5

Modeling Higher Moments

In this chapter, we investigate two important issues for the modeling of asset returns. The first issue is the modeling of the entire density of returns, so that it incorporates some of the features described in Chapter 2, in particular the asymmetry and the fat-tailedness of the distribution. In several problems in finance, it is crucial to recognize that the conditional distribution of returns is non-normal and to correctly model it. A precise knowledge of the conditional distribution is required, for instance, for asset allocation, VaR (Value-at-Risk) computation, or the valuation of contingent claims.

The second issue is the modeling of the time-variability of higher moments, which can be viewed as summaries of the density. This is a key issue for problems that involve data sampled at a rather high frequency (say daily or weekly data) or very non-normal asset returns. For instance, the modeling of conditional higher moments will be crucial in the computation of conditional VaR, which is a short-term measure typically calculated for a 1- or 10-day horizon, in which case the modeling of higher moments may substantially improve the accuracy of VaR. Let us consider now the case of asset allocation. Exactly as the model of Markovitz builds on the first two moments, mean and variance, more advanced models build on moments beyond the first two. At first glance, the existence of time-varying higher moments may be expected to strongly affect the allocation of wealth. Yet, asset allocation is often considered at a rather low frequency (with a monthly or even a quarterly horizon). At such intervals, higher moments are less likely to be time-varying. For asset allocation purpose, we may conclude that the modeling of the higher moments is crucial for asset classes with very non-normal returns, such as hedge funds or emerging market indices. A more complete description of these issues is provided in Chapters 8 and 9.

We therefore consider now the explicit modeling of the higher moments of asset returns and their conditional distributions. Since the seminal work of Engle (1982), time-varying volatility has been shown to produce fat tails in the unconditional distribution. But time-varying volatility alone is not enough to explain all the tail fatness; volatility filtered residuals still have tails fatter than

the normal distribution. Instead of considering all the possible distributions that may fit the returns data empirically, our strategy here is to focus on the two stylized facts above and set out to find distributions that can capture these two characteristics. Although there are already many fat-tailed extensions to the normal distribution (such as the Pareto and the Student t), not all of these distributions can capture asymmetry. Hence, finding a distribution with a suitable asymmetry property will be a main objective here. An important criteria in this search for a better alternative distribution is to have an as large as possible range of admissible skewness and kurtosis. Ideally, the only constraints on this domain of definition would be those ensuring that the distribution is definite. However, most distributions discussed in the literature typically impose additional restrictions on this domain of definition.

The remaining of this chapter is organized as follows. Section 5.1 presents the general problems involved in the modeling of higher moments. Section 5.2 describes a number of distributions that can be used to capture higher moments. In Section 5.3, we address the issues of specification tests and inference. We provide an illustration of some of the material presented in this Chapter in Section 5.4. Finally, Section 5.5 describes various ways to model higher moments conditional on past observations.

5.1 The general problem

Let x_t, for $t = 1, \cdots, T$, be a time series of asset returns. It is convenient to break down the complete characterization of x_t into three components: (i) the conditional mean, which contains all the information about the location of the distribution, (ii) the conditional variance, which contains the scale parameter that measures the dispersion of the distribution, and (iii) the shape parameters (e.g., skewness, kurtosis) that determine the form of a conditional distribution within the general family of distributions. Thus, we may write

$$x_t = \mu_t(\theta) + \varepsilon_t, \tag{5.1}$$

$$\mu_t(\theta) = E[x_t | \mathcal{F}_{t-1}] = \mu(\theta, \mathcal{F}_{t-1}), \tag{5.2}$$

$$\varepsilon_t = \sigma_t(\theta) z_t, \tag{5.3}$$

$$\sigma_t^2(\theta) = E\left[(x_t - \mu_t)^2 | \mathcal{F}_{t-1}\right] = \sigma^2(\theta, \mathcal{F}_{t-1}), \tag{5.4}$$

$$z_t \sim g(z_t|\eta). \tag{5.5}$$

Equation (5.1) decomposes the return at time t into a conditional mean, μ_t, and an error term, ε_t. The dynamics of the conditional mean is given by (5.2). The standardized innovation, $z_t = (x_t - \mu_t(\theta))/\sigma_t(\theta)$ has zero mean and unit variance. Equation (5.4) determines the dynamics of volatility. This may be any specification including the GARCH models. Vector θ contains all the parameters associated with the conditional mean and the conditional variance equations. Finally, equation (5.5) specifies that the standardized innovation

follows a conditional distribution g with vector of shape parameters η. In the previous chapter, the conditional distribution was assumed to be normal $\mathcal{N}(0,1)$ with no shape parameter. In the more general case we consider now, shape parameters η will generally involve parameters capturing asymmetry and fat-tailedness of the distribution.

In this section, we address several issues related to the modeling of non-normal returns:

1. If the unconditional distribution has been found to be non-normal, do we necessarily have to assume that the conditional distribution $g(.)$ is non-normal?
2. If the conditional distribution is assumed to be non-normal, do we necessarily have to model it explicitly?
3. If we have to model the conditional distribution explicitly, how far can we go in terms of asymmetry and fat-tailedness of the distribution?

5.1.1 Higher moments of a GARCH process

The idea behind the first issue is that the modeling of the conditional volatility renders the unconditional distribution of the error term ε_t non-normal, even if the conditional distribution of the innovation z_t is still assumed to be normal. This result, highlighted by Engle and González-Rivera (1991), can be of great importance for series that are characterized by a slight departure from normality. In such a case, estimating a GARCH model with a normal conditional distribution may be enough to capture the fat-tailedness of the unconditional distribution. To explain why this is the case, we need to distinguish between the conditional and the unconditional higher moments.

Conditional and unconditional skewness and kurtosis

Let σ^2 denote the unconditional variance of ε_t. The conditional skewness and kurtosis are defined conditionally to the information set $\mathcal{F}_{t-1}$

$$s_c = \frac{E\left[\varepsilon_t^3 \,|\mathcal{F}_{t-1}\right]}{\sigma_t^3},$$
$$\kappa_c = \frac{E\left[\varepsilon_t^4 \,|\mathcal{F}_{t-1}\right]}{\sigma_t^4},$$

whereas their unconditional counterparts are defined as

$$s_u = \frac{E\left[\varepsilon_t^3\right]}{\left(E\left[\sigma_t^2\right]\right)^{3/2}} = \frac{E\left[\varepsilon_t^3\right]}{\sigma^3},$$
$$\kappa_u = \frac{E\left[\varepsilon_t^4\right]}{\left(E\left[\sigma_t^2\right]\right)^2} = \frac{E\left[\varepsilon_t^4\right]}{\sigma^4}.$$

Although not shown explicitly above, the shape parameters of the unconditional distribution (s_u and κ_u) will depend on the characteristics of the conditional distribution. Indeed consider the model (5.1)–(5.5) above, so that $z_t = \varepsilon_t/\sigma_t$ is an *iid* process. Then, for any distribution such that $E\left[\varepsilon_t^3\right] < \infty$, the unconditional skewness is given by

$$s_u = \frac{E\left[\varepsilon_t^3\right]}{\left(E\left[\sigma_t^2\right]\right)^{3/2}} = \frac{E\left[E\left[\varepsilon_t^3 \,|\mathcal{F}_{t-1}\right]\right]}{\left(E\left[\sigma_t^2\right]\right)^{3/2}} = \frac{E\left[s_c\sigma_t^3\right]}{\left(E\left[\sigma_t^2\right]\right)^{3/2}},$$

so that

$$s_u = s_c \frac{E\left[\sigma_t^3\right]}{\left(E\left[\sigma_t^2\right]\right)^{3/2}}.$$

Because of Jensen's inequality, one has $E\left[\sigma_t^3\right] \geq \left(E\left[\sigma_t^2\right]\right)^{3/2}$ as $\sigma_t > 0$, and thus $|s_u| \geq |s_c|$. Therefore, the unconditional skewness has the same sign as the conditional one, but it is possible to capture a high unconditional skewness using a conditional distribution with a smaller skewness. Notice that in cases where the conditional distribution is symmetric, introducing GARCH effects would not result in an asymmetric distribution.

Assuming that $E\left[\varepsilon_t^4\right] < \infty$, the unconditional kurtosis is given by

$$\kappa_u = \frac{E\left[\varepsilon_t^4\right]}{\left(E\left[\sigma_t^2\right]\right)^2} = \frac{E\left[E\left[\varepsilon_t^4 \,|\mathcal{F}_{t-1}\right]\right]}{\left(E\left[\sigma_t^2\right]\right)^2} = \frac{E\left[\kappa_c\sigma_t^4\right]}{\left(E\left[\sigma_t^2\right]\right)^2},$$

so that

$$\kappa_u = \kappa_c \frac{E\left[\sigma_t^4\right]}{\left(E\left[\sigma_t^2\right]\right)^2}.$$

Once again, Jensen's inequality implies $E\left[\sigma_t^4\right] \geq \left(E\left[\sigma_t^2\right]\right)^2$ for $\sigma_t > 0$, and hence $\kappa_u \geq \kappa_c$. Therefore, even in the case where innovations z_t are assumed to be normal, a GARCH model yields an unconditional distribution with fatter tails than the normal distribution.

Conditional normal distribution

To be more explicit, consider an ARCH(1) model, with $\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2$, and normal innovations. Then, as shown in Chapter 4, we have

$$\kappa_u = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2},$$

which exceeds 3 for $\alpha_1 > 0$ and $3\alpha_1^2 < 1$.

For a GARCH(1, 1) model, with $\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2$, Bollerslev (1986) obtains

$$\kappa_u = 3\frac{1-\beta_1^2-2\alpha_1\beta_1-\alpha_1^2}{1-\beta_1^2-2\alpha_1\beta_1-3\alpha_1^2},$$

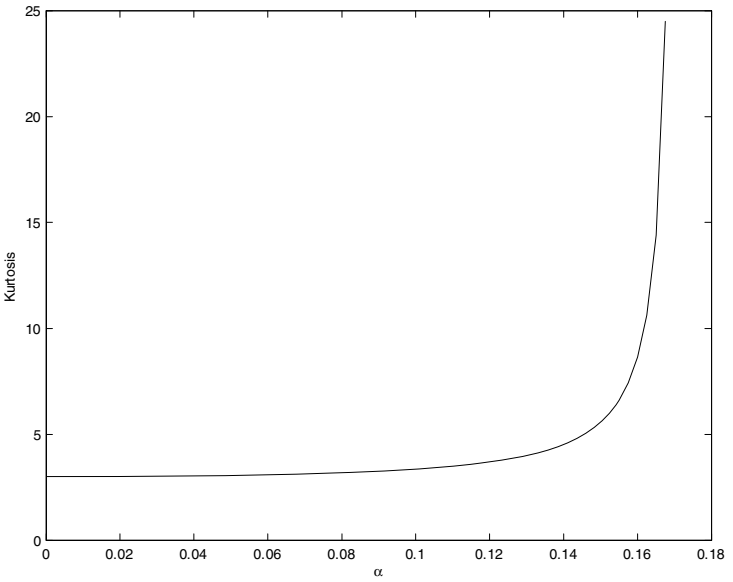

Fig. 5.1. *Unconditional kurtosis of a GARCH*$(1,1)$ *process when* α *varies (with* $\beta = 0.8$*).*

which exceeds 3 for $\alpha_1 > 0$, $\beta_1 > 0$ and $\beta_1^2 + 2\alpha_1\beta_1 + 3\alpha_1^2 < 1$.

Figure 5.1 illustrates how the unconditional kurtosis κ_u of a GARCH$(1,1)$ model varies when we increase the persistence of volatility. More precisely, we assume $\beta_1 = 0.8$ and vary α_1 between 0.05 and 0.16. Then, the unconditional kurtosis increases from 3 to virtually infinity. Controlling the persistence of the conditional volatility thus allows to control the fat-tailedness of the unconditional distribution.

Conditional non-normal distributions

Often, a GARCH model with a normal conditional distribution is not sufficient to capture the fat tails found in the unconditional distribution. This implies that the standardized innovation z_t is not normal and that we need a more general specification for the conditional distribution.

The first attempt to capture the excess kurtosis of GARCH standardized residuals is by Bollerslev (1987), who uses as conditional distribution $g(z_t|\eta)$ a Student t distribution. The density of the symmetric Student t distribution with zero mean and unit variance is given by

$$t(z_t|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{z_t^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \tag{5.6}$$

where $\Gamma(.)$ is the gamma function and ν the degree-of-freedom parameter. The kurtosis of the conditional distribution of z_t is

$$\kappa_c = 3\frac{\nu-2}{\nu-4},$$

provided that $\nu > 4$. The degree-of-freedom adjustment contributes to excess kurtosis, because it is by construction larger than one, except when $\nu \to \infty$, in which case the Student t distribution collapses to a normal distribution. κ_c is always larger than 3 and decreases toward 3 when the degree-of-freedom parameter ν decreases toward 4. The unconditional kurtosis of the error term ε_t is then given by

$$\kappa_u = 3\frac{\nu-2}{\nu-4}\frac{E\left[\sigma_t^4\right]}{\left(E\left[\sigma_t^2\right]\right)^2}.$$

The term involving the degree-of-freedom parameter contributes to capture further excess kurtosis, because it is by construction larger than one.

Nelson (1991) also proposed the GED (*Generalized Error Distribution*) defined as

$$g\left(z_t|\nu\right) = \frac{\nu}{2^{1+1/\nu}\lambda\Gamma\left[1/\nu\right]}\exp\left(-0.5\left|\frac{z_t}{\lambda}\right|^{\nu}\right) \qquad 0 < \nu < \infty, \tag{5.7}$$

with $\lambda = \left(2^{-2/\nu}\Gamma\left[1/\nu\right]/\Gamma\left[3/\nu\right]\right)^{1/2}$. When $\nu = 2$, the distribution in (5.7) reduces to the normal distribution. When $\nu < 2$ ($\nu > 2$), the conditional distribution has fatter (thinner) tails than the normal distribution. The kurtosis of the conditional distribution of z_t is

$$\kappa_c = \frac{\Gamma\left[1/\nu\right]\Gamma\left[5/\nu\right]}{\left(\Gamma\left[3/\nu\right]\right)^2},$$

so that it is larger than 3 when $\nu < 2$. The unconditional kurtosis of ε_t is then given by

$$\kappa_u = \frac{\Gamma\left[1/\nu\right]\Gamma\left[5/\nu\right]}{\left(\Gamma\left[3/\nu\right]\right)^2}\frac{E\left[\sigma_t^4\right]}{\left(E\left[\sigma_t^2\right]\right)^2}.$$

Bai, Russell, and Tiao (2003) provide the exact representation of unconditional kurtosis for GARCH and stochastic volatility models, when innovations are conditionally non-normal. They show that the unconditional kurtosis for both models is determined jointly by the conditional distribution of z_t and the volatility clustering of ε_t.

5.1.2 Quasi Maximum Likelihood Estimation

When the conditional distribution g is not normal, the ML approach described in Section 4.3.3 cannot be directly used, because this estimation procedure assumes that the true conditional distribution of innovations is normal. However, Gouriéroux, Monfort, and Trognon (1984), Weiss (1986), and Bollerslev

and Wooldridge (1992) have shown that, provided the first and second moments are correctly specified, the parameters θ pertaining to the conditional mean and the conditional variance equations can be consistently estimated by maximizing the normal likelihood function, even if the true distribution is not normal. This procedure is called the Quasi Maximum Likelihood Estimation (QMLE). The ML and QML estimators $\hat{\theta}$ of θ are the same, because they are the solutions of the same maximization problem. However, the covariance matrices of the estimators differ, because the QML covariance matrix is computed without assuming conditional normality.

The QML estimator of θ, denoted $\hat{\theta}_{QML}$, is obtained by maximizing the conditional normal log-likelihood function, defined as

$$L_T\left(\theta|\underline{x}_T\right) = \sum_{t=1}^{T} \ell_t\left(\theta\right), \tag{5.8}$$

where

$$\ell_t\left(\theta\right) = -\frac{1}{2}\log\left(\sigma_t^2\left(\theta\right)\right) + \log\left(g\left(z_t\right)\right) = -\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma_t^2\left(\theta\right)\right) - \frac{1}{2}z_t^2,$$

and $z_t = \left(x_t - \mu_t\left(\theta\right)\right)/\sigma_t\left(\theta\right)$. So the maximization problem only depends on the first and second moments. Consequently, if they are correctly specified, the QML estimation will lead to a consistent estimator $\hat{\theta}_{QML}$.

