ELSEVIER Journal of Econometrics 106 (2002) 119–142

JOURNAL OF
Econometrics

www.elsevier.com/locate/econbase

Entropy densities with an application to autoregressive conditional skewness and kurtosis

Michael Rockinger[a,*], Eric Jondeau[b]

[a]*Department of Finance, HEC-School of Management, 1, rue de la Libération, 78351 Jouy-en-Josas, France*

[b]*Banque de France, DEER, 41-1391 Centre de Recherche, 31, rue Croix des Petits Champs, 75049 Paris, France*

Received 4 January 2000; revised 4 January 2001; accepted 30 April 2001

Abstract

The entropy principle yields, for a given set of moments, a density that involves the smallest amount of prior information. We first show how entropy densities may be constructed in a numerically efficient way as the minimization of a potential. Next, for the case where the first four moments are given, we characterize the skewness–kurtosis domain for which densities are defined. This domain is found to be much larger than for Hermite or Edgeworth expansions. Last, we show how this technique can be used to estimate a GARCH model where skewness and kurtosis are time varying. We find that there is little predictability of skewness and kurtosis for weekly data. © 2002 Elsevier Science S.A. All rights reserved.

JEL classification: C40; C61; G10

Keywords: Semi-nonparametric estimation; Time-varying skewness and kurtosis; GARCH

1. Introduction

Methods based on the entropy principle of Shannon (1948), and popularized by Jaynes (1957, 1982), have made their way into econometrics, e.g. Golan et al. (1996). At a practical level, entropy-based applications still appear to be scarce but for a few exceptions such as Zellner and Highfield

* Corresponding author. Tel.: +33-1-39-677259; fax: +33-1-39-677085.
E-mail address: rockinger@hec.fr (M. Rockinger).

PII: S0304-4076(01)00092-6

(1988), Hawkins et al. (1996), Stutzer (1996), Buchen and Kelly (1996), Zellner et al. (1997), or Ormoneit and White (1999). One possible reason is that difficulties with the numerical implementation of this technique may have hindered its widespread use. The aim of this work is to develop a very fast method to obtain entropy densities and to show that entropy densities may also be used in rather complex empirical likelihood estimations.

In a numerical application, we reconsider Bollerslev's (1986) GARCH model which extends Engle (1982).[1] In typical applications of this model, the unconditional distribution is assumed to allow for some form of fat-tailedness, modeled for instance as a Student-*t* (Bollerslev, 1987), a generalized error distribution (Nelson, 1991), or as a fully nonparametric density (e.g. Engle and Gonzales-Rivera, 1991). Recent applications to finance, dealing with the issue of conditional fat-tailedness, involve models with a noncentral gamma distribution (Harvey and Siddique, 1999) or a generalized Student-*t* (Hansen, 1994), where the degree of freedom and the asymmetry parameter are time varying. There, the time variability is achieved by expressing the degree of freedom or an asymmetry parameter as a function of actual data. As an alternative to these densities we propose the use of an entropy density (ED). The advantage of the ED is that skewness and kurtosis appear directly as parameters. As a consequence, to obtain the value of skewness and kurtosis, it is not necessary to compute additional functions of more primitive parameters.

ED can also be conveniently used in nonparametric econometrics or in financial applications such as in modeling the pricing kernel arising in Euler equations.[2]

A better description of the conditional behavior of asset returns with a particular emphasis on the time-variability of skewness and kurtosis is of great relevance for risk management as well as for asset allocation problems. An econometric description involving higher moments may also have important implications for the testing of asset pricing models (e.g. Kraus and Litzenberger, 1976). Improvements to the existing econometric literature on the time-variability of higher moments are, therefore, relevant. A possible reason why little progress has been made in financial applications is that there exist only very few densities where skewness and kurtosis appear directly as parameters. An exception are the Gram–Charlier and Edgeworth expansions. The skewness and kurtosis domain of these densities has been investigated by Jondeau and Rockinger (2001) and has been found to be too small to correctly describe financial returns. In this work we demonstrate that

[1] See also Bera and Higgins (1992) or Bollerslev et al. (1994) for surveys of the large literature dealing with this type of model.

[2] See for instance Gallant and Tauchen (1989) for an application involving a nonparametric estimation of a density within an Euler equation.

entropy-based methods allow for a very large range of possible values for the parameters. This implies that greater numerical stability will be achieved during the estimation, since the latter may be performed for less restricted parameters.

This method's gain in flexibility does not come for free since the construction of an ED from its moments involves a numerical optimization. In general, numerical optimization is a very time consuming process. However, given the special nature of the entropy problem, it is possible to construct EDs with only a few numerical iterations (e.g. Alhassid et al., 1978; Agmon et al., 1979a, b as well as Mead and Papanicolaou, 1984) by mapping the problem into a minimization of a very well behaved potential function.

The structure of this paper is as follows. In the next section, we provide theoretical considerations concerning EDs. In Section 3, we introduce a model, in the spirit of Hansen (1994) where we allow for time-varying parameters. In Section 4, we present the empirical results. The last section contains a conclusion.

2. Theoretical background

2.1. The definition of entropy densities

We assume that the econometrician is seeking a probability $p(x)$ defined over some real convex domain, $\mathscr{D}$, while disposing only of information on the m first moments of the probability, written as b_i where $i = 1, \ldots, m$. The construction of a probability density defined on infinitely many points with the knowledge of only a few moments is hopeless without an additional criterion. A first possibility to obtain a density, matching the given moments, is to use ad-hoc step functions. Such an approach is implemented by Wheeler and Gordon (1969). Another criterion is given by the maximization of an entropy under the moment and density restrictions. Under this criterion one solves

$$p \in \arg\max - \int_{x \in \mathscr{D}} p(x) \ln(p(x)) \, \mathrm{d}x, \tag{1}$$

$$\text{s.t.} \quad \int_{x \in \mathscr{D}} p(x) \, \mathrm{d}x = 1, \tag{2}$$

$$\int_{x \in \mathscr{D}} x^i p(x) \, \mathrm{d}x = b_i, \quad i = 1, \ldots, m. \tag{3}$$

We will refer to a density satisfying these conditions as an Entropy Density.[3] Jaynes (1957) notices that the entropy is a criterion where the statistician

[3] Given that a log-function is involved in (1), $p(x) \geqslant 0$, $\forall x \in \mathscr{D}$.

imposes a minimum amount of information. The conventional way of solving this program is to define the Hamiltonian

$$H = -\int_{\mathscr{D}} p(x)\ln(p(x))\,\mathrm{d}x - \lambda_0'' \int_{\mathscr{D}} p(x)\,\mathrm{d}x$$
$$-\sum_{i=1}^{m} \lambda_i \left[\int_{\mathscr{D}} x^i p(x)\,\mathrm{d}x - b_i \right].$$

The λ_0'' is a Lagrange parameter[4] as are the λ_i, $i = 1,\ldots,m$.

To obtain a solution of this problem one seeks a zero for the Fréchet derivative. Defining $\lambda_0' = \lambda_0'' + 1$ we get

$$\delta H = 0 \Rightarrow p(x) = \exp\left(-\lambda_0' - \sum_{i=1}^{m} \lambda_i x^i \right). \tag{4}$$

Derivation with respect to the $m+1$ Lagrange multipliers yields the $m+1$ conditions (2)–(3).

Eq. (4) shows that the density will belong to the Pearsonian family.[5] For small values of m, it is possible to obtain explicit solutions. If $m = 0$, meaning that no information is given, beyond the fact that one seeks a density, then one obtains the uniform distribution over $\mathscr{D}$. As one adds the first and second moments, Golan (1996) recall that one obtains the exponential, and the normal density. The knowledge of the third or higher moment does not yield a density in closed form. Only numerical solutions may provide densities. In this work, we show how densities may be obtained in a numerically efficient manner if third and higher moments are given. This work extends, therefore, Zellner and Highfield (1988) as well as Ormoneit and White (1999) by providing a more efficient estimation technique.

