta_i}^{s-1} \cdot \alpha_i \log(f_\theta(r_{t-s})). \tag{5}$$

Proof. Proof by induction. See appendix A.

Proposition 2.4 *Given a MCECD model defined as in 2.2 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 B.

Remark 2.5 *The MCECD models from 2.2 can be defined for abitrary combinations of time-varying parameters. In order to specify a distinct model, we introduce the following nomenclature: The names of the time-varying moments, accounted for 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 Sections 2.3 and 2.4 we define and analyze the Vola-MCECD and Mean-Vola-MCECD models more thoroughly.

2.2 Stationarity of the MCECD model

A key feature of models for financial time series is stationarity, which claims, loosely 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 stems 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 for the remainder of this section on the parameter process.

Lemma 2.6 *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_∞

$$S_\infty = \sum_{k=0}^{\infty} \beta^k \cdot \log(f_\theta(r_{t-k-1})) \tag{6}$$

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

Proof. See appendix C.

For $\beta = 0$, the convergence is trivial. Within our paper we will 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 2.6, we define a MCECD process with infinite history, the unconditional MCECD analogous to Nelson (1990).

Definition 2.7 ***(Unconditional MCECD)*** *Let $F_{\theta_t}(x)$, ϵ_t, r_t, and θ_t be defined as in 2.2, 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*

$$ {}_{-\infty}H_t^i(\theta) = \sum_{s=1}^{\infty} {\beta_i}^{s-1} \cdot \alpha_0 \log(f_\theta(\bar{x}_i)) + \sum_{s=1}^{\infty} {\beta_i}^{s-1} \cdot \alpha_i \log(f_\theta(r_{t-s})). \tag{7} $$

The results of 2.4 and 2.6 lead us directly to the following proposition.

Proposition 2.8 *Given an unconditional MCECD model defined in 2.7, then the resulting m-dimensional parameter process $(\theta_t)_t$ is (strictly) stationary.*

Proof. See appendix D.

With this proposition, we can conclude that the return process generated by MCECD according to equation (1) is stationary.

2.3 The Vola-MCECD model

2.3.1 A comparison to GARCH

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

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 $f_{\mu,\sigma}(x)$.*

The resulting conditional Vola-MCECD model takes the form

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

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 2.7.

Proposition 2.9 *Given a Vola-MCECD model which satisfies assumption A1, then there exists an equivalent GARCH model with specification*

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

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* $\sigma_0^2 := (x_0 - \mu)^2$*.*

In particular, both models govern the same volatility process

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

Proof. See appendix E.

Remark 2.10 *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 2.9, 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 different distributions in order to account for special features of the return data. Bollerslev and Wooldridge (1992) showed that even if the assumption of normality is violated, the normal distribution can be used for inference of GARCH parameters. This procedure is called quasi-maximum likelihood (QMLE) and leads to consistent estimators. Given proposition 2.9, 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.

2.3.2 Models with non-Gaussian innovations

Since Mandelbrot (1963) and Fama (1963), it is a generally accepted fact that 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) discuss the stationarity problem for this distributional assumption within the GARCH framework and propose a solution within the empirically relevant parameter range. 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),[9] and the α-stable distribution $S_\alpha(C,\beta,\mu)$. For the SEP, we derive an explicit Vola-MCECD model and outline its differences relative to the Gaussian case. For $S_\alpha(C,\beta,\mu)$, we analyze the induced parameter process based on numerical analysis only, due to the lack of a closed-form expression for its PDF.

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

Definition 2.11 *Given a parameter process $(\theta_t)_{t\in\mathbb{Z}}$ of a MCECD model defined in 2.2. Then the i-th component of the parameter process is called linear autoregressive, if $\theta_{t,i}$ follows the equations*

$$\begin{aligned}\theta_{t,i}^{\gamma} &= \alpha_0 \cdot \bar{\theta}_i^{\gamma} + \alpha_i \cdot g_{\theta_{-i}}(r_{t-1}) + \beta_i \cdot \theta_{t-1,i}^{\gamma} \\ \theta_{1,i}^{\gamma} &= \theta_{0,i}^{\gamma},\end{aligned} \tag{8}$$

where $g_{\theta_{-i}}(x)$ is the ML 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

$$\theta_{t,i}^{\gamma} = \beta_i^{t-1} \cdot \theta_{0,i}^{\gamma} + \alpha_0 \sum_{s=1}^{t-1} \beta_i^{s-1} \cdot \bar{\theta}_i^{\gamma} + \alpha_i \sum_{s=1}^{t-1} \beta_i^{s-1} \cdot g_{\theta_{-i}}(r_{t-s}). \tag{9}$$

Our findings in propostion 2.9 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 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 $\frac{X-\mu}{\sigma}$ it holds that

$$f_\theta(x) = \frac{1}{\sigma} \cdot f_{\theta^{Std}}\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma} \cdot f\left(\frac{x-\mu}{\sigma}\right), \tag{10}$$

where θ^{Std} denotes the parameter vector of the standardized random variable.