The asymptotic distribution of $\hat{\theta}_{QML}$ is given by

$$\sqrt{T}(\hat{\theta}_{QML} - \theta_0) \Rightarrow \mathcal{N}(0, \Omega), \tag{5.9}$$

where θ_0 is the true value of the parameter vector. Under normality, the asymptotic covariance matrix Ω is simply the inverse of the information matrix, as described in Section 4.3.3. When normality is not assumed, the standard errors of the QML estimator have to be "robustified", following White (1982) and Gouriéroux, Monfort, and Trognon (1984).[1] They give asymptotically valid confidence intervals for the "pseudo-true" parameter values that minimize the information distance between the true probability measure and the normal likelihood. The robust standard errors are the square roots of the diagonal elements of the matrix

$$\Omega = A_0^{-1} B_0 A_0^{-1},$$

where A_0 is the information matrix evaluated at the true parameter vector θ_0

$$A_0 = -E\left[\frac{\partial^2 \ell_t\left(\theta_0\right)}{\partial\theta\partial\theta'}\right],$$

and B_0 is the outer product of the gradients evaluated at θ_0

[1] In the statistical literature, such standard errors are also called "sandwiched standard errors", given that a matrix is "sandwiched" between two others.

$$B_0 = \frac{1}{T}\sum_{t=1}^{T} E\left[\frac{\partial \ell_t(\theta_0)}{\partial \theta'}\frac{\partial \ell_t(\theta_0)}{\partial \theta}'\right],$$

with $\partial \ell_t(\theta_0)/\partial \theta'$ the score vector of $\ell_t(\theta)$.

In finite sample, the asymptotic covariance matrix is estimated by

$$\hat{\Omega}_T = \hat{A}_T^{-1}\hat{B}_T\hat{A}_T^{-1},$$

where

$$\hat{A}_T = -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \ell_t\left(\hat{\theta}_{QML}\right)}{\partial\theta\partial\theta'},$$

$$\hat{B}_T = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \ell_t\left(\hat{\theta}_{QML}\right)}{\partial\theta'}\frac{\partial \ell_t\left(\hat{\theta}_{QML}\right)}{\partial\theta}',$$

are evaluated at the QML estimates $\hat{\theta}_{QML}$.[2]

The QML procedure has the advantage of robustness with respect to the distributional assumption of the model. Obviously, if the distribution is not normal, the QML estimator becomes inefficient. It has been shown by Engle and González-Rivera (1991) that Gaussian QMLE has a degree of inefficiency that increases with the degree of departure from normality. Using simulation evidence, Engle and González-Rivera (1991) show that the loss of efficiency in using the Gaussian QMLE instead of the MLE with the true distribution may be as high as 84% (i.e., the variance of the QMLE can be 6.25 times larger than the variance of the MLE). This finding strongly supports the call for more efficient estimators.

A current practice, therefore, consists in using QMLE for a non-Gaussian distribution. For instance, Bollerslev (1987) estimates a GARCH model assuming a Student t distribution. In this case, the log-likelihood is maximized under the assumed distribution, using the corresponding likelihood

$$L_T(\psi|x_t, t = 1, \cdots, T) = \sum_{t=1}^{T}\ell_t(\psi),$$

where

$$\ell_t(\psi) = -\frac{1}{2}\log\left(\sigma_t^2(\theta)\right) + \log\left(g\left(z_t(\theta)|\eta\right)\right),$$

and $\psi = \left(\theta', \eta'\right)'$ is a vector of parameters. The covariance matrix of the estimator $\hat{\psi}_{QML}$ is robustified to account for departure from the assumed distribution.

[2] Notice that, if the innovation process is actually normally distributed, we have $A_0 = B_0$, so that $\Omega = A_0^{-1}$. Under normality, the asymptotic covariance matrix of the ML estimator is simply given by the inverse of the Hessian matrix.

The dilemma in abandoning the Gaussian QMLE, as Newey and Steigerwald (1997) point out, is that when an incorrect non-Gaussian specification is used, the estimators may no longer be consistent. They show that consistency of a non-Gaussian QMLE is achieved if either (i) the conditional mean is identically zero; or (ii) the assumed (theoretical) and true (empirical) error *pdf*s are symmetric about zero.

If the symmetry condition is not satisfied, the correct specification of the conditional mean and variance is no longer sufficient to ensure consistency of the QML estimators, as the mean and the variance need not correspond to the natural location and scale parameters.[3] An additional location parameter is needed in this case to position the true distribution in order to satisfy the identification condition for consistency. The location parameter accounts for the asymmetry of z_t (i.e., the discrepancy between the conditional mean and the natural location parameter) and can be introduced in either the conditional mean equation or the distribution function. Newey and Steigerwald (1997) suggest estimating the following model

$$x_t = \mu_t(\theta) + \sigma_t(\theta)(\alpha + z_t),$$

so that

$$\ell_t(\theta, \eta, \alpha) = -\frac{1}{2}\log\left(\sigma_t^2(\theta)\right) + \log\left(g\left(\frac{x_t - \mu_t(\theta) - \alpha\sigma_t(\theta)}{\sigma_t(\theta)}|\eta\right)\right).$$

Hence, a crucial issue when a non-normal likelihood is used for the QMLE is whether adequation tests confirm that the assumed distribution correctly fits the data.

5.1.3 The existence of distribution with given moments

The first issue in considering alternative distributions to the normal distribution is that not all values of skewness and kurtosis are attainable. This is closely related to conditions for the existence of moments, which was first investigated by Stiltjes (1894) for the case of a density with bounded support and Hamburger (1920) for the case of unbounded support. Widder (1946) reminds that, to ensure the existence of moments, the following sequence of inequalities, involving moments μ_j and their determinants must be satisfied

$$\mu_0 \geq 0, \qquad \begin{vmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{vmatrix} \geq 0, \qquad \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix} \geq 0, \qquad \cdots$$

[3] For a density $g(x)$, the natural location parameter μ and scale parameter σ are those that minimize

$$E\left[-\frac{1}{2}\log\left(\sigma^2\right) + \log\left(g\left(\frac{x_t - \mu}{\sigma}\right)\right)\right].$$

where $\mu_i = \int z^i f(z)\, dz$. By construction, we have $\mu_0 = 1$, because the *pdf* integrates to 1. For the case where z_t are standardized innovations, we have $\mu_1 = 0$ and $\mu_2 = 1$. This implies the following relation between skewness μ_3 and kurtosis μ_4

$$\mu_3^2 < \mu_4 - 1 \qquad \text{with} \qquad \mu_4 > 0. \tag{5.10}$$

This relation shows that, for a given level of kurtosis, only a finite range of skewness may be attained. The curve in Figure 5.2 delimits the skewness-kurtosis boundary corresponding to the domain (5.10). For all pairs of skewness and kurtosis between the upper and lower curves, a density exists.

Any proposed extension to the normal distribution with finite third and fourth moments will have a domain of definition inside the curves, and the exact coverage will depend on the given distribution and, more precisely, on the way asymmetry and fat tails are introduced in the distribution. We will illustrate this further in the next section using several alternative distributions proposed in the literature.

5.2 Distributions with higher moments

There are several ways of dealing with asymmetry or fat tails in a distribution. First, some distributions allow asymmetry or fat tails. For instance, the skewed Student t distribution or the Pearson IV distribution are directly

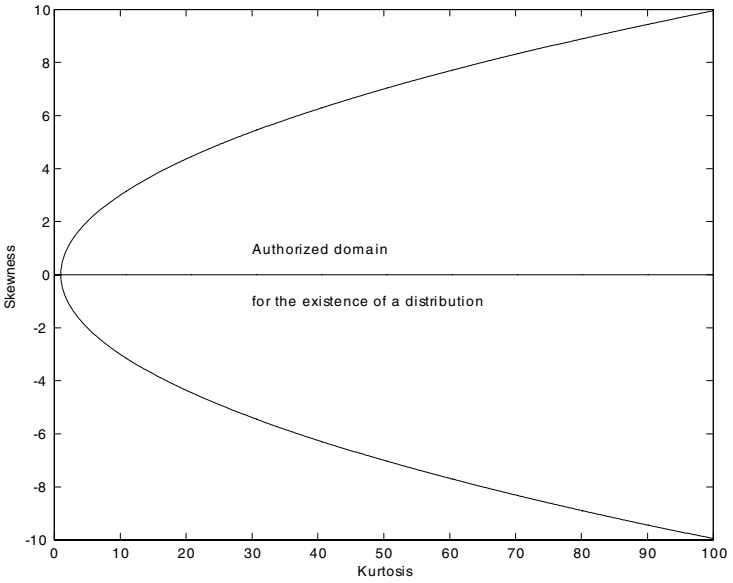

Fig. 5.2. *General skewness-kurtosis boundary.*

designed to incorporate such features. Second, asymmetry can also be introduced, using an expansion about a symmetric distribution. The main interest of the latter approach is that once the general framework has been developed, such a generalization can be applied to any symmetric distribution.

Engle and González-Rivera (1991) and Gallant, Hsieh, and Tauchen (1991) are among the first to address the issue of asymmetry and tail thickness in the conditional distribution; both attempt to extend from the GARCH type models. Engle and González-Rivera (1991) adopt a semi-parametric approach whereas Gallant, Hsieh, and Tauchen (1991) use a truncated Hermite polynomial expansion. Further work in this strand of the research includes Hansen (1994), Drost and Klaassen (1997), Harvey and Siddique (1999), Jondeau and Rockinger (2001, 2003a), Lambert and Laurent (2002), Premaratne and Bera (2000), and Rockinger and Jondeau (2002). Some of these approaches are presented in the following section.

Several alternative distributions have been proposed to capture asymmetry and fat tails. We do not present all possible extensions but rather a set of distributions, which have been found to have interesting properties. Other approaches may be found in Drost and Klaassen (1997) (who proposed an adaptive estimator), Bränas and Nordman (2003) (log-generalized gamma distribution), or in Lambert and Laurent (2002) (skewed location-stable distribution).

5.2.1 Semi-parametric approach

In their so-called semi-parametric ARCH model, Engle and González-Rivera (1991) assume that first and second moments are given by a parametric ARMA process and a parametric ARCH model, respectively, while the conditional density is approximated by a non-parametric density estimator. The overall model they consider is given by (5.1)–(5.4). However the conditional distribution $g(z_t)$ is not assumed to be known but is estimated non-parametrically in a two-step procedure. The assumption is that z_t is *iid* with zero mean and unit variance. The log-likelihood function is defined by

$$\ell(\psi | x_t, t = 1, \cdots, T) = \sum_{t=1}^{T} \ell_t(\psi), \tag{5.11}$$

where

$$\ell_t(\psi) = -\frac{1}{2} \log\left(\sigma_t^2(\theta)\right) + \log\left(g\left(z_t(\theta) | \eta\right)\right),$$

with $\psi = \left(\theta', \eta'\right)'$ the vector of unknown parameters.

To maximize the log-likelihood function (5.11), the following procedure is used:

(i) An initial estimate of the set of parameters θ is given by $\tilde{\theta}$. It may be obtained by QML estimation of (5.2) and (5.4).

(ii) The fitted residuals $\hat{\varepsilon}_t$ and the fitted variances $\hat{\sigma}_t^2(\tilde{\theta})$ are used to compute the standardized residuals $\hat{z}_t(\tilde{\theta}) = \hat{\varepsilon}_t/\hat{\sigma}_t(\tilde{\theta})$, which should have zero mean and unit variance.
(iii) The density $g(\hat{z}_t(\theta))$ is estimated using a non-parametric method. The estimated density is denoted $\hat{g}$.
(iv) The log-likelihood is computed using equation

$$L_T(\theta|x_t, t = 1, \cdots, T) = \sum_{t=1}^{T} \ell_t(\theta),$$

where

$$\ell_t(\theta) = -\frac{1}{2}\log\left(\hat{\sigma}_t^2(\theta)\right) + \log(\hat{g}).$$

The log-likelihood function is maximized, with $\hat{g}$ held fixed, iterating steps (ii)–(iv) until convergence.
(v) Although the estimated density is assumed to be known when the log-likelihood is estimated, the score vector $\partial \ell_t(\theta_0)/\partial \theta'$ (that can be used for the estimation) is computed using the derivatives of $g(\hat{z}_t(\theta))$ with respect to θ. The score is therefore given by

$$\frac{\partial \ell_t(\theta_0)}{\partial \theta'} = -\frac{1}{2}\frac{1}{\sigma_t^2}\frac{\partial \sigma_t^2}{\partial \theta'} + \frac{1}{2}\left(\frac{\partial \varepsilon_t}{\partial \theta'} - \frac{1}{2}\frac{\partial \sigma_t^2}{\partial \theta'}\frac{\varepsilon_t}{\sigma_t^2}\right)\frac{\partial g(\hat{z}_t(\theta))}{\partial \theta'}\frac{1}{g(\hat{z}_t(\theta))}.$$

Engle and González-Rivera (1991) use the discrete maximum penalized likelihood estimation technique developed by Tapia and Thompson (1978) to estimate the non-parametric density. This technique works as follows. The density is approximated using the histogram of z_t with knots $(n_1, \cdots, n_{m-1})$ and heights $(p_1, \cdots, p_{m-1})$ over an interval (a, b) divided in m sub-intervals of length q. This approximation is illustrated in Figure 5.3.

For a sample $(z_1, \cdots, z_T)$, the following optimization problem is solved

$$\max_{(p_1, \cdots, p_{m-1})} \quad \sum_{t=1}^{T} \log(g(z_t)) - \frac{\lambda}{q}\sum_{k=1}^{m-1}(p_{k+1} - 2p_k + p_{k-1})^2,$$

$$\text{subject to } q\sum_{k=1}^{m-1} p_k = 1 \qquad p_k \geq 0 \qquad \text{for } k = 1, \cdots, m-1,$$

where $p_0 = p_m = 0$ and

$$g(z) = \begin{cases} p_k + \frac{p_{k+1} - p_k}{q}(z - n_k) & \text{if } z \in [n_k; n_{k+1}), \\ 0 & \text{if } z \notin (n_0; n_m), \end{cases}$$

and λ is the penalty term to ensure smoothness of the estimate of the heights $(p_1, \cdots, p_{m-1})$.

The semi-parametric method avoids some problems of distribution misspecification, because using a non-normal distribution may lead to inconsistent parameter estimates if the distribution is incorrect. At the same time, the semi-parametric estimator may be more efficient (i.e., with smaller standard error) than a fully non-parametric method.

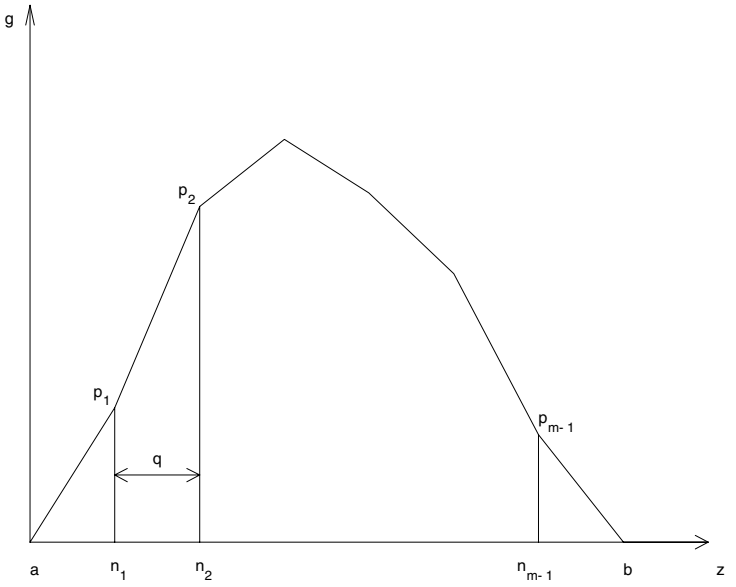

Fig. 5.3. *Example of semi-parametric distribution.*

5.2.2 Series expansion about the normal distribution

Early attempts to explicitly model departure from normality using series expansion around the normal distribution are by Gallant and Tauchen (1989), Gallant, Hsieh, and Tauchen (1991), and Lee and Tse (1991). Gallant, Hsieh, and Tauchen (1991) adopt a very general specification of the series expansion. Provided that large data sets are available, this semi-nonparametric approach may reveal a lot of information about the true distribution. However, such a parameterization is not parsimonious, and consequently estimation may be very computationally expensive. In addition, due to the number of parameters to be estimated, the characteristics of the distribution are virtually impossible to study.

A special case of a series expansion is the so-called Gram-Charlier expansion. Gallant and Tauchen (1989) and Lee and Tse (1991) use Gram-Charlier expansions to describe deviations from normality of innovations in a GARCH framework. Gram-Charlier expansions allow for additional flexibility over a normal distribution because they naturally introduce the skewness and kurtosis of the distribution as unknown parameters. However, being polynomial approximations, they have the drawback of yielding negative values for certain parameters. Moreover, there does not seem to be an easy and analytic characterization of those parameters for which the density will take positive values. Barton and Dennis (1952) establish conditions on the parameters guarantee-

ing positive definiteness of the underlying densities. Jondeau and Rockinger (2001) show that these conditions can be implemented numerically.