Substitution of (4) into (2) defines a function that turns out to be a potential function, as shown later. The expression of this function is

$$P(\lambda_1,\ldots,\lambda_m) \equiv \exp(-\lambda_0') = \int_{\mathscr{D}} \exp\left(\sum_{i=1}^{m} \lambda_i x^i \right) \mathrm{d}x \tag{5}$$

so that

$$p(x) = \exp\left(\sum_{i=1}^{m} \lambda_i x^i \right) \Big/ P(\lambda_1,\ldots,\lambda_m). \tag{6}$$

For a given set of $\lambda = (\lambda_1,\ldots,\lambda_m)'$, one could evaluate (6) and, thus, the moment restrictions (3). This suggests as a first estimation technique nonlinear least squares (NLLS) applied to (3). As we rediscovered painfully, such

[4] The double prime has been introduced for notational convenience only.

[5] See, for instance, Johnson et al. (1994).

an estimation yields multiple solutions and is rather slow.[6] As discovered by Agmon et al. (1979a, b), a faster and numerically stable procedure is available. This procedure uses the physical properties of the entropy definition. In order to use this procedure, it is convenient to introduce further results.

Since $\int_{\mathscr{D}} p(x)\,\mathrm{d}x = 1$, multiplication of the right-hand side of (3) by this integral and the grouping under one single integral yields

$$\int_{\mathscr{D}} (x^i - b_i)p(x)\,\mathrm{d}x = 0, \quad i = 1,\dots,m.$$

Furthermore, writing $p(x) = \exp(\lambda_0 + \sum_{i=1}^m \lambda_i(x^i - b_i))$, where $\lambda_0 = \lambda_0' + \sum_{i=1}^m \lambda_i b_i$ indicates that the number of computations required to evaluate (6) subject to (3) may be reduced. Also, the passage from λ_0' to λ_0 is a trivial linear transformation. Again, $p(x)$ must satisfy (3) and this yields a definition for λ_0:

$$Q(\lambda_1,\dots,\lambda_m) \equiv \exp(-\lambda_0) = \int_{\mathscr{D}} \exp\left(\sum_{i=1}^m \lambda_i(x^i - b_i)\right)\mathrm{d}x. \tag{7}$$

So that the probability can be rewritten as

$$p(x) = \exp\left(\sum_{i=1}^m \lambda_i(x^i - b_i)\right) \bigg/ Q(\lambda_1,\dots,\lambda_m). \tag{8}$$

At this point we have obtained two equivalent definitions for the density, namely Eqs. (6) and (8). Depending on the situation, one definition or the other is useful.

With the definition of (7), we obtain that

$$g_i \equiv \frac{\partial Q}{\partial \lambda_i} = 0 \Rightarrow \int_{\mathscr{D}} (x^i - b_i)p(x)\,\mathrm{d}x = 0$$

and, therefore, the zeros of the gradient of Q yield the first-order conditions. This computation validates the claim that Q defines a potential.[7] Next, we obtain that

$$G_{ij} \equiv \frac{\partial^2 Q}{\partial \lambda_i \partial \lambda_j} = \int_{\mathscr{D}} (x^i - b_i)(x^j - b_j)p(x)\,\mathrm{d}x,$$

showing that the Hessian matrix is a variance–covariance matrix.[8] As a consequence the Hessian matrix is symmetric and positive definite. An inverse

[6] The technique developed by Ormoneit and White (1999) follows, however, this approach. They show how such an NLLS algorithm may be implemented more efficiently as in Zellner and Highfield (1988), yet, they report estimations lasting several seconds whereas ours takes a fraction of a second.

[7] If U is an open subset of $\mathscr{R}^n$ a map f from U into $\mathscr{R}^n$ is called a *vector field*. For instance, if F is a scalar function from U into $\mathscr{R}$, then $f = \operatorname{grad} F$ defines a vector field. If for a given vector field f there exists a scalar function F such that $f = \operatorname{grad} F$ then F is a *potential function* and the vector field f is said to derive from a potential.

[8] See also Alhassid et al. (1978).

of the Hessian will exist if the matrix is of full rank. This last condition implies that, as long as $p(x)$ is a density, the minimization of Q has a unique solution. We write the gradient of Q as g and his Hessian matrix as G.

At this stage, we have obtained the first key result, namely that the minimization of the potential function Q will yield a density satisfying moment the conditions. We insist on the fact that the key step to obtain a solution resides in a minimization rather than in a search for a zero of a map. It turns out, that, numerically, the minimization is well defined, whereas the search for a zero may even yield multiple solutions. The problem will be numerically stable if Q is of full rank and if the solution is finite.

As Agmon et al. (1979a) point out, it is not guaranteed that the minimization of the potential function will occur at finite distance. It is possible to guarantee finiteness of the solution, but to do so it is first necessary to define how to compute the integrals involved. We turn to this issue now.

2.2. Gauss–Legendre approximation of the integrals

The construction of Q always involves the computation of an integral. For numerical purposes, it is convenient to assume finiteness of $\mathscr{D}$. Under the assumption that $\mathscr{D}$ is a finite interval $[l,u]$, the affine function

$$z=[2x-(u+l)]/(u-l)$$

will map $x\in[l,u]$ into $z\in[-1,1]$. The Jacobian is $(u-l)/2$. In this case, using a generic notation, all our integrals change from

$$\int_l^u h(x)\,\mathrm{d}x \text{ to } \int_{-1}^1 \frac{u-l}{2}h\left(\frac{1}{2}[z(u-l)+(u+l)]\right)\mathrm{d}z=\int_{-1}^1 \tilde{h}(z)\,\mathrm{d}z.$$

This last integral may now be approximated using a Gauss–Legendre quadrature (e.g. Davis and Polonsky, 1970), that is

$$\int_{-1}^1 \tilde{h}(z)\,\mathrm{d}z \sim \sum_{j=1}^n \tilde{h}(z_j)w_j,$$

w_j are the Gauss–Legendre weights and z_j are the abscissa, in $[-1,1]$, where the integrand should be evaluated. Those values are tabulated, for instance, in Abramowitz and Stegun (1970).

We may now return to the computation of the entropy density. For given $z_j\in[-1,1]$, $j=1,\ldots,n$ and boundaries l,u of the domain $\mathscr{D}$, we may use

$$x_j=[(u-l)z_j+(u+l)]/2 \tag{9}$$

to obtain the equivalent of z_j in the domain $\mathscr{D}$.

With this approximation, we face the problem of minimizing the potential

$$Q(\lambda)=\sum_{j=1}^m \exp\left(\sum_{i=1}^m \lambda_i(x_j^i-b_i)\right)w_j.$$

To guarantee that the Hessian is of full rank, given the way the (x_j, w_j) are obtained, it is necessary to have $2n > m$. Under this condition, even if the problem is symmetrical (for instance because the mean and skewness is 0), the Hessian will be well defined.

Introducing a matrix A with elements $a_{ij} \equiv x_j^i - b_i$, $i = 1, \ldots, m$; $j = 1, \ldots, n$, $n > m$, we obtain $Q(\lambda) = w' \exp(A\lambda)$ where $w' = (w_1, \ldots, w_n)$ is a row vector with the n weights.[9] This expression shows that, in numerical applications, the evaluation of Q can be vectorized and rendered very fast.

The minimization will yield a solution, since, under the stated assumptions, the matrix $A'A$ is of full rank. This follows from the fact that the transformation from z_j into x_j is not degenerate.