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

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

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 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

$$\begin{aligned}\frac{\partial H_t(\sigma)}{\partial \sigma}\Big|_{\sigma_t} &= {\alpha_2}^{t-1} \cdot \left(-\frac{1}{\sigma} - \frac{f'(\frac{x_0-\mu}{\sigma})}{f(\frac{x_0-\mu}{\sigma})} \cdot \frac{x_0-\mu}{\sigma^2}\right) + \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \cdot \alpha_0 \cdot \left(-\frac{1}{\sigma} - \frac{f'(\frac{\bar{x}-\mu}{\sigma})}{f(\frac{\bar{x}-\mu}{\sigma})} \cdot \frac{\bar{x}-\mu}{\sigma^2}\right) \\ &\quad + \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \cdot \alpha_1 \cdot \left(-\frac{1}{\sigma} - \frac{f'(\frac{r_{t-s}-\mu}{\sigma})}{f(\frac{r_{t-s}-\mu}{\sigma})} \cdot \frac{r_{t-s}-\mu}{\sigma^2}\right) \\ &= 0.\end{aligned} \tag{11}$$

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

Proposition 2.12 *Consider a distribution with PDF $f_{\theta_t}(x)$ that satisfies equation (10) and that 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*

$$-\frac{f'_{\theta_t^{Std}}(x)}{f_{\theta_t^{Std}}(x)} = k(\operatorname{sign}(x), \theta_t^{Std}) \cdot x^{\gamma-1}, \tag{12}$$

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 (11) because the equation can be solved with a linear autoregressive form for σ_t as given in equation (9), if and only if the ratio $-f'(\frac{x-\mu}{\sigma})/f(\frac{x-\mu}{\sigma})$ is ceteris paribus piecewise proportional to $\left(\frac{x-\mu}{\sigma}\right)^\gamma$ in the intervals $x < \mu$ and $x > \mu$.

We 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 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 (10). Hence, proposition 2.12 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 ratio of the first derivative and PDF satisfies equation (12)

$$-\frac{f'_{SEP}(x; \alpha, 1, \beta, 0)}{f_{SEP}(x; \alpha, 1, \beta, 0)} = \begin{cases} \frac{1}{(2\beta)^\alpha} |x|^{\alpha-1} & : \quad x < 0 \\ \frac{1}{(2(1-\beta))^\alpha} \cdot |x|^{\alpha-1} & : \quad x > 0, \end{cases} \tag{13}$$

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 join (13) with the first-order optimality in (11). Solving for σ_t then yields

$$\sigma_t^\alpha = \alpha_2^{t-1} \cdot \sigma_0^\alpha + \alpha_0 \sum_{s=1}^{t-1} \alpha_2^{s-1} \cdot \bar{\sigma}^\alpha + \alpha_1 \sum_{s=1}^{t-1} \alpha_2^{s-1} \cdot k(\text{sign}(r_{t-s} - \mu), \beta) \cdot |r_{t-s} - \mu|^\alpha, \tag{14}$$

with $\sigma_0^\alpha := k(\text{sign}(x_0 - \mu), \beta) \cdot |x_0 - \mu|^\alpha$ and $\bar{\sigma}^\alpha := k(\text{sign}(\bar{x} - \mu), \beta) \cdot |\bar{x} - \mu|^\alpha$. 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

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

The value of $k(\text{sign}(x - \mu), \beta)$ 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 $e_t < 0$. This behavior directly stems 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, but much smaller impact on the volatility estimator as the empirically observed leverage effect.[10] 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 (14) 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 (14).

The stable Paretian distribution is defined by its characteristic function $\phi(t; \alpha, \beta, C, \mu)$, a Fourier transform of its PDF

$$\phi(t; \alpha, \beta, C, \mu) = \exp\{it\mu - C|t|^\alpha (1 - i\beta \, \text{sign}(t) z(t, \alpha))\}, \tag{15}$$

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

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

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 argumenation, we use dispersion and volatility process as synonyms. It is common knowledge that the stable Paretian PDF $f(x; \alpha, \beta, C, \mu)$

satisfies the standardization condition in (10), but does not have a closed-form expression. In order to apply proposition 2.12, we need to analyze the ratio $-f'(x;\alpha,\beta,1,0)/f(x;\alpha,\beta,0,1)$ numerically.

[FIGURE 1 ABOUT HERE]

Figure 1 shows that the ratio is not proportional to x^γ, except for the Gaussian case $\alpha = 2$. Therefore the dispersion parameter process of an stable Paretian driven Vola-MCECD model is not linear autoregressive and hence equation (14) for the SEP does not describe the parameter process when the underlying distribution is stable Paretian.

Looking at the volatility dynamics under non-Gaussian assumptions, there are 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, as shown in the stable Paretian case.

2.4 The Mean-Vola-MCECD model

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

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 process for the mean component

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

$$H_1^1(\mu,\sigma) = \log(f_{\mu,\sigma}(x_{0,1}))$$

and for the volatility component

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

$$H_1^2(\mu,\sigma) = \log(f_{\mu,\sigma}(x_{0,2})).$$

$\bar{x}$ and x_0 are two-dimensional vectors. The unconditional Mean-Vola-MCECD model can be defined analogously to 2.7. With the assumption of a Gaussian conditional distribution, an explicit form for the dynamics of the parameter process θ_t is available.

Proposition 2.13 *Given a MCECD model which satisfies assumption A2, then the mean process μ_t follows the equations*

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

and the volatility process σ_t follows

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

Proof. See appendix F.

In the Mean-Vola-MCECD model, the volatility dynamics given in proposition 2.13 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.

