Expert Systems with Applications 40 (2013) 2233–2243

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Heavy-tailed mixture GARCH volatility modeling and Value-at-Risk estimation

Nikolay Y. Nikolaev [a,*], Georgi N. Boshnakov [b], Robert Zimmer [a]

[a] Department of Computing, Goldsmiths College, University of London, London SE14 6NW, UK
[b] School of Mathematics, University of Manchester, Manchester M13 9PL, UK

ARTICLE INFO

Keywords:
GARCH models
Mixture models
Student-*t* distribution
VaR estimation

ABSTRACT

This paper presents a heavy-tailed mixture model for describing time-varying conditional distributions in time series of returns on prices. Student-*t* component distributions are taken to capture the heavy tails typically encountered in such financial data. We design a mixture MT(m)-GARCH(p,q) volatility model for returns, and develop an EM algorithm for maximum likelihood estimation of its parameters. This includes formulation of proper temporal derivatives for the volatility parameters. The experiments with a low order MT(2)-GARCH(1,1) show that it yields results with improved statistical characteristics and economic performance compared to linear and nonlinear heavy-tail GARCH, as well as normal mixture GARCH. We demonstrate that our model leads to reliable Value-at-Risk performance in short and long trading positions across different confidence levels.

1. Introduction

The time-dependent variance in series of returns on prices, known also as volatility, is of particular interest in finance as it impacts the pricing of financial instruments, and it is a key concept in market regulation. Popular tools that describe the conditional distribution of returns via a latent variance process are the Generalized Autoregressive Conditional Heteroskedastic (GARCH) models (Engle, 1982; Bollerslev, 1987). The variance is represented by a function of the past squared returns and the past variances, which facilitates the model estimation and the computation of the prediction errors. Although the standard Gaussian GARCH has been used extensively in practice, it cannot capture well some empirical properties in the data, like the high kurtosis and the heavy tails of the return distribution.

One common idea in recent research into GARCH modeling is to develop specifications with non-normal distributions (Ardia, 2008; Bauwens, Hafner, & Laurent, 2012; Deschamps, 2012; Kim, Rachev, Bianchi, Mitov, & Fabozzi, 2011; Lee & Pai, 2010). The distributional assumptions are found more influential for improvements than making non-linear extensions with neural networks (Bildirici & Ersin, 2009; Wang, 2009) or kernel machines (Pérez-Cruz, Afonso-Rodriguez, & Giner, 2003; Tang, Sheng, & Tang, 2009). These non-linear extensions have been added to complement the autoregressive part of the return equation with intention to adapt the GARCH model more closely to the data. The shortcoming of these approaches is that without careful treatment of the noise they do not have enough representation potential to accommodate aberrant changes in returns arising from occasional financial events (Miazhynskaia, Dorffner, & Dockner, 2003).

The Normal mixture NM-GARCH(Alexander & Lazar, 2006; Ausin & Galeano, 2007; Haas, Mittnik, & Paolella, 2004; Schittenkopf, Dorffner, & Dockner, 2000; Wong & Li, 2001) are among the most tractable non-Gaussian volatility models that achieve better fit to time series whose data distribution deviates from the normal. These authors designed normal mixtures in order to describe flexibly return distributions under the assumption that the data may originate from alternative sources. However, financial returns tend to have more peaked (leptokurtic) distributions with longer tails while such normal mixture models often fail to account for these features together. When there are extreme observations mixture models whose components are based on the Gaussian density may not be sufficient to achieve satisfactory performance (Ardia, Hoogerheide, & van Dijk, 2009).

Financial series with heavier than normal tails are typically treated using the Student-*t* distribution (Bollerslev, 1987) and the Generalised Error Distribution (GED) (Wilmhelmsson, 2006). They can model various density shapes with accuracy controlled by a degrees of freedom parameter. When the degrees of freedom is large the distribution approaches the Gaussian. When the data require modeling by smaller degrees of freedom the Student-*t* distribution is more robust than the Gaussian, which in this case leads to distorted variance (Wong, Chan, & Kam, 2009). These problems may be addressed using mixtures of Student-*t* distributions (Ardia et al., 2009; Peel & McLachlan, 2000; Sefidpour & Bouguila, 2012; Wong et al., 2009) as they provide greater representation potential

* Corresponding author.
E-mail addresses: n.nikolaev@gold.ac.uk (N.Y. Nikolaev), georgi.boshnakov@manchester.ac.uk (G.N. Boshnakov), r.zimmer@gold.ac.uk (R. Zimmer).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.038

for dealing with densities that exhibit heavier tails and higher kurtosis. The Student-t mixtures is also attractive because it has proven universal approximation capacity in the sense that they can represent any continuous density function on a compact domain to arbitrary accuracy under certain conditions (Zeevi & Meir, 1997).

This paper develops a mixture GARCH modeling approach using Student-t components to handle time series data with abnormal behavior. There are two main contributions: (1) we design a heavy-tail mixture of Student-t distributions MT(m)-GARCH(p,q) model in which the volatility is explicitly specified as a time-dependent variable using GARCH equations (this heavy-tail model generalizes previous normal mixture volatility models as well as heavy-tail logistic mixture models with constant volatility); (2) we design an Expectation Maximisation (EM) algorithm for maximum likelihood estimation (MLE) of the mixture model parameters in the style of Peel and McLachlan (2000). This includes formulation of proper dynamic temporal derivatives for estimating the volatility parameters which reflect the time ordering of the data. Thus, the parameter fitting algorithm accounts explicitly for the time dependencies among the data. The temporal derivatives are general as they can be used also to train nonlinear volatility specifications.

The heavy-tail mixture approach is investigated empirically on simulated and real financial data in order to assess its practical usefulness. We report results from its statistical, econometric, and risk management performance. The developed MT(2)-GARCH(1,1) is compared with a standard linear GARCH, a linear TGARCH with Student-t noise, a nonlinear NGARCH with Student-t innovations, and a normal mixture NM-GARCH trained by *MCMC* sampling. The conducted studies found that the MT(2)-GARCH(1,1) achieves improved statistical characteristics on in-sample and out-of-sample data than all these models. These results are consistent and complement previous studies into normal mixture NM-GARCH models which found that they outperform GARCH and TGARCH models (Alexander & Lazar, 2006; Haas et al., 2004). The MT(2)-GARCH(1,1) model also demonstrates better out-of-sample econometric performance as it captures more precisely future directional changes.

The risk management performance is evaluated using the Value-at-Risk (VaR) measure (Christoffersen, 2003; Jorion, 1996) which makes our results comparable to these from recent relevant research (Orhan & Köksal, 2012). The experimental methodology involves bootstrapping the returns (Pascual, Romo, & Ruiz, 2006) and estimating semi-parametrically VaR. Using the obtained VaR predictions for 95% and 99% confidence levels there were computed likelihood ratio coverage tests according to the methodology of Christoffersen (2003). We found that MT(2)-GARCH(1,1) exhibits similar risk performance to the other competing models on 99% VaR, and it is only slightly better on 95% VaR estimation. On 95% VaR estimation MT(2)-GARCH(1,1) is almost equivalent to the normal mixture NM(2)-GARCH(1,1), but both of them outperform all other models. The heavy-tail mixture model seems most reliable as it generates equally good 1% and 5% VaR in long as well as short positions. The standard linear GARCH(1,1) leads to unacceptably large failure rates on 99% and 95% VaR for the left tail, so it may be inadequate for predicting risk. The linear TGARCH(1,1) also produced exceeding 95% VaR estimates but for the right tail.

The paper is organized as follows. Section 2 introduces the normal mixture NM(m)-GARCH(p,q) models, and presents the heavy-tailed mixture MT(m)-GARCH(p,q) model. Section 3 presents the EM algorithm, including the log-likelihood function computed during the expectation step, as well as the formulae and the temporal derivatives for computing the parameters during the maximization step. Section 4 gives the experimental results on volatility inference obtained using simulated and benchmark series. Finally, a conclusion is provided.

2. Mixtures of GARCH processes

2.1. Normal NM(m)-GARCH(p,q) models

The variance of returns on prices changes over time, and these changes are modeled by an unobserved process. The latent time-dependent variance of the observable returns is also known as volatility. Let us consider the log-returns from a series of prices S_t, $1 \leqslant t \leqslant T$, that is $r_t = \log(S_t/S_{t-1})$. The dynamics of these log-returns can be described by a mixture of normal heteroskedastic NM(m)-GARCH(p,q) models (Alexander & Lazar, 2006; Haas et al., 2004):

$$p(r_t|R_{t-1}) = \sum_{m=1}^{M} \rho_m \phi\left(r_t; \mu_t, \sigma^2_{m,t}\right) \tag{1}$$

$$\sigma^2_{m,t} = \alpha_{m,0} + \sum_{i=1}^{q} \alpha_{m,i} e^2_{t-i} + \sum_{j=1}^{p} \beta_{m,j} \sigma^2_{m,t-j} \tag{2}$$

where $R_{t-1} = (r_1, r_2, \ldots, r_{t-1})$ denotes the past information arrived up to time t, ρ_m are the mixing coefficients, ϕ is the Gaussian density function, μ_t is the mean of the return distribution $E[r_t|R_{t-1}] = \mu_t$, and $\sigma^2_{m,t}$ is the conditional variance [1] of the m component so that $\sigma^2_t = \sum_{m=1}^{M} \rho_m \sigma^2_{m,t}$ and $E[(r_t - \mu_t)^2 | R_{t-1}] = \sigma^2_t$. The mean is typically captured by an autoregressive AR(1) model $\mu_t = a + br_{t-1}$. The mixing coefficients should satisfy $\sum_{m=1}^{M} \rho_m = 1$ $(0 < \rho_m < 1)$. The volatility parameters are restricted to ensure positive variance ($\alpha_{m,i} \geqslant 0$, $\beta_{m,j} \geqslant 0$) and stationarity $\left(\sum_{i=1}^{q} \alpha_{m,i} + \sum_{j=1}^{p} \beta_{m,j} < 1\right)$.

Recent research found that such NM(m)-GARCH(p,q) often fail to capture well the excess kurtosis in return distributions (Wong et al., 2009). Excess kurtosis means that the tail of the distribution is longer than the normal, such distributions are said to feature heavier (fatter) tails.