Distribution

When the true *pdf* of a random variable Z is unknown, yet believed to be rather close to a normal one, it is quite natural to approximate it with a *pdf* of the form

$$g(z|\eta) = \varphi(z)\, p_n(z|\eta), \tag{5.12}$$

where $\varphi(z)$ is the standard normal density with zero mean and unit variance and where $p_n(z|\eta)$ is chosen so that $g(z|\eta)$ has the same first moments as the *pdf* of Z. A widely used approximation of the true density $g(z|\eta)$ is based on the $(n+1)$ first Hermite polynomials, i.e.,

$$p_n(z|\eta) = \sum_{i=0}^{n} c_i He_i(z), \tag{5.13}$$

where $He_i(z)$ denotes the Hermite polynomial of order i. It is defined by $He_i(z) = (-1)^i \frac{\partial^i \varphi}{\partial z^i} \frac{1}{\varphi(z)}$.[4]

Two representations have been typically adopted in the literature, the Gram-Charlier type A expansion

$$p_4(z|\eta) = 1 + \frac{\gamma_1}{6} He_3(z) + \frac{\gamma_2}{24} He_4(z), \tag{5.14}$$

and the Edgeworth expansion

$$p_6(z|\eta) = 1 + \frac{\gamma_1}{6} He_3(z) + \frac{\gamma_2}{24} He_4(z) + \frac{\gamma_1^2}{72} He_6(z), \tag{5.15}$$

with $\eta = (\gamma_1, \gamma_2)'$.

The Edgeworth expansion (5.15) involves an additional Hermite polynomial while keeping the number of parameters constant. Therefore, the range for γ_1 and γ_2 over which positivity of the approximation is guaranteed is smaller than for the Gram-Charlier one. For this reason, we focus on the first approximation.

It can be shown that the two parameters γ_1 and γ_2 coincide with the skewness and excess kurtosis of $g(z|\eta)$, respectively. Indeed, since the approximating density is assumed to have zero mean and unit variance, we have (Johnson, Kotz, and Balakrishnan, 1994)

[4] Straightforward computations yield the following expressions for the first six Hermite polynomials: $He_0(z) = 1, He_1(z) = z$, $He_2(z) = z^2 - 1$, $He_3(z) = z^3 - 3z$, $He_4(z) = z^4 - 6z^2 + 3$, $He_5(z) = z^5 - 10z^3 + 15z$, and $He_6(z) = z^6 - 15z^4 + 45z^2 - 15$.

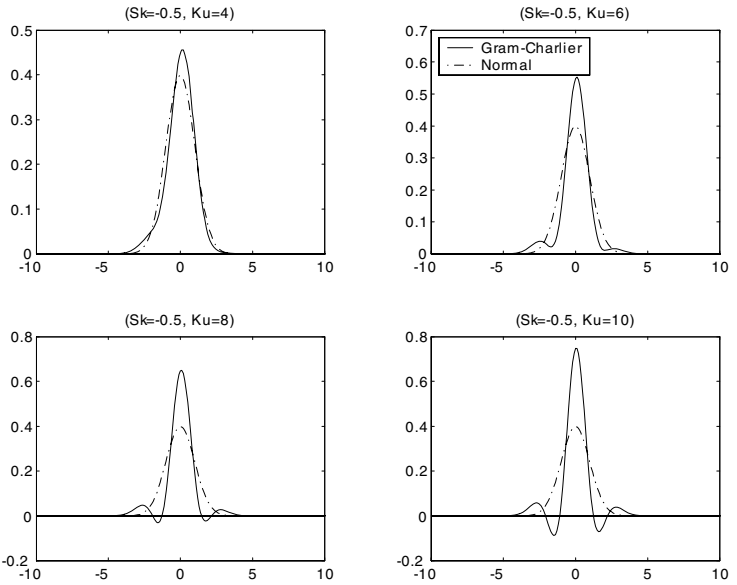

Fig. 5.4. *Pdf of the Gram-Charlier distribution for various values of skewness and kurtosis (without positivity constraint).*

$$\int_{-\infty}^{+\infty} zg(z\,|\eta)dz = 0, \qquad \int_{-\infty}^{+\infty} z^2 g(z\,|\eta)dz = 1,$$
$$\int_{-\infty}^{+\infty} z^3 g(z\,|\eta)dz = \gamma_1, \quad \text{and} \quad \int_{-\infty}^{+\infty} z^4 g(z\,|\eta)dz = 3+\gamma_2.$$

This property partly explains the success of Gram-Charlier expansions in the empirical literature, because the two additional parameters γ_1 and γ_2 are directly related to the third and fourth moments. However, Gram-Charlier expansions also have some drawbacks. A first shortcoming is that, for some (γ_1, γ_2) distant from the normal values $(0, 0)$, $g\,(z\,|\eta\,)$ can be negative for some z. For other pairs, the *pdf* $g(z\,|\eta\,)$ may be multimodal. These problems are illustrated in Figure 5.4. In addition, as will be shown below, the domain of definition for which the distribution is well defined turns out to be rather small.

Domain of definition

We focus now on implementing numerical conditions so that Gram-Charlier approximations are positive. To ensure positivity, Gallant and Tauchen (1989) suggest to square the polynomial part, $p_n\,(z\,|\eta\,)$, of (5.12). But by doing so, we lose the interpretation of the various parameters as moments of the density.

Jondeau and Rockinger (2001) provide some analytical results for computing the skewness-kurtosis boundary ensuring that the Gram-Charlier approximation is a density.

Some properties are useful to identify the region $\mathcal{D}$ in the (γ_1, γ_2)-plane for which $g(z|\eta)$ is positive. For $g(z|\eta)$ to be positive, the polynomial $p_4(z|\eta)$ is required to be positive for every z, that is

$$1 + \frac{\gamma_1}{6} He_3(z) + \frac{\gamma_2}{24} He_4(z) \geq 0, \qquad \forall z.$$

For a given value of z, the equation

$$p_4(z|\eta) = 1 + \frac{\gamma_1}{6} He_3(z) + \frac{\gamma_2}{24} He_4(z) = 0 \tag{5.16}$$

defines a straight line in the (γ_1, γ_2)-plane. A small deviation for z, while holding (γ_1, γ_2) fixed, will then yield a $p_4(z)$ of either positive or negative sign. Thus, we determine the set of (γ_1, γ_2), as a function of z, such that $p_4(z|\eta)$ remains zero for small variations of z, because this set will define the requested boundary. This set is determined by the derivative of (5.16) with respect to z

$$\frac{\gamma_1}{2} He_2(z) + \frac{\gamma_2}{6} He_3(z) = 0. \tag{5.17}$$

The set of (γ_1, γ_2) that solves (5.16) and (5.17) simultaneously, also called the *envelope* of $p_4(z|\eta)$, yields a parametric representation of the boundary where $p_4(z|\eta)$ is zero for a given z. Once this boundary is determined, it remains to find that sub-region delimited by $p_4(z|\eta) = 0$ for all z.[5]

Solving the system given by (5.16) and (5.17) yields the expression for the skewness and the excess kurtosis as functions of z

$$\gamma_1(z) = -24 \frac{He_3(z)}{d(z)},$$
$$\gamma_2(z) = 72 \frac{He_2(z)}{d(z)},$$

with $d(z) = 4(He_3(z))^2 - 3He_2(z)He_4(z)$. Straightforward computations allow us to rewrite the denominator of both expressions as $d(z) = z^6 - 3z^4 + 9z^2 + 9$. Since its minimum is attained for $z = 0$ where $d(0) = 9$, we conclude that $d(z)$ is always positive.

The sign of $\gamma_2(z)$ changes with $He_2(z) = z^2 - 1$. It is positive for $z \in (-\infty; -1] \cup [1; +\infty)$. It is negative for $z \in [-1; 1]$. Similarly, the sign of $s(z)$ changes with $He_3(z) = z^3 - 3z$. It is positive for $z \in \left(-\infty; -\sqrt{3}\right] \cup \left[0; \sqrt{3}\right]$ and negative elsewhere.

These properties are summarized in Figure 5.5. The authorized domain, ensuring the existence of the distribution, is clearly symmetric with respect

[5] This approach has been highlighted by Barton and Dennis (1952) in a slightly different context.

to the horizontal axis. We observe that the excess kurtosis γ_2 is inside the interval $[0, 4]$. Indeed, we find that $\gamma_2(\pm\infty) = 0$ and $\gamma_2\left(\pm\sqrt{3}\right) = 4$. Since γ_2 is bounded below by 0, the kurtosis of $g(\cdot)$ will always be larger than for a normal distribution. Finally, we notice that the authorized domain for skewness and kurtosis is rather small. In many empirical applications, in particular at the daily frequency, the Gram-Charlier density would not be able to capture the skewness and kurtosis found in financial returns.

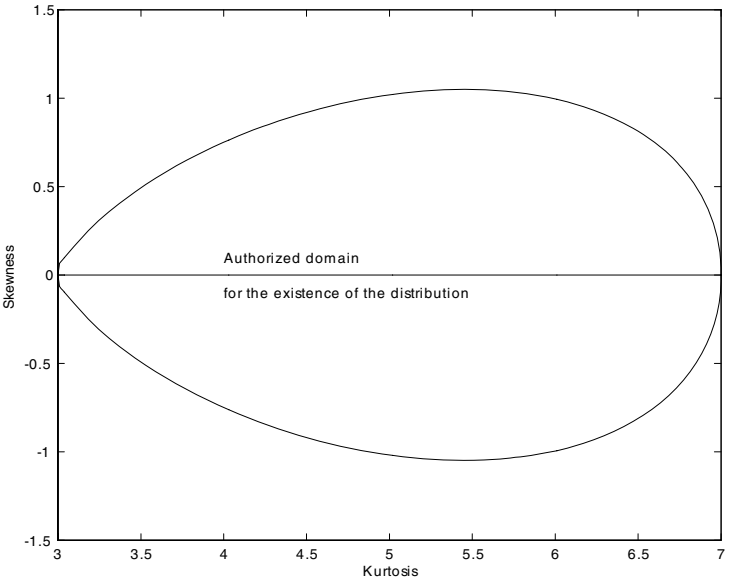

Fig. 5.5. *Skewness-kurtosis boundary for positive Gram-Charlier distribution.*

5.2.3 Skewed Student t distribution

The use of the Student t distribution to capture the fat tails of financial returns goes back to Bollerslev (1987) and Bollerslev and Wooldridge (1992). It is, however, a symmetric distribution, so that it cannot capture asymmetry. A generalization that fills this gap has been proposed by Hansen (1994), with the so-called skewed Student t distribution. He achieves this by introducing a generalization of the Student t distribution where asymmetries may occur, while maintaining the assumption of a zero mean and unit variance. He also illustrates how parameters, and subsequently higher moments, can be rendered time varying. Further extensions of this distribution are by Theodossiou (1998) and Jondeau and Rockinger (2003a, 2003b). A definite advantage of

the skewed Student t distribution is that multivariate extensions are available (See Section 6.2).

Distribution

In the following presentation, we directly define the distribution for the innovation Z, because this is the variable we are ultimately interested in. The skewed Student t distribution proposed by Hansen (1994) is defined by

$$g(z|\nu,\lambda) = b\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{\zeta^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \tag{5.18}$$

where

$$\zeta = \begin{cases} (bz+a)/(1-\lambda) \text{ if } z < -a/b, \\ (bz+a)/(1+\lambda) \text{ if } z \geq -a/b. \end{cases}$$

The constant terms a and b are defined as

$$a = 4\lambda c\frac{\nu-2}{\nu-1},$$
$$b^2 = 1+3\lambda^2-a^2,$$

and are introduced to obtain a variable Z with zero mean and unit variance,[6] where we denote

$$c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\,\Gamma\left(\frac{\nu}{2}\right)}.$$

Here, ν denotes the degree-of-freedom parameter and λ the asymmetry parameter, so that the vector of shape parameters is $\eta = (\nu,\lambda)'$. The density is defined for $2 < \nu < \infty$ and $-1 < \lambda < 1$. Furthermore, it encompasses a large set of conventional densities. For instance, if $\lambda = 0$, Hansen's distribution is reduced to the traditional Student t distribution, without asymmetry. If, in addition, $\nu = \infty$, the Student t distribution collapses to the normal density.[7]

Figure 5.6 illustrates the *pdf*s that can be obtained with the skewed Student t distribution. Upper (lower) figures represent skewed t distributions with large negative (positive) asymmetry. Left-side (right-side) figures correspond to distributions with Gaussian (fat) tails.

[6] As already mentioned, the mean of an asymmetric distribution is not the actual location parameter, a problem discussed by Newey and Steigerwald (1997).

[7] As described in Section 5.2.4, Fernández and Steel (1998) also proposed a method to make any symmetric density asymmetric by some change of variable. The link between their parameter ξ (using the notation of Section 5.2.4) and the parameter λ of Hansen (1994) is simply given by

$$\lambda = \frac{\xi^2-1}{\xi^2+1}.$$

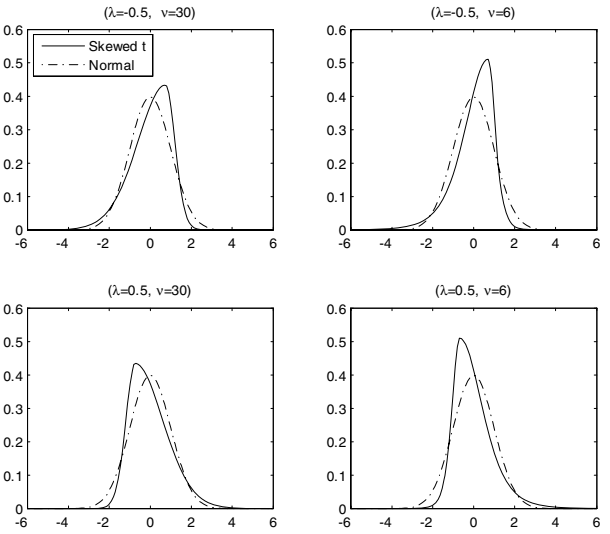

Fig. 5.6. *Pdf of the skewed Student t distribution for various values of ν and λ.*

It is well-known that a traditional Student t distribution with ν degrees of freedom will allow for the existence of up to νth moments (see, for instance, Mood, Graybill, and Boes, 1974). Therefore, given the restriction $\nu > 2$, the skewed Student t distribution is well defined and its second moment exists. The higher moments can be computed as follows. Let $\tilde{m}_r$ be the rth moment of the standard (symmetric) Student t distribution[8]

$$\tilde{m}_r = 2\int_0^\infty x^r t(x|\nu)dx = \frac{\Gamma\left(\frac{\nu-r}{2}\right)\Gamma\left(\frac{r+1}{2}\right)(\nu-2)^{\frac{r+1}{2}}}{\sqrt{\pi(\nu-2)}\,\Gamma\left(\frac{\nu}{2}\right)}.$$

Then, the rth moment of the innovation $Z^* = bZ + a$ of the skewed Student t distribution can be computed as

$$M_r = E\left[(Z^*)^r\right] = \tilde{m}_r\left[(-1)^r(1-\lambda)^{r+1} + (1+\lambda)^{r+1}\right],$$

so that we have, from Jondeau and Rockinger (2003a)

[8] This expression is obtained using the following result of Gradshteyn and Ryzhik (1994, p. 341, 3.241.4):

$$\int_0^\infty x^{\mu-1}(p+qx^\nu)^{-(n+1)}dx = \frac{1}{\nu p^{n+1}}\left(\frac{p}{q}\right)^{\frac{\mu}{\nu}}\frac{\Gamma\left(\frac{\mu}{\nu}\right)\Gamma\left(1+n-\frac{\mu}{\nu}\right)}{\Gamma(1+n)},$$

In particular, we have $\tilde{m}_1 = 2c\frac{\nu-2}{\nu-1}$, $\tilde{m}_2 = 1$, $\tilde{m}_3 = 4c\frac{(\nu-2)^2}{(\nu-1)(\nu-3)}$, and $\tilde{m}_4 = 3\frac{\nu-2}{\nu-4}$.

$$
\begin{aligned}
M_1 &= 4c\lambda\frac{\nu-2}{\nu-1} = a,\\
M_2 &= 1+3\lambda^2 = b^2+a^2,\\
M_3 &= 16c\,\lambda(1+\lambda^2)\frac{(\nu-2)^2}{(\nu-1)(\nu-3)} \qquad \text{if } \nu>3,\\
M_4 &= 3\frac{\nu-2}{\nu-4}(1+10\lambda^2+5\lambda^4) \qquad \text{if } \nu>4.
\end{aligned}
$$