3 Empirical comparison of Mean-Vola-MCECD and ARMA-GARCH

In this section, we empirically compare Mean-Vola-MCECD to its autoregression-based alternative, the ARMA-GARCH process. The dynamics of ARMA-GARCH are given by

$$r_t = a \cdot e_{t-1} + b \cdot r_{t-1} + c + e_t \tag{16}$$
$$r_1 = c,$$

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

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

We analyze the models along three dimensions: (1) simultaneously modeling of time-varying mean and volatility, (2) distinguishing time-varying from time-invariant trajectories, and (3) forecasting properties. Concerning goodness-of-fit, we 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 use Bollerslev's QMLE method, whereby the innovation process is governed by the Koponen distribution (Koponen (1995)).

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 resort to daily log-returns of U.S. stock indices and several individual U.S. stocks from the Dow Jones Average. 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. Our selection is such that it includes the Dotcom Collapse in April 2000 and the U.S. financial crisis that began in September 2008.

3.1 Simultaneous modeling of time-varying moments

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 an underlying uniformly distributed processes $(p_t^{\mu})_{t\in\mathbb{N}_{>0}}$ and $(p_t^{\sigma})_{t\in\mathbb{N}_{>0}}$ using the following specifications

$$\mu_t = \begin{cases} \mu_{t-1} + 0.001 & : \quad p^{\mu} \geq 0.9 \\ \mu_{t-1} - 0.001 & : \quad p^{\mu} \leq 0.1 \end{cases} \qquad \sigma_t = \begin{cases} \sigma_{t-1} \cdot 1.08 & : \quad p^{\sigma} \geq 0.75 \\ \sigma_{t-1} \cdot 0.925 & : \quad p^{\sigma} \leq 0.25. \end{cases}$$

[FIGURE 2 ABOUT HERE]

Figures 2 show that both models, ARMA-GARCH and Mean-Vola-MCECD are suitable for modeling time series with conditional mean and conditional volatility. Their approximation quality for the parameter trajectories is similar. This finding is supported by the goodness-of-fit analysis shown below:

	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

Both models yield an equivalent overall as well as tail fit.

3.2 Distinguishing between time-varying and time-invariant moments

In the following 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 1 suggest that the mean of the S&P 500 index returns is time-invariant and positive. This result is contrasted by the ARMA-GARCH estimates in Table 1. The ARMA parameters clearly suggest a time-varying component in the mean. Figure 3 illustrates the time-varying effect in the conditional mean.

[FIGURE 3 ABOUT HERE]

Furthermore, the goodness-of-fit results in Table 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 conclude that the data are characterized by a time-invariant mean. The Mean-Vola-MCECD indicates whether or not a parameter process is time-invariant. Therefore, it might be the preferred choice to obtain reliable parameter trajectories for the conditional density.

[TABLE 4 ABOUT HERE]

For the analyses based on the log-return data of DJA and Nasdaq 100 indices, parameter estimates and goodness-of-fit results again support our findings.

3.3 Quality of one-day forecasting

In a first step, we use classical Value-at-Risk (VaR) backtesting to evaluate the one-day forecasting quality of both models. We apply the Kupiec test[11] and the Lopez statistic[12] 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 exceedings over the specified quantile, whereas the Lopez statistics also considers the distance to the quantile.[13]

[TABLE 5 ABOUT HERE]

According to the results reported in Table 5, 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. Given the log-return process $(r_t)_t$ and the derived parameter process $(\theta_t)_t$, under a distributional assumption $F_\theta(x)$ we define

$$y_t := F_{\theta_t}(r_t). \tag{17}$$

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 assessed by analyzing the empirical distribution of y_t.

[TABLE 6 ABOUT HERE]

Table 6 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 individual stocks as shown in Table 6 for three 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) are used.

4 Conclusion

In this paper, we introduce a new time series model, minimally cross-entropic conditional density, for conditional density based on cross-entropy. Our model is a generalization of the seminal GARCH model. We show that MCECD can overcome problems associated with an autoregressive approach. In particular, it 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 distribution, we show that MCECD includes the power-ARCH model as a special case and that induced parameter dynamics can be non-linear, even for the volatility process.

The empirical analysis shows that Mean-Vola-MCECD leads to a slightly improved goodness-of-fit and forecasting quality while yielding similar results in the multiple time-varying parameter case. An advantage of the MCECD based model is that it needs fewer parameters, thus reducing the risk of overfitting. The most interesting characteristic is its capability to detect if a parameter process is time-varying. This feature is key in analyzing market dynamics in risk and portfolio management.

Footnotes

[1]See Bera and Higgins (1993), Duan (1997), and Christoffersen and Jacobs (2004a) for an overview of GARCH-like models as well as empirical studies.

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

[3]See Dark (2010) for an overview of empirical studies and model specifications on ARCD.

[4]We use GARCH and ARMA-GARCH as synonyms for GARCH(1,1) and ARMA(1,1)-GARCH(1,1).

[5]See Rachev and Mittnik (2000) for a solution in the stable Paretian case.

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

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

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

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

[10]See Black (1976).

[11]See Kupiec (1995).

[12]See Lopez (1998).

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

Figure legends

Figure 1 : Ratio of first derivative and PDF of a stable Paretian distribution with parameters $\beta = 0$, $C = 1$, and $\mu = 0$.