To obtain finiteness of the solution, Agmon et al. (1979a, b) point out that, for any direction taken, $Q(\lambda)$ should increase to infinity as λ gets large. However, this condition is difficult to implement and an alternative consists in verifying the existence of a solution to conditions (2)–(3), i.e. obtaining $p_i \geqslant 0$, $i = 1, \ldots, n$.

2.3. Numerical implementation

At this stage, we wish to show how the existence of a finite solution can be guaranteed. The discretization of (2) and (3) yields

$$\sum_{j=1}^{n} w_j p_j = 1, \tag{10}$$

$$\sum_{j=1}^{n} w_j x_j^i p_j = b_i, \quad i = 1, \ldots, m, \tag{11}$$

$$p_j \geqslant 0, \quad j = 1, \ldots, n. \tag{12}$$

This set of equations can be viewed as a linear programming problem where one seeks a solution to $m+1$ equations under positivity constraints. We solve this problem with the phase I step of the simplex algorithm (see for instance Press et al., 1999). If a solution exists, the algorithm will find it within $m+1$ and $2(m+1)$ steps.

If a solution exists, then it is known that Q will be minimized for some finite solution. The problem, then, is one of numerically minimizing $Q(\lambda)$. As pointed out by Fletcher (1994), many algorithms are available. However, if the problem is known to have a single minimum, as it will be the case in this framework, Newton's method works well. It is this method that we implement to minimize the potential.

[9] We interpret the vector λ as a column vector.

Algorithm 1.

1. We first need to define the domain $\mathscr{D}$ over which the density will be defined. We restrict ourselves to the range $[l,u]$. Below we discuss how l and u may be obtained. We use a $n=40$ point gaussian quadrature. This quadrature associates to the points $z_j \in [-1,1]$ the weights w_j, $j=1,\ldots,n$, where $n=40$.[10]

2. Using (9) we map the z_j into x_j. We also define the matrix A whose jth line contains $(y_j, y_j^2-b_1,\ldots,y_j^m-b_m)$. We recall that $Q(\lambda)=w'\exp(A\lambda)$ where $w'=(w_1,\ldots,w_n)$.

3. Set $k=0$. Use as a starting value $\lambda^{(0)}=(0,\ldots,0)$ the vector with m zeros.

4. At step k, set $g_i^{(k)}=\partial Q(\lambda^{(k-1)})/\partial\lambda_i$ and $G_{ij}^{(k)}=\partial^2 Q(\lambda^{(k-1)})/\partial\lambda_i\partial\lambda_j$. The element $g_i^{(k)}$ will be the ith element of a column vector $g^{(k)}$ with m components. Similarly, $G_{ij}^{(k)}$ is the jth element of the ith line of the matrix $G^{(k)}$.

5. Let $\delta^{(k)}$ be the solution to $G^{(k)}\delta^{(k)}=-g^{(k)}$.

6. Update the vector of Lagrange multipliers $\lambda^{(k)}=\lambda^{(k-1)}+\delta^{(k)}$.

7. Set $k=k+1$ and return to 4 unless a required accuracy has been obtained.

In step 1 of the algorithm, it is necessary to choose the bounds l and u. This choice is relatively easy if the entropy density is used in an empirical likelihood context. It suffices to choose boundaries somewhat larger than the range of studentized data. If one is interested in the general construction of an ED, a possible criterion is based on the accuracy of the approximation. This accuracy may be computed with a numerical integration of the various moments using the estimated ED. This computation is also a verification that the number of abscissa z_j used in the gaussian quadrature is sufficient.

The Newton algorithm is based on the observation that if $G^{(k)}\delta^{(k)}=-g^{(k)}$ then the approximation in a second-order Taylor expansion of Q, that is $Q(\lambda^{(k-1)}+\delta^{(k)})=Q(\lambda^{k-1})+\delta^{(k)}g^{(k)}+\frac{1}{2}\delta^{(k)\prime}G^{(k)}\delta^{(k)}$, leads to a *flat* spot of Q, that is an extremum. In step 6 of the algorithm, a typical criterion to stop iterating is given by the Euclidean norm of the vector $g^{(k)}$. For most cases considered in this work the algorithm converged within 10 iterations with a precision of the gradient $g^{(k)}$ smaller than 10^{-6}. This speed is remarkable and makes it possible to use the entropy densities in situations that were not possible before. Once the parameters λ have been obtained the value of the ED at some point x may be obtained using Eq. (8).

Agmon et al. (1979b) also suggest the use of an orthogonalized A matrix. We followed their suggestion and included in our algorithm a Gram–Schmidt

[10] The (z_j,x_j) for $j=1,\ldots,n$ are tabulated for values of n up to 96 in Abramowitz and Stegun (1970). We found that for our problems $n=40$ is sufficient.

orthogonalization. For the problem at hand, such an orthogonalization did not lead to an improvement of the speed of convergence towards an optimum.

2.4. Entropy densities for a given skewness and kurtosis

In statistical applications, it is easy to standardize a given sample r_t, $t=1,\dots,T$, by subtracting its mean and dividing by its standard deviation. For this reason, we focus now, without loss of generality, on the study of those densities that satisfy $b_1=0$, $b_2=1$, $b_3=s$, and $b_4=k$. In this case, the parameters s and k represent skewness and kurtosis, respectively. Since a solution to our problem, defined by Eqs. (10)–(12), exists only if the simplex phase I problem is well behaved, we start with a rough grid-search over a large skewness–kurtosis domain where a solution to the simplex algorithm might exist. Given the obvious symmetry of the problem, we only consider the case of positive skewness. We performed this grid-search by using values of kurtosis ranging from 0 to 15 and with step-length of 0.5. For skewness, we took a grid ranging from 0 to 6 with a step-length of 0.25. For each skewness and kurtosis pair on the bi-dimensional grid, we ran the phase I part of the simplex algorithm.[11] We found that the authorized domain will be convex; i.e. there are no disconnected regions from the one determined with high accuracy below.

Once we got an idea of the general shape of the authorized domain, we performed a search of the exact boundary for a given kurtosis using a bisection algorithm determining the boundary up to a precision of 10^{-6}. Fig. 1 displays the graph of the boundary in the kurtosis–skewness space. The actual domain over which EDs exist is symmetric with respect to the horizontal axis. For convenience we only present the upper half of the existence domain. Points located under the curve are compatible with some ED. We call this domain $\mathscr{E}$. Comparison of the possible domain with the one obtained for instance in polynomial approximations involving Hermite expansions (e.g. Barton and Dennis, 1952 or Jondeau and Rockinger, 2001) indicates that EDs are defined over a much larger set of possible values of skewness and kurtosis.[12]

In later numerical computations, it will be necessary to restrict skewness and kurtosis to the domain $\mathscr{E}$ to guarantee the existence of a density. For this reason we consider a functional description of the authorized domain.[13] An

[11] All computations in this research were done under GAUSS on a WINDOWS 98 platform.

[12] In those contributions it is shown that skewness and kurtosis must be in the interior of a domain similar to an elipse. Kurtosis may vary from 0 to 4 and the maximal allowed skewness is $\sqrt{6}/\sqrt{3+\sqrt{6}} \cong 1.04$ for a kurtosis of $\sqrt{6} \cong 2.45$.

[13] In Jondeau and Rockinger (2001), the boundary constraint was imposed-using a linear interpolation. This method of imposing the boundary conditions may be the only one available for domains that are difficult to characterize. For the problem at hand, a simpler characterization is possible.