We observe that, when $\lambda \neq 0$, Z^* does not have zero mean and unit variance anymore. This is because the transformation done to introduce asymmetry reallocates probability mass from one side of the distribution to the other. Moments of the standardized innovations, $Z = (Z^* - a)/b$, are then defined as $\mu^{(r)} = E\left[(Z)^r\right]$ with

$$E[Z] = \mu^{(1)} = 0, \tag{5.19}$$

$$V[Z] = \mu^{(2)} = 1, \tag{5.20}$$

and given that Z has zero mean and unit variance,

$$S[Z] = \mu^{(3)} = \frac{M_3 - 3aM_2 + 2a^3}{b^3}, \tag{5.21}$$

$$K[Z] = \mu^{(4)} = \frac{M_4 - 4aM_3 + 6a^2M_2 - 3a^4}{b^4}. \tag{5.22}$$

For some applications, it is very useful to have the *cdf* as well as its inverse. For instance, as will be shown in Chapter 8, VaR computation requires the expression of the inverse *cdf* of the skewed t distribution, because the VaR is based on the quantile of the distribution. It is shown in Jondeau and Rockinger (2003a) that the *cdf* is defined by

$$G(z) = \begin{cases} (1-\lambda)T\left(\frac{bz+a}{1-\lambda}\sqrt{\frac{\nu}{\nu-2}}|\nu\right) & \text{if } z < -a/b,\\ (1+\lambda)T\left(\frac{bz+a}{1+\lambda}\sqrt{\frac{\nu}{\nu-2}}|\nu\right) - \lambda \text{ if } z \geq -a/b, \end{cases}$$

where $T(x|\nu)$ is the *cdf* of the standard t distribution with ν degrees of freedom

$$T(x|\nu) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{w^2}{\nu}\right)^{-\frac{\nu+1}{2}} dw.$$

It is straightforward to show that the inverse of the *cdf* of G is written as

$$G^{-1}(y) = \begin{cases} \frac{1}{b}\left((1-\lambda)\sqrt{\frac{\nu-2}{\nu}}T^{-1}\left(\frac{y}{1-\lambda}|\nu\right) - a\right) \text{ if } y < \frac{1-\lambda}{2},\\ \frac{1}{b}\left((1+\lambda)\sqrt{\frac{\nu-2}{\nu}}T^{-1}\left(\frac{y+\lambda}{1+\lambda}|\nu\right) - a\right) \text{ if } y \geq \frac{1-\lambda}{2}. \end{cases} \tag{5.23}$$

As is well-known, the inverse of the *cdf* allows the simulation of pseudo-random variates. It suffices to generate $y \sim U(0,1)$ and to compute $z = G^{-1}(y)$. Then, z will be distributed as G.

Domain of definition

The density and the various moments do not exist for all parameters. As already mentioned, the density is defined only for $\nu > 2$ and $-1 < \lambda < 1$. Furthermore, careful scrutiny of the algebra yielding (5.21) shows that skewness exists if $\nu > 3$. Last, kurtosis in (5.22) is well defined if $\nu > 4$.

We define as $\mathcal{D}$ the domain $(\nu, \lambda) \in]2, +\infty[\times]-1, 1[$. Given these restrictions on the underlying parameters, it is clear that the range of skewness and kurtosis will also be restricted to a certain domain. In Figure 5.7, we trace various curves relating skewness to λ varying between -1 and 1 for selected values of ν. Similarly, in Figure 5.8, we trace various curves relating kurtosis to ν varying between 4.1 and 8 for selected values of λ.

Focusing on Figure 5.7, we notice that, as ν decreases, skewness can attain very large values. On the other hand, as shown by Figure 5.8, when we increase ν, say around 8, the tails of the density become thinner, even for extreme values of the λ parameter. This picture illustrates the fact that, for a given level of kurtosis, only a finite range of skewness exists. This feature raises the question of existence of a density for given moments.

Figure 5.9 displays the skewness-kurtosis boundary ensuring the existence of a density. The curve ABC corresponds to the theoretical domain of maximal size given by inequality (5.10). The curve DEF corresponds to the domain of skewness and kurtosis, which is attainable with a skewed Student t distribution, assuming $\nu > 2$. We notice that the kurtosis is bounded from below by

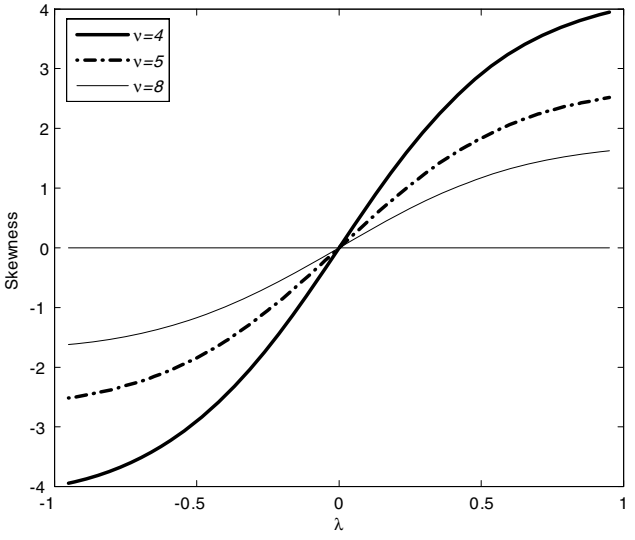

Fig. 5.7. *Skewness for various values of* ν.

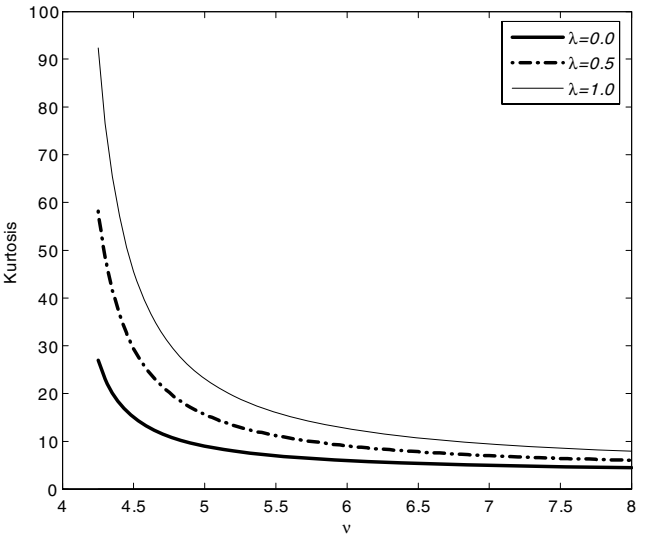

Fig. 5.8. *Kurtosis for various values of* λ.

3, indicating that the skewed Student t distribution does not allow tails to be thinner than those of the normal distribution.

The domain denoted $\mathcal{E}$ corresponding to DEF in Figure 5.9 is spanned by skewness and kurtosis if both moments exist. We notice that the relation between $\mathcal{D}$ and $\mathcal{E}$ is not one-to-one. In particular, those points that are located in $\mathcal{D}$ but where $\nu < 4$ have no counterpart in $\mathcal{E}$. The logic is that there are points in $\mathcal{D}$ where skewness or kurtosis cease to exist, whereas in $\mathcal{E}$ skewness and kurtosis are finite by construction. It is only when we reduce the domain $\mathcal{D}$ to $]4, +\infty[\times]-1, 1[$ that the relation is bijective.

Alternative specifications

After the initial description of the skewed Student t distribution by Hansen (1994), several alternative specifications for an asymmetric Student t distributions have been proposed. Some of these specifications can be viewed as particular cases of the general approaches presented in the next section to generate skewness from a symmetric distribution.

Fernández and Steel (1998) have specialized their general approach to generate asymmetric distributions (based on hidden truncation) to the case of the Student t distribution

$$g\left(z\,|\nu,\xi\right) = \frac{2}{\xi + \frac{1}{\xi}}\left[t\left(z\xi|\nu\right)1_{(-\infty,0)}\left(z\right) + t\left(\frac{z}{\xi}|\nu\right)1_{[0,\infty)}\left(z\right)\right].$$

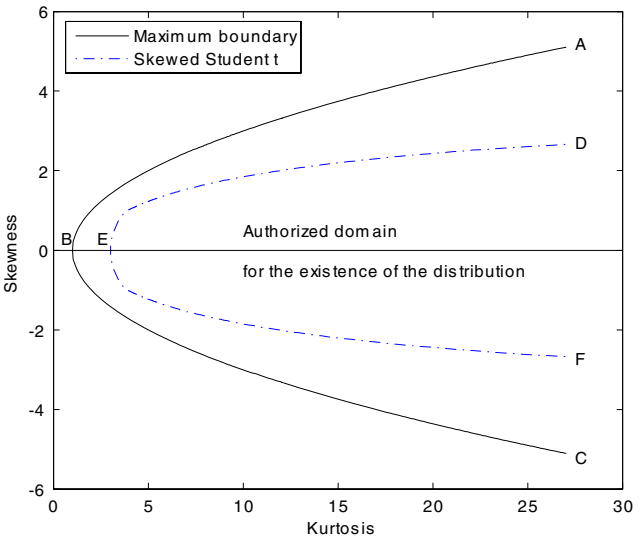

Fig. 5.9. *Skewness-kurtosis boundary for skewed t distribution.*

It can be easily seen that this distribution is directly related to the one proposed by Hansen (1994) through a mere change of notation of the asymmetry parameter. Lambert and Laurent (2002) have also extended this approach to the case of the skewed stable distribution and, more generally, to the case of the skewed location-scale distributions. Notice that the general approach of generating asymmetric distribution will be used in Section 6.2 to define multivariate skewed distributions.

Azzalini and Capitanio (2003) have proposed another approach to generate some asymmetry in (multivariate) distributions. In the case of the univariate Student t distribution, they propose the use of

$$g\left(z|\nu,\lambda\right)=2t\left(z|\nu\right)T\left(\lambda z\sqrt{\frac{\nu+1}{z^2+\nu}}|\nu+1\right), \tag{5.24}$$

where $T\left(z|\nu\right)$ is the *cdf* of the Student t distribution with ν degrees of freedom. When $\lambda=0$, $g\left(z\right)$ reduces to the standard Student t distribution. When $\lambda\rightarrow-\infty$ (∞), we obtain a Student t distribution truncated from above (below) at zero. Parameter λ plays therefore the role of shape parameter.

Jones and Faddy (2003) have constructed the following distribution

$$g\left(z|\alpha,\beta\right)=\frac{1}{2^{\alpha+\beta-1}B\left(\alpha,\beta\right)\sqrt{\alpha+\beta}}\left(1+\frac{z}{\sqrt{\alpha+\beta+z^2}}\right)^{\alpha+\frac{1}{2}}$$
$$\times\left(1-\frac{z}{\sqrt{\alpha+\beta+z^2}}\right)^{\beta+\frac{1}{2}},$$

where $B\left(\alpha,\beta\right)$ denotes the beta function. For $\alpha=\beta$, the distribution reduces to the standard Student t distribution. When $a<\beta$ (resp. $\alpha>\beta$), g is negatively (resp. positively) skewed.

Another alternative specification has been proposed by Harvey and Siddique (1999). Their so-called non-central Student t distribution allows one to model skewness but does not allow for an independent variation of skewness and kurtosis. The non-central Student t distribution is defined by two parameters: ν the degree-of-freedom parameter and δ the non-centrality parameter. This distribution is scaled to have a unit variance, while the non-centrality parameter controls the shape of the distribution. The non-central Student t distribution is defined as

$$g\left(z\,|\nu,\delta\,\right)=\frac{\nu^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\times\frac{\exp\left(-\frac{\delta^2}{2}\right)}{\sqrt{\pi}\left(\nu+z^2\right)^{\frac{\nu+1}{2}}}\times\sum_{i=0}^{\infty}\Gamma\left(\frac{\nu+i+1}{2}\right)\left(\frac{\delta^i}{i!}\right)\left(\frac{2z^2}{\nu+z^2}\right)^{\frac{i}{2}}.$$

Since one of the parameters is needed to control the centrality of the distribution, it is clear that this density is more restrictive as far as skewness and kurtosis are concerned.

5.2.4 Generating asymmetric distributions

Several ways of generating skewness in a symmetric distribution have been proposed. Without analyzing these various approaches in detail, we provide in this section a brief description of the two most well-known methods. See Ferreira and Steel (2004) for a recent review of some approaches developed to generate asymmetry in univariate distributions.

Hidden truncation

The first approach, based on conditioning (or hidden truncation), was introduced by Azzalini and Dalla Valle (1996) and Azzalini and Capitanio (1999) in a multivariate context (it will be described more deeply in Section 6.2).[9] In the case of a univariate distribution, we have

$$g\left(z|\xi\right)=2f\left(z\right)F\left(\xi z\right),$$

where f denotes a *pdf* and F its *cdf*. When $\xi=0$, $g\left(z\right)$ reduces to the symmetric distribution f. When $\xi\to-\infty$ (∞), we obtain a Student t distribution

[9] See also Arnold and Beaver (2000).

truncated from above (below) at zero. The parameter ξ plays therefore the role of shape (skewness) parameter.

A difficulty with this approach is that moments are hard to compute for general distributions f. In the case of the Student t distribution, we obtain (5.24). Then, as shown by Azzalini and Capitanio (2003), the first moments are

$$E[Z] = \mu = \delta \left(\frac{\nu}{\pi}\right)^{1/2} \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \quad (\nu > 1),$$

$$V[Z] = \sigma^2 = \frac{\nu}{\nu-2} - \mu^2 \quad (\nu > 2),$$

$$\tilde{S}[Z] = \mu \left(\frac{\nu\left(3-\delta^2\right)}{\nu-3} - \frac{3\nu}{\nu-2} + 2\mu^2 \right) \quad (\nu > 3),$$

$$\tilde{K}[Z] = \frac{3\nu^2}{(\nu-2)(\nu-4)} - \frac{4\mu^2\nu\left(3-\delta^2\right)}{\nu-3} + \frac{6\mu^2\nu}{\nu-2} - 3\mu^4 - \frac{3}{\sigma^4} \quad (\nu > 4),$$

where $\delta = \xi/\sqrt{1+\xi^2}$.

Standardized skewness and kurtosis are defined in the usual way by the expressions $S[Z] = \tilde{S}[Z]/\sigma^3$ and $K[Z] = \tilde{K}[Z]/\sigma^4$. Note that we have $E[Z] = S[Z] \gtreqless 0$ when $\xi \gtreqless 0$. Moreover $E[Z|\xi] = -E[Z|-\xi]$ and $S[Z|\xi] = -S[Z|-\xi]$. This shows that the parameter ξ directly controls the asymmetry of the distribution.

The expressions above clearly indicate that Z does not have zero mean and unit variance anymore. The reason is that the transformation done to introduce asymmetry reallocates some probability mass from one side of the distribution to the other. More precisely, non-standardized innovations Z have mean a and variance b^2. If we now define the standardized innovation $Z^* = (Z-a)/b$, we obtain $E[Z^*] = 0$, $V[Z^*] = 1$, $S[Z^*] = S[Z]$ and $K[Z^*] = K[Z]$. The new innovation Z^* has the same skewness and kurtosis as Z.

Inverse scale factor

For Bayesian estimation purpose, Fernández and Steel (1998) proposed another method for introducing skewness into any continuous unimodal distribution symmetric around the origin. The method involves changing the scale of the distribution at each side of the mode. The new asymmetric random variable Z is then distributed as[10]

[10] The distribution $g(z|\xi)$ may be viewed as a distribution of the mixture (see Lambert and Laurent, 2002) below

$$z = u\xi|x| - (1-u)\frac{1}{\xi}|x|,$$

where u is the realization of a Bernouilli process with probability of success $\xi^2/\left(1+\xi^2\right)$, and x is the realization of a process with distribution f.

$$g(z|\xi) = \frac{2}{\xi + \frac{1}{\xi}} f\left(z\xi^{-sign(z)}\right)$$
$$= \frac{2}{\xi + \frac{1}{\xi}} \left[f(z\xi)\, 1_{(-\infty,0)}(z) + f\left(\frac{z}{\xi}\right) 1_{[0,\infty)}(z) \right],$$

where f denotes any symmetric *pdf*. The asymmetry parameter $\xi > 0$ is such that the ratio of probability masses above and below the mode is

$$\frac{\Pr[Z \geq 0]}{\Pr[Z < 0]} = \xi^2. \tag{5.25}$$

Hence, parameter ξ provides an interesting indicator of the asymmetry in the distribution. The distribution $g(z|1/\xi)$ is the symmetric of $g(z|\xi)$ with respect to the mode, with $g(z|\xi) = g(-z|1/\xi)$. Therefore, inverting ξ produces the mirror image around zero.