Figure 2 : Conditional mean and volatility trajectories of simulated data compared to corresponding trajectories of Mean-Vola-MCECD (top charts) and ARMA-GARCH (bottom charts)

Figure 3 : Trajectories of conditional mean (top chart) and volatility (bottom chart) for 10 years daily log-return data of S&P 500 index

References

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.

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.

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

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. (2004a). 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.

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

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.

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

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

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

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

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.

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.

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

Rachev, S. and Mittnik, S. (2000). *Stable Paretian models in finance*. John Willey & Sons, New York.

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

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

A Proof for iterative formula of general cross-entropy process

We prove this proposition by means of induction. The base case $t = 2$ directly follows from the definition 2.2

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

For the inductive step, we assume that there exists a $t \in \mathbb{N}_{>0}$ for which the equation (5) holds and write

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

Now we expand the equation and get

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

Base case and inductive step together prove the iterative formula in (5).

B Proof for predictability of cross-entropy process

In order to prove the predictability of the cross-entropy process, we need to show that $H_t^i(\theta)$ is a deterministic function of the past innovations $\epsilon_{t-1}, ..., \epsilon_1$. We do so by induction over time t. With equation (3) the base case $t = 2$ results in

$$H_2^i(\theta) = \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(r_1)) + \beta_i \cdot H_1^i(\theta).$$

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

$$H_2^i(\theta) = \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(F_{\theta_1}^{-1}(F_{\theta_{Norm}}(\epsilon_1)))) + \beta_i \cdot \log(f_\theta(x_{0,i})).$$

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

$$\theta_{1,i} = \underset{\xi \in \Theta_i}{\operatorname{argmin}} -H_1^i(\xi, \theta_{t,-i})$$

reveals that θ_1 is deterministic. Furthermore, we know that θ_{Norm} is deterministic and hence we conclude that the randomness of $\log(f_\theta(F_{\theta_1}^{-1}(F_{\theta_{Norm}}(\epsilon_1))))$ 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 $H_t^i(\theta)$ is $\mathcal{F}_{t-1}$-measurable. In analogy to the base case we conclude from

$$H_{t+1}^i(\theta) = \alpha_0 \cdot \log(f_\theta(\bar{x}_i)) + \alpha_i \cdot \log(f_\theta(F_{\theta_t}^{-1}(F_{\theta_{Norm}}(\epsilon_t)))) + \beta_i \cdot H_t^i(\theta)$$

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

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

and by assumption $H_t^i(\theta)$ 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 $H_{t+1}^i(\theta)$ is predictable.

C Proof for 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

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

This yields

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

Furthermore 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 (6) converges as well.

For the reverse implication, we prove that if $\beta \geq 1$ then (6) diverges. According to the n-th term test, the series does not convert 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.

D Proof for stationarity condition of MCECD

Before proving the proposition, we introduce the folloing notations

Remark D.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 part. 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 2.6, 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 (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 (4). Due to proposition 2.4, ${}_{\infty}H^i_t(\theta)$ is ${}_{-\infty}\mathcal{F}_{t-1}$-predictable, where the filtration is defined by ${}_{-\infty}\mathcal{F}_t = \sigma\big(\{\epsilon_s | s \in \mathbb{Z} \text{ and } s \le t\}\big)$. Conditioning the cross-entropy process at time t by the innovation path ${}_{-\infty}(\epsilon)_{t-1} = y$ yields a deterministic term

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

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 ${}_{-\infty}H^i_t(\theta) = h$. The set of all these innovation paths y will be denoted by $Y^i_t(\theta, h) = \{y \in \mathbb{R}^\infty | {}_{-\infty}H^i_t(\theta)\big|_{{}_{-\infty}(\epsilon)_{t-1}=y} = h\}$

$$ g^i_t(h;\theta) = \int_{y \in Y^i_t(\theta,h)} f_{t-1}(y)dy, $$

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 innovations path ${}_{-\infty}(\epsilon)_{t-1} = y$ leads to ${}_{-\infty}H^i_t(\theta) = h$, then ${}_{-\infty}(\epsilon)_{t-2} = y$ leads to ${}_{-\infty}H^i_{t-1}(\theta) = h$. In other words the term

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

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,

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

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

$$G^i(h_{t_1},...,h_{t_u};\theta) = \int_X \int_Y G^i(h_{t_1},...,h_{t_u};\theta|_{-\infty}H_{t_1-1}^i(\theta) = x,\ {}_{t_1-1}(\epsilon)_{t_u-1} = y) \cdot g_{(t_1-1)}^i(x;\theta) \cdot f_{(t_u-1)}(y)\, dy\, dx.$$

If the cross-entropy ${}_{-\infty}H_{t_1-1}^i(\theta)$ 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 ${}_{-\infty}H_{t_j}^i(\theta)$ with $j \in 1,...,u$ are deterministic functions due to proposition 2.4. This yields

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

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 ${}_{-\infty}H_{t_j}^i(\theta)$ is time-invariant

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

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

$$\begin{aligned} G^i(h_{t_1},...,h_{t_u};\theta) &= \int_X \int_Y G^i(h_{t_1},...,h_{t_u};\theta|_{-\infty}H_{t_1-1}^i(\theta) = x,\ {}_{t_1-1}(\epsilon)_{t_u-1} = y) \cdot g_{(t_1-1)}^i(x;\theta) \cdot f_{(t_u-1)}(y)\, dy\, dx \\ &= \int_X \int_Y G^i(h_{t_1+k},...,h_{t_u+k};\theta|_{-\infty}H_{t_1-1+k}^i(\theta) = x,\ {}_{t_1-1+k}(\epsilon)_{t_u-1+k} = y) \\ &\quad \cdot g_{(t_1-1+k)}^i(x;\theta) \cdot f_{(t_u-1+k)}(y)\, dy\, dx \\ &= G^i(h_{t_1+k},...,h_{t_u+k};\theta). \end{aligned}$$

From equation (2), we know that the relation between the cross-entropy ${}_{-\infty}H_t^i(\theta)$ 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.