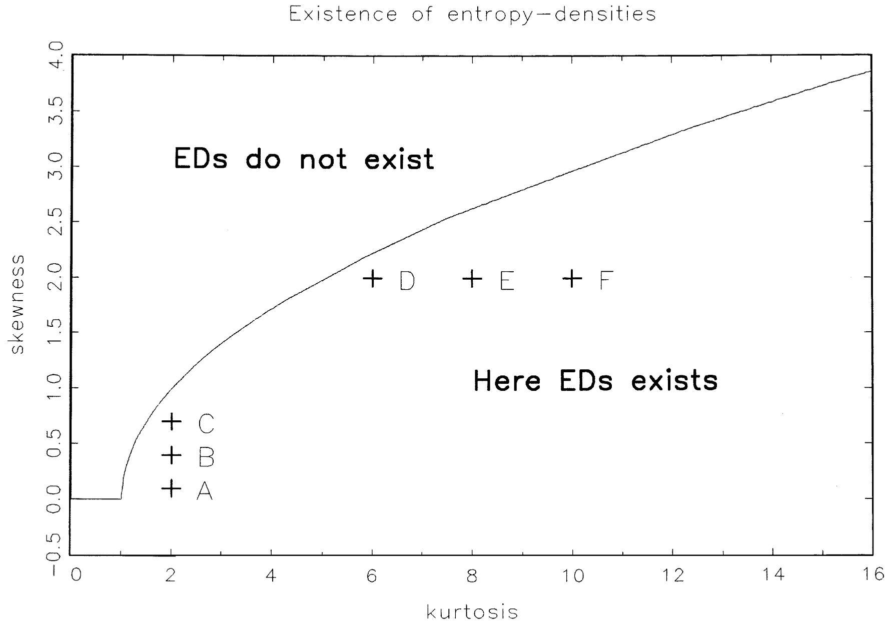

Fig. 1. Here we represent the frontier delimiting the skewness and kurtosis domain where entropy densities exist. This graph represents only the upper half of the authorized domain. The various letters A–F correspond to points for which we represent the entropy density in later figures.

OLS fit of $k = as^2 + bs + c$ indicates that for $k > 1$ the skewness range is $[-s^*(k), s^*(k)]$ where[14]

$$s^*(k) = [-b + \sqrt{b^2 - 4a(c-k)}]/(2a), \quad k > 1. \tag{13}$$

Next, we consider how the ED behaves as skewness, s, and kurtosis, k, vary. In Fig. 1, we trace various pairs of skewness and kurtosis, represented by stars, the density of which is represented in Figs. 2 and 3. An inspection of these figures reveals a rich pattern of possible densities. For densities with small kurtosis, the probability mass is squeezed towards the center. Introduction of skewness then leads to multi-modal densities. For densities with large kurtosis and skewness, given the assumed finiteness of the boundary, a small hump in the tail of the distribution will accommodate the skewness.[15] We

[14] The fit between s and k turned out to be rather good. We found the values $a = 0.9325$, $b = 0.0802$, $c = 0.9946$.

[15] For all possible skewness and kurtosis pairs chosen, our algorithm finds a density typically in a small fraction of a second. This contrasts with other methods that involve at least a few seconds for each ED evaluation. We verified that one obtains the normal density for $s = 0$ and $k = 3$.

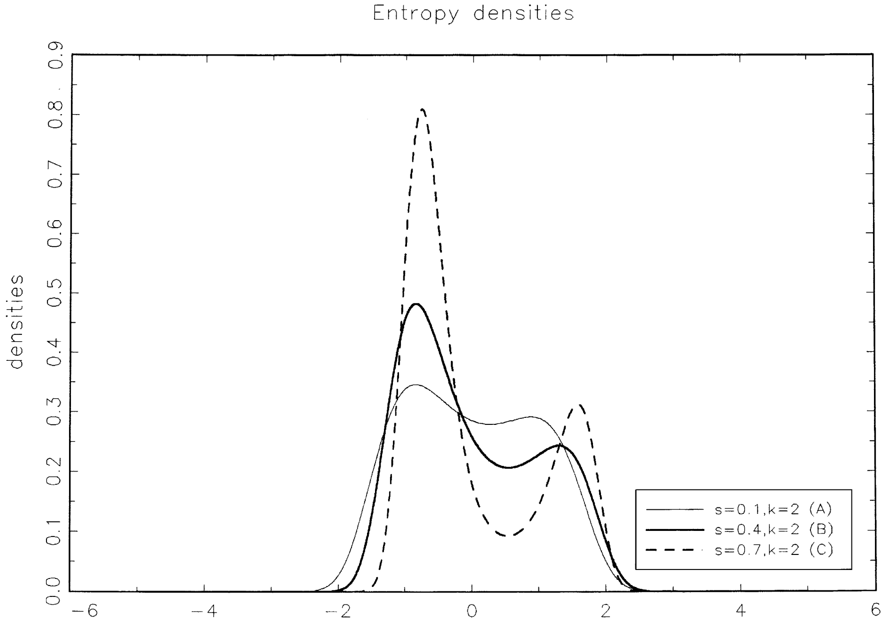

Fig. 2. Entropy densities for points A, B and C.

obtain that EDs may be of use in situations where the tails of the distributions are much thinner than the tails of the normal density. Inversely, k may become very large allowing for rather thick tails.

3. A model with autoregressive heteroskedasticity, skewness, and kurtosis

3.1. The model

In this part of the paper, we wish to illustrate the usefulness of EDs by showing how Bollerslev's (1986) GARCH model can be extended to allow for time variation in skewness and kurtosis. Hansen (1994) considers a similar model where innovations are modeled as generalized Student-t. The generalized Student-t does not allow for humps, hence, intuitively, the skewness–kurtosis range is smaller than for EDs. Moreover, a direct description of the parameters as skewness and kurtosis is not possible. Hansen's contribution also allows for an asymmetry of the density.

The general model we consider is given by

$$r_t = \mu + y_t, \tag{14}$$

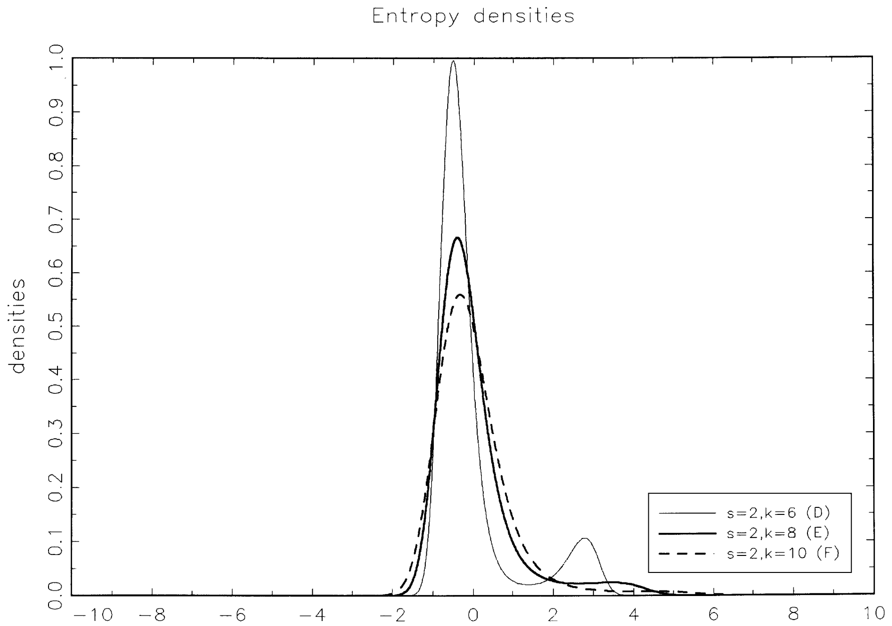

Fig. 3. Entropy densities for points D, E and F.