2.2. The Heavy-tailed MT(m)-GARCH(p,q) model

Aiming at more robust approximation of return distributions we elaborate a generalized mixture MT(m)-GARCH(p,q) model. The MT(m)-GARCH(p,q) model describes the evolution of the volatility again by the classic volatility equation (2) while the equation for returns Eq. (1) is replaced by a mixture of heavy-tailed (non-Gaussian) distributions:

$$p(r_t|R_{t-1}) = \sum_{m=1}^{M} \rho_m f\left(r_t; \mu_t, \sigma^2_{m,t}, \nu_m\right) \tag{3}$$

where each component uses the standardized Student-t probability density function $f(\cdot)$ defined as follows:

$$\begin{aligned} f(r_t; \mu_t, \sigma^2_{m,t}, \nu_m) &= \frac{C}{\sigma_{m,t}} \left(1 + \frac{(r_t - \mu_t)^2}{\sigma^2_{m,t}(\nu_m - 2)}\right)^{-(\nu_m+1)/2} \\ C &= \frac{\Gamma((\nu_m+1)/2)}{\sqrt{\pi(\nu_m - 2)}\,\Gamma(\nu_m/2)} \end{aligned} \tag{4}$$

where $\Gamma(\cdot)$ is the Gamma function and $\nu_m > 2$ is the degrees of freedom parameter. This parameter controls the thickness of the tails and the peakedness of the distribution.

The MT(m)-GARCH(p,q) defined by Eqs. (2)–(4) extends the previous non-heteroscedastic mixtures of Student-t models (Ardia et al., 2009; Wong et al., 2009) by accomodating additional latent variables $\sigma^2_{m,t}$ for describing time-varying volatilities using the recursive GARCH equation Eq. (2) instead of using constant variances.

[1] The conditional dependence of the volatility components on the inputs is omitted for clarity, $\sigma^2_{m,t} \equiv \sigma^2_{m,t}(r_{t-1})$.

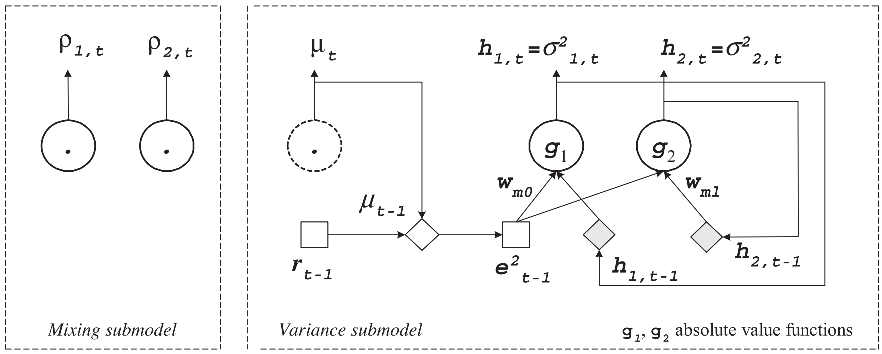

Fig. 1. MT(2)-GARCH(1, 1): an autoregressive heteroskedastic mixture model for time series distributions with a common mean μ_t and conditional volatilities (variances) $\sigma^2_{1,t}$ and $\sigma^2_{2,t}$.

This mixture model can be implemented as a hybrid structure of two submodels (Fig. 1): (1) a mixing submodel that provides the mixing coefficients ρ_m; and (2) a variance submodel that produces the volatilities $h_{m,t} = \sigma^2_{m,t}$ using as inputs the squared past errors $e^2_{t-l} = (y_{t-l} - \mu_{t-l})^2$ and the past volatilities $h_{t-l+p} = \sigma^2_{t-l+p}$. The volatilities h_{t-l+p} are temporal variables that contain information from the past and send it via the recurrent connections. We call the GARCH parameters weights. These weights correspond to the parameters mean $\alpha_{m,0}$, autoregression $\alpha_{m,l}$ and persistence $\beta_{m,j}$ in the GARCH specification Eq. (2), that is: $w_{m0} \equiv \alpha_{m,0}$, $w_{ml} \equiv \alpha_{m,l}$ $(1 \leqslant l \leqslant p)$, and $w_{ml} \equiv \beta_{m,j}$ $(p+1 \leqslant l \leqslant p+q,\ 1 \leqslant j \leqslant p)$.

Let us adopt the following notation for the inputs to the variance submodel at time t:

$$x_{t-l} = \begin{cases} 1, & \text{i.e. bias if } l = 0 \\ e^2_{t-l} = (y_{t-l} - \mu_{t-l})^2, & \text{if } 1 \leqslant l \leqslant p \\ h_{t-l+p}, & \text{if } (p+1) \leqslant l \leqslant (p+q) \end{cases} \quad (5)$$

where p is the number of lagged inputs, q is the number of recurrencies, and l enumerates the inputs.

The variance submodel generates the volatilities at its outputs by weighted summation:

$$h_{m,t} = g(y_{m,t}) = g\left(\sum_{l=0}^{p+q} w_{ml} x_{t-l}\right)$$
$$= g\left(\sum_{l=1}^{p} w_{ml} e^2_{t-l} + \sum_{l=p+1}^{p+q} w_{ml} h_{m,t-l+p} + w_{m0}\right) \quad (6)$$

where w_{ml} are the weights (parameters), and $g(\cdot)$ is the absolute value function $g(y_{m,t}) = |y_{m,t}|$.

This MT-GARCH has a common mean for all mixtures, thus it is less complex than previous NM-GARCH(Alexander & Lazar, 2006; Haas et al., 2004) and switching heavy-tail GARCH(Ardia, 2008). It is only slightly more complex than the mixture GARCH with a single volatility equation (Ausin & Galeano, 2007).

3. Estimation of the MT-GARCH model

The Expectation–Maximization (EM) algorithm is preferred for computing maximum likelihood estimates (MLE) of the parameters in mixture models (Peel & McLachlan, 2000). The rationale is that it decouples the estimation of the parameters into two steps assuming that each data point is generated by one mixture component. Since it is not known which component has generated the observation, the algorithm also learns indicator variables. EM iteratively updates the indicator variables (assuming that the parameters are known) during the expectation step, and the model parameters (using the expected values of the indicators) during the maximization step.

The EM algorithm adapts iteratively the MT-GARCH parameters $\Theta = \left\{ \left((w_{mi})_{i=0}^{W} \right)_{m=1}^{M}, (\rho_m)_{m=1}^{M}, (\nu_m)_{m=1}^{M} \right\}$ by alternating the two steps until the expected complete data likelihood $E[\log L(\Theta)|R_t]$ is maximized:

$$E\text{-}step: L(\Theta; \hat{\Theta}_{(k)}) = E[\log L(\Theta)|R_t, \hat{\Theta}_{(k)}]$$
$$M\text{-}step: \hat{\Theta}_{(k+1)} = \arg\max_{\Theta}\{L(\Theta; \hat{\Theta}_{(k)})\}$$

starting with the estimates $\hat{\Theta}_{(k)}$ from the previous kth step.

The EM is a principled algorithm that ensures convergence while seeking the mode of the total log-likelihood (Dempster, Laird, & Rubin, 1977).

3.1. Log-likelihood function

The probability that the mth mixture produces the return r_t is defined by a missing indicator variable $Z_m = \{(Z_{m,t}|r_t)\}_{t=1}^{T}$, $(1 \leqslant m \leqslant M)$. The complete (total) negative log-likelihood using such missing variables is assumed to be:

$$-\log L(\Theta) = -\sum_{t=1}^{T}\sum_{m=1}^{M} Z_{m,t} \log(\rho_m f(r_t; \mu_t, h_{m,t}, \nu_m)) \quad (7)$$

which requires taking the expectations of the hidden variables.

A tractable log-likelihood that admits solutions in closed form, is formulated by expressing the Student-t density as a continuous mixture of Normal $\mathcal{N}$ and Gamma $\mathcal{G}$ densities (Peel & McLachlan, 2000) as follows:

$$f(r_t; \mu_t, h_{m,t}, \nu_m) = \int_0^\infty \mathcal{N}\left(r_t; \mu_t, \frac{h_{m,t}}{u_{m,t}}\right) \mathcal{G}\left(u_{m,t}; \frac{\nu_m}{2}, \frac{\nu_m}{2}\right) du_{m,t}$$

where

$$\mathcal{N}\left(r_t; \mu_t, \frac{h_{m,t}}{u_{m,t}}\right) = \frac{1}{\sqrt{2\pi h_{m,t}/u_{m,t}}} \exp\left(-\frac{u_{m,t}(r_t - \mu_t)^2}{2h_{m,t}}\right)$$

$$\mathcal{G}\left(u_{m,t}; \frac{\nu_m}{2}, \frac{\nu_m}{2}\right) = \frac{(\nu_m/2)^{\nu_m/2}}{\Gamma(\nu_m/2)} u_{m,t}^{\nu_m/2-1} \exp\left(-\frac{\nu_m u_{m,t}}{2}\right)$$

and $u_{m,t}$ ($u_{m,t} > 0$) is a precision parameter defined as the expectation $u_{m,t} = E[U_t|r_t, z_{m,t} = 1]$ of another missing random variable $U = \left\{(U_t|r_t)_{t=1}^{T}\right\}$ drawn from the Gamma distribution.