E Proof for equivalence of Vola-MCECD and GARCH under normality assumption

Under the assumption of only one ($m = 1$) time-varying parameter $\theta_t = \sigma_t$ we can rewrite equation (3)

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

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 (5) yields for $t \in \mathbb{N}_{>1}$

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

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

$$\log(f_{\mu,\sigma}(x)) = -0.5 \log(2\pi) - \log(\sigma) - 0.5 \cdot \frac{(x-\mu)^2}{\sigma^2}. \tag{18}$$

We prove the proposition by applying the iterative formula in (5) to the definition of the optimal parameter process in (2)

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

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

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

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

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

After basic calculations, we derive

$$\sigma_t^2 \Big(\alpha_0 \sum_{s=1}^{t-1} {\alpha_2}^{s-1} + {\alpha_2}^{t-1} + \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \alpha_1 \Big) = \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \alpha_0 (\bar{x}-\mu)^2 + \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \alpha_1 (r_{t-s}-\mu)^2 + {\alpha_2}^{t-1} (x_0-\mu)^2. \tag{19}$$

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

$$\alpha_0 \sum_{s=1}^{t-1} {\alpha_2}^{s-1} + {\alpha_2}^{t-1} + \sum_{s=1}^{t-1} {\alpha_2}^{s-1} \cdot \alpha_1 = \alpha_0 \sum_{s=0}^{t-2} {\alpha_2}^{s} + {\alpha_2}^{t-1} + \alpha_1 \sum_{s=0}^{t-2} {\alpha_2}^{s} = \alpha_0 \frac{1-\alpha_2^{t-1}}{1-\alpha_2} + {\alpha_2}^{t-1} + \alpha_1 \frac{1-\alpha_2^{t-1}}{1-\alpha_2}.$$

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

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

With this result, we can rewrite equation (19)

$$\begin{aligned}\sigma_t^2 &= \alpha_0 \cdot \left((\bar{x} - \mu)^2 + \alpha_2 \cdot \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot (\bar{x} - \mu)^2\right) + \alpha_1 \cdot \left((r_{t-1} - \mu)^2 + \alpha_2 \cdot \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot (r_{t-s-1} - \mu)^2\right) \\ &\quad + \alpha_2 \cdot {\alpha_2}^{t-2} \cdot (x_0 - \mu)^2 \\ &= \alpha_0 \cdot (\bar{x} - \mu)^2 + \alpha_1 \cdot (r_{t-1} - \mu)^2 \\ &\quad + \alpha_2 \cdot \Big[\sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_0 (\bar{x} - \mu)^2 + \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_1 (r_{t-s-1} - \mu)^2 + {\alpha_2}^{t-2} \cdot (x_0 - \mu)^2\Big].\end{aligned} \tag{20}$$

σ_{t-1} can as well be calculated using equation (19)

$$\sigma_{t-1}^2 = \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_0 (\bar{x} - \mu)^2 + \sum_{s=1}^{t-2} {\alpha_2}^{s-1} \cdot \alpha_1 (r_{t-s-1} - \mu)^2 + {\alpha_2}^{t-2} \cdot (x_0 - \mu)^2. \tag{21}$$

Inserting equation (21) in (20) concludes the proof

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

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

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

F Proof for explicit formulas of Mean-Vola-MCECD model

Under the assumption of time-varying mean and volatility ($m = 2$, $\theta_t = (\mu_t, \sigma_t)$), we can rewrite equation (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 (5) yields for $t \in \mathbb{N}_{>1}$

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

and

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

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

We prove the proposition in analogy to proof in appendix E by applying the iterative formula in (5) to the definition of the optimal parameter process in (2)

$$\begin{aligned}\mu_t &= \underset{\mu}{\operatorname{argmin}} -H_t^1(\mu,\sigma_t) \\ \sigma_t &= \underset{\sigma}{\operatorname{argmin}} -H_t^2(\mu_t,\sigma)\end{aligned} \tag{22}$$

The first-order optimality for the mean leads to

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

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

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

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

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

In analogy to the proof for GARCH equivalence,we know that from to equation (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

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

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 E 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

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

From equation (22), it follows that σ_t is contingent on μ_t. Hence in the Mean-Vola-Model with time-varying mean parameter μ_t the optimal volatility process is