$$y_t = \sigma_t \varepsilon_t, \tag{15}$$

$$\varepsilon_t \sim \mathrm{ED}(0, 1, s_t, k_t), \tag{16}$$

$$\sigma_t^2 = a_0 + b_0 y_{t-1}^2 + c_0 \sigma_{t-1}^2, \tag{17}$$

$$s_t = a_1 + b_1 y_{t-1}, \tag{18}$$

$$k_t = a_2 + b_2 |y_{t-1}|, \tag{19}$$

$$(s_t, k_t) \in \mathscr{E}. \tag{20}$$

In Eq. (14), r_t represents $100\ln(S_t/S_{t-1})$, where S_t is the closing price of some asset at time t. Here, we assume a constant mean return, μ. The innovations, y_t, are written as a product between the conditional volatility σ_t and an innovation ε_t. In Eq. (16), we assume that ε_t follows an ED with zero mean, unit variance, skewness s_t and kurtosis k_t. Eq. (17) specifies volatility as a simple GARCH(1, 1). Eqs. (18) and (19) assume that skewness and kurtosis depend conditionally on past realizations of the residual y_t.

As usual, we impose that a_0, b_0, and c_0 be positive as well as that $b_0 + c_0 < 1$.[16] In Eq. (18), a_1 and b_1 are estimated freely. To guarantee positivity of kurtosis, one may assume a_2 and $b_2 > 0$. This constraint may, however, be overly restrictive. If the parameter a_2 is found to be large, then b_2 could be negative as long as y_{t-1} remains small. For a given sample, this may be the case. Intuitively, a negative b_2 corresponds to a situation where, after the realization of a relatively large return in absolute value, kurtosis becomes smaller than average.

In this paper, we also estimate specifications of skewness and kurtosis where

$$s_t = a_1 + b_1 \frac{y_{t-1}}{\sigma_{t-1}}, \tag{21}$$

$$k_t = a_2 + b_2 \left| \frac{y_{t-1}}{\sigma_{t-1}} \right|, \tag{22}$$

thus, involving standardized residuals.

Our specification encompasses Bollerslev's GARCH(1, 1) model with Gaussian errors. This is obtained by setting $b_1 = b_2 = 0$, $a_1 = 0$, and $a_2 = 3$ for all t. We do not encompass the case of errors following the Student-t or the generalized error distribution (GED).

We also impose that $(s_t, k_t) \in \mathscr{E}$ in the following way: If k_t is out of the authorized domain, we impose a large penalty for the log-likelihood at time t. If k_t is in $\mathscr{E}$, then, we compute using (13) the upper skewness boundary s^*. If $|s_t| > s^*$, we impose again a large penalty for the log-likelihood.

To ease the estimation, we standardize returns by computing the mean μ separately. In a preliminary step, we also divide r_t by its standard deviation.[17]

3.2. Estimation

The estimation of the parameters follows various steps. In a first step, we estimate the unconditional mean μ using as estimate $\hat{\mu} = 1/T \sum_{t=1}^{T} r_t$. This yields the innovations, defined as $y_t = r_t - \hat{\mu}$. It is with these innovations that we estimate the GARCH model with time-varying skewness and kurtosis.

The estimation of the remaining parameters is performed by maximizing the empirical likelihood. To perform a maximization, it is necessary to have an optimization routine and an objective function. We now discuss the algorithm

[16] Least stringent constraints could also be used.

[17] In other words, we estimate the model for a series of returns with mean 0 and unit standard deviation. It may not suffice to standardize returns with standard deviation because of extreme values. For series where extremely large returns are present, it may be necessary to choose a particularly large domain D and a standardization by a number larger than the standard deviation.

yielding the objective function, then we discuss the maximization algorithm used. The difficulty that we have is that we should ideally be able to impose restrictions on the parameters so that all the s_t and k_t are always in $\mathscr{E}$. That would mean imposing several thousand inequality constraints. Not having access to such code, we truncate skewness and kurtosis to the domain while imposing penalties.

The objective function will involve a parameter vector, say $\theta = (a_0, b_0, c_0, a_1, b_1, a_2, b_2)$, and the innovations y_t, $t = 1, \dots, T$. At each call to the procedure that computes the objective function, the parameter vector and the innovations have to be supplied. Within the procedure we use the following algorithm:

Algorithm 2.

1. This is an initialization step. We define $\rho = 10\,000$, some large number that will be used as a penalty. We define a lower and an upper boundary for kurtosis, that is $k_l = 1$ and $k_u = 16$, respectively. We also set the limits of the domain $\mathscr{D}$ sufficiently large to guarantee that it contains y_t/σ_t for all t. We initialize the dynamics of volatility using $\sigma_0^2 = 1/T \sum_{t=1}^{T} y_t^2$. To define the domain over which the EDs exist we set $a = 0.9325$, $b = 0.0802$, $c = 0.9946$, and $e = 0.001$.

2. We set $t = 2$ and initialize a vector l with $T - 1$ zeros that will contain the log-likelihoods.

3. Here we compute $\sigma_t^2 = a_0 + b_0 y_{t-1}^2 + c_0 \sigma_{t-1}^2$, $s_t = a_1 + b_1 y_{t-1}$, $k_t = a_2 + b_2 |y_{t-1}|$, $\varepsilon_t = y_t/\sigma_t$.

4. If $k_t > k_u$ we compute $\pi_k = (k_t - k_u)\rho$ and truncate the kurtosis by setting $k_t = k_u$. Similarly, if $k_t < k_l$ we compute $\pi_k = (k_l - k_t)\rho$ and set $k_t = k_l$.

5. Now, we compute the limit of skewness beyond which EDs no longer exist. That is $s^*(k_t) = [-b + \sqrt{b^2 - 4a(c - k_t)}]/(2a) - e$. We introduce the e to be certain that we are in the interior of the authorized domain.

6. Now we truncate skewness in case of exceedance. If $s_t > s^*$ then we set $\pi_s = (s_t - s^*)\rho$ and $s_t = s^*$. If $s_t < -s^*$ then we set $\pi_s = (-s^* - s_t)\rho$ and $s_t = -s^*$.

7. Given s_t and k_t we construct the ED using algorithm 1.

8. Now we evaluate the ED, say d_t, at the point ε_t and compute the log-likelihood for period t, $l_t = \ln(d_t) - \ln(\sigma_t)$. The term $\ln(\sigma_t)$ comes from the Jacobian of the transform from ε_t into y_t/σ_t. When a boundary on skewness or kurtosis is binding then π_s or π_k get added to the likelihood, l_t.

9. Set $t = t + 1$. Continue with step 3 until $t > T$.

After each run through the sample, the procedure exports the vector of log-likelihoods l_t that will be used by the optimization routine. Many optimization routines exist. We used the Broyden, Fletcher, Goldfarb, and Shannon (BFGS) algorithm (see Nocedal and Wright, 1999).