Plugging this indefinite scaled mixture density $f(r_t; \mu_t, h_{m,t}, \nu_m)$ into the log-likelihood function (Eq. (7)) allows us to perform efficient optimization. Taking the logarithm of the likelihood leads to three terms proportional respectively to the marginal density of the indicator variable, the conditional density of the observations, and the conditional density of the precision variable. After taking the expectation of each term, we arrive at the following expected total log-likelihood of the mixture model:

$$E[\log L(\Theta)|\hat{\Theta}_{(k)}] = L1(\hat{\Theta}_{(k)}) + L2(\hat{\Theta}_{(k)}) + L3(\hat{\Theta}_{(k)}) \tag{8}$$

The terms in this expected total log-likelihood are:

$$L1(\hat{\Theta}_{(k)}) = \sum_{t=1}^{T}\sum_{m=1}^{M} z_{m,t}\log\rho_m \tag{9}$$

$$L2(\hat{\Theta}_{(k)}) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{M} E[l2_{m,t}(\hat{\Theta}_{(k)})] \tag{10}$$

$$L3(\hat{\Theta}_{(k)}) = \sum_{t=1}^{T}\sum_{m=1}^{M} E[l3_{m,t}(\hat{\Theta}_{(k)})] \tag{11}$$

These terms are defined via the following functions:

$$l2_{m,t}(\hat{\Theta}_{(k)}) = z_{m,t}\left[\log h_{m,t} - \log u_{m,t} + \frac{u_{m,t}(r_t-\mu_t)^2}{h_{m,t}}\right]$$

$$l3_{m,t}(\hat{\Theta}_{(k)}) = z_{m,t}\left[\frac{v_m}{2}\log\left(\frac{v_m}{2}\right) - \log\Gamma\left(\frac{v_m}{2}\right) + \frac{v_m}{2}(\log u_{m,t} + y_m - u_{m,t}) - \log u_{m,t}\right]$$

where the conditional expectations are $z_{m,t} = E[Z_{m,t}|r_t]$, $u_{m,t} = E[U_t|r_t, z_{m,t} = 1]$ and $\log u_{m,t} + y_m = E[\log U_t|r_t, z_{m,t} = 1]$ of the hidden random variables respectively $\{Z_{m,t}|r_t\}$, $\{U_t|r_t, z_{m,t} = 1\}$ and $\{\log U_t|r_t, z_{m,t} = 1\}$.

3.2. Expectation step

During the expectation step the total expected log-likelihood $E[\log L(\Theta)|R_t, \hat{\Theta}_{(k)}]$ is obtained by replacing the missing variables $Z_m = \{(Z_{m,t}|r_t)\}_{t=1}^{T}$ $(1 \leqslant m \leqslant M)$ and $U = \{(U_t|r_t)\}_{t=1}^{T}$ with their expected values (conditioned on the observed data up to this moment in time) based on the current parameters $\hat{\Theta}_{(k)}$. The expected missing variables $E[Z_{m,t}|r_t]$, the expected precisions $E[U_t|r_t, z_{m,t} = 1]$ and $E[\log U_t|r_t, z_{m,t} = 1]$ are computed as follows:

$$z_{m,t} = E[Z_{m,t}|r_t] = \frac{\rho_m f(r_t;\mu_t, h_{m,t}, v_m)}{\sum_{j=1}^{M}\rho_j f(r_t;\mu_t, h_{j,t}, v_j)} \tag{12}$$

$$u_{m,t} = E[U_t|r_t, z_{m,t} = 1] = \frac{v_m + 1}{v_m + \delta_{m,t}} \tag{13}$$

$$\log u_{m,t} + y_m = E[\log U_t|r_t, z_{m,t} = 1] = \log u_{m,t} + \psi\left(\frac{v_m+1}{2}\right) - \log\left(\frac{v_m+1}{2}\right) \tag{14}$$

where $\psi(\cdot)$ is the Digamma function, and $\delta_{m,t}$ is the squared Mahalanobis distance $\delta_{m,t} = (r_t - \mu_t)^2/h_{m,t}$.

3.3. Maximisation step

The maximization step involves computing the mixing coefficients, the volatility parameters, the mean return and the degrees of freedom parameter. Each of these computations are performed in a different manner.

3.3.1. Finding the mixing coefficients

The first term $L1(\hat{\Theta}_{(k)})$ measures the cross-entropy between the missing variables $z_{m,t}$ and the mixing coefficients ρ_m, and penalizes sharing of data points by the components. Taking the derivatives of this term with respect to the mixing coefficients and using Lagrange multipliers yields the following equation (see Appendix):

$$\rho_m = \frac{1}{T}\sum_{t=1}^{T} z_{m,t} \tag{15}$$

3.3.2. Estimating the volatility parameters

The second term $L2(\hat{\Theta}_{(k)})$ in the expected total log-likelihood Eq. (8) accounts for the effect of the volatility on the errors. The impact of the volatility on the errors is however scaled by the mixing coefficients. This indicates that finding of a more probable mixture component will be affected more by the particular data point. The derivatives of the second term can be used for finding the variance parameters by plugging them into an optimizer.

We propose here to compute the derivatives of $E[l2_{m,t}(\hat{\Theta}_{(k)})]$ with respect to the variance weights w_{mj} $(1 \leqslant m \leqslant M,\ 0 \leqslant j \leqslant (p+q))$ following the real-time learning algorithm (Williams & Zipser, 1995) in order not only to asjust the weights using the instantaneous error, but also to take advantage of the full error flow through time. These dynamic, temporal derivatives are obtained using the chain rule as follows (see Appendix):

$$\frac{\partial E[l2_{m,t}(\hat{\Theta}_{(k)})]}{\partial w_{mj}} = \left(\frac{z_{m,t}}{h_{m,t}}\left[1 - \frac{u_{m,t}(r_t-\mu_t)^2}{h_{m,t}}\right]\right)\frac{\partial h_{m,t}}{\partial w_{mj}} \tag{16}$$

The temporal derivatives $\partial h_{m,t}/\partial w_{mj}$ of an output $h_{m,t}$ with respect to a weight w_{mj} $(1 \leqslant m \leqslant M,\ 0 \leqslant j \leqslant (p+q))$ are evaluated recursively with stored partial derivatives from the past by (see Appendix):

$$\frac{\partial h_{m,t}}{\partial w_{mj}} = g'(y_{m,t})\left(\sum_{l=p+1}^{p+q}\left[w_{ml}\frac{\partial h_{m,t-l+p}}{\partial w_{mj}}\right] + x_{t-j}\right) \tag{17}$$

where $g'(y_{m,t}) = y_{m,t}/|y_{m,t}|$.

These temporal derivatives are more general than the previous analytical GARCH derivatives (Fiorentini, Calzolari, & Panattoni, 1996) as they can be used also to train nonlinear volatility specifications (Nikolaev, Tino, & Smirnov, 2011).

3.3.3. Estimating the mean return

The MT(m)-GARCH(p,q) has a common mean $\mu_t = E[r_t|R_{t-1}]$ $(1 \leqslant t \leqslant T)$ for both mixture components. A simple way to determine the mean is to perform OLS fitting of the autoregressive AR(1) model $\mu_t = a + br_{t-1}$ to the given training data.

3.3.4. Evaluating the degrees of freedom

The degrees of freedom parameters v_m $(v_m > 0)$ of the Student-t component distributions however cannot be obtained in closed-form (Peel & McLachlan, 2000), but may be found as solutions of the corresponding equations:

$$\log\left(\frac{v_m}{2}\right) + 1 - \psi\left(\frac{v_m}{2}\right) + \frac{1}{T}\sum_{t=1}^{T} z_{m,t}(\log u_{m,t} - u_{m,t}) + y_m = 0 \tag{18}$$

where the expected values of the hidden variables are computed by the above equations. This paper assumes for simplicity a common degrees of freedom parameter v for both mixture components (although having distinct parameters v_m $(v_m > 0)$ for each component provides more flexibility).

4. Applications to volatility inference

Experiments in volatility inference were conducted using simulated series, and a benchmark series of DEM/GBP exchange rates. The research was carried out to examine the influence of heavy-tailed models and their dynamic training on the performance of MT-GARCH models, and to relate them to NM-GARCH models, as well as to conventional linear and nonlinear TGARCH models based on the Student-t distribution.

4.1. Processing simulated series

4.1.1. Heteroskedastic mixture model

The ability of the proposed approach to learn descriptions of stochastic processes was tested by simulating series with an MT(2)-GARCH(1,1) volatility model with a common mean defined by:

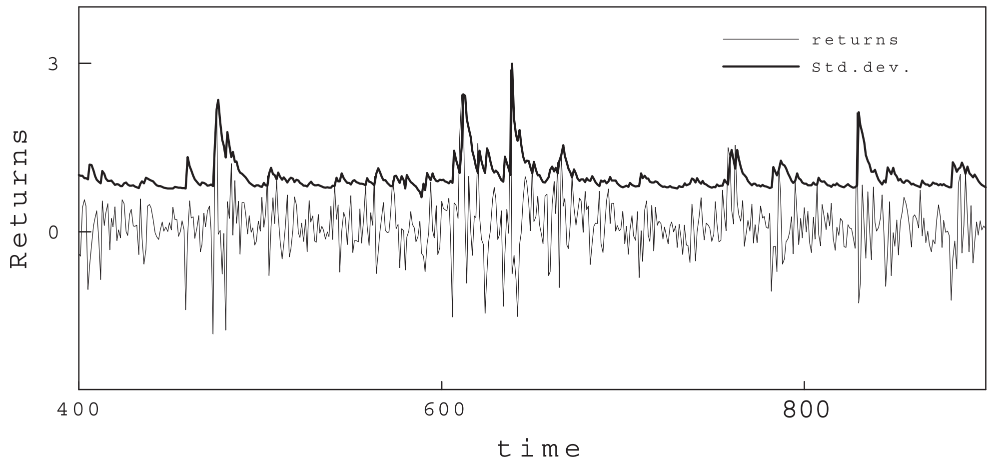

Fig. 2. Generated series with the selected MT(2)-GARCH(1, 1) mixture model wrapped by its time-varying standard deviation $\pm 2\sqrt{\sigma_t^2}$, computed using the mixed volatility $\sigma_t^2 = \sum_{m=1}^{M} \rho_m \sigma_{m,t}^2$.

$$\begin{aligned} p(r_t|R_{t-1}) &= \rho f(\mu, h_{t,1}, \nu) + (1-\rho) f(\mu, h_{t,2}, \nu) \\ h_{t,1} &= \alpha_{10} + \alpha_{11} e_{t-1}^2 + \beta_{11} h_{t-1} \\ h_{t,2} &= \alpha_{20} + \alpha_{21} e_{t-1}^2 + \beta_{22} h_{t-1} \end{aligned} \tag{19}$$

where ρ is the mixing coefficient, f is the Student-t pdf (Eq. (4)), and ν is the degrees of freedom parameter. The selected true values for these parameters were: $\mu = 0.1,\ \alpha_{10} = 0.05,\ \alpha_{11} = 0.2,\ \beta_{11} = 0.6,\ \alpha_{20} = 0.1,\ \alpha_{21} = 0.1,\ \beta_{21} = 0.8,\ \rho = 0.9,\ \nu = 5,\ e_0^2 = 0.0$ and $h_0 = 1.0$.