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

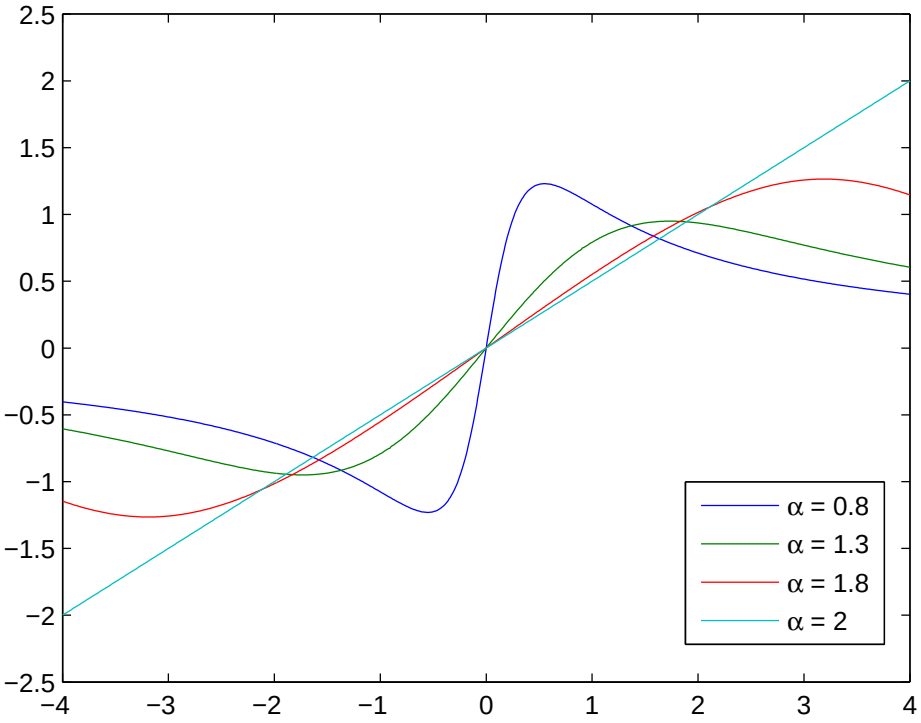

Figure 1

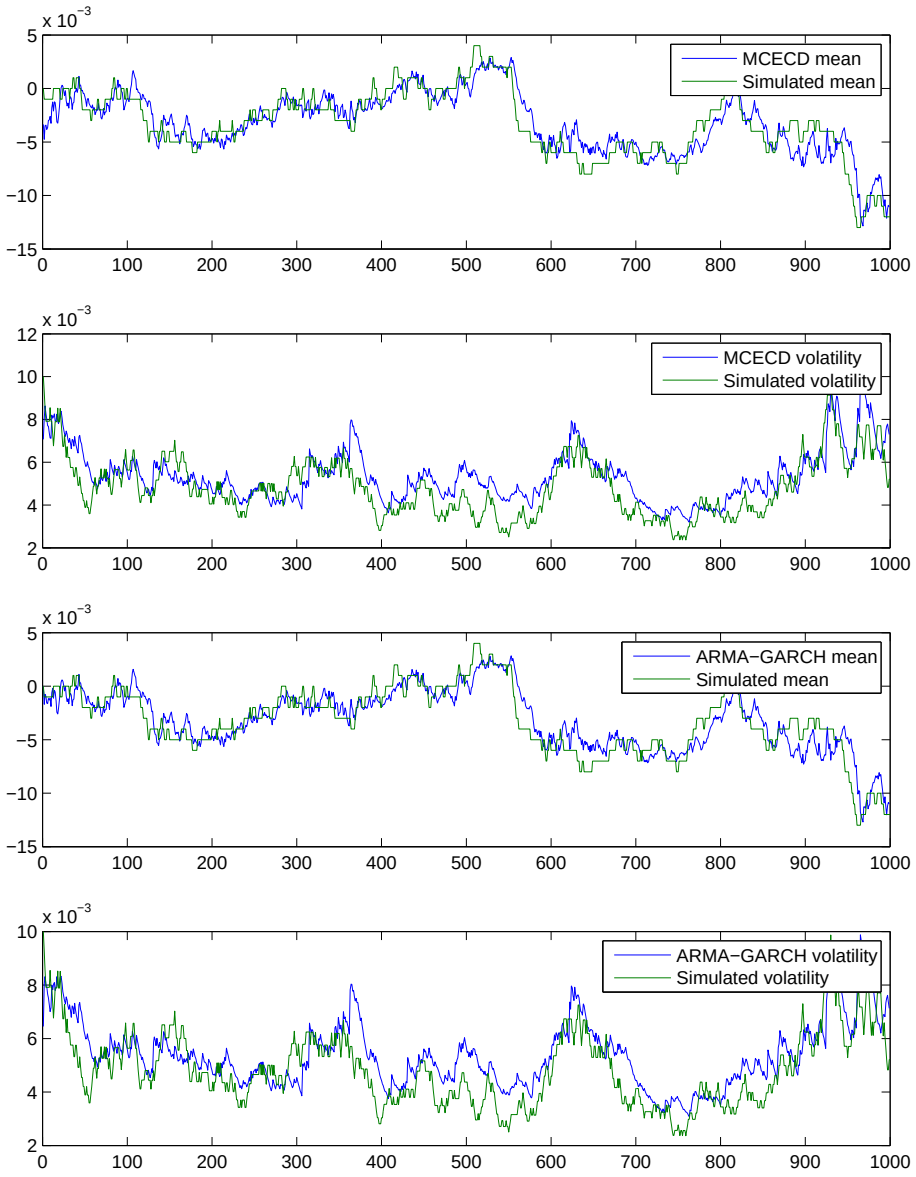
x 10^−3
MCECD mean
Simulated mean
x 10^−3
MCECD volatility
Simulated volatility
x 10^−3
ARMA−GARCH mean
Simulated mean
x 10^−3
ARMA−GARCH volatility
Simulated volatility