3.3. Properties of the estimates

We will estimate Eq. (14)–(20), or alternatively the same model but with (18) and (19) replaced by (21) and (22). Given that we assume for the errors a certain semi-nonparametric representation it follows that our estimation becomes one of empirical maximum likelihood. This raises the issue of the type of standard errors to use.[18]

We note as $L_Q^{\mathrm{E}}(\theta; y)$ the quasi likelihood obtained by using the entropy density for the residuals. We let $y=(y_1,\ldots,y_T)'$ the vector of innovations and θ the vector of all parameters. The quasi-maximum likelihood (QML) estimate $\hat{\theta}$ is obtained as solution to

$$\hat{\theta} \in \arg\max_{\theta\in\Omega} [\ln L_Q^{\mathrm{E}}(\theta; y)].$$

The limit distribution is given by

$$\sqrt{T}(\hat{\theta}-\theta_0) \Rightarrow \mathcal{N}(0, h(\theta_0)^{-1}\Sigma h(\theta_0)^{-1}), \tag{23}$$

where θ_0 is the true value of θ, and where $\Rightarrow$ indicates convergence in distribution. The matrices $h(\theta_0)$, respectively, Σ may be estimated using

$$\hat{h}(\theta_0) = T^{-1} \left. \frac{\partial^2 L_Q^{\mathrm{E}}(\hat{\theta}; y)}{\partial\theta\partial\theta'} \right|_{\hat{\theta}},$$

$$\hat{\Sigma} = T^{-1} \sum_{t=1}^{T} \left[\left. \frac{\partial L_Q^{\mathrm{E}}(\hat{\theta}; y)}{\partial\theta} \right|_{\hat{\theta}} \left. \frac{\partial L_Q^{\mathrm{E}}(\hat{\theta}; y)}{\partial\theta'} \right|_{\hat{\theta}} \right].$$

If for given parameters the entropy density correctly specifies the true density of the ε_t then, White (1994) has shown that $\Sigma = -h(\theta_0)$ and the maximum likelihood has the familiar asymptotic normality

$$\sqrt{T}(\hat{\theta}-\theta_0) \Rightarrow \mathcal{N}\left(0, \operatorname*{plim}_{T\to\infty}\left(-\frac{1}{T}\left[\left.\frac{\partial^2 \ln L_Q^{\mathrm{E}}(\theta; y)}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1}\right)\right).$$

Even though we believe that the modeling of innovations with a more general density allowing for time varying moments is a step towards the correct description of innovations, given the complexity of financial data, it is wise to assume that there is still mis-specification in our model. For this reason, we recommended the use of the robust formulas in (23). In the empirical work we will only have these types of standard errors.

[18] The following discussion is inspired by Mittelhammer et al. (2000, p. 248–249).

Table 1
Descriptive statistics. This table represents moments computed with GMM and a correction for heteroskedasticity.[1]

	SP 500	FT 100	NIKKEI
Mean	0.1774	0.1844	0.1267
(s.e.)	0.056[a]	0.069[a]	0.060[a]
Var	2.1479	2.678	2.3169
(s.e)	0.063[a]	0.113[a]	0.072[a]
Skew	−0.3243	−0.3874	−0.3557
(s.e.)	0.238	0.591	0.217
Kurt	3.2179	8.7147	3.7658
(s.e.)	0.965[a]	3.987[a]	0.647[a]
Normality	12.08	5.64	36.50
p-value	0.00	0.06	0.00
Engle	79.05	48.81	52.95
p-value	0.00	0.00	0.00

[1] The numbers in parenthesis, i.e. s.e., are the standard errors of the statistics. Normality corresponds to the Jarque–Bera test of normality. This statistics is obtained as the sum of the squared standardized skewness and the squared standardized excess-kurtosis. Engle is the Lagrange-multiplier statistic TR^2 of joint significance of the regressors in an OLS regression of squared centered returns on their lags. There are 1500 weekly observations in the sample.
Note: In this and the following tables the superscripts a and b indicate statistical significance at the 5% respectively the 10% level.

4. Empirical results

4.1. The data used

Out of Datastream, we extracted daily closing prices for the S&P 500 Composite Index, the FT 100 Share Index, and the Nikkei 225 Stock Index. Using closing prices, sampled for each Tuesday (or the day closest to it), we constructed weekly returns. The sample covers the period from August 27, 1971 through May 31, 2000. Our database, therefore, consists of 1500 observations. Table 1 provides sample statistics where all moments are computed in the GMM setting of Richardson and Smith (1993), thus, controlling for heteroskedasticity. All the series under consideration are negatively skewed and fat-tailed. Furthermore, the Engle statistics reveals the presence of conditional heteroskedasticity in the data.

4.2. Preliminary estimation

In order to get an idea of the unconditional behavior of the model, we start with the estimation of traditional models assuming for the residuals normality, a Student-t and a generalized error distribution (GED). This means that we consider Eqs. (14)–(17), where we replace the entropy density either with a normal density, $(2\pi)^{-1/2}\exp(-0.5x^2)$, a Student-$t$ with v degrees of freedom, $\Gamma((v+1)/2)/\Gamma(v/2)\ (v\pi)^{-1/2}(1+x^2/v)^{(v+1)/2}$, or the GED with parameter η. The density of the GED is given by $(\eta\exp[-0.5|x/\lambda|^{\eta}])/(\lambda 2^{((\eta+1)/\eta)}\Gamma(1/\eta))$ where $\lambda=\{2^{-2/\eta}\Gamma(1/\eta)/\Gamma(3/\eta)\}^{1/2}$.

The estimates are reported in Table 2. We obtain for the parameters typical values. The parameter b_0 oscillates around 0.1 and c_0 around 0.88 suggesting that there is a fair amount of persistence in volatility. When we inspect the parameters for given data of various models we notice that the estimates remain very similar. Next, we may inspect the standard errors. We find that the robust standard errors are similar across the various models.

It is further possible to compare the gaussian model with the Student-t and GED since the Student-t encompasses the normal case for $v=\infty$ and the GED does the same for $\eta=2$. To perform the test, we may either directly use the Wald t-test associated with the parameters v and η or the likelihood ratio test of the Gaussian restriction. For both types of tests, we notice that we always soundly reject the gaussian restriction. When we consider the SP 500 we find that the parameter v is relatively large taking the value 11.57 and the η the value 1.66. This suggests that this series has innovations that are close to the gaussian case.

Now, we turn to a model where errors follow an unconditional ED obtained by setting $b_1=b_2=0$, i.e. skewness and kurtosis are constant. For this estimation, we use as starting values for skewness and kurtosis the estimates reported in Table 1 and as starting values of the volatility equation, (17), the ones of the GARCH(1,1) model. Convergence was achieved after a few seconds. The results are presented in Table 3. We notice values for the estimates of the volatility equation (17) that are close to the values reported in Table 2.

Turning to skewness and kurtosis, for the SP 500 we obtained in Table 1 the values -0.32 and 3.22. Now we obtain -0.28 and 3.87, thus, the parameters are relatively close. However, for the FT 100, this is not the case. Skewness took the value -0.38 in Table 1 but now takes the value -0.78. Since it is known that conditional volatility creates fat-tails, this suggests that the filtering by the conditional volatility amplifies the tail behavior, questioning the specification of GARCH models for certain series.

Even though the ED does not encompass the Student-t nor the GED, one may ask which model we should select. Several selection criteria may be used such as the Akaike or Schwarz criterion. Since the number of parameters is equal in all these models, the selection among the Student-t, the GED or the ED boils down to the choice of the model with the largest likelihood. Both for the SP 500 and the FT 100 we find that the entropy performs best. For the Nikkei we find that the Student-t has a higher likelihood than the ED which in turn performs better than the GED. This observation suggests that there are extreme realizations in the Nikkei that the ED has difficulties to capture.