There were generated four groups of series with increasing length as follows: $T = 500$, $T = 1000$, $T = 1500$, and $T = 2000$. Each group included 1000 simulated series of the same size. Each run of the optimizer started with the same plausible initial parameters for the mixture model: $\alpha_{10} = 0.01$, $\alpha_{11} = 0.1$, $\beta_{11} = 0.7$, $\alpha_{20} = 0.01$, $\alpha_{21} = 0.1$, $\beta_{21} = 0.90$.

4.1.2. Experimental results

Table 1 gives the averaged parameter estimates and their Root Mean Squared Errors (RMSE) in parentheses. These parameters are actually averages over 10,000 runs starting with the same initial values but each run on a different sampled series. One can see in Table 1 that the proposed estimation algorithm infers correctly the mixing coefficient $\rho_1 \equiv \rho$, which is close to the unknown true value of 0.9 indeed. The other learned parameters are also adequate, although their precisions vary with the sample sizes. Most precisely seem learnable the volatility persistence parameters β_{11} and β_{21}, and next the volatility moving average coefficients α_{11} and α_{21}. The means α_{10} and α_{20} in the volatility equations are found slightly more distant from the true values.

Fig. 2 illustrates that the inferred volatility of a particular series wraps closely the generated returns. We plot the time-varying standard deviation $\pm 2\sqrt{\sigma_t^2}$, which is obtained from the common variance calculated by $\sigma_t^2 = \sum_{m=1}^{M} \rho_m \sigma_{m,t}^2$. The next Fig. 3 plots the difference between the low and high regime components of the mixture model, and shows that the identified low and high volatility components from the same series are very close to the given true components.

Fig. 4 offers a plot of the probability density of one particular simulated series of size 500. The bimodal distribution given with thin curve was obtained with the true density of the generative mixture model (Eq. (19)). The bold curve was obtained using MT(2)-GARCH(1, 1), and the dotted curve was produced with the starting parameters. The true density of the generative mixture model has been developed by random selection of the next component to produce the specific observation, while the density inferred by MT(2)-GARCH(1, 1) has been made using the estimated regime probabilities. This Fig. 4 illustrates that the inferred distribution approximates closely the true, unknown distribution of the generated data.

Table 1
Estimated average MT(2)-GARCH(1, 1) parameters and their RMSE errors in parentheses from 1000 simulated series of increasing sizes T generated with the chosen true parameters for the two-component mixture model (Eq. (19)) using a degrees of freedom parameter $\nu = 4.92$ computed in advance.

Param.	True	$T = 500$	$T = 1000$	$T = 1500$	$T = 2000$
μ	0.1	0.0995	0.0996	0.1008	0.1005
		(0.0124)	(0.0122)	(0.0115)	(0.0107)
α_{10}	0.05	0.0469	0.0428	0.0416	0.0408
		(0.0223)	(0.0197)	(0.0168)	(0.0142)
α_{11}	0.2	0.1802	0.1845	0.1785	0.1763
		(0.0745)	(0.0672)	(0.0551)	(0.0529)
β_{11}	0.6	0.5972	0.5937	0.5946	0.5981
		(0.0991)	(0.0965)	(0.0912)	(0.0824)
α_{20}	0.1	0.1107	0.1114	0.1142	0.1154
		(0.0735)	(0.0658)	(0.0572)	(0.0543)
α_{21}	0.1	0.0973	0.0969	0.0923	0.0912
		(0.0954)	(0.0948)	(0.0929)	(0.00915)
β_{21}	0.8	0.8024	0.8036	0.8041	0.8055
		(0.1016)	(0.0983)	(0.0972)	(0.0971)
ρ_1	0.9	0.9047	0.9008	0.9003	0.9004
		(0.0036)	(0.0029)	(0.0024)	(0.0023)
ρ_2	0.1	0.0992	0.0995	0.0996	0.0995
		(0.0038)	(0.0022)	(0.0019)	(0.0018)

4.2. Processing financial series

4.2.1. Implemented models and algorithms

GARCH Models. Several models were programmed and processed on the same platform in order to gain insights about the differences in their behavior, although these models are based on different distributional assumptions. We implemented: a standard normal linear GARCH(1, 1) (Engle, 1982), a TGARCH(1, 1) (Bollerslev, 1987) model with heavy-tailed Student-t innovations, a nonlinear NGARCH(1, 1) (Nikolaev et al., 2011) with Student-t noise, a normal mixture NM-GARCH(1, 1) (Ausin & Galeano, 2007) with a single volatility equation, a normal mixture with two volatility equations NM(2)-GARCH(1, 1) given by Eqs. (1) and (2), and a heavy-tailed mixture with two volatility equations MT(2)-GARCH(1, 1) given by Eqs. (2)–(4).

Estimation algorithms. The NM-GARCH(1, 1) model was tackled with the original MCMC sampling algorithm developed by Ausin and Galeano (Ausin & Galeano, 2007), using 100 burn-in iterations

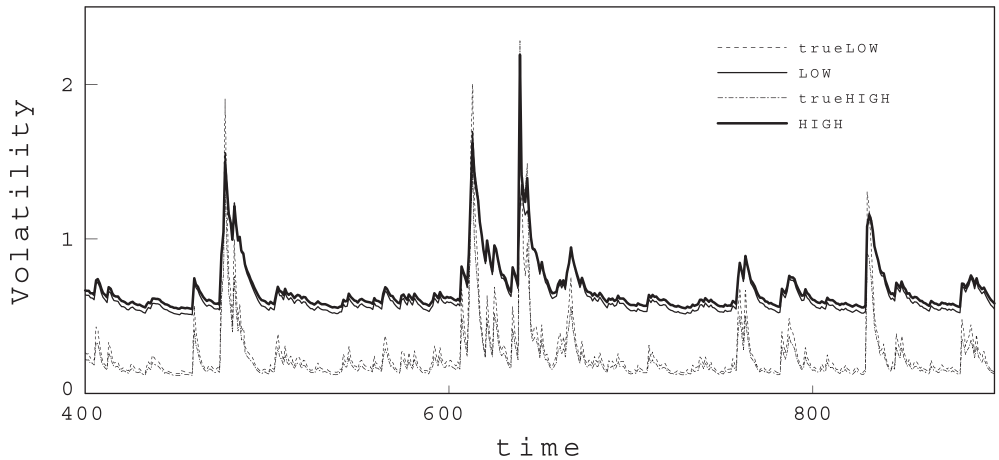

Fig. 3. High and low volatility regimes extracted as components of the common heteroskedastic volatility from Fig. 2 obtained by EM estimation of the MT(2)-GARCH(1,1) mixture model.

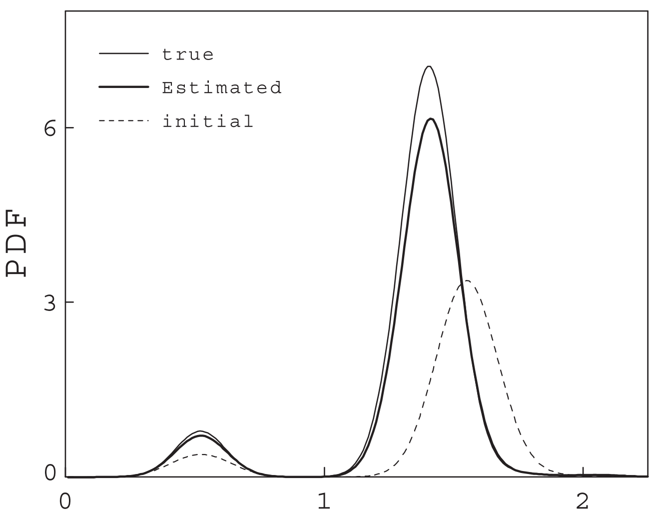

Fig. 4. Probability density of data from a simulated series, obtained with inferred MT(2)-GARCH(1,1) parameters, plotted with the initial density computed with the starting parameters, and the true density of the generative mixture model. All curves are developed by nonparametric density estimation.

and 100 training iterations. The parameters of all other models were found by Maximum Likelihood Estimation (*MLE*) by iterating the EM algorithm for up to 5 cycles. Special versions MLE_t were made for processing the non-normal GARCH models using the Student-*t* likelihood function. In all experiments the same BFGS optimizer (Fletcher, 1987) was executed with settings: *Tolerance* = $1.0e^{-10}$, *MaxIterations* = 10^2 and *FunctionEvaluations* = 10^2.

The optimization of the standard GARCH(1,1) and TGARCH(1,1) models was carried out using numerical derivatives, while the remaining NGARCH(1,1), NM(2)-GARCH(1,1) and MT(2)-GARCH(1,1) were treated using dynamic analytical derivatives plugged into the same BFGS optimizer. Our research found that processing of the last three models using numerical derivatives leads to suboptimal results, which actually motivated the development of the analytical derivatives presented here. The degree of freedom parameter was precomputed in advance for all studied models.

4.2.2. Experimental methodology

Performance measures. Investigations were conducted to analyse the accuracy of the learned parameters, and to examine the statistical characteristics of the inferred models using standardized residuals. First, we studied the in-sample performance by calculating the coefficients of skewness and kurtosis, autocorrelation statistics, and the log-likelihood. Second, the predicted volatilities were taken to compute several out-of-sample performance measures, namely the normalized mean squared error (*NMSE*), the normalized mean absolute error (*NMAE*) and the hit rate (*HR*) (Schittenkopf et al., 2000). Since in practice the returns are typically noisy, we also considered two other measures for testing the predictability of GARCH models, these are the logarithmic loss (*LLOS*) (Pagan and Schwert, 1990) and the Gaussian maximum likelihood (*GMLE*) (Bollerslev, Engle, & Nelson, 1994).

Technology. The given series of returns was split into training and testing subseries. The training series was used to infer the model parameters and their standard errors. The out-of-sample accuracy was investigated by computing one-step-ahead volatility forecasts, and rolling sequentially by one-step foreword over the testing subseries. After predicting the future volatility the model was retrained, and this algorithm repeated till the end of the testing series. The training subseries were overlapping but their size was constant.

4.2.3. Processing the benchmark DEM/GBP series

The benchmark DEM/GBP currency exchange rates series was taken to facilitate comparisons with relevant research (Ardia, 2008; Brooks, Burke, & Persand, 2001). The series contains 1974 daily observations recorded from 3/1/1984 to 31/12/1991, which were divided into 1500 data for training and 474 for testing.