Figure 2

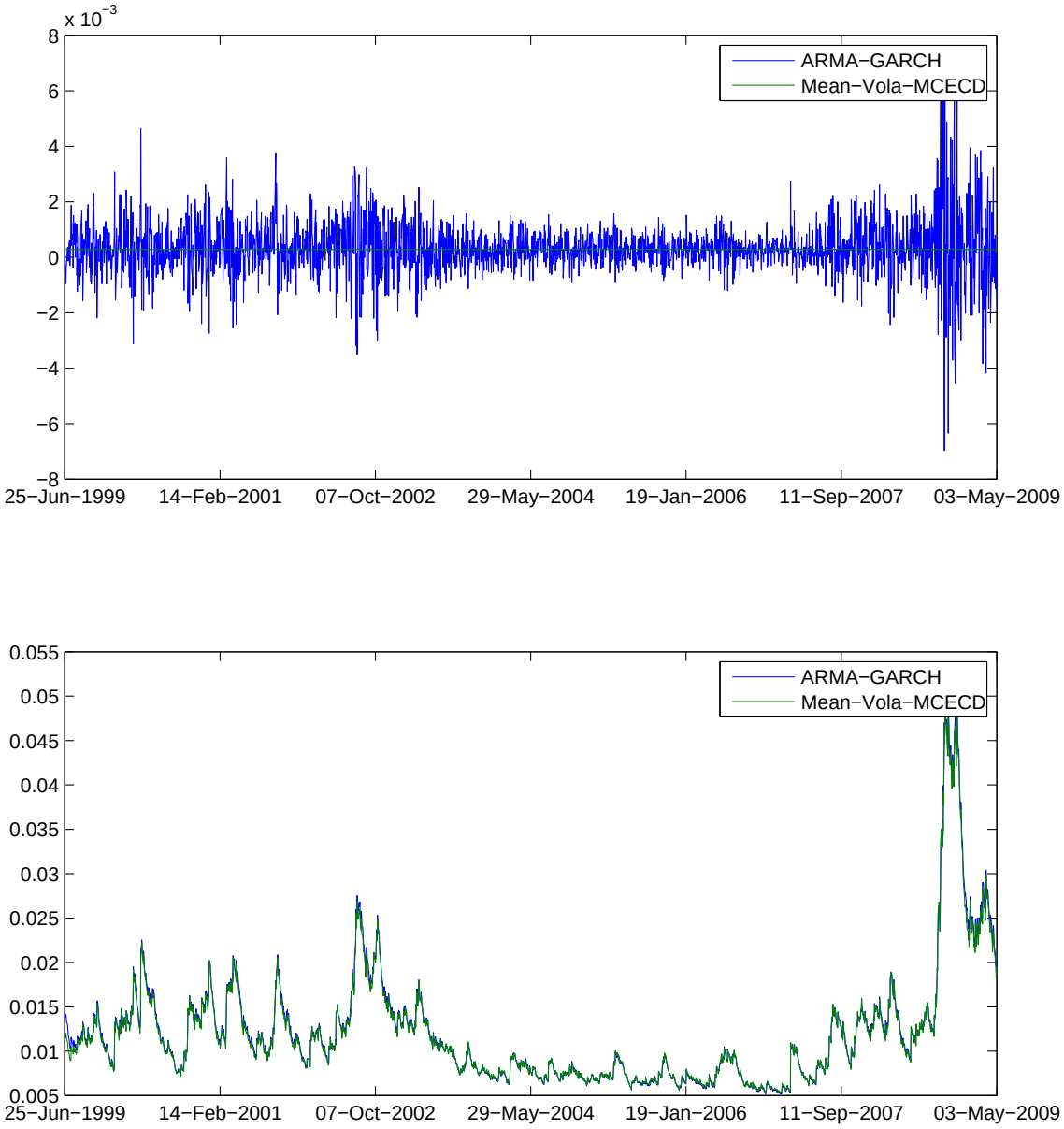
x 10^−3
8
6
4
2
0
−2
−4
−6
−8
ARMA−GARCH
Mean−Vola−MCECD
25−Jun−1999
14−Feb−2001
07−Oct−2002
29−May−2004
19−Jan−2006
11−Sep−2007
03−May−2009
0.055
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
ARMA−GARCH
Mean−Vola−MCECD
25−Jun−1999
14−Feb−2001
07−Oct−2002
29−May−2004
19−Jan−2006
11−Sep−2007
03−May−2009