Table 2
GARCH estimates with traditional densities. This table displays the results of the estimation of traditional models.[1]

	Gaussian model			Student-t			GED		
	SP 500	FT 100	NIKKEI	SP 500	FT 100	NIKKEI	SP 500	FT 100	NIKKEI
a_0	0.0292	0.0264	0.0143	0.0223	0.0224	0.0124	0.0283	0.0282	0.0159
	0.0147[a]	0.0100[a]	0.0091	0.0115[b]	0.0072[a]	0.0060[a]	0.0143[a]	0.0094[a]	0.0082[b]
b_0	0.1078	0.0937	0.1148	0.0826	0.0651	0.0921	0.1045	0.0883	0.1272
	0.0242[a]	0.0191[a]	0.0345[a]	0.0196[a]	0.0136[a]	0.0216[a]	0.0239[a]	0.0177[a]	0.0306[a]
c_0	0.8667	0.8836	0.8807	0.8757	0.8814	0.8553	0.8704	0.8835	0.8670
	0.0323[a]	0.0203[a]	0.0387[a]	0.0324[a]	0.0210[a]	0.0335[a]	0.0324[a]	0.0208[a]	0.0330[a]
v	—	—	—	11.5751	8.2593	6.0466	—	—	—
				2.8544[a]	2.0155[a]	0.9527[a]			
η	—	—	—	—	—	—	1.6635	1.4506	1.3292
							0.0925[a]	0.1420[a]	0.0826[a]
Lik	−2014.76	−1990.22	−1958.18	−2002.75	−1939.58	−1904.62	−2007.62	−1957.01	−1914.56
LRT	—	—	—	24.03	101.27	107.13	14.29	66.41	87.25

[1] We describe the innovations $y_t = r_t - 1/T \sum_{t=1}^{T} r_t$ by assuming that $y_t = \sigma_t \varepsilon_t$ where $\sigma_t = a_0 + b_0 y_{t-1}^2 + c_0 \sigma_{t-1}^2$ and ε_t is either modeled with the standard normal, the Student-t, or the generalized error distribution (GED).

Table 3
GARCH estimates with entropy density and constant skewness and kurtosis. This table represents the parameters of the GARCH regressions where innovations are assumed to be distributed as an entropy density.[1]

	SP 500	FT 100	NIKKEI
a_0	0.0233	0.0329	0.0221
	0.0117[a]	0.0109[a]	0.0099[a]
b_0	0.0967	0.0965	0.1384
	0.0216[a]	0.0219[a]	0.0324[a]
c_0	0.8823	0.8710	0.8435
	0.0283[a]	0.0252[a]	0.0365[a]
s	−0.2839	−0.7907	−0.4453
	0.1041[a]	0.5405	0.1516[a]
k	3.8767	9.1575	5.2828
	0.2469[a]	5.2757	0.7179[a]
Lik	−2001.31	−1932.80	−1910.01
LRT	26.91	114.84	96.34

[1] Here skewness s and kurtosis k are supposed to be constants.

In Fig. 4 we display the shape of the various distributions of ε_t for the FT 100. The choice between a normal and Student-t is not an obvious one since their shape is very close. The GED, on the other hand, differs from the normal and the Student-t. Its peak is much more pronounced. One of the disadvantages of these distributions is that they are symmetric and, therefore, they do not allow for an asymmetry. An inspection of the ED reveals that there is a strong asymmetry in the data. This skewness is due to large negative realizations.

4.3. Estimation of the general model

A first remark is that in optimization problems of this kind, the choice of initial values is quite important. Even though the code has been written in such a manner that the parameters end up in the authorized domain, the estimation is sensibly faster if one starts with an interior parameter, i.e. all the constraints are satisfied. We will use in our estimations the parameter estimates obtained from the unconditional entropy density estimation.

The left part of Table 4 corresponds to specification I, that is skewness and kurtosis are described by Eqs. (18) and (19). The right part of the table corresponds to the specification involving Eqs. (21) and (22).

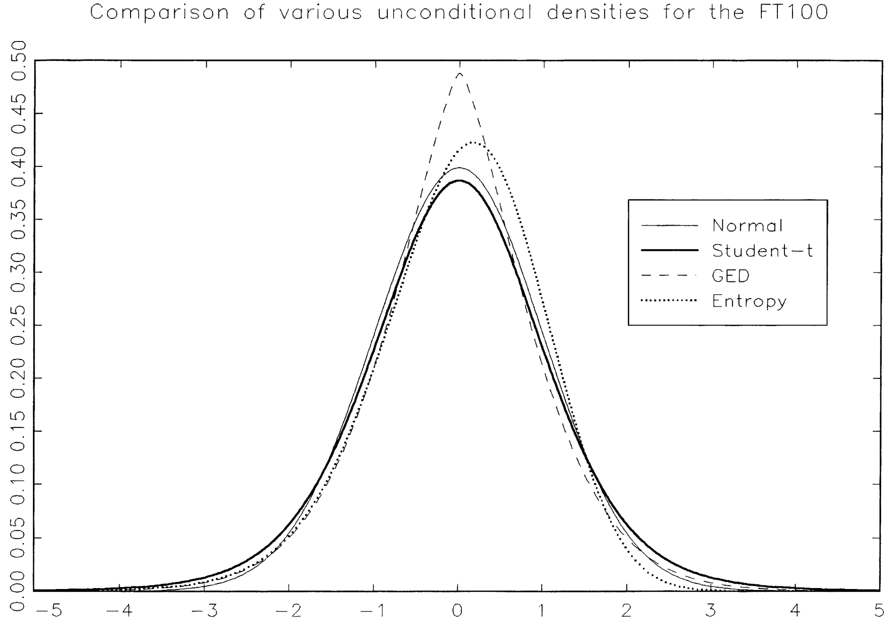

Fig. 4. Various unconditional densities obtained in a GARCH estimation. This graph is for the FT100.

A first comparison between the volatility parameters of Table 4 with those of Table 3 reveals that these parameters are essentially unaffected by introducing time-varying skewness and kurtosis. For instance, the parameter b_0 of the FT 100 took the value 0.0219 when skewness and kurtosis were held constant, whereas now it becomes 0.0167.

An inspection of the constant in the skewness equation with s of Table 3 indicates that for certain specifications this parameter is rather unstable. For the FT 100, the constant markedly decreases in absolute value from −0.79 to −0.31. Interestingly, the parameter a_1 is now very close to the unconditional skewness estimated in Table 1.

Using the likelihoods of the various models it is possible to test the restriction $b_1 = b_2$. We notice that, for the Nikkei, specification I appears as an improvement over the model with constant skewness and kurtosis, and similarly specification II for the SP 500. At first glance, for the other estimations, the model with time varying skewness and kurtosis does not bring much improvement from a statistical point of view. A more careful inspection of the t-statistics of the parameters shows that the lagged parameter b_1 in the skewness dynamic of the FT 100 is statistically different from 0. Also from an economic point of view, an inspection of the magnitude of the point

Table 4
GARCH estimates with entropy density and time-varying skewness and kurtosis.[1]

	Specification I			Specification II		
	SP 500	FT 100	NIKKEI	SP 500	FT 100	NIKKEI
a_0	0.0203	0.0308	0.0253	0.0203	0.0314	0.0230
	0.0120^{b}	0.0093^{a}	0.0102^{a}	0.0120^{b}	0.0096^{a}	0.0099^{a}
b_0	0.0987	0.0823	0.1597	0.0959	0.0834	0.1396
	0.0234^{a}	0.0167^{a}	0.0577^{a}	0.0217^{a}	0.0175^{a}	0.0322^{a}
c_0	0.8837	0.8777	0.8045	0.8862	0.8838	0.8467
	0.0296^{a}	0.0209^{a}	0.0824^{a}	0.0287^{a}	0.0226^{a}	0.0321^{a}
a_1	−0.2832	−0.3175	−0.3298	−0.3131	−0.3386	−0.6859
	0.0895^{a}	0.2148	0.1304^{a}	0.0923^{a}	0.2669	0.3037^{a}
b_1	0.1202	0.2337	−0.0892	0.1119	0.2495	−0.1967
	0.0907	0.0817^{a}	0.1084	0.0812	0.0867^{a}	0.1907
a_2	4.0575	4.3433	4.8576	4.2132	4.1606	5.2450
	0.3168^{a}	1.7948	0.5115^{a}	0.3476^{a}	4.3792	1.4504
b_2	−0.0310	0.6884	0.0125	−0.1664	1.4494	1.4055
	0.2124	0.5354	0.1534	0.2188	3.6701	1.8705
Lik	−1995.16	−1931.74	−1901.27	−1994.68	−1932.05	−1909.67
LRT	12.30	2.11	17.50	13.26	1.50	0.69