Table 2 offers estimates of the main parameters of the studied GARCH models. Their parameters were computed after starting from the following initial values: $\alpha_{10} = 0.01$, $\alpha_{11} = 0.05$, $\beta_{11} = 0.5$, $\alpha_{20} = 0.05$, $\alpha_{21} = 0.1$, $\beta_{21} = 0.9$ (the single equation models used only one triple). These estimates are very close to the parameters reported by previous research: $\alpha_0 = 0.011$, $\alpha_1 = 0.1531$ and $\beta = 0.8059$ (Brooks et al., 2001). It should be noted that the mixture NM(2)-GARCH(1,1) and MT(2)-GARCH(1,1) yield two descriptions of the volatility process. These two processes are used to obtain the common volatility by: $\sigma_t^2 = \sum_{m=1}^{M} \rho_m \sigma_{m,t}^2$ (Fig. 5). We found that stable convergence during the optimization of NM(2)-GARCH(1,1) and MT(2)-GARCH(1,1) can only be achieved using the temporal derivatives, while using numerical differences instead did not produce satisfactory results.

Table 3 gives the results from statistical tests with the in-sample standardized residuals: $\hat{\epsilon} = (r_t - \mu_t)/\sqrt{h_t}$, which are recommended for volatility models (Kim, Shephard, & Chib, 1998). One can see that NM(2)-GARCH(1,1) and MT(2)-GARCH(1,1) lead to models with close low skewness and reduced remaining kurtosis. The MT(2)-GARCH(1,1) achieves lowest kurtosis. The

Table 2
Estimated parameters of linear and non-linear, normal and heavy-tailed, as well as mixture GARCH(1,1) volatility models and their standard errors in parentheses, obtained via optimization and sampling over the series of 1500 daily returns on DEM/GBP currency exchange rates.

	a	*b*	α_0	α_1	β
GARCH(1,1)	0.00001	0.0312	0.0089	0.1467	0.8102
	(0.0003)	(0.0058)	(0.0043)	(0.0382)	(0.0441)
TGARCH(1,1)	0.00006	0.0268	0.0097	0.1195	0.8102
	(0.0004)	(0.0065)	(0.0030)	(0.0279)	(0.0441)
NGARCH(1,1)	0.0001	0.0211	0.0196	0.1375	0.8593
	(0.0003)	(0.0048)	(0.0041)	(0.0273)	(0.0382)
NM-GARCH(1,1)	0.0084		0.0198	0.2573	0.6984
	(0.0058)		(0.0086)	(0.0918)	(0.0911)
NM(2)-GARCH(1,1)	0.0004	0.0152	0.0083	0.1245	0.7312
	(0.0002)	(0.0037)	(0.0006)	(0.0321)	(0.0413)
			0.0175	0.1560	0.8412
			(0.0044)	(0.0319)	(0.0452)
MT(2)-GARCH(1,1)	0.0002	0.0199	0.0078	0.1027	0.7105
	(0.0001)	(0.0041)	(0.0005)	(0.0304)	(0.0425)
			0.0213	0.1662	0.8014
			(0.0041)	(0.0378)	(0.0466)

Notes:
- The mean was found by fitting the AR(1) model: $\mu_t = a + br_{t-1}$.
- A common degrees of freedom parameter for all models was pre-computed to facilitate comparisons $\nu = 7.48$.
- The NM-GARCH(1,1) used mixture Gaussian noise with variance $\sigma^2 = 1/(\rho + (1-\rho)/\lambda)$ which is scaled with parameters ρ and λ, so its sampling algorithm inferred also: $\rho = 0.6925$, $\lambda = 0.1994$.
- The models NM(2)-GARCH(1,1) and MT(2)-GARCH(1,1) used a single mean for both of their components.
- The mixing probabilities in NM(2)-GARCH(1,1) were respectively $\rho_1 = 0.74$, and $\rho_2 = (1-\rho_1)$.
- The mixing probabilities in MT(2)-GARCH(1,1) were respectively $\rho_1 = 0.71$, and $\rho_2 = (1-\rho_1)$.

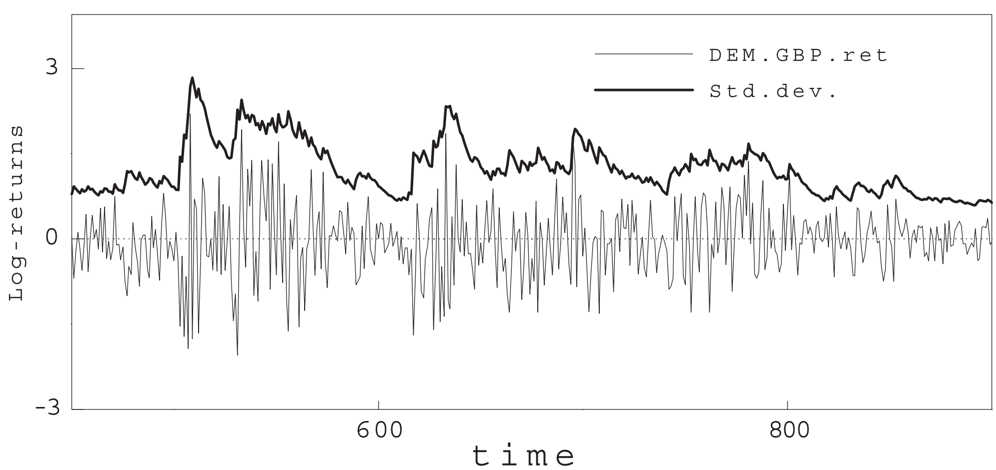

Fig. 5. A segment of the log-returns from the DEM/GBP exchange rates series, recorded from 3/1/1984 to 31/12/1991, and their mixed volatility $\sigma_t^2 = \sum_{i=1}^{N} \rho_i \sigma_{i,t}^2$ obtained using the MT(2)-GARCH(1,1) model.

Table 3
Statistical in-sample diagnostics computed with standardized residuals, obtained by fitting the studied GARCH(1,1) models to the series of daily returns on DEM/GBP using the studied algorithms.

	Skewness	Kurtosis	B–P	L–B (30)	Log-lik.
GARCH(1,1)	−0.3924	5.3105	0.0187	26.4551	−913.42
TGARCH(1,1)	−0.3821	5.2118	0.0185	27.6285	−878.65
NGARCH(1,1)	−0.3845	5.2570	0.0173	27.3067	−869.74
NM-GARCH(1,1)	−0.3966	5.2164	0.0176	26.5763	−857.22
NM(2)-GARCH(1,1)	−0.3889	5.1165	0.0182	27.2915	−848.76
MT(2)-GARCH(1,1)	−0.3902	5.0127	0.0181	27.2854	−782.74

Table 4
Out-of-sample errors of the studied GARCH(1,1) models calculated with one-step-ahead forecasts, obtained via rolling regression over the testing subseries of 474 future unseen DEM/GBP exchange rates data.

	NMSE	*NMAE*	*HR*	*LLOS*	*GMLE*
				$(*10^{-1})$	$(*10^{-1})$
GARCH(1,1)	0.7497	0.8817	0.6751	0.9121	−0.1001
TGARCH(1,1)	0.7511	0.8558	0.6817	0.8461	−0.0976
NGARCH(1,1)	0.7508	0.8744	0.6803	0.8904	−0.0988
NM-GARCH(1,1)	0.7547	0.8722	0.6805	0.8506	−0.1002
NM(2)-GARCH(1,1)	0.7461	0.8913	0.6816	0.9421	−0.1014
MT(2)-GARCH(1,1)	0.7459	0.7639	0.7028	0.8261	−0.0943

values of the Box–Pierce (B–P) and Ljung–Box (L–B) statistics indicate that there is no significant autocorrelation in the residuals. Both NM(2)-GARCH(1,1) and MT(2)-GARCH(1,1) exhibit slightly better statistical characteristics and higher log-likelihoods than the other models. Overall the results in Table 3 indicate that the proposed MT(2)-GARCH(1,1) shows improved distributional behavior. The out-of-sample errors in Table 4 show that the accuracy of the predictions vary across the models.

Our findings can be summarized as follows: (1) the nonlinear NGARCH(1,1) outperforms both linear models GARCH(1,1) and

TGARCH(1,1) with respect to log-likelihood and several statistical characteristics; (2) all mixture models, including NM(2)-GARCH(1,1) and MT(2)-GARCH(1,1), are more likely than the other models as they achieve lower log-likelihoods; (3) the MT(2)-GARCH(1,1) mixture model shows lowest *NMSE* and *NMAE* errors, as well as highest *HR*.

Fig. 5 shows that the deviations of the mixed volatility obtained by MT(2)-GARCH(1,1) wrap closely the observed returns. Fig. 6(a) illustrates the correlogram of the standardized residuals which demonstrates that the remaining autocorrelation in the residuals up to 30 lags is small and close to zero. Fig. 6(b) presents a Gaussian approximation to the histogram of the standardized residuals.

Notes for Table 4:

- The errors were computed with the following formulae (Schittenkopf et al., 2000):

$$NMSE = \sqrt{\sum_{t=1}^{T}(r_t^2 - h_t)^2} \Big/ \sqrt{\sum_{t=1}^{T}(r_t^2 - r_{t-1}^2)^2}$$
$$NMAE = \sum_{t=1}^{T}|r_t^2 - h_t| \Big/ \sum_{t=1}^{T}|r_t^2 - r_{t-1}^2| \quad (20)$$

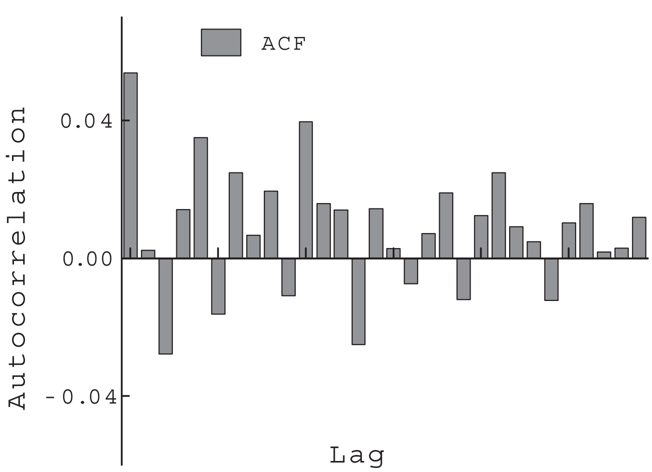

Fig. 6a. Correlogram of the standardized residuals.