Moments of the asymmetric distribution $g(z|\xi)$ are easily deduced from those of the symmetric one $f(z)$. If the rth moment of a r.v. with distribution $f(.)$ exists, then the associated random variable Z with distribution $g(.)$ also has a finite rth moment, defined as

$$M_r = \tilde{m}_r \frac{\xi^{r+1} + \frac{(-1)^r}{\xi^{r+1}}}{\xi + \frac{1}{\xi}},$$

where

$$\tilde{m}_r = 2E[Z^r | Z > 0] = 2\int_0^\infty u^r f(u)\, du$$

is the rth moment of $f(.)$ truncated to the positive real values. Provided that they exist, the first four moments are then obtained as

$$E[Z] = M_1 = \tilde{m}_1 \left(\xi - \frac{1}{\xi}\right) \equiv a,$$
$$V[Z] = M_2 - M_1^2 = \left(\tilde{m}_2 - \tilde{m}_1^2\right)\left(\xi^2 + \frac{1}{\xi^2}\right) + 2\tilde{m}_1^2 - \tilde{m}_2 \equiv b^2,$$
$$\tilde{S}[Z] = M_3 - 3M_1M_2 + 2M_1^3$$
$$= \left(\xi - \frac{1}{\xi}\right)\left[\left(\tilde{m}_3 + 2\tilde{m}_1^3 - 3\tilde{m}_1\tilde{m}_2\right)\left(\xi^2 + \frac{1}{\xi^2}\right) + 3\tilde{m}_1\tilde{m}_2 - 4\tilde{m}_1^3\right],$$
$$\tilde{K}[Z] = M_4 - 4M_1M_3 + 6M_2M_1^2 - 3M_1^4.$$

Note that we have $S[Z|\xi] = -S[Z|1/\xi]$ and $S[Z] = 0$ when $\xi = 1$.

As before, Z does not have zero mean and unit variance anymore, but can be standardized to yield innovations $Z^* = (Z - a)/b$, with $E[Z^*] = 0$, $V[Z^*] = 1$, $S[Z^*] = S[Z]$, and $K[Z^*] = K[Z]$.

5.2.5 Pearson IV distribution

Premaratne and Bera (2000) and Brännäs and Nordman (2001) have proposed the Pearson type IV distribution to model the dynamic of asset returns. The main interest of this distribution is that the three parameters can be interpreted as the variance, skewness, and kurtosis of the process, so that it is possible to directly model these moments if required.

Distribution

This distribution is a member of the Pearson family, which is characterized as the distributions $g(z^* | \eta)$ satisfying

$$\frac{d \log (g(z^* | \eta))}{dz} = \frac{z^* - \beta}{b_0 + b_1 z^* + b_2 z^{*2}},$$

where β, b_0, b_1, and b_2 are parameters to be estimated. Inside this group, assuming some specific values of the parameters, we obtain the normal, gamma, Student t, or Pareto distributions. In the general case where $b_1 \neq 0$ and $b_2 \neq 0$, and where the roots of $b_0 + b_1 z^* + b_2 z^{*2}$ are imaginary, say $b + ia$ and $b - ia$, the resulting distribution is given as

$$g(z^* | \eta) = c^{-1} \left(1 + \left(\frac{z^*}{a}\right)^2\right)^{-m} \exp\left(\delta \arctan\left(\frac{z^*}{a}\right)\right), \tag{5.26}$$

where

$$\begin{aligned} m &= -1/(2b_2), \\ \delta &= (b - \beta)/(ab_2), \end{aligned}$$

and

$$c = a \int_{-\pi/2}^{\pi/2} \cos^r(\omega) \exp(-\delta\omega)\, d\omega$$

is the integration constant that ensures that the density integrates to one. We also define $r = 2(m - 1)$. The vector of shape parameters is thus $\eta = (a, r, \delta)'$. This distribution is known as the Pearson IV distribution. It is of particular interest because it allows for both asymmetry and fat tails. It has been little used in empirical studies, however, mainly because its estimation raises some technical problems.

Domain of definition

We note that Z^* does not have zero mean and unit variance, because we have the following moments (see Kendall and Stuart, 1977, Premaratne and Bera, 2000)

$$
\begin{aligned}
E\left[Z^*\right] &= \frac{\delta a}{r} \qquad (r>0)\,,\\
V\left[Z^*\right] &= \frac{a^2}{r^2\left(r-1\right)}\left(r^2+\delta^2\right) \qquad (r>1)\,,\\
\tilde{S}\left[Z^*\right] &= \frac{4a^3\delta\left(r^2+\delta^2\right)}{r^3\left(r-1\right)\left(r-2\right)} \qquad (r>2)\,,\\
\tilde{K}\left[Z^*\right] &= 3\frac{a^4\left(r^2+\delta^2\right)\left[\left(r+6\right)\left(r^2+\delta^2\right)-8r^2\right]}{r^4\left(r-1\right)\left(r-2\right)\left(r-3\right)} \qquad (r>3)\,,
\end{aligned}
$$

with standardized skewness and kurtosis

$$
\begin{aligned}
S\left[Z^*\right] &= s = \frac{4\delta}{r-2}\sqrt{\frac{r-1}{r^2+\delta^2}},\\
K\left[Z^*\right] &= \kappa = 3\frac{\left(r-1\right)\left[\left(r+6\right)\left(r^2+\delta^2\right)-8r^2\right]}{\left(r-2\right)\left(r-3\right)\left(r^2+\delta^2\right)}.
\end{aligned}
$$

Parameter δ can be viewed as an asymmetry parameter, because it controls the sign of skewness. When $\delta = 0$, with $r > 2$, we obtain $S\left[Z^*\right] = s = 0$, implying symmetry of the distribution. In addition, $\delta > 0$ (< 0) entails a positive (negative) skewness. Parameter r can be viewed as a degree-of-freedom parameter, which controls for fat tails. To see this, note that we have under symmetry ($\delta = 0$)

$$
\kappa = 3\frac{r-1}{r-3},
$$

so that lower values of r imply thicker tails. We obtain normality when $r \to \infty$. However, δ and r cannot be directly interpreted as a skewness and kurtosis parameter, because skewness as well as kurtosis depend on both δ and r.

Figure 5.10 displays the domain of definition of the Pearson type IV distribution. We also report the boundary for skewness and kurtosis ensuring the existence of the distribution $\kappa - s^2 - 1 = 0$ already described (see Section 5.1.3).

Estimation issue

Let us consider now the issue of estimating this model in a time-series context. It should be noted that, since the innovation Z_t^* does not have zero mean, the conditional mean of the return process $X_t = \mu_t\left(\theta\right) + \sigma_t\left(\theta\right) Z_t^*$ is given by

$$
E\left[X_t\left|\mathcal{F}_{t-1}\right.\right] = \mu_t\left(\theta\right) + \sigma_t\left(\theta\right)\frac{\delta a}{r}.
$$

For symmetric unimodal distributions, mean, median, and mode coincide, and therefore correct specification of the conditional mean is sufficient to ensure consistent estimates of the conditional mean parameters. For asymmetric distributions, the mean and variance may not be its natural location and scale

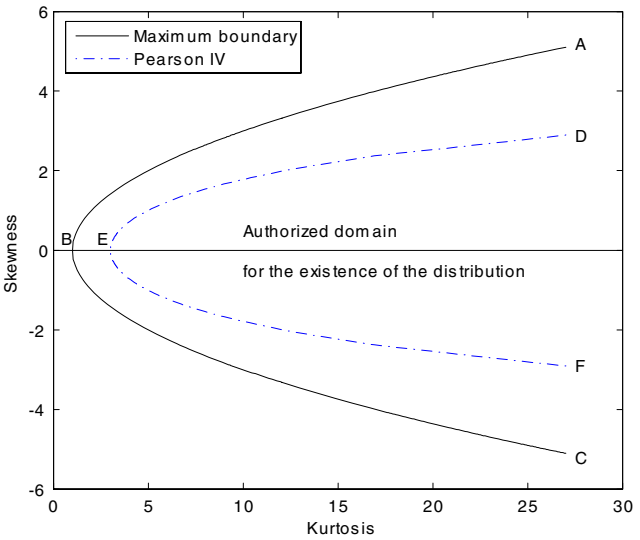

Fig. 5.10. *Skewness-kurtosis boundary for the Pearson IV distribution.*

parameters (Newey and Steigerwald, 1997). In this case, we have to introduce an additional parameter $\tilde{\mu}$ in the density of the innovation process to capture the fact that the mean may not be the natural location parameter. The density function (5.26) therefore becomes

$$g\left(z_t^* \left| \eta, \tilde{\mu} \right.\right) = c^{-1} \left(1 + \left(\frac{z_t^* - \tilde{\mu}}{a}\right)^2\right)^{-m} \exp\left(\delta \arctan\left(\frac{z_t^* - \tilde{\mu}}{a}\right)\right).$$

In this case, we have

$$E\left[Z_t^*\right] = \tilde{\mu} + \frac{\delta a}{r}.$$

Consequently, the conditional mean of the return process X_t becomes

$$E\left[X_t \left| \mathcal{F}_{t-1} \right.\right] = \mu_t\left(\theta\right) + \sigma_t\left(\theta\right)\left(\tilde{\mu} + \frac{\delta a}{r}\right).$$

With this specification, as shown by Newey and Steigerwald (1997), the consistency of the QML estimator is ensured. Finally, the log-likelihood at time t to be maximized is given by

$$\ell_t\left(\psi\right) = -\frac{1}{2}\log\left(\sigma_t^2\left(\theta\right)\right) + \log\left(g\left(\frac{x_t - \mu_t\left(\theta\right) - \sigma_t\left(\theta\right)\tilde{\mu}}{\sigma_t\left(\theta\right)} \left| \eta \right.\right)\right),$$

with $\psi = \left(\theta', \eta', \tilde{\mu}\right)'$ the vector of unknown parameters.

5.2.6 Entropy distribution

Methods based on the entropy principle of Shannon (1948), and popularized by Jaynes (1957, 1982), have made their way into econometrics, e.g., Golan, Judge, and Miller (1996). At a practical level, entropy-based applications still appear to be scarce but for a few exceptions such as Zellner and Highfield (1988), Stutzer (1996), Buchen and Kelly (1996), or Ormoneit and White (1999). The difficulties with the numerical implementation of this technique may have hindered its widespread use. Yet, Rockinger and Jondeau (2002), based on Agmon, Alhassid, and Levine (1979a, 1979b), have shown that it is possible to develop a very rapid method to obtain entropy densities and that entropy densities may also be used in rather complex empirical likelihood estimations. It is worth emphasizing that this approach allows one to estimate distributions with the largest domain of definition possible compatible with the bound found in Figure 5.2.

Distribution

We assume that the econometrician is seeking a probability $p(z)$ defined over some real convex domain, $\mathcal{D}$, while disposing only of information on the m first moments of the probability, written as μ_i where $i = 1, \cdots, m$. 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

$$p \in \arg\max_{\{p\}} \quad -\int_{z\in\mathcal{D}} p(z)\log(p(z))dz, \tag{5.27}$$

$$\text{s.t.} \quad \int_{z\in\mathcal{D}} p(z)dz = 1, \tag{5.28}$$

$$\int_{z\in\mathcal{D}} z^i p(z)dz = \mu_i, \quad i = 1, \cdots, m. \tag{5.29}$$

We refer to a density satisfying these conditions as an Entropy Density.[11] Jaynes (1957) notices that the entropy is a criterion where the statistician imposes a minimum amount of information. The conventional way of solving this program is to define the Hamiltonian

$$H = -\int_{\mathcal{D}} p(z)\log(p(z))dz - \lambda_0^*\left(\int_{\mathcal{D}} p(z)dz - 1\right) - \sum_{i=1}^{m}\lambda_i\left(\int_{\mathcal{D}} z^i p(z)dz - \mu_i\right),$$

[11] Given that a log-function is involved in (5.27), $p(z) \geq 0$, $\forall z \in \mathcal{D}$.

where the λ_0^* is a Lagrange parameter as are the λ_i, $i = 1, \ldots, m$.

To obtain a solution of this problem, we seek a zero for the Fréchet derivative. Defining $\lambda_0' = \lambda_0^* + 1$, we get

$$\delta H = 0 \qquad \Rightarrow \qquad p(z) = \exp\left(-\lambda_0' - \sum_{i=1}^{m} \lambda_i z^i\right). \tag{5.30}$$

Derivation with respect to the $m + 1$ Lagrange multipliers yields the $m + 1$ conditions (5.28) to (5.29).

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 we obtain the uniform distribution over $\mathcal{D}$. As we add the first and second moments, Golan, Judge, and Miller (1996) recall that we obtain 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. We focus now on how densities may be obtained in a numerically efficient manner if third and higher moments are given. Some results can be found in Zellner and Highfield (1988) and Ormoneit and White (1999), while Rockinger and Jondeau (2002) describe a more efficient estimation technique.

Estimation

Substitution of (5.30) into (5.28) defines a function that, as shown later, turns out to be a *potential function*. The expression of this function is

$$P(\lambda_1, \cdots, \lambda_m) \equiv \exp\left(-\lambda_0'\right) = \int_{\mathcal{D}} \exp\left(\sum_{i=1}^{m} \lambda_i z^i\right) dz, \tag{5.31}$$

so that

$$p(z) = \exp\left(\sum_{i=1}^{m} \lambda_i z^i\right) / P(\lambda_1, \cdots, \lambda_m). \tag{5.32}$$

For a given set of $\lambda = (\lambda_1, \cdots, \lambda_m)'$, one could evaluate (5.32) and, thus, the moment restrictions (5.29). This result suggests as a first estimation technique non-linear least squares applied to (5.29). Such an estimation yields multiple solutions and is rather slow.[12] As found by Agmon, Alhassid, and Levine (1979a, 1979b), 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 some further results.

Since $\int_{\mathcal{D}} p(z)dz = 1$, multiplication of the right-hand side of (5.29) by this integral and the grouping under one single integral yields

[12] The technique developed by Ormoneit and White (1999) follows this approach. They show how such an algorithm may be implemented more efficiently as in Zellner and Highfield (1988), yet, they report estimations lasting several seconds whereas the ones of Rockinger and Jondeau (2002) take a fraction of a second.

$$\int_{\mathcal{D}} (z^i - \mu_i) p(z) dz = 0, \quad i = 1, \cdots, m.$$

Furthermore, writing $p(z) = \exp\left(\lambda_0 + \sum_{i=1}^m \lambda_i (z^i - \mu_i)\right)$, where $\lambda_0 = \lambda_0' + \sum_{i=1}^m \lambda_i \mu_i$ indicates that the number of computations required to evaluate (5.32) subject to (5.29) may be reduced. Also, the passage from λ_0' to λ_0 is a trivial linear transformation. Again, $p(z)$ must satisfy (5.29), and this yields a definition for λ_0

$$Q(\lambda_1, \cdots, \lambda_m) \equiv \exp(-\lambda_0) = \int_{\mathcal{D}} \exp\left(\sum_{i=1}^m \lambda_i (z^i - \mu_i)\right) dz, \tag{5.33}$$

so that the probability can be rewritten as

$$p(z) = \exp\left(\sum_{i=1}^m \lambda_i (z^i - \mu_i)\right) / Q(\lambda_1, \cdots, \lambda_m). \tag{5.34}$$

At this point, we have obtained two equivalent definitions for the density, namely (5.32) and (5.34). Depending on the situation, one definition or the other is useful. With the definition of (5.33), we obtain that

$$g_i \equiv \frac{\partial Q}{\partial \lambda_i} = 0 \Rightarrow \int_{\mathcal{D}} (z^i - \mu_i) p(z) dz = 0,$$

and, therefore, the zeros of the gradient of Q yield the first-order conditions. This computation validates the claim that Q defines a potential.[13] Next, we obtain that

$$G_{ij} \equiv \frac{\partial^2 Q}{\partial \lambda_i \partial \lambda_j} = \int_{\mathcal{D}} (z^i - \mu_i)(z^j - \mu_j) p(z) dz,$$

showing that the Hessian matrix is a variance-covariance matrix. As a consequence, the Hessian matrix is symmetric and positive definite. An inverse of the Hessian will exist if the matrix is of full rank. This last condition implies that, as long as $p(z)$ is a density, the minimization of Q has a unique solution. We write the gradient of Q as g and its 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 solution. This solution, in turn, will define an entropy distribution. 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 is numerically stable if Q is of full rank and if the solution is finite.