[1] In this table, we estimate a GARCH model on the full sample allowing for time-varying skewness and kurtosis. In specification I, skewness and kurtosis are modeled as $s_t = a_1 + b_1 y_{t-1}$ and $k_t = a_2 + b_2|y_{t-1}|$. In specification II, the model is $s_t = a_1 + b_1 y_{t-1}/\sigma_{t-1}$ and $k_t = a_2 + b_2|y_{t-1}/\sigma_{t-1}|$. The label LRT corresponds to a likelihood ratio test statistics of the restricted model where $b_1 = b_2$.

estimates of b_1 and b_2 shows that these coefficients are relatively large. One possible interpretation of these results is that skewness and kurtosis measure extreme realizations that occur only seldomly. Due to this rare occurrence, statistical tests will have little power.

To sum up, our model of conditional skewness and kurtosis reveals some conditional behavior at a weekly frequency, however, this dynamics is rather difficult to interpret.

5. Conclusion

In this paper, we have first shown how entropy densities can be estimated in an efficient manner. We characterize the skewness and kurtosis domain over which entropy densities will be well defined while retaining the mean equal to zero and the variance equal to one. In a numerical application, involving series of weekly stock returns, we show that the entropy density is of value in traditional GARCH models, i.e. where skewness and kurtosis is not time variant. Turning to the model allowing for time varying parameters we show that the estimation of a model involving a time-varying skewness and kurtosis is possible.

A further contribution is that we show how skewness and kurtosis may be rendered time varying using entropy densities. We find that from a statistical point of view there is little evidence that skewness and kurtosis are dependent on past returns. One possible reason for this finding is that these moments are driven by extreme realizations that occur only infrequently. Due to this rare occurrence statistical test may lack power.

Acknowledgements

Michael Rockinger is also CEPR, and scientific consultant at the Banque de France. He acknowledges help from the HEC Foundation, the Fondation Nationale pour l'Enseignement de la Gestion des Entreprises (FNEGE), and the European Community TMR Grant: "Financial Market Efficiency and Economic Efficiency". This paper was finished while the first author was visiting UC San Diego. We thank Noam Agmon for his help. We are grateful to Robert Engle, Hal White, and Arnold Zellner (the editor) for precious comments. The usual disclaimer applies. The Banque de France does not necessarily endorse the views expressed in this paper.

References

Abramowitz, M., Stegun, I.A., 1970. Handbook of Mathematical Functions. Dover Publications, Inc., New York.

Alhassid, Y., Agmon, N., Levine, R.D., 1978. An upper bound for the entropy and its applications to the maximal entropy problem. Chemical Physics Letters 53 (1), 22–26.

Agmon, N., Alhassid, Y., Levine, R.D., 1979a. An algorithm for finding the distribution of maximal entropy. Journal of Computational Physics 30, 250–259.

Agmon, N., Alhassid, Y., Levine, R.D., 1979b. The maximum entropy formalism. In: Levine, R.D., Tribus, M. (Eds.), The Maximum Entropy Formalism. MIT Press, Cambridge, MA, pp. 207–209.

Barton, D.E., Dennis, K.E.R., 1952. The conditions under which Gram–Charlier and Edgeworth curves are positive definite and unimodal. Biometrika 39, 425–427.

Bera, A.K., Higgins, M.L., 1992. A class of nonlinear ARCH models. International Economic Review 33 (1), 137–158.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31 (3), 307–328.

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: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Vol. 4. Elsevier Science, Amsterdam, pp. 2959–3038.

Buchen, P., Kelly, M., 1996. The maximum entropy distribution of an asset inferred from option prices. Journal of Financial and Quantitative Analysis 31 (1), 143–159.

Davis, P., Polonsky, I., 1970. Numerical interpolation, differentiation and integration. In: Abramowitz, M., Stegun, I.A. (Eds.), Handbook of Mathematical Functions. U.S. Govt. Printing Office, Washington, pp. 875–924.

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

Engle, R.F., Gonzales-Rivera, G., 1991. Semiparametric ARCH models. Journal of Business and Economic Statistics 9 (4), 345–359.

Fletcher, R., 1994. An overview of unconstrained optimization. In: Spedicato, E. (Ed.), Algorithms for Continuous Optimization, Vol. XX. Kluwer Academic, Dordrecht, Netherland, pp. 109–143.

Gallant, A.R., Tauchen, G., 1989. Semi non-parametric estimation of conditionally constrained heterogenous processes: asset pricing applications. Econometrica 57 (5), 1091–1120.

Golan, A., Judge, G., Miller, D., 1996. Maximum Entropy Econometrics: Robust Estimation with Limited Data. Wiley, Chichester.

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

Harvey, C.R., Siddique, A., 1999. Autoregressive conditional skewness. Journal of Financial and Quantitative Analysis 34 (3), 465–487.

Hawkins, R.J., Rubinstein, M., Daniell, G.J., 1996. Reconstruction of the probability density function implicit in option prices from incomplete and noisy data. In: Hanson, K.M., Silver, R.N. (Eds.), Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers, Netherlands, pp. 1–8.

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

Jaynes, E.T., 1982. On the rationale of maximum-entropy methods. Proceedings of the IEEE 70 (9), 939–952.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous univariate distribution. Vol. 1, 2nd Edition. Wiley, New York.

Jondeau, E., Rockinger, M., 2001. Gram–Charlier densities. Journal of Economic Dynamics and Control 25 (10), 1457–1483.

Kraus, A., Litzenberger, R.H., 1976. Skewness preference and the valuation of risk assets. Journal of Finance 31 (4), 1085–1100.

Mead, L.R., Papanicolaou, N., 1984. Maximum entropy in the problem of moments. Journal of Mathematical Physics 25 (8), 2404–2417.

Mittelhammer, R.C., Judge, G.G., Miller, D.J., 2000. Econometric Foundations. Cambridge University Press, Cambridge, UK.

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

Nocedal, J., Wright, S.J., 1999. Numerical Optimization. Springer, New York, USA.

Ormoneit, D., White, H., 1999. An efficient algorithm to compute maximum entropy densities. Econometric Reviews 18 (2), 127–140.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1999. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, UK.

Richardson, M., Smith, T., 1993. A test for multivariate normality in stock returns. Journal of Business 66 (2), 295–321.

Shannon, C.E., 1948. The mathematical theory of communication. Bell Systems Technical Journal 27, 379–423; 623–656.

Stutzer, M., 1996. A simple nonparametric approach to derivative security valuation. Journal of Finance 51 (5), 1633–1652.

Wheeler, J.C., Gordon, R.G., 1969. Rigorous bounds for thermodynamic properties of harmonic solids. The Journal of Chemical Physics 51 (12), 5566–5583.

White, H., 1994. Estimation, Inference, and Specification Analysis. Cambridge University Press, Cambridge, UK.

Zellner, A., Highfield, R.A., 1988. Calculation of maximum entropy distributions and approximation of marginal posterior distributions. Journal of Econometrics 37 (2), 195–209.

Zellner, A., Tobias, J., Ryu, H., 1997. Bayesian method of moments analysis of time series models with an application to forecasting turning points in output growth rates. (Estadística, 49–51, 152–157), pp. 3–63.