Figure 3

Model	Years	Koponen parameters			Model parameters					
		α	β	λ	c	a	b	α_0	α_1	β_1
(MCECD parameters)					$(\bar{\mu})$	$(\bar{\sigma})$		(α_1)	(α_2)	(α_3)
Mean-Vola-MCECD	10	0.5	-0.21603	1.94671	0.00028	0.01142		0	0.07205	0.91795
	8	0.5	-0.23240	1.96651	0.00033	0.01097		0	0.07189	0.91811
	6	0.5	-0.24596	1.67385	0.00038	0.01011		0	0.06734	0.92266
	4	0.5	-0.19440	1.40703	0.00036	0.01009		0	0.07243	0.91590
ARMA-GARCH	10	0.5	-0.22762	1.84969	0.00032	0.09940	-0.15657	1.01634E-06	0.07223	0.92245
	8	0.5	-0.29014	1.87581	0.00016	-0.59241	0.50889	9.84114E-07	0.07105	0.92219
	6	0.5	-0.30533	1.64002	0.00019	-0.61388	0.51790	1.02280E-06	0.06922	0.92117
	4	0.5	-0.23902	1.38241	0.00022	-0.55839	0.43330	1.38219E-06	0.08970	0.90336
GARCH	10	0.5	-0.21100	1.89557	0.00027			1.04256E-06	0.07250	0.92197
	8	0.5	-0.22725	1.93171	0.00032			9.61595E-07	0.07036	0.92319
	6	0.5	-0.24337	1.64121	0.00038			1.02343E-06	0.06962	0.92084
	4	0.5	-0.18043	1.30169	0.00036			1.38844E-06	0.08957	0.90362

Table 1 Parameter estimates for Mean-Vola-MCECD, ARMA-GARCH, and GARCH models on daily S&P 500 log-return data

Model	Years	Koponen parameters			Model parameters					
		α	β	λ	c	a	b	α_0	α_1	β_1
(MCECD parameters)					$(\bar{\mu})$	$(\bar{\sigma})$		(α_1)	(α_2)	(α_3)
Mean-Vola-MCECD	10	1.90000	-0.92373	0.30000	0.00036	0.01146		0	0.08069	0.90668
	8	1.90000	-0.92137	0.30000	0.00047	0.01193		0	0.07943	0.91057
	6	0.66342	-0.33352	2.00000	0.00047	0.01002		0	0.06467	0.92501
	4	0.50000	-0.32586	2.00000	0.00044	0.01081		0	0.07193	0.91786
ARMA-GARCH	10	1.87732	-0.98000	0.30000	0.00003	-0.95641	0.93039	1.58942E-06	0.08606	0.90440
	8	1.88033	-0.98000	0.30000	0.00010	-0.84693	0.79675	1.35616E-06	0.08315	0.90893
	6	0.57598	-0.41773	2.00000	0.00009	-0.87118	0.81179	1.06254E-06	0.06966	0.92121
	4	0.50000	-0.41114	1.93424	0.00015	-0.74448	0.64966	1.29188E-06	0.08548	0.90849
GARCH	10	1.88412	-0.82495	0.30000	0.00037			1.58838E-06	0.08501	0.90553
	8	1.90000	-0.93343	0.30000	0.00046			1.29530E-06	0.08017	0.91231
	6	0.59390	-0.31835	2.00000	0.00047			1.04849E-06	0.06919	0.92198
	4	0.50000	-0.31495	1.90562	0.00042			1.29717E-06	0.08500	0.90912

Table 2 Parameter estimates for Mean-Vola-MCECD, ARMA-GARCH, and GARCH models on daily DJA log-return data

Model	Years	Koponen parameters			Model parameters					
		α	β	λ	c	a	b	α_0	α_1	β_1
(MCECD parameters)					$(\bar{\mu})$	$(\bar{\sigma})$		(α_1)	(α_2)	(α_3)
Mean-Vola-MCECD	10	1.88702	-0.12024	0.46212	0.00052	0.01531		0	0.06519	0.92481
	8	1.88044	-0.24528	0.38668	0.00046	0.01518		0	0.06285	0.92715
	6	0.67704	-0.25896	2.00000	0.00047	0.01390		0	0.06036	0.92964
	4	0.50000	-0.22928	1.96141	0.00051	0.01276		0	0.06472	0.92395
ARMA-GARCH	10	1.90000	-0.34882	0.37928	0.00020	-0.63450	0.56638	9.87612E-07	0.06026	0.93826
	8	1.89388	-0.42427	0.34361	0.00018	-0.65020	0.58643	9.88719E-07	0.05373	0.94308
	6	0.69290	-0.33409	2.00000	0.00026	-0.50520	0.43325	1.48657E-06	0.05723	0.93522
	4	0.50000	-0.29847	1.85861	0.00023	-0.61819	0.54327	2.06122E-06	0.07743	0.91511
GARCH	10	1.90000	-0.09657	0.33958	0.00048			9.72177E-07	0.05997	0.93868
	8	1.90000	-0.25093	0.31324	0.00043			9.82425E-07	0.05341	0.94347
	6	0.69415	-0.26707	2.00000	0.00046			1.48308E-06	0.05741	0.93511
	4	0.50000	-0.21939	1.81670	0.00049			2.06147E-06	0.07712	0.91549

Table 3 Parameter estimates for Mean-Vola-MCECD, ARMA-GARCH, and GARCH models on daily Nasdaq 100 log-return data

Data	Years	Method	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
DJA	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
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 4 Goodness-of-Fit for different models on daily log-return data from different U.S. stock indices

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

Table 5 One-year VaR backtesting results for stock indices from 06/26/2008 to 06/24/2009 based on 0.01 and 0.05 confidence levels

Data	Method	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 6 One-year CDF forecasting results for stocks and stock indices based on daily log-return data from 06/26/2008 to 06/24/2009