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

As Agmon, Alhassid, and Levine (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, and we show how the existence of a finite solution can be guaranteed. The discretization of the potential (5.33) and the constraints (5.28) and (5.29) can be obtained by a Gauss-Legendre approximation (see, for instance, Davis and Polonsky, 1970) that yields

$$Q(\lambda) = \sum_{j=1}^{n} \exp\left(\sum_{i=1}^{m} \lambda_i (x_j^i - \mu_i)\right) w_j, \tag{5.35}$$

$$\sum_{j=1}^{n} w_j p_j = 1, \tag{5.36}$$

$$\sum_{j=1}^{n} w_j z_j^i p_j = \mu_i, \qquad i = 1, \cdots, m, \tag{5.37}$$

$$p_j \geq 0, \qquad j = 1, \cdots, n, \tag{5.38}$$

where $w = (w_1, \cdots, w_n)'$ is a vector of weights where the integral is evaluated. Those values are tabulated, for instance, in Abramowitz and Stegun (1970). 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.

This set of equations (5.36)–(5.38) can be viewed as a linear programming problem where we seek a solution to $m + 1$ equations under positivity constraints. If a solution exists, the algorithm will find it within $m+1$ and $2(m+1)$ steps. Now, if a solution exists, then it is known that Q will be minimized for some finite solution. The problem thus consists in numerically minimizing $Q(\lambda)$. As pointed out by Fletcher (1994), many algorithms are available. If the problem is known to have a single minimum, Newton's method works well. We may start with an initial value $\lambda^{(0)} = (0, \cdots, 0)$ the vector with m zeros. At step k, we update the vector as

$$\lambda^{(k)} = \lambda^{(k-1)} + \delta^{(k)},$$

where $\delta^{(k)}$ is solution to $G^{(k)}\delta^{(k)} = -g^{(k)}$. This latter condition guarantees that the kth approximation in a second-order Taylor expansion of Q, that is $Q(\lambda^{(k-1)} + \delta^{(k)}) = Q(\lambda^{(k-1)}) + \delta^{(k)} g^{(k)} + \frac{1}{2}\delta^{(k)'} G^{(k)} \delta^{(k)}$, leads to a *flat* spot of Q, which is an extremum.

Domain of definition

We focus now, without loss of generality, on the study of the densities of a standardized variable Z that satisfy $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = s$, and $\mu_4 = \kappa$. In

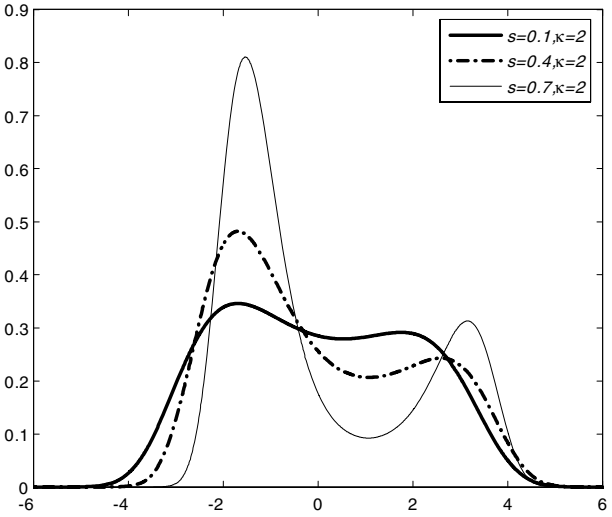

Fig. 5.11. *Entropy distributions for various levels of skewness.*

this case, the parameters s and κ represent skewness and kurtosis, respectively. A solution to the minimization of the potential (5.33) subject to constraints (5.36)–(5.38) is guaranteed only if parameters s and κ are in the domain of definition $\mathcal{D}$. We may determine numerically the domain of definition for the entropy distribution. This may be done, for instance, using a grid-search over a large skewness-kurtosis domain where a solution to the simplex algorithm might exist. Doing so, we verify numerically that the boundary for the entropy distribution in fact corresponds to the maximum boundary $\mu_3^2 < \mu_4 - 1$.

Next, we consider how the entropy density behaves as skewness, s, and kurtosis, κ, vary. An inspection of Figures 5.11 and 5.12 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 multimodal 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. We obtain that entropy densities may be of use when the tails of the distributions are much thinner than the tails of the normal density. Inversely, κ may become very large allowing for fat tails.

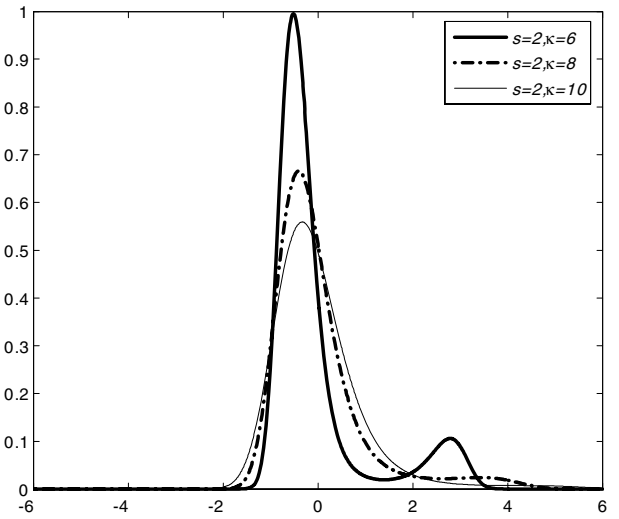

Fig. 5.12. *Entropy distributions for various levels of kurtosis.*

5.3 Specification tests and inference

5.3.1 Moment specification tests

The use of distributions allowing higher moments is primarily intended to capture some relevant statistical features of the data. Therefore, some conditions between theoretical and sample moments should hold. These tests imply non-linear restrictions on the parameter vector and can be viewed as mis-specification tests.[14] They are related to the test already presented in Chapter 3 to test the mixture of distributions hypothesis.

To briefly describe the moment specification tests, let θ denote the $(n, 1)$ vector of parameters and $r(\theta)$ denote the $(J, 1)$ restriction functions (or moment conditions) (examples of such restrictions are given below). The null hypothesis is $H_0 : r(\theta) = 0$. The Wald test statistic is defined by $W = r(\hat{\theta})' \hat{\Omega}^{-1} r(\hat{\theta})$, where $\hat{\Omega}$ is an estimate of the (J, J) covariance matrix Ω of $r(\hat{\theta})$. Under the null hypothesis, W is asymptotically distributed as a $\chi^2(J)$.

Let $m_i(z_t, \theta)$ be the ith moment condition for the tth observation. Then, the ith restriction is defined as $r_i(\theta) = \frac{1}{T} \sum_{t=1}^{T} m_i(z_t, \theta)$. The empirical coun-

[14] See Newey (1985), Nelson (1991), Harvey and Siddique (1999), and Brooks, Burke, and Persand (2005).

terpart is simply $r_i(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \hat{m}_i\left(z_t, \hat{\theta}\right)$, where $\hat{m}_i\left(z_t, \hat{\theta}\right)$ is the estimated moment. We note $\hat{m}_j = \left[\hat{m}_j\left(z_1, \hat{\theta}\right) \cdots \hat{m}_j\left(z_T, \hat{\theta}\right)\right]'$.

The covariance matrix Ω may be estimated using the derivatives of the log-likelihood function

$$\hat{\Omega} = \frac{1}{T^2}\left(M'M - M'D\left(D'D\right)^{-1}D'M\right),$$

with $M = \left(\hat{m}_{t,j}\right)_{t,j}$ the (T, J) matrix containing the empirical moment conditions

$$M = \left[\hat{m}_1 \cdots \hat{m}_J\right] = \begin{bmatrix} \hat{m}_1\left(z_1, \hat{\theta}\right) & \cdots & \hat{m}_J\left(z_1, \hat{\theta}\right) \\ \vdots & \ddots & \vdots \\ \hat{m}_1\left(z_T, \hat{\theta}\right) & \cdots & \hat{m}_J\left(z_T, \hat{\theta}\right) \end{bmatrix}.$$

Each element $\hat{m}_{t,j}$ is not required to be identically zero, but the jth moment restriction states that the sum of the $\hat{m}_{t,j}$ across observations should be zero. Finally, the (T, n) matrix of derivatives D is defined as

$$D = \left[\hat{d}_1 \cdots \hat{d}_n\right] = \begin{bmatrix} \hat{d}_{1,1} & \cdots & \hat{d}_{1,n} \\ \vdots & \ddots & \vdots \\ \hat{d}_{T,1} & \cdots & \hat{d}_{T,n} \end{bmatrix},$$

where the (t, j)th element $\hat{d}_{t,j}$ is the derivative of the log-likelihood at date t with respect to the jth parameter θ_j

$$\hat{d}_{t,j} = \hat{d}_{t,j}\left(\hat{\theta}\right) = \left.\frac{\partial \ell_t(\theta)}{\partial \theta_j}\right|_{\theta=\hat{\theta}}.$$

The only remaining issue now is to determine the appropriate form of the standardized residual that should be selected as the basis for the test. Under the correct model specification, standardized residuals should be *iid* with the postulated distribution as conditional distribution. Since we consider standardized residuals, they should have zero mean and unit variance, resulting in the following orthogonality conditions for the first and the second moments

$$\begin{aligned} E\left[z_t\right] &= 0, \\ E\left[z_t^2 - 1\right] &= 0. \end{aligned}$$

Additional conditions depend on the postulated distribution, and, in particular, on whether asymmetry and fat tails are allowed. For example, a conditional normal distribution would imply that skewness and excess kurtosis are zero

$$\begin{aligned} E\left[z_t^3\right] &= 0, \\ E\left[z_t^4 - 3\right] &= 0, \end{aligned}$$

while assuming a conditional Student t distribution would imply

$$E\left[z_t^3\right] = 0,$$
$$E\left[z_t^4 - 3\frac{\nu-2}{\nu-4}\right] = 0.$$

Additional orthogonality conditions can be imposed on the serial correlation of a given moments or on even higher moments. For instance, in the normal case, we have

$$\begin{aligned} E\left[z_t z_{t-j}\right] &= 0 \qquad j = 1, 2, \cdots, \\ E\left[(z_t^2 - 1)(z_{t-j}^2 - 1)\right] &= 0, \\ E\left[z_t^3 z_{t-j}^3\right] &= 0, \\ E\left[(z_t^4 - 3)(z_{t-j}^4 - 3)\right] &= 0. \end{aligned}$$

These moment conditions translate into the restrictions that the first four powers of the series should not be serially correlated.

5.3.2 Adequacy tests based on density forecasts

As pointed out by Engle and González-Rivera (1991), the QMLE has a degree of inefficiency that increases with the degree of departure from normality. Maximizing the log-likelihood using the correct distribution of Z_t is therefore likely to improve significantly the efficiency of the estimator. Unfortunately, if we use an incorrect non-normal specification, this estimator is not consistent (Newey and Steigerwald, 1997). Therefore, it appears crucial to check the validity of the distributional assumption by carrying out adequacy tests. Some of these tests are presented below.

A first natural way to test the adequacy of the estimated distribution to the data is to measure the distance between the empirical (unknown) distribution and the assumed (estimated) distribution. This type of test is similar to the Kolmogorov-Smirnov test widely used to test normality (see Section 2.2.3).

Let $f_t(z_t)$ denote the empirical (unknown) *pdf*, which we try to describe as accurately as possible using the parametric distribution $g_t(z_t)$. For instance, f_t may be the actual distribution of the SP500, and g_t may be one of the distributions described in Section 5.2. Now, define the probability integral transform as $u_t = \int_{-\infty}^{z_t} g_t(y_t)\, dy_t$ (Rosenblatt, 1952). We may view u_t as the value taken by the *cdf* at z_t. Then, the crucial result of this approach is the following.

Proposition 5.1. *(Diebold, Gunther, and Tay, 1998). Suppose* $\{z_t\}_{t=1}^T$ *is generated from the distribution* $\{f_t(z_t|I_t)\}_{t=1}^T$ *where* $I_t = \{z_{t-1}, z_{t-2}, \cdots\}$. *If a sequence of density forecasts* $\{g_t(z_t)\}_{t=1}^T$ *coincides with* $\{f_t(z_t|I_t)\}_{t=1}^T$,

then under the usual condition of a non-zero Jacobian with continuous partial derivatives, the sequence of probability integral transforms of $\{z_t\}_{t=1}^T$ *with respect to* $\{g_t(z_t)\}_{t=1}^T$ *is iid* $U(0,1)$*, that is*

$$\{u_t\}_{t=1}^T \sim iid\ U(0,1).$$

In other words, if the postulated distribution $g_t(z_t)$ is correct, then its probability integral transform u_t is *iid* $U(0,1)$. Diebold, Gunther, and Tay (1998) propose a two-step test. First, we test the null hypothesis that u_t is serially uncorrelated. It can be done using a standard LM test. For this purpose, Diebold, Gunther, and Tay suggest examining the autocorrelations of $(u_t - \bar{u})^i$, for $i = 1, \cdots, 4$, by regressing $(u_t - \bar{u})^i$ on K lags of the variable.[15] The LM test statistic is defined as $(T-K)R^2$, where R^2 is the coefficient of determination, and is distributed, under the null of no serial correlation, as a χ^2 with K degrees of freedom.

Second, we test the null hypothesis that u_t is distributed as a uniform $U(0,1)$. For this purpose, Diebold, Gunther, and Tay divide the empirical and theoretical distribution into N cells and test whether the two distributions significantly differ on each cell. Such an approach allows a graphical representation that can be used to identify areas where the theoretical distribution fails to fit the data. The test statistic

$$DGT(N) = \sum_{n=1}^{N} \frac{(F_n - T/N)^2}{T/N},$$

where F_n is the number of observations in the cell n and T/N is the expected number of observations under the null. Under the null of correct specification, the test statistic $DGT(N)$ is asymptotically distributed as a $\chi^2(N-1)$. Vlaar and Palm (1993) note that when the test statistic depends on p estimated parameters, the asymptotic distribution of $DGT(N)$ is in fact bounded between a $\chi^2(N-1)$ and a $\chi^2(N-p-1)$.

5.3.3 Adequacy tests based on interval forecasts

A very promising area of research for adequacy tests is the "hit" test proposed by Christoffersen (1998). He generalizes the approach based on density forecasts by focusing on conditional interval forecast. The idea is to test if the probability of being in a given interval, based on the given distribution, is compatible with the actual data.

Define the indicator variable Hit_t for a given interval forecast $(L_{t/t-1}(p), U_{t/t-1}(p))$ at time t as

$$Hit_t = \begin{cases} 1 & \text{if } z_t \in \left[L_{t/t-1}(p), U_{t/t-1}(p)\right], \\ 0 & \text{if } z_t \notin \left[L_{t/t-1}(p), U_{t/t-1}(p)\right], \end{cases}$$

[15] Zero correlation is equivalent to independence only under normality. The correlogram is, therefore, only suggestive of possible independence.

where $L_{t/t-1}(p)$ and $U_{t/t-1}(p)$ denote, respectively, for the lower and upper limits of the interval.

In an unconditional setup, a natural goodness-of-fit test for a given interval forecast consists in comparing the nominal coverage $\sum_{t=1}^{T} Hit_t/T$ with the true coverage p. This is the test proposed by Diebold, Gunther, and Tay (1998). In a conditional setup, Christoffersen (1998) shows that testing

$$E[Hit_t|\mathcal{F}_{t-1}] = E[Hit_t|Hit_{t-1}, Hit_{t-2}, \cdots] = p,$$

for all t is equivalent to testing if the sequence $\{Hit_t\}_{t=1}^{T}$ is *iid* Bernoulli(p). The test for conditional coverage is, thus, performed in two steps. First, we test the *unconditional coverage*, i.e., the null hypothesis that $E[Hit_t] = p$. It can be done using the likelihood-ratio test statistic

$$LR_{unc} = -2\log\left(\frac{L(p|Hit_t, t = 1, \cdots, T)}{L(\hat{\pi}|Hit_t, t = 1, \cdots, T)}\right),$$

where

$$L(p|Hit_t, t = 1, \cdots, T) = (1-p)^{n_0} p^{n_1},$$
$$L(\hat{\pi}|Hit_t, t = 1, \cdots, T) = (1-\hat{\pi})^{n_0} \hat{\pi}^{n_1},$$

with $\hat{\pi} = n_1/(n_0 + n_1)$ is the MLE of π the true coverage probability. Clearly, n_0 is the number of outcomes in the interval and $n_1 = T - n_0$. Under the null hypothesis, the test statistic LR_{unc} is asymptotically distributed as a $\chi^2(1)$.