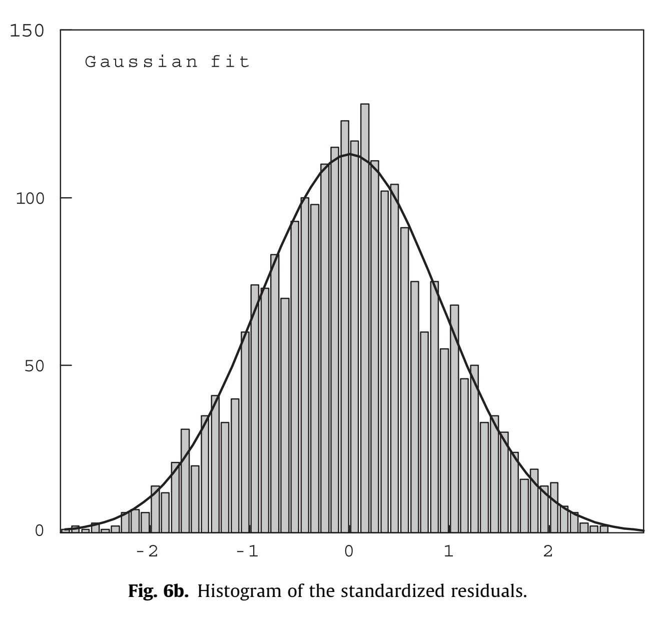

Fig. 6b. Histogram of the standardized residuals.

which relate the volatility estimate h_t to the target return r_t^2 assuming that it is the true volatility from the naive model.

- The hit rate is measured as follows (Schittenkopf et al., 2000):

$$HR = (1/T)\sum_{t=1}^{T}\theta_t$$
$$\theta_t = \begin{cases} 1 & (h_t - r_{t-1}^2)(r_t^2 - r_{t-1}^2) \geqslant 0 \\ 0 & \end{cases} \quad (21)$$

- The logarithmic loss (*LLOS*) function (Pagan and Schwert, 1990) and the Gaussian maximum likelihood (*GMLE*) were obtained by the formulae (Bollerslev et al., 1994):

$$LLOS = (1/T)\sum_{t=1}^{T}(\log e_t^2 - \log h_t)^2$$
$$GMLE = (1/T)\sum_{t=1}^{T}(\log h_t + e_t^2/h_t) \quad (22)$$

where $e_t = (y_t - \mu_t)$ is the error at time *t*, and $h_t = \sigma_t^2$ is the volatility.

4.2.4. Value-at-Risk performance

The usefulness of MT(2)-GARCH(1,1) predictions for managing risk exposure was examined using the Value-at-Risk (VaR) measure (Christoffersen, 2003; Jorion, 1996).

Value-at-Risk measure. The VaR is a risk measure which tells us the maximum loss that may happen with a certain confidence over a given time period because of the fluctuations of market prices. The VaR is associated with the time-varying quantiles of the return distribution (that is, the probability mass of the tails). The quantile of the distribution is the value that will not be exceeded with certain probability. Assuming that the returns come from a heavy-tail Student-*t* distribution, the left quantile at *q*% confidence level (and respectively the right quantile at $(1-q)$%) is defined by:

$$\Pr(r_t \leqslant VaR_t(q)) = q\% = \int_{-\infty}^{VaR} f(r_t)dr_t \quad (23)$$

where $f(\cdot)$ denotes the conditional density of the returns.

VaR values for heavy-tail returns can be computed using the mean μ_t and volatility σ_t forecasts (which in the mixture models are computed by: $\sigma_t^2 = \sum_{m=1}^{M}\rho_{m,t}\sigma_{m,t}^2$) as follows:

$$VaR_t(q) = \mu_t - \sigma_t t_{q,v}/\sqrt{v/(v-2)} \quad (24)$$

where $t_{q,v}$ is the critical value of the Student-*t* distribution with *v* degrees of freedom.

Bootstrapping VaR forecasts. This research performs VaR estimation via bootstrapping [2] to compensate for the uncertainty in the parameters of the GARCH model (Pascual et al., 2006). Bootstrap replicates of the returns $\{r_1^*, r_2^*, \ldots, r_T^*\}$ are produced with the equation: $r_t^* = \mu_t^* + \hat{\epsilon}^*\sqrt{h_t^*}$, $1 \leqslant t \leqslant T$, by drawing samples $\hat{\epsilon}^* = \hat{\epsilon} - mean(\hat{\epsilon})$ from the standardized residuals $\hat{\epsilon}_t = (r_t - \mu_t)/\sqrt{h_t}$. At each time step replicates of the common volatility $h_t^* = \sum_{m=1}^{M}\rho_m h_{m,t}^*$ are made after recursive evaluation of the two component volatilities $h_{1,t}^*$ and $h_{2,t}^*$ with the estimated parameters $\left\{\hat{\alpha}_{m,0}, (\hat{\alpha}_{m,i})_{i=1}^{q}, (\hat{\beta}_{m,j})_{j=1}^{p}\right\}_{m=1}^{M}$. The process starts with $\mu_1^* = \hat{\mu}_1$ and $h_m, t^* = \hat{h}_m, t$. There are also computed the means by $\mu_t^* = \hat{a} + \hat{b}r_{t-1}^*$. Next, the parameters are reestimated over the replicated returns, which yields a new set of adapted

[2] Although the bootstrapped VaR approach overcomes the restrictions of the distributional assumptions and it is numerically reliable (Hartz, Mittnik, & Paolella, 2006), when the dependencies between the returns are more complex alternative Monte Carlo simulations using stochastic sampling (Ausin & Galeano, 2007) may be applied for estimating VaR.

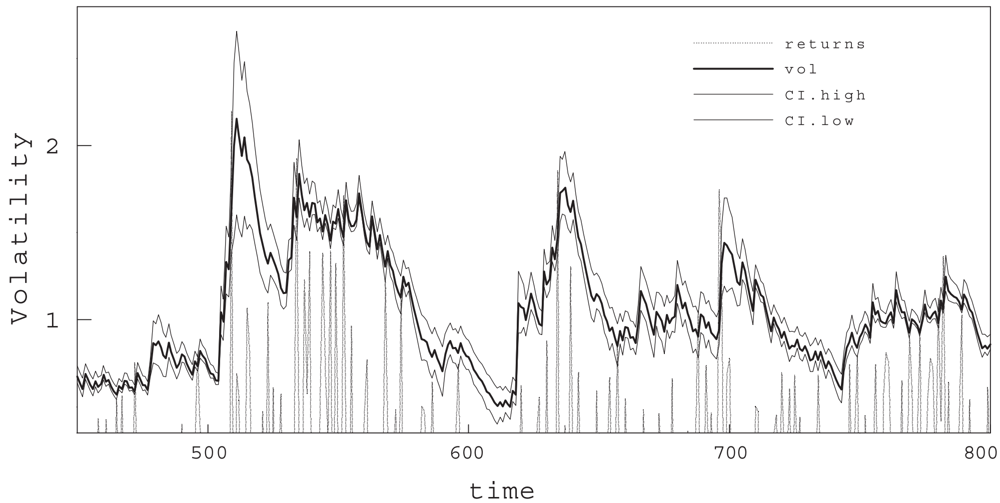

Fig. 7. Bootsrapped 95% confidence intervals averaged over 100 replicas of the common (mixed) volatility inferred by the MT(2)-GARCH(1,1) model over the DEM/GBP exchange rates series.

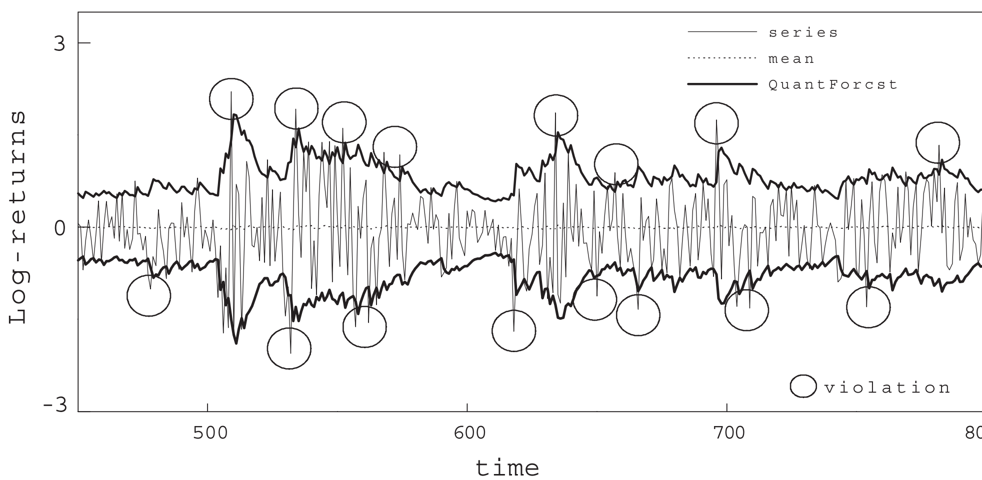

Fig. 8. A segment of the log-returns from the DEM/GBP exchange rates series with the bootstrapped 95% quantile, and some of the violations exceeding the return values (enclosed in circles).

parameters $\hat{a}^*$, $\hat{b}^*$, $\left\{\hat{\alpha}^*_{m,0}, \left(\hat{\alpha}^*_{m,i}\right)_{i=1}^{q}, \left(\hat{\beta}^*_{m,j}\right)_{j=1}^{p}\right\}_{m=1}^{M}$. This process is repeated for a sufficiently large number of times, and from each reestimated model we generate predictions of the mean return and the volatility. Thus, we obtain samples from the entire empirical predictive return distribution at each time step, which allows us to calculate more precisely VaR as the average from sampled empirical quantiles.