Next, we test the *independence of the Hit_ts*. Under the alternative, the indicator variable Hit_t is assumed to be a binary first-order Markov chain, with transition probability matrix

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix},$$

where $\pi_{ij} = \Pr[Hit_t = j|Hit_{t-1} = i]$. Therefore, the likelihood function for this process is

$$L(\Pi_1|Hit_t, t = 1, \cdots, T) = (1-\pi_{01})^{n_{00}} (\pi_{01})^{n_{01}} (1-\pi_{11})^{n_{10}} (\pi_{11})^{n_{11}},$$

where n_{ij} is the number of observations with value i followed by value j. The probability π_{ij} is estimated by $\hat{\pi}_{ij} = n_{ij}/(n_{i0} + n_{i1})$. Under the null of independence, the transition probability matrix reduces to

$$\Pi_2 = \begin{bmatrix} 1-\pi_2 & \pi_2 \\ 1-\pi_2 & \pi_2 \end{bmatrix},$$

with likelihood function

$$L(\Pi_2|Hit_t, t = 1, \cdots, T) = (1-\pi_2)^{n_{00}+n_{10}} (\pi_2)^{n_{01}+n_{11}}.$$

The probability π_2 is estimated by $\hat{\pi}_2 = (n_{01} + n_{11})/T$. Then, the test statistic is

$$LR_{ind} = -2\log\left(\frac{L\left(\hat{\Pi}_2|Hit_t, t = 1, \cdots, T\right)}{L\left(\hat{\Pi}_1|Hit_t, t = 1, \cdots, T\right)}\right).$$

Under the null hypothesis, the test statistic LR_{ind} is asymptotically distributed as a $\chi^2(1)$.

Finally, the joint test of coverage and independence, which corresponds to the *test of conditional coverage*, is given by the test statistic

$$LR_{cc} = -2\log\left(\frac{L(p|Hit_t, t = 1, \cdots, T)}{L\left(\hat{\Pi}_1|Hit_t, t = 1, \cdots, T\right)}\right).$$

which is asymptotically distributed as a $\chi^2(2)$ under the null hypothesis.

This adequacy test has been extended in a number of directions. The most interesting extension is probably the modeling of the probability of being in a given interval. Such an expansion of the information set has been proposed by Christoffersen (1998). The idea is to incorporate some additional variables in the information set such that $\mathcal{F}_{t-1} = \{X_{t-1}, \cdots, X_1\}$, where X_t is a vector of observed variables. Then, if we run the regression

$$Hit_t = \alpha + \beta' f(X_{t-1}) + \varepsilon_t,$$

with f a function of past variables X_{t-1}, the test of interval forecast efficiency with respect to the information set $\mathcal{F}_{t-1}$ is a joint test of independence ($\beta = 0$) and correct unconditional coverage ($\alpha = p$). Since the error term ε_t can be shown to be *iid*, standard OLS technique can be applied to this regression. Such a regression-based approach was also followed by Clements (2002) and Wallis (2003). Engle and Manganelli (2004) adopt a similar approach for evaluating the VaR of a portfolio in the context of quantile regressions (with their CAViaR model).

5.4 Illustration

To illustrate some aspects of the modeling of non-normal distributions, we consider the dynamic of the same four stock market indices as in previous sections, namely the SP500, the DAX, the FT-SE, and the Nikkei indices. We use daily returns over the period from January 1980 to August 2004. We remove the day corresponding to the October 1987 crash. Although it would not affect markedly the estimation of the GARCH models, it would have important consequences for the test of adequacy of the assumed distribution to the empirical distribution.

In order to provide a complete diagnosis on the series at hand, we estimate a model that is designed to capture the serial correlation and heteroskedasticity found in the daily returns. In addition, we look for a conditional distribution that is able to adjust to the empirical distribution of returns. We begin with the following AR(p)–GARCH(1, 1) model with conditionally normal innovations. The number of lags in the AR(p) process is selected in such a way that residuals ε_t can be viewed as serially uncorrelated. Consequently, the optimal number of lags is likely to depend on the return series. Table 5.1 reports the parameter estimates of the AR(p) process. It also presents the LM test for ARCH effects. We notice that the autoregressive parameters for daily returns are not very large but that some of them are strongly significant. Therefore, there is some amount of information in past returns to predict future returns. This autocorrelation may be induced by assets of different liquidity incorporating news at different speed. It should be noticed, however, that the main source of dependency in returns does not come from the correlation between returns but from the correlation between squared returns. As highlighted in the LM test, squared residuals are extremely correlated, indicating that the variability of asset returns changes over time.

Table 5.2 reports the parameter estimates of the GARCH(1, 1) process for residuals. We notice a large persistence in volatility as shown by the sum

Table 5.1. *Parameter estimates of the AR(p) process for daily returns*

	SP500	DAX	FT-SE	Nikkei
μ	0.043	0.037	0.035	0.012
	(0.0130)	(0.0171)	(0.0118)	(0.0165)
ρ_1	0.017	0.008	0.061	0.010
	(0.0127)	(0.0127)	(0.0127)	(0.0128)
ρ_2	−0.016	−0.024	0.010	−0.066
	(0.0127)	(0.0127)	(0.0127)	(0.0128)
ρ_3	−0.051	−0.014	−0.037	–
	(0.0127)	(0.0127)	(0.0127)	
ρ_4	–	0.020	0.042	–
		(0.0127)	(0.0127)	
ρ_5	–	−0.008	−0.019	–
		(0.0127)	(0.0127)	
ρ_6	–	−0.051	−0.023	–
		(0.0127)	(0.0127)	
ρ_7	–	0.018	0.000	–
		(0.0127)	(0.0127)	
ρ_8	–	–	0.048	–
			(0.0127)	
LM test	1710.20	1123.75	1107.65	1125.22
$(T \times R^2)$	(0.000)	(0.000)	(0.000)	(0.000)

$\alpha_1 + \beta_1$.[16] The table also provides results of the adequacy test proposed by Diebold, Gunther, and Tay (1998). In this case, the proposed model is rejected for two reasons. First, the margin u_t is not *iid*. This suggests that the AR(p)–GARCH(1, 1) is not able to capture all the time dependency of returns. Second, the margin is not uniformly distributed. This indicates that the normal distribution is not able to capture the features of the empirical distribution, in particular its asymmetry and fat-tailedness. To get a better identification of the source of rejection of the normal distribution, we compare, for various parts of the distribution of standardized residuals, the expected number of observations under the normal hypothesis and the observed number of observations under the empirical distribution. In Figure 5.13, we plot, for 50 cells of equal probability, the expected number of observations under normality (the medium horizontal line with 125 observations) as well as its confidence interval. We also plot, with diamonds, the observed number of observations, that corresponds to the empirical distribution. The figure suggests that the rejection of the normal distribution for SP500 daily returns comes from the extremes. Crashes are much more frequent and booms are less frequent in the empirical distribution than in the assumed normal distribution.

Now, we assume that innovations are not normally but t distributed, $z_t \sim t_\nu$. Thus we also estimate the degree-of-freedom parameter ν. Table 5.3 reports the corresponding parameter estimates. The degree of freedom lies within the range $[6, 12]$, suggesting that the tails of the empirical distribution are fatter than those implied by the normal distribution. Yet, the fat-tailedness of the empirical distribution is not as extreme as the one we would have had with a stable distribution. Indeed, because our estimates of ν are clearly larger than 4, we can admit that, at least, the first four moments of the distribution do exist.[17] The adequacy test reveals that the t distribution captures most features of the empirical distribution for the SP500 and the DAX indices, as illustrated in Figure 5.14. However, this distribution still fails to adequately represent the distribution of the FT-SE and Nikkei indices, because it implies too many positive extremes to be consistent with the empirical distribution.

Finally, we assume that innovations are distributed as a skewed t distribution: $z_t \sim g(\nu, \lambda)$. We, therefore, estimate the degree-of-freedom parameter ν and the asymmetry parameter λ. Table 5.4 reports the estimation of the GARCH(1, 1) model with skewed t innovation. All asymmetry parameters are negative, although it is barely significant for the SP500. For other indices, we obtain a pronounced leftward asymmetry. The adequacy test reveals that the skewed t distribution captures most features of the empirical distribution of the FT-SE and the Nikkei. Figure 5.15 shows that the fit of the empirical distribution of the SP500 by the skewed t distribution is rather good.

[16] Parameter estimates slightly differ from those reported in Chapter 4 for the same series, because we now take the serial correlation in return into account.

[17] An interesting property of the t distribution with ν degrees of freedom is that all moments of order above ν do not exist.

Table 5.2. *Parameter estimates of the GARCH*$(1,1)$ *process under normality*

	SP500	DAX	FT-SE	Nikkei
ω	0.007	0.030	0.022	0.013
	(0.0014)	(0.0041)	(0.0032)	(0.0021)
α	0.048	0.099	0.099	0.114
	(0.0043)	(0.0081)	(0.0085)	(0.0079)
β	0.945	0.885	0.876	0.886
	(0.0047)	(0.0088)	(0.0100)	(0.0070)
$\log(L_{norm})$	−8318.10	−9543.70	−7646.58	−9126.98
Time independency				
$(u_t - \bar{u})$	7.959	11.213	25.567	37.148
	(0.6329)	(0.3412)	(0.0044)	(0.0001)
$(u_t - \bar{u})^2$	38.857	39.416	20.656	13.661
	(0.0000)	(0.0000)	(0.0236)	(0.1890)
$(u_t - \bar{u})^3$	24.419	27.216	34.567	47.055
	(0.0066)	(0.0024)	(0.0001)	(0.0000)
$(u_t - \bar{u})^4$	22.805	27.371	12.443	13.231
	(0.0115)	(0.0023)	(0.2565)	(0.2111)
Uniform(0,1)	167.758	86.166	88.887	161.102
	(0.000)	(0.001)	(0.000)	(0.000)

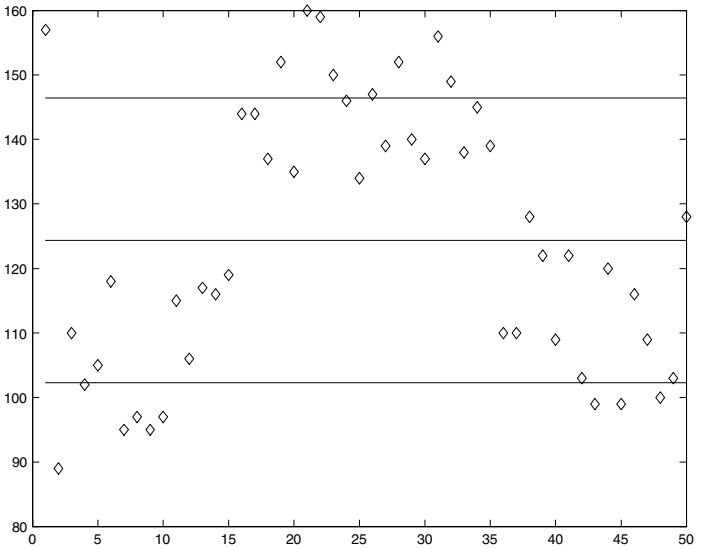

Fig. 5.13. *Goodness-of-fit test to the* $\mathcal{N}(0,1)$ *for SP500 daily returns.*

Table 5.3. *Parameter estimates of the GARCH*$(1,1)$ *process under t innovations*

	SP500	DAX	FT-SE	Nikkei
ω	0.005	0.015	0.015	0.008
	(0.0015)	(0.0034)	(0.0030)	(0.0021)
α	0.044	0.083	0.082	0.092
	(0.0055)	(0.0085)	(0.0085)	(0.0093)
β	0.952	0.909	0.899	0.908
	(0.0059)	(0.0090)	(0.0104)	(0.0087)
ν	6.707	9.808	12.079	6.677
	(0.5508)	(0.9687)	(1.4013)	(0.5256)
$\log(L_{stud})$	−8161.40	−9391.84	−7539.09	−8956.95
Time independency				
$(u_t - \bar{u})$	7.034	10.624	25.270	34.984
	(0.7223)	(0.3875)	(0.0049)	(0.0001)
$(u_t - \bar{u})^2$	40.848	38.271	18.747	16.639
	(0.0000)	(0.0000)	(0.0436)	(0.0827)
$(u_t - \bar{u})^3$	22.235	26.082	34.346	46.900
	(0.0140)	(0.0036)	(0.0002)	(0.0000)
$(u_t - \bar{u})^4$	28.221	28.634	13.482	22.549
	(0.0017)	(0.0014)	(0.1979)	(0.0125)
Uniform(0,1)	60.217	57.974	89.595	68.668
	(0.131)	(0.178)	(0.000)	(0.033)

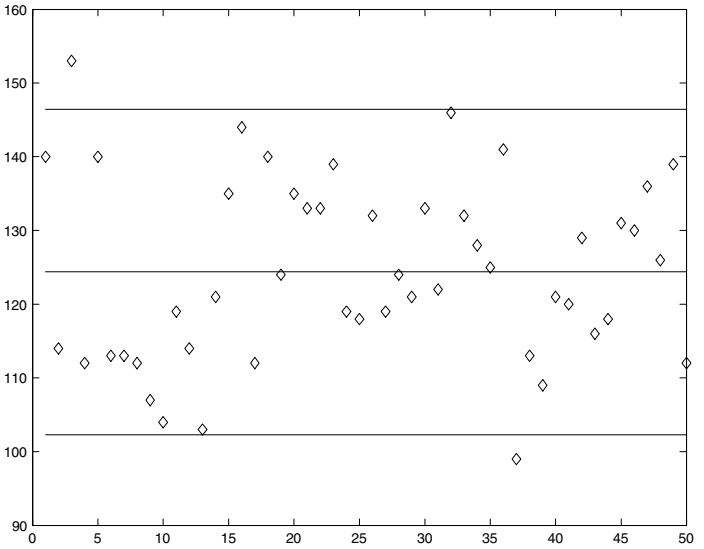

Fig. 5.14. *Goodness-of-fit test to the* t_ν *for SP500 daily returns.*

Table 5.4. *Parameter estimates of the GARCH*$(1,1)$ *process with a skewed t*

	SP500	DAX	FT-SE	Nikkei
ω	0.005	0.015	0.015	0.009
	(0.0015)	(0.0033)	(0.0030)	(0.0022)
α	0.045	0.083	0.082	0.094
	(0.0056)	(0.0085)	(0.0084)	(0.0093)
β	0.952	0.910	0.899	0.906
	(0.0060)	(0.0090)	(0.0102)	(0.0087)
ν	6.764	9.994	12.583	6.629
	(0.5636)	(1.0046)	(1.5042)	(0.5183)
λ	-0.027	-0.084	-0.115	-0.084
	(0.0149)	(0.0189)	(0.0189)	(0.0174)
$\log(L_{skstud})$	-8160.07	-9381.96	-7520.76	-8945.10
Time independency				
$(u_t - \bar{u})$	7.073	10.148	24.773	32.945
	(0.7186)	(0.4276)	(0.0058)	(0.0003)
$(u_t - \bar{u})^2$	41.303	38.789	18.952	13.910
	(0.0000)	(0.0000)	(0.0409)	(0.1771)
$(u_t - \bar{u})^3$	21.679	23.177	30.186	42.200
	(0.0168)	(0.0101)	(0.0008)	(0.0000)
$(u_t - \bar{u})^4$	28.038	28.565	13.580	15.837
	(0.0018)	(0.0015)	(0.1931)	(0.1044)
Uniform(0,1)	66.520	49.938	47.594	34.671
	(0.048)	(0.436)	(0.530)	(0.939)

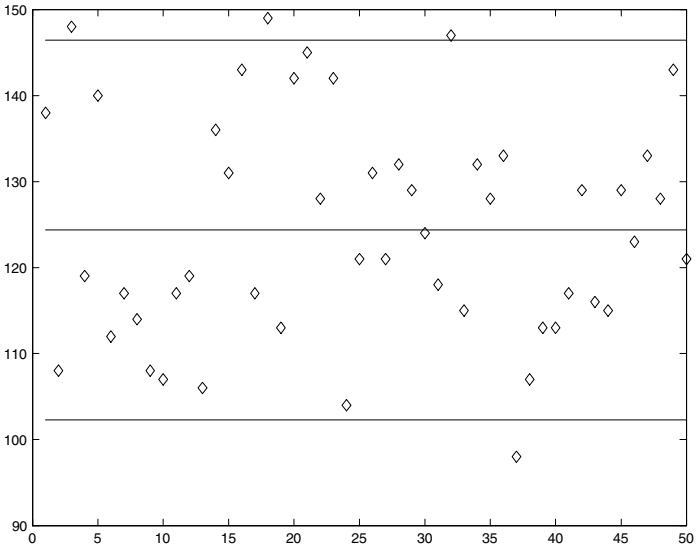

Fig. 5.15. *Goodness-of-fit test to the skewed t for SP500 daily returns.*

To conclude, we observe that the GARCH model is unable to capture all the serial correlation in second and higher moments. Two alternative approaches may improve this disappointing result: (i) estimate an asymmetric GARCH model (such as the GJR or TARCH models), which would probably help capture at least the serial correlation in the second moment; (ii) model the dynamic of the third and fourth moments, as initially proposed in Hansen (1994) and extended in Jondeau and Rockinger (2003a) (See Section 5.5).

5.5 Modeling conditional higher moments

It is possible and sometimes necessary to go one step further and generalize the model with non-normal distribution. For a number of applications, allowing higher moments to vary over time can be very useful. The idea behind this approach is simply that the asymmetry and the fat-tailedness of returns may change over time. For asset allocation, the changing characteristics of the return distribution have to be taken into account by the investors in their decision process. In a non-normal environment, we generally recognize that investors like positive skewness and dislike kurtosis (see Chapter 9). In this context, when higher moments are time-varying, it is important to forecast their future path in order to improve the allocation of wealth. Another instance is the VaR computation of a portfolio with non-normal assets. If the distribution of asset returns varies over time, it is obviously crucial to update the measure of VaR when the asymmetry and the fat-tailedness of the distribution change.

Now, we assume that the characteristics of the shape of the distribution vary over time. Therefore, the parameter vector η is itself allowed to vary over time. Hence, we generalize model (5.1)–(5.5) as follows

$$x_t = \mu_t(\theta) + \sigma_t(\theta) z_t, \tag{5.39}$$

$$\mu_t(\theta) = E[x_t | \mathcal{F}_{t-1}] = \mu(\theta, \mathcal{F}_{t-1}), \tag{5.40}$$

$$\sigma_t^2(\theta) = E\left[(x_t - \mu_t)^2 | \mathcal{F}_{t-1}\right] = \sigma^2(\theta, \mathcal{F}_{t-1}), \tag{5.41}$$

$$z_t \sim g(z_t | \eta_t), \tag{5.42}$$

$$\eta_t = \eta(\theta, \mathcal{F}_{t-1}), \tag{5.43}$$

where the innovation, z_t, follows a conditional distribution g with time-varying parameters. The conditional dependency of shape parameters is given by (5.43). For instance, in the case of the skewed Student t distribution, η_t includes the degree-of-freedom and the asymmetric parameters.

Since the GARCH approach has been empirically successful for modeling the conditional variance, it may seem natural to extend it to include the dynamics of the shape parameters η_t as follows

$$\eta_t = \eta(z_{t-1}, z_{t-2}, \cdots).$$

An important issue is to constrain the dynamics in order to ensure that the function g is always a definite distribution.

Hansen (1994), Harvey and Siddique (1999), Lambert and Laurent (2002), and Jondeau and Rockinger (2003a) have all modeled skewness and distribution parameters as extended GARCH models. This sort of extension is not as straightforward as it may seem. First it requires, at each date and each step of the optimization, to solve the non-linear problem that links parameters and moments. The second problem, as mentioned before, is that a lot of constraints have to be imposed on the time-varying higher moments to make sure that the conditional distribution is well defined. The complex boundary conditions and complex non-linear relationships between parameters make the estimation very tricky, computationally intensive, and time consuming.

5.5.1 Tests for autoregressive conditional higher moments

It is possible to test for autoregressive conditional higher moments in a similar manner to Engle's $T \times R^2$ (Lagrange Multiplier, LM) test for conditional heteroskedasticity. We recall that the latter is performed by regressing the square of non-standardized residuals $\hat{\varepsilon}_t^2$ on p lagged values. Under the null of no conditional heteroskedasticity, the LM statistics $T \times R^2$ is distributed as a $\chi^2(p)$.

Similar tests for autoregressive conditional skewness and kurtosis can be performed. They are based on the fact that, because innovations z_t are standardized, we have $E\left[z_t^3\right] = s$ and $E\left[z_t^4\right] = \kappa$. The idea is then to regress estimates of z_t^3 and z_t^4 on lagged values respectively. Therefore, if we denote $\hat{z}_t$ the estimated standardized residual, we have the two regressions

$$\begin{aligned} &\text{for skewness:} && \hat{z}_t^3 = a_0 + a_1 \hat{z}_{t-1}^3 + \cdots + a_p \hat{z}_{t-p}^3 + u_t, \\ &\text{for kurtosis:} && \hat{z}_t^4 = b_0 + b_1 \hat{z}_{t-1}^4 + \cdots + b_p \hat{z}_{t-p}^4 + u_t. \end{aligned}$$

Under the null of no autoregressive conditional skewness or kurtosis, the LM statistic $T \times R^2$ is distributed as a $\chi^2(p)$.[18]

Note that the test of constancy of higher moments should be performed on standardized residuals z_t instead of the non-standardized error term ε_t. The reason is that the skewness and kurtosis of the non-standardized error term (i.e., $s\sigma_t^3$ and $\kappa\sigma_t^4$) are directly related to volatility, so that they will be time-varying in case of heteroskedasticity.

5.5.2 Modeling higher moments directly

Several studies have directly considered the modeling of higher moments. Harvey and Siddique (1999) focus on the modeling of the dynamics of skewness in a non-central Student t distribution. Rockinger and Jondeau (2002) model

[18] This result holds if moments up to the eighth one exist. For financial data, this assumption may be wrong.

skewness and kurtosis in the context of an entropy distribution. Brooks, Burke, and Persand (2005) consider an autoregressive conditional kurtosis in the context of the standard Student t distribution.

In Harvey and Siddique (1999), the dynamics of skewness is given by

$$s_t = \alpha_0 + \alpha_1 s_{t-1} + \alpha_2 z_{t-1}^3, \tag{5.44}$$

where z_t is the standardized innovation, with zero mean and unit variance. They placed the following constraints on the parameters: $-1 < \alpha_1 < 1$, $-1 < \alpha_2 < 1$, and $-1 < \alpha_1 + \alpha_2 < 1$. Since (5.44) is modeled within a GARCH context, it is called a GARCHS$(1, 1, 1)$ (i.e., GARCH with skewness).

Using a very similar approach, Brooks, Burke, and Persand (2002) adopt the following dynamics for kurtosis

$$\kappa_t = \beta_0 + \beta_1 \kappa_{t-1} + \beta_2 z_{t-1}^4, \tag{5.45}$$

with the following constraints on the parameters: $-1 < \beta_1 < 1$, $-1 < \beta_2 < 1$, and $-1 < \beta_1 + \beta_2 < 1$. This model is named by the authors a GARCHK$(1, 1, 1)$ (GARCH with kurtosis).

These approaches present the advantage that higher moments are directly estimated in the course of the optimization. Therefore, higher moments are modeled in a way that is similar to the variance in the GARCH framework. However, this does not come without a price. In order to estimate the model parameters, we have to derive the relationship between the higher moments s_t and κ_t and the (time-varying) shape parameters η_t. This relationship is very likely to be non-linear, except in very special cases such as Gram-Charlier expansions or entropy densities. In the context of the Student t distribution, which is adopted by Harvey and Siddique (1999) as well as by Brooks, Burke, and Persand (2005), the relationship between the higher moments and the estimated parameters is highly non-linear. Therefore, modeling higher moments directly requires, at each date and each step of the optimization, to solve the non-linear problem that links parameters and moments. This renders the estimation computationally very intensive.

Another drawback of this approach is that the constraints, which should be imposed on the higher moments, to ensure positivity of the distribution, are not clearly established. As seen before, skewness and kurtosis cannot be freely estimated. In the specifications above (equations (5.44) and (5.45)), the only constraints on the dynamics of skewness and kurtosis are to ensure stationarity of the process, not positivity of the distribution. It turns out that the constraints, presented in Section 5.2, to ensure that the density is defined, cannot be easily imposed in this context. As shown in the context of the skewed Student t distribution (Section 5.2.3), the constraints on parameters λ and ν to ensure that the distribution is definite are easy to establish. In contrast, the constraints on s and κ are not analytically known. A similar result holds for the Gram-Charlier series expansion.

In a slightly different context, Jondeau and Rockinger (2003a) have shown that an autoregressive structure such as presented above suffers from a severe

drawback. Consider the postulated dynamics (5.44) for conditional skewness. For data with sufficient variability, as the sample increases, the model is likely to degenerate to a solution where $\alpha_2 = 0$. To understand why this is so, we may write, if t is sufficiently large and $|\alpha_1| < 1$, that

$$s_t = \alpha_0/(1-\alpha_1) + \alpha_2 \sum_{s=0}^{\infty} \alpha_1^s z_{t-1-s}^3.$$

From this expression, we see that the mean of s_t is $\alpha_0/(1-\alpha_1)$ and its variance is $\alpha_2^2 V\left[\sum_{s=0}^{\infty} \alpha_1^s z_{t-1-s}^3\right]$. This shows that the restriction $\underline{s} < s_t < \overline{s}$ will be satisfied for $\alpha_2 = 0$ only. Otherwise, there is a non-zero probability that the constraint may be violated for some observations.[19] Consequently, an autoregressive process for skewness or kurtosis is very unlikely to be a fruitful model.

In Rockinger and Jondeau (2002), the dynamics of skewness and kurtosis are also directly modeled in the context of the entropy distribution. To avoid the problem described above, a very simple specification is adopted, that does not resort to an auto-regressive process

$$\begin{aligned} s_t &= a_0 + a_1 z_{t-1}, \\ \kappa_t &= b_0 + b_1 \left| z_{t-1} \right|. \end{aligned}$$

These dynamics are estimated while numerically imposing that skewness and kurtosis be inside the domain of definition described in section 5.2.6.

5.5.3 Modeling the parameters of the distribution

In his 1994 paper, Hansen extends the model with skewed Student t distribution (cf. section 5.2.3) to the case where parameters associated with the distribution (the degree of freedom ν and the asymmetry parameter λ) are also time varying. This yields the concept of AutoRegressive Conditional Density (ARCD).

A similar approach has been adopted by Jondeau and Rockinger (2003a) and Lambert and Laurent (2002). Jondeau and Rockinger (2003a) have discussed several possible specifications for the dynamics of the degree-of-freedom parameter ν_t and the asymmetry parameter λ_t. In Table 5.5, we display these specifications. Many other specifications could be designed, involving further lags or less linear relations. We emphasize these specifications, because they highlight some difficulties that may be encountered. We assume that the coefficients are such that stationarity is guaranteed.

Model M1 specifies directly ν_t and λ_t as functions of past realizations. The advantage of this specification is that no further non-linear map is required to

[19] Such a problem does not occur in the GARCH context, because the volatility at time t, σ_t, is only required to be positive. This constraint is always fulfilled because the ε_t terms are squared.

Table 5.5. *Possible specifications of the model*

$$\text{Model M1:} \begin{cases} \nu_t = a_1 + b_1^+ z_{t-1}^+ + b_1^- z_{t-1}^-, \\ \lambda_t = a_2 + b_2^+ z_{t-1}^+ + b_2^- z_{t-1}^-. \end{cases}$$

$$\text{Model M2:} \begin{cases} \tilde{\nu}_t = a_1 + b_1^+ z_{t-1}^+ + b_1^- z_{t-1}^-, \\ \tilde{\lambda}_t = a_2 + b_2^+ z_{t-1}^+ + b_2^- z_{t-1}^-, \\ \nu_t = \mathcal{L}_{[2,30[}(\tilde{\nu}_t), \qquad \lambda_t = \mathcal{L}_{]-1,1[}(\tilde{\lambda}_t). \end{cases}$$

$$\text{Model M3:} \begin{cases} \nu_t = a_1 + b_1^+ z_{t-1}^+ + b_1^- z_{t-1}^- + c_1 \nu_{t-1}, \\ \lambda_t = a_2 + b_2^+ z_{t-1}^+ + b_2^- z_{t-1}^- + c_2 \lambda_{t-1}. \end{cases}$$

$$\text{Model M4:} \begin{cases} \tilde{\nu}_t = a_1 + b_1^+ z_{t-1}^+ + b_1^- z_{t-1}^- + c_1 \nu_{t-1}, \\ \tilde{\lambda}_t = a_2 + b_2^+ z_{t-1}^+ + b_2^- z_{t-1}^- + c_2 \tilde{\lambda}_{t-1}, \\ \nu_t = \mathcal{L}_{[2,30[}(\tilde{\nu}_t), \qquad \lambda_t = \mathcal{L}_{]-1,1[}(\tilde{\lambda}_t). \end{cases}$$

$$\text{Model M5:} \begin{cases} \tilde{\mu}_{3t} = a_1 + b_1 z_{t-1}^3, \\ \tilde{\mu}_{4t} = a_2 + b_2 z_{t-1}^4, \\ (\mu_{3t}, \mu_{4t}) = G(\tilde{\mu}_{3t}, \tilde{\mu}_{4t}), \\ (\nu_t, \lambda_t) = F^{-1}(\mu_{3t}, \mu_{4t}). \end{cases}$$

$$\text{Model M6:} \begin{cases} \tilde{\mu}_{3t} = a_1 + b_1 z_{t-1}^3 + c_1 \tilde{\mu}_{3t-1}, \\ \tilde{\mu}_{4t} = a_2 + b_2 z_{t-1}^4 + c_2 \tilde{\mu}_{4t-1}, \\ (\mu_{3t}, \mu_{4t}) = G(\tilde{\mu}_{3t}, \tilde{\mu}_{4t}), \\ (\nu_t, \lambda_t) = F^{-1}(\mu_{3t}, \mu_{4t}). \end{cases}$$

obtain a description of the parameters. A drawback of this specification is that its estimation is cumbersome because the constraints $2 < \nu_t$ and $-1 < \lambda_t < 1$ must be numerically imposed. Furthermore, nothing guarantees that ν_t and λ_t will be well defined out of sample. Ad hoc techniques, such as truncation at the boundaries, could be devised for forecasting purposes.

Model M2 specifies a dynamic for unconstrained $\tilde{\nu}_t$ and $\tilde{\lambda}_t$. These unrestricted parameters get mapped into the authorized domain $\mathcal{D}$ via the logistic map $\mathcal{L}_{[a,b[}(x) = a + \exp(x) / (1 + \exp(x))(b - a)$. Many of the drawbacks of model M1 disappear. However, one consequence is that the impact of extreme realizations gets dampened because the logistic map tends to flatten the response of variables located in its tails.

Model M3 specifies parameters ν_t and λ_t as an autoregressive structure. As discussed above, such a specification suffers from a severe drawback, because, for data with sufficient variability, as the sample increases, the model is likely to degenerate to a solution where $b_2 = 0$.

Model M4 is similar to M3, yet it uses a non-linear map to constrain the parameters to $\mathcal{D}$. The model could still be estimated, but some care is needed in the interpretation of the estimates. For instance, if one estimates M4 and finds that b_1 is not statistically different from 0, then the model reduces to

$$\tilde{\nu}_t = a_1 + c_1 \tilde{\nu}_{t-1}.$$

At this stage, c_1 may be statistically significant. This may lead to the conclusion that there is persistence in the $\tilde{\nu}_t$. Such a conclusion would be erroneous, however. Indeed, if actual observations y_{t-1} do not matter, then, starting from some initial value $\tilde{\nu}_0$, the series of $\tilde{\nu}_t$ will quickly converge to its stationary level given by

$$\tilde{\nu}^* = a_1/(1 - c_1).$$

In other words, the model where we would have estimated $\tilde{\nu}_t = \tilde{\nu}^*$ (with $b_1 = c_1 = 0$) could not be distinguished from the one obtained earlier. This implies that there exists an entire class of parameters (a_1, c_1), all satisfying $(1 - c_1)\tilde{\nu}^* = a_1$, for which the model's characteristics are indistinguishable. The algorithm converges to one solution at random. To avoid this type of spurious finding, it is recommended to estimate M2 before M4 and to verify that past observations affect $\tilde{\nu}_t$ or $\tilde{\lambda}_t$. In no way should we trust in an estimation where c_1 or c_2 is statistically significant, yet, the parameters on the lagged innovations are not statistically significant. A further diagnostic to detect this behavior consists in changing the value of $\tilde{\nu}_1$ or of the initial value of the parameters in the numerical estimation. If the algorithm converges to significantly different values, then we should be careful about the estimated parameters.

In specification M5, the third and fourth non-central moments, s_t and κ_t, get specified using actual observations. For the model to be well defined, it must be that s_t and κ_t belong to the domain $\mathcal{E}$. This implies a potentially highly non-linear map G that maps some unrestricted $\tilde{s}_t$ and $\tilde{\kappa}_t$ into $\mathcal{E}$. Furthermore, in order to obtain λ_t and ν_t from s_t and κ_t, it is necessary to invert a highly non-linear map that we call F in the table. Even though such an inversion could be done in theory, it will lead to a slow algorithm and also to a rather unstable estimation because the analytic computation of gradients may not be feasible. With such a specification, it is implicitly assumed that skewness and kurtosis are finite at each point of time. This observation is at odds with results from extreme value theory.

Model M6 presents the same difficulties as M5 with the added complication already discussed for model M3, in that one may find spurious dependence of skewness due to a lack of significant b_1 or b_2 estimates.