Estimating VaR. The effectiveness of the competing models in risk measurement was investigated using bootstrapped one-step ahead predictions of the mean $\{mean(\mu^*_{T+1})\}$ and volatility $\{mean(h^*_{T+1})\}$ of the DEM/GBP series, and calculating semi-parametric VaR forecasts for 99% and 95% confidence levels as recommended (Christoffersen, 2003). The objective was to find out how many of the predicted losses are correct. When the number of exceptions (violations) is too small it is assumed that the model overestimates the risk, and when the number of exceptions is too large then the model underestimates the risk. Fig. 7 illustrates the 95% confidence intervals of the bootsrapped common volatility inferred by the MT(2)-GARCH(1,1) model, which are made from 100 replicates of the return series. The incorrect predictions, called also violations, that these forecasts produce are illustrated in Fig. 8.

We followed the three step approach of Christoffersen (2003). It requires to precompute the failure rate (*FR*) as the proportion of the number of exceptions for the left and the right tails of the return distribution that exceed the corresponding Value-at-Risk. The *FR* for the left tail (long positions) is the proportion of returns smaller than $FR_L = (1/T)\sum_{t=1}^{T} I(r_t \leqslant VaR_t(q))$, respectively the *FR* for the right tail (short positions) is the proportion of returns larger than $FR_S = (1/T)\sum_{t=1}^{T} I(r_t \leqslant VaR_t(1-q))$, where $I(\cdot)$ is the indicator function for membership in a subset. A model is considered adequate when its failures are close to the desired q% level.

First, we computed likelihood ratio statistics (LR_{uc}) to test the unconditional coverage, that is to check whether the number of exceptions is sufficiently close to the predefined significance level. Second, we computed likelihood ratio statistics (LR_{ind}) to test the independence of the exceptions, that is to find out as to what degree they are serially uncorrelated. Third, we computed likelihood ratio statistics (LR_{cc}) to test the conditional coverage, that is to find out whether the model accounts for volatility clustering. The adequacy of the studied volatility models was judged by checking whether these coverage estimates are within the nominal error margin for their corresponding distribution. The first two statistics are distributed as $\chi^2(1)$, with 1% critical value of 6.63 and 5%

Table 5
Results from likelihood-ratio coverage tests for adequacy of the 99% bootsrtapped VaR estimates of the given (in-sample) DEM/GBP series data using one-step-ahead forecasts averaged from 100 replicates. A plus + indicates that the exceptions are out of the nominal margins for the 1% level.

99% daily VaR forecasts				
	FR_L	LR_{un}	LR_{ind}	LR_{cc}
Long position				
GARCH(1,1)	2.0013	11.7621$^+$	2.1489	13.9552$^+$
TGARCH(1,1)	1.2675	0.9985	1.3907	2.4156
NGARCH(1,1)	1.2684	1.0044	1.2294	2.2338
NM-GARCH(1,1)	1.5344	3.7164	0.5171	4.2335
NM(2)-GARCH(1,1)	1.6672	1.9008	2.4274	4.3175
MT(2)-GARCH(1,1)	1.4018	2.1723	0.9479	3.1202
	FR_S	LR_{un}	LR_{ind}	LR_{cc}
Short position				
GARCH(1,1)	0.7341	1.1882	3.3542	4.5503
TGARCH(1,1)	0.4677	5.3625	0.0658	5.3454
NGARCH(1,1)	0.5344	3.9561	0.0863	4.0531
NM-GARCH(1,1)	0.7672	0.2794	3.7163	4.0058
NM(2)-GARCH(1,1)	0.6671	1.9081	2.4274	4.3283
MT(2)-GARCH(1,1)	0.6671	1.9081	2.4274	4.3283

Table 6
Results from likelihood-ratio coverage tests for the adequacy of the 95% bootstrapped VaR estimates of the given (in-sample) DEM/GBP series data using one-step-ahead forecasts averaged from 100 replicates. A plus + indicates that the exceptions are out of the nominal margins for the 5% level.

95% daily VaR forecasts				
	FR_L	LR_{un}	LR_{ind}	LR_{cc}
Long position				
GARCH(1,1)	6.4712	6.2738$^+$	1.8358	8.1096$^+$
TGARCH(1,1)	4.7364	0.2297	2.9419	3.1648
NGARCH(1,1)	4.8699	0.0538	1.6386	1.6925
NM-GARCH(1,1)	5.2734	3.8192	1.9241	5.7434
NM(2)-GARCH(1,1)	6.1374	3.8192	1.9468	5.7657
MT(2)-GARCH(1,1)	5.9373	2.6225	2.1218	5.7114
	FR_S	LR_{un}	LR_{ind}	LR_{cc}
Short position				
GARCH(1,1)	4.2028	2.1155	0.1110	2.2254
TGARCH(1,1)	2.9353	5.6963$^+$	1.4439	7.1402$^+$
NGARCH(1,1)	4.8699	7.2478$^+$	1.0255	8.2735$^+$
NM-GARCH(1,1)	3.8693	4.3612	0.0469	4.4083
NM(2)-GARCH(1,1)	3.4692	5.2478	0.0213	5.2891
MT(2)-GARCH(1,1)	3.8025	4.9125	0.0216	5.2891

Notes for Tables 5 and 6:
- The bootstrapping of all models was carried out using the standardized residuals computed as follows: $\hat{\epsilon}_t = (r_t - \mu_t)/\sqrt{h_t}$.
- The returns were replicated by drawing randomly from the centered residuals $\hat{\epsilon}^* = \hat{\epsilon} - mean(\hat{\epsilon})$ as follows: $r_t^* = \mu_t^* + \hat{\epsilon}^*\sqrt{h_t^*}$, where: $h_t^* = \sum_{m=1}^{M} \rho_m h_{m,t}^*$, starting with initial values $\mu_1^* = \hat{\mu}_1$ and $h_1^* = \hat{h}_1$.
- The means were recursively evaluated by: $\mu_t^* = \hat{a} + \hat{b}r_{t-1}^*$.
- Bootstrap replicates of the volatility were generated using the estimated parameters $\left\{\hat{\alpha}_{m,0}, (\hat{\alpha}_{m,i})_{i=1}^{q}, (\hat{\beta}_{m,j})_{j=1}^{p}\right\}_{m=1}^{M}$ in the volatility equation: $h_{m,t}^* = \hat{\alpha}_{m,0} + \sum_{i=1}^{q} \hat{\alpha}_{m,i}\left(r_{t-i}^* - \mu_{t-i}^*\right)^2 + \sum_{j=1}^{p} \hat{\beta}_{m,j} h_{m,t-j}^*$.
- The bootstrapped VaR were obtained after re-estimating the model parameters: $\hat{a}^*,\ \hat{b}^*,\ \left\{\hat{\alpha}_{m,0}^*, \left(\hat{\alpha}_{m,i}^*\right)_{i=1}^{q}, \left(\hat{\beta}_{m,j}^*\right)_{j=1}^{p}\right\}_{m=1}^{M}$ over the replicated return series $\{r_1^*, r_2^*, \ldots, r_T^*\}$, and computing the average of the one-step ahead forecasts produced with them over the given time interval.

critical value of 3.84. The third statistic is distributed as $\chi^2(2)$, with 1% critical value of 9.21 and 5% critical value of 5.99.

Tables 5 and 6 report the empirical results from such likelihood-ratio coverage tests, which are useful for analyzing whether the produced VaR represent adequately the selected quantiles. These results are (in-sample) bootstrapped VaR predictions obtained over the DEM/GBP series using averaged one-step-ahead forecasts of the mean and the volatility. The heavy-tailed MT(2)-GARCH(1,1) shows quite similar performance, but it does not outperform significantly the other models on 99% VaR estimation, while it is only slightly better on 95% VaR estimation. The heavy-tail mixture MT(2)-GARCH(1,1) is almost equivalent to the normal mixture NM(2)-GARCH(1,1) model on 95% VaR estimation, and both of them outperform all remaining models. The advantage of MT(2)-GARCH(1,1) is its reliability as it generates almost equally good 1% and 5% VaR estimates in long as well as short trading positions. One reason for having better MT(2)-GARCH(1,1) results with less bias in VaR estimation is their flexibility for capturing better the tails of the return distribution.

The results summarized in Tables 5 and 6 indicate that standard linear GARCH(1,1) does not pass successfully some of the tests. The linear GARCH(1,1) models seem inadequate on predicting risk as they lead to unacceptably large failure rates on 99% and 95% VaR estimation for the left tail of the return distribution. More precisely, the independent likelihood ratio statistic LR_{ind} = 11.7621 and the conditional coverage statistic LR_{cc} = 11.7621 for 99% VaR, as well as the LR_{ind} = 6.2738 and LR_{cc} = 8.1096 for 95% VaR are outside of the acceptance boundaries. We are inclined to think that the Gaussian GARCH(1,1) models tend to overestimate the VaR. The linear TGARCH(1,1) model is also not always acceptable, as it has been found that it produced exceeding 95% VaR estimates but for the right tail, showing respectively LR_{un} = 7.2478 and LR_{cc} = 8.2735 for 95% time-varying VaR.

5. Conclusion

This paper presented a heavy-tailed mixture density approach to GARCH modeling of the dynamic evolution of the conditional variance in financial series of returns on prices. We demonstrated empirically that the EM algorithm can infer flexible MT-GARCH models of the time-varying conditional moments of the returns without imposing restrictions on the returns. The proposed heavy-tail mixture model MT(2)-GARCH(1,1) outperforms in statistical terms typical linear and nonlinear, as well as normal mixture GARCH models. The MT(2)-GARCH(1,1) model generates reliable VaR forecasts and seems a promising tool for assessment of the risk exposure and, thus, mitigating future financial risks.

The research continues with explorations multi-step predictions (Boshnakov, 2009), and with investigations into whether the MT(2)-GARCH(1,1) may capture better the conditional heteroskedasticity than the similar nonlinear Markov Regime-Switching MRS-GARCH in which the structural breaks in the parameters are also controlled through constant transition probabilities but with Markovian evolution.

Acknowledgments

The authors thank Professor Pedro Galeano from University of Santiago de Compostela, Spain for providing his original implementation of the *MCMC* sampling algorithm for estimating the NM-GARCH(1,1) model. We also thank Dr. Peter Tino from The University of Birmingham for the helpful discussions on the real-time training algorithm for recurrent networks.

Appendix A. Likelihood calculations

The formulation of the estimation algorithm for finding the MT(m)-GARCH(p,q) mixture model parameters $\Theta = \left\{\left((w_{mi})_{i=0}^{W}\right)_{m=1}^{M}, (\rho_m)_{m=1}^{M}, (v_m)_{m=1}^{M}\right\}$ requires to determine the gradients of the expected total log-likelihood $E[\log L(\Theta)|R_t, \hat{\Theta}_{(k)}]$ with respect to each of them.

Solving for the mixing coefficients. The derivative of the expected log-likelihood with respect to a mixing coefficient ρ_m $(1 \leqslant m \leqslant M)$ is computed via a technique using Lagrange multipliers λ with constraint $\sum_{m=1}^{M}\rho_m = 1$ as follows:

$$\frac{\partial L1_{m,t}(\hat{\Theta}_{(k)})}{\partial \rho_m} = \frac{\partial\left[\sum_{t=1}^{T}\sum_{m=1}^{M} z_{m,t}\log\rho_m - \lambda\left(\sum_{m=1}^{M}\rho_m - 1\right)\right]}{\partial \rho_m} = \sum_{t=1}^{T}\frac{z_{m,t}}{\rho_m} - \lambda \tag{A.1}$$

The Lagrange multiplier λ can be derived by summation $\lambda = \sum_{t=1}^{T}\sum_{m=1}^{M} z_{m,t} = \sum_{t=1}^{T} 1 = T$. After setting to zero and solving, we obtain Eq. (17) for the mixing coefficient.

Likelihood derivatives of the volatility parameters. The optimization of the volatility involves finding the first-order derivatives of the likelihood term $E[l2_{m,t}(\hat{\Theta}_{(k)})]$ with respect to each parameter w_{mj} $(1 \leqslant m \leqslant M, 0 \leqslant j \leqslant (p+q))$. We apply the chain rule and calculate these derivatives following the real-time recurrent learning algorithm (Williams & Zipser, 1995) in the following way:

$$\frac{\partial E[l2_{m,t}(\hat{\Theta}_{(k)})]}{\partial w_{mj}} = \frac{\partial E[l2_{m,t}(\hat{\Theta}_{(k)})]}{\partial h_{m,t}}\frac{\partial h_{m,t}}{\partial w_{mj}}$$

The differentiation of the this log-likelihood term with respect to the regime volatilities $h_{m,t}$ at a particular time step is carried out as follows:

$$\begin{aligned}\frac{\partial E[l2_{m,t}(\hat{\Theta}_{(k)})]}{\partial h_{m,t}} &= \frac{z_{m,t}\partial[\log h_{m,t} + u_{m,t}(r_t - \mu_t)^2/h_{m,t}]}{\partial h_{m,t}} \\ &= z_{m,t}\left[\frac{1}{h_{m,t}} - \frac{u_{m,t}(r_t-\mu_t)^2}{h_{m,t}^2}\right] \\ &= \frac{z_{m,t}}{h_{m,t}}\left[1 - \frac{u_{m,t}(r_t-\mu_t)^2}{h_{m,t}}\right]\end{aligned} \tag{A.2}$$

The derivative of the volatility $h_{m,t}$ with respect to a weight w_{mj} $(1 \leqslant m \leqslant M, 0 \leqslant j \leqslant (p+q))$ is also computed using the chain rule:

$$\begin{aligned}\frac{\partial h_{m,t}}{\partial w_{mj}} &= \frac{\partial h_{m,t}}{\partial y_{m,t}}\frac{\partial y_{m,t}}{\partial w_{mj}} = g'(y_{m,t})\frac{\partial\left[\sum_{l=0}^{p+q} w_{ml}x_{t-l}\right]}{\partial w_{mj}} \\ &= g'(y_{m,t})\sum_{l=0}^{p+q}\left[w_{ml}\frac{\partial x_{t-l}}{\partial w_{mj}} + x_{t-l}\frac{\partial w_{ml}}{\partial w_{mj}}\right] \\ &= g'(y_{m,t})\left(\sum_{l=1}^{p}\left[w_{ml}\frac{\partial e_{t-l}^2}{\partial w_{mj}}\right] + \sum_{l=p+1}^{p+q}\left[w_{ml}\frac{\partial h_{m,t-l+p}}{\partial w_{mj}}\right] + x_{t-j}\right) \\ &= g'(y_{m,t})\left(\sum_{l=p+1}^{p+q}\left[w_{ml}\frac{\partial h_{m,t-l+p}}{\partial w_{mj}}\right] + x_{t-j}\right)\end{aligned} \tag{A.3}$$

where w_{ml} are the weights on the recurrent connections, and we have used the fact that $\partial e_{t-l}^2/\partial w_{mj} = 0$.

References

Alexander, C., & Lazar, E. (2006). Normal mixture GARCH(1,1): Applications to exchange rate modelling. *Journal of Applied Econometrics, 21*, 307–336.

Ardia, D. (2008). *Financial risk management with Bayesian estimation of GARCH models: Theory and applications*. Berlin: Springer-Verlag.

Ardia, D., Hoogerheide, L. F., & van Dijk, H. K. (2009). Adaptive mixture of student-*t* distributions as a flexible candidate distribution for efficient simulation. *Journal of Statistical Software, 29*, 1–32.

Ausin, M. C., & Galeano, P. (2007). Bayesian estimation of the Gaussian mixture GARCH model. *Computational Statistics and Data Analysis, 51*, 2636–2652.

Bauwens, L., Hafner, C. M., & Laurent, S. (2012). *Handbook of volatility models and their applications*. Hoboken, NJ: John Wiley and Sons.

Bildirici, M., & Ersin, O. O. (2009). Improving forecasts of GARCH family models with the artificial neural networks. *Expert Systems with Applications, 36*, 7355–7362.

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

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). ARCH models. In R. F. Engle & D. McFadden (Eds.). *The handbook of econometrics* (Vol. 4, pp. 2959–3038). Amsterdam: North-Holland.

Boshnakov, G. N. (2009). Analytic expressions for predictive distributions in mixture autoregressive models. *Statistics and Probability Letters, 79*, 1704–1709.

Brooks, C., Burke, S. P., & Persand, G. (2001). Benchmarks and the accuracy of GARCH model estimation. *Journal of Forecasting, 17*, 45–56.

Christoffersen, P. (2003). *Elements of financial risk management*. San Diego, CA: Academic Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B, 39*, 1–38.

Deschamps, P. J. (2012). Bayesian estimation of generalized hyperbolic skewed student GARCH models. *Computational Statistics and Data Analysis, 56*, 3035–3054.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica, 50*, 987–1007.

Fiorentini, G., Calzolari, G., & Panattoni, L. (1996). Analytical derivatives and the computation of GARCH estimates. *Journal of Applied Econometrics, 11*, 399–417.

Fletcher, R. (1987). *Practical methods for optimization* (2nd ed.). New York: John Wiley and Sons.

Haas, M., Mittnik, S., & Paolella, M. S. (2004). Mixed normal conditional heteroskedasticity. *Journal of Financial Econometrics, 2*, 211–250.

Hartz, C., Mittnik, S., & Paolella, M. S. (2006). Accurate value-at-risk forecasts with the normal-GARCH model. *Computational Statistics and Data Analysis, 51*, 2295–2312.

Jorion, P. (1996). *Value at risk: The new benchmark for managing financial risk*. New York: The McGraw-Hill Co..

Kim, Y. S., Rachev, S. T., Bianchi, M. L., Mitov, I., & Fabozzi, F. J. (2011). Time series analysis for market meltdowns. *Journal of Banking and Finance, 35*, 1879–1891.

Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies, 65*, 361–393.

Lee, Y-S., & Pai, T-Y. (2010). REIT volatility prediction for skew-GED distribution of the GARCH model. *Expert Systems with Applications, 37*, 4737–4741.

Miazhynskaia, T., Dorffner, G., & Dockner, E. J. (2003). Risk management application of the recurrent mixture density network models. In *International conference on artificial neural networks, ICANN-2003. LNCS* (Vol. 2714, pp. 589–596). Berlin: Springer.

Nikolaev, N., Tino, P., & Smirnov, E. N. (2011). Time-dependent series variance estimation via recurrent neural networks. In T. Honkela et al. (Eds.), *Proceedings of international conference on artificial neural networks, ICANN-2011. LNCS* (Vol. 6971, pp. 176–184). Berlin: Springer.

Orhan, M., & Köksal, B. (2012). A comparison of GARCH models for VaR estimation. *Expert Systems with Applications, 39*, 3582–3592.

Pascual, L., Romo, J., & Ruiz, E. (2006). Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics and Data Analysis, 50*, 2293–2312.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the *t* distribution. *Statistics and Computing, 10*, 339–348.

Pérez-Cruz, F., Afonso-Rodriguez, J. A., & Giner, J. (2003). Estimating GARCH models using support vector machines. *Journal of Quantitative Finance, 3*, 163–172.

Schittenkopf, C., Dorffner, G., & Dockner, E. J. (2000). Forecasting time-dependent conditional densities: A semi non-parametric neural network. *Journal of Forecasting, 19*, 355–374.

Sefidpour, A., & Bouguila, N. (2012). Spatial color image segmentation based on finite non-Gaussian mixture models. *Expert Systems with Applications, 39*, 8993–9001.

Tang, L. B., Sheng, H. Y., & Tang, L. X. (2009). Forecasting volatility based on wavelet support vector machine. *Expert Systems with Applications, 36*, 2901–2909.

Wang, Y-H. (2009). Nonlinear neural network forecasting model for stock index option price: Hybrid GJR–GARCH approach. *Expert Systems with Applications, 36*, 564–570.

Williams, R. J., & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y. Chauvin & D. E. Rumelhart (Eds.), *Back-propagation: Theory, architectures and applications* (pp. 433–486). Hillsdale, NJ: Lawrence Erlbaum Publications.

Wilmhelmsson, A. (2006). GARCH forecasting performance under different distribution assumptions. *Journal of Forecasting, 25*, 561–578.

Wong, C. S., Chan, W. S., & Kam, P. L. (2009). A student-*t* mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika, 96*, 751–760.

Wong, C. S., & Li, W. K. (2001). On a mixture autoregressive conditional heteroskedastic model. *Journal of the American Statistical Association, 96*, 955–982.

Zeevi, A. J., & Meir, R. (1997). Density estimation through convex combinations of densities: Approximation and error bounds. *Neural Networks, 10*, 99–109.